Introducing Ensemble Methods to Predict the Performance of Waste Water Treatment Plants (WWTP)

Bharat B. Gulyani and Arshia Fathima

***Abstract*—Optimization and control of waste water treatment plants (WWTP) is an ongoing effort to make the process more efficient and cost-effective. As found in literature, data mining models such as neural networks have been applied to simulate and model various aspects of the plant such as performance, quality parameters and process parameters. In this paper, we introduce bagging model, an ensemble data mining model, to predict the performance of the WWTP. Ensemble models have been shown to stabilize the base classifier used and avoid overfitting the data. Bagging was used to predict the performance of individual units (primary settler and secondary settler) and the global plant performance. The predicted performance of individual units was also used as inputs to predict the global performance thereby enabling good process control via predictive data models. Upon application to the WWTP dataset, it was found that bagging models perform at par or even better than ANN or SVM for the prediction and hence are suitable models that can be implemented for process control of the water treatment plants.**

***Index Terms*—Waste water treatment plant (WWTP), ensemble models, bagging, process control.**

I. Introduction

With the ever-increasing demand for water, research efforts are being made to enhance the water treatment process and designs to enable cost-effective and sustainable technology development for the future. One of the focus areas in water-related research is to cut down costs via the optimization of the waste water treatment plant (WWTP). Studies have been conducted on better operational control and maintenance of the water treatment plants using intelligent process control methods including neural networks. Data mining methods such as neural networks offer the advantage of simulation and modelling of complex and multi-variable dependent behaviors such as those found in water treatment. Some of these studies have focused on the use of Artificial Neural Networks (ANN) for coagulant dosage control [1], [2], simulation and modelling of filtration and osmotic processes [3], [4], and also model for UV-disinfection control [5]. In addition to modelling and simulation of the water treatment plants, few studies have also predicted the performance of these plants [4], by the application of the data mining algorithms based on the effluent quality parameters such as biochemical oxygen demand (BOD), chemical oxygen demand (COD) and suspended solids (SS). This paper illustrates the use of bagging, an ensemble data mining method, to develop a prediction model for the performance of WWTP and compares the stability of such models with ANN.

Manuscript received August 25, 2016; revised February 12, 2017.

Bharat B. Gulyani is with the Department of Chemical Engineering at BITS Pilani, Dubai Campus, Academic City, Dubai 345055, UAE (e-mail: gulyanibb@gmail.com).

Arshia Fathima is with Nanolabs, Alfaisal University, Saudi Arabia (e-mail: arshiafathima92@gmail.com).

II. Literature Review

Data mining methods have been employed to predict the performance of WWTP as a means to have enhanced process control and efficient operation of the plant. The non-linear complex behavior of these plants have also been captured effectively using data mining methods [6]. According to the literature, mostly ANN were used to modeling and simulation of various aspects of the water treatment plants. One such study predicted the long-term membrane fouling in order to capture the effects of influent water quality changes thereby providing for better operational control of the process [7]. Studies have also used ANN as intelligent controls to model and control anaerobic digesters as well as control the chlorination in disinfection process [8].

Besides ANN, some other data mining methods such as fuzzy networks have been used in conjunction with ANN to develop robust water treatment models. An example of such studies is the fuzzy neural network that was developed to control the coagulant addition process for wastewater from paper mill. This model aided the real-time control and optimization of coagulant dosage with excellent efficiency as the error was almost zero [9]. Similar studies have been reported for coagulant dosage with good performance of the ANN [2].

Another focus of neural network models has been to predict the performance of WWTP based on the quality of the influent and effluent water. These ANN models have been developed using algorithms such as back-propagation, Levenberg–Marquardt algorithm and fuzzy models. A recent study used ANN model to predict the WWTP performance in terms of BOD, COD and total suspended solids (TSS). The plant modeled was a sequencing batch reactor and it was found that the ANN model was able to predict the performance with a correlation coefficient of 0.90, hence establishing its potential to simulate the non-linear behavior of the WWTP by data mining models [6]. Another recent study used feed-forward back propagation ANN to model the reverse osmosis units in a wastewater treatment plant. The model was based on a small dataset and described the permeate flow profiles for the reverse osmosis (RO) units with a high correlation coefficients up to 0.99 with minimum error [3].

Other data mining techniques that were used to assess the WWTP performance with regards to organic matter removal

doi: 10.18178/ijesd.2017.8.7.1004

include self-organizing maps (SOM), principal component analysis (PCA), parallel factor analysis (PARAFAC), partial least squares (PLS) and regression techniques which were also used in combination with ANN. It was shown that the best results were given by the combined models namely PARAFAC/PLS and SOM/ANN combination with a correlation coefficient of 0.93 for both models and RMSE values less than 0.6 [10]. In the present paper, we have used bagging that has been employed in other fields such as bioinformatics but has not been widely studied for WWTP performance modeling. The ensemble methods such as bagging have shown to work well with small data and also avoid the problem of over-fitting the data by averaging the results. These models work by combining multiple base classifiers with multiple starting points and averaging their predictions thereby reducing the risk of choosing wrong classifier. As such ensemble methods will help to stabilize base classifiers being used [11]. Based on this merit of ensemble methods, bagging models with ANN and Support Vector Machines (SVM) as base classifiers have been assessed for the prediction of WWTP performance.

III. Methodology

A. Dataset Processing

The waste water treatment dataset was obtained from UCI Machine Learning Repository [12] which was obtained via daily measurement using sensors for the primary and secondary settlers in the plant for 1990-91. The dataset had a total of 38 attributes in addition to the date of measurement. For our purpose, the dates were converted to an attribute called number of days in operation with the earliest date considered as Day 1 (i.e. 1 Jan 1990 as Day 1). This was done to model the performance with respect to time in days to account for measurements that were missing in the time series. Then dataset reduction was done to deal with any missing values for the attributes. All rows with any missing data were removed from the dataset thereby resulting in a dataset with 380 instances. The list of attributes includes pH, conductivity, Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), suspended solids (SS), sediments, volatile suspended solids, local performance of the settlers based on input BOD/COD/SS and the global performance of the plant based on the input BOD/COD/SS.

As the BOD and COD measurements are time-consuming and costly, BOD and COD related attributes were removed if they didn't affect the model performance. The input attributes zinc in flow to plant (attribute #2) was not considered at all for any models. The input BOD and COD to plant (attribute #4 and 5) were not considered for any of the models except for the global performance models. The output attributes BOD and COD to the plant (attribute #24 & 25) were also not considered for any of the models except for the secondary settler performance models for they were shown to drastically improve the model performance as shown in the results section.

B. Bagging Model for Data Mining

An ensemble model, including bagging, random forests and boosting, simultaneously trains multiple base classifiers and averages their results to give the final output for prediction or classification (depending upon the application). Bootstrap Aggregating or bagging is an ensemble method that selects instances by using bootstrap sampling for getting the training and testing sets from feed data. Bootstrap sampling involves sampling 'n' instances 'n' times with replacement. In this way, all the data will be used for training and validating the data giving a generalized model thereby avoiding the issues of errors and overfitting [13]. Hence, ensemble models can be used with ANN or SVM as base classifier for the prediction of WWTP performance as highlighted in this paper thereby enabling better process control.

C. Model Performance Measures

The performance measures used in the present study were Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Relative Absolute Error (RAE) and correlation coefficient.

1) The RMSE is calculated by the following formula:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{N} \frac{(p_i - a_i)}{N}} \quad (1)$$

where p_i is the predicted value for the *ith* instance, a_i is the actual value for the *ith* instance and N is the total number of instances in the given dataset. The smaller the RMSE, the better the performance of the model [14]. The RMSE tends to have a bias towards larger events [15], so other performance measures need to be evaluated for model selection.

2) Mean Absolute Error (MAE) is the average of the absolute values of the difference between the predicted and actual values. It reduces the bias towards large events unlike RMSE. The equation for MAE [15] is:

$$\mathrm{MAE} = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} |a_i - p_i| \quad (2)$$

where p_i is the predicted value for the *ith* instance, a_i is the actual value for the *ith* instance and N is the total number of instances in the given data set.

3) Relative Absolute Error (RAE): It is the relative equivalent of MAE [15] and is given by:

$$\mathrm{RAE} = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} \frac{|a_i - p_i|}{a_i} \quad (3)$$

where p_i is the predicted value for the *ith* instance, a_i is the actual value for the *ith* instance and N is the total number of instances in the given data set.

4) Correlation Coefficient (R^2): It measures the degree of linear relation between two variables. A correlation coefficient of 0 implies no correlation between variables while a value of 1 implies perfect correlation. The correlation coefficient between actual and predicted variables enables us to get the accuracy of the prediction model. This measure is calculated by [14]:

$$R^2 = S_{pa} / \left(\sqrt{S_p S_a} \right) \quad (4)$$

where $\bar{a}$ and $\bar{p}$ are the averages respectively, and

$$S_{pa} = \sum_{i=1}^{N} (p_i - \bar{p})(a_i - \bar{a})(N-1)$$

$$S_p = \frac{\sum_{i=1}^{N} (p_i - \bar{p})^2}{N-1}$$

$$S_a = \frac{\sum_{i=1}^{N} (a_i - \bar{a})^2}{N-1}$$

IV. Results

The data mining models were developed using the open source software Waikato Environment for Knowledge Analysis (Weka) [16]. The models were developed with 10-fold cross validation and default parameters as defined in Weka for ANN (multilayer perceptron), SVM, Bagging with ANN and Bagging with SVM. The only parameter changed was for kernel type in SVM. The kernel was changed from polykernel (default) to normalized polykernel as it was found to enhance performance in all models except for global model performance. Individual prediction models were built from these data mining algorithms to predict the primary and secondary settler performance and the global plant performance. The details on the attributes used for input and output are given in the appendixes. The results obtained for the above mentioned models are discussed below.

A. Prediction of Primary Settler Performance Based on Input BOD

The performance predictions of bagging with ANN were found to be better than that of ANN. Though the correlation coefficients of both these models were same, the RMSE and MAE values were lower for bagging with ANN, showing that bagging stabilizes the ANN, hence lowering the errors. The SVM based models also have higher accuracy (>95%) but their error was higher than those of ANN models. The results are given in Table I.

TABLE I: Model Comparison for Prediction of Primary Settler Performance Based on Input BOD (Attribute # 30)

Data Mining Model	Correlation Coefficient	RMSE	MAE	RAE (%)
ANN	0.99	1.6	0.84	7.05
SVM (Normalized polykernel)	0.95	4.5	2.88	24.05
Bagging with ANN (10 iterations)	0.99	1.4	0.59	4.93
Bagging with SVM (Normalized polykernel &10 iterations)	0.95	4.6	2.96	24.71

B. Prediction of Primary Settler Performance Based on Input SS

The performance predictions of bagging with ANN were found to be better than that of ANN as shown in Table II. The SVM based models had an acceptable accuracy (correlation coefficient of 0.90) but their error was higher than that of ANN models and hence they can't be used for process control.

TABLE II: Model Comparison for Prediction of Primary Settler Performance Based on Input SS (Attribute # 31)

Data Mining Model	Correlation Coefficient	RMSE	MAE	RAE (%)
ANN	0.97	3.15	1.48	14.60
SVM (Normalized polykernel)	0.91	5.57	4.06	39.97
Bagging with ANN (10 iterations)	0.99	2.2	1.14	11.18
Bagging with SVM (Normalized polykernel &10 iterations)	0.92	5.33	3.81	37.52

C. Prediction of Secondary Settler Performance Based on Input BOD and COD

The secondary settler performance was best predicted by ANN model though bagging model also showed similar performance. However, bagging with SVM did not perform as well as SVM even though its RMSE is similar to that of ANN.

TABLE III: Model Comparison for Prediction of Secondary Settler Performance Based on BOD (Attribute # 33) without Using Input Attributes (#24 & 25)

Data Mining Model	Correlation Coefficient	RMSE	MAE	RAE (%)
ANN	0.68	5.6	3.92	85.99
SVM (Normalized polykernel)	0.64	5.39	3.27	71.73
Bagging with ANN (10 iterations)	0.67	5.83	3.88	85.08
Bagging with SVM (Normalized polykernel &10 iterations)	0.59	5.59	3.28	71.8

TABLE IV: Model Comparison for Prediction of Secondary Settler Performance Based on BOD (Attribute # 33 with Model Using Input Attributes (#24 & 25)

Data Mining Model	Correlation Coefficient	RMSE	MAE	RAE (%)
ANN	0.98	1.48	0.79	17.32
SVM (Normalized polykernel)	0.92	3.06	1.05	22.94
Bagging with ANN (10 iterations)	0.99	1.06	0.45	9.94
Bagging with SVM (Normalized polykernel &10 iterations)	0.89	3.37	1.04	22.69

It was also observed (as seen from Tables III - VI) that the prediction of the secondary settler performance drastically improved with the inclusion of the output BOD and COD (attributes # 24 & 25). As the dataset considered was based on 2 settlers, these results confirm the strong dependence of the secondary settler performance on the effluent quality.

TABLE V: MODEL COMPARISON FOR PREDICTION OF SECONDARY SETTLER PERFORMANCE BASED ON COD (ATTRIBUTE # 34) WITHOUT USING INPUT ATTRIBUTES (#24 & 25)

Data Mining Model	Correlation Coefficient	RMSE	MAE	RAE (%)
ANN	0.16	16.74	10.93	141.43
SVM (Normalized polykernel)	0.49	9.23	6.83	88.3
Bagging with ANN (10 iterations)	0.37	11.82	8.54	110.48
Bagging with SVM (Normalized polykernel &10 iterations)	0.49	9.17	6.76	87.51

TABLE VI: MODEL COMPARISON FOR PREDICTION OF SECONDARY SETTLER PERFORMANCE BASED ON COD (ATTRIBUTE # 34) WITH MODEL USING INPUT ATTRIBUTES (#24 & 25)

Data Mining Model	Correlation Coefficient	RMSE	MAE	RAE (%)
ANN	0.99	1.13	0.60	7.80
SVM (Normalized polykernel)	0.97	2.56	1.38	17.85
Bagging with ANN (10 iterations)	0.99	1.22	0.50	6.44
Bagging with SVM (Normalized polykernel &10 iterations)	0.97	2.70	1.38	17.88

D. Prediction of Global Performance

The three global performance attributes from the dataset were based on input BOD, COD, and SS, respectively. These performance attributes were successfully predicted by ANN and bagging with ANN models as shown in Tables VII - IX.

TABLE VII: MODEL COMPARISONS FOR PREDICTION OF GLOBAL PERFORMANCE BASED ON INPUT BOD (ATTRIBUTE # 35)

Data Mining Model	Correlation Coefficient	RMSE	MAE	RAE (%)
ANN	0.98	1.24	0.58	17.97
SVM (normalized kernel)	0.78	3.52	1.80	55.70
SVM (polykernel)	0.95	1.75	0.93	28.79
Bagging with ANN (10 iterations)	0.96	1.70	0.46	14.22
Bagging with SVM with polykernel (10 iterations)	0.95	1.81	0.92	28.37

TABLE VIII: MODEL COMPARISONS FOR PREDICTION OF GLOBAL PERFORMANCE BASED ON INPUT COD (ATTRIBUTE # 36)

Data Mining Model	Correlation Coefficient	RMSE	MAE	RAE (%)
ANN	0.98	1.40	0.75	12.26
SVM (normalized kernel)	0.94	2.78	2.0	32.83
SVM (polykernel)	0.96	2.23	1.32	21.72
Bagging with ANN (10 iterations)	0.99	0.86	0.48	7.88
Bagging with SVM with polykernel (10 iterations)	0.96	2.26	1.34	22.08

The SVM based models also showed similar correlation coefficients like ANN but their RMSE values were much higher thereby indicating lower accuracy. For SVM model, the polykernel performed better than the normalized kernel thereby showing a better fit of global performance data with polynomial function.

TABLE IX: MODEL COMPARISONS FOR PREDICTION OF GLOBAL PERFORMANCE BASED ON INPUT SS (ATTRIBUTE # 37)

Data Mining Model	Correlation Coefficient	RMSE	MAE	RAE (%)
ANN	0.98	1.28	0.77	18.66
SVM (normalized kernel)	0.96	1.66	0.93	22.72
SVM (polykernel)	0.96	1.77	1.04	25.25
Bagging with ANN (10 iterations)	0.98	1.12	0.49	12.05
Bagging with SVM with polykernel (10 iterations)	0.96	1.78	1.06	25.83

E. Prediction of Global Performance Using Previously Predicted Individual Performance Values

Besides individual prediction models, models were also built using previously predicted attributes as inputs for the prediction of the global performance of the plant. Building sequential predictors by using previously predicted performance of the individual settlers as inputs for the global performance prediction, we can also develop feedback control on the input streams. For example, if the predicted performance of the secondary settler based on input BOD/COD values is good but the corresponding predicted global performance is low, then controller can adjust the process parameters of the settlers accordingly. For building these models, as ingle algorithm was used throughout the sequential predictors. For example, to predict the global performance based on BOD (attribute #36) using ANN, the corresponding predicted settler performance values were obtained from the individual ANN models. For secondary settler, best predicted values for ANN (from Table IV) were used.

According to the results as given in Table X-XII, it was observed that using predicted performance values did lower the prediction performance, however ANN and bagging with ANN did give acceptable results with an average correlation coefficient of 0.95 for the global performance parameters.

TABLE X: MODEL COMPARISONS FOR PREDICTION OF GLOBAL PERFORMANCE BASED ON INPUT BOD (ATTRIBUTE # 35) USING PREDICTED PERFORMANCE DATA OF SETTLERS AS INPUTS

Data Mining Model	Correlation Coefficient	RMSE	MAE	RAE (%)
ANN	0.94	1.91	1.12	34.8
SVM (polykernel)	0.82	3.19	1.05	32.39
Bagging with ANN (10 iterations)	0.93	2.11	0.65	20.05
Bagging with SVM (polykernel & 10 iterations)	0.75	3.64	1.08	33.45

The SVM and bagging with SVM models were only able to give a correlation coefficient of 0.82 and 0.75 respectively for global performance based on BOD, but were able to give higher correlation coefficients for global performance based on COD and SS. This difference can be explained based on results from Tables I and IV, which show that the predicted outputs based on BOD had larger RMSE values. This error in prediction of performance in individual settlers was carried forward into the global performance prediction models

thereby affecting the accuracy of the models.

TABLE XI: MODEL COMPARISONS FOR PREDICTION OF GLOBAL PERFORMANCE BASED ON INPUT COD (ATTRIBUTE # 36) USING PREDICTED PERFORMANCE DATA OF SETTLERS AS INPUTS

Data Mining Model	Correlation Coefficient	RMSE	MAE	RAE (%)
ANN	0.97	1.99	1.04	17.10
SVM (polykernel)	0.95	2.64	1.51	24.87
Bagging with ANN (10 iterations)	0.99	1.28	0.66	10.9
Bagging with SVM (polykernel & 10 iterations)	0.94	2.88	1.54	25.3

TABLE XII: MODEL COMPARISONS FOR PREDICTION OF GLOBAL PERFORMANCE BASED ON INPUT SS (ATTRIBUTE # 36) USING PREDICTED PERFORMANCE DATA OF SETTLERS AS INPUTS

Data Mining Model	Correlation Coefficient	RMSE	MAE	RAE (%)
ANN	0.97	1.39	0.78	19.03
SVM (polykernel)	0.95	2.0	1.13	27.61
Bagging with ANN (10 iterations)	0.98	1.1	0.5	12.08
Bagging with SVM (polykernel & 10 iterations)	0.95	1.88	1.06	25.72

V. CONCLUSIONS

Data mining algorithms such as ANN and bagging offer the capability for better process control using predicted performance based on input quality parameters that can be easily measured. This provides for a cost-effective, timely and efficient way to operate and maintain the WWTP. In this paper we have introduced bagging, an ensemble model, for accurate predictions of the WWTP performance which has shown to perform at par with neural networks while avoiding overfitting. Global performance prediction models based on previously predicted individual performance values were also developed. These models based on ANN and bagging with ANN also had acceptable prediction capabilities which will enable for enhanced feedback control of the WWTP. A series of predictive models based on ANN or Bagging with ANN have shown to predict the plant performance satisfactorily thereby providing a model for feedback control based on predicted performance. Future optimization studies can be done on using a combination of data mining models to predict the intermediate output/performance parameters and also develop further models based on these intermediate results for global output performance prediction.

APPENDIX

The following appendixes give the details on the input and output attributes used for developing the models. Appendix A gives the attribute list while Appendix B gives attributes used for specific models.

APPENDIX A: ATTRIBUTE DETAILS FROM UCI REPOSITORY

Attribute Number	Attribute Name	Input or Output for Model
1	Input flow to plant	Input
2	Input Zinc to plant	Input
3	Input pH to plant	Input
4	Input BOD to plant	Input
5	Input COD to plant	Input
6	Input SS to plant	Input
7	Input volatile SS to plant	Input
8	Input sediments to plant	Input
9	Input conductivity to plant	Input
10	Input pH to primary settler	Input
11	Input BOD to primary settler	Input
12	Input SS to primary settler	Input
13	Input volatile SS to primary settler	Input
14	Input sediments to primary settler	Input
15	Input conductivity to primary settler	Input
16	Input pH to secondary settler	Input
17	Input BOD to secondary settler	Input
18	Input COD to secondary settler	Input
19	Input SS to secondary settler	Input
20	Input volatile SS to secondary settler	Input
21	Input sediments to secondary settler	Input
22	Input conductivity to secondary settler	Input
23	Output pH of plant	Input
24	Output BOD of plant	Input
25	Output COD of plant of plant	Input
26	Output SS	Input
27	Output volatile SS	Input
28	Output sediments	Input
29	Output conductivity	Input
30	Performance based on input BOD in primary settler	Output
31	Performance based on input SS to primary settler	Output
32	Performance based on input sediments to primary settler	Output
33	Performance based on input BOD to secondary settler	Output
34	Performance based on input COD to secondary settler	Output
35	Global performance based on input BOD	Output
36	Global performance based on input COD	Output
37	Global performance based on input SS	Output
38	Global performance based on input sediments	Output

APPENDIX B: INPUT AND OUTPUT ATTRIBUTE DETAILS FOR MODELS

Model Name	Input Attribute #	Output Attribute #
Prediction of primary settler performance based on input BOD	1, 3, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 26, 27, 28, 29 and number of days in operation	30
Prediction of primary settler performance based on input SS	1, 3, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 26, 27, 28, 29 and number of days in operation	31
Prediction of secondary settler performance based on input BOD	Without O/P BOD & COD: 1, 3, 6, 7, 8, 9, 10, 11, 12, 13, 14,15, 16, 17, 18, 19, 20, 21, 22, 23, 26, 27, 28, 29 and number of days in operation.	33 & 34 (based on BOD & COD respectively)

Model Name	Input Attribute #	Output Attribute #
	With O/P BOD & COD: 1, 3, 6, 7, 8, 9, 10, 11, 12, 13, 14,15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 and number days in operation.	
Prediction of global performance	Based on input BOD: 1, 3, 4,5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 26, 27, 28, 29, 30, 33 and number of days in operation	35, 36 & 37 (based on BOD, COD and SS respectively)
	Based on input COD: 1, 3, 4,5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 26, 27, 28, 29, 30, 34 and number of days in operation	
	Based on input SS: 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 26, 27, 28, 29, 31 and # days in operation.	
Prediction of global performance using previously predicted individual performance values	Based on input BOD - This was carried out with input parameters # 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 26, 27, 28, 29 and # days in operation. Predicted parameters – 30 and 33.	35, 36 & 37 (based on BOD, COD and SS respectively)
	COD based perf- This was carried out with input parameters # 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 26, 27, 28, 29 and # days in operation. Predicted parameters – 30 and 34.	
	SS based perf- This was carried out with input parameters # 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 26, 27, 28, 29 and # days in operation. Predicted parameter- 31.	

REFERENCES

[1] S. Kriti and J. Smita, "Artificial neural network modelling of shyamala water works, Bhopal MP, India: A green approach towards the optimization of water treatment process," *Research Journal of Recent Sciences*, vol. 2, pp. 26–28, 2013.

[2] G.-D. Wu and S.-L. Lo, "Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system," *Engineering Applications of Artificial Intelligence*, vol. 21, no. 8, pp. 1189–1195, Dec. 2008.

[3] A. Salgado-Reyna, E. Soto-Regalado, R. Gómez-González, F. J. Cerino-Córdova, R. B. García-Reyes, M. T. Garza-González, and M. M. Alcalá-Rodríguez, "Artificial neural networks for modeling the reverse osmosis unit in a wastewater pilot treatment plant," *Desalination and Water Treatment*, vol. 53, no. 5, pp. 1177–1187, Jan. 2015.

[4] Y. Zhao, J. S. Taylor, and S. Chellam, "Predicting RO/NF water quality by modified solution diffusion model and artificial neural networks," *Journal of membrane science*, vol. 263, no. 1, pp. 38–46, 2005.

[5] C.-H. Lin, R.-F. Yu, W.-P. Cheng, and C.-R. Liu, "Monitoring and control of UV and UV-TiO2 disinfections for municipal wastewater reclamation using artificial neural networks," *Journal of Hazardous Materials*, vol. 209, pp. 348–354, Mar. 2012.

[6] M. S. Nasr, M. A. E. Moustafa, H. A. E. Seif, and G. E. Kobrosy, "Application of artificial neural network (ANN) for the prediction of EL-AGAMY wastewater treatment plant performance-EGYPT," *Alexandria Engineering Journal*, vol. 51, no. 1, pp. 37–43, Mar. 2012.

[7] G. R. Shetty and S. Chellam, "Predicting membrane fouling during municipal drinking water nanofiltration using artificial neural networks," *Journal of Membrane Science*, vol. 217, no. 1, pp. 69–86, 2003.

[8] M. M. Hamed, M. G. Khalafallah, and E. A. Hassanien, "Prediction of wastewater treatment plant performance using artificial neural networks," *Environmental Modelling & Software*, vol. 19, no. 10, pp. 919–928, Oct. 2004.

[9] H. Mingzhi, Y. Ma, W. Jinquan, and W. Yan, "Simulation of a paper mill wastewater treatment using a fuzzy neural network," *Expert Systems with Applications*, vol. 36, no. 3, Part 1, pp. 5064–5070, Apr. 2009.

[10] M. Bieroza, A. Baker, and J. Bridgeman, "New data mining and calibration approaches to the assessment of water treatment efficiency," *Advances in Engineering Software*, vol. 44, no. 1, pp. 126–135, 2011.

[11] Z. Zheng and B. Padmanabhan, "Constructing Ensembles from Data Envelopment Analysis," *INFORMS Journal on Computing*, vol. 19, no. 4, pp. 486–496, 2007.

[12] UCI machine learning repository: Water treatment plant data set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Water+Treatment+Plant

[13] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, 1st ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005.

[14] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000.

[15] N. D. Bennett, B. F. W. Croke, G. Guariso, J. H. A. Guillaume, S. H. Hamilton, A. J. Jakeman, S. Marsili-Libelli, L. T. H. Newham, J. P. Norton, C. Perrin, S. A. Pierce, B. Robson, R. Seppelt, A. A. Voinov, B. D. Fath, and V. Andreassian, "Characterising performance of environmental models," *Environmental Modelling & Software*, vol. 40, pp. 1–20, Feb. 2013.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorer Newsletters*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

Bharat B. Gulyani had his bachelor, masters and doctoral degrees from University of Roorkee, India (now IIT Roorkee) in the field of chemical engineering. He has more than 20 years of research experience ad had published and presented at various conferences more than 30 research papers. He is currently associate professor at Birla Institute of Technology and Science at their Dubai campus.

Arshia Fathima has received her bachelors in chemical engineering and computer science from BPDC, UAE in 2014 and her masters in chemical engineering with specialization in product development from UC Berkeley, USA in 2015. Currently, she is pursuing research at the Nanolabs, Alfaisal University under Dr. Edreese Alsharaeh. Actively involved in research, she has published 5 papers till date in conferences and journals.