Discover Artificial Intelligence

RESEARCH Open Access

# Harnessing machine learning to mitigate water pollution in support of climate action

Bestami Özkaya[1], Faruk Dikmen[2], Ahmet Demir[2], Muhammad Owais Raza[3], Shtwai Alsubai[4], Onur Osman[5] and Jawad Rasheed[3,6,7,8*]

*Correspondence:
Jawad Rasheed
jawad.rasheed@izu.edu.tr
[1]Department of Civil Engineering, Istinye University, 34396 Istanbul, Turkey
[2]Department of Environmental Engineering, Yildiz Technical University, 34220 Istanbul, Turkey
[3]Department of Computer Engineering, Istanbul Sabahattin Zaim University, 34303 Istanbul, Turkey
[4]Department of Computer Science, College of Computer Engineering and Sciences in Al- Kharj, Prince Sattam Bin Abdulaziz University, P.O. Box 151, 11942 Al-Kharj, Saudi Arabia
[5]Department of Electrical and Electronics Engineering, Istanbul Topkapi University, Istanbul, Turkey
[6]Department of Software Engineering, Istanbul Nisantasi University, 34398 Istanbul, Turkey
[7]Research Institute, Istanbul Medipol University, 34810 Istanbul, Turkey
[8]Applied Science Research Center, Applied Science Private University, Amman, Jordan

**Abstract**

Wastewater treatment plants (WWTPs) are crucial in protecting public health and the environment by reducing pollutants before discharge into water bodies. This research presents a data-driven approach to enhance wastewater monitoring, ensuring compliance with environmental regulations by evaluating the predictive accuracy of several machine learning models in assessing effluent quality and categorizing effluent threats. In the first task, regression models such as Decision Tree, Random Forest, AdaBoost, and Support Vector Machine (SVM) were applied to predict Biochemical Oxygen Demand (BOD) and Chemical Oxygen Demand (COD), with Mean Absolute Error (MAE) and R-squared ($R^2$) used as evaluation metrics. In the second task, the same models were utilized to categorize effluent threat levels, and their performance was measured through accuracy, precision, recall, and F1-score. The results demonstrate that Gradient Boosting Regressor (GBR) and AdaBoost performed well in COD prediction, achieving the lowest MAE of 6.11 and the highest $R^2$ of 0.81. At the same time, Random Forest obtained the lowest MAE of 1.61 for BOD prediction. In the classification task, the Gradient Boosting Classifier (GBC) and AdaBoost achieved superior precision, recall, and F1 Scores, with all models attaining an overall accuracy of 97%. According to these results, machine learning methods, particularly GBC and AdaBoost, can significantly enhance prediction and classification accuracy for effluent quality, thereby improving WWTP management. This study contributes to climate resilience and sustainability by applying AI to minimize wastewater pollution, supporting SDG 6 (Clean Water and Sanitation), SDG 13 (Climate Action), SDG 14 (Life Below Water), and SDG 9 (Industry, Innovation, and Infrastructure).

**Keywords** Climate mitigation, Climate resilience, Climate-smart water management, Green infrastructure, Sustainable wastewater systems, Clean water and sanitation

## 1 Introduction

Treating wastewater to eliminate dangerous materials before it is discharged into water bodies, wastewater treatment plants (WWTPs) [1] play a critical role in protecting the environment and public health. These plants are built to reduce pollutants and maintain the water quality standards established by the authorities [2]. However, due to the

complexity of wastewater treatment processes, ongoing management and monitoring are necessary to ensure the plant operates effectively and within the required discharge limits. Predicting potential disturbances in discharge limits and which operational parameter variations can cause them is one of the main problems WWTPs face. Two of the most important markers of water quality among these parameters are Chemical Oxygen Demand (COD) and Biochemical Oxygen Demand (BOD) [3]. Wastewater treatment facilities must proactively anticipate and mitigate risks to discharge limits due to the potential repercussions of such disruptions. Plant operators can proactively avoid non-compliance by precisely predicting when COD and BOD values may surpass regulatory thresholds [4]. These actions may include modifying treatment procedures, optimizing resource utilization, or implementing early interventions. Therefore, improving the effectiveness of wastewater treatment operations and ensuring compliance with environmental safety standards requires developing sophisticated predictive models that can analyze historical data and forecast potential disruptions to discharge limits.

Ensuring the sustainable management of water resources is a critical challenge amid rapid urbanization, industrialization, and climate change. WWTPs are key infrastructures that mitigate water pollution, safeguard ecosystems, and contribute to public health. However, conventional monitoring and management techniques often fail to predict and prevent violations of effluent discharge limits, leading to environmental degradation. With the introduction of machine learning (ML), predictive modeling has entered a new era, offering sophisticated tools for understanding and mitigating risks in complex systems, such as environmental modeling. ML and DL models have demonstrated remarkable success in predicting environmental parameters, identifying pollution sources, and optimizing treatment processes under variable climatic conditions [5, 6]. By leveraging large datasets and learning from historical and Real-time information, ML DL algorithms enable adaptive decision-making. Furthermore, model's ability to model complex, nonlinear relationships makes them highly effective in assessing the impact of climate variability such as temperature, rainfall, and wind speed fluctuations.

Integrating these ML-driven methods into WWTP operations not only enhances predictive accuracy but also contributes to long-term sustainability by reducing energy consumption, optimizing chemical usage, and minimizing the carbon footprint of treatment processes. It is done by utilizing past data, such as laboratory readings, environmental conditions, and facility information, and training machine learning models on that data. Models can predict potential exceedances of discharge limits, enabling early intervention and improved resource management [7]. These capabilities enable plant operators to take early preventive actions, ensuring regulatory compliance and minimizing environmental impact.

Climate change and wastewater treatment are deeply interconnected, as wastewater is both a contributor to and a victim of climate impacts. Untreated or poorly managed wastewater releases significant quantities of greenhouse gases such as methane ($CH_4$) and nitrous oxide ($N_2O$), both of which have much higher global warming potentials than carbon dioxide ($CO_2$). Additionally, effluent discharge into water bodies accelerates eutrophication and degrades aquatic ecosystems, undermining climate resilience. Conversely, the operation of WWTPs is increasingly challenged by climate-induced variability in inflow volumes, temperature fluctuations, and extreme weather events. In this context, AI-driven wastewater optimization offers a powerful tool for both climate

mitigation and adaptation. By enabling real-time monitoring, early threat detection, and more efficient process control, machine learning models can reduce energy use, chemical consumption, and pollutant discharge, thereby lowering greenhouse gas emissions and operational risks. This proactive, data-informed management approach supports the development of climate-smart water infrastructure that is better equipped to withstand and respond to climate stressors.

This study aims to enhance the efficiency of WWTPs by developing ML models for predicting effluent quality and classifying threats. Beyond improving operational performance, this research directly supports global sustainability initiatives. By reducing pollutant discharge into water bodies, this study aligns with SDG 6 (Clean Water and Sanitation) by ensuring access to safe and clean water, as well as with SDG 13 (Climate Action). It also contributes to SDG 14 (Life Below Water) by preventing the contamination of aquatic ecosystems and SDG 9 (Industry, Innovation, and Infrastructure) by integrating advanced AI-driven solutions into industrial water management.

The novel aspect of this study is its dual methodology, which uses machine learning models to classify effluent threats and predict important effluent indicators (COD and BOD) while concurrently completing regression and classification of threat and no threat. This integrated framework offers a more thorough grasp of wastewater quality dynamics than previous research that concentrates on a single prediction type. The following are the key contributions of this study:

- Implemented various ML algorithms to predict COD, including Decision Tree, Random Forest, AdaBoost, Support Vector Machine (SVM) KNN, and Gradient Boosting.
- Compared 6 ML models for the prediction of BOD and COD prediction.
- Developed a classification model to categorize effluent threat levels as "no threat", "potential threat," and "threat" based on COD and BOD levels.
- Provided insights into class imbalance challenges and the importance of handling misclassifications.

The rest of the paper is organized as follows: Sect. 2 highlights prior studies published in reputable journals and conferences. Section 3 provides insight into exploited methodologies, Sect. 4 outlines the regression engines used, Sect. 5 details the experimental results, and Sect. 6 concludes the study.

## 2 Literature review

Pollution is one of the key global problems. One modern technique for addressing pollution is machine learning [8], which has broad applicability [9]. Various studies apply machine learning to the domain of global pollution; for instance, [10] analyzes global pollution trends, emphasizing the impacts of the electricity and industrial sectors. [10] used a feedforward neural network (FFNN) to predict contamination levels over the next 3 years, the model achieved MSE and RMSE of 79.6% and 89.2% respectively. [11] shows the effectiveness of hybrid ML DL models for air quality prediction. Using WAQI and SOGA datasets, a hybrid Polynomial Regression Fully Connected Neural Network (PR-FCNN) achieved the best accuracy for PM2.5. If we zoom in on the scope of pollution, water contamination is one of the biggest challenges and has become a severe global concern that stifles human development [12]. Recent studies have applied Machine

Learning (ML) to predict water quality, demonstrating high accuracy and adaptability. For instance, in [13] researchers used Haridwar's Water Quality Index (WQI) and developed models such as SVR, RFR, XGBR, and RNN, with the RNN achieving an $R^2$ of 0.96 and an RMSE of 0.75. WWTPs are key to combating water pollution, as they help treat wastewater so it can be returned to the environment safely. WWTP optimization and management have been the subject of much research due to their significance in preserving environmental safety and complying with regulatory requirements.

The application of predictive modeling to estimate probable disruptions in WWTP discharge limits has been investigated in several studies [7, 14]. Since COD and BOD levels are important markers of water quality, predicting them has received significant attention. Because of their ability to analyze vast amounts of data and identify patterns that may not be immediately obvious with conventional approaches, ML techniques have attracted significant attention [15]. One such study [16] demonstrated that ML can accurately and reliably predict COD levels from historical WWTP data. Researchers in [5] applied AI and ML models to predict final effluent BOD (F-BOD) in industrial wastewater facilities using 19 years of operational data. The Extra Trees model achieved the highest accuracy ($R^2 \approx 0.98$), while Neural Networks performed best for simulations. Their results imply that machine learning models can efficiently identify early indicators of operational inefficiencies, enabling prompt action to prevent discharge violations.

Accurately predicting COD and BOD levels remains difficult despite the encouraging results of ML-based techniques [17]. Predictive models ensure that WWTPs operate within reasonable bounds when controlling key parameters such as COD and BOD. Managing the inherent variability in wastewater composition and treatment plant operations is a major challenge [16]. Seasonal variations, operational changes, and fluctuating inflow characteristics are a few factors that can introduce noise into the data and hinder the generalization of predictive models [7, 18]. Researchers in [19] investigated the effects of data variability and quality on prediction model performance, proposing that feature engineering and data normalization could lessen these difficulties.

Work presented in [16] investigated some AI-based techniques, such as reinforcement learning, for optimizing energy use and reducing operating costs while abiding discharge regulations. They found that AI-based optimization methods could significantly improve WWTP efficiency, highlighting the potential to integrate these technologies into existing plant management systems.

Most earlier research focuses on either classification (classifying effluent quality) or regression (predicting COD or BOD), but rarely both. By completing both tasks simultaneously, predicting BOD and COD levels while categorizing effluent threats, this study fills that gap and provides a more comprehensive and useful framework for WWTP monitoring and management.

## 3 Methodology

In this study, we are developing machine learning models for predicting potential disruptions in the discharge limits of a wastewater treatment plant. The system has four main components: 1) Data Engine, 2) Preprocessing Dataset, 3) Regression Engine, and 4) Classification Engine. In the data engine, the raw dataset is labeled into a labeled dataset based on the discharge limit values. The regression engine predicts the discharge

limits for COD and BOD. The classification engine predicts three classes (threat, no threat, or any potential threat). Figure 1 shows the detailed methodology for the study.

### 3.1 Data engine

#### *3.1.1 Dataset*

The dataset contains environmental conditions, facility data, laboratory data, and critical facilities. The data is collected daily between 01.01.2022 to 08.12.2024, thus contains 1075 Rows. It includes 65 Features and two target variables: COD and BOD. The features are categorized into four major groups relevant to wastewater treatment analysis: Important Facility Parts (e.g., diffusers and aeration tanks), Laboratory Data (e.g., organic and chemical oxygen indicators), Facility Information (e.g., energy consumption, chemical usage, sludge management, and flow rates), and Environmental Conditions (e.g., temperature, humidity, and weather factors influencing treatment performance).

#### *3.1.2 Data annotation*

To annotate the dataset into three categories —Threat (2), Potential Threat (1), and No Threat (0) —predefined thresholds for COD and BOD are used. The thresholds for COD and BOD are 125 mg/l and 25 mg/l, respectively. Labeling is done by comparing effluent values to regulatory thresholds and accounting for potential threats by setting a buffer zone at 90% of the threshold. This annotation enables preventive action before

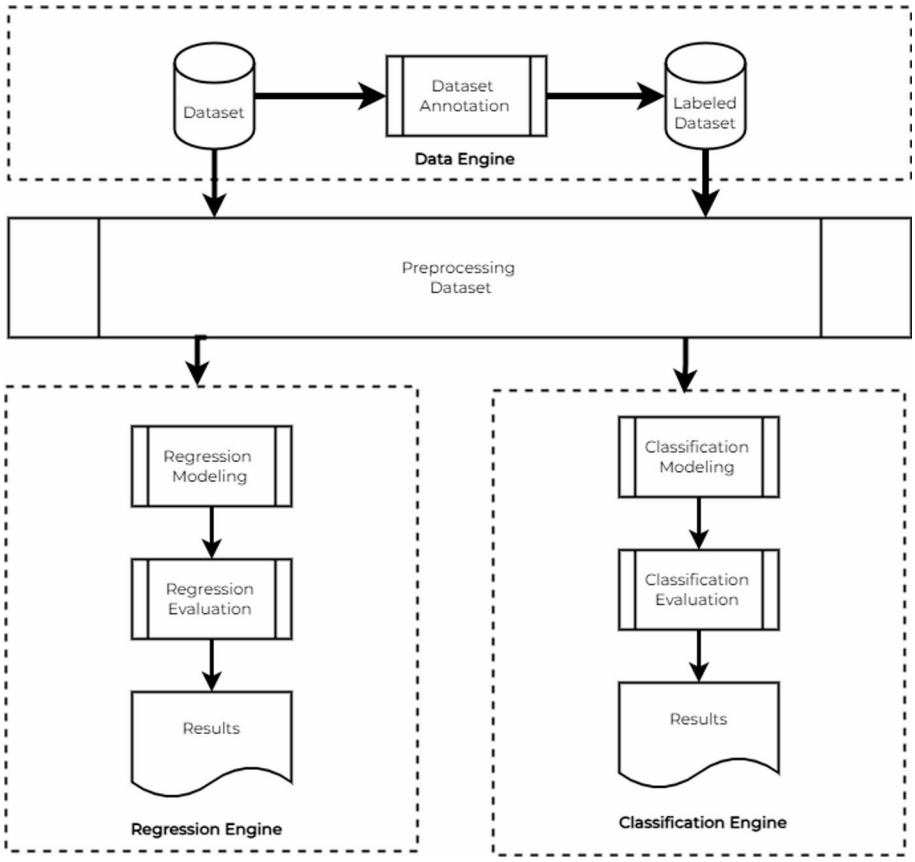


**Fig. 1** A schematic representation of the machine learning-based wastewater treatment plant management approach. It includes four main components: data engine, preprocessing, regression engine, and classification engine

regulatory violations occur by providing early warnings when effluent levels approach critical values.

#### 3.1.3 Labeled dataset

This dataset is the same as the one discussed in the previous sections. It has only one additional column with the label that shows whether there is a threat, potential threat, or no threat. This dataset will be useful in the classification engine.

### 3.2 Preprocessing

To ensure the dataset's reliability and suitability for model training, a systematic preprocessing pipeline was implemented. The process involved Handling Missing Values (HMV), Normalizing Data (ND), Removing Outliers (RO), and splitting the dataset into training and validation subsets in an 80–20 ratio. Missing values in numerical columns were imputed with the mean of each feature, preserving the overall data distribution. The Interquartile Range (IQR) method was applied to detect and remove extreme outliers, ensuring that anomalous readings did not distort model performance. After outlier removal, all numerical features were normalized to a standard scale (typically 0–1) to improve model convergence and prevent bias toward higher-magnitude variables. This comprehensive preprocessing ensured that the dataset remained balanced, consistent, and well-prepared for robust predictive modeling. It is represented by (1), where the 1st and 3rd quartiles are denoted by $Q_1$ and $Q_3$, respectively. Values that fall outside of this range are eliminated as outliers.

$$IQR = Q_3 - Q_1 \tag{1}$$

The dataset is finally split into 80% for training and 20% for testing to ensure successful model training and evaluation. A dataset with size N has $N \times 0.8$ training samples and $N \times 0.2$ testing samples. These preprocessing steps improve predictive performance and generalizability.

## 4 Regression engine

This section discusses the regression mechanism employed in this study. It has two parts: preprocessing, modeling, and evaluation. Let's discuss each in detail.

### 4.1 Regression modeling

Once the data is preprocessed, the next step is applying machine learning modeling. 80% of the training data extracted in the last step is used for regression modeling in this step. The selected algorithms are decision tree, random forest regression, AdaBoost regression, and support vector machine. All of these algorithms are discussed as follows.

#### 4.1.1 Decision tree regression

Using a tree-like structure, the Decision Tree Regressor is a non-linear model that predicts outcomes based on input features. To create branches that ultimately lead to predicted continuous outcomes at the leaves, it first divides the data into subsets by choosing the feature that best separates the data at each node. This model helps to understand how features contribute to predictions and is easy to interpret because it visualizes the decision-making process. It can handle linear and non-linear relationships

but is prone to overfitting, particularly when working with deep trees or complicated datasets. Pruning methods and hyperparameter adjustments, such as setting a minimum number of samples per leaf or a maximum depth, can reduce overfitting and improve generalization. Decision tree regressions are frequently employed when interpretability is crucial and when there are non-linear relationships between features and results [20]. In the context of wastewater prediction, Decision Trees help identify key operational variables that strongly influence COD and BOD variations; Decision Tree Regression provides clear interpretability for plant operators.

#### *4.1.2 Random forest regression*

An ensemble learning method called Random Forest builds multiple decision trees and aggregates their predictions to improve prediction accuracy. Compared to linear regression, it manages nonlinear relationships and feature interactions better. It is immune to overfitting and data noise because it combines predictions from several trees. Because Random Forest can handle complex dependencies and estimate feature importance without extensive preprocessing, it is especially helpful for high-dimensional data and missing values [21]. Random Forest improves Decision Trees by combining multiple decision paths, reducing overfitting, and effectively handling noisy COD and BOD variations.

#### *4.1.3 Adaboost regression*

An adaptive boosting technique (AdaBoost) improves predictive performance by gradually training weak learners (typically shallow decision trees) and assigning misclassified instances higher weights. This iterative correction process allows the model to concentrate on hard-to-predict patterns in effluent quality data. AdaBoost works well with moderately complex wastewater treatment systems. Still, it is sensitive to outliers and noisy data, which can make predictions unstable if not adjusted [21]. AdaBoost performs well when effluent behavior changes dynamically, as it adaptively focuses on difficult to predict samples crucial for early identification of potential discharge violations.

#### *4.1.4 Support vector regression (SVR)*

Support Vector Machines (SVMs) are the source of SVR, a regression method for continuous-valued prediction. SVR operates by locating the hyperplane in a high-dimensional space that best fits the data while keeping an acceptable error margin (epsilon). It seeks to ensure that the model generalizes well to unseen data by minimizing model complexity while allowing slight prediction errors. SVR requires careful parameter selection, including the kernel, epsilon, and penalty parameter (C), and can be computationally costly, especially with large datasets. Nevertheless, it can produce reliable predictions [20]. SVR is particularly suited for distinguishing between borderline effluent quality classes, where minor fluctuations in parameters can shift compliance outcomes. SVR is particularly suited for distinguishing between borderline effluent quantity, where minor fluctuations in parameters can shift the BOD and COD values.

#### *4.1.5 Gradient boosting regressor (GBR)*

The Gradient Boosting Regressor builds a series of weak learners iteratively to minimize prediction errors, thereby capturing complex, nonlinear dependencies between

treatment plant variables and effluent indicators such as COD and BOD. Its ability to model subtle interactions among features such as pH, inflow rate, and suspended solids makes it particularly effective for accurate effluent-quality forecasts under dynamic operating conditions.

#### *4.1.6 K-nearest neighbors (KNN) regressor*

The KNN Regressor estimates COD and BOD by comparing current operational states to historical observations with similar process conditions. Its instance-based learning approach enables it to adapt quickly to seasonal fluctuations and site-specific treatment behaviors without extensive model retraining.

### 4.2 Regression evaluation

To evaluate the trained machine learning models, 20% of the dataset is used. The evaluation metrics used in this study are Mean Absolute Error (MAE) and Coefficient of Determination ($R^2$). The following are the details of each metric.

#### *4.2.1 Mean absolute average (MAE)*

It shows the average absolute difference between real and expected values. It helps understand the average deviation between predictions and observed values, as it provides an easy-to-understand interpretation of prediction errors in a unit similar to the target. A lower MAE shows better model performance. MAE is mathematically demonstrated in (2), where $y_i$ is the actual value, n is the total number of observations, and $\hat{y}_i$ is the predicted value [5]. MAE provides a direct measure of average prediction deviation for COD and BOD, reflecting how much predicted values differ from laboratory results in real-world WWTP scenarios.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

#### *4.2.2 Co-efficient determination ($R^2$)*

In the R2 measure, perfect predictions are indicated by values near 1. In contrast, values close to 0 indicate that the model does not explain a sizable portion of the variance in the predicting variable [5]. The R2 is displayed in (3), where the actual value is $y_i$, the predicted value is $\hat{y}_i$, and the mean of the actual values is $\overline{y}$. R2 quantifies how effectively the model captures variation in COD/BOD due to process or seasonal changes, reflecting the reliability of predictions for operational adjustments.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{3}$$

### 4.3 Classification engine modeling

Once we have preprocessed the labeled dataset, the next step is to train classification models on 80% of the training data. This study uses four classification algorithms: SVMs, random forests, AdaBoost, and decision trees. Each of these models is discussed in detail below.

#### 4.3.1 Decision tree

A supervised learning algorithm called Decision Tree Classification uses a tree-like structure to classify data. It chooses the feature that best separates the data at each node by splitting the data based on feature values. The leaves stand in for the anticipated class labels, while each branch represents a decision rule. Recursively dividing the dataset to lower impurity —often using metrics like Gini Impurity or Information Gain (entropy) —builds the tree. Decision trees are easy to understand because their structure makes it clear how predictions are made. They can, however, overfit, particularly when dealing with noisy data or deep trees. Overfitting can be avoided by using pruning strategies, establishing a maximum depth, and requiring a minimum number of samples per leaf. Despite their simplicity, decision trees work well for issues where interpretability is crucial and can handle both numerical and categorical data [20]. In the context of wastewater prediction, the Decision Trees classifier helps identify key features that strongly influence threat classification.

#### 4.3.2 Random forest

An ensemble learning technique called Random Forest builds several decision trees and aggregates their predictions to increase precision and decrease overfitting. It works especially well when dealing with non-linearity in data. The algorithm uses the Bootstrap Aggregation (Bagging) principle, which randomly samples subsets of the dataset with replacement to train multiple trees. Large datasets are easily handled by Random Forest, which also offers feature importance scores to aid in feature selection. However, training can be slow, particularly when working with many trees or features, and is computationally costly [22]. Random Forest classifiers improve Decision Trees by combining multiple decision paths, reducing overfitting, and effectively handling noisy COD and BOD variations, and providing more robust classification of threats.

#### 4.3.3 Adaboost

An ensemble of weak learners, typically simple decision trees known as stumps, is constructed using the AdaBoost boosting technique, thereby enhancing their performance by focusing on instances incorrectly classified. In contrast to bagging techniques like Random Forests, AdaBoost assigns misclassified samples greater weights in each iteration, ensuring that weaker learners focus on these challenging cases. This process is iteratively carried out, with each new student fixing the errors of their predecessor. Despite this, AdaBoost is still a strong classification algorithm, particularly in fields where improving performance for weak learners can have a substantial impact [21]. AdaBoost classifier performs well when effluent behavior changes dynamically, as it adaptively focuses on difficult-to-predict samples for threat classification.

#### 4.3.4 Support vector classifier

Support Vector Classifier (SVC) is a supervised learning algorithm that categorizes data into different classes. It belongs to the Support Vector Machine (SVM) family. SVC is susceptible to parameter tuning, which affects the model's performance and generalization. This is especially true for the kernel function and the penalty parameter (C). SVC can be computationally intensive, even with its power, particularly when dealing with large datasets [20]. It is computationally intensive, though, and to prevent overfitting,

hyperparameters (such as the learning rate and the number of trees) must be carefully adjusted. Because it is sequential, the training process is slower than that of Random Forest [21]. SVC is particularly suited for distinguishing between borderline effluent threat classes, where minor fluctuations in parameters can shift compliance outcomes, i.e., threat and no threat.

#### *4.3.5 Gradient boosting classifier*

For effluent threat classification, the Gradient Boosting Classifier aggregates multiple shallow models to enhance predictive power and resilience against noise in wastewater data. This enables the system to accurately distinguish between threat and non-threat discharge levels, supporting proactive compliance management across variable climatic and industrial conditions.

#### *4.3.6 K-nearest neighbors (KNN) classifier*

In effluent threat classification, the KNN Classifier relies on the similarity between new and past data points to categorize effluent quality. This simplicity makes it highly interpretable for plant operators, providing intuitive insights into how current readings align with known threat or non-threat patterns in the plant's operation.

### 4.4 Classification evaluation

The classification models that were trained in the previous section are evaluated in this section using 20% of the dataset. The study employed Accuracy, precision, recall, and F1-Score. Let's go over each metric in more detail.

#### *4.4.1 Accuracy*

Although accuracy offers a broad indicator of correctness, it can be misleading when datasets are unbalanced. For example, a model that predicts all samples as "normal" would still achieve 90% accuracy even though it is ineffective at detecting actual violations if the dataset contains 90% normal effluent samples and only 10% violations. Consequently, other metrics ought to be used in addition to accuracy. Accuracy is calculated as given in (4). For classification, accuracy indicates how well models identify threat versus non-threat effluent measure.

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{4}$$

#### *4.4.2 Precision*

Positive Predictive Value (PPV), also known as precision, measures how reliable positive predictions are. Precision is computed by (5). It is measured by the proportion of predicted violations that were actually violations. Precision ensures that flagged "Threat" samples are genuinely problematic, minimizing false alarms that could waste resources.

$$Precision = \frac{T_p}{T_p + F_p} \tag{5}$$

#### 4.4.3 Recall

The model's recall, also known as sensitivity or true positive rate (TPR), gauges its ability to identify actual infractions. "Out of all actual violations, how many were correctly predicted?" is the question it answers. The calculation is done by (6). High recall reflects the model's ability to detect most actual threat, a vital factor for preventive environmental management.

$$Recall = \frac{T_p}{T_p + F_n} \tag{6}$$

#### 4.4.4 F1-Score

The F1 Score offers a balanced metric when precision and recall are equally significant. It takes the harmonic mean of the two. This is especially helpful when dealing with unbalanced datasets, where depending only on accuracy can be deceptive. The F1 Score is calculated by (7). F1 balances precision and recall, making it especially meaningful for wastewater monitoring, where missing true threats is as costly as overestimating them.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

## 5 Experimental results

This section shows the empirical results from multiple machine learning models. Effluent COD and BOD prediction models were tested in the first task, which focused on effluent prediction. For this task, models such as SVMs, Random Forest, AdaBoost, and Decision Tree were tested. $R^2$ and MAE were the evaluation metrics employed for this task.

Table 1 shows that the GBR ($R^2 = 0.81$, MAE = 6.11) and AdaBoost ($R^2 = 0.80$, MAE = 7.84) both performed best for Effluent COD, successfully capturing about 80% of the variance in the target variable. While SVR struggled greatly, producing the lowest $R^2$ score (0.20) and the highest MAE (11.73), suggesting poor predictive power, the Decision Tree Regressor performed moderately well ($R^2 = 0.59$, MAE = 9.05).

It is evident from Table 2 that GBR ($R^2 = 0.74$, MAE = 1.64 and) AdaBoost ($R^2 = 0.74$, MAE = 1.73) performed better than the other models and produced the most accurate predictions for Effluent BOD. Compared to ensemble models, the Decision Tree Regressor performed poorly ($R^2 = 0.26$, MAE = 2.49), while SVR performed marginally better ($R^2 = 0.49$, MAE = 1.94). These results imply that the best techniques for predicting COD and BOD in wastewater are ensemble methods, especially GBR and AdaBoost.

**Table 1** Performance evaluation of different regression models in predicting effluent COD

| Model | MAE | $R^2$ |
|---|---|---|
| Decision tree regressor | 9.05 | 0.59 |
| Random forest | 6.22 | 0.79 |
| AdaBoost | 7.84 | 0.80 |
| SVR | 11.73 | 0.20 |
| KNN | 10.06 | 0.61 |
| GBR | 6.11 | 0.81 |

It presents the mean absolute error (MAE) and $R^2$ scores for decision tree, random forest, adaboost, and support vector regression (SVR) KNN and GBR,. GB and AdaBoost exhibited superior performance in COD prediction

**Table 2** Performance evaluation of different regression models in predicting effluent BOD

| Model | MAE | $R^2$ |
|---|---|---|
| Decision tree regressor | 2.49 | 0.26 |
| Random forest | 1.61 | 0.73 |
| AdaBoost | 1.73 | 0.74 |
| SVR | 1.94 | 0.49 |
| KNN | 1.97 | 0.62 |
| GBR | 1.64 | 0.74 |

It presents the mean absolute error (MAE) and $R^2$ scores for decision tree, random forest, adaboost, and support vector regression (SVR), KNN and GBR. AdaBoost and GBR exhibited superior performance in BOD prediction

**Table 3** Comparison of classification models (Decision tree, Support vector machine, Random forest, and AdaBoost) in predicting effluent threat levels

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|---|
| Decision tree | 96 | 96 | 96 | 96 |
| Random forest | 97 | 94 | 97 | 96 |
| AdaBoost | 97 | 97 | 97 | 97 |
| Support vector machine | 97 | 95 | 97 | 96 |
| KNN | 97 | 95 | 97 | 95 |
| GBC | 97 | 97 | 97 | 97 |

It presents accuracy, precision, recall, and F1 scores, with AdaBoost achieving the best overall performance

Table 3 presents the classification metrics for all the evaluated models. Overall, all models demonstrated strong performance, achieving high accuracy rates. Decision Tree, Support Vector Machine, Random Forest, and AdaBoost each attained an accuracy of 96% or higher. GBC and AdaBoost stood out with consistent scores of 97% in precision, recall, and F1-score, indicating balanced predictive power. The Support Vector Machine and Random Forest models also performed well, with F1 scores of 96%, despite a slight drop in Precision for Random Forest (94%). These results indicate that all models are well-suited for the classification task, with GBC and AdaBoost displaying the most stable performance across all metrics.

The confusion matrices in Fig. 2 depict the classification performance of various models and highlight misclassification regions in three classes.

The Decision Tree model misclassifies two as Class 1 and two as Class 2, but it performs well on Class 0, correctly classifying 201 instances. Nonetheless, Class 1 and Class 2 have greater misclassification rates, with numerous instances being placed in the wrong classes. The SVM model incorrectly classifies four Class 1 instances as Class 0 and one as Class 2, demonstrating its difficulty with minority classes. It correctly classifies all 205 instances of Class 0. Class 2, on the other hand, incorrectly classifies one instance as Class 0 despite having four correct classifications. All 205 instances were correctly classified by Random Forest, demonstrating high accuracy for Class 0. Nevertheless, Class 1 continues to present issues because four cases are incorrectly classified as Class 0 and one as Class 2. Class 2 performs well overall, with three Class 0 classifications accurate and two incorrect.

The Precision-Recall (PR) Curve in Fig. 3 shows the trade-off between precision and recall for six models: AdaBoost, Random Forest, SVM, Decision Tree, KNN, and GBC.

SVM (orange) performs well because it continuously maintains high precision across a range of recall values, whereas Random Forest (green) is stable, particularly at higher recall levels. As recall rises, precision declines rapidly, and the Decision Tree (blue)

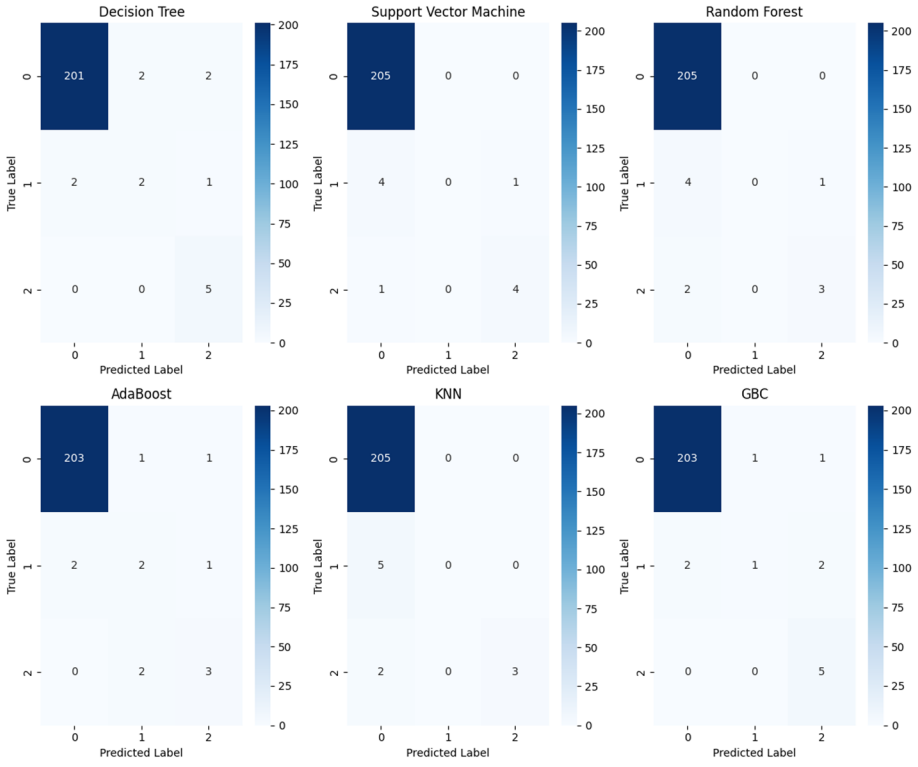


**Fig. 2** Confusion matrices illustrating the classification performance of decision tree, support vector machine, random forest, adaboost, knn, and gbc models to highlight misclassification patterns across different threat categories

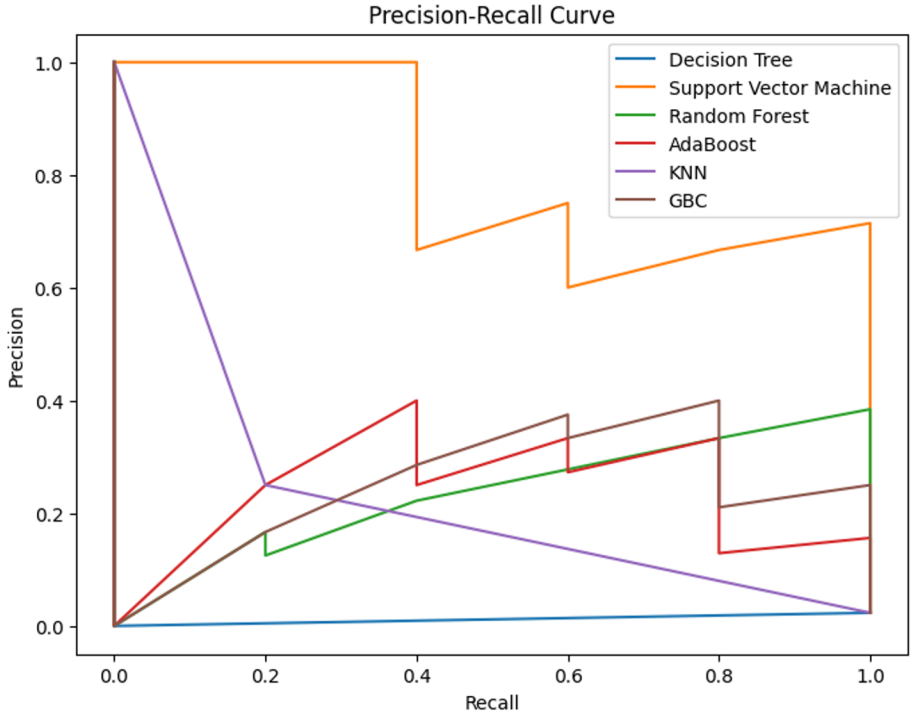


**Fig. 3** A precision-recall (PR) curve comparing the trade-off between precision and recall for KNN, GBC, AdaBoost, Random forest, Support vector machine, and Decision tree classifiers. SVM maintains high precision at different recall values, while the decision tree shows the least stable performance

performs least steadily, while AdaBoost (red) shows varying precision. Because of misclassifications, precision tends to decline as recall increases, although the sharp vertical jumps at low recall indicate that models predict a small number of positives with high precision. SVM performs better overall than the other models, with Decision Tree being the least dependable.

The results of this study highlight the potential of ML-based predictive models to transform wastewater management by improving efficiency and sustainability. The high accuracy achieved by AdaBoost and Random Forest models demonstrates the feasibility of data-driven effluent monitoring. More importantly, applying these models can significantly reduce WWTPs' environmental footprint by enabling proactive measures to prevent effluent violations.

A key implication of this research is its contribution to sustainability through improved wastewater quality monitoring. By accurately predicting COD and BOD levels, WWTPs optimize treatment processes, reducing chemical usage, lowering energy consumption, and minimizing environmental contamination. These advancements support SDG 6 by enhancing water safety and SDG 14 by mitigating the impact of untreated effluents on aquatic life. Furthermore, implementing AI-driven wastewater treatment aligns with SDG 9 by fostering innovation in industrial water management. The insights from this research can be leveraged to develop automated, AI-powered monitoring systems, thereby enabling more sustainable water resource management globally.

## 6 Discussion

This study demonstrates how machine learning (ML) can improve operations at wastewater treatment plants (WWTPs). The models that performed best in both regression and classification tasks were Gradient Boosting and AdaBoost. Gradient Boosting showed strong predictive accuracy for COD ($R^2 = 0.81$, MAE = 6.11) and BOD ($R^2 = 0.74$, MAE = 1.64), while AdaBoost achieved the best classification metrics, with 97% precision, recall, and F1-score. Because ensemble models are robust and can capture complex, non-linear relationships, they performed better than simpler models like KNN and Decision Tree. These results show how ML can support real-time decision-making, streamline treatment procedures, and ensure regulatory compliance, all of which advance SDGs 6 (clean water) and 9 (innovation). The dual-task approach offers a useful, data-driven framework for more intelligent and sustainable wastewater management.

## 7 Conclusion

The study's empirical findings demonstrate the potential of machine learning models for wastewater treatment plant management, particularly for identifying potential effluent threats and forecasting effluent quality. For the regression task, Random Forest performed well in BOD prediction, while GBC, AdaBoost, and Random Forest produced the best results for COD prediction. The models demonstrated remarkable performance in classification tasks, with all models attaining 96% accuracy: GBC, KNN, AdaBoost, and Decision Tree, and Random Forest exhibiting superior precision, recall, and F1-Score. Random Forest's bias towards the majority class was evident in its limited ability to classify the minority class, despite strong overall performance. GBC and AdaBoost's superior performance, especially when handling class imbalance, was further supported by the confusion matrix and Precision-Recall curves. The findings emphasize

that integrating AI into wastewater treatment enhances efficiency and promotes global sustainability efforts. To increase model robustness, future research will concentrate on broadening the dataset across various WWTPs and conditions. Sophisticated deep learning models, such as Transformers or LSTMs, can be investigated to capture temporal patterns. To further improve model interpretability and facilitate intelligent, sustainable wastewater management, explainable AI (XAI) and optimization frameworks will be incorporated.

**Author contributions**

Conceptualization: Bestami Özkaya, Faruk Dikmen, and Ahmet Demir; methodology: Faruk Dikmen, Muhammad Owais Raza, Shtwai Alsubai, Onur Osman, and Jawad Rasheed; software: Faruk Dikmen, Muhammad Owais Raza, and Jawad Rasheed; validation: Bestami Özkaya, Faruk Dikmen, Ahmet Demir, Muhammad Owais Raza and Jawad Rasheed; formal analysis: Faruk Dikmen, Muhammad Owais Raza, and Jawad Rasheed; investigation: Bestami Özkaya, Faruk Dikmen, Ahmet Demir, and Muhammad Owais Raza; resources: Faruk Dikmen, Muhammad Owais Raza, Shtwai Alsubai, and Onur Osman; data curation: Faruk Dikmen; writing-original draft: Bestami Özkaya, Ahmet Demir, Muhammad Owais Raza, Shtwai Alsubai, Onur Osman, and Jawad Rasheed; writing-review and editing: Faruk Dikmen, Muhammad Owais Raza, and Jawad Rasheed; visualization: Faruk Dikmen, Muhammad Owais Raza, Shtwai Alsubai, and Onur Osman; supervision: Bestami Özkaya, and Jawad Rasheed. All authors have read and agreed to the published version of the manuscript.

**Funding**

The authors declare that this research received no external funding.

**Data availability**

The data supporting the findings of this study are available from the Istanbul Water and Wastewater Administration; however, restrictions apply to the availability of these data, which were used under license for the current study and so are not publicly available. Data are, however, available from the author (Faruk Dikmen, email: faruk.dikmen@std.yildiz.edu.tr) upon reasonable request and with permission of the Istanbul Water and Wastewater Administration.

## Declarations

**Ethics approval and consent to participate**

Not applicable

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare no competing interests.

Received: 23 July 2025 / Accepted: 27 November 2025

Published online: 02 December 2025

**References**

1. Skoczko I, Struk-Soko lowska J, Ofman P. Seasonal changes in nitrogen, phosphorus, BOD and COD removal in bystre wastewater treatment plant. J Ecol Eng. 2017;18(4):185–91.
2. Oliveira SC, Von Sperling M. Reliability analysis of wastewater treatment plants. Water Res. 2008;42(4–5):1182–94.
3. Prambudy H, Supriyatin T, Setiawan F. The testing of chemical oxygen demand (cod) and biological oxygen demand (bod) of river water in cipager cirebon. J Phys: Conf Ser. 2019;1360:012010
4. Çağatay Çetinkaya M, Üstün GE. Monitoring and evaluation of the efficiency of a mixed textile-domestic wastewater treatment plant for 3 years. Environ Monit Assess. 2022;194(6):430.
5. Dawar I, Singal M, Singh V, Lamba S, Jain S. Predicting air quality index using machine learning: a case study of the Himalayan city of Dehradun. Nat Hazards. 2025;121(5):5821–47.
6. Jing Z, Zhang Yi, Liu X, Li Q, Hao Y, Li Y, et al. Identifying human activities causing water pollution based on microbial community sequencing and source classifier machine learning. Environ Int. 2025;195:109240.
7. Inbar O, Shahar M, Avisar D. Predictive modeling of bod throughout wastewater treatment: a generalizable machine learning approach for improved effluent quality. Environ Sci Water Res Technol. 2024;10(10):2577–88.
8. Rautela KS, Goyal MK. Transforming air pollution management in India with AI and machine learning technologies. Sci Rep. 2024;14(1):20412.
9. Inam SA, Hashim H, Awan AM, Rajput H, Umer S. A novel approach toward windspeed forecasting using an advanced deep learning framework with explainable AI. VFAST Trans Softw Eng. 2025;13(3):198–210.
10. Inam, Azeem S, Zaidi SMH, Khan AA, Ullah S. A neural network approach to carbon emission prediction in industrial and power sectors. Discov Appl Sci. 2025;7(6):640.
11. Inam SA, Khan AA, Mazhar T, Ahmed N, Shahzad T, Khan MA, et al. PR-FCNN: a data-driven hybrid approach for predicting PM2. 5 concentration. Discov Artif Intel. 2024;4(1):75.
12. Chen L, Wang J, Zhu M, He R, Mu H, Ren H, et al. Quality evaluation parameter and classification model for effluents of wastewater treatment plant based on machine learning. Water Res. 2025;1(268):122696.
13. Lamba S, Dawar I, Singal M, Singh J. Predicting water quality index using machine learning techniques: a case study of river Ganga in Haridwar, India. Earth Sci Inform. 2025;18(2):1–20.

14. El-Rawy M, Abd-Ellah MK, Fathi H, Ahmed AKA. Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques. J Water Process Eng. 2021;44:102380.
15. Shaikha, S.S., Shahapurkara, R. Predicting cod and bod parameters of greywater using multivariate linear regression. This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4(10.3233) 2021
16. Gholizadeh M, Saeedi R, Bagheri A, Paeezi M. Machine learning-based prediction of effluent total suspended solids in a wastewater treatment plant using different feature selection approaches: a comparative study. Environ Res. 2024;246:118146.
17. Jadhav AR, Pathak PD, Raut RY. Water and wastewater quality prediction: current trends and challenges in the implementation of artificial neural network. Environ Monit Assess. 2023;195(2):321.
18. Qambar AS, Al Khalidy MMM. Development of local and global wastewater biochemical oxygen demand real-time prediction models using supervised machine learning algorithms. Eng Appl Artif Intell. 2023;118:105709.
19. Mahanna H, El-Rashidy N, Kaloop MR, El-Sapakh S, Alluqmani A, Hassan R. Prediction of wastewater treatment plant performance through machine learning techniques. Desalin Water Treat. 2024;319:100524.
20. Somvanshi, M., Chavan, P., Tambade, S., Shinde, S.: A review of machine learning techniques using decision tree and support vector machine. In: 2016 International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 1–7 (2016). IEEE
21. Su B, Zhang W, Li R, Bai Y, Chang J. En-wbf: a novel ensemble learning approach to wastewater quality prediction based on weighted boostforest. Water. 2024;16(8):1090.
22. Kheimi M, Almadani MA, Zounemat-Kermani M. Simulating wastewater treatment plants for heavy metals using machine learning models. Arab J Geosci. 2022;15(17):1458.

## Publisher's note