JOURNAL OF WATER AND LAND DEVELOPMENT
e-ISSN 2083-4535
Polish Academy of Sciences (PAN) Institute of Technology and Life Sciences – National Research Institute (ITP – PIB)

JOURNAL OF WATER AND LAND DEVELOPMENT
DOI: 10.24425/jwld.2025.155310
2025, No. 66 (VII–IX): 137–151

Machine learning algorithms in wastewater technology: Predicting treatment quality and efficiency

Sihame Barahi*[1)], Amina Azizi[2)], Mohamed Taky[1)], Sakina Belhamidi[1), 3)]

[1)] Laboratory of Advanced Materials and Process Engineering, Faculty of Sciences, Ibn Tofail University, P. O. Box 1246, Kenitra, Morocco
[2)] Independent Water and Electricity Company, B.P. 229, 14001 Kenitra, Morocco
[3)] Ibn Tofail University, Superior School of Technology, B.P. 1246, Kenitra, Morocco

* Corresponding author

RECEIVED 13.11.2024 ACCEPTED 16.07.2025 AVAILABLE ONLINE 17.09.2025

Abstract: Wastewater treatment is essential for protecting both the environment and public health. With a growing global population and concerns about water shortages, wastewater must be treated effectively to meet the increasing demand for drinking water. Wastewater treatment plants (WWTP) that use innovative technologies, such as machine learning (ML), are playing a leading role in addressing this challenge. This study aims to use advanced ML algorithms to predict parameters in WWTP Kenitra, such as total suspended solids (TSS), chemical oxygen demand (COD), and biological oxygen demand (BOD). Four ML models were evaluated, including random forest (RF), decision tree regressor (DTR), Gaussian process regressor (GPR), and adaptive boosting regression (AdaBoost-R). The coefficient of determination (R^2) and accuracy were used to evaluate the algorithm's efficiency, R^2 values of 0.99, 0.93, and 0.96 were obtained by the DTR, reflecting exceptional performance with *RMSE* values of 1.33 mg·dm^{-3} for TSS, 3.85 mg·dm^{-3} for COD, and 2.32 mg·dm^{-3} for BOD. The GPR demonstrated strong predictive capability, achieving R^2 values of 0.92 for TSS and 0.97 for BOD, with corresponding *RMSE* values of 3.12 mg·dm^{-3}, and 2.67 mg·dm^{-3}, respectively. These results indicate that the DTR and GPR learning models provide better algorithms for evaluating wastewater parameters. In particular, the study demonstrates the main benefits of using ML algorithms to predict the parameters of WWTP. This study illustrates that the DTR optimises treatment solutions and monitors the treatment process. The proposed method outperforms other algorithms in terms of efficiency and provides an accurate way to improve the performance of WWTP.

Keywords: AdaBoost regressor, decision tree regressor, Gaussian process regressor, machine learning, performances, random forest, wastewater treatment plants (WWTP)

INTRODUCTION

Water is a vital resource for the sustainability of both human and natural ecosystems (Daud, 2023). The escalating demand for water has emerged as a critical global challenge exacerbated by increasing industrial pollutants, rapid population growth, and intensive agricultural practices, highlighting the need for effective water management strategies (Santos, Carvalho and Martins, 2023). Wastewater treatment technology is essential to eliminate contaminants from water systems. Using advanced treatment methods enhances wastewater management and plays a key role in protecting the environment and public health by removing harmful substances (Rousis *et al.*, 2024). Among various wastewater treatment technologies the activated sludge process treats and maintains water quality by removing pollutants (Ayyoub *et al.*, 2022). Biological wastewater treatment using rich and diverse microbial communities significantly reduces the levels of pollutants and nutrients (Ayyoub *et al.*, 2023). Treatment conditions and wastewater characteristics influence the size and composition of microbial communities, affecting their role in treatment processes and their overall efficiency (Lukyanova, Golodov and Kirilenko, 2024). However, complex, nonlinear process variables and fluctuating intake parameters present challenges for wastewater treatment plants (WWTP), as they

require comprehensive and continuous effluent quality monitoring (Dey *et al.*, 2024).

Traditional wastewater monitoring techniques are costly and labour-intensive for real-time applications, and they cannot maintain effluent quality standards due to increasing industrialisation and urbanisation. Although the application of wireless sensor networks is limited by their cost and availability, such networks can offer some online measurement capabilities. However, these traditional methods face significant limitations (Duarte *et al.*, 2024). Traditional methods for monitoring biochemical oxygen demand (BOD_5), total suspended solids (TSS), and chemical oxygen demand (COD) in wastewater treatment are very costly and resource-intensive; standard BOD testing takes five days, and COD and TSS assessments require specialised equipment.

These challenges, combined with the non-linear correlation of these parameters, limit the effectiveness of traditional methods, underscoring the need for advanced real-time monitoring solutions to maximise performance (Asteris *et al.*, 2022). To overcome these challenges and improve wastewater management operations, machine learning (ML) must be integrated. ML techniques provide WWTPs with a deeper analytical understanding of key processes and parameter interactions, and serve as a powerful tool for predicting key wastewater parameters, which in turn enables operational improvements (Aghdam *et al.*, 2023). By using ML, we can model complex nonlinear relationships without defining the treatment process through chemical or mathematical equations. The integration of ML techniques into wastewater treatment operations can lead to several benefits, including decreased energy consumption and maintenance costs, enhanced plant efficiency, process optimisation, and environmental conservation (Qambar and Khalidy, 2022). The most common ML models used to predict, evaluate, and diagnose WWTP include artificial neural networks (ANNs), fuzzy logic, genetic algorithms (GA), adaptive-network-based fuzzy inference system (ANFIS), and hybrid models that include ANN-GA. The ANNs are especially widely used to model and predict the performance of biological treatment processes in WWTP. Duarte *et al.* (2024) used three machine learning models, i.e. random forest (RF), support vector machine (SVM), and multilayer perceptron (MLP) to predict wastewater quality parameters in wastewater treatment plants. Shingare *et al.* (2024) demonstrated how ML can be applied to WWTP to ensure high-quality effluent, maximise energy efficiency, detect issues, and monitor sensors. Nasir Bin and Li (2024) used three machine learning models, i.e. RF, gradient boosting machine (GBM), and gradient boosting tree (GBT) to predict the amount of sludge produced in wastewater treatment facilities. With the lowest error measures and highest coefficient of determination (R^2) values, RF outperformed the other models, especially when feature selection techniques were used to improve it. All models achieved high prediction accuracy. Cechinel *et al.* (2024) predicted the COD in wastewater treatment plants using ML models, including RF, multilayer receptors, long short-term memory, and SVM. The results demonstrated the effectiveness of the models and provided guidance for improving wastewater treatment procedures, with MLP performing best with daily data, long short-term memory (LSTM) outperforming with hourly data, and SVM outperforming other models when using actual waste measurements.

Gholizadeh *et al.* (2024) predicted the TSS in wastewater treatment facility using three machine learning algorithms: adaptive boosting (AdaBoost), *k*-nearest neighbours (KNN), and artificial neural network-multilayer (ANN-MLP). The fourth scenario, which used the sequential backward selection feature selection method, was the most effective among the five scenarios. The ANN-MLP performed the best proving its reliability in predicting TSS.

Aghilesh *et al.* (2023) optimised the biosorption process for the removal of methylene blue dye from textile wastewater by using sugarcane residue and groundnut shells in combination with response surface methodology, ANN, and ANFIS. Fouchal *et al.* (2025) created two novel hybrid machine learning models, neural architecture search–deep neural network (NAS-DNN) and neural architecture search–random forest regression (NAS-RFR), that outperformed traditional wastewater quality assessment and treatment plant optimisation decision-making processes. These models achieved correlation coefficient (R) values of 0.953 and 0.934 at WWTPs and predicted BOD_5 with remarkable accuracy. Tan, Arumugasamy and Teo (2025) employed ANN models to predict and forecast the water quality index (*WQI*) using key water quality parameters. The prediction model outperformed the forecasting model in terms of accuracy. Rashidi-Khazaee, Rezvantalab and Kheshti Monasebi (2024) demonstrated the efficiency of ensemble machine learning algorithms, including AdaBoost, RF, GB, in accurately predicting key wastewater quality indicators such as COD, BOD, TSS, TN, and TP. Notably, the AdaBoost model outperformed the others, achieving the lowest mean absolute error for pH, BOD, and COD.

Wastewater treatment processes face numerous challenges, particularly in predicting key parameters such as COD_{eff}, BOD_{eff}, and TSS_{eff}. This issue is particularly relevant for the Kenitra WWTP, where local wastewater characteristics and environmental conditions pose unique challenges. Furthermore, few studies have explored the environmental impacts of discharging treated wastewater into natural water bodies – a major concern for Kenitra. Our study aims to fill these gaps by assessing the predictive accuracy of machine learning algorithms like RF, decision tree regressor (DTR), Gaussian process regressor (GPR), and AdaBoost-R. These algorithms also improve the precision and effectiveness of wastewater treatment assessments, support compliance with environmental regulations, and promote more environmentally friendly wastewater management techniques.

MATERIALS AND METHODS

STUDY AREA

The wastewater treatment plant is located northeast of the city of Kenitra in the Rabat-Sale-Kenitra region of Morocco. The system is designed to treat pollutants generated by the equivalent of 350,000 inhabitants, treating approximately 60,000 m^3 of wastewater per day on average. The wastewater treated at the Kenitra plant is classified as urban and is treated using an activated sludge system. The plant operates two main treatment lines, each addressing different aspects of the wastewater treatment process. The lines are divided into three functional components, as illustrated in Figure 1; this figure shows a schematic representation of the study area, which was designed using Microsoft PowerPoint:

– "water" line is in charge of biologically treating urban wastewater to ensure it satisfies environmental regulations before it is released into the Sebou River;

– "sludge" line: the secondary sludge produced during the biological treatment process is treated and processed in this section;
– "biogas" line: this section focuses on the anaerobic digestion of sludge to create biogas, which is a renewable energy source that improves the plant's sustainability and energy efficiency. In this study, we focused on the "water" line.

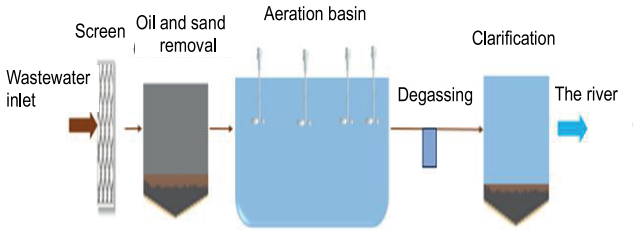

Fig. 1. Wastewater treatment line diagram; source: own elaboration

As shown in Figure 1, wastewater arriving at the pumping station undergoes screening followed by to primary treatment to remove coarse matter. This includes fine screening and combined grit and grease removal. The primary laminar sedimentation process then effectively removes the majority of the suspended solids. The partially treated water is subsequently directed to an aeration tank, where activated sludge is used to reduce carbon pollution. In the final stage, a purification process separates the treated water from the sludge, after which water is discharged into the Sebou River, one of Morocco's main rivers. Historically, water quality issues in the Sebou River have been impacted by untreated industrial and domestic effluents. The current plant ensures that the effluent complies with environmental standards, helping to improve water quality. Daily monitoring between 2022 and 2023 produced 390 data points. The dataset includes input parameters, such as pH, electrical conductivity (*EC*), chemical oxygen demand (COD), biological oxygen demand (BOD_5), total suspended solids (TSS), and output parameters, such as COD, BOD_5, and TSS.

ANALYTICAL METHODS

Daily influent and effluent samples from the Kenitra wastewater treatment plant were collected in strict adherence to Moroccan standards outlined in the Ministère de l'Environnement du Maroc (2002). Critical efficiency parameters measured included pH, COD ($mg \cdot dm^{-3}$), BOD_5, ($mg \cdot dm^{-3}$), and *EC* ($\mu S \cdot cm^{-1}$). This daily sampling ensured the acquisition of a thorough and high-resolution dataset, providing a solid basis for assessing treatment effectiveness.

Statistical analysis

The influent and effluent concentrations of pH, BOD_5, *EC*, COD, and TSS were examined using analysis of variance (ANOVA) in OriginPro 2018. The results indicated statistically significant differences between values.

Machine learning models

Supervised learning is an ML approach in which algorithms are trained on labelled datasets, enabling them to classify data or predict outcomes with known accuracy. Over time, the model can be trained and tested for accuracy using labelled inputs and outputs. Generally, supervised learning problems can be divided into two categories: classification algorithms, which attempt to group data into distinct categories, and regression algorithms, which are useful for applications and predict numerical values by figuring out how dependent and independent variables relate to one another (Wang, Cui and Ke, 2023). In this study, we used four different ML algorithms to predict wastewater quality effluent at the Kenitra WWTPs: RF, DTR, GPR, and AdaBoost-R.

The flowchart outlines a systematic approach using machine learning to predict wastewater parameters in the Kenitra WWTPs (Fig. 2). The influent and effluent parameters are collected, and the data is pre-processed for analysis. Four ML models were selected for prediction, including GPR, RF, DTR, and AdaBoost-R. The dataset is split into training (70%) and testing (30%) subsets and the models are trained and tested. The model performance was evaluated using R^2, root mean square error (*RMSE*), Nash–Sutcliffe efficiency (*NSE*), Kling–Gupta efficiency (*KGE*), mean square error (*MSE*), and mean absolute error (*MAE*) metrics. The results are compared to assess the model performance, and conclusions are drawn, offering insights and recommendations. This approach ensures a comprehensive evaluation of the role of machine learning in predicting and improving wastewater treatment plants.

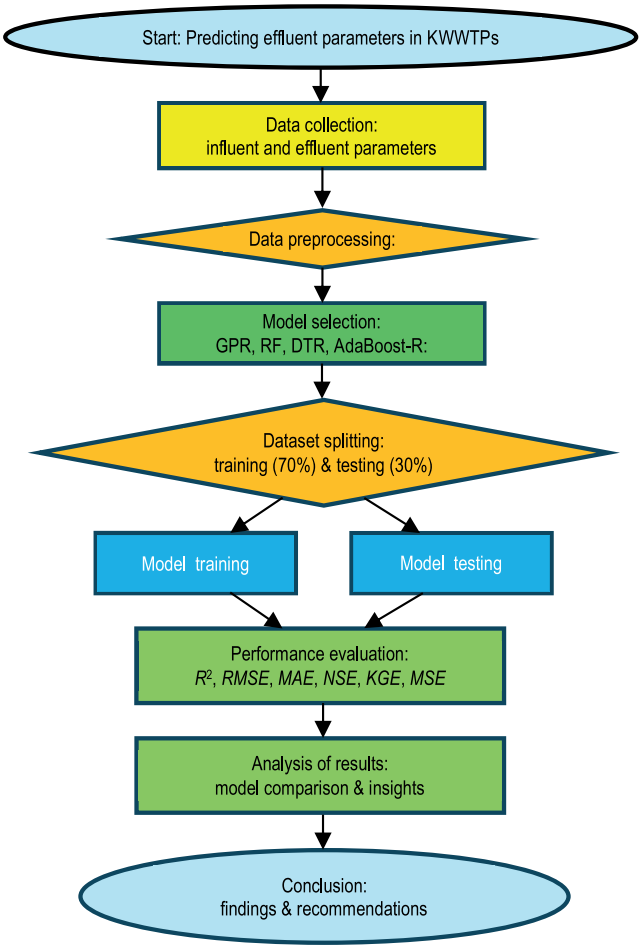

Fig. 2. Study flow chart; KWWTPs = Kenitra wastewater treatment plants, GPR = Gaussian process regressor, RF = random forest, DTR = decision tree regressor, AdaBoost-R = adaptive boosting regression, R^2 = coefficient of determination, *RMSE* = root mean square error, *MAE* = mean absolute error, *NSE* = Nash–Sutcliffe efficiency, *KGE* = Kling–Gupta efficiency, *MSE* = mean square error; source: own elaboration

• **Random forest (RF)**

The random forest model is an ML technique with high adaptive learning capability and nonlinear mapping features and has emerged as an important focus of AI research (Cheng, Chunhong and Qianglin, 2023). It has become useful in many fields, including water resource applications. To improve prediction accuracy and reduce overfitting, a large number of decision trees were constructed during the training phase, and their outputs were collected. To enhance the diversity among the trees, the algorithm employs an ensemble strategy, where each tree is trained on a random subset of the data. By selecting a random subset of features for each tree, a random forest also introduces more randomness, which reduces correlation and enhances the robustness of the model (Tyralis, Papacharalampous and Langousis, 2019). The final prediction in the RF dataset was calculated by averaging the predictions of each tree. The proposed algorithm is well known for its capacity to control large, high-dimensional datasets and avoid overfitting. The accuracy and robustness of the RF algorithms make them suitable for various real-world water resource problems.

• **Decision tree regressor**

The decision tree regression is a supervised machine learning algorithm that iteratively splits the input space according to specific feature values (Saleem, Harrou and Sun, 2024). The model is structured as a binary tree, where each internal node represents a decision based on a binary yes/no criterion for a given feature. The algorithm begins with a root node, which is often referred to as the parent node, from which subsequent nodes are recursively split according to specified feature thresholds. The splits generate new left and right child nodes. This iterative process of splitting continues until the pre-specified stopping criteria are met, forming leaf nodes of the tree, and representing the final predictive values. At each split, a decision is made based on the feature values, which results in data partitioning, with each division contributing to the model's prediction (Yazdani, Doostizadeh and Aminifar, 2023). Finally, the leaf nodes hold the predicted output values, determined by the splitting performed throughout the decision-making process (Atanasova and Kompare, 2002).

• **Gaussian process regressor**

The Gaussian regression is a supervised learning technique that solves probabilistic regression and classification problems (Ivan and Ivan, 2023). The model is nonparametric and used for regression in which the target function is a stochastic process. As a Gaussian distribution, it is defined by mean and covariance functions. GPR provides predictions with associated uncertainty by modelling the underlying function and observational noise. The computational practicality has improved with recent advances such as GPy-Torch. In complicated systems, GPR is useful for fault detection because it can identify and isolate operational issues and sensor failures (Ivan and Ivan, 2023).

• **Adaptive boosting regression (AdaBoost-R)**

The AdaBoost-R is an ensemble ML technique that uses adaptive resampling to iteratively correct the errors of the underlying model, which improves prediction accuracy. The proposed model builds a set of models, fixes the errors in previous models, and then uses a weighted sum to combine all models into a final predictive model with modifications. The proposed method, which works particularly well when combined with DTs, excels at optimising performance on complex datasets by focusing on hard-to-predict situations. The AdaBoost-R is a useful tool for many applications, such as wastewater treatment and other predictive tasks, due to its ability to handle noisy data, and its focus on difficult predictions (Nguyen and Seidu, 2022).

THE QUALITY OF THE DATASET AND ITS SPLITTING IN MACHINE LEARNING PERFORMANCE

In complex domains like wastewater treatment, the ML model performance is heavily influenced by the completeness and quality of a dataset. The model is guaranteed to capture the variability found in real-life situations when based on a representative dataset. Data from 2022 to 2023 covering 390 samples and key influent characteristics including *EC*, pH, TSS, COD, and BOD_5 were used to predict effluent concentrations TSS_{eff}, COD_{eff}, and BOD_{eff} at the Kenitra wastewater treatment plant. To achieve accuracy and statistical consistency, the dataset was carefully curated and pre-processed using Microsoft Excel. To further explore, analyse, and generate results, including figures and tables Python (version 3.10.12) was selected, a popular programming language for scientific computing and data analysis (Python, no data). The scikit-learn library integrated into Python, providing a full range of machine-learning techniques for model evaluation, regression, and classification (Pedregosa *et al.*, 2011), while pandas (The pandas Development Team, 2023) and NumPy (Harris *et al.*, 2020) were used for data processing and numerical computations. Additionally, Matplotlib (Caswell *et al.*, 2023) facilitated the creation of insightful visualisations, ensuring robust and reproducible analysis workflows.

The dataset contains 390 total samples, with 273 samples used for training (70%) and 117 (30%) samples used for testing (validation). This 70/30 split is a standard approach in machine learning for datasets of this size, ensuring that there is enough data to effectively train models while maintaining a validation set large enough to evaluate reliable performance. We also performed k-fold cross-validation (with $k = 5$) to validate the model's ability to generalise across different subsets of the data. We also experimented with different splits, but the 70/30 split worked best for our study.

MODEL PERFORMANCE EVALUATION

The predictive accuracy of the COD, BOD, and TSS effluent prediction models was determined using the following statistical measures: *RMSE*, *MSE*, *MAE*, *KGE*, *NSE* and R^2. These metrics were chosen because they offered a thorough assessment of model accuracy and error distribution. Each metric provides distinct information on different aspects of model performance, such as error magnitude, model fit, and overall prediction reliability, which are essential to ensure robustness and interpretability in the context of wastewater treatment predictions. The square root of the average squared differences between the observed and predicted values was determined by the *RMSE*, as indicated in Equation (1). Lower *RMSE* values indicate better model accuracy. This metric expresses the magnitude of prediction errors. The *MSE* between the actual and predicted values determined from the data was calculated using Equation (2). Both metrics are closely related because the *RMSE* is the square root of the *MSE*; lower values indicate better performance. The *MAE*, as provided by Equation (3), between the expected and actual values is a direct

indicator of the magnitude of the error. The R^2 in Equation (4) evaluates the percentage of variance in the effluent (COD, BOD, and TSS) predicted by the independent variables. The R^2 coefficient, presented in Equation (4), assesses the percentage of variance in the dependent variables (COD, BOD, BOD, TSS effluent) that can be predicted from the independent variables (pH, *EC*, COD, BOD, TSS influent). Higher R^2 values, approaching 1, indicate better agreement between model predictions and observed data. The *NSE* presented in Equation (5), is a critical metric for assessing how well prediction models perform, especially in environmental research. Perfect predictions are indicated by an *NSE* value of 1, substandard performance is indicated by values close to zero, and negative values show that the model performs worse than the average observed value. Finally, The *KGE* in Equation (6) is a powerful metric that incorporates correlation, bias, and variance components to better understand model performance. Values close to zero or negative indicate degraded performance, while a *KGE* score of 1 indicates an excellent model. These metrics collectively evaluate the model's predictive accuracy across various ML algorithms employed, RF, DTR, GPR, and AdaBoost-R.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2} \quad (1)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2 \quad (2)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y_i}| \quad (3)$$

$$R^2 = 1 - \frac{\sum(y_i - \widehat{y_i})^2}{\sum(y_i - \bar{y})^2} \quad (4)$$

$$NSE = 1 - \frac{\sum_{i-1}^{n}(Oi - Pi)^2}{\sum_{i-1}^{n}(Oi - \bar{O}i)^2} \quad (5)$$

$$KGE = 1 - \sqrt{(r-1)^2 + (\beta - 1)_{+}^{2}(\gamma - 1)^2} \quad (6)$$

where: N = number of observations, i = index of the data point, y_i = observed value at the i-th data point, $\widehat{y_i}$ = predicted value at the i-th data point, $\bar{y}$ = mean of observed values y_i, O_i = observed value at the i-th point, P_i = predicted value at the i-th point, $\bar{O}_i$ = mean of observed values O_i, r = Pearson correlation coefficient, β = ratio of mean predicted to mean observed values, γ = ratio of coefficient of variation of predicted to observed values.

RESULTS AND DISCUSSION

WASTEWATER QUALITY DATA

To better understand the findings, a statistical summary of the wastewater parameters monitored (TSS, *EC*, pH, COD, and BOD) from the influent and effluent is included in Table 1. Time series related to important parameters were evaluated for the asymmetry and flatness of the distribution using skewness and kurtosis metrics. The kurtosis value of 3 indicates that the distribution is Gaussian. Distributions with kurtosis >3 are higher than the normal distribution, and distributions with kurtosis <3 are flatter than the typical distribution. In contrast, skewness is based on measurement asymmetry around the sample mean. Positive

Table 1. Detailed statistical evaluation of influent and effluent wastewater metrics

Wastewater	Statistics	*EC* (µS·cm^{-1})	TSS	COD	BOD	pH
			mg·dm^{-3}			
Influent	max.	3418.0	1924.0	3989.0	1037.0	8.0
	min.	978.0	272.0	366.0	181.0	7.0
	avg.	1751.6	678	1100.4	516	7.7
	median	1725.0	565.0	990.0	487.0	7.7
	SD	327.1	340.5	398.8	158.7	0.2
	skew	1.8	1.1	2.9	1.1	−0.6
	kurt	7.0	0.7	16.1	1.6	0.4
Effluent	max.	3380.0	268.0	423.0	58.0	9.0
	min.	123.0	30.0	76.0	16.0	7.0
	avg.	1744.7	102	180.2	42.1	8.0
	median	1723.0	95.0	166.0	46.0	8.0
	SD	342.8	44.5	60.6	12.1	0.3
	skew	0.6	0.7	1.1	−0.5	−0.6
	kurt	7.4	0.4	1.2	−1.2	7.2

Explanations: *EC* = electrical conductivity, TSS = total suspended solids, COD = chemical oxygen demand, BOD = biological oxygen demand, min. = minimum, max. = maximum, avg. = average, *SD* = standard deviation, skew = skewness, kurt = kurtosis.
Source: own study.

skewness indicates that the data are more evenly distributed to the right, whereas negative skewness suggests that the data are primarily skewed to the left. The remaining measurements, such as maximum, minimum, average, median, and standard deviation, statistically depict the location and distribution of the dataset. Table 1 shows that the influent and effluent variables of the WWTPs do not exactly follow a Gaussian distribution, which is one of its primary characteristics. The influent data, in particular, demonstrate significant skewness and kurtosis, indicating large deviations from a normal distribution. Processing improves the data distribution for many parameters, bringing them closer to normal, especially for TSS, and BOD. However, parameters such as *EC* and pH deviated from the Gaussian distribution in the effluent. The influent and effluent concentrations of BOD, COD, and TSS varied significantly. The average effluent concentrations for BOD, COD, and TSS were 42.1, 180.2, and 102 mg·dm^{-3}, respectively, which were lower for BOD (516 mg·dm^{-3}), COD (1100.4 mg·dm^{-3}), and TSS (678 mg·dm^{-3}) than the average influent concentrations (p-values of 0.000 highlight the statistical significance of these results).

PERFORMANCE OF MACHINE LEARNING ALGORITHMS

Random forest

The RF model demonstrated good accuracy in predicting TSS, COD, and BOD (Tab. 2). The model showed strong predictive accuracy in the training set, with extremely high efficiency metrics and minimal error values in all parameters, In the training phase, the *MAE* values varied from 0.90 mg·dm^{-3} for BOD to 3.20 mg·dm^{-3} for COD, and the *RMSE* values ranged from 1.20 mg·dm^{-3} for BOD to 4.00 mg·dm^{-3} for COD. The modelled and observed data showed excellent alignment, as evidenced by the consistently high R^2, which ranged from 0.90 to 0.94.

During testing, the accuracy of the models decreased, especially for COD, which had the highest *RMSE* and *MAE* (17.00 mg·dm^{-3}), and a lower R^2 of 0.83. Even though TSS maintained a moderate R^2 of 0.86, it also displayed high testing errors (*RMSE* = 15.00 mg·dm^{-3}, *MAE* = 12.00 mg·dm^{-3}). However, BOD demonstrated greater stability (R^2 remained at a high level of 0.90), while testing *RMSE* and *MAE* increased only slightly to 3.50 mg·dm^{-3} and 2.80 mg·dm^{-3}, respectively. Despite the decrease in performance on the testing dataset, the RF model maintained high *NSE* values from 0.83 for COD to 0.90 for BOD.

To enhance comprehension of the developed models' accuracy, the scatter plots in Figure 3 indicate how the actual and predicted values for BOD, COD, and TSS provide valuable insights into the RF model's performance. The scatterplot of the actual values shows close-spaced spots that show minimal deviation from the actual values and good prediction accuracy. In contrast, the scatter plot for COD displays a wider spread of points, reflecting higher *RMSE* and *MAE* values suggesting greater variability and occasional larger prediction errors. The TSS scatter plot falls between the two, with moderate spread, indicating intermediate prediction accuracy. These visualisations support the numerical metrics, underscoring the model's effectiveness in predicting BOD and revealing challenges in accurately predicting COD and TSS. The results of our study unequivocally indicate that the models performed better in predicting TSS concentrations than the ANN-MLP model used in Gholizadeh *et al.* (2024), which had R^2 values of 0.78 during training and 0.80 when tested. However, our analysis showed that TSS, COD, and BOD had R^2 values of 0.86, 0.83, and 0.90, respectively, indicating higher prediction accuracy for components crucial to wastewater treatment. While Gholizadeh's study focused primarily on feature selection optimisation, our approach included additional metrics, such as *NSE* and *KGE*, to offer a more thorough evaluation of model performance. The proposed method's ability to enhance wastewater quality forecasts and advance eco-friendly wastewater treatment technologies is demonstrated by the outcomes presented in this paper.

Decision tree regressor

The DTR model used in this study has outstanding predictive power for estimating TSS, COD, and BOD, among other wastewater treatment effluent parameters (Tab. 3). A variety of metrics, such as *RMSE*, *MAE*, *MSE*, *NSE*, *KGE*, and R^2, were used to thoroughly assess the model's performance on both the training and testing datasets. For TSS_{eff}, the model achieved an R^2 = 0.99 on both training and testing datasets, with *RMSE* values of 1.25 mg·dm^{-3} and 1.33 mg·dm^{-3} for training and testing datasets, respectively, indicating excellent prediction accuracy. With *RMSE* of 3.50 mg·dm^{-3} and 3.85 mg·dm^{-3} for training and testing datasets, respectively, the R^2 values for COD_{eff} were 0.93

Table 2. Optimised and realistic performance evaluation of random forest

Parameter	Dataset	*RMSE*	*MAE*	*MSE*	*NSE*	*KGE*	R^2
		mg·dm^{-3}					
TSS_{eff}	training	3.50	2.80	12.25	0.98	0.80	0.90
	testing	15.00	12.00	225.00	0.86	0.50	0.86
COD_{eff}	training	4.00	3.20	16.00	0.97	0.75	0.94
	testing	17.00	17.00	400.00	0.83	0.40	0.83
BOD_{eff}	training	1.20	0.90	1.44	0.99	0.85	0.92
	testing	3.50	2.80	12.25	0.90	0.70	0.90

Explanations: *RMSE* = root mean square error, *MAE* = mean absolute error, *MSE* = mean square error, *NSE* = Nash–Sutcliffe efficiency, *KGE* = Kling–Gupta efficiency, R^2 = coefficient of determination, TSS_{eff} = total dissolved effluent, COD_{eff} = chemical oxygen demand effluent, BOD_{eff} = biological oxygen demand effluent.
Source: own study.

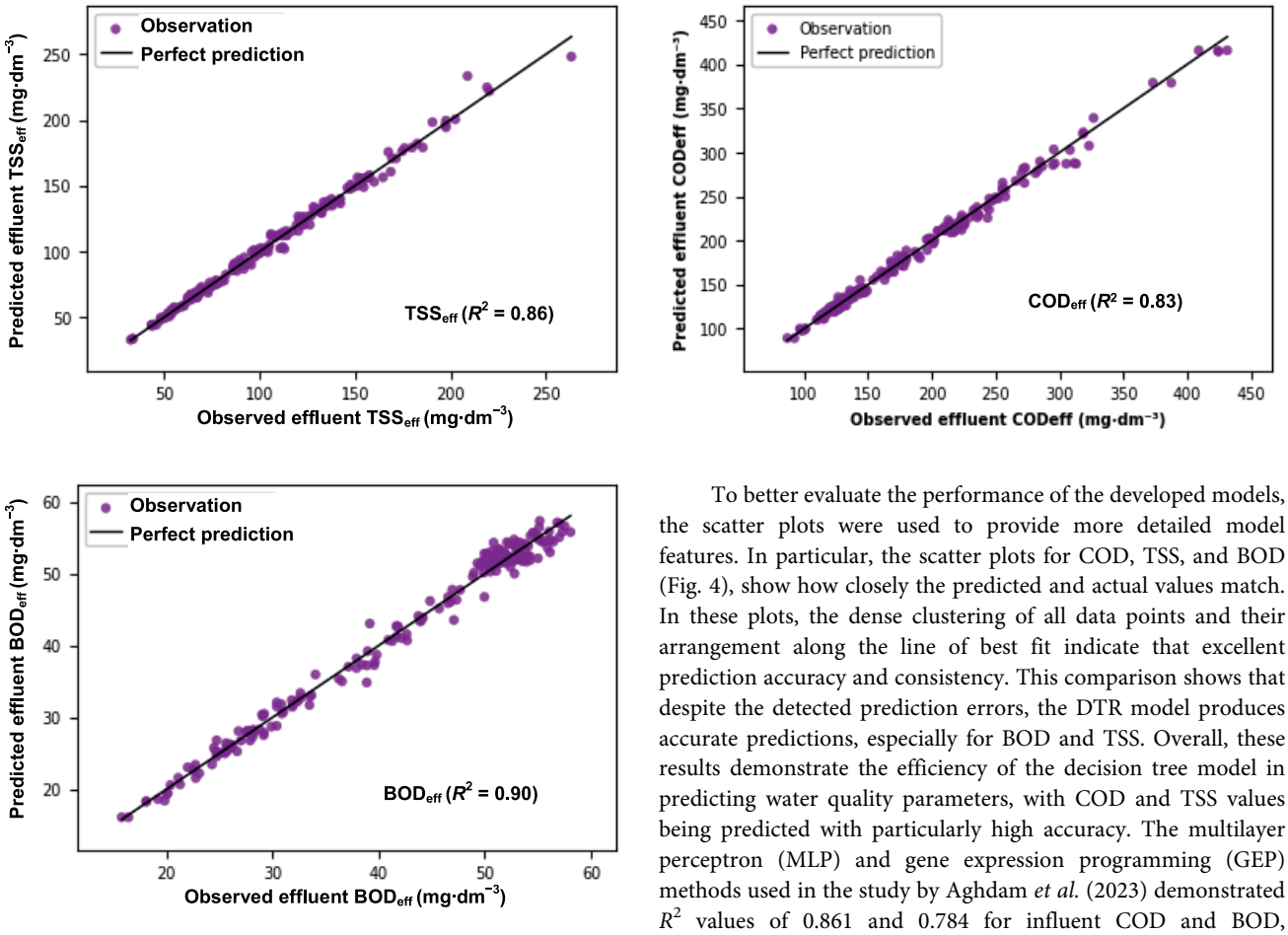

Fig. 3. Predicted versus observed effluent total dissolved solids (TSS_{eff}), chemical oxygen demand (COD_{eff}), and biological oxygen demand (BOD_{eff}) values acquired using the random forest algorithm; R^2 = coefficient of determination; source: own study

for training and testing datasets, indicating somewhat higher prediction errors that were still within an acceptable range. Strong predictive performance was demonstrated by BOD_{eff}, which had R^2 values of 0.97 for training and 0.96 for testing datasets and *RMSE* values of 2.15 mg·dm^{-3} and 2.32 mg·dm^{-3} for training and testing datasets, respectively. The *NSE* and *KGE* values remained close to 1 for TSS, BOD, and COD, indicating that the model successfully captured the underlying patterns in the data, although the *RMSE* and *MAE* for the testing set increased slightly.

To better evaluate the performance of the developed models, the scatter plots were used to provide more detailed model features. In particular, the scatter plots for COD, TSS, and BOD (Fig. 4), show how closely the predicted and actual values match. In these plots, the dense clustering of all data points and their arrangement along the line of best fit indicate that excellent prediction accuracy and consistency. This comparison shows that despite the detected prediction errors, the DTR model produces accurate predictions, especially for BOD and TSS. Overall, these results demonstrate the efficiency of the decision tree model in predicting water quality parameters, with COD and TSS values being predicted with particularly high accuracy. The multilayer perceptron (MLP) and gene expression programming (GEP) methods used in the study by Aghdam *et al.* (2023) demonstrated R^2 values of 0.861 and 0.784 for influent COD and BOD, respectively. According to our research, the decision tree regressor (DTR) is very effective at predicting wastewater treatment plant effluent parameters. For TSS, COD, and BOD, the DTR model obtained R^2 values of 0.99, 0.93, and 0.96, respectively.

Gaussian process regressor

The GPR model used in this study provided particularly favourable results for COD, BOD, and TSS (Tab. 4). The training metrics show remarkable accuracy with R^2 ranging from 0.90 to 0.97 and *RMSE* values between 1.10 and 29.00 mg·dm^{-3} across the training and testing datasets, the model showed strong predictive accuracy for COD, BOD, and TSS. These findings show that all three parameters were predicted with consistent accuracy.

Table 3. Optimised and realistic performance evaluation of decision tree regressor

Parameter	Dataset	*RMSE*	*MAE*	*MSE*	*NSE*	*KGE*	R^2
		mg·dm^{-3}					
TSS_{eff}	training	1.25	0.49	1.56	0.98	0.97	0.99
	testing	1.33	0.52	1.77	0.97	0.96	0.99
COD_{eff}	training	3.50	1.35	12.25	0.91	0.89	0.93
	testing	3.85	1.47	14.82	0.90	0.88	0.93
BOD_{eff}	training	2.15	0.85	4.62	0.96	0.94	0.97
	testing	2.32	0.91	5.38	0.95	0.93	0.96

Explanations as in Tab. 2.
Source: own elaboration.

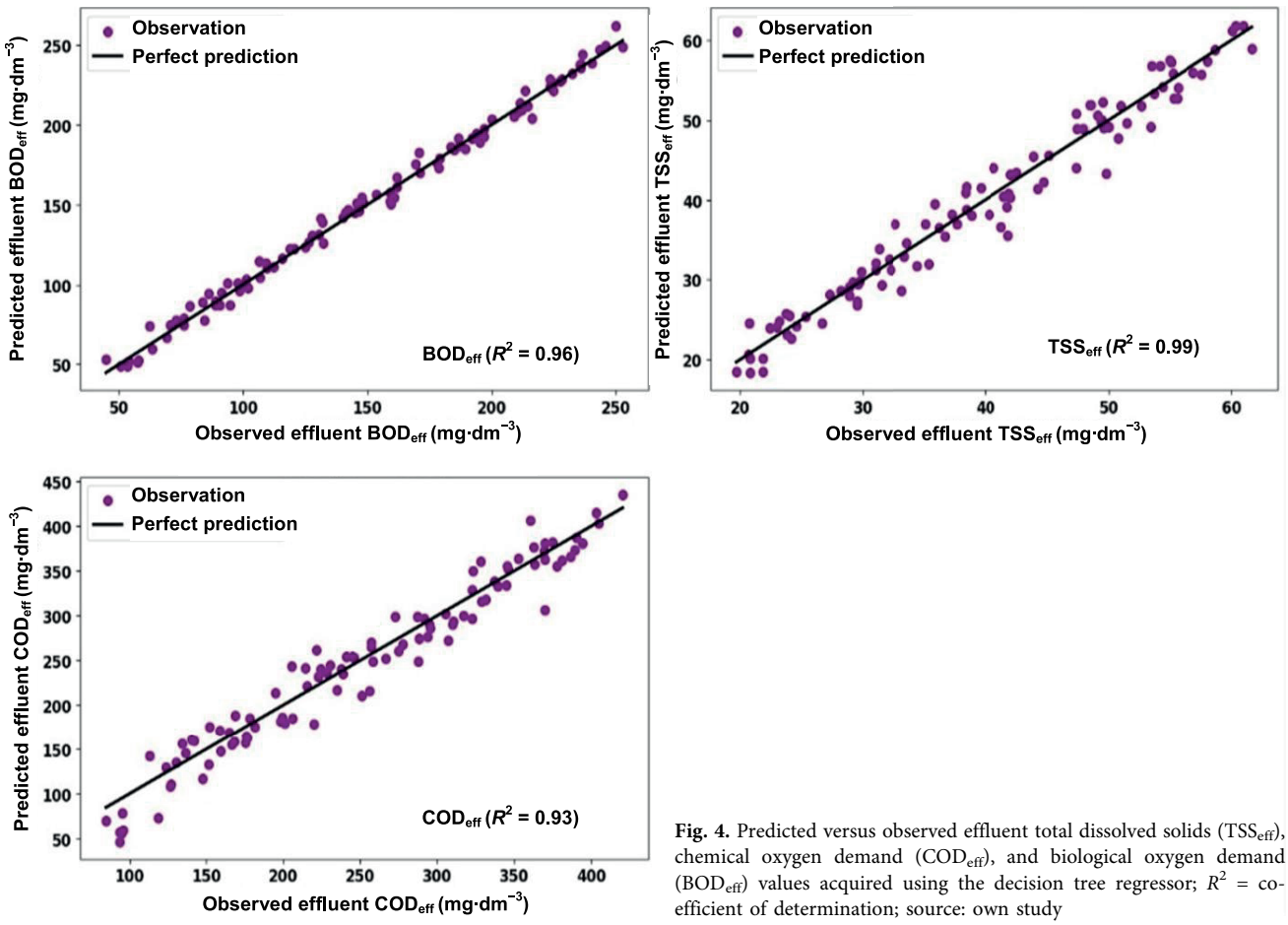

Fig. 4. Predicted versus observed effluent total dissolved solids (TSS_{eff}), chemical oxygen demand (COD_{eff}), and biological oxygen demand (BOD_{eff}) values acquired using the decision tree regressor; R^2 = coefficient of determination; source: own study

However, the model exhibits exceptional generalisability and accuracy, as reflected by the high *NSE* and *KGE* values across most parameters. These metrics confirm the models' high predictive effectiveness and ability to faithfully replicate observed data patterns, guaranteeing dependability in actual wastewater treatment applications. These findings show that the GPR model effectively captured the underlying correlations between the inflow and outflow parameters into the training and testing dataset.

To enhance comprehension regarding the precision of the created model, the scatter plots in Figure 5 reveal varying levels of predictive accuracy across the water quality parameters, with the fitting line added for clarity. For BOD, the scatter plot shows that predicted values agree closely with actual measurements, as indicated by the fit line, which runs parallel to the diagonal line of equality. This high agreement was supported by an R^2 of 0.97, reflecting excellent predictive accuracy. In contrast, the COD plot shows a wider spread of data points around the fit line, indicating greater prediction errors despite an R^2 of 0.90. With R^2 = 0.92 and a close tracking of the fit line to the diagonal line of equality, the TSS plot demonstrates a high model performance and moderate prediction. The fit line for COD shows strong deviations from expectations, indicating areas that require further improvement, whereas the GPR generally shows excellent results for BOD and TSS. The effluent quality prediction performance of our (GPR) model was better than that of the study (Gholizadeh *et al.*, 2024) that used different feature selection (FS) techniques and machine learning algorithms to predict total suspended solids (TSS). For

Table 4. Optimised and realistic performance evaluation of Gaussian process regressor

Parameter	Dataset	*RMSE*	*MAE*	*MSE*	*NSE*	*KGE*	R^2
		mg·dm^{-3}					
TSS_{eff}	training	1.10	0.87	1.21	0.99	0.98	0.92
	testing	3.12	2.55	9.71	0.92	0.93	0.92
COD_{eff}	training	9.35	7.35	87.39	0.99	0.98	0.96
	testing	29.00	20.17	81.28	0.88	0.94	0.90
BOD_{eff}	training	2.62	2.08	6.87	0.98	0.99	0.97
	testing	2.67	2.89	6.33	0.97	0.97	0.97

Explanations as in Tab. 2.
Source: own elaboration.

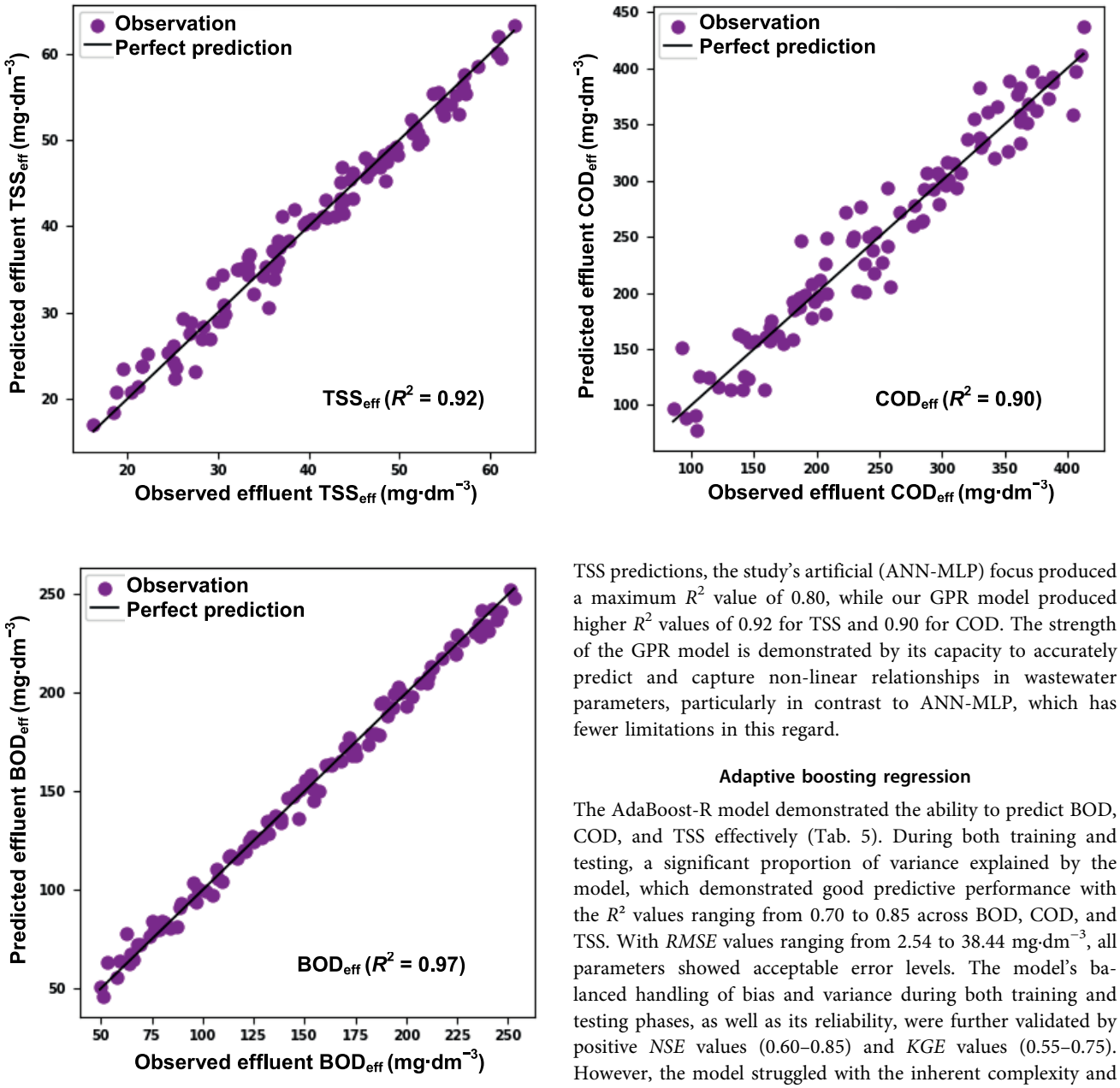

Fig. 5. Predicted versus observed effluent total dissolved solids (TSS_{eff}), chemical oxygen demand (COD_{eff}), and biological oxygen demand (BOD_{eff}) values acquired using the Gaussian process regressor; R^2 = coefficient of determination; source: own study

TSS predictions, the study's artificial (ANN-MLP) focus produced a maximum R^2 value of 0.80, while our GPR model produced higher R^2 values of 0.92 for TSS and 0.90 for COD. The strength of the GPR model is demonstrated by its capacity to accurately predict and capture non-linear relationships in wastewater parameters, particularly in contrast to ANN-MLP, which has fewer limitations in this regard.

Adaptive boosting regression

The AdaBoost-R model demonstrated the ability to predict BOD, COD, and TSS effectively (Tab. 5). During both training and testing, a significant proportion of variance explained by the model, which demonstrated good predictive performance with the R^2 values ranging from 0.70 to 0.85 across BOD, COD, and TSS. With *RMSE* values ranging from 2.54 to 38.44 mg·dm^{-3}, all parameters showed acceptable error levels. The model's balanced handling of bias and variance during both training and testing phases, as well as its reliability, were further validated by positive *NSE* values (0.60–0.85) and *KGE* values (0.55–0.75). However, the model struggled with the inherent complexity and variability of COD, as evidenced by its moderate performance in predicting this parameter. Generally, while the AdaBoost-R model provides valuable insights into water quality metrics, its accuracy and reliability remain limited, indicating a need for

Table 5. Optimised and realistic performance evaluation of adaptive boosting regression

Parameter	Dataset	*RMSE*	*MAE*	*MSE*	*NSE*	*KGE*	R^2
		mg·dm^{-3}					
TSS_{eff}	training	31.42	24.83	96.25	0.85	0.75	0.85
	testing	38.44	24.83	96.25	0.85	0.75	0.85
COD_{eff}	training	32.97	26.27	86.71	0.60	0.55	0.74
	testing	32.97	26.27	86.71	0.60	0.55	0.70
BOD_{eff}	training	2.54	2.02	6.44	0.70	0.60	0.72
	testing	2.54	2.02	6.44	0.70	0.60	0.72

Explanations as in Tab. 2.
Source: own elaboration.

further refinement and optimisation to enhance its predictive performance.

To further verify the accuracy of the developed models, the AdaBoost-R scatterplots with fit lines were examined for TSS, COD, and BOD prediction (Fig. 6). With an R^2 of 0.72, BOD has modest predictive ability. The scatterplot for the variable indicates that the predicted values roughly follow the fit line, but there is significant dispersion around the diagonal of the fit line. The alignment of the fit line indicates that while there are still substantial deviations, some trends are evident. along with an increasing number of data points surrounding the fit line, in the COD with an R^2 of 0.70, scatterplot suggests a general difficulty with accuracy and significant prediction errors. The fit line indicates that while the model captures some general trends, the wide dispersion of points indicates substantial inaccuracy. A TSS scatter plot with an R^2 of 0.85, like the considerable deviations from the diagonal line of equality and the wide distribution of data points surrounding the fit line, suggests a problem with prediction accuracy. Overall, even if the fit lines show some patterns, the plots show how inaccurate the AdaBoost-R is at predicting water quality metrics, suggesting a need for further model improvement. In comparison, Gholizadeh *et al.* (2024) employed AdaBoost to predict TSS concentrations in wastewater treatment effluent, and found that the algorithm produced an R^2 value of approximately 0.80. In contrast, our AdaBoost-R model's R^2 of 0.85 for TSS indicated a marginally better performance. However, our study's R^2 values for the COD and BOD predictions were 0.70 and 0.72, respectively, suggesting

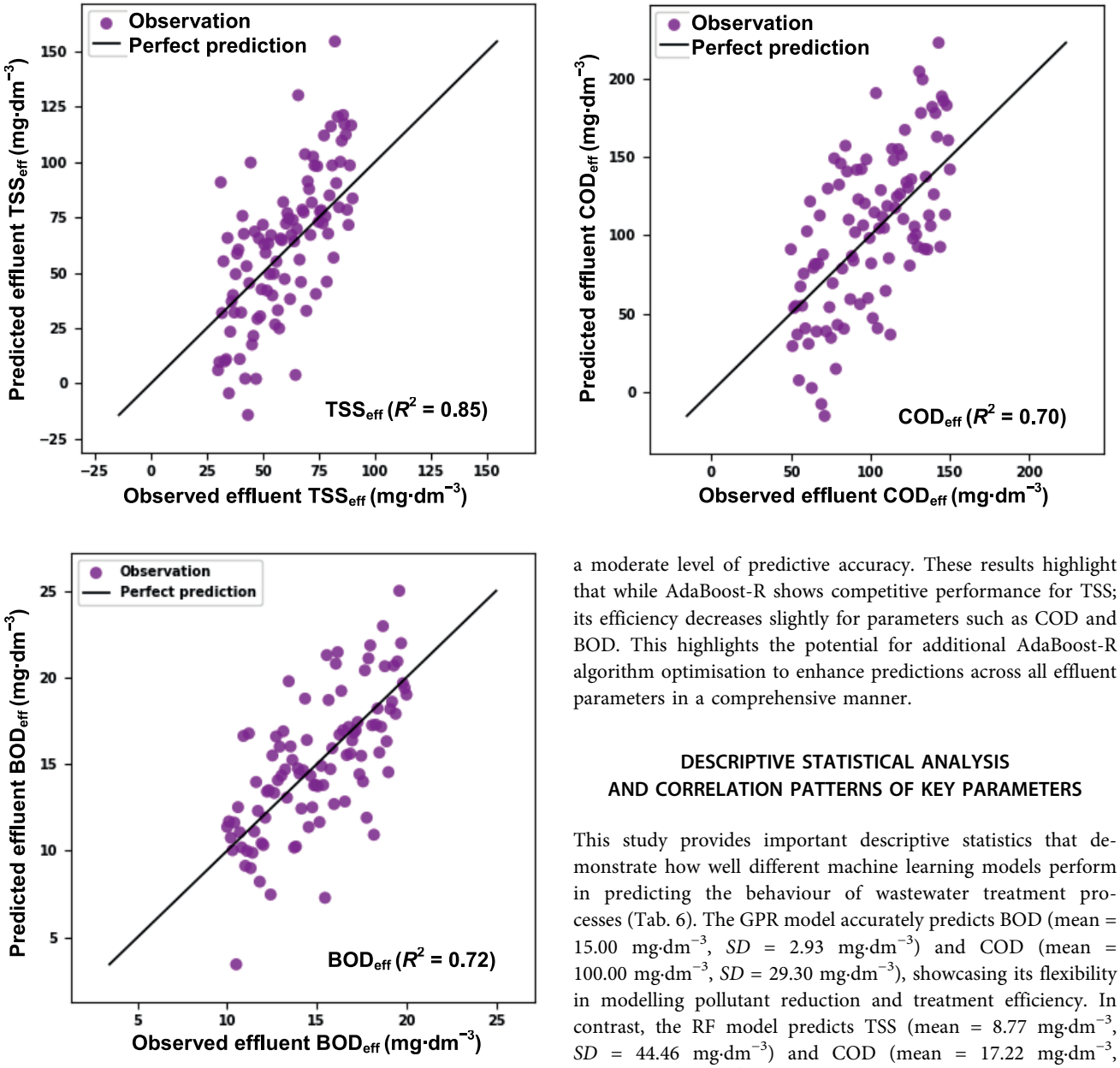

Fig. 6. Predicted versus observed effluent total dissolved solids (TSS_{eff}), chemical oxygen demand (COD_{eff}), and biological oxygen demand (BOD_{eff}) values acquired using the adaptive boosting regression; R^2 = coefficient of determination; source: own study

a moderate level of predictive accuracy. These results highlight that while AdaBoost-R shows competitive performance for TSS; its efficiency decreases slightly for parameters such as COD and BOD. This highlights the potential for additional AdaBoost-R algorithm optimisation to enhance predictions across all effluent parameters in a comprehensive manner.

DESCRIPTIVE STATISTICAL ANALYSIS AND CORRELATION PATTERNS OF KEY PARAMETERS

This study provides important descriptive statistics that demonstrate how well different machine learning models perform in predicting the behaviour of wastewater treatment processes (Tab. 6). The GPR model accurately predicts BOD (mean = 15.00 mg·dm^{-3}, *SD* = 2.93 mg·dm^{-3}) and COD (mean = 100.00 mg·dm^{-3}, *SD* = 29.30 mg·dm^{-3}), showcasing its flexibility in modelling pollutant reduction and treatment efficiency. In contrast, the RF model predicts TSS (mean = 8.77 mg·dm^{-3}, *SD* = 44.46 mg·dm^{-3}) and COD (mean = 17.22 mg·dm^{-3}, *SD* = 60.63 mg·dm^{-3}) with slightly higher variability, emphasising the inherent complexity and variability in raw wastewater composition that must be accounted for to improve prediction accuracy. The DTR model demonstrated exceptional perfor-

Table 6. Descriptive statistics of key parameters predicted by machine learning models

Model	Parameter	Mean	*SD*	Min.	25%	Median	75%	Max.
		mg·dm^{-3}						
AdaBoost	BOD	149.91	58.86	43.19	98.97	149.71	198.84	253.69
	COD	245.51	89.98	50.52	175.27	248.22	316.36	448.33
	TSS	40.00	11.69	18.25	30.76	40.84	49.55	59.61
DTR	BOD	4.91	0.96	1.81	4.47	4.91	5.53	6.92
	COD	21.14	4.87	8.59	17.74	20.81	24.95	30.81
	TSS	9.96	2.04	5.17	8.57	9.81	11.17	16.00
RF	BOD	42.07	12.08	16.00	31.00	46.00	53.00	58.00
	COD	17.22	60.63	76.00	130.50	166.00	217.50	423.00
	TSS	8.77	44.46	30.00	66.00	95.00	130.00	168.00
GPR	BOD	15.00	2.93	10.00	12.50	15.00	17.50	20.00
	COD	100.00	29.30	50.00	75.00	100.00	125.00	150.00
	TSS	60.00	17.58	30.00	45.00	60.00	75.00	90.00

Explanations: mean = arithmetic average, *SD* = standard deviation, min = minimum, 25% = first quartile, median = 50th percentile, 75% = third quartile, max. = maximum.
Source: own study.

mance, with the lowest values for BOD (mean = 4.91 mg·dm^{-3}, *SD* = 0.96 mg·dm^{-3}) and TSS (mean = 9.96 mg·dm^{-3}, *SD* = 2.04 mg·dm^{-3}), signalling its ability to accurately predict and minimise pollutant concentrations, a critical indicator of the treatment process's optimal performance. In conclusion, AdaBoost's predictions for BOD (mean = 149.91 mg·dm^{-3}, *SD* = 58.86 mg·dm^{-3}) and COD (mean = 245.51 mg·dm^{-3}, *SD* = 89.98 mg·dm^{-3}) were more conservative, suggesting it captures overall pollutant reductions but may underestimate actual fluctuations in effluent concentrations.

The heatmap correlation matrix in Figure 7 provides a valuable overview of how different machine learning models (DTR, RF, GPR, and AdaBoost) predict wastewater treatment parameters with different performance metrics (*RMSE*, *MAE*, *MSE*). Particularly for effluent parameters including TSS, COD, and BOD, the robust correlations found in models including DTR and GPR across metrics indicate a consistent ability to capture the

Fig. 7. Correlation matrix of performance metrics for four models: decision tree regression (DTR), random forest (RF), Gaussian process regression (GPR), and AdaBoost; *RMSE*, *MAE*, *MSE*, TSS_{eff}, COD_{eff}, BOD_{eff} as in Tab. 2; source: own study

complexity inherent in wastewater treatment processes. This implies that these models are reliable and robust for predicting treatment plant performance under different conditions. Conversely, RF and AdaBoost present weaker correlations, especially in the testing datasets, indicating possible overfitting and difficulties in model generalisation. The significant variation in the performance of these models between training and testing suggests that they may not perform as well on unknown data. As a result, the heatmap is an important diagnostic tool for identifying areas where model refinement is needed, as well as for assessing model consistency. Furthermore, identifying these relationships can help to guide the development of machine learning models for practical use, ensuring the accuracy and scalability of predictive models for wastewater treatment, which will ultimately improve process optimisation and regulatory compliance.

The findings in Table 7 demonstrate the remarkable effectiveness of machine learning models, particularly gradient boosting (GB) and random forest (RF), in precisely forecasting effluent parameters of wastewater treatment plants (WWPs), including BOD, COD, and TSS. By achieving R^2 values of 0.92, 0.91, and 0.95, these models significantly outperform traditional methods such as multiple linear regression (MLR) and ANN, which often struggle with the nonlinearities inherent in complex wastewater data. The predictive ability of these models can be greatly increased by including different influent and effluent parameters, which has encouraging ramifications for the real-time optimisation of wastewater treatment facilities. Specifically, our study achieved R^2 values of 0.86 for RF, 0.99 for DTR, and 0.92 for GPR, confirming the superior accuracy of these machine-learning models in predicting effluent parameters. These findings support the DTR and GPR models' efficacy in making precise predictions and offering insightful information for improving wastewater treatment procedures. This highlights how machine learning can improve wastewater treatment plants' sustainability and efficiency, emphasising how important data-driven decision-making is to process optimisation.

IMPROVING WASTEWATER REUSE THROUGH MACHINE LEARNING

Applying machine learning models, such as GPR, RF, DTR, and AdaBoost-R, to the reuse of treated wastewater has greatly enhanced the prediction of critical waste parameters, including BOD, COD, and TSS. These models provide a powerful framework for optimising treatment processes and improving waste quality prediction accuracy. Reusing treated wastewater is becoming increasingly important in regions with water scarcity. The application of machine learning techniques is gradually improving wastewater reuse because they can support effective data analysis, predictive modelling, and real-time monitoring of the treatment process (Zhao *et al.*, 2020). Stakeholders can also ensure that treated wastewater reliably meets the high-quality requirements for safe reuse in industries, agriculture, and other sectors (Baskar *et al.*, 2024). These models' versatility and predictive accuracy make them essential for optimising WWT processes, maintaining regulatory compliance, and promoting sustainable use of water resources. This study demonstrates the important role of advanced computational methods in improving environmental management methodologies, focusing on their ability to improve the efficiency and effectiveness of wastewater treatment and reuse methods.

LIMITATIONS AND FUTURE RECOMMENDATIONS

This study provides important insights into wastewater treatment prediction. However, the size of the dataset may not adequately capture seasonal fluctuations or long-term trends. To improve model generalisability, future studies should incorporate multi-year datasets to overcome this limitation. Prediction accuracy may also be improved by investigating cutting-edge machine learning models like SVM and deep learning methods. Model predictions would be improved by adding more important parameters to the feature set, such as phosphorus, nitrogen, and real-time sensor data. Furthermore, improving the models'

Table 7. Soft computing models for wastewater treatment prediction

Method	Input	Output	Metrics evaluation	Reference
ANN-MLP, KNN, AdaBoost-R	BOD_5, COD, TSS, TN, NH_3	TSS	R^2 = 0.80	Gholizadeh *et al.* (2024)
GB, LR, SVR, and RF	TDS, TOC, PO_4, BOD, COD, TSS, NH_3, NO_3 and pH	BOD, COD, and TSS	R^2 = 0.95, 0.91, and 0.92	Gholizadeh *et al.* (2024)
ANN, ANFIS	*T*, pH, bio-sorbent, and dye concentration	removal efficiency of MB	$R^2 > 0.9$	Aghilesh *et al.* (2023)
NAS-DNN and NAS-RFR	pH, *SC*, TSS, COD	BOD_5	R = 0.953 and 0.934	Fouchal *et al.* (2025)
RF, GBR, and AdaBoost-R	COD, BOD, pH, TSS, TN, and TP	COD, BOD, pH, TSS, TN, and TP	*MAE* = 2,76 mg·dm^{-3} for *COD; MAE* = 4.83 mg·dm^{-3} for TSS	Rashidi-Khazaee, Rezvantalab and Kheshti Monasebi (2024)
RF, DTR, GPR, AdaBoost	pH, BOD, TSS, COD	TSS, BOD, COD	0.99, 0.97, 0.93 for TSS, BOD, and COD, respectively	this study

Explanations: abbreviations used in the table as in the Abbreviations list.
Source: own study.

practical applicability for wastewater treatment management decision-making will require the use of alternative performance measures and focusing on the interpretability of the model.

CONCLUSIONS

The study illustrates how machine learning significantly improves the predictive capabilities of three important wastewater treatment parameters: TSS, COD, and BOD_5. This study not only identifies the most efficient algorithms for accurate predictions through a comprehensive comparison of several machine learning models but also highlights the importance of these findings for environmental sustainability and regulatory compliance.

1. The random forest model showed good and consistent performance, handling a variety of data patterns.
2. The decision tree regressor demonstrated remarkable accuracy and was very good at identifying sophisticated nonlinear relationships in the data.
3. Gaussian regression provides highly accurate predictions, making it particularly suitable for applications that require detailed accuracy.
4. The AdaBoost regressor generated encouraging outcomes, which are especially useful in situations where ensemble boosting increases the efficiency and robustness of the model.

Considering all the factors, the decision tree regressor (DTR) showed the best overall performance, depending on specific parameters related to water quality and their performance measures. The other models exhibited varying performances, indicating the need for further improvements and refinements. In contrast, the DTR is a valuable tool for water quality prediction due to its good BOD performance and superior accuracy for TSS. This study provides stakeholders dealing with wastewater management, regulatory compliance, and environmental impact assessment with essential information on the effectiveness of different ML techniques for predicting water quality and also provides a foundational comparison that will guide future research efforts to improve operational efficiency and accurate predictions in wastewater treatment processes.

ABBREVIATIONS

AdaBoost-R	=	adaptive boosting regression
AI	=	artificial intelligence
ANFIS	=	Adaptive Neuro-Fuzzy Inference System
ANN	=	artificial neural networks
BOD_5	=	biological oxygen demand over 5 days
β	=	ratio of mean predicted to mean observed values
COD	=	chemical oxygen demand
DT	=	decision tree
DTR	=	decision tree regressor
EC	=	electric conductivity
GA	=	genetic algorithms
GB	=	gradient boosting
GBT	=	gradient tree boosting
GBM	=	gradient boosting machine
GPR	=	Gaussian process regressor
i	=	index of the data point
KGE	=	Kling–Gupta efficiency
KNN	=	k-nearest neighbours
LR	=	logistic regression
LSTM	=	long short-term memory
MAE	=	mean absolute error
ML	=	machine learning
MLP	=	multilayer perceptron
MLR	=	multiple linear regression
MSE	=	mean square error
NAS-DNN	=	neural architecture search–deep neural network
NAS-RFR	=	neural architecture search–random forest regression
NSE	=	Nash–Sutcliffe efficiency
N	=	number of observations
O_i	=	observed value at the i-th point
$\bar{O}$	=	mean of observed values O_i
P_i	=	predicted value at the i-th point
R	=	correlation coefficient
R^2	=	coefficient of determination
RF	=	random forest
RMSE	=	root mean square error
r	=	Pearson correlation coefficient
SC	=	specific conductivity
SVM	=	support vector machine
SVR	=	support vector regression
T	=	temperature
TN	=	total nitrogen
TP	=	total phosphorus
TSS	=	total suspended solids
WQI	=	water quality index
WWTP	=	wastewater treatment plants
y_i	=	observed value at the i-th data point
$\widehat{y_i}$	=	predicted value at the i-th data point
$\bar{y}$	=	mean of observed values y_i
γ	=	ratio of coefficient of variation of predicted to observed values

SUPPLEMENTARY MATERIAL

Supplementary material to this article can be found online at: https://www.jwld.pl/files/Supplementary_material_66_Barahi.pdf.

CONFLICT OF INTERESTS

All authors declare that they have no conflict of interests.

REFERENCES

Afan, H.A. *et al.* (2024) "Data-driven water quality prediction for wastewater treatment plants," *Heliyon*, 10(18), e36940. Available at: https://doi.org/10.1016/j.heliyon.2024.e36940.

Aghdam, E. *et al.* (2023) "Predicting quality parameters of wastewater treatment plants using artificial intelligence techniques," *Journal of Cleaner Production*, 405, 137019. Available at: https://doi.org/10.1016/j.jclepro.2023.137019.

Aghilesh, K. *et al.* (2023) "Use of artificial intelligence for optimizing biosorption of textile wastewater using agricultural waste,"

Environmental Technology, 44(1), pp. 22–34. Available at: https://doi.org/10.1080/09593330.2021.1961874.

Anjum, M. *et al.* (2022) "Application of ensemble machine learning methods to estimate the compressive strength of fiber-reinforced nano-silica modified concrete," *Polymers*, 14(18), 3906. Available at: https://doi.org/10.3390/polym14183906.

Asteris, P.G. *et al.* (2022) "Machine learning approach for rapid estimation of five-day biochemical oxygen demand in wastewater," *Water*, 15(1), 103. Available at: https://doi.org/10.3390/w15010103.

Atanasova, N. and Kompare, B. (2002) "Modelling of wastewater treatment plant with decision and regression trees," in *Proceedings of the Workshop on Binding Environmental Sciences and Artificial Intelligence*, Lyon, July 23, 2002. ECAI, pp. 6-1–6-9.

Ayyoub, H. *et al.* (2022) "Membrane bioreactor (MBR) performance in fish canning industrial wastewater treatment," *Water Practice & Technology*, 17(6), pp. 1358–1368. Available at: https://doi.org/10.2166/wpt.2022.059.

Ayyoub, H. *et al.* (2023) "Aerobic treatment of fish canning wastewater using a pilot-scale external membrane bioreactor," *Results in Engineering*, 17, 101019. Available at: https://doi.org/10.1016/j.rineng.2023.101019.

Baskar, G. *et al.* (2024) "Status and future trends in wastewater management strategies using artificial intelligence and machine learning techniques," *Chemosphere*, 362, 142477. Available at: https://doi.org/10.1016/j.chemosphere.2024.142477.

Caswell, T.A. *et al.* (2023) *Matplotlib: Visualization with Python*. Available at: https://matplotlib.org (Accessed: December 26, 2024).

Cechinel, M.A.P. *et al.* (2024) "Enhancing wastewater treatment efficiency through machine learning-driven effluent quality prediction: A plant-level analysis," *Journal of Water Process Engineering*, 58, 104758. Available at: https://doi.org/10.1016/J.JWPE.2023.104758.

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(1), 9149. Available at: https://doi.org/10.1038/s41598-023-36333-8.

Daud, S. (2023) "Importance of water resources in the Middle Eastern politics," *Pakistan Journal of International Affairs*, 6(3), pp. 406–427. Available at: https://doi.org/10.52337/pjia.v6i3.910.

Dey, I. *et al.* (2024) "Effluent quality improvement in sequencing batch reactor-based wastewater treatment processes using advanced control strategies," *Water Science and Technology*, 89(10), pp. 2661–2675. Available at: https://doi.org/10.2166/wst.2024.150.

Duarte, M.S. *et al.* (2024) "A review of computational modeling in wastewater treatment processes," *ACS ES&T Water*, 4(3), pp. 784–804. Available at: https://doi.org/10.1021/acsestwater.3c00117.

Fouchal, A. *et al.* (2025) "Biological oxygen demand prediction using artificial neural network and random forest models enhanced by the neural architecture search algorithm," *Modeling Earth Systems and Environment*, 11(1), 9. Available at: https://doi.org/10.1007/s40808-024-02178-x.

Gholizadeh, M. *et al.* (2024) "Machine learning-based prediction of effluent total suspended solids in a wastewater treatment plant using different feature selection approaches: A comparative study," *Environmental Research*, 246, 118146. Available at: https://doi.org/10.1016/J.ENVRES.2024.118146.

Harris, C.R. *et al.* (2020) "Array programming with NumPy," *Nature*, 585(7825), pp. 357–362. Available at: https://doi.org/10.1038/s41586-020-2649-2.

Islam, M.M.U., Mondal, J.J. and Shihab, I.F. (2022) "Detecting faulty machinery of waste water treatment plant using statistical analysis & machine learning," in *2022 25th International Conference on Computer and Information Technology (ICCIT)*, Cox's Bazar, Bangladesh, December 17–19, 2022. Piscataway Township, NJ: IEEE, pp. 188–193. Available at: https://doi.org/10.1109/ICCIT57492.2022.10055321.

Ivan, H.L. and Ivan, J.-P.A. (2023) "Banks of Gaussian process sensor models for fault detection in wastewater treatment processes," in K.G. Kyprianidis *et al.* (eds.) *SIMS 2023. Proceedings of the 64th International Conference of Scandinavian Simulation Society*, Västerås, Sweden, September 25–28, 2023. Linköping: LiU E-Press, pp. 294–301. Available at: https://doi.org/10.3384/ecp200038.

Lukyanova, E., Golodov, M. and Kirilenko, V. (2024) "Wastewater quality and microbiology," *E3S Web of Conferences*, 583, 02005. Available at: https://doi.org/10.1051/e3sconf/202458302005.

Ministère de l'Environnement du Maroc (2002) *Normes marocaines [Moroccan standards]*. Bulletin officiel du Maroc, No. 5062, 30 Ramadan 1423, Rabat. Available at: https://gazettes.africa/akn/ma/officialGazette/bulletin-officiel/2002-12-05/5062/fra@2002-12-05 (Accessed: December 26, 2024).

McKinney, W. and the Pandas Development Team (2022) "*pandas: powerful Python data analysis toolkit* – Release 1.4.4.," Available at: https://pandas.pydata.org/pandas-docs/version/1.4/pandas.pdf (Accessed: December 26, 2024).

Nasir Bin, F. and Li, J. (2024) "Understanding machine learning predictions of wastewater treatment plant sludge with explainable artificial intelligence," *Water Environment Research*, 96(10), e11136. Available at: https://doi.org/10.1002/wer.11136.

Nguyen, L.V. and Seidu, R. (2022) "Application of regression-based machine learning algorithms in sewer condition assessment for Ålesund City, Norway," *Water*, 14(24), 3993. Available at: https://doi.org/10.3390/w14243993.

Pedregosa, F. *et al.* (2011) "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, 12, pp. 2825–2830. Available at: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf (Accessed: December 26, 2024).

Python (no date) *Python 3.10 Documentation: The Python Language Reference*. Available at: https://docs.python.org/3.10/reference/index.html (Accessed: December 26, 2024).

Qambar, A.S. and Khalidy, M.M.A. (2022) "Prediction of municipal wastewater biochemical oxygen demand using machine learning techniques: A sustainable approach," *Process Safety and Environmental Protection*, 168, pp. 833–845. Available at: https://doi.org/https://doi.org/10.1016/j.psep.2022.10.033.

Rashidi-Khazaee, P., Rezvantalab, S. and Kheshti Monasebi, A. (2024) "Predicting wastewater treatment plant effluent quality using ensemble learning," *Green Technologies*, 02(01), pp. 35–43. Available at: https://gt.uut.ac.ir/article_210703.html (Accessed: December 26, 2024).

Rousis, N.I. *et al.* (2024) "Chapter 19 – Removal of emerging contaminants from wastewater by various treatment technologies in wastewater treatment plants," in M. Hadi Dehghani, R.R. Karri and I. Tyagi (eds.) *Sustainable remediation technologies for emerging pollutants in aqueous environment*. Amsterdam: Elsevier, pp. 389–409. Available at: https://doi.org/10.1016/B978-0-443-18618-9.00020-6.

Saleem, M.A., Harrou, F. and Sun, Y. (2024) "Explainable machine learning methods for predicting water treatment plant features

under varying weather conditions," *Results in Engineering*, 21, 101930. Available at: https://doi.org/https://doi.org/10.1016/j.rineng.2024.101930.

Santos, E., Carvalho, M. and Martins, S. (2023) "Sustainable water management: understanding the socioeconomic and cultural dimensions," *Sustainability*, 15(17), 13074. Available at: https://doi.org/10.3390/su151713074.

Shingare, S.P. *et al.* (2024) "Applicability of machine learning in waste water quality detection," in *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, Greater Noida, India, February 09–10, 2024. Piscataway Township, NJ: IEEE, pp. 525–530. Available at: https://doi.org/10.1109/IC2PCT60090.2024.10486814.

Tan, J.J., Arumugasamy, S.K. and Teo, F.Y. (2025) "Water quality index prediction using artificial neural network: A case study of Selangor River, Malaysia," *International Journal of Sustainable Agricultural Management and Informatics*, 11(1), pp. 48–71. Available at: https://doi.org/10.1504/IJSAMI.2025.143101.

Tyralis, H., Papacharalampous, G. and Langousis, A. (2019) "A brief review of random forests for water scientists and practitioners and their recent history in water resources," *Water*, 11(5), 910. Available at: https://doi.org/10.3390/w11050910.

Wang, Y., Cui, Z. and Ke, R. (2023) "Chapter 3 – Machine learning basics," in *Machine learning for transportation research and applications*. Amsterdam: Elsevier, pp. 25–40. Available at: https://doi.org/https://doi.org/10.1016/B978-0-32-396126-4.00008-4.

Yang, T. *et al.* (2024) "The LPST-Net: A new deep interval health monitoring and prediction framework for bearing-rotor systems under complex operating conditions," *Advanced Engineering Informatics*, 62, 102558. Available at: https://doi.org/10.1016/j.aei.2024.102558.

Yazdani, H., Doostizadeh, M. and Aminifar, F. (2023) "Forecast-aided power and flexibility trading of prosumers in peer to peer markets," *IET Renewable Power Generation*, 17(4), pp. 920–934. Available at: https://doi.org/10.1049/rpg2.12645.

Zhao, L. *et al.* (2020) "Application of artificial intelligence to wastewater treatment: A bibliometric analysis and systematic review of technology, economy, management, and wastewater reuse," *Process Safety and Environmental Protection*, 133, pp. 169–182. Available at: https://doi.org/https://doi.org/10.1016/j.psep.2019.11.014.