Journal of Cleaner Production 405 (2023) 137019

Contents lists available at ScienceDirect

Journal of Cleaner Production

journal homepage: www.elsevier.com/locate/jclepro

Predicting quality parameters of wastewater treatment plants using artificial intelligence techniques

Ehsan Aghdam [a], Saeed Reza Mohandes [b,*], Patrick Manu [b], Clara Cheung [b], Akilu Yunusa-Kaltungo [b], Tarek Zayed [a]

[a] Department of Building and Real Estate (BRE), Faculty of Construction and Environment (FCE), The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[b] Department of Mechanical, Aerospace and Civil Engineering, School of Engineering, The University of Manchester, M13 9PL, Manchester, United Kingdom

ARTICLE INFO

Handling editor: Jin-Kuk Kim

Keywords:
Gene expression programming
AI models
Wastewater quality parameters
Artificial neural networks
Monte Carlo simulation

ABSTRACT

Estimating wastewater treatment plants' (WWTPs) influent parameters such as 5-day biological oxygen demand (BOD_5) and chemical oxygen demand (COD) is vital for optimizing electricity and energy consumption. Against this backdrop, the existing body of knowledge is bereft of a study employing Artificial Intelligence-based techniques for the prediction of BOD_5 and COD. Thus, in this study, Gene expression programming (GEP), multilayer perception neural networks, multi-linear regression, k-nearest neighbors, gradient boosting, and regression trees -based models were trained for predicting BOD_5 and COD, using monthly data collected from the inflow of 7 WWTPs over a three-year period in Hong Kong. Based on different statistical parameters, GEP provides more accurate estimations, with R^2 values of 0.784 and 0.861 for BOD_5 and COD respectively. Furthermore, results of sensitivity analysis undertaken by monte Carlo simulation revealed that both BOD_5 and COD were mostly affected by concentrations of total suspended solids, and a 10% increase in the value of TSS resulted in a 7.94% and 7.92% increase in the values of BOD_5 and COD, respectively. It is seen that the GEP modeling results complied with the fundamental chemistry of the wastewater quality parameters and can be further applied on other sewage sources such as industrial sewage and leachate. The promising results obtained pave the way for forecasting the operational parameters during sludge processing, leading to an extensive energy savings during the wastewater treatment processes.

1. Introduction

Wastewater treatment plants (WWTPs) are an essential part of the societies' infrastructure, which play a significant role in improving the quality of life by treating the municipal and industrial sewage and discharging the treated effluent into the receiving water bodies. On the other hand, WWTPs are one of the major sources of energy consumption and the production of greenhouse gases (Wang et al., 2016). The performance of the WWTPs is affected by several parameters, such as the variations of the influent and shocking loads (Ansari et al., 2020). Acquiring the influent characteristics is essential for designing WWTPs and adjusting the operational parameters such as the amount of aeration, which consumes a huge amount of electricity in the WWTPs (Wang et al., 2022). The major influent parameters affecting the performance of the WWTP at different stages are 5-day biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), ammoniacal nitrogen (NH_3), organic nitrogen content (OrgN), organic phosphorous (OrgP) and inorganic phosphorous (InorgP) content of the influent (Henze et al., 2008). Due to the high fluctuations of the WWTPs' influent and their impact on the performance of the treatment processes, predicting the influent characteristics or obtaining their real-time values play a significant role in optimizing the operation of WWTP (Abouzari et al., 2021; Asami et al., 2021; Cheng et al., 2018; Kim et al., 2016; Li and Vanrolleghem, 2022; Wodecka et al., 2022). For example, BOD_5 is representative of the biodegradable content of the wastewater and determines the amount of aeration, which is the most energy-consuming step in the WWTPs (Wang et al., 2019; Luo et al., 2019). Therefore, the overestimation of BOD_5 leads to higher energy consumption and cost. On the other hand, COD is indicative of the chemical compounds which are non-biodegradable and detrimental to the micro-organisms in the wastewater. Moreover, the phosphorous and nitrogenous

* Corresponding author.
E-mail addresses: ehsan.aghdam@polyu.edu.hk (E. Aghdam), saeedreza.mohandes@manchester.ac.uk (S.R. Mohandes), patrick.manu@manchester.ac.uk (P. Manu), clara.cheung@manchester.ac.uk (C. Cheung), akilu.kaltungo@manchester.ac.uk (A. Yunusa-Kaltungo), tarek.zayed@polyu.edu.hk (T. Zayed).

https://doi.org/10.1016/j.jclepro.2023.137019
Received 10 October 2022; Received in revised form 14 March 2023; Accepted 28 March 2023
Available online 31 March 2023

Abbreviations

AI	Artificial intelligence
ARIMA	Autoregressive Integrated Moving Average
BOD_5	5-day biochemical oxygen demand
COD	Chemical oxygen demand
ELM	Extreme learning machine
GB	Gradient boosting
GEP	Gene expression programming
GHGs	Greenhouse gases
InorgP	Inorganic phosphorous compounds
kNN	k-Nearest Neighbor
MARS	multi-variate adaptive regression spline
MCS	Monte Carlo simulation
Multilayer perception neural networks	MLP-NN
NN	Neural Network
OP	Orthophosphates
ORELM	Outlier robust extreme learning machine
OrgN	Organic nitrogenous compounds
OrgP	Organic phosphorous compounds
PDF	Probability density function
RF	Random forest
RT	Regression tree
TKN	Total Kjeldahl nitrogen
TP	Total phosphorous
TDS	Total dissolved solids
TSS	Total suspended solids
WWTP	Wastewater treatment plant

compounds may interfere with the biological processes during the wastewater treatment processes, considering different treatment methods (Metcalf, 2003).

Measuring all of the influent parameters is time-consuming and requires conducting complicated tests and consuming hazardous materials as fully explained in the standard methods for the examination of water and wastewater (Baird et al., 2017). To address this issue, some electrical sensors have recently been developed to provide the real-time values of the influent's quality parameters (Haimi et al., 2015; Olsson, 2012; Olsson et al., 2014; Vanrolleghem and Lee, 2003). However, some important parameters, such as BOD_5 and COD is difficult and expensive to obtain by sensors (Jouanneau et al., 2014; Kim et al., 2020), which necessitates the development of mathematical predictive models of estimating their values using historical data (Ching et al., 2022; Foschi et al., 2021). The major advantage of using mathematical and statistical methods to predict wastewater characteristics and WWTPs' performance is to fully understand the physicochemical and biological processes behind them. On the other hand, integrating the models with chemical reactions provides insights into the interpretations of the results. Predicting the quality parameters of wastewater, especially BOD_5 and COD is of great importance due to: (1) high cost and time consumption of measuring them; (2) providing the opportunity for the decision-makers to adjust the operational parameters to optimize the energy consumption; and (3) spotting any abnormal variations in the WWTP's influent due to intrusion of any discharge in the upstream.

Table S1 presents some examples of AI methods employed in the literature to predict the parameters of WWTPs. As can be seen, artificial neural networks (ANNs), random forest (RF), multi-linear regression (MLR) and fuzzy models are the most common methods to predict wastewater quality parameters. However, these methods suffer from some drawbacks. For instance, the ANN models are opaque and difficult to interpret (Dürrenmatt and Gujer, 2012).

Among AI methods, gene expression programming (GEP) has recently gained interest in the environmental engineering and wastewater treatment area due to its effectiveness and accuracy over other machine learning methods in predicting complex parameters. Noman et al. (2022) and Yaqub et al. (2022a) developed GEP-based models to respectively predict the inactivation of several types of bacteria in wastewater and the performance of micellar-enhanced ultrafiltration for removal of mercury and arsenic from water. Shishegaran et al. (2020) investigated process optimization of the surfactants' removal by ultrafiltration using MLR, multiple Ln-equation regression (MLnER), and GEP, while Seckin et al. (2011) evaluated the prediction ability of GEP and Stover–Kincannon model in estimating methane yield and effluent substrate of upflow anaerobic filters. Shah et al. (2021a) applied GEP and ANN to predict the surface water quality.

Besides estimating the target parameters, assessing its sensitivity to the variations of other parameters is of great importance because it provides the most influential parameters. The Monte Carlo Simulation (MCS) technique provides insights into the uncertainty of estimating a parameter by a developed model using other uncertain surrogate parameters. Moreover, the MCS generates several data, which can be used to analyze the influence of the input parameters on the target parameter by sensitivity analysis. Quantifying the model output resulting from the lack of knowledge about input data provides essential information for the decision-makers to take appropriate actions in times of severe fluctuations in the input parameters. Using MC analysis to assess the sensitivity of the target parameters to the variations of input parameters embodies several advantages such as: (1) allocating the uncertainties of the unknown parameters to the uncertainties in the input parameters; (2) determining the most influential parameters on the variations of the target parameters; and (3) scrutinizing the interactions between wastewater quality parameters.

Though some ML-based techniques have been used in the literature, there has been dearth of a study presenting a mathematical representation for calculating BOD5 and COD in the existing body of knowledge. The lack of such an inclusive formula demands a lot of efforts, resources, and expenses to be incurred at undertaking the related experiments for measuring BOD5 and COD in laboratories. This knowledge gap along with the mentioned need has given impetus to the authors of this study for coming up with a lucid mathematical formula for estimating the said parameters.

In light of the points mentioned above, there are three major knowledge gaps to be addressed as follows.

- Lack of robust mathematical expressions to correlate the difficult-to-measure parameters such as BOD_5 and COD to easy-to-measure parameters such as TSS, ammonia, OrgN, OrgP and InorgP content of wastewater, which can be obtained by online sensors,
- Sparse research in investigating the most influential parameters on predicting BOD_5 and COD and
- Lack of knowledge in evaluating the uncertainty of predicting BOD_5 and COD values.

The integration of GEP modelling and MCS to study wastewater characteristics provides a strong tool for environmental engineers to monitor the biological and chemical properties of wastewater in real-time, which then enables them to optimize the operational parameters. Thus, the objectives of this study are to: (1) develop models by GEP, MLP-NN, kNN, GB, RT and RF to predict BOD_5 and COD in the influent of the WWTPs according to the surrogate parameters of TSS, NH_3, OrgN, InorgP and OrgP content of the wastewater influent samples, (2) compare different AI models regarding the accuracy of predicting BOD_5 and COD, (3) evaluate the sensitivity of BOD_5 and COD to each of the input parameters through MC, and (4) interpret the modelling results according to the chemistry of the wastewater constituents.

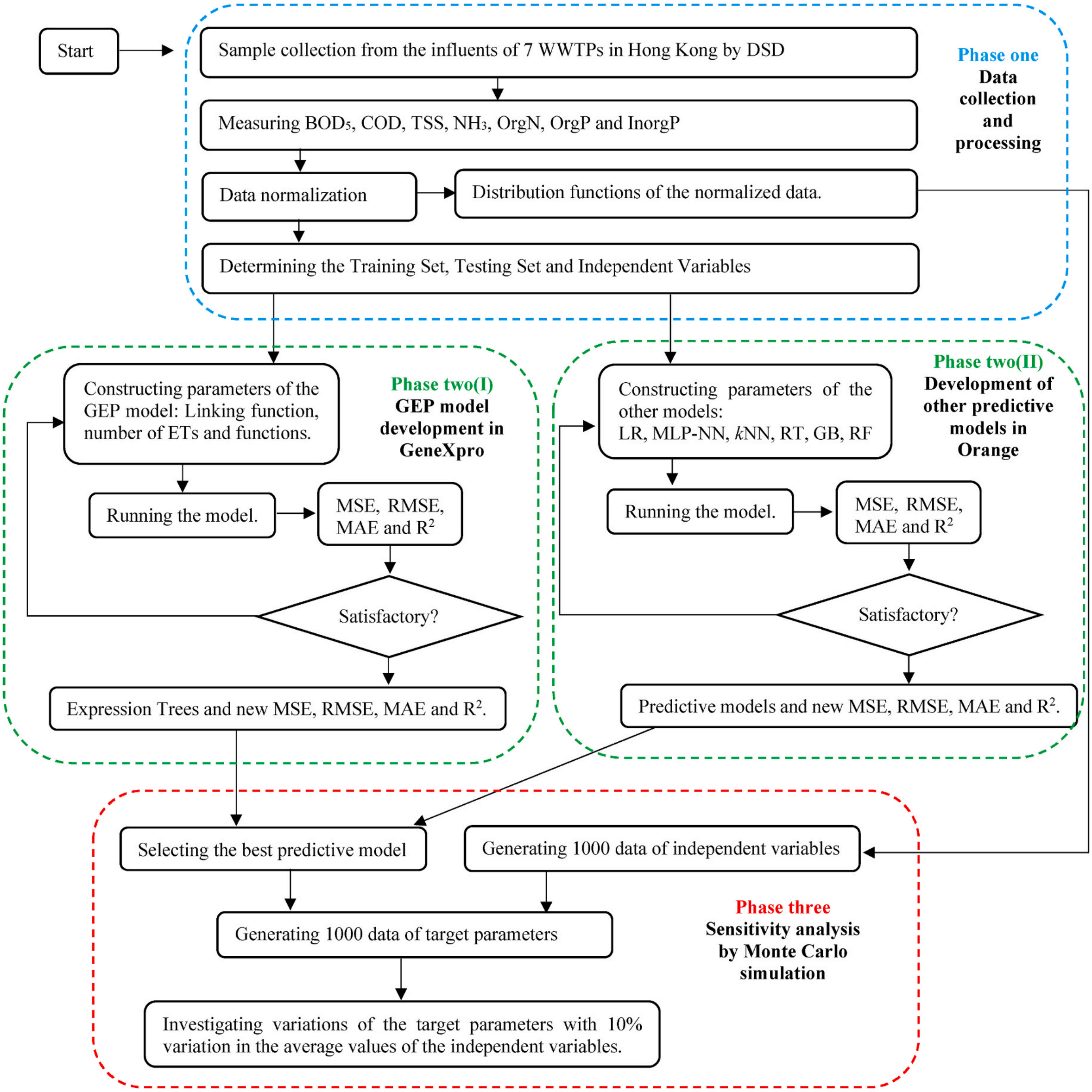
The findings of this study provide the environmental decision-makers and policymakers with solutions to deal with the pollution produced from wastewater associated with sewer pipelines, which in turn, culminates in improving public health and taking a giant leap towards realizing a smart and sustainable built environment. In addition, the obtained results can be used to adjust the chemical dosage as well as estimate the aeration demand of the consecutive wastewater treatment stages, with a view to minimizing the energy cost and production of greenhouse gases.

2. Methodology

The general framework of this research is presented in Fig. 1, which encompasses the following five major steps: (1) data collection; (2) normalization of data; (3) developing the different AI models with the pre-processed data to predict BOD_5 and COD; (4) validation of the models; and (5) assessing the sensitivity of the target parameters toward the predictors.

2.1. Phase one: Data collection and preparation

Samples were collected from the influents of 7 WWTPs in Hong Kong, namely Sha Tin Sewage Treatment Works (ST-STW), Sai Kung Sewage Treatment Works (SK-STW), Sham Tseng Sewage Treatment Works (SmT-STW), Yuen Long Sewage Treatment Works (YL-STW), Tai Po Sewage Treatment Works (TP-STW), Shek Wu Hui Sewage Treatment Works (SWH-STW) and Stonecutters Island Sewage Treatment Works (SCISTW) (HKDSD, 2018). The locations of these WWTPs are shown in Fig. 2. Their design flow and main treatment processes are listed in Table 1 (HKDSD, 2018). During the study period (i.e. from Jan 2018 to Dec 2020), composite samples of the WWTPs' influents were grabbed monthly from the manholes near the treatment plants. The quality parameters of the samples, including BOD_5, COD, TSS, NH_3, OrgN, OrgP and InorgP were measured according to the standard methods mentioned in Table 2. Notably, they were characterized by the Hong Kong Drainage Service Department (DSD). To better understand the data, the statistical parameters of the data and their probability density function (PDF) are determined with Origin Pro 2021 software, and accordingly, the outliers were removed from the initial data. It is



Start
Sample collection from the influents of 7 WWTPs in Hong Kong by DSD
Measuring BOD5, COD, TSS, NH3, OrgN, OrgP and InorgP
Data normalization
Distribution functions of the normalized data.
Determining the Training Set, Testing Set and Independent Variables
Phase one
Data collection and processing
Constructing parameters of the GEP model: Linking function, number of ETs and functions.
Phase two(I)
GEP model development in GeneXpro
Running the model.
MSE, RMSE, MAE and R2
Satisfactory?
Expression Trees and new MSE, RMSE, MAE and R2.
Constructing parameters of the other models: LR, MLP-NN, kNN, RT, GB, RF
Phase two(II)
Development of other predictive models in Orange
Running the model.
MSE, RMSE, MAE and R2
Satisfactory?
Predictive models and new MSE, RMSE, MAE and R2.
Selecting the best predictive model
Generating 1000 data of independent variables
Generating 1000 data of target parameters
Phase three
Sensitivity analysis by Monte Carlo simulation
Investigating variations of the target parameters with 10% variation in the average values of the independent variables.

Fig. 1. The research framework.

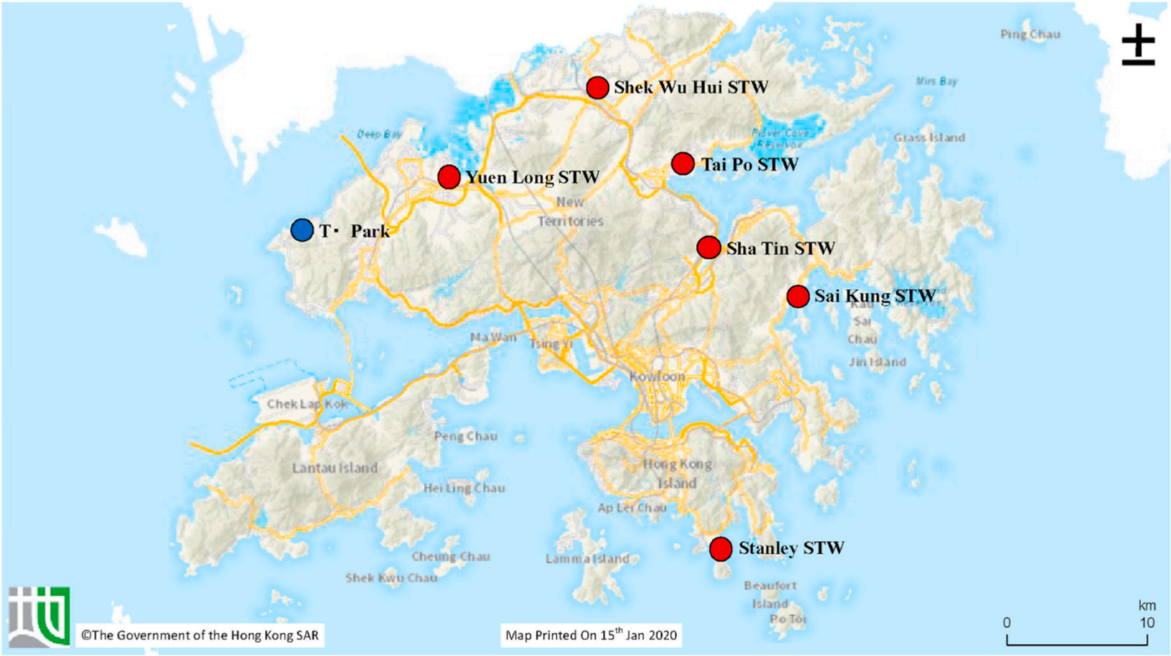


Fig. 2. The locations of 7 WWTPs, from which the data was collected (Map source: www.map.gov.hk).

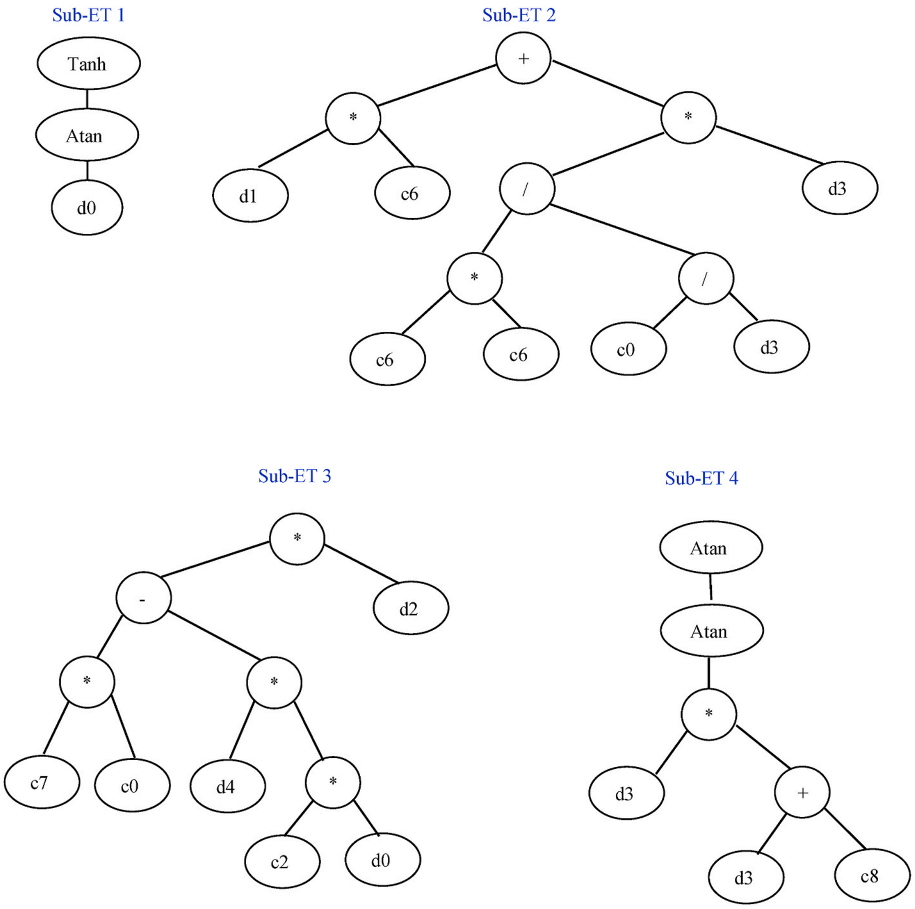


Fig. 3. GEP expression tree for predicting COD from TSS, NH_3, OrgN, OrgP and InorgP.(Note: d_0: concentration of TSS, d_1: concentration of NH_3, d_2: concentration of OrgN, d_3: concentration of OrgP, d_0: concentration of InorgP. The numbers of c_i are constants generated by GeneXpro.).

Table 1
The basic information regarding the WWTPs in Hong Kong.

WWTP	Designed flow rate (m^3/d)	Actual flow rate (m^3/d)
Sai Kung Sewage Treatment Works (SK-STW)	22000	8000
Sham Tseng (SmT-STW)	5600	–
Yuen Long Sewage Treatment Works (YL-STW)	70000	16,000
Tai Po Sewage Treatment Works (TP-STW)	120000	95,000
Shek Wu Hui Sewage Treatment Works (SWH-STW)	93000	81,000
Sha Tin Sewage Treatment Works (ST-STW)	340000	250,000
Stonecutters Island Sewage Treatment Works (SCISTW)	2,450,000	1,900,000

Table 2
The wastewater parameters and the analytical methods of them.

Parameters	Analytic methods	units
BOD_5	5210 B.	mg/L as O_2
COD	5220 B.	mg/L as O_2
TSS	2540 D.	mg/L
NH_3	4500-NH_3 F.	mg/L as N
OrgN	4500-N_{org} C.	mg/L as N
OrgP	4500-P.	mg/L as P
InorgP	4500-P.	mg/L as P

noteworthy that all the wastewater treatment plants were municipal wastewater treatment plants, and the initial conditions and operational conditions were similar. On the other hand, since the data were collected from the influent of the treatment plants before the treatment processes, the “design conditions” of them did not affect the data analysis. Moreover, since all treatment plants were in Hong Kong, the geographical conditions of all of them were similar and may not affect the analysis.

Before developing the models for BOD_5 and COD prediction, the input data were normalized to prevent overfitting and minimize the errors associated with the large differences between the covered ranges of the different parameters. To this purpose, all data are normalized within the range of [0, 1] according to Eq. (1) (Tijani and Zayed, 2022):

$$a = \frac{(x - x_{min})}{(x_{max} - x_{min})} \tag{1}$$

Where a is the normalized value of the input parameter, x, x_{max} and x_{min} are the value of the respective variable, the maximum value of all the inputs for a particular variable, and the minimum value of all the inputs for a particular variable, respectively.

With the above in mind, BOD5 is indicative of the oxygen requirement by microorganisms to stabilize the organic matter in the wastewater. It is one of the major wastewater parameters due to its application in: (1) determining the required oxygen in the aeration tanks; (2) designing the wastewater treatment facilities; (3) evaluating the performance of the wastewater treatment processes; and (4) assessing the quality of the discharged wastewater (Eddy et al., 2014; Gerardi, 2015). However, measuring BOD5 is cumbersome, time-consuming and cannot be obtained in real-time. Therefore, developing a correlation between BOD5 and easy-to-find parameters simplifies decision making regarding operational works in the WWTPs. On the other hand, COD is the amount of oxygen for oxidizing organic compounds in the presence of dichromate in acidic environment. COD value is always higher than BOD5 due to: (1) presence of non-biodegradable compounds; (2) presence of inorganic compounds; which can be oxidized in the presence of dichromate in the COD measurement test; and (3) presence of some compounds toxic to the microorganisms in the BOD5 measurement test. Therefore, predicting the amount of COD and COD/BOD5 ratio is advantageous for the decision makers in terms of adjusting the operational parameters of a WWTP (Jouanneau et al., 2014). Moreover, COD/BOD5 ratio indicates the amount of non-biodegradable compounds in the wastewater inflow, and consequently the likelihood of industrial sewage intrusion into the municipal wastewater network (Eddy et al., 2014; Henze et al., 2008). Considering the mentioned reasonings, it can be inferred that there are potential relationships among BOD5, COD, TSS, NH3, OrgN, OrgP and InorgP.

2.2. Phase two (I): Gene expression programming method

GEP method, one of the newly developed supervised machine learning techniques, is a combination of genetic algorithm (GA) and genetic programming (GP). The basic principle of GEP, as proposed by Ferreira (2002), resembles that of human’s genetic system. In GEP modelling, the observed and modelled data are considered as chromosomes, which comprise of several genes embodying the protein coding regions. The main purpose of GEP is to find a mathematical function, which correlates several independent variables to a target parameter according to the actual observed data. The relationship embodies several mathematical functions, including the four principal operations, trigonometric functions and their inverse and hyperbolic counterparts, logarithmic, power, inverse, root and so on (Noman et al., 2022). Developing a GEP model requires five basic elements including: (1) set of mathematical functions to be used in building a correlation between the independent variables and the target parameter; (2) a set of terminals including the independent variables and constants; (3) fitness function; (4) control parameters; and (5) conditions for stopping the program (Kayadelen et al., 2009). To grasp more understanding in GEP utilization, readers are referred to the following references: (Tijani and Zayed, 2022), (Noman et al., 2022), and (Islam et al., 2022).

In the current study, the GEP algorithm was employed to predict the BOD_5 and COD of the WWTPs’ influent as a function of concentrations of five independent variables of TSS (mg/L), NH_3 (mg/L as N),OrgN (mg/L as N), InorgP (mg/L as P) and OrgP (mg/L as P). GeneXpro Tools 5.0 software was employed to develop the GEP model. 70% of the total data was used for training the model and the remaining was used for model validation.

A wide range of mathematical functions including $+, -, \times, \div, \mathit{Tanh}, \mathit{Arctan}, \mathit{exp}, \mathit{power}, \mathit{logharithm}, \mathit{Ln}$ were selected as the function set. However, the GEP may not use all of them. GEP uses Roulette wheel to select the data and simultaneously reproduce them by genetic operators to find the most appropriate function for predicting the target parameter. The inappropriate data are eliminated, and the appropriate data are kept and eventually conveyed to the next generation, just as is the case with genetic mutation in the human genome. The goal of this mutation in GEP modelling is to randomly optimize the chromosomes (Shishegaran et al., 2020).

2.3. Model performance evaluation

Five statistical metrics, Root Mean Square Error (RMSE), Mean Square Error (MSE), coefficient of determination (R^2), RAE and MRRE as expressed in Eq. (2)–6 respectively, were used to evaluate the performance of the developed model. RMSE implies the average error between the predicted and observed values as expressed in Eq. (2). MSE as presented in Eq. (3), measures the average difference between the modelled values and the values acquired from the field. It should be noted that RMSE is the square root of MSE and both parameters are within the range of [0, 1] for the normalized data. The closer these values are to 0, the more accurate the model is (Shah et al., 2021b). On the other hand, the R^2 given in Eq. (4), expresses the degree of fitness between the predicted and observed values. Higher values of R^2 closer to 1, indicates higher accuracy of the model. Relative absolute error (RAE) and mean root relative error (MRRE) are given by Eqs. (5) and (6), respectively.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_p - y_m\right)^2} \quad (2)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_p - y_m\right)^2 \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_p - y_m\right)^2}{\sum_{i=1}^{n}\left(y_p - \overline{y_m}\right)^2} \quad (4)$$

$$RAE = \frac{\sum\left|\left(y_p - y_m\right|\right.}{\sum\left|\left(y_p - \frac{1}{n}\sum y_m\right)\right|} \quad (5)$$

where n is the number of the inputs, y_m is the measured or observed values, y_p is the predicted or modelled values, and $\overline{y_m}$ is the average of the measured values of an input parameter.

2.4. Phase two (II): Comparative analysis

Once the results of GEP have been obtained for the prediction of BOD_5 and COD, an exhaustive comparative analysis against the other ML-based techniques, LR, NN, *k*NN, RT, GB and RF is undertaken. This is of paramount importance, since the accuracy and performance of various ML-based techniques need to be gauged with respect to each other. LR modelling assumes a linear relationship between the target parameter (BOD_5 or COD) and the predictors, TSS, NH_3, OrgN, InorgP and OrgP. On the other hand, MLP-NN includes five independent variables neurons (TSS, NH_3, OrgN, InorgP and OrgP), each linked to the output variables(BOD_5 or COD). In the *k*NN method, "*k*" numbers of the input dataset such as TSS_i, $NH_{3(i)}$, $OrgN_i$, $InorgP_i$, $OrgP_i$ from the training dataset were determined which are the closest ones to another input data in the validation dataset such as TSS_v, $NH_{3(v)}$, $OrgN_v$, $InorgP_v$, and $OrgP_v$. The predicted value of the target parameter corresponding to TSS_v, $NH_{3(v)}$, $OrgN_v$, $InorgP_v$, and $OrgP_v$ is the average of the target parameters corresponding to TSS_i, $NH_{3(i)}$, $OrgN_i$, $InorgP_i$, and $OrgP_i$, in which "i" is in the range of [1, k] (R. Wang et al., 2022). It should be noted that although kNN and NN methods consider the nonlinearity of the correlations, they do not provide any specific equation. In the RT algorithm, for predicting the target parameter, the whole dataset is considered as root of a tree and split into subsets. Meanwhile, a multivariate linear regression is used to predict the target in each subset (Breiman et al., 2017; Granata et al., 2017). In the RF, several trees (forest) were built in parallel to predict the target parameter and the mean value of the predictions is recognized as the predicted value (Bunce and Graham, 2019; Dürrenmatt and Gujer, 2012; Lakshmanaprabu et al., 2019; Mateo Pérez et al., 2021; Sharafati et al., 2020). In the GB algorithm, trees were generated consecutively considering the error between the target value and the new value predicted by the new tree. The error of the estimation decreased with increasing number of trees.

2.5. Phase three: Monte Carlo simulation (MCS)

The MCS was used to conduct the sensitivity analysis of the best predictive model, which was found according to the statistical metrics of the models. Firstly, the probability density functions (PDFs) of the independent variables, TSS, NH_3, OrgN, InorgP and OrgP were defined according to the historical data collected by HK DSD. It should be noted that the normalized data were used for conducting MCS since the models were developed based on the normalized data. The goodness of each distribution function was evaluated by Kolmogorov-Smirnov test (KS test) including the p-value parameter. Higher than 0.05 of p-value indicates that the assumed distribution function can represent the distribution of the observed data. For MCS, 1000 random values of TSS, NH_3, OrgN, InorgP and OrgP were generated based on the parameters of their distribution function. Then, the corresponding BOD_5 and COD values for each set were found according to the model-generated equation. Afterwards, the sensitivity of BOD_5 and COD to ±10%, ±20% and ±50 change in one of the parameters was found, while the other four parameters were kept constant.

3. Results and discussion

3.1. Wastewater quality parameters and statistical metrics

Based on the three years of historical data, which were collected monthly from the inlets of seven WWTPs all over Hong Kong, 11 individual variables were obtained. A statistical summary of these 11 influent variables is presented in Table 3. In addition, the dataset of each parameter was fitted to a probability distribution function (PDF). It was observed that NH_3, OrgP and InorgP followed log-normal distribution, but the other variables do not follow any of the known distribution functions. These results showed significant differences in the distributions of the concentrations of the influent's parameters, and this necessitates normalization of the data before applying GEP. On the other hand, all dataets possessed high skewness, indicating that most data values are concentrated within a certain range. The presence of high values of contaminants in some cases may be due to the intrusion of industrial wastewater into the municipal sewage networks. Despite their rare occurrence, they were not removed from the dataset to account for their effect on the developed model. To remove the effect of the large range of some of the data on the PDFs, the distribution of the normalized data was obtained. Moreover, the distribution function of the normalized data will form the basis of the MCS-based sensitivity analysis. All parameters except OrgN fit the log-normal distribution. In another research, Oliveira et al. (2012) found the log-normal distribution to be the most appropriate for describing the distribution of the wastewater parameters. On the other hand, InorgP and NH_3 values fitted to the Gamma distribution better. The best fits of the normalized input data and the parameters of the distribution functions are tabulated in Table 4.

BOD_5 and COD were selected as the target parameters, since they can be used as surrogate parameters to estimate the energy consumption, as well as sludge production of a WWTP (Benedict et al., 1979.). On the other hand, measuring BOD_5 is tedious and prone to several errors, which is even made more challenging by the scarcity of real-time data (Basant et al., 2010). Moreover, COD measurement requires skillful laboratory workers and precautionary measurements (Baird et al., 2017). Among the other parameters, TSS, OrgN, NH_3, InorgP and OrgP were selected as the predictors of BOD_5 and COD. The reasoning for selecting the said parameters lies in the fact that BOD5 and COD are representatives of the biodegradable and non-biodegradable compounds in wastewater. On the other hand, biodegradable and non-biodegradable compounds in wastewater consist of insoluble and soluble fractions, and the insoluble compounds are considered TSS. Therefore, BOD5 and COD may have correlations with the TSS value of the wastewater. Moreover, the biodegradable and non-biodegradable compounds consist of nitrogenous and phosphorous compounds, which are represented by OrgN and NH3 (for nitrogenous compounds) and InorgP and OrgP (for phosphorous compounds). Following this, the correlation of BOD_5 and COD with each of the said variables was determined, as well as the sensitivity of BOD_5 and COD to each of them.

3.2. Predictive equations of BOD_5 and COD

To develop a GEP model, several parameters such as the number of the chromosomes, number of the genes, head size, functions linking the genes and other functions relevant to the equations should be adjusted to increase the R^2 while reducing the values of MSE and RMSE. The predicted results and the corresponding experimental values were compared to each other by linear regression. After several trial runs with different parameters, a GEP model was developed with a head size of 7 and 3 genes linked by the addition function and consisting of constants

Table 3
The statistical data of the parameters of the WWTPs' influent.

Variable	Unit	SD	Mean	Min	Max	Skewness	Kurtosis
BOD_5	mg/L as O_2	86.95	217.25	540.00	86.73	1.17	0.78
COD	mg/L as O_2	223.39	523.54	1300.00	145.08	0.95	0.40
TSS	mg/L	168.99	327.83	880.00	91.43	1.10	0.49
NH_3	mg/L as N	6.19	23.75	50.00	11.00	0.68	0.89
OrgN	mg/L as N	9.08	16.93	59.00	2.00	1.58	2.80
OrgP	mg/L as P	0.77	2.54	6.30	0.96	1.16	2.37
InrgP	mg/L as P	2.14	3.07	14.41	0.27	1.61	3.31

Table 4
The distribution functions of the normalized data.

Variable	Distribution fit	*p*-value	Parameters of the distribution
BOD_5	Lognormal	0.146	$\mu = -1.430$ $\sigma = 0.639$
COD	Lognormal	0.149	$\mu = -1.266$ $\sigma = 0.589$
TSS	Lognormal	0.193	$\mu = -1.425$ $\sigma = 0.707$
NH_3	Gamma	0.699	$\alpha = 4.21$ $\theta = 0.079$
OrgN	Lognormal	0.080	$\mu = -1.485$ $\sigma = 0.573$
OrgP	Gamma	0.159	$\alpha = 4.571$ $\theta = 0.063$
InorgP	Lognormal	0.635	$\mu = -1.864$ $\sigma = 0.746$

and independent variables of TSS, NH_3, OrgN, OrgP and InorgP. The expression trees of the genes for predicting BOD_5 and COD are illustrated in Figs. 3 and 4, respectively. The predictive equations of COD and BOD_5 including the functions and the parameters are also presented as Eq. (7) and Eq. (8), respectively.

$$COD = \sum_{i=1}^{4} Sub - ET\ i \tag{7}$$

In which:

$$Sub - ET\ 1 = Tanh(Arctan(TSS))$$

$$Sub - ET\ 2 = 0.377 \times NH_3 - 6.825 \times OrgP^2$$

$$Sub - ET\ 3 = OrgN \times (0.043 + 0.346 \times TSS \times InorgP)$$

$$Sub - ET\ 4 = Arctan(Arctan(OrgP \times (OrgP - 0.615)))$$

$$BOD_5 = \sum_{i=1}^{4} Sub - ET\ i \tag{8}$$

In which:

$$Sub - ET\ 1 = (-0.164 - TSS) \times (OrgP - OrgN) \times (Arctan(OrgN) - OrgP)$$

$$Sub - ET\ 2 = 0.023$$

$$Sub - ET\ 3 = TSS \times OrgN^{\left(-0.799 \times OrgN^2\right)}$$

$$Sub - ET\ 4 = Tanh(Arctan(TSS)) \times Arctan(-0.679 \times TSS)$$

As Eq. (8) shows, three parameters of TSS, OrgN, and OrgP are involved in the predictive equation of BOD_5. However, the correlations between BOD_5 and the above-mentioned parameters are nonlinear, which indicates the complicated nature of the relationships between these parameters. Due to the complexity of the equation and taking its derivative according to the independent parameters, the sensitivity of BOD_5 is calculated by MCS, which will be discussed later.

BOD_5 is comprised of particulate and non-particulate fractions, including colloidal and soluble compounds such as ammoniacal and carbonaceous ones. Particulate matter constitutes approximately 75% of BOD_5, therefore, it was expected that TSS would contribute more than other parameters in the developed BOD_5 prediction model. Non-particulate fraction of BOD_5 mostly consists of proteins or amino-acids which can be considered as organic nitrogenous compounds (Gerardi, 2015). It is noteworthy that ammonia can be oxidized by the nitrifying bacteria to nitrite and consequently to nitrate according to Eq. (9):

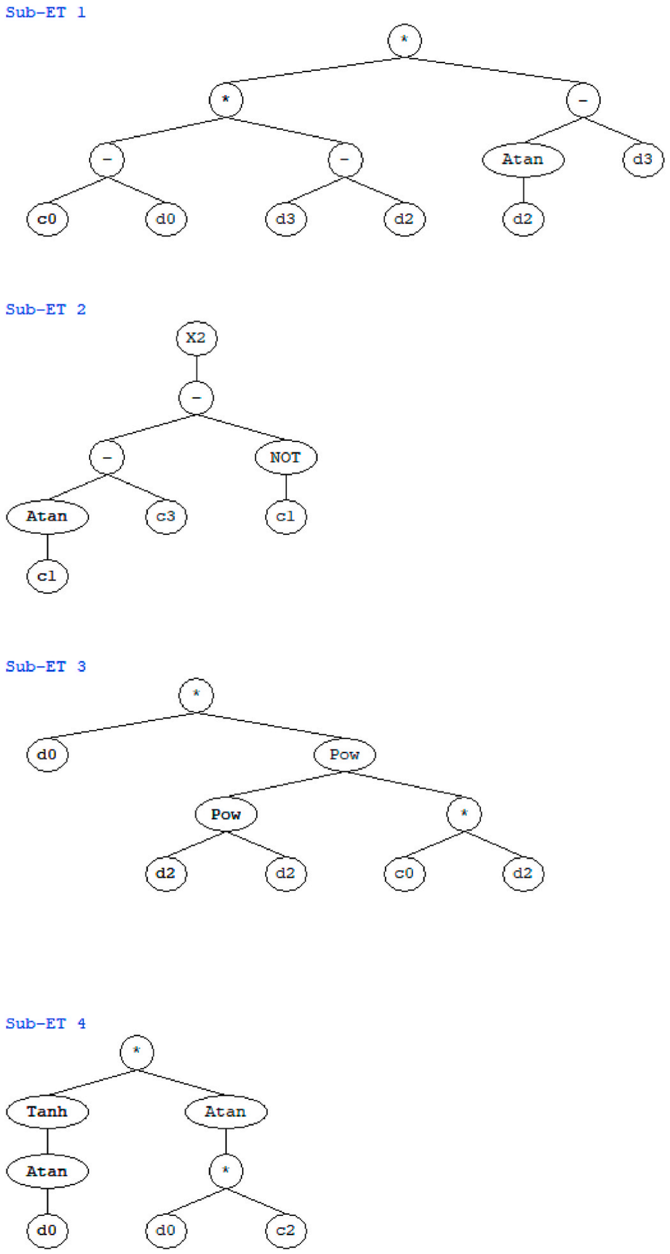


Fig. 4. GEP expression tree for predicting BOD_5 from TSS, NH_3, OrgN, OrgP and InorgP. (Note: d_0: concentration of TSS, d_1: concentration of NH_3, d_2: concentration of OrgN, d_3: concentration of OrgP, d_0: concentration of InorgP. The numbers of c_i are constants generated by GeneXpro.).

$$NH_3 + 2O_2 \xrightarrow{nitrifying\ bacteria} HNO_3 + H_2O \tag{9}$$

The ammonia oxidation by the nitrifying bacteria consumes oxygen, however, it generally occurs after 6 days. Therefore, ammonia is expected to have negligible effect on BOD_5, unless a considerable amount of groundwater containing nitrifying bacteria infiltrates into the sewage conveyance network. The absence of NH_3 variable in the GEP model for predicting BOD_5 complies with the negligible role of NH_3 hypothesis. On the other hand, inorganic phosphorous compounds, mostly comprised of phosphate salts, do not affect the BOD_5 value. On the other hand, involvement of OrgP term in the equations indicates that the biodegradable species in the wastewater can partially be comprised of organic phosphorous compounds.

On the other hand, the proposed equation for measuring COD is depicted in Eq. (10):

$$C_nH_aO_bN_c + dCr_2O_7^{2-} + (8d+c)H^+ \rightarrow nCO_2 + \frac{a+8d-3c}{2}H_2O + cNH_4^+ + 2dCr^{3+} \tag{10}$$

where $d = \frac{2n}{3} + \frac{a}{6} - \frac{b}{3} - \frac{c}{2}$.

According to Eq. (10), phosphorous compounds, theoretically, are not involved in the COD measurement, however, in reality, they may have negligible contributions to the process. As Eq (8) shows, ammoniacal compounds are not involved in COD measurement either. However, the presence of ammonia may interfere with COD measurement, since one of the final products of COD measurement is NH_4^+. It should be noted that the GEP modelling only considers the correlation between the parameters based on the training data but not the chemical reactions that occurs between them. Therefore, increasing ammonia concentration may be indicative of increasing the contaminants and therefore increasing COD value. Hence, the results of GEP modelling and theoretical chemical reactions should complement each other when interpreting the equations.

4. Comparative analysis

The accuracy and statistical metrics of the different models to predict BOD_5 and COD in the WWTPs are tabulated in Tables 5 and 6, respectively. The R^2 of the different models for predicting BOD_5 are in the order of GEP > RF > kNN > LR > MLP-NN > GB > RT. On the other hand, the accuracies of the employed AI algorithms for predicting COD are in the order of GEP > LR > MLP-NN > GB > RF > RT. Therefore, according to the statistical parameters of the AI models, the GEP model predicts BOD_5 and COD of the WWTPs' influent more efficiently than the other models. Moreover, the GEP model residuals are lower than those of the other models. In addition, GEP provides the relationship between the input and output parameters with specific mathematical functions, thereby allowing for a better understanding of the interactions that exists between the parameters. However, the other AI models serve as black box models due to their lack of derived equations.

As can be seen from Tables 5 and 6, the prediction of BOD_5 and COD with RT is the least accurate with R^2 values of 0.565 and 0.646, for BOD_5 and COD predictions, respectively. The low accuracy of RT can be attributed to the occurrence of overfitting during the development one tree. For prediction of BOD_5, GB has higher accuracy than RT, with R^2 value of 0.639. This is due to employing more trees in prediction of the target parameter. On the other hand, for predicting COD, GB, RF and kNN showed comparable accuracies with respective R^2 values of 0.807, 0.790 and 0.773. It should be recalled that in RF, GB and kNN, the value of the target parameter corresponding to the input vector (TSS_v, $NH_{3(v)}$, $OrgN_v$, $InorgP_v$, $OrgP_v$) is the mean value of the target parameters corresponding to some input vectors (TSS_i, $NH_{3(i)}$, $OrgN_i$, $InorgP_i$, $OrgP_i$), which are the closest values to the TSS_v, $NH_{3(v)}$, $OrgN_v$, $InorgP_v$, $OrgP_v$ in the structure formed during the training step. The similar accuracies of RF, GB and kNN in predicting COD may be due to the similar structure of the trees in RF and GB, as well as the neighboring values in kNN. This can be attributed to the presence of a high portion of data in a limited range and similar shape of the PDF of the normalized data. In predicting COD, the performance of kNN was weaker than that of GB and RF. This implies that the more organized structure of the trees in GB and RF modelling performs better than the structure of locating neighbors in kNN for predicting COD. This may be due to lower errors and more accuracy of COD measurement, which makes RF and GB better predictive models than kNN. For predicting COD, MLP-NN and LR perform better than RF, GB and kNN, but their accuracies are very similar. Also, MLP-NN and LR depicted comparable accuracies, which are slightly higher than those observed for other models, except GEP. The comparable performances of LR and MLP-NN implies that the feed-forward and backward propagation mechanisms of the MLP-NN conclude a semi-linear relationship between the target parameters and the input independent variables.

For predicting BOD_5, GB performs weaker than RF, implying that the GB algorithm for generating trees is not suitable for BOD_5 prediction. However, RF algorithm for BOD_5 estimation showed a better performance level. For COD estimation, LR and MLP-NN similarly showed comparable accuracies in BOD_5 estimation. However, LR has lower accuracy than RF, indicating that correlation of BOD_5 with the independent variables tend to be more non-linear than that of COD.

Since RT, RF, GB, kNN and MLP-NN do not provide specific relationship function, the linear correlation provided by LR is further investigated. Eq. (11) and Eq. (12) show the linear correlation of COD and BOD_5, respectively, with the input parameters.

$$COD = 0.511 \times TSS + 0.332 \times NH_3 + 0.174 \times OrgN - 0.076 \times OrgP + 0.159 \times InorgP \tag{11}$$

$$BOD_5 = 0.605 \times TSS + 0.046 \times NH_3 + 0.142 \times OrgN + 0.095 \times OrgP + 0.116 \times InorgP \tag{12}$$

As Eqs. (11) and (12), show, TSS contribute higher than 50% to predicting BOD_5 and COD. This is due to the fact that suspended solids of wastewater contain a large amount of organic compounds, which are represented by BOD_5 and COD (Veronez et al., 2018). However, NH_3 plays a negligible role in predicting BOD_5 as earlier highlighted, owing

Table 5
Performance parameters of the BOD_5 predictive models.

Target variable	model	stage	MSE	RMSE	MAE	R^2
BOD_5	GEP	Training	0.009	0.095	0.067	0.727
		Validation	0.008	0.092	0.068	0.749
	RF	Validation	0.008	0.088	0.068	0.708
	kNN	Validation	0.009	0.093	0.07	0.675
	LR	Validation	0.009	0.094	0.071	0.669
	MLP-NN	Validation	0.009	0.096	0.072	0.655
	GB	Validation	0.01	0.098	0.073	0.639
	RT	Validation	0.012	0.108	0.082	0.565

Table 6
Performance parameters of the COD predictive models.

Target variable	model	stage	MSE	RMSE	MAE	R^2
COD	GEP	Training	0.009	0.095	0.067	0.816
		Validation	0.008	0.092	0.068	0.862
	LR	Validation	0.005	0.072	0.054	0.841
	MLP-NN	Validation	0.005	0.073	0.056	0.836
	GB	Validation	0.006	0.08	0.061	0.807
	RF	Validation	0.007	0.083	0.063	0.790
	kNN	Validation	0.007	0.086	0.062	0.773
	RT	Validation	0.012	0.108	0.082	0.646

to the fact that biologically degradable content of wastewater mostly contain organic compounds and not ammoniacal compounds. On the other hand, NH_3 has significant role (up to 33%) in predicting COD. This may be due to the hidden correlation between the concentrations of NH_3 and other toxic or non-biodegradable compounds in the wastewater samples, which lead to higher COD values. From the practical point of view, it can be inferred that higher ammonia concentrations in the WWTPs' inflows may be indicative of the presence of toxic compounds, which may disturb the reactions in the biological treatment stages (Henze et al., 2008).

It is noteworthy that the coefficients of NH_3 in the predictive models of BOD_5 and COD significantly differ from each other. Meanwhile, the coefficients of the other independent variables such as TSS, OrgN, OrgP and InorgP in the predictive models of BOD_5 and COD are comparable. Since the performance of the GEP model was superior to the other ones in estimation of BOD_5 and COD, its performance will be investigated in more detail and the sensitivity analysis will be conducted based on the GEP model in the following sections.

5. Performance of the GEP model

The performance of the GEP model for predicting BOD_5 and COD for the training and validation was evaluated according to the statistical indicators of R^2, MSE, RMSE, MAE, RSE, and RRSE as shown in Tables 5 and 6 The results show that the developed predictive model has satisfactory performance for both training and validation of the dataset. The plots of the predicted versus observed BOD_5 and COD values for both training and validation dataset are presented in Figs. 5 and 6, respectively. The R^2 of the training and validation datasets are 0.749 and 0.727, respectively, for predicting BOD_5 and 0.816 and 0.861 for COD, indicating sufficiency of the GEP model for predicting BOD_5 and COD in the influent of the WWTPs, based on the values of TSS, NH_3, OrgN, OrgP and InorgP. It is noteworthy that the accuracy of the model for predicting COD is higher than that of BOD_5, since measuring BOD_5 is prone to higher errors. Measuring BOD_5 requires several steps and each of the steps has its own intrinsic errors. On the other hand, the data were acquired over three years from different areas of HK, which may be affected by intrusion of the industrial wastewater. Therefore, the water quality input parameters were non-homogenous, which may have affected the accuracy of the correlations between the parameters (Yaqub et al., 2022b). Notably, the normalization was done for developing the models as expressed in the methodology, based on which the correlation between the parameters were found. However, for plotting Figs. 5 and 6, the "predicted" values, which are normalized values, were converted to the original values to obtain more realistic comparison with the observed values.

Although GEP model exhibits strong fitness with the actual data, digging deeper into the errors and investigating the data with higher errors provided insights on minimizing their influence on predictions. To this purpose, the differences between the predicted and actual values as well as the error percentage are obtained. Based on the results, for 46% of the cases, the predicted BOD_5 was lower than the actual BOD_5. These cases may lead to underestimating the BOD_5 and consequently underestimating the aeration amount. However, a more accurate look at the results reveal that these errors can be avoided. For example, 84% of the underestimation cases showed that the actual BOD_5 is 20% higher than the predicted ones. Therefore, by applying 20% safety factor, the underestimation error can be avoided. On the other hand, 16% of the underestimation cases showed that the actual BOD_5 is 20% higher than the predicted ones, which may be indicative of erroneous measurements of parameters. Therefore, frequent inspection and calibration of the sensors installed in the WWTP's inflow could be pertinent to minimizing the errors associated with the measured parameters. On the other hand, the errors in the overestimation of BOD_5 follows similar trend as the underestimation one.

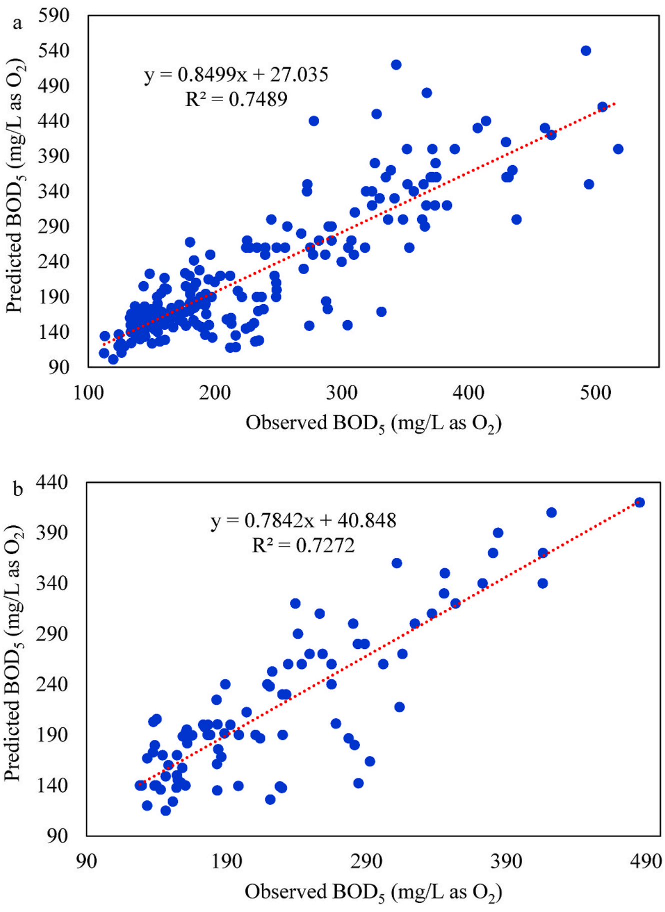


Fig. 5. The scatter plots of the observed BOD_5 values versus predicted ones for (a) training and (b) validation by the GEP model.

6. Sensitivity analysis using MCS

In order to investigate the most influential parameters, their effects and the interactions between them, sensitivity analysis was conducted by employing MCS. The MCS can provide insights on the uncertainty of the parameters and the major parameters affecting the change in the target variables (Golzar et al., 2020). This requires random generation of a dataset of the input variables and determining the corresponding BOD_5 for each set of inputs by applying the GEP derived equation to them. In order to generate the random input variables, the PDF of normalized values for each parameter was initially defined based on the actual data. The PDF of the normalized values of all variables with the p-value of the fitted distribution function on the actual data are tabulated in Table 4.

After obtaining 1000 data points for the predictors and the target parameter, the sensitivity of BOD_5 to each independent variable is found by changing that variable by ±10%, while keeping the other variables constant, as depicted in Table 7. As per Table 7, TSS is the most influential parameter for BOD_5 and COD estimation. 10% increase in the average TSS value of 309.11nullmg/L at the average values of the other independent variables, leads to a 7.94% increase in the BOD_5 value from 212 to 222nullmg/L as O_2. Likewise, 10% decrease in the average TSS value leads to 7.88% decrease in the average value of BOD_5, which corresponds to 14nullmg/L as O_2 decrease in BOD_5 concentration. On the other hand, 10% increase or decrease in the value of average TSS results in 7.918% increase and 8.269% decrease in the mean value of COD, respectively. The strong effect of TSS on the BOD_5 may be due to the similar sources of TSS and BOD_5. According to Henze et al. (2008), BOD_5 is representative of the biodegradable compounds by the microorganisms, and these compounds can be in either soluble or particulate

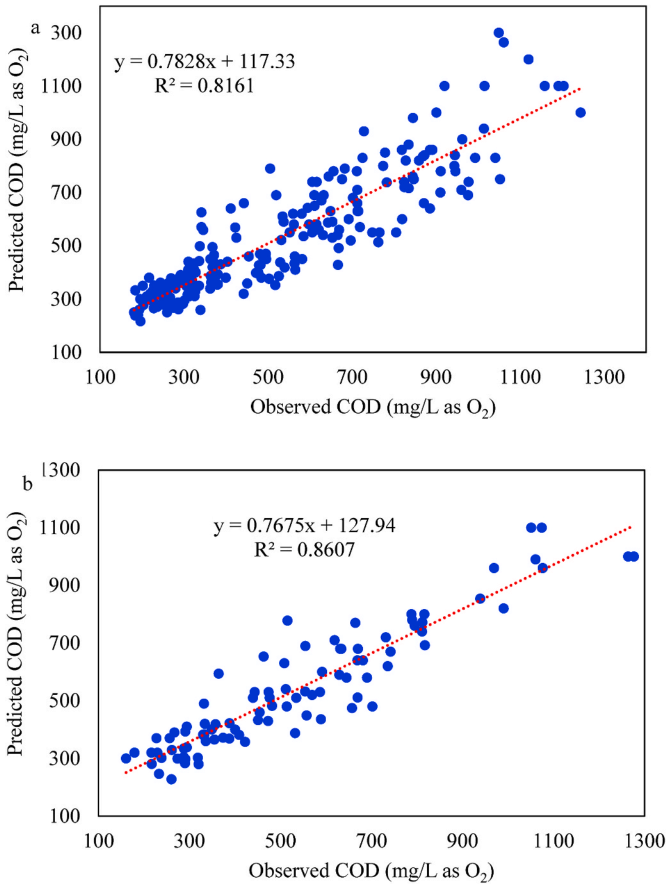


Fig. 6. The scatter plots of the observed COD values versus predicted ones for (a) training and (b) validation by the GEP model.

Table 7
Variations in the BOD_5 and COD values by changing the values of the independent variables by ±10%.

Parameter	Average value	Variation in the average value	BOD_5 (mg/L as O_2) change (%) compared with the average value	COD (mg/L as O_2) change (%) compared with the average value
TSS	309.11 mg/L	+10% −10%	7.94 −7.88	7.918 −8.269
NH_3	23.42 mg/L as N	+10% −10%	0 0	5.176 −5.176
OrgN	16.28 mg/L as N	+10% −10%	−0.448 0.585	0.415 −0.415
OrgP	2.49 mg/L as P	+10% −10%	0.358 0.077	−0.723 1.580
InorgP	2.92 mg/L as P	+10% −10%	0 0	0.142 −0.142

forms; thus, an increase in the amount of the particulate compounds (which are represented by TSS) implies an increase in the value of BOD_5.

Having said that, the BOD_5 value is not affected by NH_3 and InorgP, due to the fact that the biodegradable compounds in the wastewater are not comprised of ammonical and inorganic phosphorous compounds. The main reason for this is that the ammonical compounds and phosphate salts have non-biodegradability characteristics (Hou et al., 2022). It is noteworthy that according to Eq. (12), NH_3 and InorgP slightly contribute to the value of BOD_5 in the LR model. However, this does not seem reasonable. Therefore, the GEP model represents more accurate predictive model in terms of the interactions of various wastewater parameters. Moreover, 10% increase or decrease in the values of organic nitrogenous or organic phosphorous compounds resulted in less than 1% of change in BOD_5 value. The low contribution of OrgN and OrgP to the estimation of BOD_5 lies in the following reasons: the GEP algorithm considers OrgN and OrgP as part of TSS, or negligible contribution of organic nitrogenous and phosphorous compounds to the compounds which require oxygen in the aeration process. Either of this indicates that intrusion of industrial sewages containing large amounts of organic nitrogenous and phosphorous to the municipal wastewater may not increase BOD_5 (Chen et al., 2023). On the other hand, if they contain toxic and non-biodegradable compounds (that are represented by COD/BOD_5 ratio), they negatively impact the BOD_5 value.

On the other hand, NH_3 is the second most influential factor contributing to COD estimation after TSS, while OrgN, OrgP and InorgP has negligible effect on COD estimated value. High contribution of NH_3 to the predicted value of COD is compatible with the LR model of COD estimation and indicates similar sources of ammoniacal and non-biodegradable compounds. This is of great importance from a practical point of view in the sense that detecting high values of NH_3 in the WWTPs' inflow may lead to disturbances in the biological treatment stages and pre-cautionary actions are required to be taken.

7. Environmental and engineering implications

In the first glance, all wastewater parameters may have a positive correlation with each other because the parameters are representative of the wastewater pollution levels, whereby an increase in one of them implies an increase in the other parameters. However, the statistical analysis opposes this.

It was determined that the most influential parameter for predicting BOD_5 and COD is TSS. To explain the correlation between BOD_5 and TSS, it should be noted that BOD consists of soluble BOD_5 and particulate BOD_5. On the other hand, TSS consists of biodegradable and non-biodegradable portions. The strong correlation between BOD_5 and TSS implies that major proportions of the biodegradable materials are in the particulate form. On the other hand, the major proportion of the TSS is biodegradable. On the other hand, BOD_5 has no correlation with $N–NH_{3,}$ indicating that ammoniacal nitrogen comprises of a negligible proportion of BOD_5, and increasing BOD_5 may not lead to an increase in $N–NH_3$. On the other hand, BOD_5 has good correlation with OrgN. This implies that the organic nitrogenous compounds significantly contribute to the oxygen demand of a wastewater sample. Surprisingly, conductivity of the wastewater, which is representative of the concentrations of the ions in the wastewater, has no correlation with BOD_5. This lack of correlation indicates that the variations in the concentrations of ions in the wastewater may not affect the oxygen demand and the activity of the aerobic bacteria.

GEP provides a mathematical relationship between the inputs and outputs, thereby facilitating the analyses of the interactions between the input parameters as well as the effects of their variations on the output results under certain conditions. The availability of a function that can capture the complex relationships that exists between the parameters is crucial for developing insights that will facilitate accurate estimations of the target parameters. Such include BOD_5 and COD in the case of fluctuations in a WWTP's influent. On the other hand, the target parameters for this study, namely BOD_5 and COD can be associated with the energy consumption and aeration amount during the biological treatment. In addition, given the target BOD_5 and TSS removal value, the sludge production can be estimated, which could pave way for forecasting the operational parameters during sludge processing. This leads to extensive energy saving during the wastewater treatment processes.

8. Conclusions

This paper examines the utilization of several AI-based algorithms for predicting BOD_5 and COD of the influents of WWTPs. The data were collected from the inflow of 7 municipal wastewater treatment plants in Hong Kong over a period of three years. Based on the various statistical tests undertaken, it was seen that the GEP-based model provides more accurate results as compared to the other AI-based techniques. Moreover, the sensitivity of the target parameters to the variations of the input parameters was assessed through Monte Carlo simulation. This study has led to three major contributions as follows. Firstly, a comprehensive mathematical expression for predicting BOD_5 and COD from TSS, NH_3, OrgN, OrgP and InorgP was developed and validated; the specific mathematical equation given by the GEP model provides insights on the contribution of each independent variable on estimation of BOD_5 and COD. Secondly, the most influential parameters on estimating BOD_5 and COD were revealed using the mathematical expressions developed by the GEP model; it was seen that TSS was the most influential parameter for the estimation of BOD_5 and COD. Thirdly, the uncertainty of the BOD_5 and COD values were assessed using MCS; it was observed that TSS had the highest level of sensitivity to the BOD_5 and COD, while InorgP was the most insensitive parameter in the developed model to predict both BOD_5 and COD.

The results of this work can be used to adjust the chemical dosage as well as estimate the aeration demand of the consecutive wastewater treatment stages in order to minimize the energy cost and production of greenhouse gases. Aside from the above-mentioned contributions, this study has some limitations which necessitate the corresponding future works as follows. Firstly, this study was unable to generate a comprehensive investigation of the correlations between the influent parameters and the energy and cost-related parameters in a wastewater treatment plants, due to the lack of data related to the energy and cost of the WWTPs' operation; thus, future endeavors need to develop a specific mathematical model for predicting energy saving of WWTPs' operation. Secondly, the lack of data on industrial wastewater in Hong Kong also impeded the ability of this study to compare the performances of predictive AI-based models for municipal and industrial wastewater treatment plants and water treatment plants. The third limitation of this study is associated with the WWTPs' prediction under different processes; thus, future research need to develop an AI-based predictive model by taking into account the activated sludge and other nitrogen and phosphorous removal processes. Finally, the quantitative evaluation of the impact of predicting wastewater parameters on reducing the energy consumption in the wastewater treatment plants is another limitation of this work, which needs to be considered in future research.

CRediT authorship contribution statement

Ehsan Aghdam: Data curation, Methodology, Software, Formal analysis, Investigation, Visualization, Validation, Writing – original draft, Writing – review & editing. **Saeed Reza Mohandes:** Conceptualization, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Patrick Manu:** Writing – original draft, Writing – review & editing. **Clara Man Cheung:** Writing – original draft, Writing – review & editing. **Akilu Yunusa-Kaltungo:** Writing – original draft, Writing – review & editing. **Tarek Zayed:** Writing – original draft, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

The authors gratefully acknowledge the support from the Environment Conservation Fund (ECF) under grant number ECF/058/2019 and the Drainage Services Department (DSD) of the Government of Hong Kong for providing the required data and case study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclepro.2023.137019.

References

Abouzari, M., Pahlavani, P., Izaditame, F., Bigdeli, B., 2021. Estimating the chemical oxygen demand of petrochemical wastewater treatment plants using linear and nonlinear statistical models–A case study. Chemosphere 270, 129465.

Ansari, M., Othman, F., El-Shafie, A., 2020. Optimized fuzzy inference system to enhance prediction accuracy for influent characteristics of a sewage treatment plant. Sci. Total Environ. 722, 137878.

Asami, H., Golabi, M., Albaji, M., 2021. Simulation of the biochemical and chemical oxygen demand and total suspended solids in wastewater treatment plants: data-mining approach. J. Clean. Prod. 296 https://doi.org/10.1016/j.jclepro.2021.126533.

Baird, R., Rice, E.W., Eaton, A.D., Bridgewater, L., Federation, W.E., 2017. Standard Methods for the Examination of Water and Wastewater. American Public Health Association.

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. Chemometr. Intell. Lab. Syst. 104, 172–180.

Benedict, A.H., Merrill, M.S., Mauseth, G.S., 1979. Sludge production, waste composition, and BOD loading effects for activated sludge systems. Journal (Water Pollution Control Federation) 2898–2915.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 2017. Classification and Regression Trees. Routledge.

Bunce, J.T., Graham, D.W., 2019. A simple approach to predicting the reliability of small wastewater treatment plants. Water (Basel) 11, 2397.

Chen, H., Wang, K., She, S., Yu, X., Yu, L., Xue, G., Li, X., 2023. Insight into dissolved organic nitrogen transformation and characteristics: focus on printing and dyeing wastewater treatment process. J. Hazard Mater., 131086

Cheng, Z., Li, X., Bai, Y., Li, C., 2018. Multi-Scale fuzzy inference system for influent characteristic prediction of wastewater treatment. CLEAN–Soil, Air, Water 46, 1700343.

Ching, P.M.L., Zou, X., Wu, D., So, R.H.Y., Chen, G.H., 2022. Development of a wide-range soft sensor for predicting wastewater BOD5 using an eXtreme gradient boosting (XGBoost) machine. Environ. Res. 210, 112953.

Dürrenmatt, D.J., Gujer, W., 2012. Data-driven modeling approaches to support wastewater treatment plant operation. Environ. Model. Software 30, 47–56.

Eddy, M., Abu-Orf, M., Bowden, G., Burton, F.L., Pfrang, W., Stensel, H.D., Tchobanoglous, G., Tsuchihashi, R., Firm), A., 2014. Wastewater Engineering: Treatment and Resource Recovery. McGraw Hill Education.

Ferreira, C., 2002. Gene expression programming in problem solving. In: Soft Computing and Industry. Springer, pp. 635–653.

Foschi, J., Turolla, A., Antonelli, M., 2021. Soft sensor predictor of E. coli concentration based on conventional monitoring parameters for wastewater disinfection control. Water Res. 191, 116806.

Gerardi, M.H., 2015. The Biology and Troubleshooting of Facultative Lagoons. John Wiley & Sons.

Golzar, F., Nilsson, D., Martin, V., 2020. Forecasting wastewater temperature based on artificial neural network (ANN) technique and Monte Carlo sensitivity analysis. Sustainability 12, 6386.

Granata, F., Papirio, S., Esposito, G., Gargano, R., de Marinis, G., 2017. Machine learning algorithms for the forecasting of wastewater quality indicators. Water (Basel) 9, 105.

Haimi, H., Corona, F., Mulas, M., Sundell, L., Heinonen, M., Vahala, R., 2015. Shall we use hardware sensor measurements or soft-sensor estimates? Case study in a full-scale WWTP. Environ. Model. Software 72, 215–229.

Henze, M., van Loosdrecht, M.C.M., Ekama, G.A., Brdjanovic, D., 2008. Biological Wastewater Treatment. IWA publishing.

HKDSD, 2018. Sewerage Strategy [WWW Document]. URL: https://www.dsd.gov.hk/EN/Sewerage/Sewerage_Strategy/index.html.

Hou, Z., Zhou, X., Zhao, Z., Dong, W., Wang, H., Liu, H., Zeng, Z., Xie, J., 2022. Advanced aromatic organic compounds removal from refractory coking wastewater in a step-feed three-stage integrated A/O bio-filter: spectrum characterization and biodegradation mechanism. J. Environ. Manag. 322, 116140.

Islam, M.S., Mohandes, S.R., Mahdiyar, A., Fallahpour, A., Olanipekun, A.O., 2022. A Coupled Genetic Programming Monte Carlo Simulation–Based Model for Cost Overrun Prediction of Thermal Power Plant Projects. J. Construct.Eng. Manag. 148 (8), 04022073.

Jouanneau, S., Recoules, L., Durand, M.J., Boukabache, A., Picot, V., Primault, Y., Lakel, A., Sengelin, M., Barillon, B., Thouand, G., 2014. Methods for assessing biochemical oxygen demand (BOD): a review. Water Res. 49, 62–82. https://doi.org/10.1016/j.watres.2013.10.066.

Kayadelen, C., Günaydın, O., Fener, M., Demir, A., Özvan, A., 2009. Modeling of the angle of shearing resistance of soils using soft computing systems. Expert Syst. Appl. 36, 11814–11826. https://doi.org/10.1016/j.eswa.2009.04.008.

Kim, M., Kim, Y., Kim, H., Piao, W., Kim, C., 2016. Evaluation of the k-nearest neighbor method for forecasting the influent characteristics of wastewater treatment plant. Front. Environ. Sci. Eng. 10, 299–310.

Kim, S., Alizamir, M., Zounemat-Kermani, M., Kisi, O., Singh, V.P., 2020. Assessing the biochemical oxygen demand using neural networks and ensemble tree approaches in South Korea. J. Environ. Manag. 270, 110834.

Lakshmanaprabu, S.K., Shankar, K., Ilayaraja, M., Nasir, A.W., Vijayakumar, V., Chilamkurti, N., 2019. Random forest for big data classification in the internet of things using optimal features. International journal of machine learning and cybernetics 10, 2609–2618.

Li, F., Vanrolleghem, P.A., 2022. An Essential Tool for WRRF Modelling: a Realistic and Complete Influent Generator for Flow Rate and Water Quality Based on Data-Driven Methods. Water Science and Technology.

Luo, L., Dzakpasu, M., Yang, B., Zhang, W., Yang, Y., Wang, X.C., 2019. A novel index of total oxygen demand for the comprehensive evaluation of energy consumption for urban wastewater treatment. Appl. Energy 236, 253–261.

Mateo Pérez, V., Mesa Fernández, J.M., Villanueva Balsera, J., Alonso Álvarez, C., 2021. A random forest model for the prediction of fog content in inlet wastewater from urban wwtps. Water (Basel) 13, 1237.

Metcalf, P., 2003. *They lie, we lie: Getting on with anthropology*. Routledge.

Noman, E.A., Al-Gheethi, A.A., Saphira, R.M.R.M., Talip, B.A., Al-Sahari, M., Ismail, N., 2022. Mathematical prediction models for inactivation of antibiotic-resistant bacteria in kitchen wastewater by bimetallic bionanoparticles using machine learning with gene expression programming. J. Clean. Prod. 333, 130131.

Oliveira, S.C., Souki, I., von Sperling, M., 2012. Lognormal behaviour of untreated and treated wastewater constituents. Water Sci. Technol. 65, 596–603.

Olsson, G., 2012. ICA and me – a subjective review. Water Res. 46, 1585–1624. https://doi.org/10.1016/j.watres.2011.12.054.

Olsson, G., Carlsson, B., Comas, J., Copp, J., Gernaey, K. v, Ingildsen, P., Jeppsson, U., Kim, C., Rieger, L., Rodriguez-Roda, I., 2014. Instrumentation, control and automation in wastewater–from London 1973 to Narbonne 2013. Water Sci. Technol. 69, 1373–1385.

Seckin, G., Yilmaz, T., Guven, A., Yuceer, A., Basibuyuk, M., Ersu, C.B., 2011. Modeling the performance of upflow anaerobic filters treating paper-mill wastewater using gene-expression programming. Ecol. Eng. 37, 523–528.

Shah, M.I., Alaloul, W.S., Alqahtani, A., Aldrees, A., Musarat, M.A., Javed, M.F., 2021a. Predictive modeling approach for surface water quality: development and comparison of machine learning models. Sustainability 13, 7515.

Shah, M.I., Javed, M.F., Abunama, T., 2021b. Proposed formulation of surface water quality and modelling using gene expression, machine learning, and regression techniques. Environ. Sci. Pollut. Control Ser. 28, 13202–13220.

Sharafati, A., Asadollah, S.B.H.S., Hosseinzadeh, M., 2020. The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty. Process Saf. Environ. Protect. 140, 68–78.

Shishegaran, A., Boushehri, A.N., Ismail, A.F., 2020. Gene expression programming for process parameter optimization during ultrafiltration of surfactant wastewater using hydrophilic polyethersulfone membrane. J. Environ. Manag. 264, 110444.

Tijani, I.A., Zayed, T., 2022. Gene expression programming based mathematical modeling for leak detection of water distribution networks. Measurement 188, 110611.

Vanrolleghem, P.A., Lee, D.S., 2003. On-line monitoring equipment for wastewater treatment processes: state of the art. Water Sci. Technol. 47, 1–34.

Veronez, Camila Lopes, Moreno, Adriana S., Constantino-Silva, Rosemeire Navickas, Maia, Luana SM, Ferriani, Mariana PL, Castro, Fábio FM, Valle, Solange Rodrigues, et al., 2018. Hereditary angioedema with normal C1 inhibitor and F12 mutations in 42 Brazilian families. *The Journal of Allergy and Clinical Immunology: In Practice* 6 4, 1209–1216.

Wang, H., Yang, Y., Keller, A.A., Li, X., Feng, S., Dong, Y., Li, F., 2016. Comparative analysis of energy intensity and carbon emissions in wastewater treatment in USA, Germany, China and South Africa. Appl. Energy 184, 873–881. https://doi.org/10.1016/j.apenergy.2016.07.061.

Wang, J.-H., Zhao, X.-L., Guo, Z.-W., Yan, P., Gao, X., Shen, Y., Chen, Y.-P., 2022. A full-view management method based on artificial neural networks for energy and material-savings in wastewater treatment plants. Environ. Res. 211, 113054.

Wang, R., Yu, Y., Chen, Y., Pan, Z., Li, X., Tan, Z., Zhang, J., 2022. Model construction and application for effluent prediction in wastewater treatment plant: data processing method optimization and process parameters integration. J. Environ. Manag. 302, 114020.

Wang, Z., Man, Y., Hu, Y., Li, J., Hong, M., Cui, P., 2019. A deep learning based dynamic COD prediction model for urban sewage. Environ. Sci. J. Integr. Environ. Res.: Water Research and Technology 5, 2210–2218. https://doi.org/10.1039/c9ew00505f.

Wodecka, B., Drewnowski, J., Białek, A., Łazuka, E., Szulżyk-Cieplak, J., 2022. Prediction of wastewater quality at a wastewater treatment plant inlet using a system based on machine learning methods. Processes 10, 85.

Yaqub, M., Lee, S.H., Lee, W., 2022a. Investigating micellar-enhanced ultrafiltration (MEUF) of mercury and arsenic from aqueous solution using response surface methodology and gene expression programming. Separ. Purif. Technol. 281, 119880.

Yaqub, M., Park, S., Alzahrani, E., Farouk, A.E.A., Lee, W., 2022b. Application of data-driven machine learning to predict propranolol and trimethoprim removal using a managed aquifer recharge system. J. Environ. Chem. Eng. 10 https://doi.org/10.1016/j.jece.2021.106847.