CHEMICAL ENGINEERING TRANSACTIONS
VOL. 106, 2023
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Jeng Shiun Lim, Nor Alafiza Yunus, Peck Loo Kiew, Hon Huin Chin

ISBN 979-12-81206-05-2; **ISSN** 2283-9216
DOI: 10.3303/CET23106117

Integration of Machine Learning Methods for Effluent Quality Prediction in Moving Bed Biofilm Reactor System and Chlorination Wastewater Treatment

Andrei Fryle I. Jaluague, Arnel B. Beltran*, Kathleen B. Aviso

Department of Chemical Engineering, De La Salle University, 2401 Taft Ave, Malate, Manila, Philippines
arnel.beltran@dlsu.edu.ph

Wastewater treatment ensures that effluent water may safely be disposed of in bodies of water. During extreme weather conditions, the influent characteristics (ICs) stray from their typical behavior, affecting the performance of wastewater treatment plants (WWTPs). These points, referred to as anomalies, are discarded when modelling the WWTP performance. Considering this, an effluent quality prediction model, for a chlorination WWTP with moving bed biofilm reactor systems as secondary treatment, was developed. The ICs—flow rate, biological oxygen demand (BOD), chemical oxygen demand (COD), total coliform (TC), pH, and total suspended solids—were analyzed to determine the Mahalanobis distance (MD) and identify anomalies. The classification of MD was used to develop a support vector machine (SVM) model. Artificial Neural Networks (ANNs) were developed for both anomaly and non-anomaly points to predict the pre-chlorination and effluent BOD, COD, and TC. The optimal SVM for anomaly detection, modelled using 266 datapoints, 201 non-anomalies and 65 anomalies, used a fine Gaussian SVM architecture with an area under the receiver operating characteristic curve (AUC-ROC) of 0.90. The developed optimal ANNs exhibited correlation values ranging from 0.863 to 0.972. The generalization ability of the integrated SVM-ANN model was evaluated using a new set of data from the same WWTP. The mean absolute error values for the effluent BOD, COD, and TC value prediction were above average with values of 4.95 ppm, 12.18 ppm, and 26.29 MPN/100 mL. The overall model captures the trend of the test datasets, allowing the accurate forecast of effluent parameters and informing future modifications on WWTP design.

1. Introduction

Wastewater treatment plants (WWTPs) have been designed to improve wastewater quality such that it does not contribute to pollution when discharged to a receiving water body (Ratola et al., 2012). The inward stream is referred to as the influent, while the outgoing stream is its effluent. WWTPs employ treatment strategies to meet effluent standards. Its secondary treatment stage removes organic material from wastewater, while the tertiary treatment prevents human exposure to waterborne microorganisms (Ramírez-Castillo et al., 2015).

Wastewater quality may be measured through biological oxygen demand (BOD), and chemical oxygen demand (COD), which signal the presence of organic matter (Li and Liu, 2019), total suspended solids (TSS), which are dissolved solids coming from erosion, organic, and inorganic matter (Scholz, 2016), and the total coliform (TC) content, which characterizes the sanitary condition of the water supply. Mathematical modelling and machine learning tools have been employed to accurately model the WWTP's effluent characteristics given the quality of its influent stream; but these did not consider differing influent conditions (ICs), caused by environmental factors such as heavy rainfall, seawater intrusion, and water supply shutdown (Hernandez-Ramirez et al., 2019).

These may be addressed by employing black box models, which consider complex and nonlinear relationships of input and output data based on specific algorithms (Mundi et al., 2021). Anomaly detection is used to identify patterns of a dataset, and data points that deviate from the expected behavior. Among these, the computation of the Mahalanobis distance (MD) identifies the typical behavior of a given dataset containing multiple variables. This will be coupled with a two-class support vector machine (SVM) tool for a more robust identification of anomalies within a given dataset (Zhao et al., 2013). This is similar to Khawaga et al.'s (2019) approach of using a situation algorithm and classification algorithm for anomaly detection. When compared with purely

Paper Received: 14 June 2023; Revised: 02 July 2023; Accepted: 18 July 2023
Please cite this article as: Jaluague A.F.I., Beltran A.B., Aviso K.B., 2023, Integration of Machine Learning Methods for Effluent Quality Prediction in Moving Bed Biofilm Reactor System and Chlorination Wastewater Treatment, Chemical Engineering Transactions, 106, 697-702 DOI:10.3303/CET23106117

classification algorithms, these garnered a better performance in their correlation coefficients and area under the receiver operating characteristic curve (AUC-ROC). Artificial Neural Networks (ANN) can predict output values, given a set of input variables by utilizing modelling techniques through historical data. While ANN has been used to model a WWTP's performance, no attempts have been made to reconcile the effects of anomaly ICs to the prediction of effluent wastewater quality.

This study will integrate anomaly detection and ANN techniques to utilize the capacity of anomaly detection to describe the behavior of the WWTP's influent streams and identify points with anomalous behavior. ANN models will be modelled to datasets of anomalies and non-anomalies to identify the typical behavior of the influent stream characteristics. When determining the action plans of WWTP operators, the performance of treatment also varies for different ICs. This model would produce accurate predictions for effluent characteristics and would be useful and easily modelling WWTPs that would inform the improvement of design and specifications of the WWTP (Khawaga et al., 2019).

2. Case Study

This study analyzed a WWTP employing a moving bed biofilm reactor (MBBR) system for the secondary treatment, and chlorination for the tertiary treatment. This type of setup has been widely utilized especially in developing countries because both treatment processes require low capital and maintenance costs, show flexibility for use in multiple tank sizes, and high treatment efficiency (Ali et al., 2021). The MBBR system has 266 datapoints whose characteristics are described in Table 1, which shows the maximum and the minimum values of each parameter considered for this study. A value of any parameter that is outside the range implies that the resulting prediction for effluent BOD, COD, and TC will have a huge error.

Table 1: Data Description for WWTP with MBBR-Chlorination System

Parameter	Mean	Standard Deviation	Skewness	Range	Minimum	Maximum
Influent Flow Rate (m^3/d)	4,985.52	2,281.64	0.04	10,930.00	217.00	11,147.00
Influent Total Coliform (MPN/100 mL)	$2.95x10^8$	7.40×10^8	3.82	5.20×10^9	1.60×10^3	5.20×10^9
Influent BOD (ppm)	151.82	149.31	4.87	1,417.00	8.00	1,425.00
Influent COD (ppm)	320.13	512.51	11.76	7,721.00	13.00	7,734.00
Pre-Chlorination TSS (ppm)	16.83	21.15	3.59	163.50	0.50	164.00
Pre-Chlorination pH	7.17	0.34	-0.47	3.10	5.28	8.38
Pre-Chlorination BOD (ppm)	10.92	11.52	4.49	118.20	1.00	119.20
Pre-Chlorination COD (ppm)	44.80	36.20	3.03	309.00	5.00	314.00
Effluent Total Coliform (MPN/100 mL)	$9.07x10^5$	1.05×10^7	12.20	1.44×10^8	0.00	1.44×10^8

3. Methodology

3.1 Development of Anomaly Detection Framework

The anomaly detection using MD-based Two-Class SVM was used in differentiating the anomaly and non-anomaly ICs. The parameters considered were flow rate, pH, TC, BOD, COD, TSS, and served as the model's independent variables. The anomaly detection was comprised of two parts: the computation of MD, and the training and validation of the two-class SVM. The MD was determined via regression algorithms. The probability that a datapoint is an outlier is determined assuming a chi-square distribution. If the computed probability is less than an arbitrary 5 % level of significance, it will be considered an anomaly. These values were tallied and were noted for each datapoint.

The anomaly and non-anomalies were considered as classes for the two-class SVM; a hyperspace was created using the determined anomalies and non-anomalies using MD to define the constraints of the data that is considered normal. A feature selection process, to determine which parameters are most significant for the SVM, was utilized to improve the performance and efficiency of the SVM algorithm. A training dataset, which constituted 60 % of the dataset, was utilized in building the proposed SVM models. The validation data, which encompassed 30 % of the data, was employed to validate the generalization of the models to decrease the likelihood of overfitting. Finally, the testing dataset, which comprised 10 % of the dataset, was used in the simulation of the whole framework to individually gauge its performance. Different types of SVM, namely linear, quadratic, cubic, fine Gaussian, medium Gaussian, and coarse Gaussian, were tested to see which would fit the dataset the best, determined through the identification of the values for the AUC-ROC.

3.2 Development of Artificial Neural Network Models

The training for the ANN models was performed after the anomaly data and the non-anomaly data were identified using the MD-based SVM. Datasets were split into the training, validation, and testing datasets for all ANNs developed in the study; a ratio of 6:3:1 was employed for training, validation, and testing. Various network architectures were assessed based on the number of neurons per hidden layer, ranging from 7 to 10, and the number of hidden layers, ranging from 1 to 3. The ANN architectures investigated were based on the models utilized by Tümer and Edebali (2015). The characteristics of the ANN models are summarized in Table 2.

Table 2: ANN Model Network Properties

Property of ANN Model	Secondary Treatment	Tertiary Treatment
Network Inputs	Influent flow rate, pH, TC, BOD, COD, TSS	Pre-chlorination flow rate, pH, TC, BOD, COD, TSS
Network Outputs	Pre-chlorination BOD, COD	Effluent TC
Network Type	Feed-forward ANN	
Training Algorithm	Levenberg-Marquardt Backpropagation	
Adaption Algorithm	Gradient descent with momentum weight and bias learning function	
Input Layer Activation Function	Tan-sigmoid, or Log-sigmoid	
Output Layer Activation Function	Pure Linear	

Pre-chlorination BOD and COD were assumed to be equal to effluent levels, because chlorination does not deal with treatment of organic material. The effluent TSS concentration was assumed to be constant throughout, while the influent TC was assumed to be equal to the pre-chlorination TC, because it is not significantly reduced before disinfection (Chiemchaisri et al., 2022).

3.3 Simulation for Overall Framework Evaluation

The optimal anomaly detection and ANN models were assembled as in Figure 1. Because simulation measures the network's capacity to precisely forecast the outputs of new input data, it enables performance evaluation of the chosen optimal network outside of training data. Eighteen data points were pre-selected as input data for the simulation phase at the beginning of the approach. It is significant to note that the chosen datapoints were selected so that each simulation dataset for each case study contains both anomaly and non-anomaly ICs. For each datapoint, the absolute error was determined. The mean absolute error was identified to determine the performance of the overall framework that integrated the MD-based SVM with the ANN modelling.

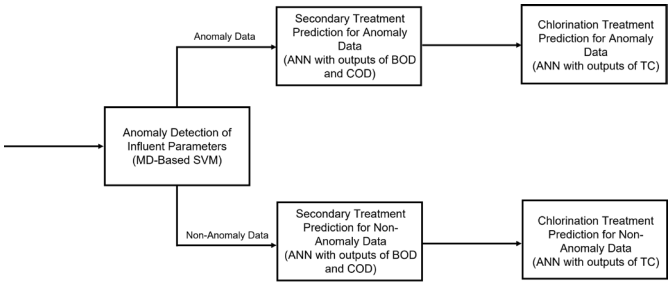

Figure 1: Structure of Study's Overall Model

4. Results and Discussion

Using the computation for MD, 201 datapoints were determined to be non-anomalies, while 65 were anomalies. Using these characteristics as the basis for SVM, its different types were compared in terms of accuracy of the testing and validation datasets, and the AUC-ROC, as seen in Figure 2. The SVM model with the best performance is the fine Gaussian SVM, mainly due to the high AUC-ROC score (0.90), as the fine Gaussian SVM's accuracy scores (training: 0.99, validation: 0.76) were second only to that of the cubic SVM (training: 0.99, validation: 0.79). The performance of SVM shows its potential as a classification algorithm for anomaly detection techniques. The use of SVM in classifying anomalies and non-anomalies in the ICs of a WWTP would be helpful for planning necessary adjustments towards the operating conditions in a WWTP (Waqas et al. 2022). The instances of misclassifications from the SVM model may be attributed to the lack of datapoints considered for this study.

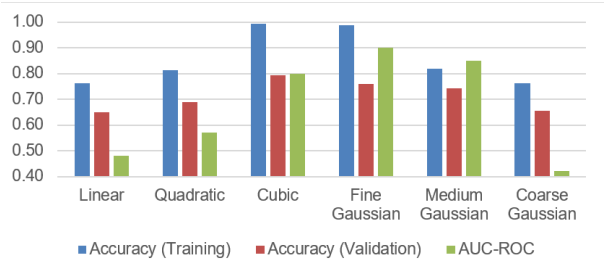

Figure 2: Summary of SVM performance for each type

Two sets of optimized ANNs that will be trained based on the characteristics of the datasets for anomalies and non-anomalies were developed. There were 159 data points used in developing the ANN model for non-anomaly ICs, while 59 datapoints were used in modelling the effluent characteristics for anomaly ICs. A stepwise regression algorithm was utilized to demonstrate the independence of the behavior of the input variables. All input variables considered–influent flow rate, TC, BOD, COD, TSS, and pH–have a significant effect on the effluent BOD, COD, and TC in non-anomaly ICs, and the effluent BOD and TC for anomaly ICs. Only the influent BOD, COD, and TSS were significant for the prediction model for effluent COD for anomaly ICs. A trial-and-error method was employed to compare models with varying architectures. Figure 3 presents a summary of the performance of the network architectures with their corresponding transfer functions for modelling the effluent BOD, COD, and TC for non-anomaly ICs and anomaly ICs.

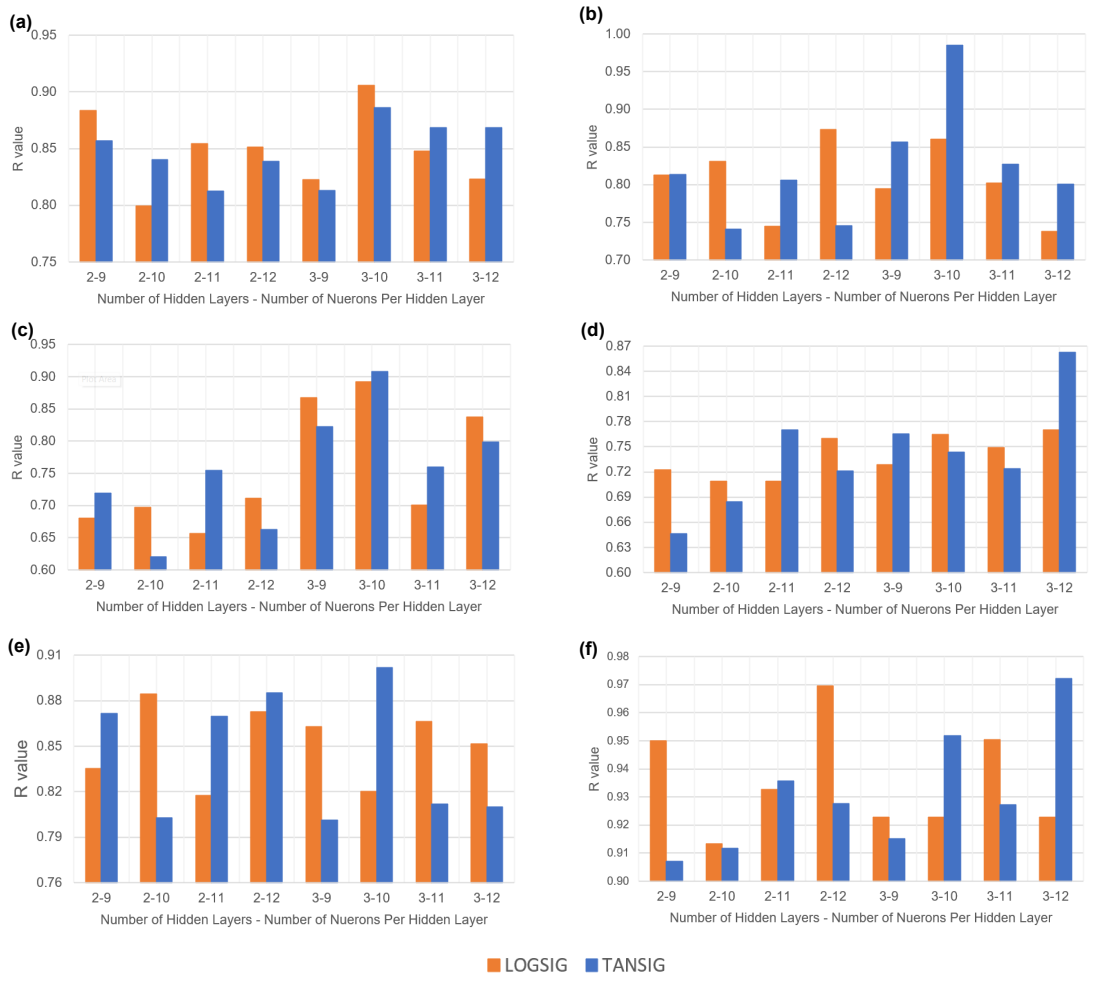

Figure 3: Summary of Network Performances in Modelling (a) Effluent BOD, (b) Effluent COD, and (c) Effluent TC for Non-Anomaly ICs, and (d) Effluent BOD, (e) Effluent COD, and (f) Effluent TC for Non-Anomaly ICs

The optimal network for modelling effluent BOD, COD, and TC for non-anomaly ICs have the same network architecture with six inputs, three hidden layers with three nodes each, then one output, while the network architecture for modelling effluent BOD and TC has six inputs, three hidden layers with 12 nodes each, and one output. The model for COD had three inputs, three hidden layers with then nodes each and one output layer. The process demonstrated in the sample simulation (Figure 1) was executed for all datapoints. During the training phase, the ANN models were able to produce R values ranging from 0.80 to 1.00, at par with values obtained by Alsulaili and Refaie (2021), El-Rawy et al. (2021), and Mundi et al. (2021). The mean absolute error (MAE) for the BOD ANN model was found to be lower than the one obtained in the training, while the ones for the COD and TC had a value that was between the training MAE for the anomaly and non-anomaly IC. In this manner, the actual vs. simulated values are presented in Figure 4.

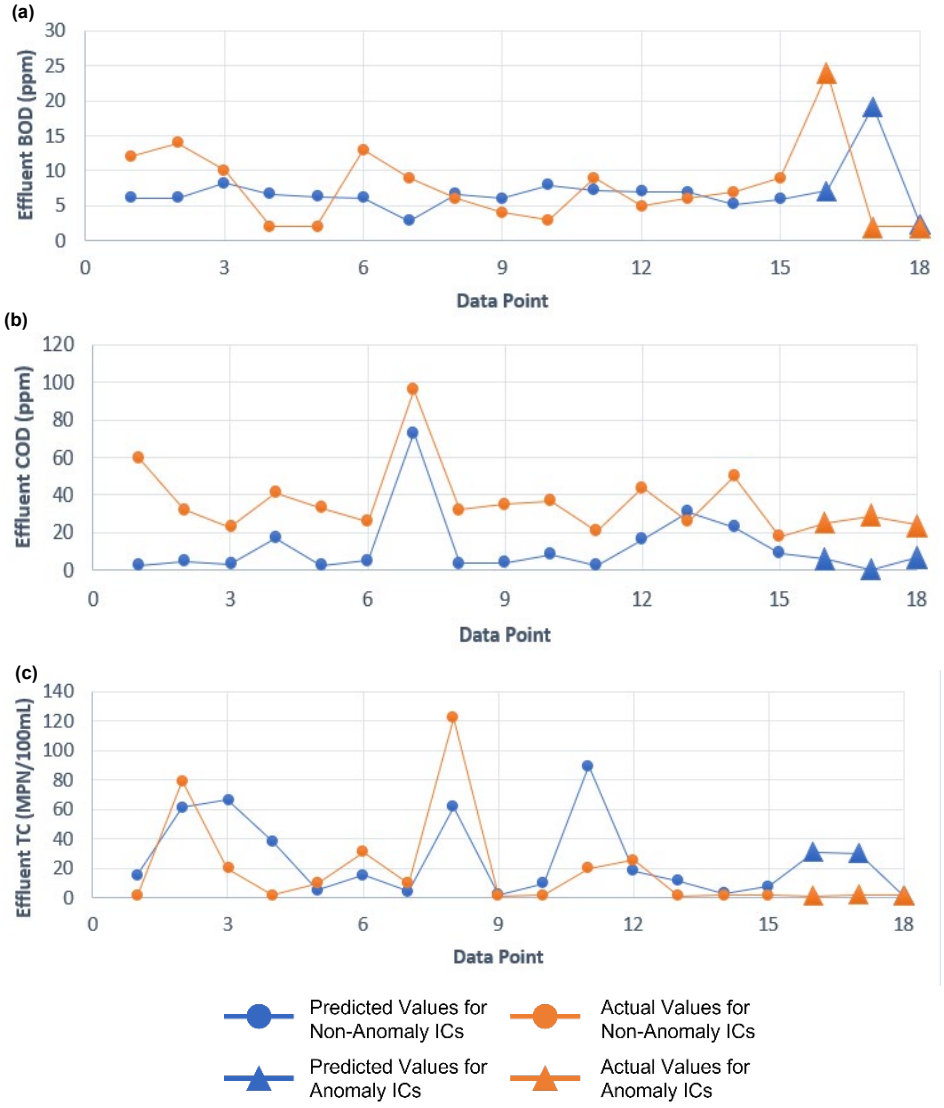

Figure 4: Summary of Data Points for Actual vs. Simulated Effluent (a) BOD, (b) COD, and (c) TC

In Figure 4, the test datapoints included 15 non-anomalies, and three anomalies determined through their Mahalanobis distance. From this figure, it can be deduced that the overall framework used in the study tends to underestimate the effluent BOD, COD, and TC. Despite this, the network was able to capture the trend with regards to the increase and decrease in the actual values. The errors in this network may be attributed to the false classifications for the anomaly detection algorithm. Two out of the three anomalies in the test dataset were identified correctly. All non-anomalies were identified correctly. The misclassification garnered a high absolute error for the dataset; BOD had 16.87 ppm, COD had 5.93 ppm, and TC had 30.95 MPN/100 mL. These errors have been attributed to few datapoints to represent anomaly ICs, not incorporating specific phenomena that

result from extreme external factors, overfitting, small overall dataset, infrequency of and method for sampling, and lack of process parameters in secondary and tertiary treatment (chlorination) incorporated in the models. To address issues on the framework's generalization ability, improvements must be made when it comes to the training data. On-site recording of data may be improved by using composite sampling, instead of grab sampling, keeping records of extreme environmental conditions occurring near the WWTP, and recording more water quality and process parameters. Future research may also use more feature selection tools, network optimization algorithms, development of interface for eased accessibility of the framework presented in this study and testing more machine learning tools for the same purpose indicated in the study.

5. Conclusion

An anomaly detection algorithm comprising a support vector machine model based on results from Mahalanobis distance was developed to determine anomaly and non-anomaly ICs. Based on these classifications, ANNs were modelled to cater towards the specific behavior of treatment performance for different types of ICs. The optimal type of SVM was the fine Gaussian SVM, with an AUC-ROC value of 0.90. The training and validation dataset accuracy for the MBBR system were 0.99 and 0.76, the best among all considered types of SVM. The optimal ANN models had three hidden layers, and 10 or 12 neurons per hidden layer depending on the projected network output. The MAE values for the ANN models were 4.95 ppm for BOD, 12.18 ppm for COD, and 26.29 MPN/100 mL for TC, when a fresh data set was used as the input to the optimal network. Overall, the study presented a viable tool to improve the modelling of WWTP performance that will aid in developing informed plans of action for its design and operation, particularly in determining possible dimensions best suitable for the specific conditions for a specific WWTP.

References

Ali F., Salim C., Lestari D.L., Azmi K.N., 2021, Challenges of moving bed biofilm reactor and integrated fixed-film activated sludge implementation for wastewater treatment in Indonesia, Chemical Engineering Transactions, 83, 223-228.

Alsulaili A., Refaie A., 2021, Artificial neural network modeling approach for the prediction of five-day biological oxygen demand and wastewater treatment plant performance, Water Supply, 21, 1861–1877.

Chiemchaisri C., Chiemchaisri W., Dachsrijan S., Saengam C., 2022, Coliform removal in membrane bioreactor and disinfection during hospital wastewater treatment, Journal of Engineering and Technological Sciences, 54, 220401.

El-Rawy M., Abd-Ellah M. K., Fathi H., Ahmed A. K. A., 2021, Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques, Journal of Water Process Engineering, 44, 102380.

Hernandez-Ramirez A. G., Martinez-Tavera E., Rodriguez-Espinosa P. F., Mendoza-Pérez J. A., Tabla-Hernandez J., Escobedo-Urías, D. C., Jonathan, M. P., Sujitha, S. B., 2019, Detection, provenance and associated environmental risks of water quality pollutants during anomaly events in River Atoyac, Central Mexico: A real-time monitoring approach. Science of the Total Environment, 669, 1019–1032.

Khawaga R. I., Abdel Jabbar N., Al-Asheh S., Abouleish M., 2019, Model identification and control of chlorine residual for disinfection of wastewater. Journal of Water Process Engineering, 32, 100936.

Li, D., Liu, S., 2019, Water Quality Monitoring and Management: Basis, technology and case studies, Water Quality Monitoring in Aquaculture, Academic Press, London, UK, 303–328.

Mundi G., Zytner R. G., Warriner K., Bonakdari H., Gharabaghi B., 2021, Machine learning models for predicting water quality of treated fruit and vegetable wastewater. Water (Switzerland), 13, 1–17.

Ramírez-Castillo F. Y., Loera-Muro A., Jacques M., Garneau P., Avelar-González F. J., Harel J., Guerrero-Barrera A. L., 2015, Waterborne pathogens: Detection methods and challenges. Pathogens, 4, 307–334.

Ratola N., Cincinelli A., Alves A., Katsoyiannis A., 2012, Occurrence of organic microcontaminants in the wastewater treatment process: A mini review. Journal of Hazardous Materials, 239–240, 1–18.

Scholz M., 2016, Wetlands for water pollution control, Constructed wetlands, Elsevier, Amsterdam, Netherlands, 137–155.

Tümer A. E., Edebali S., 2015, An artificial neural network model for wastewater treatment plant of Konya. International Journal of Intelligent Systems and Applications in Engineering, 3, 131–135.

Waqas S., Harun N. Y., Sambudi N. S., Arshad U., Nordin N. A. H. M., Bilad M. R., Saeed A. A. H., Malik, A. A., 2022, SVM and ANN modelling approach for the optimization of membrane permeability of a membrane rotating biological contactor for wastewater treatment, Membranes, 12, 81.

Zhao W., Tao T., Zio E., 2013, Parameters tuning in support vector regression for reliability forecasting, Chemical Engineering Transactions, 33, 523-528.