12th IFAC International Symposium on
Advanced Control of Chemical Processes
July 14-17, 2024. Toronto, Canada

Bacteria cells estimation in wastewater treatment plants using data-driven models ⋆

Fahad Aljehani * **Ibrahima N'Doye** **,* **Pei-Ying Hong** **
Mohammad K. Monjed *** **Taous-Meriem Laleg-Kirati** ****

* *Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia, (e-mail: fahad.aljehani@kaust.edu.sa, ibrahima.ndoye@kaust.edu.sa, taousmeriem.laleg@kaust.edu.sa).*
** *Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia, (e-mail: peiying.hong@kaust.edu.sa).*
*** *Department of Biology, Faculty of Science, Umm Al-Qura University, Makkah, 21961, Saudi Arabia, (e-mail: mkmonjed@uqu.edu.sa)*
**** *National Institute for Research in Digital Science and Technology (INRIA), Saclay, France, (e-mail: taousmeriem.laleg@inria.fr)*

Abstract: Estimating or predicting the concentrations of bacteria cells is crucial for achieving better control of the operation of wastewater treatment plants. However, measuring the bacterial concentration along the influent wastewater stage to the treated effluent process is challenging as it involves lab access and trained personnel. Additionally, wastewater plants are generally nonlinear systems involving time-varying physical and biological characteristics, increasing the difficulty in estimating the bacterial concentration from a model-based approach. This paper proposes data-driven models based on four machine-learning models to estimate the bacterial cell density with a limited dataset in a wastewater treatment plant. The performance results demonstrate that the machine-learning models (i.e., K-Nearest Neighbour (kNN), Random Forest (RF), Gradient Boosting Regression (GBR), Extreme Gradient Boosting (XGB)) have the potential to estimate accurately the bacterial concentration. RF displays better bacteria estimation in the influent by 10.7% compared to GBR and 7.4% compared to XGB and kNN. Whereas for the effluent, XGB improved the estimation by 12.8%, 2.4%, 14.6% compared to GBR, RF, and kNN, respectively. Also, results show that conductivity as a single feature is the most significant parameter affecting the bacterial cell estimation in the influent stage for the four machine learning algorithms. Similarly, the chemical oxygen demand (COD) and turbidity have pronounced effects in the effluent stage. These results reveal potential signs of designing a universal data-driven model-based approach applicable for bacteria estimation at influent and effluent based on the minimum feature combinations (conductivity, COD, and turbidity).

Keywords: Wastewater treatment plant, data-driven models, bacterial estimation models, machine learning algorithms

1. INTRODUCTION

The primary objective of a wastewater treatment plant (WWTP) is to reduce nutrient and pollutant concentrations as vectors for spreading the transmission of bacterial and viral contaminants. WWTPs are generally nonlinear systems involving time-varying physical and biological characteristics and exhibiting significant input variability in the influent stage. Therefore, estimating bacterial concentration is challenging due to the difficulty of building predictive mathematical models of the bacterial output and the non-availability of variables of interest in WWTP. Sample taking and laboratory analysis are the conventional methods to measure bacterial concentrations, which are time-consuming and do not reflect the state of the process in real-time (Manti et al., 2008). Hence, designing algorithms to estimate accurately and monitor the bacterial concentration through a soft sensor model is crucial to alleviate this problem and support the plant operators in WWTPs.

⋆ This work has been supported by the King Abdullah University of Science and Technology (KAUST) baseline research fund (BAS/1/1627-01-01) awarded to Taous-Meriem Laleg-Kirati, baseline research fund (BAS/1/1033-01-01) awarded to Pei-Ying Hong and NGTC-AI fund (REI/1/5233-01-01) awarded to Pei-Ying Hong and Taous-Meriem Laleg-Kirati.

Model-based estimation approaches provide an efficient strategy for assessing the variables of interest, including state, parameters, and unknown faults and disturbances (Dochain, 2003). However, the identification and model representation of accurate process models, along with the assumptions in the model derivation, for example, microorganism concentrations and reaction rates of WWTP systems, are the bottlenecks of the model-based approach. In this context, data-driven methods can circumvent analytical and model-based design methods by characterizing dominant underlying patterns in WWTPs (see, for instance, (Farhi et al., 2021; Pisa et al., 2019; Mokhtari et al., 2020; Alharbi et al., 2022b; Ekundayo et al., 2023; Wang et al., 2022; Alharbi et al., 2022a) and references therein). Data-driven models is currently revolutionizing how we model, predict and control complex systems. For instance, a key benefit of decision tree models is identifying models that contain the most required nonlinear terms (Cheng et al., 2023).

Different from our previous effective results on bacterial sensing for large data based on sliding window neural network (Alharbi et al., 2022b), the data-driven based estimation models here do not consider input-output data generation and work effectively with limited data, which is crucial for rapid identification and estimation of the bacterial concentration model. Furthermore, the work in (Alharbi et al., 2022b) relies inherently on the dependence on the last three water quality samples (during three months) to estimate bacteria, which is challenging to capture the relationship between water quality and bacteria without using data generation. More importantly, the synergy between data-driven based identification and model-based estimation approach promotes the generalization of the bacterial concentration beyond the training data. In line with this, we propose data-driven models for bacterial concentration estimation by considering two datasets from two WWTPs in Saudi Arabia. We evaluate four models (K-Nearest Neighbour (kNN), Random Forest (RF), Gradient Boosting Regression (GBR), Extreme Gradient Boosting (XGB)), and compare their performances. These data-driven models are mainly based on a decision tree framework and have the advantage of capturing the dominant underlying features of the influent and effluent bacterial concentration within limited samples. Subsequently, they have the merit to prevent overfitting and the ability to encompass good generalization when combined with a model-based approach. The main contributions of this paper are summarized as follows:

- Develop four data-driven models for bacterial concentration estimation on a combined dataset of two WWTPs.
- Estimate the concentrations of influent and effluent bacterial biomass based on the minimum feature combinations (conductivity, COD, turbidity).
- Reveal the importance of developing a universal data-driven model-based estimation based on these dominant features.

The paper's outline is organized as follows: Section 2 describes the process of the wastewater treatment plant, including the water quality factors. Section 3 provides the data preprocessing and introduces the machine-learning algorithms. Section 4 discusses the bacteria estimation results. Finally, the article summarizes the main contributions and future works in Section 5.

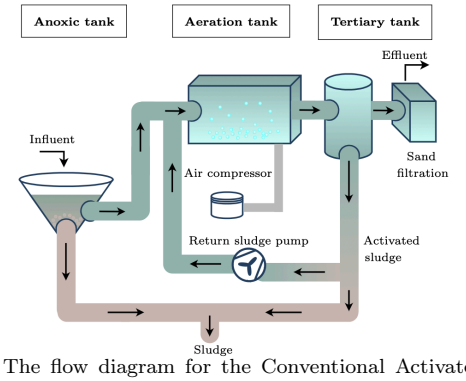


Fig. 1. The flow diagram for the Conventional Activated Sludge (CAS) process with tertiary treatment located in Saudi Arabia illustrates the process of untreated water flowing in the anoxic tank as influent and treated in the tertiary tank as effluent. Furthermore, the air compressor diffuses bubbles in the aeration tank, and the activated sludge recycles to the aeration tank using the pump.

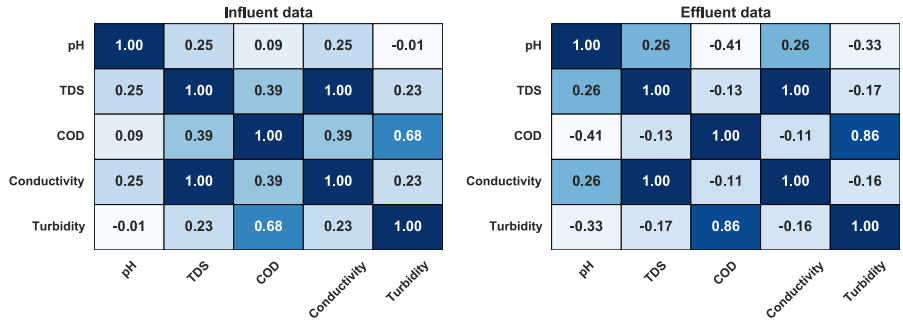
Table 1. Statistics of water quality factors and bacteria concentrations

Variables	Mean ± Standard deviation (SD)	
	Influent	Effluent
pH	7.28 ± 0.22	7.67 ± 0.46
TDS (mg/L)	876.27 ± 419.43	755.56 ± 372.57
Conductivity (μS/cm)	1239.36 ± 592.43	1068.08 ± 526.84
COD (mg/L)	135.84 ± 66.29	12.51 ± 5.56
Turbidity (NTU)	57.14 ± 34.08	2.10 ± 1.30
Bacteria levels (cell/L)	8.31 ± 0.55	7.52 ± 0.86

2. MATERIALS

This study aims to predict bacteria concentrations based on accessible measurements of influent and effluent in wastewater treatment plants in Saudi Arabia. The plant has three stages: anoxic, aeration, and tertiary tanks. Fig. 1 presents the process of wastewater as influent to treated water as effluent. The influent enters an anoxic tank where the objective is to reduce nitrate (NO_3^-) and nitrite (NO_2^-) by denitrification process. Then, the water that has reduced nitrogen content flows to an aeration tank, which is high in dissolved oxygen due to the air compressor and reduces Biochemical Oxygen Demand (BOD). Once the water exits the aeration tank, it flows to the tertiary tank, removing solids and polishing the water as effluent. The remaining activated sludge is recycled into the aeration tank using a pump, and the process is repeated.

The main monitoring water quality factors are pH (potential of hydrogen), COD (chemical oxygen demand), TDS (total dissolved solids), turbidity, and conductivity. Table 1 summarizes the statistics of these factors including bacteria concentration, which has been determined in the lab by flow cytometry protocols as in (Timraz et al., 2017). The data were collected from July 1, 2020, to July 20, 2022. The total samples of these data is 64 samples. All the values of the bacteria cells are converted to a logarithmic scale.



(a) (b)

Fig. 2. Correlation matrix between water quality features. Fig. 2(a) influent feature correlation where (TDS, Conductivity) present high correlation. Fig. 2(b) effluent features correlation, which also illustrates high correlation between (TDS, Conductivity)

Table 2. Optimal hyper-parameter tuning using Optuna for the four machine-learning algorithms at the influent and efflunt stages.

Algorithm		**Best input features**	**Parameters**
kNN	Influent	(Conductivity)	n_neighbors: 1, weights: uniform
	Effluent	(COD, Conductivity, Turbidity)	n_neighbors: 1, weights: uniform
RF	Influent	(Conductivity)	n_estimators: 60, max_depth: 10, min_samples_split: 2
	Effluent	(COD, Turbidity)	n_estimators: 50, max_depth: 9, min_samples_split: 2
GBR	Influent	(Conductivity)	n_estimators: 150, learning_rate: 0.197, max_depth: 10
	Effluent	(pH, COD, Conductivity)	n_estimators: 133, learning_rate: 0.178, max_depth: 10
XGB	Influent	(pH, Conductivity)	n_estimators: 128, learning_rate: 0.096, max_depth: 10
	Effluent	(COD, Conductivity, Turbidity)	n_estimators: 108, learning_rate: 0.149, max_depth: 5

3. METHODS

3.1 Data preparation

The objective is to build sufficient models to predict bacteria concentrations based on these measurements. Investigating the correlation is essential to achieve this objective and reduce feature redundancy. The correlation aids in understating how two variables/features are linearly close. The correlation is formulated as follows

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}, \quad (1)$$

where x_i and y_i are the samples of features x and y, and $\bar{x}$ and $\bar{y}$ are the mean values of the features x and y. Highly correlated two features result in r being near to 1. Fig. 2 illustrates the correlation between the features of the influent and effluent. The correlation between conductivity and TDS is very high for influent and effluent datasets, suggesting that one of these features should be eliminated. Therefore, TDS is excluded, and the utilized features are pH, COD, conductivity, and turbidity for the rest of our analysis.

3.2 Machine learning algorithms

In the model development process, machine learning algorithms such as K-Nearest Neighbour (kNN), Random Forest (RF), Gradient Boosting Regression (GBR) and Extreme Gradient Boosting (XGB) have been implemented for regression model screening. Note that other machine learning algorithms, such as support vector regression (SVR), multivariate adaptive regression spline (MARS), and multi-layer perception (MLP), have been evaluated in our dataset. However, their estimation performances were not accurate during the testing stage.

3.3 Performance measures

To evaluate the performance during training and testing, four quantitative error metrics are calculated: Root-Mean-Square-Error (RMSE), Mean-Squared-Error (MSE), Mean Absolute Error (MAE), and Mean-Absolute-Percentage-Error (MAPE) as follows:

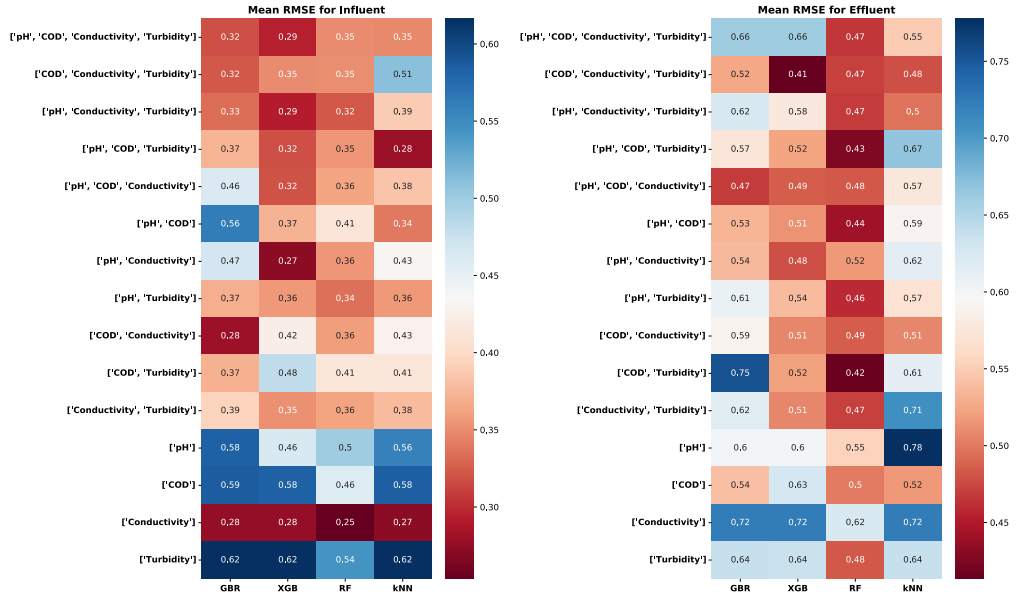
$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}}, \quad \text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad \text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

where y_i refers to the real output at sample i, $\hat{y}_i$ denotes the predicted output at sample i.

4. RESULTS AND DISCUSSION

The four machine-learning algorithms have been implemented using limited data samples where the data was split into (80%, 20%) for training and testing. The first



(a) (b)

Fig. 3. Mean RMSE for influent and effluent using five K-folds for the four machine-learning algorithms. Fig.3(a) suggests the optimal feature combinations that resulted in minimum mean RMSE for GBR is (Conductivity), XGB is (pH, Conductivity), RF is (Conductivity), and kNN is (Conductivity). Fig. 3(b) suggests the optimal feature combinations that resulted in minimum mean RMSE for GBR is (pH, COD, Conductivity), XGB is (COD, Conductivity, Turbidity), RF is (COD, Turbidity), and kNN is (COD, Conductivity, Turbidity). This study shows that conductivity and COD/turbidity are crucial in estimating the bacterial concentration in the influent and effluent stages, respectively.

objective is determining each algorithm's best hyper-parameter and feature combinations. This step has been done using a well-known library in Python called *Optuna* Akiba et al. (2019). Table 2 presents the optimal feature combinations and hyper-parameter tuning using 1000 trials. It is noticeable that conductivity is an important feature, which appears alone or in combination with other features in most of the influent and effluent, except for RF effluent. The most essential feature combinations for the effluent include COD and turbidity.

For further analysis, we utilized the concept of K-folds to generalize the performance of the algorithms. In addition, we focused on the RMSE metric to evaluate the performance of the models. Fig. 3 illustrates mean RMSE for influent and effluent using five K-folds for the four machine-learning algorithms where Figs. 3(a) and (b) refer to influent and effluent, respectively. Fig. 4 shows the results of bacteria estimation for the four algorithms in the training and testing stages. Figs. 4(a), (c), (e), and (g) illustrate the training stage with used feature combinations for influent using GBR, XGB, RF, and kNN, respectively. One the other hand, Figs. 4(b), (d), (f), and (h) present the testing stage. Similarly, Fig. 5 shows the training and testing results for the effluent. In the influent case, the minimum mean RMSEs for each algorithm are 0.28, 0.27, 0.25, and 0.27 for GBR, XGB, RF, and kNN, respectively. Suggesting that RF resulted in minimum RMSE which has improved the estimation error by 10.7% compared to GBR and 7.4% compared to XGB and kNN. The feature conductivity is distinguishable and performs well alone in GBR, RF, and kNN and combination with pH in XGB. For the effluent case, the minimum RMSE for GBR is 0.47, XGB is 0.41, RF is 0.42, and kNN is 0.48. XGB present good estimation compared to GBR by 12.8%, RF by 2.4%, and kNN by 14.6%. COD and turbidity features resulted in minimum RMSE in three algorithms whether in combination with each other alone as in RF, or with conductivity as in GBR and kNN.

5. CONCLUSION

In this work, we proposed data-driven models based on K-Nearest Neighbour (kNN), Random Forest (RF), Gradient Boosting Regression (GBR), and Extreme Gradient Boosting (XGB) models to estimate the bacterial cell density in a wastewater treatment plant. The four machine learning models displayed good bacteria estimation performance. We also demonstrated that conductivity as a single feature is the most significant parameter affecting the bacterial cell estimation in the influent stage. Similarly, the chemical oxygen demand (COD) and turbidity have pronounced effects in the effluent stage. The proposed data-driven models based primarily on traditional machine learning models could provide a timely and rapid assessment of the rate of the bacterial concentration density. These results also implied that leveraging data-driven models with dominant underlying conductivity, COD and turbidity patterns could benefit from model-based methods to provide an evidence-based strategy to tackle the universality of the proposed models, which are essential to the operation of wastewater treatment plants.

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter

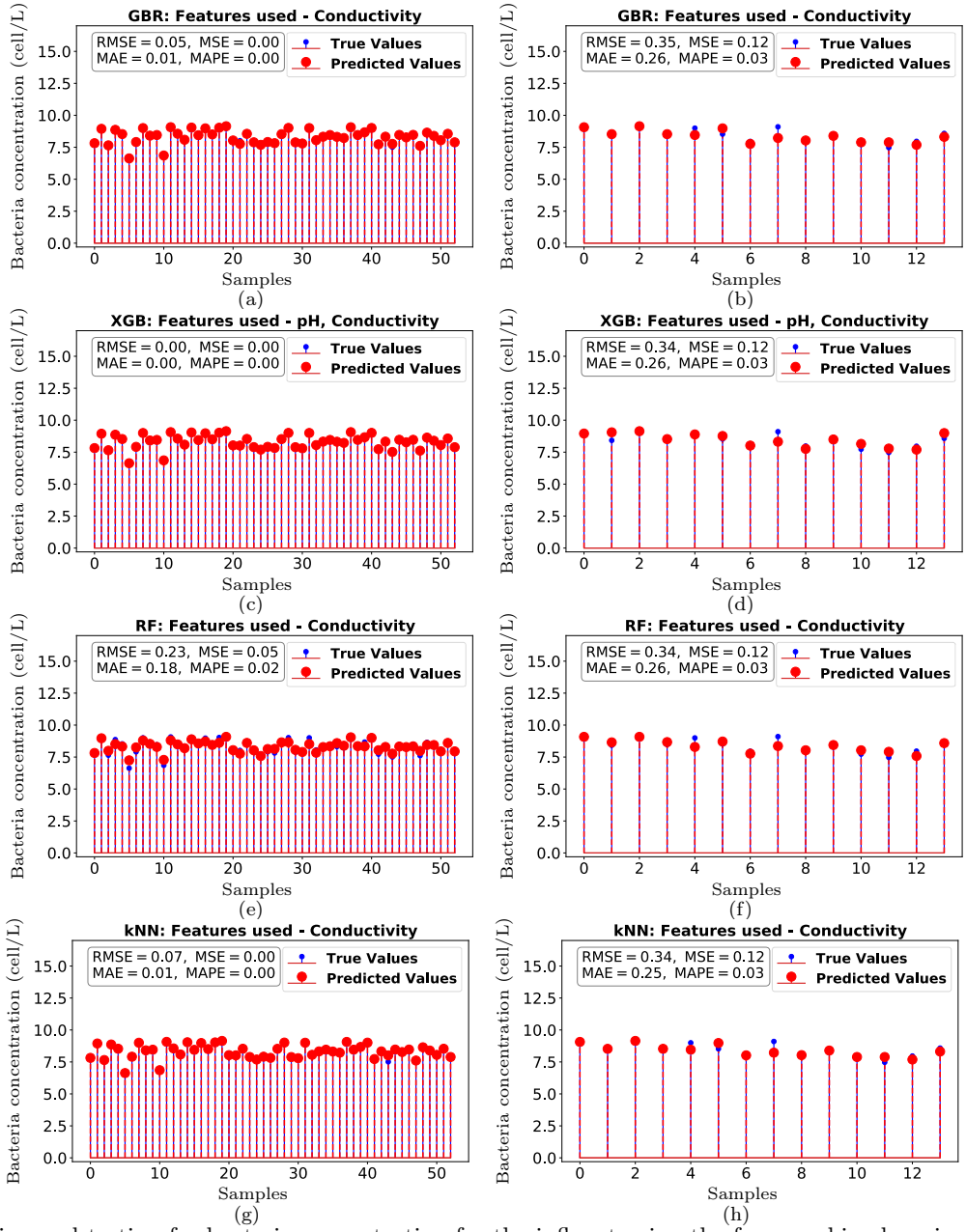


Fig. 4. Training and testing for bacteria concentration for the influent using the four machine-learning algorithms. On the left column of Figs. 4 (a, c, e, g) is the training stage for GBR, XGB, RF, and kNN. Whereas on the right column of Figs. 4 (b, d, g, h) is testing stage. RF shows better bacteria estimation in the influent by 10.7% compared to GBR and 7.4% compared to XGB and kNN in the mean RMSE criteria (See heatmaps in Fig. 3).

optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Alharbi, M.S., Hong, P.Y., and Laleg-Kirati, T.M. (2022a). Adaptive neural network based monitoring of wastewater treatment plants. In *American Control Conference (ACC)*, 3204–3211.

Alharbi, M., Hong, P.Y., and Laleg-Kirati, T.M. (2022b). Sliding window neural network based sensing of bacteria in wastewater treatment plants. *Journal of Process Control*, 110, 35–44.

Cheng, Q., Chunhong, Z., and Qianglin, L. (2023). Development and application of random forest regression soft sensor model for treating domestic wastewater in a sequencing batch reactor. *Scientific reports*, 13, 9149.

Dochain, D. (2003). State and parameter estimation in chemical and biochemical processes: a tutorial. *Journal of Process Control*, 13(8), 801–818.

Ekundayo, T.C., Adewoyin, M.A., Ijabadeniyi, O.A., Igbinosa, E.O., and Okoh, A.I. (2023). Machine learning-guided determination of acinetobacter density in waterbodies receiving municipal and hospital wastewater effluents. *Scientific Reports*, 13(1), 7749.

Farhi, N., Kohen, E., Mamane, H., and Shavitt, Y. (2021). Prediction of wastewater treatment quality using LSTM neural network. *Environmental Technology and Innova-*

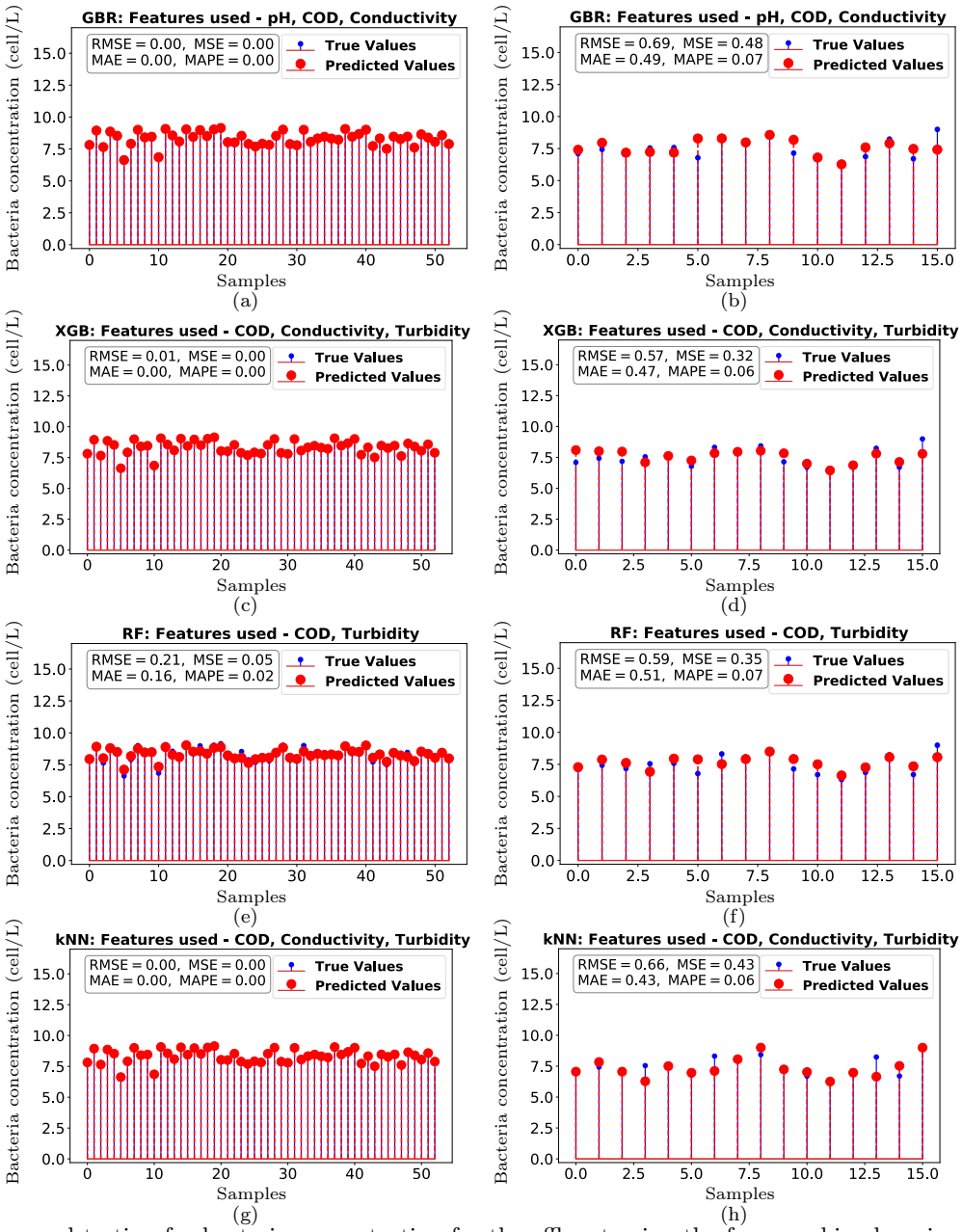


Fig. 5. Training and testing for bacteria concentration for the effluent using the four machine-learning algorithms. On the left column of Figs. 5 (a, c, e, g) are the training stage for GBR, XGB, RF, and kNN, respectively. Whereas on the right column of Figs. 5 (b, d, f, h) are the testing stage. XGB improved the estimation by 12.8%, 2.4%, 14.6% compared to GBR, RF, and kNN in the effluent case in the mean RMSE criteria (See heatmaps in Fig. 3).

tion, 23, 101632.

Manti, A., Boi, P., Falcioni, T., Canonico, B., Ventura, A., Sisti, D., Pianetti, A., Balsamo, M., and Papa, S. (2008). Bacterial cell monitoring in wastewater treatment plants by flow cytometry. *Water Environ. Res.*, 80(4), 346–354.

Mokhtari, H.A., Bagheri, M., Mirbagheri, S.A., and Akbari, A. (2020). Performance evaluation and modelling of an integrated municipal wastewater treatment system using neural networks. *Water and Environment Journal*, 34, 622–634.

Pisa, I., Santin, I., Morell, A., Vicario, J.L., and Vilanova, R. (2019). LSTM-based wastewater treatment plants operation strategies for effluent quality improvement. *IEEE Access*, 7, 159773–159786.

Timraz, K., Xiong, Y., Al Qarni, H., and Hong, P.Y. (2017). Removal of bacterial cells, antibiotic resistance genes and integrase genes by on-site hospital wastewater treatment plants: Surveillance of treated hospital effluent quality. *Environ. Sci.: Water Res. Technol.*, 3(2), 293–303.

Wang, R., Yu, Y., Chen, Y., Pan, Z., Li, X., Tan, Z., and Zhang, J. (2022). Model construction and application for effluent prediction in wastewater treatment plant: Data processing method optimization and process parameters integration. *Journal of Environmental Management*, 302, 114020.