Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2018.22095

117 (2018) 78–87
June

Activated sludge process modelling using selected machine learning techniques

Bartosz Szeląg[a,*], Krzysztof Barbusiński[b], Jan Studziński[c]

[a]*Kielce University of Technology, Tysiąclecia Państwa Polskiego 7 Av., 25-314 Kielce, Poland, Tel. +41-342-47-35; email: bszelag@tu.kielce.pl*
[b]*Silesian University of Technology, Konarskiego 18 Street, 44-100 Gliwice, Poland, Tel. +32-237-11-94; email: krzysztof.barbusinski@polsl.pl*
[c]*Systems Research Institute PAN, Newelska 6 Street, 00-001 Warszawa, Poland, Tel. +22-381-02-75; email: janstudzinski@ibspan.waw.pl*

Received 20 December 2017; Accepted 20 February 2018

ABSTRACT

An approach to forecast the mixed liquor suspended solids (MLSS) and food-to-mass ratio (F/M) of the activated sludge in bioreactor using some methods of statistical modelling has been proposed. The impact of explanatory variables used in the models on the exactness of the models developed has also been analyzed. Those variables are wastewater quality indicators and parameters of activated sludge chambers while the modelling methods used are the support vectors machines, cascade neural networks and boosted trees. Moreover, the possibility of modelling those variables based on the measurements of wastewater flow and temperature in the wastewater inflow to the wastewater treatment plant has been investigated. It was concluded that the MLSS as well as the F/M could be successfully forecasted by variety of statistical models in which the wastewater quality indicators are not measured but modelled. The method is very useful operationally because it makes possible to monitor and correct the values of MLSS and F/M quickly and efficiently while only a limited access to the wastewater quality measurements is available.

Keywords: Wastewater treatment plant; Food-to-mass ratio; Mixed liquor suspended solids; Cascade neural network; Support vector machines; Boosted trees

1. Introduction

Owing to changes in amount and quality of raw wastewater inflow, online control of parameters of activated sludge chambers (ASC) is needed to maintain the stable conditions of the wastewater treatment process. Such permanent control is aimed to assert steady values of the sludge age, food-to-mass ratio (F/M) and the biomass (or sludge) mixed liquor suspended solids (MLSS). In typical technological systems to remove the carbon, nitrogen and phosphorous compounds from the wastewater, it is recommended that F/M value cannot be higher than 0.10 g biological oxygen demand (BOD_5)/g MLSS·d and it cannot be lower than 0.05 g BOD_5/g MLSS·d; then the problems concerning the sludge sedimentation induced by filamentous bacteria can be eliminated [1–5]. Nowadays a correction of ASC parameters is performed ad hoc in real time what is usually neither economically nor ecologically efficient. Hence, the loads of the biodegradable wastewater compounds as well as other wastewater quality indicators are suggested to be based on statistical models. Such approach would help the treatment plant operator to predict the correction values of ASC parameters and to maintain the F/M and MLSS values in the really optimal range. It has been recently shown [6–9] that statistical or physical models are widely developed in order to improve the efficiency of ASC operation. However, the use of physical models is connected with many calculation troubles concerning their calibration and the complexity of description of biochemical reactions. Therefore, the statistical models have been increasingly applied when their structure is generated on the stage of model learning and when the exactness of their prediction ability is checked on the stage of model testing. Those models

* Corresponding author.

Presented at the 13th Conference on Microcontaminants in Human Environment, 4–6 December 2017, Czestochowa, Poland.

are created by means of certain data mining methods, such as neural networks and their modifications, support vectors machines (SVM) or method of random trees. The methods are commonly used to optimize the processes of sewage nitrification, denitrification, dephosphatation and aeration, and also of sludge sedimentation or BOD_5 and chemical oxygen demand (COD) reduction in the wastewater [10–13]. The models could be applied in the prediction processes and they could improve essentially the efficiency of wastewater treatment plants (WWTP). Nevertheless, to our knowledge, the importance of controlling F/M parameter in WWTP operation has been described very rarely [14]. Moreover, in the majority of the models developed, the wastewater quality indicators and ASC operational parameters are considered as the input variables. Such approach is unaccepted from economic and operating points of view. Therefore, it is strongly advised to develop some sophisticated models addressed to predict the difficult measurable variables on the basis of other variables which are fast and easily measured.

A novel route to model activated sludge concentration and substrate loading has been described in this paper. The models could be applied in a lack of input data concerning the wastewater quality or reactor operating parameters. To avoid the costs resulted from online measurements of wastewater input data, it is recommended to predict those values on the basis of temperature and wastewater inflow measurements using the selected data mining methods. The goal of the mentioned analyses is to develop a method to model and control a biological reactor working in a continuous manner.

2. Object of investigation

The measurement data used in the following for modelling and forecasting have been collected from the WWTP Sitkówka-Nowiny, located in Kielce in central part of Poland. The nominal capacity of that WWTP is 72.000 m^3/d what corresponds to the population equivalent of 275.000 p.e. The flowing wastewater is pretreated mechanically on stepped bars and in the aerated grits (the fats removal included) and then it is primarily clarified. After that it flows into the biological reactor supplied with BARDENPHO system. Then, it is transported into four secondary clarifiers where the process of a final clarification occurs and after that the treated wastewater is discharged into Bobrza river.

3. Methodology

Some selected statistical models to predict the MLSS and F/M were applied. The impact of individual input variables (e.g., amount and quality of wastewater inflow) on the exactness of prediction of modelled bioreactor parameters has been analyzed. The MLSS prediction is described by the formula:

$$\mathrm{MLSS} = f\left(x_1, x_2, x_3, x_i \ldots, x_n\right) \tag{1}$$

where n is the number of input variables concerned in the models (i = 1,2,3,…,n); x_i is the variable values concerning the wastewater inflow (Q), the wastewater quality indicators in the inflow ($BOD_{5,in}$, COD_{in}, TSS_{in}, TN_{in} and $N–NH_{4,in}^+$), operational parameters (Z_t, where t = 1,2,3, …6) of ASC (pH, temperature in bioreactor (T_{sl}), oxygen concentration in nitrification chamber (dissolved oxygen [DO]), return activated sludge (RAS) pumping rate in % of daily flow and amount of excessive sludge (waste activated sludge [WAS]) directed to anaerobic digesters, methanol added (m_{met})).

The possibility to predict the concentrations of $BOD_{5,in}$ and COD_{in}, total suspended solids (TSS_{in}), total nitrogen (TN_{in}) and ammoniacal nitrogen ($N–NH_{4,in}^+$) based on the measurements of daily wastewater inflow (Q) and of its temperature (T_{in}) can be described by the following relation [15,16]:

$$C\left(t\right)_j = f\left(\begin{matrix} Q\left(t-1\right), Q\left(t-2\right), Q\left(t-m\right), \ldots, \\ T_{in}\left(t-1\right), T_{in}\left(t-2\right), T_{in}\left(t-k\right) \end{matrix}\right) \tag{2}$$

where m, k is the time gaps in the measurements of concerned variables, j is the number of a wastewater quality indicator analyzed. The stationarity of the time series collected was checked using Mann Kendall test for trend analysis prior to determining the model prognoses of wastewater quality and quantity.

Eq. (2) is based on the assumption that the wastewater quality characterized by the pollution load is predictable and the dilution of the wastewater inflow and the corresponding pollution degradation influence the values of respective wastewater quality indicators. That assumption has been confirmed [15–17] for the distribution wastewater network supplying the WWTP investigated.

To identify the explanatory variables used for modelling the wastewater quality indicators the method of boosted trees (BT) has been used [18]. In practical considerations this is a frequently used approach what allows to limit the number of inputs (explanatory variables) in mathematical models [18,19]. So, the importance indicators for particular predictors have been calculated and the ranking list of them was prepared. The variables taken into account in Eq. (2) are based on that list. Based on the models the F/M has been calculated using the formula:

$$\frac{F}{M} = \frac{Q_{\mathrm{cal}}(t) \cdot \mathrm{BOD}_{5,\mathrm{in,cal}}(t)}{V_{\mathrm{ASC}} \cdot \mathrm{MLSS}_{\mathrm{cal}}(t)} \tag{3}$$

where V_{ASC} is the capacity of ASC, $BOD_{5,in,cal}(t)$ is the BOD_5 value modelled by means of Eq. (2), $Q_{cal}(t)$ is the predicted value of wastewater inflow based on the $Q(t - p)$ measurements (where p means the time gap between the modelled and measured variable values), $MLSS_{cal}(t)$ is the predicted value of MLSS calculated from Eq. (1) or using the combination of Eqs. (1) and (2).

Using the diagram in Fig. 1, the operational parameters of bioreactor can be modelled and adjusted so that the F/M value will be placed in the optimal range even if the measuring devices installed on the wastewater inflow and in the bioreactor are out of order.

To model the wastewater quality indicators and MLSS and F/M values, the methods of SVM, cascade neural networks (CNN) and BT were used. The usefulness of the SVM, CNN and BT methods for forecasting the operation of WWTP has been demonstrated in many studies [6–8,12].

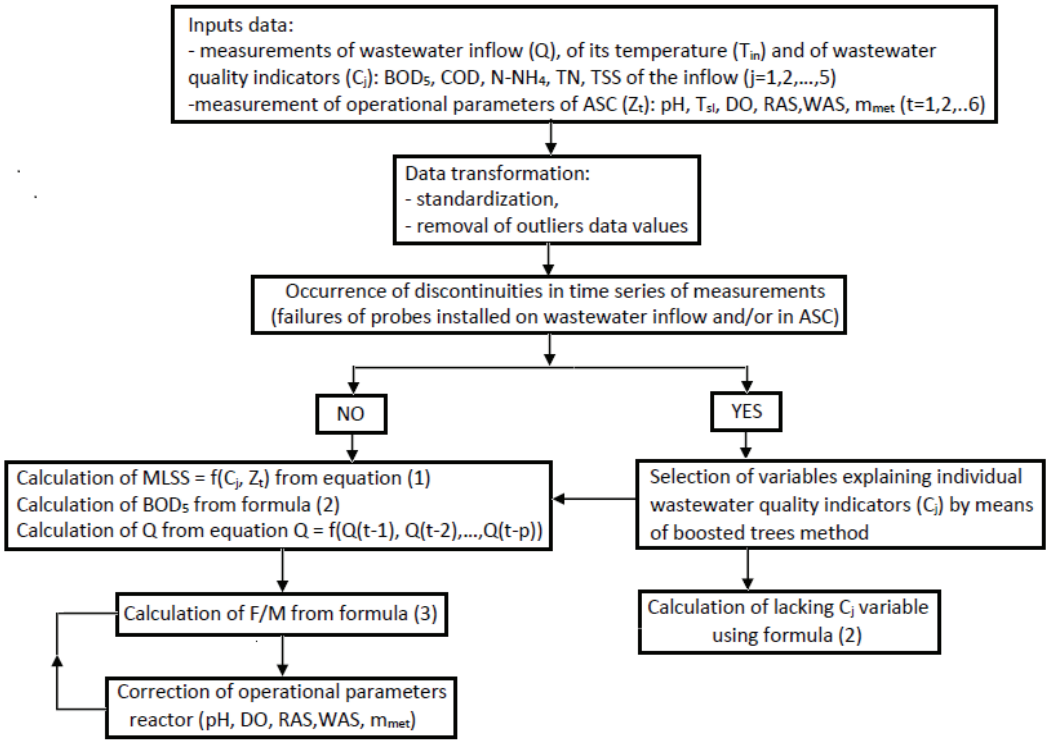

Fig. 1. Diagram of calculation and control of MLSS and F/M values.

In order to properly carry out the learning process and then evaluate the performance of the obtained statistical models, a fivefold cross-validation was performed, dividing the available measurements data into a learning set (75%) and a test set (25%). It was found that the learning and testing data set included separately 250 values of MLSS, F/M and the following wastewater quality indicators: $BOD_{5,in}$, COD_{in}, TN_{in}, $N\text{–}NH_{4,in}^{+}$ and TSS_{in}. To develop the models for forecasting the daily inflow of wastewater a set of data 1,250 values was used. The data for the learning and testing sets were randomly selected. All data have been standardized by means of the min–max transformation before they were used to calculate the models.

To simulate complex environmental processes neural networks are commonly used and then the multilayer perceptron (MLP) are mostly applied [6,7,20,21]. Neural networks of MLP type are usually made up of three layers (input, hidden and output layer). At the model learning stage, for the adopted activation function and the number of neurons on the hidden layer, the weights values for individual neurons are estimated on the basis of some numerical algorithms. Detailed information about the MLP networks and learning algorithms can be found in papers [19,20]. One of MLP modifications is the CNN where some additional connections between the neurons situated on the input layer and the neurons placed on the following layers are created. A cascade network having a higher number of layers than three can be used to model complex nonlinear dependences [22,23]. In the paper, to find the optimal model structure by modelling the variables Q, COD_{in}, $BOD_{5,in}$, $N\text{–}NH_{4,in}^{+}$, TN_{in}, TSS_{in} and MLSS the number of neurons on the single hidden layer has been changed from 3 up to 2·S + 1, where S is the number of model inputs [24,25]; the activation functions between the input and hidden layer were linear, exponential, sinusoidal, sigmoidal and tangent-hyperbolic and they have been taken subsequently; the models calculated were assessed by means of the standard quality measures like mean absolute error (MAE), mean absolute percentage error (MAPE) and correlation coefficient (R). Based on the papers [26,27] concerning the CNN, three additional connections between the input neurons and the following network layers have been defined. The networks calculated have got two hidden layers [26]. To learn the models the Broyden–Goldfarb–Shanno algorithm was used [22]. The CNN models forecasting the amount of inflowing wastewater, the wastewater quality indicators and parameters of the biological reactor have been developed using the MATLAB program (toolbox Neural Network). For the adopted activation function the number of neurons on two hidden layers has been changed by trial and error method until the maximum R-value and the minimum MAE and MAPE values were obtained.

In the following method of SVM there is admissible that the relations between the model output and the input variables are nonlinear; then a nonlinear transformation of n-dimensional space of input variables into K-dimensional space of variable features using a kernel function is executed (where $K > n$). The SVM networks do not show the typical drawback of MLP networks what is a frequent case of breaking the learning modelling stage in one of the local minima of the minimized criteria function; in SVM method a special learning algorithm developed by Vapnik [28] is used to improve the calculation features of the network. In that method the prediction abilities of the obtained

models depend on the values of the following parameters: capacity (C), kernel function (γ) and threshold of insensitivity (ε) [26,28,29]. In the paper to forecast the wastewater inflow, MLSS and the wastewater quality indicators the regression SVM method with a radial kernel function has been used [12].

A significantly simpler than CNN and SVM method is the BT method being an implementation of the method of stochastic gradient strengthening [30]. The main idea of the method consists in a creation of a sequence of regression trees where by means of each of the following tree the rests generated by the preceding tree are calculated. A clear advantage of the method are relatively simple structures of the models developed and as a result lower numbers of model parameters compared with CNN and SVM models. The choice of the trees number (N) in the models developed was done by a successive approximation and by taking different tree numbers, so the limit of N = 200 trees was not exceeded in order to avoid a model overlearning.

In the case of SVM and BT models forecasting the quantity and quality of the wastewater inflow and the parameters of the biological reactor, the parameters C and γ were changed by means of the trial-and-error method using STATISTICA program (toolbox Data Mining) until the assumed modelling effects as in the CNN models was achieved.

4. Results

Based on the measurements concerning the wastewater inflow and the wastewater quality indicators and the parameters of the WWTP bioreactor the ranges of variation of those variables have been calculated (Table 1).

It is easy to notice that the pollution loads flowing to the ASC are highly variable what is followed by a large variation of the bioreactor parameters including MLSS and F/M values. Using the wastewater inflow, the wastewater quality indicators and bioreactor parameters, the statistical models to forecast the MLSS have been calculated. In those models the possibility of modelling those variables by using the inflow loads of organic compounds, suspended solids and the nitrogen and the ASC parameters (RAS, WAS, pH, T_{sl}, DO and m_{met}) were considered.

To predict the wastewater quality, expressed by the values of BOD, COD, TSS, TN and $N–NH_4$, and to determine MLSS and F/M, several various methods have been already applied (Table 2). In contrast to those activities, the application of data mining methods such as CNN, BT and SVM to model quantity and quality indicators of the wastewater has not been fully examined so far. It confirms therefore an innovative character of our current approach.

In Table 3, the results of MLSS calculation using the methods of SVM, CNN and BT have been listed. To calculate

Table 1
Variation ranges of parameters describing the wastewater inflow, the wastewater quality indicators and bioreactor parameters [9,20]

Variables	Minimum	Average	Maximum	Standard deviation
Q, m^3/d	32,564	40,698	86,592	8,088
T_{in}, °C	8.40	16.60	20.90	2.64
T_{sl}, °C	10.00	15.90	23.00	3.58
pH	7.20	7.60	7.80	0.20
MLSS, mg/L	2,010	4,260	6,520	1,040
RAS, %	44.60	90.70	167.60	23.71
m_{met}, m^3/d	0.00	1.35	4.56	1.00
WAS, kg MLSS/d	3,489	11,123	19,194	3,950
DO, mg/L	0.55	2.56	5.78	1.03
F/M, g BOD_5/g MLSS·d	0.03	0.07	0.15	0.02
HRT, d	0.85	1.98	3.54	0.32
SRT, d	10.00	16.25	22.35	5.12
$BOD_{5,in}$, mgO_2/L	127	309	557	86
$BOD_{5,eff}$, mgO_2/L	2.00	4.59	10.00	1.35
COD_{in}, mgO_2/L	384	791	1250	174
COD_{eff}, mgO_2/L	23.00	35.4	50.10	6.35
TSS_{in}, mg/L	126	329	572	77
TSS_{eff}, mg/L	1.00	5.60	14.00	4.79
$N–NH_{4,in}^+$, mg/L	24.40	37.8	65.90	7.10
$N–NH_{4,eff}^+$, mg/L	0.20	2.14	9.25	3.77
TN_{in}, mg/L	33.91	77.73	124.09	10.62
TN_{eff}, mg/L	4.38	7.38	20.10	3.47

HRT, hydraulic retention time; SRT, sludge retention time.

Table 2
Currently applied methods to predict selected wastewater quality parameters ($BOD_{5,in}$, COD_{in}, TSS_{in}, TN_{in} and $N–NH_{4,in}^{+}$) and the reactor operational parameters (MLSS, F/M)

Variable	Method
$BOD_{5,in}$	MLP [31,33], MLR [16,31], MARS [15], RF [15], RF + SOM [15]
COD_{in}	k-NN [32], MARS [15], MLP [31,33], MLR [16,31]
TSS_{in}	k-NN [18,32], MARS [15,18,32], MLP [18,32], RF + SOM [15], RF [15,18], SVM [18,32]
TN_{in}	k-NN [32], MARS [15], RF + SOM [15]
$N–NH_{4,in}$	k-NN [32], MARS [15], RF [15], RF + SOM [15]
MLSS	GP [10], MLP [10,11,14,34], MLR [10,14]
F/M	MLR [14], MLP [14]

MLP, multilayer perceptron; MLR, multilinear regression; GP, genetic programming; k-NN, k-nearest neighbour method; RF, random forest method; SVM, support vectors machines; SOM, self-organizing map; MARS, multivariate adaptive regression splines.

Table 3
Calculation results concerning the SVM, BT and CNN models forecasting the MLSS described by Eq. (1)

Model	Variables	CNN			SVM			BT		
		MAE (mg/L)	MAPE (%)	*R*	MAE (mg/L)	MAPE (%)	*R*	MAE (mg/L)	MAPE (%)	*R*
1	1,2	839	19.86	0.37	889	21.85	0.28	892	21.53	0.11
2	1,2,3	775	18.72	0.55	840	19.96	0.46	845	20.25	0.34
3	1,2,3,4	728	17.33	0.65	795	18.27	0.57	816	19.89	0.40
4	1,2,3,4,5	685	16.39	0.68	745	17.74	0.59	814	19.40	0.45
5	1,2,3,4,5,6	637	15.10	0.73	725	18.25	0.58	783	19.05	0.49
6	1,2,3,4,5,6,7	598	14.13	0.80	685	17.93	0.64	735	17.75	0.58
7	1,2,3,4,5,6,7,8	552	12.86	0.82	640	15.05	0.72	696	17.21	0.66
8	1,2,3,4,5,6,7,8,9	500	11.73	0.87	590	14.04	0.77	657	16.00	0.67
9	1,2,3,4,5,6,7,8,9,10	445	10.52	0.89	553	13.36	0.79	616	15.18	0.74
10	1,2,3,4,5,6,7,8,9,10,11	373	8.75	0.92	485	12.59	0.82	565	13.98	0.77
11	1,2,3,4,5,6,7,8,9,10,11,12	291	6.89	0.95	395	9.62	0.89	495	12.15	0.83

1, *Q*; 2, COD_{in}; 3, $BOD_{5,in}$; 4, $N–NH_{4,in}^{+}$; 5, TN_{in}; 6, TSS_{in}; 7, T_{sl}; 8, pH; 9, m_{met}; 10, DO; 11, RAS; 12, WAS.

the models using SVM method the values for its characteristic parameters *C* and γ have been taken from the ranges 5÷12 and 0.25÷0.65, respectively. In the models numbered from 1 to 3 and from 4 to 6 the neuron numbers on the singular hidden layers were 5 and 6, respectively, and in the models from 7 to 8 and from 9 to 10 the relevant neuron numbers were 8 and 9. By the BT models the number of trees created changed between *N* = 80÷100 what resulted that the models received were not overlearned. Table 3 also shows that the best prediction values of MLSS assessed by means of MAE and MAPE values were received by CNN method and the worst MLSS prediction has been obtained by BT method. The received results show that the prediction of MLSS has been essentially impacted by the COD_{in} load (L_{COD}) that is present in all models analyzed. The conclusion is also confirmed by the investigations of communal treatment plants performed by Hong and Bhamidimarri [10] and Güçlü and Dursun [11].

On the other hand, when COD_{in} is the only predictor parameter, the modelling leads to the models with the worst MLSS prediction, that is, with the highest values of MAE and MAPE errors. If some additional predictors such as $BOD_{5,in}$, TSS_{in}, TN_{in} and $N–NH_{4,in}^{+}$ are considered, then the ability of MLSS prediction improves essentially and the error values regarding, for example, MAE decline by 29% in CNN models, by 23% in SVM models and by 14% in BT models. This observation confirms substantial influence of those variables on the exactness of MLSS prediction. Table 3 also shows that the errors of MLSS prediction obtain their minimal value when the wastewater inflow, the wastewater quality indicators and operational parameters of ASC (pH, T_{sl}, WAS, RAS, DO and m_{met}) are considered as the explanatory variables. This observation is supported by the calculations of Güçlü and Dursun [11] who developed a MLSS model of satisfied prediction ability (*R* = 0.88) using MLP method and based on *Q*, COD_{in}, TSS_{in}, L_{COD}, TKN_{in} and DO as the model inputs. The data listed in Table 3 are also confirmed by the simulation results done by Hong and Bhamidimarri [10] who elaborated accurate MLSS models (*R* = 0.91÷0.92) by means of MLP

method and taking into account the ASC parameters (pH, RAS and MLSS(t – 1)), the wastewater inflow and rainfall depth as the input variables.

In order to model the wastewater quality indicators from Eq. (1) ($BOD_{5,in}$, COD_{in}, TSS_{in}, TN_{in} and $N–NH_{4,in}^{+}$) the relevant explanatory values displayed in Eq. (2) have been identified by means of BT method. Additionally, the predictors describing the daily wastewater inflow (Q) were defined and the importance indicators (IMP) have been calculated for all variables concerned (Table 4). Table 4 shows that the IMP values are changing in wide range (IMP = 0.56÷1.00). The greatest range occurs for COD_{in} variable and the lowest one for $BOD_{5,in}$. The results listed in Table 4 show that the amount and temperature of the wastewater inflow have got an essential impact on the values of the concerned wastewater quality indicators. Moreover, the Mann Kendall test calculation at the level of statistical significance $p = 0.05$ shows no trends in the time series data what makes them steady-state.

Those results are confirmed by other calculations done by Jurik et al. [16], Rousseau et al. [17] and Szeląg et al. [15]. For further modelling analyses the predictor variables attributed to the calculated IMP values higher than 0.90 have been selected [18,19]. To forecast the daily wastewater inflow to WWTP, the values of Q measurements in three following time steps, that is, of $Q(t-1)$, $Q(t-2)$ and $Q(t-3)$, were taken into account. Based on the results achieved (Table 4) the models predicting the wastewater quality indicators have been developed by means of CNN, SVM and BT methods. The modelling results received are shown in Table 5 while the results concerning the inflow prediction are listed in Table 6.

In the models predicting the wastewater quality indicators (Table 5) and developed by SVM method the used C value was in the range of 8–15 and the γ value was between 0.64 and 1.00. The number of neurons in individual hidden layers in the models received by CNN method was between 7 and 10 and the activation function applied mostly in those models was the tangent-hyperbolic function. The number of trees in the models calculated by BT method was between 70 and 105 what assures that the models have not been overlearned.

In the model predicting inflow Q received by SVM method the C value applied was 7 and in the model received by CNN method the neurons number on the singular hidden layer was 4 while the activation function used was sigmoidal. For BT model the number of the trees concerned did not exceed $N = 100$.

The modelling results concerning the TSS_{in} prediction by CNN method (Table 5) are slightly worse than the similar results obtained by Verma et al. [18] ($R = 0.91$) who calculated the COD_{in} concentration and the wastewater inflow Q. By the models predicting COD_{in} while using CNN method the calculated R value is very similar to the results achieved by Abyaneh [31] ($R = 0.81$) who used the MLP method for this purpose. However, the MAPE value received in our COD_{in} modelling is higher than that the one obtained by Minsoo et al. [32] who applied the k-NN method (MAPE = 7.35%). On the other hand, the R value received here by BOD_5 prediction while using CNN method is identical with the result obtained by Dogan et al. [33] by means of MLP method and basing on the measurements data of Q, SS_{in}, TN_{in} and TP_{in}. Comparing the results of TN_{in} modelling computed by Minsoo et al. [32] for dry weather using k-NN method it is easy to notice that they are slightly better (MAPE = 4.54%) than those once received by us using CNN method. The values of R and MAPE (Table 4) computed by the models predicting COD_{in}, $BOD_{5,in}$, TN_{in}, $N–NH_{4,in}^{+}$ and TSS_{in} while using CNN method are only marginally different from the results received by other investigators. It should be pointed out that Q and T_{in} variables have only been used in our analyses as the model predictors. That confirms a very high ability of CNN to model the wastewater quality indicators with limited access to the explained variables.

Table 4
Calculated importance indicators of predictors explaining the wastewater quality indicators

$BOD_{5,in}$		COD_{in}		TSS_{in}		$N–NH_{4,in}^{+}$		TN_{in}		Q_{in}	
Variable	IMP	Variable	IMP	Variable	IMP	Variable	IMP	Variable	IMP	Variable	IMP
$Q(t-1)$	1.00	$Q(t-1)$	1.00	$Q(t-1)$	1.00	$Q(t-1)$	1.00	$Q(t-1)$	1.00	$Q(t-1)$	1.00
$Q(t-2)$	0.97	$Q(t-2)$	0.94	$Q(t-6)$	0.97	$Q(t-2)$	0.98	$Q(t-2)$	0.96	$Q(t-2)$	0.93
$T(t-3)$	0.95	$Q(t-6)$	0.93	$Q(t-5)$	0.95	$Q(t-3)$	0.94	$Q(t-3)$	0.95	$Q(t-3)$	0.92
$Q(t-5)$	0.94	$Q(t-3)$	0.92	$Q(t-4)$	0.94	$T_{in}(t-1)$	0.93	$T_{in}(t-3)$	0.94	$Q(t-4)$	0.78
$Q(t-6)$	0.93	$Q(t-4)$	0.92	$T_{in}(t-3)$	0.93	$T_{in}(t-2)$	0.92	$T_{in}(t-2)$	0.93	$Q(t-5)$	0.69
$T_{in}(t-7)$	0.92	$Q(t-5)$	0.92	$Q(t-3)$	0.92	$T_{in}(t-3)$	0.91	$Q(t-6)$	0.92	$Q(t-6)$	0.67
$T_{in}(t-1)$	0.92	$Q(t-7)$	0.91	$T_{in}(t-1)$	0.91	$Q(t-4)$	0.90	$T_{in}(t-1)$	0.91	$Q(t-7)$	0.66
$T_{in}(t-5)$	0.91	$T_{in}(t-7)$	0.90	$Q(t-2)$	0.90	$Q(t-5)$	0.90	$Q(t-7)$	0.90	$Q(t-8)$	0.66
$Q(t-4)$	0.90	$T_{in}(t-3)$	0.74	$Q(t-7)$	0.81	$Q(t-6)$	0.81	$T_{in}(t-6)$	0.73	$Q(t-11)$	0.57
$Q(t-2)$	0.82	$T_{in}(t-6)$	0.71	$T_{in}(t-5)$	0.08	$T_{in}(t-4)$	0.80	$Q(t-4)$	0.66	$Q(t-12)$	0.55
$T(t-6)$	0.81	$T_{in}(t-2)$	0.07	$T_{in}(t-4)$	0.78	$T_{in}(t-5)$	0.80	$T_{in}(t-4)$	0.65	$Q(t-9)$	0.53
$Q(t-7)$	0.79	$T_{in}(t-4)$	0.68	$T_{in}(t-6)$	0.74	$T_{in}(t-6)$	0.77	$Q(t-5)$	0.64	$Q(t-14)$	0.50
$T_{in}(t-4)$	0.75	$T_{in}(t-1)$	0.58	$T_{in}(t-2)$	0.73	$T_{in}(t-7)$	0.76	$T_{in}(t-5)$	0.63	$Q(t-10)$	0.48
$T_{in}(t-2)$	0.72	$T_{in}(t-5)$	0.56	$T_{in}(t-7)$	0.66	$Q(t-7)$	0.75	$T_{in}(t-7)$	0.62	$Q(t-13)$	0.47

The calculation results of quality indicators obtained by means of SVM, CNN and BT methods based on Eq. (2) were included into Eq. (1) and thus MLSS value may be appointed (Table 7). The quality indicators values of Q, $BOD_{5,in}$ and MLSS obtained while using the mentioned methods were substituted in Eq. (3) and the F/M was then determined (Table 8). The calculation results regarding MLSS and F/M for different combinations of variables (whose numbering is explained in Table 2) are shown in Tables 7 and 8, respectively. Comparing the data from Tables 7 and 8 one can say that higher errors of matching MLSS measurements with calculation results were observed for the model in which the wastewater quality indicators were computed based on the measurements of temperature and wastewater inflow to WWTP. The modelling results listed in Table 7 compared with the data from Table 2 show that the prediction errors regarding the wastewater quality indicators influence vitally the exactness of MLSS prediction. The highest errors of MLSS prediction computed by CNN, SVM and BT methods (Table 7) were received for the models in which the predictors were represented by the wastewater inflow Q and the values of $BOD_{5,in}$ and COD_{in}. Whereas the lowest prediction errors of MLSS (Table 7) were received, Table 2, using as the predictors the variables describing the inflow Q, the wastewater quality ($BOD_{5,in}$, COD_{in}, TSS_{in}, TN_{in} and $N–NH_{4,in}^{+}$) and the bioreactor parameters (pH, DO, m_{met}, RAS and WAS). Making a comparison between the results of MLSS modelling shown in Table 6 and the results from other papers, for example, Hong and Bhamidimarri [10], Güçlü and Dursun [11] and Rustum [34] should be noted that the prediction errors achieved here are a little higher but it does not exclude a practical application of the models.

Table 5
Results of modelling the wastewater quality indicators ($BOD_{5,in}$, COD_{in}, TSS_{in}, $N–NH_{4,in}^{+}$ and TN_{in}) by means of SVM, CNN and BT methods

Quality indicators	CNN			SVM			BT		
	MAE (mg/L)	MAPE (%)	R	MAE (mg/L)	MAPE (%)	R	MAE (mg/L)	MAPE (%)	R
$BOD_{5,in}$	25.88	8.88	0.92	42.88	14.94	0.78	52.29	18.50	0.63
COD_{in}	67.31	9.29	0.82	89.77	12.62	0.74	99.93	13.22	0.60
$N–NH_{4,in}^{+}$	2.19	4.51	0.90	4.00	8.25	0.66	4.44	9.24	0.58
TSS_{in}	25.41	8.47	0.90	39.03	13.27	0.78	46.92	16.33	0.63
TN_{in}	3.73	4.75	0.88	5.74	7.42	0.66	6.25	8.23	0.59

Table 6
Results of modelling the daily wastewater inflow Q by means of SVM, CNN and BT methods

CNN			SVM			BT		
MAE (m^3/d)	MAPE (%)	R	MAE (m^3/d)	MAPE (%)	R	MAE (m^3/d)	MAPE (%)	R
2,541	2.54	0.86	2885	6.84	0.75	3129	7.46	0.70

Table 7
Results of MLSS calculation based on the combinations of Eqs. (1) and (2)

Variables	CNN			SVM			BT		
	MAE (mg/L)	MAPE (%)	R	MAE (mg/L)	MAPE (%)	R	MAE (mg/L)	MAPE (%)	R
1,2	836	19.72	0.42	893	21.66	0.17	926	22.40	0.08
1,2,3	880	21.08	0.15	924	22.29	0.04	961	23.17	0.03
1,2,3,4	826	20.15	0.37	872	21.74	0.30	932	22.27	0.16
1,2,3,4,5	770	19.31	0.52	842	20.84	0.35	891	21.90	0.31
1,2,3,4,5,6	721	17.65	0.65	832	20.33	0.35	865	21.19	0.36
1,2,3,4,5,6,7	665	16.56	0.72	798	19.23	0.42	840	20.39	0.40
1,2,3,4,5,6,7,8	615	16.16	0.70	791	19.26	0.46	830	20.15	0.47
1,2,3,4,5,6,7,8,9	570	15.90	0.78	768	19.31	0.52	784	19.60	0.53
1,2,3,4,5,6,7,8,9,10	520	14.02	0.79	730	19.05	0.54	769	18.96	0.56
1,2,3,4,5,6,7,8,9,10,11	450	12.35	0.83	640	17.90	0.67	743	18.24	0.58
1,2,3,4,5,6,7,8,9,10,11,12	363	8.32	0.93	530	13.50	0.84	758	17.40	0.63

Table 8
Results of F/M calculation by means of Eq. (3)

Variables	CNN			SVM			BT		
	MAE	MAPE (%)	*R*	MAE	MAPE (%)	*R*	MAE	MAPE (%)	*R*
1,2	0.0145	21.38	0.66	0.0181	26.92	0.28	0.0181	28.13	0.18
1,2,3	0.0160	23.20	0.57	0.0183	27.20	0.24	0.022	30.29	0.05
1,2,3,4	0.0153	22.12	0.63	0.018	26.56	0.26	0.019	29.78	0.08
1,2,3,4,5	0.0134	19.52	0.71	0.0179	26.29	0.25	0.0195	29.29	0.10
1,2,3,4,5,6	0.0132	18.89	0.71	0.0176	25.72	0.28	0.0193	28.37	0.16
1,2,3,4,5,6,7	0.0126	18.13	0.75	0.0171	25.10	0.31	0.0184	27.37	0.19
1,2,3,4,5,6,7,8	0.0117	16.74	0.78	0.0176	25.92	0.26	0.0194	28.50	0.15
1,2,3,4,5,6,7,8,9	0.0119	17.11	0.77	0.0177	25.90	0.29	0.0180	28.25	0.23
1,2,3,4,5,6,7,8,9,10	0.0112	16.01	0.80	0.0166	24.27	0.40	0.0175	26.90	0.32
1,2,3,4,5,6,7,8,9,10,11	0.0108	15.95	0.81	0.0133	19.28	0.67	0.0173	27.04	0.32
1,2,3,4,5,6,7,8,9,10,11,12	0.0087	13.35	0.86	0.0118	16.58	0.81	0.0171	25.43	0.46

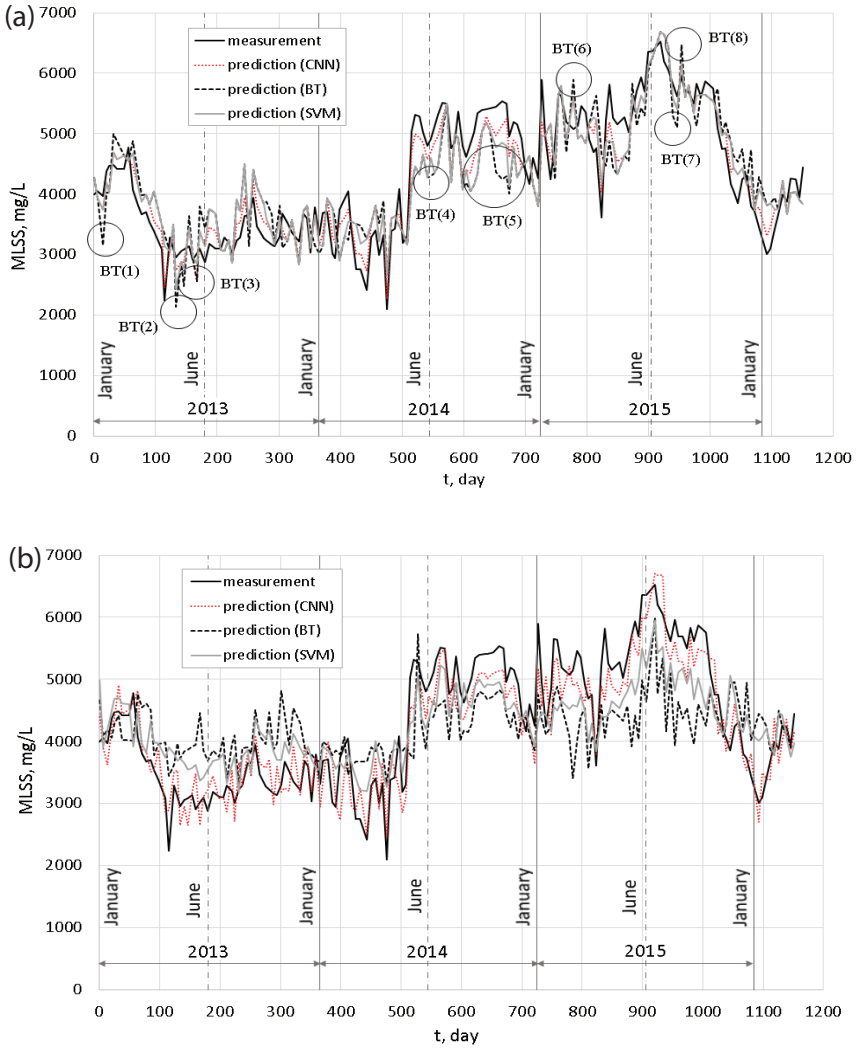

Fig. 2. Comparison of the measurements data with the results of MLSS calculation done by CNN, SVM and BT methods while using Eq. (1): (a), or Eqs. (1) and (2): (b).

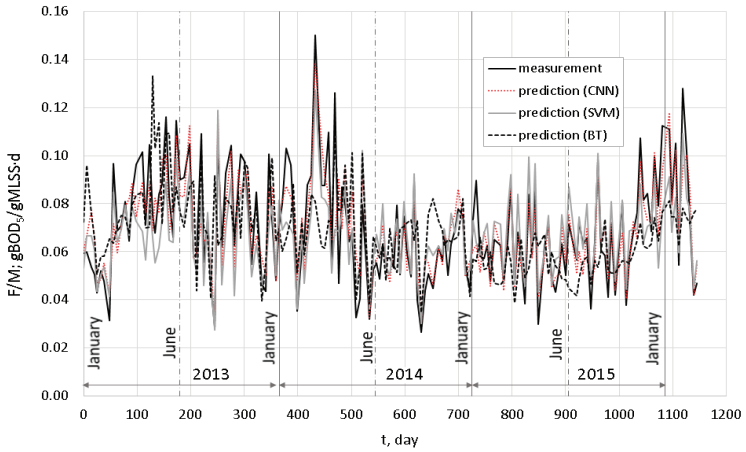

Fig. 3. Comparison of the measurements data with the results of F/M modelling by means of CNN, SVM and BT methods.

Looking at the data in Table 7, it can be concluded that the lowest errors of F/M prediction were obtained by CNN method whereas the highest one by BT method is compatible with the results of *Q* and MLSS modelling and with the modelling results concerning the wastewater quality indicators.

As for the MLSS prediction, the lowest prediction errors of F/M have been obtained with the model in which the wastewater quality indicators ($BOD_{5,in}$, COD_{in}, TSS_{in}, TN_{in} and $N–NH_{4,in}^{+}$) and the ASC parameters have been taken into account as the predictors. In order to visualize received modelling results they are shown graphically in Figs. 2 and 3, where the results of MLSS and F/M simulation carried out by the models computed are reflected.

The MLSS values predicted by means of BT method are in many cases overestimated or underestimated compared with the measurements that could be affected by inapplicable technological decisions by WWTP operation – such exemplary cases are marked in Fig. 2(a) as BT(u) for u = 1–8. Much better fitting of simulation results to the measurements data was received using SVM and CNN methods. Moreover, based on the results shown in Fig. 2(b), it can be noted that MLSS values modelled only by CNN method are mostly relevant to the measurements data. On the contrary, the MLSS values received by SVM and BT methods up to day number 500 are overestimated and after that they are mostly underestimated. Those results confirm an essential impact of prediction errors of singular wastewater quality indicators on the calculated MLSS values.

Based on the data reproduced in Fig. 3, it can be concluded that F/M values predicted by BT method are underestimated whereas the values computed by SVM method are in many cases overestimated. The results of F/M simulation computed by CNN method are comparable with the measurements data what creates the possibility to apply the elaborated model to control and monitor the F/M parameter in case of failures of probes measuring the wastewater quality indicators.

5. Conclusions

It is clear that the modelling methods of CNN, SVM and BT can be successfully applied to model wastewater quality indicators as well as MLSS. The best matching of the calculation results to the measurements data was obtained by the method of CNN and the highest error values were obtained by the method of BT. Furthermore, it should be pointed out that particular wastewater quality indicators can be predicted based on the degree of wastewater dilution that is determined by the amount of wastewater inflow and its temperature; those parameters determine the rate of biochemical processes occurring in the wastewater.

Some limitations in respect of the modelling methodology presented in the paper may refer to separate sewer systems of a small scale. In general, the possibility of prediction of the wastewater quality exclusively on the base of the wastewater inflow and temperature could be widely used in engineering practice. It enables to substitute the measured values of the wastewater quality indicators by corresponding calculation data in the models forecasting the MLSS or F/M what has been demonstrated in the paper. The lowest prediction errors while modelling the MLSS and F/M variables were received for the models in which the explained variables were the inflow of the wastewater, its quality indicators and the operational parameters of ASC. In turn, the largest prediction errors of MLSS and F/M were obtained by the models in which the pollution loads BOD_5 and COD have been taken into account as the predictors.

The presented models could be successfully applied in the practice to reduce the extent of measurements of the wastewater quality indicators and bioreactor parameters which are commonly used to conduct the operation of WWTP. That is operationally important because a sustained monitoring and control of MLSS and F/M values determines the effectiveness of ASC operation and by using the models those actions can be carried out even when the relevant measuring probes are defected or when some technical problems arise while measuring the wastewater quality indicators.

Taking into account the modelling results obtained some further analyses are to be done to assess the possibility of optimization of WWTP operation, regarding the improvement of the quality of the wastewater outflow and the settings of the bioreactor parameters with limited access to the wastewater inflow measurements.

Acknowledgement

The authors would like to thank Dr. Krzysztof Filipek for correcting the English language of this publication.

References

[1] K. Barbusiński, H. Kościelniak, Influence of substrate loading intensity on floc size in activated sludge process, Water Res., 29 (1995) 1703–1710.
[2] H. Chua, P.H. Yu, S.N. Sin, K.N. Tan, Effect of food: microorganism ratio in activated sludge foam control, Appl. Biochem. Biotechnol., 84–86 (2000) 1127–1135.
[3] M. Henze, P. Harremoes, E. Arvin, J. Lacour, Wastewater Treatment, Biological and Chemical Processes, Springer-Verlag, Berlin, 2002.
[4] K.-U. Do, R.J. Banu, D.-H. Son, I.-T. Yeom, Influence of ferrous sulphate on thermochemical sludge disintegration and on performance of wastewater treatment in an anoxic/oxic MBR, Biochem. Eng. J., 66 (2012) 20–26.
[5] K.-U. Do, I.-T. Yeom, P. Arulazhagan, J.R. Banu, Effects of sludge pretreatment on sludge reduction in a lab-scale anaerobic/anoxic/oxic system treating domestic wastewater, Int. J. Environ. Sci. Technol., 10 (2013) 495–502.
[6] S.A. Dellana, D. West, Predictive modeling for wastewater applications: linear and nonlinear approaches, Environ. Modell. Software, 24 (2009) 96–106.
[7] H. Liu, M. Huang, C.K. Yoo, A fuzzy neural network-based soft sensor for modeling nutrient removal mechanism in full-scale wastewater treatment system, Desal. Wat. Treat., 51 (2013) 5184–5193.
[8] H. Boztopak, Y. Őzbay, D. Güçlü, M. Küçükhemek, Prediction of sludge volume index bulking using image analysis and neural network at full-scale activated sludge plant, Desal. Wat. Treat., 57 (2016) 17195–17205.
[9] B. Szeląg, P. Siwicki, Application of the Selected Classification Models to the Analysis of the Settling Capacity of the Activated Sludge – Case Study, B. Kaźmierczak, M. Kutyłowska, K. Piekarska, A. Trusz-Zdybek, E3S Web of Conferences, Vol. 17, Boguszów-Gorce, Poland, 2017, pp. 1–7.
[10] Y.S.T. Hong, R. Bhamidimarri, Evolutionary self-organising modelling of a municipal wastewater treatment plant, Water Res., 37 (2003) 1199–1212.
[11] D. Güçlü, Ş. Dursun, Artificial neural network modelling of a large-scale wastewater treatment plant operation, Bioprocess Biosyst. Eng., 33 (2010) 1051–1058.
[12] D. Ribeiro, A. Sanfins, O. Belo, ICDM'13 Proceedings of the 13th International Conference on Advance in Data Mining: Applications and Theoretical Aspects, Wastewater Treatment Plant Performance Prediction with Support Vector Machines, New York, 2013, pp. 99–111.
[13] K.-U. Do, R.J. Banu, S. Kaliappan, Y. Tae, Influence of the thermochemical sludge pretreatment on nitrification of A/O reactor removing phosphorus simultaneous precipitation, Biotechnol. Bioprocess Eng., 18 (2013) 313–320.
[14] K. Yetilmezsoy, Modeling Studies for the Determination of Completely Mixed Activated Sludge Reactor Volume: Steady-State, Empirical and ANN Applications, Q. Ashton, Advance in Machine Learning Research and Application, Atlanta, 2012, pp. 559–589.
[15] B. Szeląg, L. Bartkiewicz, J. Studziński, Black-box forecasting of selected indicator values for influent wastewater quality in municipal treatment plant, Environ. Prot., 38 (2016) 39–46 (in Polish).
[16] L. Jurik, T. Kaletova, M. Sedmakova, P. Balazova, A. Cervenanska, Comparison of service characteristics of two town's WWTP, J. Ecol. Eng., 18 (2017) 61–67.
[17] D. Rousseau, F. Verdanck, D. Moerman, R. Carrette, C. Thoeye, J. Meirlaen, P.A. Venrolleghem, Development of a risk assessment based technique for design/retroffing WWTP, Water Sci. Technol., 43 (2001) 287–294.
[18] A. Verma, X. Wei, A. Kusiak, Predicting the total suspended solids in wastewater: a data-mining approach, Eng. Appl. Artif. Intell., 26 (2013) 1366–1372.
[19] A. Kusiak, H. Zheng, Z. Zhang, Virtual wind speed sensor for wind turbines, J. Energy Eng., 37 (2011) 59–69.
[20] B. Szeląg, J. Gawdzik, Assessment of the effect of wastewater quantity and quality, and sludge parameters on predictive abilities of non-linear models for activated sludge settleability predictions, Pol. Environ. Stud., 26 (2017) 315–322.
[21] L. Bartkiewicz, B. Szeląg, J. Studziński, Impact assessment of input variables and ANN model structure on forecasting wastewater inflow into sewage treatment plants, Environ. Prot., 38 (2016) 29–36 (in Polish).
[22] L. Rutkowski, Artificial Intelligence Methods and Techniques: Computational Intelligence, PWN, Warszawa, 2006 (in Polish).
[23] S. Ossowski, Neural Networks for Information Processing, Publishing House of the Warsaw University of Technology, Warszawa, 2013.
[24] H.R. Maier, A. Jain, G.C. Dandy, K.P. Sudheer, Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions, Environ. Modell. Software, 25 (2010) 891–909.
[25] I. Lou, Y. Zhao, Sludge bulking prediction using principle component regression and artificial neural network, Math. Prob. Eng., 2012 (2012) 1–17.
[26] G. Capizzi, G.L. Sciuto, P. Monforte, C. Napoli, Cascade feed forward neural network-based model for air pollutants evaluation of single monitoring stations in urban areas, Int. J. Electron. Telecommun., 61 (2015) 327–332.
[27] M.S. Al-batah, M.S. Alkhasawneh, L.T. Tay, U.K. Ngah, H.H. Lateh, N.A.M. Isa, Landslide occurrence prediction using trainable cascade forward network and multilayer perceptron, Math. Prob. Eng., 2015 (2015) 1–9.
[28] V. Vapnik, Statistical Learning Theory, John Wiley and Sons, New York, 1998.
[29] C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, U. Fayyad, Knowledge Discovery and Data Mining, Kluwer, 1998, pp. 1–43.
[30] J.H. Friedman, Stochastic gradient boosted, Comput. Stat. Data Anal., 38 (2002) 367–378.
[31] H.Z. Abyaneh, Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters, J. Environ. Health Sci., 12 (2014) 1–8.
[32] K. Minsoo, K. Yejin, K. Hyosoo, P. Wenhua, K. Changwon, Evaluation of the k-nearest neighbour method for forecasting the influent characteristics of wastewater treatment plant, Front. Environ. Sci. Eng., 10 (2016) 299–310.
[33] E. Dogan, A. Ates, E.C. Yilmaz, B. Eren, Application of artificial neural networks to estimate wastewater treatment plant inlet biochemical oxygen demand, Environ. Prog., 27 (2008) 439–446.
[34] R. Rustum, Modelling Activated Sludge Wastewater Treatment Plants Using Artificial Intelligence Techniques (Fuzzy Logic and Neural Networks), Doctor of Philosophy, Heriot, 2009.