Predicting effluent BOD_5 in wastewater treatment using machine learning

NASSIMA LEKOUAGHET, LAKHDAR BOURAGBI, BRAHIM MAHFOUD

Keywords: Biochemical Oxygen Demand (BOD_5) — wastewater treatment — machine learning — soft sensor — XGBoost — BOD_5 prediction — process optimization

ABSTRACT

Accurate and quick prediction of effluent Biochemical Oxygen Demand (BOD_5) is important for maintaining regulatory standards and improving the operation of wastewater treatment (WWT) plants. Traditional laboratory tests for BOD_5 take several days, causing delays in assessing water quality and adjusting plant performance. To overcome this issue, this study develops a machine learning model to estimate effluent BOD_5 using easily available plant data. A dataset from a large-scale wastewater treatment plant was used, comprising 12 variables from both influent and effluent data, such as pH, Chemical Oxygen Demand (COD), conductivity, Total Suspended Solids (TSS), BOD_5, and temperature, with effluent BOD_5 as the target variable. The data were cleaned and processed using the Interquartile Range (IQR) method to remove outliers, and new features were created to better reflect process behavior. Among several tested models, XGBoost showed the best performance, achieving an R^2 of 0.9695 and a Mean Squared Error (MSE) of 0.05, based on a time-based train/test split (80 % / 20 %). The study demonstrates that machine learning can serve as a powerful predictive tool for BOD_5 estimation, significantly reducing the five-day analytical lag of laboratory testing and supporting faster, data-driven decision-making in WWT management.

INTRODUCTION

The proper management and treatment of municipal and industrial wastewater are essential to preserve water resources and protect human and ecosystem health. Among the various parameters used to evaluate effluent quality, the Five-Day Biochemical Oxygen Demand (BOD_5) is one of the most widely adopted indicators. It reflects the amount of oxygen required by microorganisms to biologically decompose organic pollutants in water over a five-day period [1]. Because BOD_5 provides a direct measure of biodegradable organic load, it serves as a critical benchmark for assessing treatment efficiency and environmental impact. Consequently, regulatory authorities across the globe have established strict discharge limits to maintain acceptable BOD_5 levels in treated effluents [2].

Despite its importance, the conventional BOD_5 test has significant limitations. The standard method requires a five-day incubation period, delaying feedback on treatment performance and limiting the ability to make timely operational decisions [3]. Moreover, factors such as sample toxicity, nitrification effects, variability in microbial seeding, and preservation challenges can compromise measurement accuracy and reproducibility [4–7]. These drawbacks highlight the need for rapid, reliable, and continuous BOD_5 estimation methods.

Machine learning (ML) and soft-sensor modeling have emerged as effective tools for rapid BOD_5 estimation, enabling predictions from routinely measured operational parameters such as COD, pH, conductivity, temperature, and TSS [8, 9]. Unlike empirical or deterministic models [7, 8], which often fail under variable conditions due to the complex biochemical and hydrodynamic processes in full-scale wastewater treatment plants (WWTPs) [10, 11], ML approaches can capture nonlinear interactions and system dynamics, supporting proactive process control, early anomaly detection, and more sustainable plant operation.

Recent studies show that modern machine-learning methods can achieve very strong predictive performance for BOD_5 and other water-quality parameters. For example, Dantas et al. [12] demonstrated that artificial neural networks (ANN), which are computational models inspired by the way biological neurons process information, perform well when forecasting BOD_5 and COD in full-scale treatment plants. Zhang et al. [13] used a Random Forest model, an ensemble of many decision trees that collectively improve prediction accuracy, to estimate BOD_5 in river water, achieving an R^2 of approximately 0.89. Jafar et al. (2022) [14] employed deep neural networks (DNN), a more complex form of neural-network architecture capable of capturing highly non-linear relationships and reported R^2 values between 0.89 and 0.93. Other studies have shown comparable results: XGBoost (short for Extreme Gradient Boosting) models, which use gradient-boosted decision trees, have reached R^2 values around 0.92 in ecological applications [15], while RVFL (Random Vector Functional Link) networks have achieved $R^2 \approx 0.924$ in WWTP settings [16]. In contrast, simpler models often provide lower accuracy. For instance, Ali et al. [17] reported an R^2 of about 0.88 using CatBoost (Categorical Boosting), which efficiently handles categorical and numerical data with minimal preprocessing. It should be noted that linear regression models are not robust under varying wastewater treatment conditions, as they often fail to capture nonlinear interactions and complex system dynamics [18].

Accurate and timely estimation of the biochemical oxygen demand (BOD_5) remains a persistent challenge in WWT, as traditional laboratory analyses require a five-day incubation period. This inherent delay limits the speed of process evaluation and prevents proactive operational control. To address this challenge, this study proposes a predictive framework leveraging the XGBoost algorithm, selected for its superior ability to handle high-dimensional data and non-linear relationships. By integrating a robust data-driven methodology with domain-informed feature engineering, the model seeks to accurately capture the complex dynamics of the treatment process under realistic operating conditions. The primary objective of this work is to provide a reliable alternative for rapidly assessing wastewater quality, thereby facilitating faster decision-making and enhancing wastewater management response. The following sections detail the methodology and model development process used to achieve these objectives.

METHODOLOGY AND MODEL DEVELOPMENT

Presentation site and data collection, area characteristics

The dataset used in this study was collected from the municipal wastewater treatment plant (WWTP) of Skikda Province, located in north-eastern Algeria (36.89°N, 7.01°E). The facility operates a medium-load activated sludge process designed to handle variable organic loads from domestic and municipal sources. The plant has a design capacity of 46,000 m^3/day and serves approximately 279,700 population equivalents.

The treatment process includes preliminary screening, grit removal, primary sedimentation, biological oxidation in aeration tanks, secondary clarification, and final disinfection (see *Fig. 2*). The system is designed for an average influent BOD_5 load of about 280 mg/L, corresponding to 14,950 kg of BOD_5 per day, 34,383 kg of COD per day, and 17,250 kg of suspended solids (TSS) per day.

Treated effluent is discharged into the Oued Safsaf River, a sensitive water body that plays an important ecological role in the region. The produced excess sludge is dewatered and reused as an agricultural fertilizer, which highlights the need for reliable effluent quality control. Incomplete organic removal increases the amount of residual biodegradable matter, raising the moisture content and hindering drying, thereby reducing the quality of the final biosolid product.

Samples used in this study (2015–2024) were digitized from the WWTP's historical archives. The dataset represents continuous monitoring of influent and effluent streams across all 12 months of each year. The monitored parameters included biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), total suspended solids (TSS), pH, temperature, Electrical Conductivity, and dissolved oxygen (DO). All analyses were performed in accordance with the Standard Methods for the Examination of Water and Wastewater (APHA, 2017) [4].

Workflow and modeling strategy

This section delineates the development of a high-accuracy model for predicting effluent BOD_5 in a WWTP. The approach uses real operational dataset we applied the XGBoost algorithm to optimize performance.

Domain-informed feature engineering was applied to enhance model accuracy, and a time-based train/test split (80 % / 20 %) ensured realistic evaluation. The pipeline includes data preprocessing (handling missing values and removing outliers), feature creation, model training, and evaluation using R^2, MSE, and MAE. *Fig. 1* illustrates the complete workflow, from raw data to final prediction, highlighting a data-driven framework for effluent BOD_5 estimation.

Data collection, cleaning, and preprocessing

Data collection and initial screening

The dataset spans nine years (2015–2024) and includes 410 valid records, corresponding to an average of (3–4) measurements per month due to irregular sampling frequency. Each record contains 15 operational parameters measured at both the influent (inlet) and effluent (outlet) of the treatment process. These variables were selected based on their relevance to biological treatment performance and the availability of consistent data.

Tab. 1. Water treatment process monitoring data

Features	Unit	Inflow	Outflow
Date	dd/mm/yy	Date	Date
Potential of hydrogen (pH)	-	pH_int	pH_out
Biochemical Oxygen Demand at 5 days (BOD_5)	mg/L	BOD_5_int	BOD_5_out
Chemical Oxygen Demand (COD)	mg/L	COD_int	COD_out
Electrical Conductivity	µS/cm	Cond_int	Cond_out
Total Suspended Solids (TSS)	mg/L	TSS_int	TSS_out
Temperature (T)	°C	Temp_int	Temp_out
Dissolved Oxygen (DO)	mg/L	O_2_int	O_2_out

BOD_5_out was selected as the target variable because it is the primary measure of wastewater quality. The standard 5-day test is considered too slow for active management; however, rapid estimation is facilitated by predictive modeling compared to conventional measurement methods. This allows problems to be caught early and the plant to be run more efficiently.

Fig. 2 illustrates the temporal trends for influent and effluent BOD_5 concentrations between 2015 and 2025. While influent levels (BOD_5_int, blue) show significant volatility, peaking during 2017–2018 due to seasonal factors or hydraulic surges, the effluent levels (BOD_5_out, green) remain stable and low from 2019 onward. This stability confirms consistent treatment efficacy despite variable influent loads. Data gaps in 2016, 2018, and 2022 resulting from sensor or logging issues were managed during the preprocessing phase.

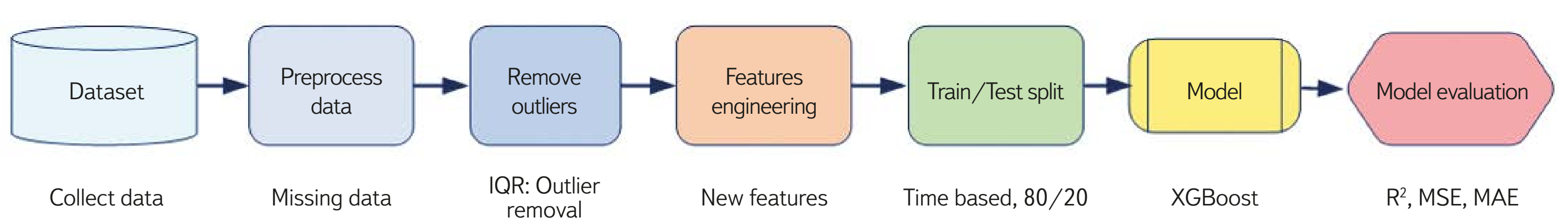

Fig. 1. Flowchart of the modeling pipeline for predicting effluent BOD_5

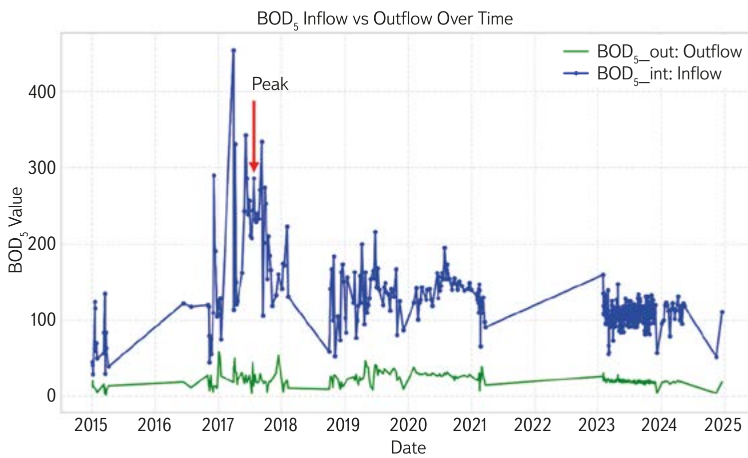

Fig. 2. Time series of BOD_5 inflow and outflow showing consistent treatment performance despite significant fluctuations

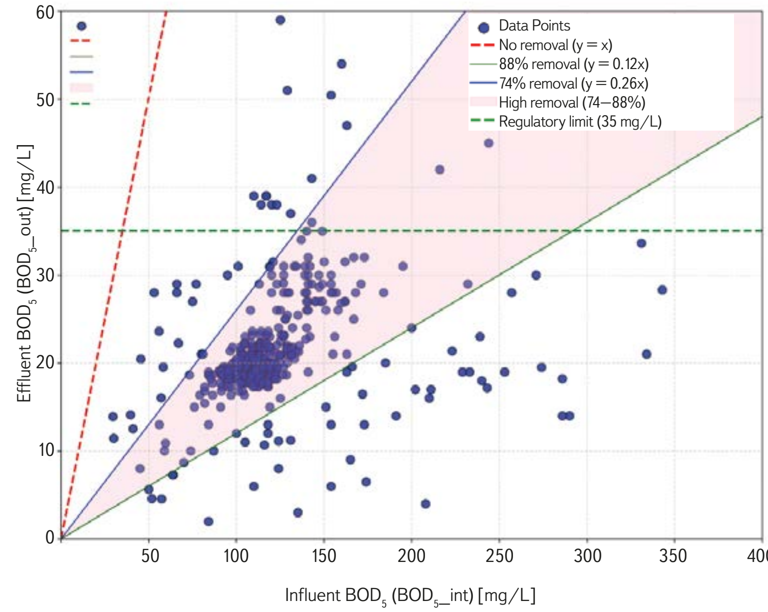

Fig. 3. Scatter plot of effluent versus influent BOD_5

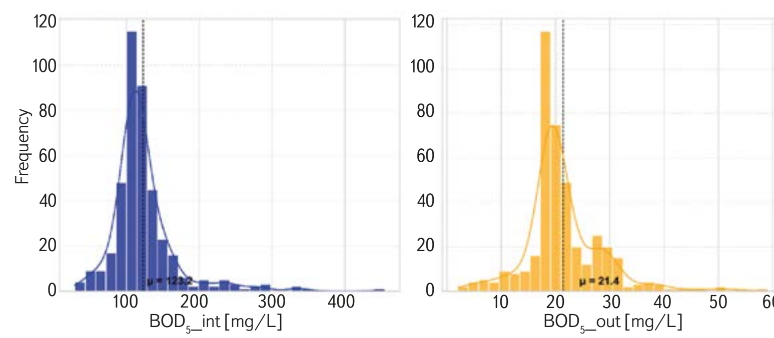

Fig. 4. Distribution of influent and effluent BOD_5 values between 2015 and 2024; there is a clustering around central values with right-skewed tails, which indicates the occurrence of rare, high-load events

Fig. 3 illustrates the relationship between influent and effluent BOD_5 values, highlighting the overall removal performance of the treatment system across a wide range of influent loads. A dense cluster of data points, where influent BOD_5 ranges from 80 to 180 mg/L and effluent concentrations remain between 10 and 47 mg/L, indicates consistent and effective organic load reduction across operational cycles. This corresponds to an estimated removal efficiency of approximately 74–88 %, as illustrated in *Fig. 3* by the region bounded between the lines (y = 0,26x) and (y = 0,12x), which represent 74 % and 88 % removal, respectively. Notably, the majority of effluent BOD_5 values remain below the regulatory discharge limit of 35 mg/L, in accordance with the discharge limits established by Algerian environmental regulations [19]. The observed non-linear distribution reflects the inherent complexity of biological treatment processes, thereby supporting the application of advanced machine learning approaches for performance modeling. As shown in *Fig. 3*, several isolated data points deviate from the main cluster, exhibiting relatively high effluent BOD_5 concentrations despite moderate influent loads. These deviations are not random but likely correspond to transient operational disturbances, such as temperature fluctuations or hydraulic shocks (sudden increases in flow rate or organic loading), that affect microbial activity and treatment efficiency. They may also reflect measurement errors or temporary equipment malfunctions. Confirming the exact cause of these events is challenging, however, due to the absence of supplementary data such as detailed flow rates, rainfall records, ambient temperature, or on-site operational logs. Despite this limitation, the presence of such anomalies underscores the importance of rigorous data preprocessing, particularly for effluent BOD_5. It also highlights the need for adaptive modeling strategies capable of maintaining predictive accuracy under variable and non-ideal operating conditions.

The BOD_5_int and BOD_5_out histograms in *Fig. 4* show right-skewed distributions, with most values clustered around central peaks (e.g. ~200 mg/L for BOD_5_int) and a few high values extending to the right (> 400 mg/L). These extremes are not necessarily errors; they may reflect real events like industrial discharges or stormwater inflows. However, they require careful handling during preprocessing stage, since they can distort model training and affect its accuracy and credibility. A balanced approach was used: extreme values linked to actual disturbances were retained; isolated outliers were removed to ensure data quality.

Data preprocessing

Raw data exhibited missing values, inconsistencies, and outliers due to measurement and logging issues. A structured preprocessing pipeline was applied to enhance data quality, covering missing data treatment, outlier detection, and temporal alignment.

Handling missing data

The initial examination of the dataset revealed that missing values were represented by the symbol '/', particularly in the dissolved oxygen variables (O_2_int and O_2_out). More than 62 % of the observations in these columns were missing, likely due to sensor downtime, maintenance activities, or intermittent data recording issues, rendering them unsuitable for reliable modeling. Keeping such variables would introduce significant bias or require heavy imputation, potentially distorting the true process behavior. Therefore, to preserve data integrity, both variables were removed from the analysis.

After replacing all '/' entries with NaN and converting the remaining inlet (int) and outlet (out) measurements to numeric format, the other variables (pH, BOD_5, COD, TSS, temperature, and the recording date) were found to be complete with no missing values. It is evident that the number of variables has been reduced from 15 (the sampling date, 7 influent variables, and 7 effluent variables, as listed in *Tab. 1*) to 13 following the exclusion of Dissolved Oxygen (DO) from both the influent and effluent data. Moreover, no missing values were identified that required imputation. These predictors were therefore used directly in the modeling process.

It should be emphasized that dissolved oxygen (DO) is a fundamental operational parameter in activated sludge systems, as it directly governs microbial activity and nitrification efficiency [11]. Its exclusion from the present analysis was driven by substantial data loss exceeding 60 %, rather than by any diminished relevance of (DO) to the underlying treatment processes.

Outlier detection and removal

Outliers in the target variable, effluent BOD_5 (BOD_5_out), were identified using the Inter-Quartile Range (IQR) method, a robust statistical approach commonly applied in environmental data analysis [20]. The IQR is defined as:

$$IQR = Q_3 - Q_1 \qquad (1)$$

where:

Q_1 and Q_3 represent the 25th and 75th percentiles, respectively.

Observations outside the following bounds were considered outliers:

$$\text{Lower Bound} = Q_1 - 1.5 \times \text{IQR} \qquad (2)$$
$$\text{Upper Bound} = Q_3 + 1.5 \times \text{IQR} \qquad (3)$$

For BOD_5_out, Q_1 = 18.18 mg/L and Q_3 = 28.63 mg/L, yielding an IQR of 10.45 mg/L. Accordingly, the lower and upper thresholds were calculated to be 2.50 mg/L and 44.30 mg/L, respectively. Applying this criterion resulted in the removal of 53 extreme observations, including unusually high values (e.g., 59 mg/L) that were likely attributable to measurement errors or occasional process disturbances. This reduced the dataset from 410 to 357 samples.

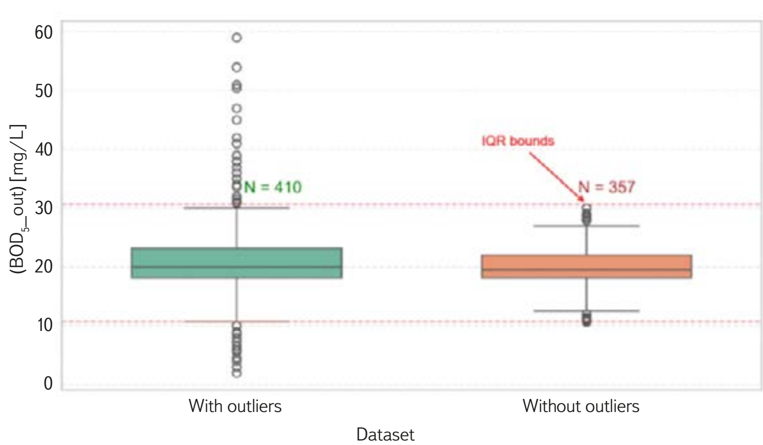

Fig. 5. Box plot comparison of BOD_5_out before and after outlier filtering

Next, rows containing missing values in either the target variable or key predictors were excluded. The final curated dataset for modeling consisted of N = 357 complete samples. As shown in *Fig. 5*, these preprocessing steps improved the symmetry and reduced the spread of the distribution, thereby strengthening the robustness and generalizability of the predictive models.

Temporal Alignment of Data

The data column was parsed into a datetime format and used to sort the dataset chronologically. This approach was adopted to ensure the preservation of temporal dependencies, thereby facilitating the validation of the model's realism.

Feature expansion and target selection for improved prediction

When developing the machine learning model, it became apparent that the initial dataset contained insufficient features to achieve high predictive accuracy. To address this limitation, we expanded the feature space through targeted engineering, carefully selecting the most relevant input and output features. This section outlines the process of generating new features and defining the target variable to improve model performance.

Selection of feature and target variable

In predictive modelling, all available system parameters may be considered as candidate inputs, but not all contribute equally to performance. Selecting and engineering informative features is therefore essential, as irrelevant variables can reduce accuracy [21]. To improve predictive capacity, the effluent outlet parameters were also integrated as inputs, a strategy shown to enhance accuracy in similar applications.

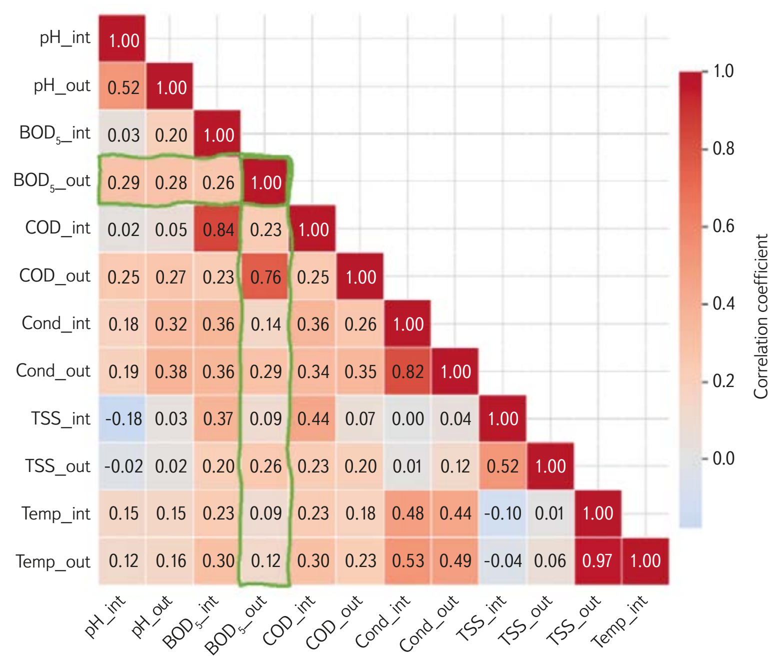

Fig. 6. Pearson correlation matrix

Fig. 6 displays the Pearson correlation matrix, showing the linear relationships among operational variables in the dataset. Strong positive correlations (red) indicate variables that increase together (e.g., COD_int and BOD_5_int), while negative correlations (blue) show opposite trends. As expected, influent BOD_5 and COD display a strong positive correlation (r = 0.84), and effluent BOD_5 and COD are also highly correlated (r = 0.76). Temperature and TSS_int show weak correlations with other variables, suggesting they have little impact on short-term BOD_5 changes. Temperature (Temp_int, Temp_out) and suspended solids (TSS_int) exhibit weak correlations with the remaining features (*Fig. 6*). However, these variables were retained to capture subtle influences on biological oxygen demand (BOD_5) and enhance model precision. This brings the total feature count to 13 variables for the final model architecture.

Feature engineering: creating new predictive input

Feature engineering transforms raw data into informative variables that enhance model interpretability, predictive performance, and generalization [20, 21]. Guided by WWT principles, domain-informed features were derived to better capture treatment dynamics. Six additional features were engineered: removal efficiencies, ratios, and rate-of-change indicators. This expanded the initial 13-variable set to a total of 18 variables, excluding the sampling date. Preliminary evaluation confirmed that these domain-specific variables improved model accuracy, particularly for process control tasks. All engineered features are summarized in *Tab. 2,* including their mathematical definitions and physical interpretations. Each feature was validated for physical plausibility and analyzed for its correlation with the target variable, BOD_5_out.

Tab. 2. Engineered features and their mathematical descriptions

Engineered Feature	Formula		Notes
COD_removal_eff	$\frac{COD\_int - COD\_out}{COD\_int}$	(4)	Efficiency of COD removal; higher values mean better pollutant reduction.
pH_change	$pH\_out - pH\_int$	(5)	Difference in pH between influent and effluent, showing neutralization effects.
$COD\_BOD_5\_ratio\_int$	$\frac{COD\_int}{BOD_5\_int}$	(6)	COD/BOD5 ratio at the inlet; high values indicate low biodegradability.
$COD\_out\_BOD_5\_int\_ratio$	$\frac{COD\_out}{BOD_5\_int}$	(7)	The COD/BOD_5 outlet/inlet ratio indicates how treatment affects biodegradability and how much pollution remains compared to the input.
$BOD_5\_int\_pH\_int$	$BOD_5\_int \times ph\_int$	(8)	Models how pH levels influence the ability of bacteria to break down organic matter (BOD5).
Cond_ratio	$\frac{Cond\_out - Cond\_int}{Cond\_int}$	(9)	The total increase or decrease in conductivity (Cond_int, Cond_out) from inlet to outlet.

Train/test split strategy

Because the dataset spans nearly a decade (2015–2024) and contains sequential time-stamped measurements, a time-based train/test split was applied to preserve temporal order and prevent data leakage. Unlike random splitting, which risks introducing temporal bias by allowing future data to influence training, this approach chronologically allocated the first 80 % of samples (286 observations, January 2015–December 2022) for training/validation and the remaining 20 % (71 observations, January 2023-May 2024) for testing, thereby simulating real-world deployment where models are trained on historical data evaluated on feature periods [22, 23].

This approach evaluates the model under realistic forecasting conditions that capture both seasonal variation and long-term operational trends. The performance metrics used in this study, including R^2 (which indicates how much of the variability in effluent BOD_5 the model can explain), MSE (which highlights the presence of large prediction errors), and MAE (which reflects the average deviation between predicted and measured values), therefore provide a meaningful picture of the model's true predictive ability. By strictly separating the test data from the training and tuning phases, the evaluation remains unbiased and representative of real-world deployment.

Selected model: XGBoost architecture and hyperparameters

XGBoost (Extreme Gradient Boosting) was selected for its high computational efficiency and superior performance in handling non-linear biological data. It is a highly optimized implementation of gradient boosting that combines decision trees in an additive manner, using gradient descent to minimize prediction error. It enhances training efficiency and model performance through:

- Second-order gradients (Hessian-based optimization), enabling faster convergence,
- Built-in L1 (Lasso) and L2 (Ridge) regularization to control complexity and reduce overfitting,
- Column and row subsampling to improve generalization.

Due to its robustness, speed, and high predictive accuracy, XGBoost has become a leading algorithm in machine learning competitions (e.g., Kaggle) and real-world applications across engineering, finance, and environmental sciences [24–25].

In this study, the XGBoost model was trained using the following hyperparameters:

- n_estimators = 200: number of boosting rounds (trees),
- learning_rate = 0.05: step size shrinkage to prevent overfitting,
- max_depth = 5: limits tree depth to balance model flexibility and interpretability,
- subsample = 0.8: uses 80 % of samples per tree to introduce randomness and improve robustness,
- colsample_bytree = 0.8: randomly selects 80 % of features at each split, enhancing diversity,
- reg_alpha = 0.1, reg_lambda = 0.1: applies mild L1 and L2 regularization to penalize large coefficients.

These parameters were selected based on prior experimentation and align with best practices for regression tasks. The model was implemented using the xgboost Python library (version 1.7.6).

Performance evaluation metrics

The predictive performance of the models was assessed using three standard regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and the Coefficient of Determination (R^2). The mathematical definitions are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (10)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (11)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \quad (12)$$

where:

y_i is the observed BOD^5_out value
$\hat{y}_i$ the predicted value
$\bar{y}_i$ the mean of the observed values
n the number of samples in the test set

MAE measures the average magnitude of prediction errors, treating all deviations equally and providing a clear and intuitive indicator of overall accuracy. It is widely used for its interpretability, particularly in environmental and geospatial modeling. MSE, by squaring each error before averaging, assigns greater weight to large deviations, making it effective for detecting models prone to producing substantial outliers [26]. R^2 complements these metrics by indicating the proportion of variance in the target variable explained by the model, a value close to 1 reflects a strong fit, whereas lower values indicate weaker explanatory capability. All performance metrics were calculated on the held-out test set (January 2023–May 2024), ensuring an unbiased assessment of the model's generalization ability.

RESULTS AND DISCUSSION

This section presents the performance of the individual base models and the stacked ensemble in predicting effluent BOD_5 (BOD_5_out). The analysis includes quantitative results, visualizations, and contextual interpretation to highlight the model's accuracy, robustness, and practical relevance.

Model performance analysis

The predictive performance of the XGBoost model was evaluated after rigorous data cleaning and feature engineering. From the initial dataset, 357 high-quality samples remained for analysis, which were partitioned into a training set (n = 286) and a test set (n = 71).

The model was constructed using 18 features, comprising both raw sensor measurements and engineered indicators such as removal efficiencies and COD/BOD_5 ratios. Of these, 17 features served as input variables, including 11 original influent and effluent parameters and 6 engineered variables. The remaining variable (BOD_5_out) was designated as the target output. Performance metrics for the test dataset are summarized in *Tab. 3*.

The results demonstrate a high degree of predictive accuracy. An R^2 of 96.95 % suggests that the model effectively captures the non-linear dynamics of the treatment process. The low MAE (0.146) is particularly significant, as it indicates that the model's average prediction error is negligible relative to the regulatory limits of BOD_5.

Tab. 3. XGBoost performance metrics for BOD5 prediction

Metric	**Value**
Mean Squared Error (MSE)	0.05
Mean Absolute Error (MAE)	0.146
Coefficient of Determination (R^2)	**96.95 %**

Fig. 7 shows the close alignment between actual and predicted BOD_5 levels. The model accurately tracks both stable periods and fluctuations, maintaining high precision even during high-load events. This performance exceeds results in previous studies [13–16]. Additionally, our use of time-based validation ensures the model is reliable for real-world operations and free from data leakage.

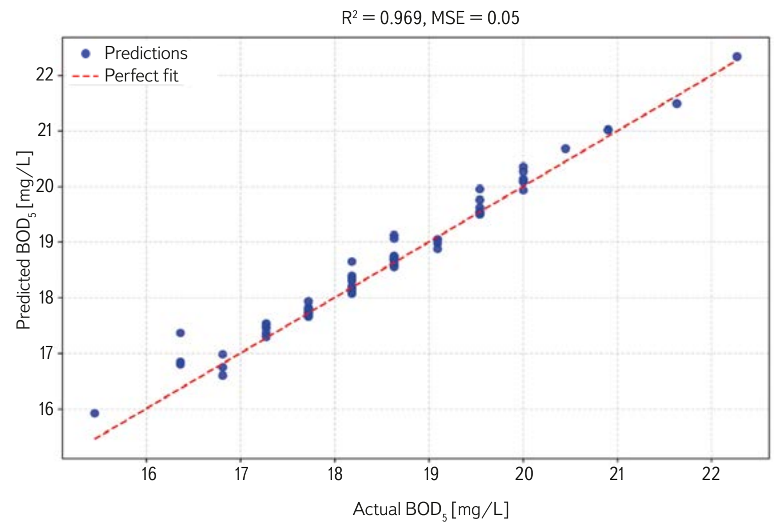

Fig. 7. Actual vs. predicted BOD_5_out values; the red dashed line represents perfect prediction (y = x)

Feature importance and process insights

The inclusion of a total of 17 features was pivotal to the model's success. Specifically, the integration of domain-informed features, such as COD_removal_eff, COD_out_BOD5_int_ratio, BOD5_int_x_pH and the net change in conductivity Cond_ratio, (equation 4, 7, 8 and 9 in *Tab. 2*), provided the model with deeper context regarding the biological health of the system. The low MSE (0.05) further confirms the model's stability, showing that it does not produce large, erratic errors (outliers).

This robustness is critical for dependable monitoring under operational conditions, as it allows plant operators to trust the predicted effluent values for proactive process adjustments.

Temporal prediction accuracy

Fig. 8 illustrates the temporal evolution of actual versus predicted BOD_5 over the test period. The model accurately tracks the dynamic behavior of the treatment process, including short-term spikes and gradual trends. Despite minor fluctuations, the model swiftly recovers and continues to provide precise tracking. Residual analysis further confirms stable performance. As shown in *Fig. 9*, residuals are randomly distributed around zero, with no visible trend or increasing variance over time. This indicates homoscedastic errors and the absence of structural bias, both of which are essential properties for reliable deployment in real-world monitoring systems.

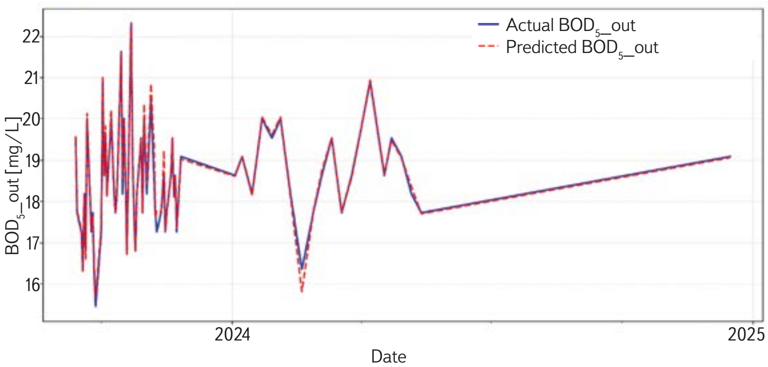

Fig. 8. Temporal comparison of actual and predicted effluent BOD_5 over the test period

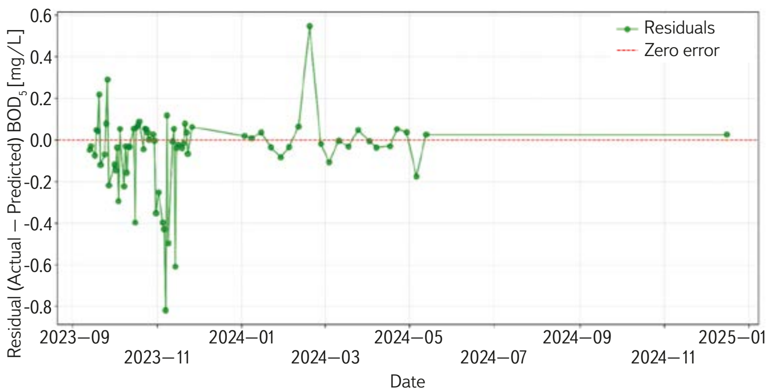

Fig. 9. Residuals over the test period; errors are centered around zero, with no systematic drift

Comparison with literature

Tab. 4 compares the performance of this study with recent works on BOD_5/COD prediction in WWTPs. Our model outperforms several existing approaches in terms of R^2. Notably, many previous studies used random train/test splits, which may lead to inflated performance estimates due to temporal leakage. In contrast, this study employs a strict chronological split (80 % / 20 %), preserving the time-series nature of the data and ensuring its generalizability under real deployment scenarios. In contrast, our model provides high accuracy while using strict validation procedures and ensuring full reproducibility.

Tab. 4. Comparative performance of BOD_5 prediction models in recent literature

Study	**Method**	**R^2 [%]**
Zhang et al. (2025) [13]	Random Forest	~89 %
Jafar et al. (2022) [14]	Deep Neural Network	89–93 %
Sethy et al. (2025) [15]	XGBoost	~93 %
Ali et al. (2025) [17]	CatBoost	~88 %
This study	XGBoost (time-split)	**96.95 %**

Generalizability and transferability considerations

Despite the strong performance achieved at the study site (R^2 = 0.9695), the model's applicability beyond this context remains constrained by its reliance on site-specific training data. Predictive performance is strongly influenced by local influent characteristics, operational conditions, and environmental factors embedded in the training dataset. Consequently, reliable performance is expected primarily for facilities with similar process configurations and operating regimes.

Factors limiting the model's transferability include variations in influent composition, hydraulic loading, and process design (e.g., aeration strategy, sludge age, and nutrient removal stages), as well as regional conditions such as climate variability and industrial discharge profiles. Applying the model to plants with substantially different characteristics without recalibration may result in reduced accuracy or increased uncertainty. Additionally, the model is primarily suitable for medium-load operations and may not be directly applicable to low-load treatment plants.

Future work should focus on multi-site validation and the integration of domain adaptation or transfer learning techniques to enhance model robustness and enable broader deployment across diverse wastewater treatment systems.

CONCLUSION

This study demonstrates that a standalone XGBoost model can achieve highly accurate prediction of effluent BOD_5 (BOD_5_out) using nearly a decade of operational data (2015–2024) from a full-scale WWTP. With an R^2 of 96.95 %, MSE of 0.05, and MAE of 0.146 mg/L on a chronologically held-out test set, the model delivers exceptional performance while avoiding the complexity of ensemble fusion strategies.

Crucially, this level of accuracy was achieved through a combination of domain-informed feature engineering, including COD removal efficiency, pH variation, and COD/BOD_5 ratios, together with rigorous time-based validation that ensures realistic evaluation under forecasting conditions.

Despite the absence of dissolved oxygen from the dataset, the model maintained strong predictive performance, indicating that routinely monitored variables such as effluent COD, conductivity, and influent pH captured sufficient information about process conditions and biological activity. Consequently, while the fundamental importance of DO in treatment kinetics remains undisputed, these findings indicate that highly correlated operational parameters can partially mitigate the absence of direct DO measurements in predictive modeling.

More importantly, the model acts as a practical sensor, providing nearly instantaneous estimates of BOD_5 from direct measurements. By replacing the 5-day laboratory delay with instant predictions, it turns BOD_5 from a static compliance value into a useful control variable. This allows:

- Early detection of process upsets: Continuous BOD_5 prediction enables early detection of abnormal biological performance caused by load variations or hydraulic disturbances, allowing timely operational interventions before effluent quality deteriorates.
- Optimization of aeration energy use: Predicted BOD_5 allows for precise aeration control, ensuring oxygen supply matches organic load. This eliminates energy waste from over-aeration while maintaining high process performance.
- Proactive adjustments to maintain regulatory compliance: Anticipatory BOD_5 forecasting enables preventive operational adjustments when predicted values approach discharge limits. This ensures consistent regulatory compliance, thereby avoiding heavy fines as well as environmental and economic damage.

In summary, this study shows that a straightforward model can still achieve excellent accuracy. With proper data handling and domain knowledge, a tuned XGBoost model can equal or surpass more complex methods while remaining faster, easier to interpret, and simpler to deploy. This makes it a strong option for scalable wastewater management applications.

References

[1] METCALF & EDDY, Inc., TCHOBANOGLOUS, G., BURTON, F. L., STENSEL, H. D. *Wastewater Engineering: Treatment and Resource Recovery*. 5th ed. New York, NY: McGraw-Hill Education, 2014. ISBN 978-0-07-340118-8. Available at: https://www.mheducation.com/highered/product/wastewater-engineering-treatment-and-resource-recovery-metcalf-and-eddy.html?viewOption=student

[2] YOUNG, J. C., CLARK, J. W. Second-Order Analysis of BOD Curve. *Journal of the Sanitary Engineering Division*. 1965, 91(1). Available at: https://doi.org/10.1061/JSEDAI.0000532

[3] PAINTER, H. A. A Review of Literature on Inorganic Nitrogen Metabolism in Microorganisms. *Water Research*. 1970, 4(6), pp. 393–450. ISSN 0043-1354. Available at: https://doi.org/10.1016/0043-1354(70)90051-5

[4] APHA. *Standard Methods for the Examination of Water and Wastewater*. 23rd ed. Washington, DC: American Public Health Association, 2017. ISBN 9780875532875.

[5] JOUANNEAU, S., RECOULES, L., DURAND, M. J., BOUKABACHE, A., PICOT, V., PRIMAULT, Y., LAKEL, A., SENGELIN, M., BARILLON, B., THOUAND, G. Methods for Assessing Biochemical Oxygen Demand (BOD): A Review. *Water Research*. 2014, 49, pp. 62–82. ISSN 0043-1354. Available at: https://doi.org/10.1016/j.watres.2013.10.066

[6] OLSSON, G., NIELSEN, M., YUAN, Z., LYNGGAARD-JENSEN, A., STEYER, J.-P. *Instrumentation, Control and Automation in Wastewater Systems*. London: IWA Publishing, 2005. ISBN 9781780402680. Available at: https://doi.org/10.2166/9781780402680

[7] CHING, P. M. L., SO, R. H. Y, MORCK, T. Advances in Soft Sensors for Wastewater Treatment Plants: A Systematic Review. *Journal of Water Process Engineering*. 2021, 44, 102367. ISSN 2214-7144. Dostupné z: https://doi.org/10.1016/j.jwpe.2021.102367

[8] SHYU, H. Y., CASTRO, C. J., BAIR, R. A., LU, Q., YEH, D. H. Development of a Soft Sensor Using Machine Learning Algorithms for Predicting the Water Quality of an Onsite Wastewater Treatment System. *ACS Environmental Au*. 2023, 3(5), pp. 308–318. Available at: https://doi.org/10.1021/acsenvironau.2c00072

[9] PATTNAIK, B. S., PATTANAYAK, A. S., UDGATA, S. K., PANDA, A. K. Machine Learning Based Soft Sensor Model for BOD Estimation Using Intelligence at Edge. *Complex & Intelligent Systems*. 2021, 7, pp. 961–976. ISSN 2198-6053. Available at: https://doi.org/10.1007/s40747-020-00259-9

[10] JU, F., GUO, F., YE, L., XIA, Y., ZHANG, T. Metagenomic Analysis on Seasonal Microbial Variations of Activated Sludge from a Full-Scale Wastewater Treatment Plant over 4 Years. *Environmental Microbiology Reports*. 2014, 6(1), pp. 80–89. ISSN 1758-2229. Available at: https://doi.org/10.1111/1758-2229.12110

[11] HENZE, M., GUJER, W., MINO, T., VAN LOOSDRECHT, M. *Activated Sludge Models ASM1, ASM2, ASM2d and ASM3*. London: IWA Publishing, 2000, 5. ISBN 1900222248. Available at: https://doi.org/10.2166/9781780402369

[12] DANTAS, M. S., CHRISTOFARO, C., OLIVEIRA, S. C. Artificial Neural Networks for Performance Prediction of Full-Scale Wastewater Treatment Plants: A Systematic Review. *Water Science and Technology*. 2023, 88(6), s. 1 447–1 470. ISSN 0273-1223. Available at: https://doi.org/10.2166/wst.2023.276

[13] ZHANG, X., ZHANG, Y., YANG, X., WANG, Z., LIU, X. Biochemical Oxygen Demand Prediction Based on Three-Dimensional Fluorescence Spectroscopy and Machine Learning. *Sensors*. 2025, 25(3), 711. ISSN 1424-8220. Available at: https://doi.org/10.3390/s25030711

[14] JAFAR, R., AWAD, A., JAFAR, K., SHAHROUR, I. Predicting Effluent Quality in Full-Scale Wastewater Treatment Plants Using Shallow and Deep Artificial Neural Networks. *Sustainability*. 2022, 14(23), 15598. ISSN 2071-1050. Available at: https://doi.org/10.3390/su142315598

[15] SETHY, K., MUDHOLKAR, M., MUDHOLKAR, P. K., BEHERA, P. K., KARTHIK, B., BALAMURALITHARAN, S. Predicting Water Quality Using Ensemble Machine Learning Models and Remote Sensing Data. *International Journal of Environmental Sciences*. 2025, 11(16s), pp. 166–175. ISSN 2229-7359. Available at: https://doi.org/10.64252/cw5pq178

[16] ELMAADAWY, K., ABD EL-AZIZ, M., ELSHEIKH, A. H., MOAWAD, A., LIU, B., LU, S. Utilization of Random Vector Functional Link Integrated with Manta Ray Foraging Optimization for Effluent Prediction of Wastewater Treatment Plant. *Journal of Environmental Management*. 2021, 298, 113520. ISSN 0301-4797. Available at: https://doi.org/10.1016/j.jenvman.2021.113520

[17] ALI, A., ZHOU, G., TAN, Y., ANTEZANA LOPEZ, F. P. Biochemical Oxygen Demand Estimation Using Explainable Ensemble Learning Methods. *Remote Sensing Applications: Society and Environment*. 2025, 41, 101835. ISSN 2352-9385. Available at: https://doi.org/10.1016/j.rsase.2025.101835

[18] GULSHIN, I., KUZINA, O. Machine Learning Methods for the Prediction of Wastewater Treatment Efficiency and Anomaly Classification with Lack of Historical Data. *Applied Sciences*. 2024, 14(22), 10689. ISSN 2076-3417. Available at: https://doi.org/10.3390/app142210689

[19] Algerian Executive Decree No. 06-141 of 19 April 2006 establishing the limit values for liquid industrial effluent discharges (official French text: *Décret exécutif n° 06-141 du 19 avril 2006 définissant les valeurs limites des rejets d'effluents liquides industriels*). *Journal Officiel de la République Algérienne Démocratique et Populaire (JORA)*, No. 26, 23 April 2006, pp. 4–6 (in French). Available at: https://www.joradp.dz/FTP/jo-francais/2006/F2006026.pdf

[20] KOUKARAS, P., TJORTJIS, C. Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices. *AI*. 2025, 6(10), 257. ISSN 2673-2688. Available at: https://doi.org/10.3390/ai6100257

[21] ZHENG, A., CASARI, A. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. Sebastopol, CA: O'Reilly Media, 2018. ISBN 978-1491953242. Available at: https://www.oreilly.com/library/view/feature-engineering-for/9781491953235/

[22] BÜYÜKKEÇECI, M., OKUR, M. C. A Comprehensive Review of Feature Selection and Feature Selection Stability in Machine Learning. *Gazi University Journal of Science and Technology*. 2023, 36(4), pp. 1 506–1 520. Dostupné z: https://doi.org/10.35378/gujs.993763

[23] HYNDMAN, R. J., ATHANASOPOULOS, G. *Forecasting: Principles and Practice*. 3rd ed. Melbourne, Australia: OTexts, 2021. ISBN 9780987507136. Available at: https://otexts.com/fpp3/ffp2/

[24] GÉRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. 2nd ed. Sebastopol, CA: O'Reilly Media, 2019. ISBN 978-1492032649. Available at: https://www.amazon.com/Hands-Machine-Learning-Scikit-Learn-TensorFlow/dp/1492032646

[25] CHEN, T., GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: ACM, 2016, pp. 785–794. ISBN 978-1450342322. Available at: https://doi.org/10.1145/2939672.2939785

[26] ROBESON, S. M., WILLMOTT, C. J. Decomposition of the Mean Absolute Error (MAE) into Systematic and Unsystematic Components. *PLOS ONE*. 2023, 18(2), e0279774. ISSN 1932-6203. Available at: https://doi.org/10.1371/journal.pone.0279774

Authors

Dr. Nassima Lekouaghet[1]
ORCID: 0000-0003-1694-1430

Dr. Lakhdar Bouragbi[2, 3]
ORCID: 0009-0002-3170-1275

Dr. Brahim Mahfoud[3]
✉ b.mahfoud@univ-bouira.dz
ORCID: 0000-0002-0709-8958

[1]Laboratory of Civil Engineering and Hydraulics, University of 8th May 1945, Guelma (Algeria)

[2]Department of Process Engineering, Faculty of Technology, University 20 aout 1955, Skikda (Algeria)

[3]Energy and Mechanics Laboratory (LEM), Faculty of Applied Science, University of Bouira (Algeria)

This paper has been peer-reviewed.

DOI: 10.46555/VTEI.2026.05.003