Chemical Process Design Vol.5 No.2 (2026) 160-183

University of Sistan and Baluchestan

Chemical Process Design

Available online at http://cpd.usb.ac.ir/

Real-Time COD Prediction in Full-Scale Wastewater Treatment Using Ensemble Learning Models and Systematic Feature Selection

Maryam Pahlavan[1], Bahareh Bidar[2]

[1] Department of Chemical Engineering, Shahid Nikbakht School of Engineering, University of Sistan and Baluchestan, Zahedan, Iran. Email: maryam.pahlavan@gmail.com

[2] Corresponding Author, Department of Chemical Engineering, Shahid Nikbakht School of Engineering, University of Sistan and Baluchestan, Zahedan, Iran. Email: b.bidar@eng.usb.ac.ir

ARTICLE INFO

Article type:
Research Article

Article history:
Received: 2026-07-02
Received in revised form: 2026-07-28
Accepted: 2026-08-18
Available online: 2026-08-18

Keywords: *Chemical oxygen demand; Wastewater treatment plant; Ensemble machine learning; Feature selection; LightGBM*

ABSTRACT

Accurate and timely prediction of Chemical Oxygen Demand (COD) is essential for efficient wastewater treatment plant (WWTP) operation; however, conventional laboratory-based COD measurements are time-consuming and unsuitable for real-time process monitoring. This study develops and compares three ensemble machine learning models—Random Forest (RF), XGBoost, and LightGBM—for COD prediction using six years of operational data collected from the full-scale Melbourne Eastern Treatment Plant. A multi-stage feature selection framework combining Average Rank aggregation, Pearson correlation analysis, and Sequential Backward Elimination was employed to reduce the original 19-variable input space to an optimal subset of five predictors: Total Nitrogen (TN), Biological Oxygen Demand (BOD), year, mean air temperature, and minimum air temperature. Model hyperparameters were optimized using Randomized Search with Time Series Split cross-validation, and predictive performance was evaluated using a chronological train–test split. The results demonstrated that feature selection improved model generalization while reducing model complexity. Among the evaluated algorithms, LightGBM achieved the highest predictive performance on the test dataset (R^2 = 0.716, RMSE = 76.30 mg/L, and MAE = 56.59 mg/L), outperforming XGBoost (R^2 = 0.682) and RF (R^2 = 0.653). TN and BOD consistently emerged as the most influential predictors across all models, highlighting their strong relationship with COD dynamics in wastewater treatment processes. Overall, the findings indicate that LightGBM combined with a systematic feature selection strategy provides an efficient and practically deployable framework for near-real-time COD prediction, offering a promising alternative to conventional laboratory analyses for WWTP monitoring and control.

Cite this article: Pahlavan, M., Bidar, B., (2026), Real-Time COD Prediction in Full-Scale Wastewater Treatment Using Ensemble Learning Models and Systematic Feature Selection, *Chemical Process Design*, 5(2), 160-183. http//doi.org/10.22111/cpd.2026.56069.1095

DOI: http//doi.org/10.22111/cpd.2026.56069.1095

Publisher: University of Sistan and Baluchestan.

1. Introduction

Clean water is one of the most critical natural resources underpinning ecological health, public wellbeing, and sustainable development. The rapid expansion of global industrialization, urban population growth, and intensified agricultural activity have substantially increased the discharge of organic pollutants into surface water bodies and municipal sewer networks, placing growing pressure on wastewater treatment plants (WWTPs) to consistently meet stringent effluent quality standards [1,2]. Among the suite of water quality parameters routinely monitored in WWTP operations, Chemical Oxygen Demand (COD) occupies a central position. Defined as the mass of oxygen required to chemically oxidize organic compounds present in a water sample, COD serves as the principal aggregate indicator of organic contamination in both influent and effluent streams, and regulatory compliance with COD discharge limits is legally mandated in most national and regional jurisdictions [3, 4].

Despite its fundamental importance, the reliable and timely determination of COD remains a significant operational challenge. The standard analytical method—based on dichromate oxidation under acidic conditions at elevated temperature—requires a minimum digestion period of two hours, involves the use of toxic and environmentally hazardous reagents including potassium dichromate and mercuric sulfate, and imposes a substantial measurement lag that is incompatible with the demands of real-time process monitoring and adaptive control [5,6]. This temporal delay is particularly consequential during episodes of influent shock loading, where rapid fluctuations in organic concentration require prompt operational response to prevent effluent quality violations. Furthermore, the cost and complexity of conventional COD analysis constrain the frequency of measurements at many facilities, especially in developing-country contexts where monitoring resources are limited [7].

These limitations have motivated extensive research into alternative approaches for COD estimation, broadly divisible into two streams: (i) soft sensor-based methods that infer COD from correlated, readily measurable process variables without direct chemical analysis [8,9,10]; and (ii) data-driven machine learning (ML) models that predict COD from operational parameters—such as influent flow rate, temperature, biological oxygen demand (BOD), total nitrogen (TN), and ammonia—that are routinely logged in SCADA and laboratory information management systems [11,12]. The present study belongs to the latter category and focuses on developing and comparing ensemble ML models for COD prediction from operational WWTP data.

The application of artificial intelligence and ML to WWTP performance modeling has expanded rapidly over the past two decades, driven by the increasing availability of high-frequency operational data and advances in computational methods [2,7]. Early studies established the feasibility of feed-forward backpropagation artificial neural networks (ANNs) for predicting effluent quality parameters including COD, with reported coefficients of determination (R^2) typically reaching 0.85–0.99 under stable operating conditions [3,13,14]. Abba and Elkiran [15] employed ANNs to predict effluent COD using influent parameters including BOD, pH, conductivity, TN, and TSS, demonstrating that ANNs substantially outperformed multiple linear regression. Ruben et al. [16] further confirmed the effectiveness of multi-layer perceptron networks for COD prediction in river restoration contexts, achieving correlation coefficients up to 0.978 with sensitivity analysis revealing dissolved oxygen and temperature as secondary predictors. A comprehensive systematic review by Dantas et al. [17] confirmed that well-configured ANN models consistently achieve high predictive accuracy for WWTP effluent quality, though with a recognized vulnerability to sudden influent changes and distribution shifts between training and operational conditions.

Soft sensor architectures have emerged as a parallel and practically motivated approach to real-time COD monitoring. Wang et al. [8] developed interpretable nonlinear soft sensor models using multivariate adaptive regression splines (MARS) for influent COD prediction at a full-scale WWTP, demonstrating that explicit, interpretable models can achieve satisfactory predictive accuracy while explaining the nonlinear input–output relationships governing influent variability. Cong et al. [9] proposed an integrated soft sensor combining a simplified activated sludge process (ASP) mechanism model with a variable-structure RBF neural network for effluent COD estimation, demonstrating that hybrid physics-informed architectures can improve real-time prediction reliability under varying operating conditions. Nair et al. [10] implemented and validated a hybrid soft sensor for simultaneous TP and COD prediction in the influent and effluent streams of a full-scale WWTP, deploying the sensor via an industrial IoT infrastructure and confirming real-time applicability with acceptable accuracy. Pattanayak et al. [18] developed an IoT-based COD soft sensor using ML algorithms including K-Nearest Neighbour, Random Forest, and SVM, evaluated on over 16,000 samples from a river Ganga monitoring network, with KNN achieving the best performance in terms of response time and RMSE. More recently, Liu et al. [19] proposed a hybrid soft sensor integrating PLS regression with Kalman filtering-based time-series forecasting for simultaneous COD and dissolved oxygen prediction, achieving a 46.5% reduction in COD prediction MAE after incorporating autoregressive time-series features.

Support vector regression (SVR) and kernel-based methods subsequently emerged as alternatives to ANNs, offering improved behavior in high-dimensional feature spaces and better resistance to overfitting on limited datasets [20,21]. Liu et al. [22] employed a least-square support vector machine (LS-SVM) combined with principal component analysis (PCA) for online effluent COD prediction in an anaerobic wastewater treatment system, achieving correlation coefficients exceeding 0.975 under steady-state conditions and above 0.99 under dynamic shock-loading scenarios. Pörhö et al. [23] systematically compared multiple linear regression, PLSR, and LASSO regression with nonlinear scaling for COD prediction at a paper mill WWTP, identifying autoregressive models with exogenous inputs as most accurate for influent COD (correlation 0.89) and PLSR as most effective for effluent COD prediction over a 20-hour prediction horizon.

More recently, ensemble decision tree methods—including Random Forest (RF), Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (XGBoost)—have attracted increasing attention owing to their ability to handle heterogeneous tabular operational data without extensive preprocessing, their robustness to outliers, and their capacity to provide intrinsic feature importance estimates that support process understanding [24,25]. In a directly relevant study, Bagherzadeh et al. [11] applied four ML algorithms—ANN, LSTM, RF, and GBM—to predict daily energy consumption of the Melbourne East WWTP using six years of operational data (2014–2019). Feature selection analysis in that study revealed that COD, together with total nitrogen and influent flow, exerted the highest influence on WWTP energy consumption, while GBM demonstrated the best prediction performance among all competing algorithms, highlighting the suitability of gradient boosting approaches for complex, nonlinear WWTP operational data. Yang et al. [26] recently demonstrated that RF, XGBoost, and LightGBM can predict waste activated sludge generation at a full-scale WWTP with high accuracy (XGBoost: $R^2 = 0.911$), with SHAP analysis identifying effluent and influent COD as the most influential predictors—directly confirming the central role of COD in ensemble ML-based WWTP process modeling.

Deep learning architectures have further expanded the methodological repertoire. Wang et al. [27] proposed a hybrid CNN-LSTM model for dynamic COD prediction of urban sewage, demonstrating that the combined architecture outperformed standalone CNN or LSTM models in capturing the complex temporal dynamics of fluctuating COD concentrations. Xu et al. [28] developed LSTM-based soft sensors to monitor and forecast effluent COD, NH_4^+, and TN in a two-stage A/O treatment system, using Pearson correlation for feature selection and demonstrating that LSTM consistently outperformed multiple linear regression across all target variables. Advancing wastewater treatment further, Jarmondi et al. [12] applied LSTM, GRU, Transformer, and a hybrid GRU-Transformer to simultaneously predict BOD and COD at the Melbourne ETP—the same facility used in the present study—achieving the best COD prediction accuracy with the GRU-Transformer hybrid (R^2 = 0.96, RMSE = 11.824 mg/L). SHAP analysis in that study identified total nitrogen (mean |SHAP| = 0.036), BOD (0.016), and temperature-related factors (0.006) as the dominant COD predictors. Li et al. [29] introduced the TAED-TCN model—a triple attention-enhanced encoder-decoder temporal convolutional network—for multi-step and synchronous soft sensing of COD and NH_3, demonstrating R^2 improvements of over 26% relative to baseline models for 1-hour ahead COD prediction and achieving a 79% reduction in monitoring costs over the equipment lifecycle.

For COD prediction in a full-scale municipal WWTP, Samadianfard et al. [30] proposed a CEEMDAN–GRU hybrid framework that decomposed the COD time series using complete ensemble empirical mode decomposition with adaptive noise before feeding intrinsic mode functions to recurrent learners, achieving near-perfect accuracy (R^2 = 0.9999, RMSE = 2.84 mg/L) on a held-out test set from the Gebze WWTP in Türkiye. Qiu et al. [31] compared linear regression, neural network, and random forest regression soft sensors for NH_3-N prediction in a sequencing batch reactor WWTP, finding that the recurrent neural network variant achieved the best performance (R^2 = 0.921, RMSE = 6.110 mg/L), and noting that RNN-based models are particularly effective for capturing temporal dependencies in wastewater treatment data. Hassnain et al. [32] demonstrated that Extra Trees and neural network models trained on nineteen years of industrial WRRF data could predict final effluent BOD with $R^2 \approx 0.98$, with SHAP analysis identifying effluent COD as the most influential predictor—further corroborating the tight coupling between COD and BOD that motivates the present study.

However, the computational cost and interpretability challenges of deep learning architectures have sustained continued interest in ensemble tree-based methods, particularly for operational deployments where model transparency, computational efficiency, and maintainability are valued alongside predictive accuracy. The role of feature selection in COD prediction models has received growing attention. Correlation-based, mutual information-based, recursive elimination, and tree-based importance ranking methods have all been employed to identify the most informative input variables and remove redundant predictors that may induce multicollinearity and reduce generalizability [11, 33]. Across the literature, influent BOD concentrations, total nitrogen, hydraulic loading, and treatment process parameters recur as the dominant predictors of COD in municipal and industrial WWTPs [5, 12]. Meteorological variables—particularly temperature—have been shown to exert secondary but significant effects on treatment efficiency and effluent quality, consistent with the known sensitivity of biological treatment processes to thermal conditions [11, 34].

Despite the substantial body of published work on ML-based COD prediction, several important gaps persist. The majority of published models are developed and validated on data from a single WWTP using random train-test splitting, which overestimates real-world generalizability by allowing future data to inform training. Furthermore, systematic head-to-head comparisons of RF, XGBoost, and LightGBM—the three dominant ensemble tree-based algorithms—under identical data preprocessing, feature selection, and chronological evaluation conditions for COD prediction specifically have not been reported. LightGBM, which offers significant computational advantages through histogram-based leaf-wise tree growth and Gradient-based One-Side Sampling [35], has been increasingly applied in environmental monitoring [26] but remains underrepresented in WWTP COD prediction literature relative to its demonstrated performance in comparable tabular data tasks. Additionally, the effect of systematic multi-stage feature selection on the relative performance of these algorithms—and on the physical interpretability of the resulting models—has received limited attention. These gaps collectively motivate a rigorous comparative evaluation under realistic operational conditions.

The present study addresses these research gaps through the development and systematic comparison of RF, XGBoost, and LightGBM models for COD prediction using six years of daily operational data (2014–2019) from the Melbourne Eastern Treatment Plant, a full-scale activated sludge facility. The dataset, originally compiled by Bagherzadeh et al. [11] and publicly available via the Mendeley Data repository, comprises 1,382 observations covering water quality, hydraulic, and meteorological variables. The main novelty of this study lies in combining a systematic multi-stage feature selection framework with a strictly chronological validation strategy to develop robust and interpretable ensemble machine learning models under realistic operational conditions. Unlike many previous studies that relied on random train–test partitioning or focused primarily on maximizing predictive accuracy, this work emphasizes model generalization, practical deployability, and computational efficiency. Accordingly, the objectives are to: (*i*) identify the most informative predictors of COD using a consensus feature selection strategy; (*ii*) develop and optimize RF, XGBoost, and LightGBM models under identical experimental conditions; (*iii*) compare their predictive performance and generalization capability using chronological validation; and (*iv*) evaluate the practical applicability of the best-performing model for near-real-time operational monitoring in full-scale WWTPs. The proposed framework provides both a practical methodology for developing reliable COD soft sensors and engineering insights into the key factors governing COD variability in activated sludge systems.

2. Material and methods

2.1. Case study and data description

The dataset employed in this study was originally compiled and published by Bagherzadeh et al. [11] and is publicly accessible through the Mendeley Data repository (https://data.mendeley.com/datasets/pprkvz3vbd/1). The data were collected from the Eastern Treatment Plant (ETP) of Melbourne, Australia, one of two principal wastewater treatment facilities operated by Melbourne Water serving the south-eastern metropolitan region. The ETP is a large-scale activated sludge facility that processes approximately 330 million liters of sewage per day from over 2.5 million residents, producing Class A recycled water for reuse or discharge [36].

The dataset spans six consecutive years of daily operational records (2014–2019), yielding a total of 1382 observations following data engineering procedures that included removal of records with null values, outliers, and physically implausible measurements (comprising less than 5% of the raw dataset). The dataset encompasses 20 variables classified into three major categories: (1) water quality parameters, (2) operational system parameters, and (3) meteorological variables. Chemical Oxygen Demand (COD) was designated as the target output variable. The complete specification of all input and output variables is presented in Table 1.

Table 1. Dataset specifications: input variables, output variable, units, and descriptions [36]

No.	Variable Name	Symbol	Variable Type	Unit	Description
1	Recording Year	year	Temporal input	—	Year of data collection; captures long-term operational trends
2	Recording Month	month	Temporal input	—	Month index; reflects seasonal fluctuations in water quality
3	Recording Day	day	Temporal input	—	Day index; defines temporal sequence of observations
4	Average Inflow	avg_inflow	Operational input	m^3/day	Mean daily inflow rate to the treatment system
5	Average Outflow	avg_outflow	Operational input	m^3/day	Mean daily outflow rate from the treatment system
6	Total Grid Power	total_grid	Operational input	kWh	Total daily electrical energy consumption of the WWTP
7	Ammonia	Am	Operational input	mg/L	Concentration of oxidizable nitrogenous ammonia compounds
8	Biological Oxygen Demand	BOD	Water quality input	mg/L	Indicator of biodegradable organic matter content
9	**Chemical Oxygen Demand**	**COD**	**Output (target)**	**mg/L**	**Primary organic pollution indicator; model target variable**
10	Total Nitrogen	TN	Water quality input	mg/L	Total nitrogen compounds in the water
11	Air Temperature	T	Meteorological input	°C	Daily mean air temperature
12	Maximum Temperature	TM	Meteorological input	°C	Daily maximum recorded temperature
13	Minimum Temperature	Tm	Meteorological input	°C	Daily minimum recorded temperature
14	Sea-Level Pressure	SLP	Meteorological input	kPa	Atmospheric pressure at sea level
15	Relative Humidity	H	Meteorological input	%	Daily mean relative humidity
16	Precipitation	PP	Meteorological input	mm	Daily recorded rainfall
17	Wind Speed	VV	Meteorological input	m/s	Mean wind speed
18	Average Wind Speed	V	Meteorological input	km/h	Daily mean wind speed (alternative measurement)
19	Maximum Wind Speed	VM	Meteorological input	km/h	Daily maximum recorded wind speed
20	Average Visibility	VG	Meteorological input	km	Daily mean horizontal visibility

2.2. Data preprocessing and feature selection

2.2.1. Data preprocessing

Prior to model development, all data were screened for missing values and data consistency. No missing observations were detected in the dataset; therefore, no data imputation or removal procedures were required. The original measurements were directly used for feature selection and model training. The dataset was partitioned chronologically into a training set comprising the first 970 records (e.g. 70%) and a temporally disjoint test set comprising the subsequent (30%, n = 412) records, Chronological splitting—as opposed to random splitting—was deliberately employed to reflect realistic operational deployment conditions, wherein the model must generalize to future time periods not represented in the training data, and to preserve the temporal autocorrelation structure of the time series.

2.2.2. Feature selection

Given the high dimensionality of the input space comprising 19 candidate predictors, and the potential for multicollinear and redundant variables to inflate model complexity and reduce generalizability, a systematic multi-stage feature selection framework was implemented prior to final model development. The framework combined model-based importance ranking with statistical correlation analysis and sequential elimination, proceeding in two successive stages. In first stage, all 19 candidate predictors were retained and used to independently train machine learning models. Feature importance scores obtained from the ML algorithms were subsequently aggregated using an Average Rank approach. For each predictor, its rank position within each model-specific importance list was determined and the mean rank was calculated. Variables with consistently high importance across all models achieved lower Average Rank values and were considered more informative. Based on this consensus ranking procedure, the highest-ranked variables were retained for further analysis. In the second stage, the reduced feature subset was refined through the combined application of Pearson correlation analysis and Sequential Backward Elimination (BE). Pearson correlation analysis was employed to evaluate the relationships between candidate predictors and COD, as well as potential redundancy among predictors. Subsequently, BE was applied iteratively by removing the least informative variables and retraining the models after each elimination step. The final selected features were then used as inputs for ML models.

To objectively identify the optimal feature subset across different subset sizes, a composite Overall Score was defined by integrating the predictive performance of all three machine learning models (RF, XGBoost, and LightGBM). First, the performance metrics of each model were independently normalized using min-max normalization. Since higher coefficient of determination (R^2) values indicate better performance, R^2 was normalized as Eq. (1):

$$R_{norm}^2 = \frac{R^2 - R_{min}^2}{R_{max}^2 - R_{min}^2} \tag{1}$$

In contrast, because lower Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) values represent better predictive performance, inverse min-max normalization was applied:

$$RMSE_{norm}^2 = 1 - \frac{RMSE - RMSE_{min}}{RMSE_{max} - RMSE_{min}} \tag{2}$$

$$MAE_{norm}^2 = 1 - \frac{MAE - MAE_{min}}{MAE_{max} - MAE_{min}} \tag{3}$$

where the minimum and maximum values were calculated separately for each metric across all evaluated feature subsets. The normalized scores were then averaged across the three machine learning models (RF, XGBoost, and LightGBM), and the final Overall Score was computed as Eq. (4):

$$Overall\ Score = \frac{\sum_{m=1}^{3} R_{norm}^2 + \sum_{m=1}^{3} RMSE_{norm} + \sum_{m=1}^{3} MAE_{norm}}{9} \tag{4}$$

m denotes the three machine learning models (RF, XGBoost, and LightGBM). The Overall Score is computed as the equally weighted average of the normalized R^2, RMSE, and MAE values, with higher Overall Score values indicating better overall predictive performance.

2.3. Machine learning models

This study employed three widely used ensemble tree-based machine learning algorithms for COD prediction, namely Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). These algorithms were selected because they represent the most widely used ensemble learning approaches for regression problems while employing distinct learning strategies. Random Forest is a bagging-based algorithm recognized for its robustness and resistance to overfitting, whereas XGBoost and LightGBM are gradient boosting algorithms designed to improve predictive accuracy through sequential learning. In particular, LightGBM employs a leaf-wise tree growth strategy that offers higher computational efficiency for large-scale datasets. By comparing these algorithms under an identical experimental framework, this study aims to evaluate the trade-offs among predictive accuracy, generalization capability, model interpretability, and computational efficiency, thereby providing practical guidance for selecting suitable ensemble learning models for COD prediction in full-scale wastewater treatment plants.

The proposed machine learning frameworks were developed based on the following assumptions: (*i*) the historical operational dataset is representative of the normal operating conditions of the Melbourne Eastern Treatment Plant during the study period (2014–2019); (*ii*) after preprocessing, the retained measurements are sufficiently reliable for model development; (*iii*) Chronological train–test partitioning is assumed to provide a realistic assessment of model generalization by preserving the temporal evolution of WWTP operating conditions and preventing information leakage between the training and testing datasets.

All machine learning models were implemented in Python 3.13 within the Spyder integrated development environment. The Scikit-learn, XGBoost, and LightGBM libraries were used for model development, hyperparameter optimization, and performance evaluation.

2.3.1. Random Forest (RF)

Random Forest is an ensemble learning algorithm based on the bagging principle, where multiple decision trees are independently trained using randomly sampled subsets of the training data. The final prediction is obtained by averaging the outputs of all trees. This approach reduces model variance, improves generalization ability, and minimizes the risk of overfitting. In addition, Random Forest provides an intrinsic measure of feature importance, making it a useful tool for identifying influential predictors in environmental and water quality applications.

2.3.2. Extreme Gradient Boosting (XGBoost)

XGBoost is an advanced gradient boosting algorithm that builds decision trees sequentially, with each new tree attempting to correct the prediction errors of the previous ensemble. The algorithm incorporates regularization mechanisms and efficient optimization strategies that improve predictive accuracy while reducing overfitting. Due to its ability to model complex nonlinear relationships and interactions among variables, XGBoost has become one of the most widely used machine learning methods for environmental prediction tasks.

2.3.3. Light Gradient Boosting Machine (LightGBM)

LightGBM is a gradient boosting framework designed to achieve high predictive performance with improved computational efficiency. Compared with conventional boosting algorithms, LightGBM employs optimized sampling and tree-growth strategies that significantly reduce training time and memory consumption. These characteristics make it particularly suitable for large tabular datasets while maintaining strong predictive capability. Owing to its speed and scalability, LightGBM has been increasingly applied in environmental monitoring and water quality prediction studies.

2.4. Hyperparameter optimization

The hyperparameters of the all three models were optimized using the Randomized Search cross-validated algorithm coupled with a five-fold Time Series Split cross-validation strategy. The model performance evaluation was employed as the optimization criterion. The optimized configurations were subsequently used to develop the final COD prediction models.

2.5. Model performance evaluation

The predictive performance of the developed machine learning models was evaluated using three widely adopted statistical indicators, namely RMSE, MAE, and R^2. These metrics were calculated for both the training and testing datasets to assess model accuracy and generalization capability.

RMSE quantifies the average magnitude of prediction errors while assigning greater weight to larger deviations between observed and predicted values (Eq. (5)). Consequently, lower RMSE values indicate better model performance. MAE measures the average absolute difference between observed and predicted values and provides a direct interpretation of the prediction error in the original units of the target variable (Eq. (6)). Similar to RMSE, lower MAE values correspond to higher predictive accuracy.

The coefficient of determination (R^2) was used to evaluate the proportion of variance in the observed COD values that could be explained by the model predictions (Eq. (7)). R^2 values range from 0 to 1, with values closer to 1 indicating a stronger agreement between predicted and observed values and, therefore, superior model performance. Collectively, these metrics provide a comprehensive assessment of model accuracy, robustness, and predictive capability.

$$\text{RMSE} = \sqrt{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 / N} \tag{5}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{7}$$

where $i = 1, 2, \ldots, N$ is the number of observations, and N is the total number of records. Considering $\hat{y}_i$ for model output (prediction) and y_i as real values.

3. Results and discussion

3.1. Statistical characterization of the dataset

Prior to model development, the statistical distribution of the primary variables was examined to characterize the inherent structure of the operational dataset (Fig. 1). The distribution of COD values exhibited a positive skew toward higher concentrations, reflecting episodic periods of elevated organic loading that are characteristic of real full-scale WWTP influent streams [7]. This right-skewed behavior indicates that simple linear models would be structurally inadequate for capturing the nonlinear and non-Gaussian dynamics governing COD variability. BOD exhibited a similar distributional pattern, which is physically consistent with the well-established biochemical coupling between biodegradable organic matter and total oxidizable carbon in wastewater [11]. In contrast, operational variables such as average inflow rate displayed more concentrated, near-symmetric distributions, reflecting the relatively stable hydraulic regime of the large-scale ETP facility. Temperature followed a unimodal seasonal pattern, consistent with the temperate climate of Melbourne, Australia. The non-normal, nonlinear distributional behavior observed across the dataset collectively justified the application of ensemble tree-based machine learning methods, which make no parametric assumptions about the underlying data distribution.

To further investigate the relationships among the dataset variables, a Pearson correlation analysis was performed and visualized as a correlation heat map (Fig. 2). The heat map provides an overview of both the strength and direction of pairwise linear associations among water quality, operational, and meteorological variables. As expected, COD exhibited the strongest positive correlations with TN and BOD, indicating their potential importance as predictive variables. Several meteorological and operational parameters showed comparatively weaker correlations with COD, suggesting a more indirect influence on effluent quality. In addition, the presence of moderate correlations among certain predictors highlighted potential redundancy within the feature space, thereby supporting the subsequent application of feature selection procedures. Overall, the correlation analysis provided an initial assessment of variable relevance and interdependence prior to machine learning model development.

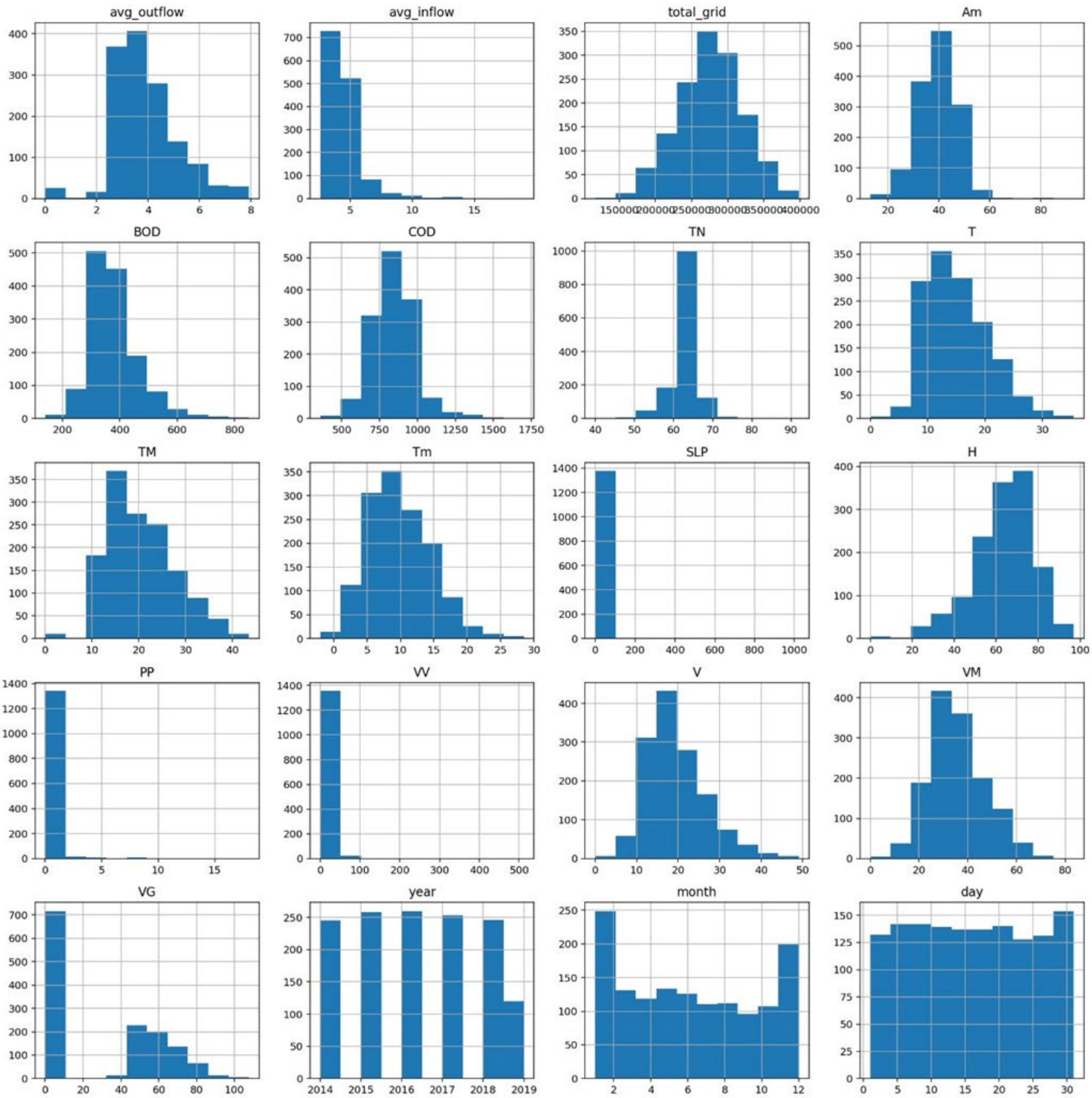


Fig. 1. Histograms illustrating the distributions of the 19 predictor variables and the target COD variable, including water quality, operational, and meteorological parameters used in this study

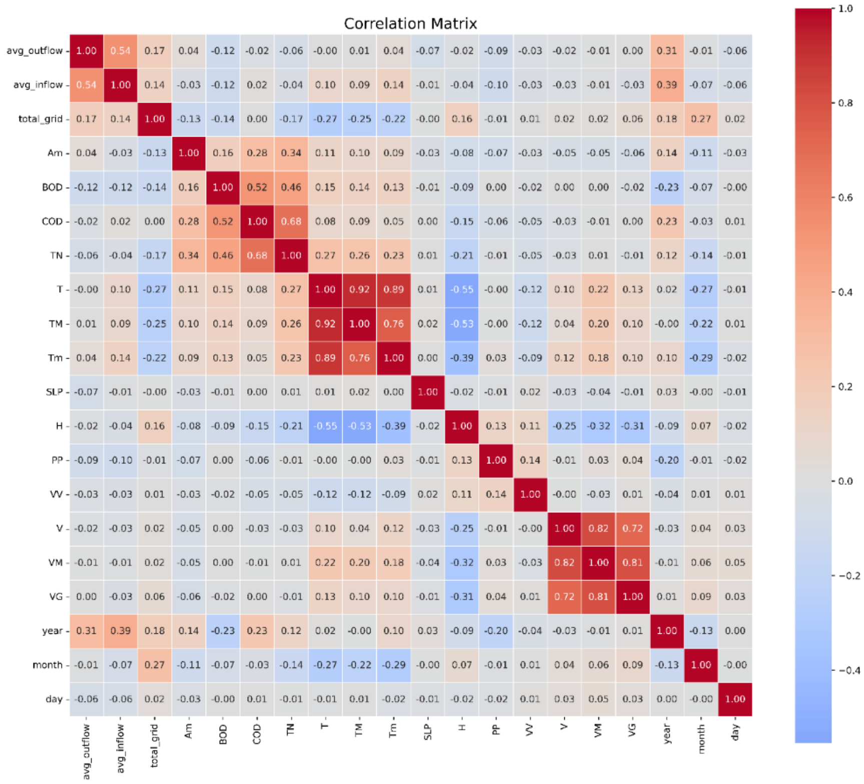


Fig. 2. Pearson correlation matrix (heat map) of the 19 predictor variables and the target COD variable

3.2. Feature selection results

3.2.1. Average rank aggregation

The feature importance scores derived from independent training of the RF, XGBoost, and LightGBM models on the complete 19-variable dataset are presented in Table 2 and visualized in Fig. 3(a–c). Although the relative importance values differ among the algorithms due to their distinct learning mechanisms, a number of predictors consistently emerged as dominant contributors to COD prediction, providing the basis for the subsequent feature selection procedure.

Table 2. Feature importance scores from RF, XGBoost, and LightGBM, and average rank aggregation results

Feature	RF Importance	RF Rank	XGB Importance	XGB Rank	LGBM Importance	LGBM Rank	Average Rank
TN	0.3866	1	0.4221	1	1225	2	**1.33**
BOD	0.2084	2	0.1167	2	1353	1	**1.67**
year	0.0345	4	0.0989	3	503	3	**3.33**
Tm	0.0257	11	0.0429	4	450	4	**6.33**
avg_inflow	0.0328	5	0.0214	10	338	5	**6.67**
total_grid	0.0294	8	0.0296	8	279	7	**7.67**
T	0.0297	7	0.0323	6	172	10	**7.67**
TM	0.0263	10	0.0359	5	162	11	**8.67**
avg_outflow	0.0304	6	0.0185	14	288	6	**8.67**
Am	0.0568	3	0.0211	13	130	13	**9.67**
month	0.0191	14	0.0298	7	197	8	**9.67**
V	0.0290	9	0.0212	12	149	12	**11.00**
H	0.0250	12	0.0213	11	126	14	**12.33**
day	0.0207	13	0.0163	16	181	9	**12.67**
PP	0.0056	18	0.0242	9	3	18	**15.00**
VM	0.0178	15	0.0183	15	52	17	**15.67**
VG	0.0114	16	0.0157	17	108	16	**16.33**
VV	0.0107	17	0.0139	18	108	15	**16.67**
SLP	~0.000	19	0.000	19	0	19	**19.00**

***Lower Average Rank = higher consensus importance across all three models.*

Across all three algorithms, Total Nitrogen (TN) and Biological Oxygen Demand (BOD) consistently emerged as the two highest-importance predictors, with TN ranking first in both RF (importance = 0.387) and XGBoost (importance = 0.422), and BOD ranking first in LightGBM (importance = 1,353) with TN second (1,225). The dominance of TN reflects the strong biochemical interdependency between total nitrogenous compounds and organic load in activated sludge treatment systems, where nitrification and carbon oxidation processes are tightly coupled [12,15]. The consistently high importance of BOD aligns with the fundamental relationship between biodegradable and total chemical oxygen demand, as BOD represents the fraction of organic matter susceptible to biological oxidation.

At the opposite extreme, Sea-Level Pressure (SLP) received near-zero importance scores in all three models (RF: 3.21×10^{-5}; XGBoost: 0; LightGBM: 0), indicating negligible predictive relevance for COD estimation. Precipitation (PP) showed inconsistent importance across algorithms (low in RF and LightGBM, moderate in XGBoost), reflecting its indirect and irregular influence on influent organic loading.

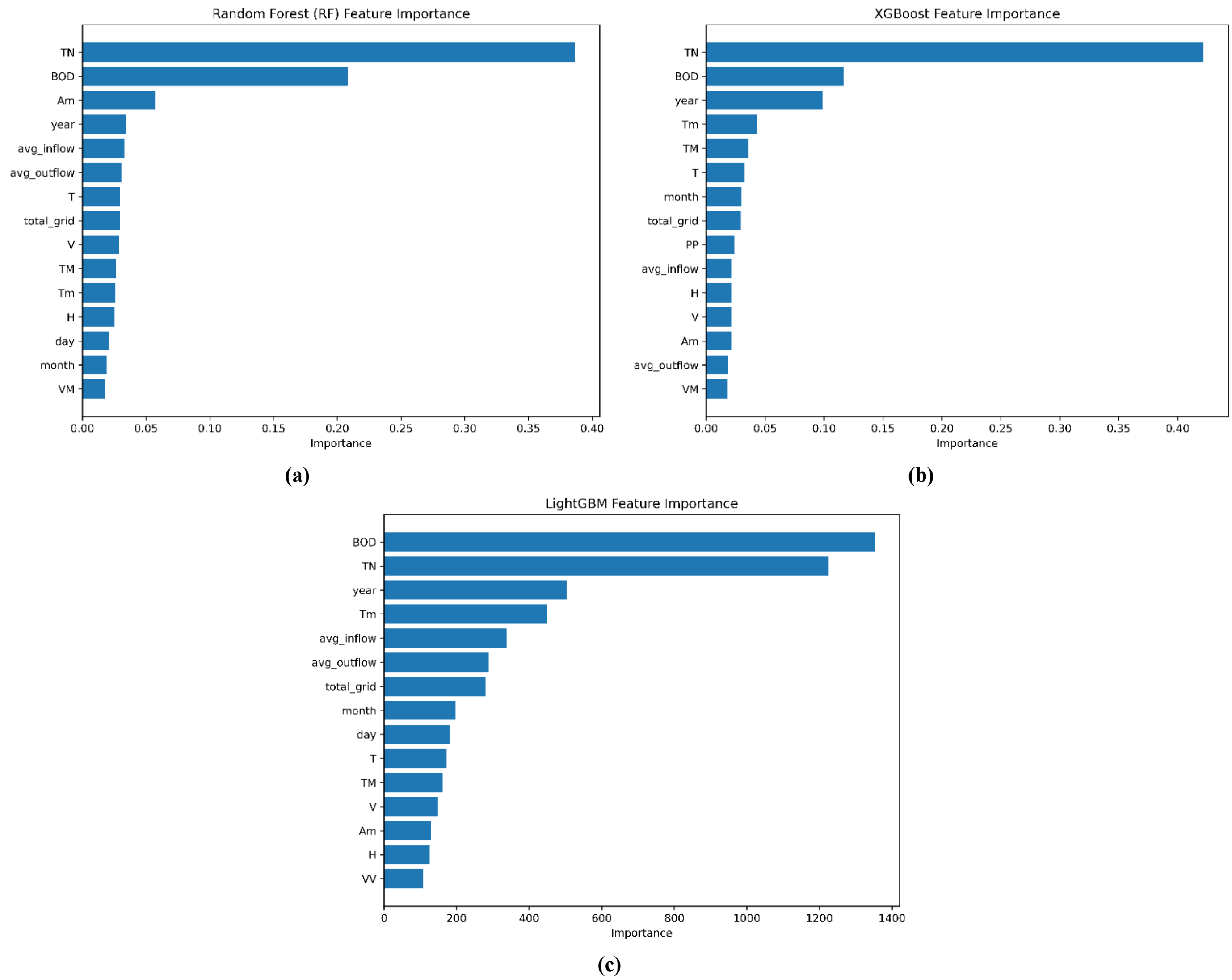


Fig. 3. Feature importance scores generated by (a) Random Forest (RF), (b) XGBoost, and (c) LightGBM using all 19 features

Based on the Average Rank aggregation results, the ten highest-ranked features (Average Rank $\leq$ 9.67) were selected as the initial reduced feature subset (n=10 features): TN, BOD, year, Tm, avg_inflow, total_grid, T, TM, avg_outflow, and Am.

3.2.2. Backward elimination and final feature subset

The ten consensus-selected features were subsequently subjected to a combined Pearson correlation and Sequential Backward Elimination procedure. The Pearson correlation analysis revealed strong collinearity among temperature variables (r(T, TM) > 0.90; r(T, Tm) > 0.88), indicating redundant information content. The correlation of TN with COD was the strongest among all predictors (r = 0.68), followed by BOD (r = 0.52), confirming their primacy as informative inputs.

The results of the Backward Elimination procedure across all feature subset sizes are presented in Table 3. Model performance was evaluated using R^2, RMSE, and MAE for all three algorithms at each elimination step. The Overall Score (Section 2.2.2) was then used to objectively rank the different feature subsets.

Table 3. Backward Elimination results: model performance metrics across all feature subset sizes

No. Features	Features	R^2			RMSE			Overall Score
		Random Forest	XGBoost	LightGBM	Random Forest	XGBoost	LightGBM	
19	All features	0.574	0.668	0.622	93.41	82.44	87.92	0.343
10	TN, BOD, year, Tm, avg_inflow, total_grid, T, TM, avg_outflow, Am	0.615	0.649	0.616	88.73	84.73	88.61	0.422
9	TN, BOD, year, Tm, avg_inflow, total_grid, T, TM, avg_outflow	0.622	0.629	0.620	87.99	87.10	88.24	0.402
8	TN, BOD, year, Tm, avg_inflow, total_grid, T, TM	0.626	0.658	0.706	87.54	83.71	77.54	0.779
7	TN, BOD, year, Tm, avg_inflow, total_grid, T	0.620	0.628	0.671	88.18	87.27	82.02	0.523
6	TN, BOD, year, Tm, avg_inflow, T	0.623	0.636	0.717	87.79	86.32	76.15	0.724
5	**TN, BOD, year, Tm, T**	**0.654**	**0.682**	**0.716**	**84.22**	**80.69**	**76.30**	**0.946**
4	TN, BOD, year, T	0.669	0.680	0.686	82.26	80.93	80.21	0.860
3	TN, BOD, year	0.619	0.648	0.689	88.32	84.91	79.83	0.629
2	TN, BOD	0.606	0.539	0.628	89.84	97.11	87.30	0.097

The overall score analysis clearly identifies **5 features** (TN, BOD, year, Tm, T) as the optimal input subset, achieving the highest composite score (0.946) across all three models simultaneously. Notably, the full 19-feature set achieved the lowest Overall Score (0.343), demonstrating that indiscriminate inclusion of all available predictors does not improve and in fact deteriorates generalization performance—a manifestation of the curse of dimensionality in high-dimensional, moderately-sized datasets. The marked improvement from 19 to 5 features confirms that the multi-stage feature selection framework successfully removed noise-contributing and redundant predictors.

The progressive elimination also reveals an important finding (Fig. 2(a-c)): reducing from 5 to 4 features (removing minimum temperature, Tm) led to a drop in LightGBM performance (R^2 declining from 0.716 to 0.686 and RMSE increasing from 76.30 to 80.21 mg/L), and further reduction to 3 variables produced consistent degradation across all models. This indicates that the 5-feature subset represents the inflection point beyond which information loss becomes detrimental. The inclusion of temporal (year) and meteorological (T, Tm) variables in the optimal subset reflects the long-term operational trends and seasonal thermal effects on biological treatment efficiency documented in the literature [11, 34].

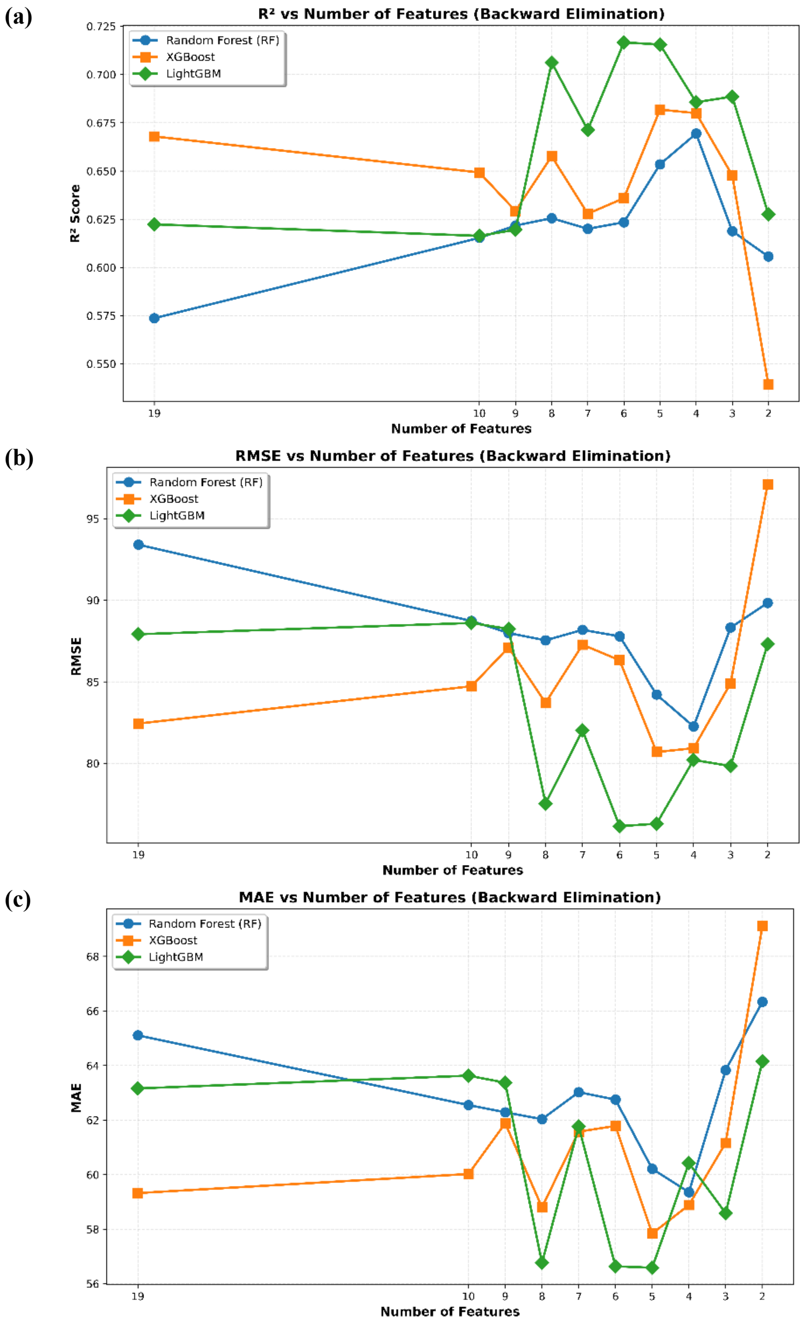


Fig. 2. Performance trends of the RF, XGBoost, and LightGBM models with varying numbers of selected features, (a) The coefficient of determination (R^2), (b) root mean square error (RMSE), and (c) mean absolute error (MAE)

3.3. Hyper parameter optimization

The hyper parameters of all three models were tuned using Randomized Search with 5-fold Time Series Split cross-validation on the training dataset. The search space and identified optimal configurations are summarized in Table 4.

Table 4. Hyper parameter search space used during Randomized Search CV, and Optimal hyper parameter configurations of the developed models based on 5 selected features

Model	Hyper parameter	Search values	Optimal value
Random Forest (RF)	n_estimators	200, 300, 500	500
	max_depth	10, 15, 20, 30	10
	min_samples_split	2, 5, 10	5
	min_samples_leaf	1, 2, 4	1
	max_features	sqrt, log2, None	sqrt
XGBoost	n_estimators	200, 300, 500	300
	max_depth	3, 4, 5, 6	3
	learning_rate	0.01, 0.03, 0.05, 0.10	0.05
	subsample	0.8, 0.9, 1.0	0.8
	colsample_bytree	0.8, 0.9, 1.0	0.8
LightGBM	n_estimators	200, 300, 500	500
	max_depth	3, 5, 7, 10	3
	num_leaves	15, 31, 63	15
	learning_rate	0.01, 0.03, 0.05, 0.10	0.05

The shallow tree depth selected for both XGBoost (max_depth = 3) and LightGBM (max_depth = 3, num_leaves = 15) is consistent with regularization-focused boosting practice, in which individual weak learners are intentionally constrained to prevent overfitting during iterative ensemble construction [35, 37]. The larger ensemble size for RF and LightGBM (n_estimators = 500) relative to XGBoost (n_estimators = 300) is also consistent with the known behavior of bagging-based versus boosting-based methods, wherein bagging benefits from larger forests while regularized boosting typically saturates at fewer trees.

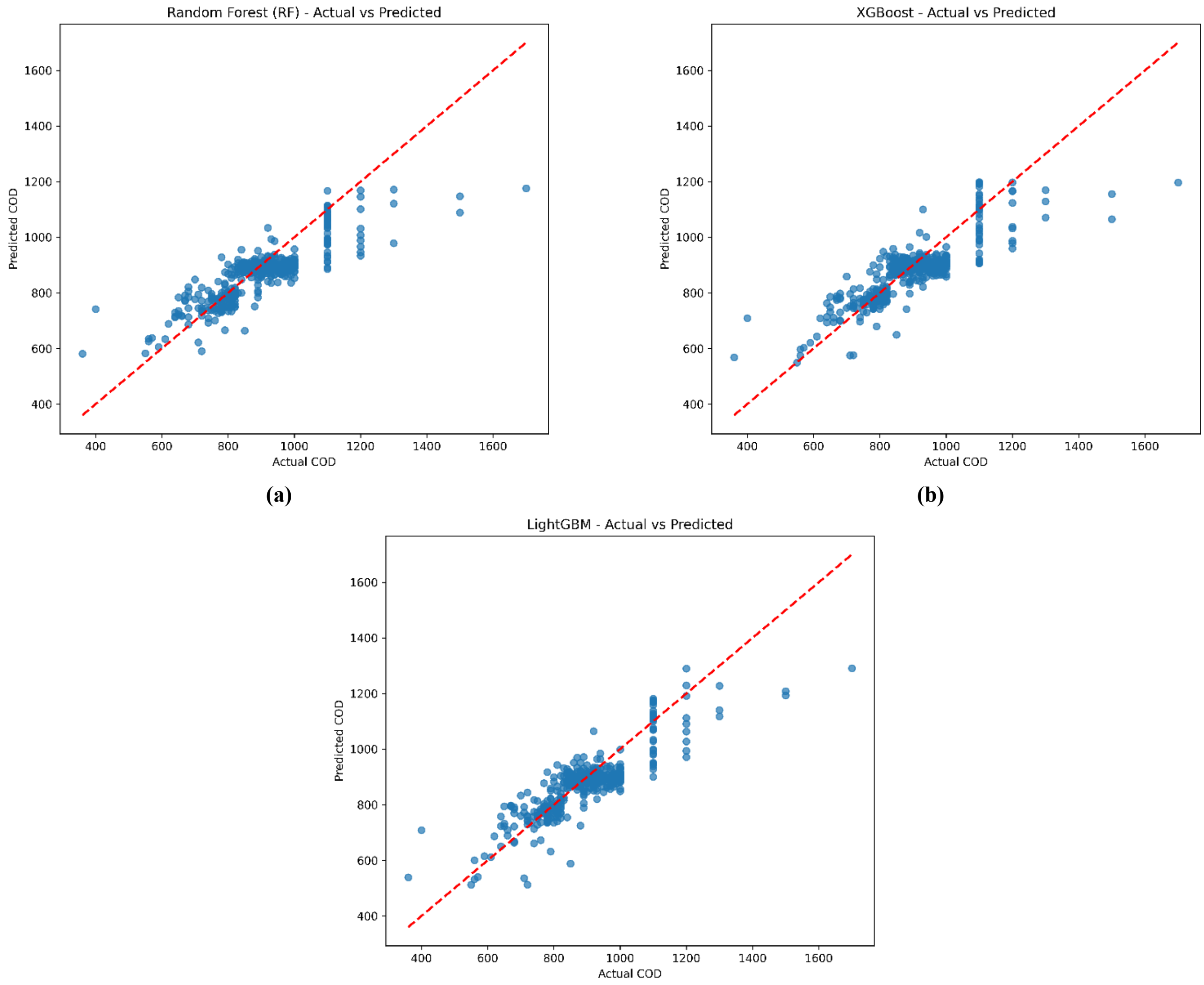
3.4. Model performance evaluation

The predictive performance of the three optimized models, trained on the 5-feature optimal subset, is summarized in Table 5. Performance metrics are reported for both the training and test datasets, along with cross-validated R^2 statistics.

Table 5. Predictive performance of RF, XGBoost, and LightGBM models for COD prediction (5-feature optimal subset)

Model	Training dataset			Testing dataset		
	R^2	RMSE (mg/L)	MAE (mg/L)	R^2	RMSE (mg/L)	MAE (mg/L)
Random Forest	0.923	39.33	26.74	0.653	84.22	60.21
XGBoost	0.896	45.80	33.23	0.682	80.69	57.84
LightGBM	**0.880**	**49.15**	**34.45**	**0.716**	**76.30**	**56.59**

Fig. 3(a–c) presents the Actual vs. Predicted scatter plots obtained for the RF, XGBoost, and LightGBM models on the independent test dataset. In all three models, the majority of data points are concentrated around the 1:1 reference line, indicating a strong agreement between predicted and measured COD values. However, noticeable differences can be observed in the degree of dispersion around the diagonal line. The model exhibiting the tightest clustering of points around the reference line demonstrates superior predictive accuracy and lower prediction error. The results confirm that the selected machine learning algorithms successfully captured the complex nonlinear relationships governing COD variability, while maintaining satisfactory generalization performance on previously unseen data.



(c)

Fig. 3. Scatter plots of actual versus predicted COD values obtained on the testing dataset using (a) Random Forest (RF), (b) XGBoost, and (c) LightGBM

Fig. 4(a–c) illustrates the time-series comparison between observed and predicted COD concentrations during the testing period (2018–2019). The predicted values closely follow the temporal fluctuations of the measured COD series, demonstrating the ability of the models to capture both long-term trends and short-term variations in wastewater characteristics. Although minor deviations are observed during periods of abrupt COD changes, the overall agreement between observed and predicted values remains high throughout the evaluation period. This behavior indicates that the developed models are capable of reproducing the dynamic behavior of the treatment system and can provide reliable COD predictions under varying operational and environmental conditions.

3.4.1. Test set performance

On the held-out test set, LightGBM achieved the best overall generalization performance, attaining an R^2 of 0.716, RMSE of 76.30 mg/L, and MAE of 56.59 mg/L. XGBoost ranked second (R^2 = 0.682, RMSE = 80.69 mg/L, MAE = 57.84 mg/L), while RF demonstrated the weakest test performance (R^2 = 0.653, RMSE = 84.22 mg/L, MAE = 60.21 mg/L). As visually confirmed in Fig. 3, LightGBM predictions cluster most tightly around the 1:1 line, with fewer systematic deviations at extreme COD values compared to RF and XGBoost.

(a)
(b)
(c)

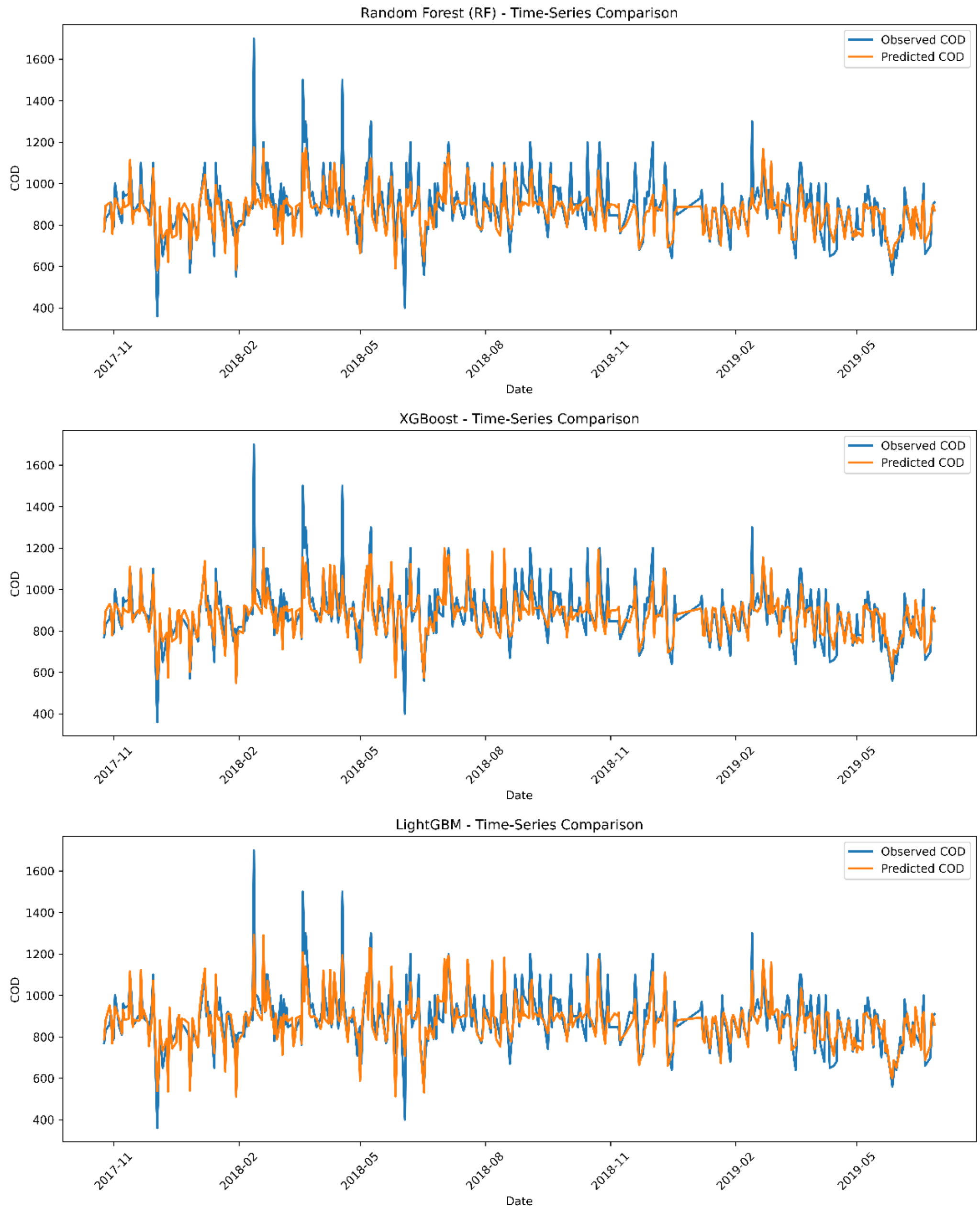


Fig. 4. Time-series comparison of observed and predicted COD concentrations on the testing dataset (2018–2019) using (a) Random Forest (RF), (b) XGBoost, and (c) LightGBM

The superior test performance of LightGBM is consistent with its optimized leaf-wise tree growth strategy and Gradient-based One-Side Sampling (GOSS) mechanism, which together reduce overfitting of the training distribution while preserving sensitivity to informative high-gradient samples [35]. The performance differential between LightGBM and RF on the test set—approximately 9.8% improvement in RMSE (76.30 vs. 84.22 mg/L)—is practically significant in the context of WWTP operations, where COD measurement accuracy directly affects regulatory compliance assessments.

3.4.2. Training performance and overfitting analysis

On the training dataset, the ranking is reversed: RF attained the highest training R^2 (0.923) and lowest training RMSE (39.33 mg/L), followed by XGBoost (R^2 = 0.896, RMSE = 45.80 mg/L) and LightGBM (R^2 = 0.880, RMSE = 49.15 mg/L). The substantial gap between training and test performance observed for RF (ΔR^2 = 0.270) indicates moderate overfitting, a well-known limitation of bagging ensembles when the training data exhibit complex interactions and the test period involves distributional shifts—as is inherent to chronological splitting of WWTP time series [17]. XGBoost exhibited a similar but smaller train-test gap ($\Delta R^2 = 0.214$), while LightGBM demonstrated the most favorable balance between training accuracy and generalization ($\Delta R^2 = 0.164$), reflecting the regularization benefits of its constrained shallow-tree architecture with num_leaves = 15. The scatter plots in Fig. 3 further corroborate this finding: RF predictions show greater dispersion around the 1:1 line, particularly at higher COD concentrations, indicating reduced reliability during peak organic loading events.

This pattern is further corroborated by the cross-validation results. XGBoost achieved the highest mean cross-validated R^2 (CV $R^2 = 0.530 \pm 0.261$), confirming its robust learning across different temporal folds. The high standard deviation in CV R^2 for all three models (0.218–0.269) reflects the inherent temporal non-stationarity of WWTP operational data, wherein treatment performance, influent composition, and environmental conditions vary substantially across seasons and years.

This variability is likely associated with temporal changes in influent composition and seasonal operating conditions over the six-year monitoring period. From a chemical engineering perspective, this temporal distributional shift can be interpreted as a manifestation of process drift in activated sludge systems. Full-scale municipal WWTPs continuously experience fluctuations in influent characteristics, hydraulic loading, microbial community dynamics, sludge age, aeration requirements, and seasonal temperature, all of which gradually modify the relationships between routinely monitored process variables and COD. Consequently, the statistical properties of the process evolve over time, making future operating conditions inherently different from those represented in the training dataset. The selection of year, mean temperature, and minimum temperature among the optimal predictors further supports the presence of long-term operational trends and seasonal effects in the dataset. Therefore, some cross-validation folds correspond to substantially different process conditions, resulting in increased variability in predictive performance. These findings highlight that the observed training–testing performance gap reflects not only the generalization challenge of machine learning models but also the inherently dynamic behavior of activated sludge processes, emphasizing the importance of chronological validation and periodic model recalibration for reliable long-term deployment in full-scale WWTPs [11].

3.4.3. Comparison with baseline (Full feature set)

A comparison of model performance using the optimal 5-feature subset versus the full 19-feature set (Table 3) reveals consistent and meaningful improvements for all three models upon feature selection. For RF, Test R^2 improved from 0.574 (19 features) to 0.653 (5 features), representing a 13.7% relative improvement, while RMSE decreased from 93.41 to 84.22 mg/L. For XGBoost, Test R^2 improved marginally from 0.668 to 0.682, and for LightGBM the improvement was most pronounced, with Test R^2 increasing from 0.622 to 0.716 (a 15.1% relative gain) and RMSE decreasing from 87.92 to 76.30 mg/L. These results confirm that the multi-stage feature selection framework substantially enhanced generalization, particularly for LightGBM, which is architecturally more sensitive to input dimensionality through its GOSS sampling mechanism.

3.5. Discussion

3.5.1. Interpretation of feature importance and selected predictors

The identification of TN and BOD as the two most influential predictors of COD is strongly grounded in the biogeochemistry of wastewater treatment. In activated sludge systems, COD represents the aggregate oxygen demand of all oxidizable organic compounds, whereas BOD captures the biodegradable fraction and TN reflects the nitrogenous oxygen demand associated with nitrification of ammonia to nitrate—a process that constitutes a significant portion of total oxygen consumption in biological treatment [11, 38]. The mutual information between these parameters and COD is therefore not merely statistical but mechanistically rooted. The additional predictive value provided by year, T (mean temperature), and Tm (minimum temperature) in the 5-feature optimal subset is consistent with the established temperature dependence of microbial metabolism: lower minimum temperatures reduce nitrification and organic degradation rates, leading to elevated effluent and residual influent COD concentrations [34]. The inclusion of the year variable captures long-term operational trends associated with progressive infrastructure upgrades and growing service population at the ETP [11].

Notably, meteorological variables such as sea-level pressure (SLP) and precipitation (PP) exhibited negligible importance for COD prediction. Although the ETP primarily serves a separated sewerage catchment, rainfall-dependent inflow and infiltration (I&I) may still occur during heavy precipitation events. Therefore, the low importance assigned to precipitation in the present study should not be interpreted as evidence of the complete absence of rainfall effects. Rather, it suggests that rainfall-related impacts were either limited during the analyzed period or indirectly captured by other operational variables included in the models.

3.5.2 Model comparison with Previous Studies

The relative ranking of the three models—LightGBM > XGBoost > RF on test set performance—is broadly consistent with recent comparative studies in WWTP machine learning applications. Nourani et al. [24] and Dantas et al. [17] reported that gradient boosting methods generally outperform bagging-based ensembles for wastewater quality prediction tasks, particularly when the training dataset size is moderate. The margin of superiority of LightGBM over XGBoost observed in the present study, however, contrasts with some earlier reports that identified XGBoost as the best-performing model for WWTP COD prediction [25, 11]. This discrepancy may be attributable to the relatively small number of selected input features ($n = 5$) and the moderate dataset size ($n \approx 970$ training records), conditions under which LightGBM's aggressive regularization through shallow leaf-wise trees provides a stronger bias-variance tradeoff than XGBoost's slightly deeper default architecture.

The absolute test R^2 values attained in this study (0.653–0.716) are somewhat lower than those reported in several single-site WWTP COD prediction studies in the literature (e.g., $R^2 \approx 0.85$–0.99 in [3]; $R^2 \approx 0.90$ in [13]). This discrepancy is likely attributable to three factors: (*i*) the present study employed strictly chronological train-test splitting, which is methodologically more rigorous but inherently more challenging than the random splitting used in many published studies; (*ii*) the 6-year dataset spans multiple operational phases and seasonal cycles that increase predictive difficulty; and (*iii*) the use of only 5 parsimonious predictors introduces an intentional constraint on model expressivity in exchange for greater generalizability. These methodological choices prioritize operational realism over inflated benchmark metrics, and the resulting performance levels remain practically meaningful for near-real-time operational monitoring and decision support in full-scale municipal WWTPs.

To further contextualize the predictive performance of the proposed model, Table 6 compares the present study with representative COD prediction studies reported in the recent literature. As shown, direct comparison of reported R^2

values should be interpreted cautiously because the studies differ substantially in dataset characteristics, validation strategies, treatment plant scale, and modeling objectives. Studies reporting exceptionally high predictive accuracy frequently relied on random train–test partitioning, laboratory-scale systems, or relatively homogeneous industrial wastewater datasets, all of which generally represent less demanding prediction scenarios than long-term forecasting using unseen future operational data.

Table 6. Table 6. Comparison of the proposed models with representative COD soft-sensing studies, highlighting differences in validation strategy, dataset characteristics, and practical applicability

Reference	Model	Test R^2	Dataset Description	Train/Test Split	WWTP Type	Remarks
Liu et al. (2019) [22]	PCA-LSSVM	>0.99 (under shock loads)	Laboratory-scale IC anaerobic reactor; 6 input variables (pH, ORP, biogas, temperature, etc.)	Random	Pilot-scale (Laboratory)	Laboratory-scale reactor with controlled conditions; tested under three shock load scenarios; random split may overestimate real-world performance
Cong et al. (2022) [9]	ASP + VSRBFNN (Hybrid)	Not reported	Full-scale WWTP in China; 250 samples; 5 input variables	Random	Municipal (Medium-scale)	Hybrid mechanism-data-driven approach; VSRBFNN compensates ASP model error; Small dataset (250 samples), limiting external generalizability
Çimen Mesutoğlu & Gok (2024) [3]	FFBANN (ANN)	0.9997	9-month daily data from industrial WWTP; 19 input parameters	Random	Industrial	Industrial WWTP with homogeneous pollution load; Short monitoring period (9 months) and random data partition may overestimate generalization performance.
Jarmondi et al. (2025) [12]	Hybrid GRU-Transformer	0.960	5-year data from Melbourne ETP; multi-domain (water quality, flow, energy, weather)	Chronological	Municipal (Large-scale)	Same Melbourne ETP dataset; chronological validation; deep learning with extensive feature engineering and SHAP interpretation; higher predictive accuracy achieved at the expense of increased model complexity.
Present Study	LightGBM XGBoost Random Forest	0.716 0.682 0.653	6-year (1382 days) from Melbourne ETP; 19 variables (quality, operational, meteorological)	Chronological	Municipal (Large-scale)	Chronological validation using completely unseen future operational data; systematic feature selection (19→5 variables); interpretable ensemble learning with low computational cost suitable for real-time operational decision support.

Note: Direct comparison of reported R^2 values should be interpreted cautiously because the cited studies differ substantially in dataset characteristics, validation strategy, treatment plant scale, and modeling objectives.

In contrast, both the present study and the work of Jarmondi et al. [12], which were conducted using chronological validation on the Melbourne Eastern Treatment Plant dataset, demonstrate that predictive performance under realistic deployment conditions is inherently more challenging, particularly when long-term operational variability is considered. Furthermore, while the Hybrid GRU–Transformer model achieved higher predictive accuracy, it required a substantially more complex deep learning architecture with extensive feature engineering. By comparison, the proposed LightGBM model achieved satisfactory predictive performance using only five carefully selected predictors, providing a favorable balance between prediction accuracy, model interpretability, computational efficiency, and ease of practical deployment.

From an engineering perspective, the primary objective of the proposed model is not to replace standardized laboratory COD analyses required for regulatory compliance, but rather to function as a soft sensor for continuous operational monitoring and early warning. Municipal wastewater treatment processes are inherently influenced by stochastic variations in influent composition, weather conditions, hydraulic loading, and biological activity, making perfect prediction unattainable using routinely measured variables alone. Under such realistic operating conditions, a model explaining approximately 72% of the temporal variability can still provide valuable operational insight by identifying abnormal loading trends, supporting timely process adjustments, and reducing reliance on labor-intensive laboratory measurements between routine sampling events. Therefore, the achieved predictive performance should be interpreted in the context of practical decision support rather than direct replacement of regulatory monitoring procedures.

3.5.3 Practical implications for WWTP operations

The ML models developed in this study offer a viable pathway toward near-real-time COD estimation from routinely logged operational parameters, without the 2-hour digestion delay of conventional dichromate analysis. In practical deployment, the 5-feature input set (TN, BOD, year, T, Tm) is readily measurable at most modern WWTPs through continuous online sensors (temperature) and daily laboratory measurements (TN, BOD), making the model operationally tractable without investment in specialized instrumentation. The LightGBM model's test MAE of 56.59 mg/L implies that real-time COD estimates would typically deviate from laboratory measurements by approximately 6.7% of the mean COD value (mean COD ≈ 846 mg/L, Table 2 of Bagherzadeh et al. [11]), which is within an operationally acceptable margin for process monitoring and early warning of compliance risk. The high CV standard deviation, however, cautions against deploying the model as a standalone regulatory compliance instrument without periodic recalibration to account for temporal distributional shifts in influent characteristics.

Furthermore, it is important to distinguish operational monitoring from regulatory compliance assessment. Regulatory discharge reporting continues to rely on standardized laboratory analyses performed according to established protocols. The proposed LightGBM model is intended to complement—not replace—these laboratory measurements by providing rapid estimates of COD between sampling events. Such real-time predictions enable earlier detection of process upsets, improved operator awareness, and more responsive process control, thereby reducing the risk of delayed corrective actions in large-scale wastewater treatment facilities.

Beyond its predictive performance, the proposed LightGBM framework also provides meaningful engineering insights into the factors governing COD variability in full-scale activated sludge systems. The integrated feature selection procedure consistently identified TN and BOD as the two most influential predictors across all ensemble models, highlighting their strong association with organic matter degradation and biological treatment performance. In addition, the selection of temperature-related variables (mean and minimum temperature) confirms the important role of seasonal conditions in influencing microbial activity and treatment efficiency. Therefore, the principal advantage of the proposed framework lies not only in achieving the lowest prediction error among the evaluated models, but also in providing an interpretable representation of the key operational and environmental drivers affecting COD dynamics.

4. Conclusion

This study presented a practical machine learning framework for predicting chemical oxygen demand (COD) in a full-scale municipal wastewater treatment plant using long-term operational data from the Melbourne Eastern Treatment Plant. By integrating systematic feature selection with ensemble learning algorithms, the proposed approach demonstrated that reliable COD prediction can be achieved using only five routinely monitored variables, thereby reducing model complexity while maintaining robust predictive performance under realistic operating conditions.

Among the evaluated models, LightGBM provided the best balance between predictive accuracy, computational efficiency, and model interpretability. The use of chronological validation, rather than conventional random data partitioning, enabled a more realistic assessment of model generalization under future operational conditions and highlighted the importance of considering temporal non-stationarity in wastewater treatment processes. These findings suggest that parsimonious ensemble learning models can serve as practical soft sensors for continuous operational monitoring, early warning of process disturbances, and decision support in large-scale municipal wastewater treatment plants.

From an industrial perspective, the proposed framework has the potential to complement conventional laboratory analyses by providing rapid estimates of COD between routine sampling events, thereby supporting more responsive process control and improving operational efficiency. Nevertheless, the proposed model is intended to assist operational decision-making rather than replace laboratory measurements required for regulatory compliance.

This study has several limitations. The developed models were trained and validated using data from a single full-scale wastewater treatment plant, and therefore their generalizability to facilities with different treatment configurations, climatic conditions, or influent characteristics requires further investigation. In addition, only routinely available operational variables were considered, and no uncertainty quantification or online adaptive learning strategy was incorporated.

Future research should focus on validating the proposed framework across multiple wastewater treatment plants, investigating transfer learning and domain adaptation techniques, incorporating uncertainty estimation and adaptive model updating, and exploring hybrid data-driven and physics-informed machine learning approaches to further enhance prediction reliability under dynamic operating conditions.

References

[1] Hernández-Chover, V., Bellver-Domingo, Á., Hernández-Sancho, F., 2018. Efficiency of wastewater treatment facilities: the influence of scale economies, Journal of Environmental Management, 228, 77–84. http://dx.doi.org/10.1016/j.jenvman.2018.09.014

[2] Zhao, L., Dai, T., Qiao, Z., Sun, P., Hao, J., Yang, Y., 2020. Application of artificial intelligence to wastewater treatment: a bibliometric analysis and systematic review of technology, economy, management, and wastewater reuse, Process Safety and Environmental Protection, 133, 169–182. http://dx.doi.org/10.1016/j.psep.2019.11.014

[3] Çimen Mesutoğlu, Ö., Gök, O., 2024. Prediction of COD in industrial wastewater treatment plant using an artificial neural network, Scientific Reports, 14, 13750. http://dx.doi.org/10.1038/s41598-024-64634-z

[4] Balogun, S., Ogwueleka, T.C., 2023. Performance prediction for wastewater treatment plant effluent COD using artificial neural network, International Journal of Environmental Science and Technology, 20, 12659–12668. http://dx.doi.org/10.1007/s13762-023-04823-x

[5] An, T., Feng, K., Cheng, P., Li, R., Zhao, Z., Xu, X., Zhu, L., 2024. Adaptive prediction for effluent quality of wastewater treatment plant: improvement with a dual-stage attention-based LSTM network, Journal of Environmental Management, 359, 120887. http://dx.doi.org/10.1016/j.jenvman.2024.120887
[6] Xie, Y., Chen, Y., Wei, Q., Yin, H., 2024. A hybrid deep learning approach to improve real-time effluent quality prediction in wastewater treatment plant, Water Research, 250, 121092. http://dx.doi.org/10.1016/j.watres.2023.121092
[7] Newhart, K.B., Holloway, R.W., Hering, A.S., Cath, T.Y., 2019. Data-driven performance analyses of wastewater treatment plants: a review, Water Research, 157, 498–513. http://dx.doi.org/10.1016/j.watres.2019.03.030
[8] Wang, X., Kvaal, K., Ratnaweera, H., 2019. Explicit and interpretable nonlinear soft sensor models for influent surveillance at a full-scale wastewater treatment plant, Journal of Process Control, 77, 1–6. http://dx.doi.org/10.1016/j.jprocont.2019.03.005
[9] Cong, Q.-M., Bo, G.-H., Shi, H.-Y., 2022. Integrated soft sensor of COD for WWTP based on ASP model and RBF neural network, Measurement and Control, 56, 1–12. http://dx.doi.org/10.1177/00202940221089272
[10] Nair, A., Hykkerud, A., Ratnaweera, H., 2022. Estimating phosphorus and COD concentrations using a hybrid soft sensor: a case study in a full-scale WWTP, Water, 14, 332. http://dx.doi.org/10.3390/w14030332
[11] Bagherzadeh, F., Shojaei Nouri, A., Mehrani, M.-J., Thennadil, S., 2021. Prediction of energy consumption and evaluation of affecting factors in a full-scale WWTP using a machine learning approach, Process Safety and Environmental Protection, 154, 458–466. http://dx.doi.org/10.1016/j.psep.2021.08.040
[12] Jarmondi, K.D.S., Tarinejad, R., Pourali, H., Hosseinpour Salehi, A., Nourani, V., 2025. Advancing wastewater treatment plant for monitoring BOD and COD through explainable AI and feature engineering techniques, Journal of Water Process Engineering, 80, 109237. http://dx.doi.org/10.1016/j.jwpe.2025.109237
[13] Nasr, M., Moustafa, M., Seif, H., El Kobrosy, G., 2012. Application of artificial neural network (ANN) for the prediction of EL-AGAMY wastewater treatment plant performance-Egypt, Alexandria Engineering Journal, 51, 37–43. http://dx.doi.org/10.1016/j.aej.2012.07.005
[14] Lou, I., Zhao, Y., 2012. Sludge bulking prediction using principal component regression and artificial neural network, Mathematical Problems in Engineering, 2012, 917071. http://dx.doi.org/10.1155/2012/237693
[15] Abba, S.I., Elkiran, G., 2017. Effluent prediction of chemical oxygen demand from the wastewater treatment plant using artificial neural network application, Procedia Computer Science, 120, 156–163. http://dx.doi.org/10.1016/j.procs.2017.11.223
[16] Ruben, G.B., Zhang, K., Bao, H., Ma, X., 2018. Application and sensitivity analysis of artificial neural network for prediction of chemical oxygen demand, Water Resources Management, 32, 273–283. http://dx.doi.org/10.1007/s11269-017-1809-0
[17] Dantas, M.S., Christofaro, C., Oliveira, S.C., 2023. Artificial neural networks for performance prediction of full-scale wastewater treatment plants: a systematic review, Water Science and Technology, 88, 1447–1470. http://dx.doi.org/10.2166/wst.2023.276
[18] Pattanayak, A.S., Pattnaik, B.S., Udgata, S.K., Panda, A.K., 2020. Development of chemical oxygen demand (COD) soft sensor using edge intelligence, IEEE Sensors Journal, 20, 14912–14921. http://dx.doi.org/10.1109/JSEN.2020.3010134
[19] Liu, H.-I., Liu, K.F.-R., Chen, S.-H., 2024. Hybrid soft sensing model for accurate COD and DO prediction: integrating PLS regression and time-series forecasting with Kalman filtering, Proceedings of the 2024 IEEE 16th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), 2024. http://dx.doi.org/10.1109/HNICEM64917.2024.11258650
[20] Mahanna, H., El-Rashidy, N., Kaloop, M.R., El-Sapakh, S., Alluqmani, A., Hassan, R., 2024. Prediction of wastewater treatment plant performance through machine learning techniques, Desalination and Water Treatment, 319, 100524. http://dx.doi.org/10.1016/j.dwt.2024.100524
[21] Sahu, P., Londhe, S.N., Kulkarni, P.S., 2023. Prediction of water quality parameters using support vector regression, Innovative Infrastructure Solutions, 8, 229. http://dx.doi.org/10.1007/s41062-023-01195-6
[22] Liu, Z.-J., Wan, J.-Q., Ma, Y.-W., Wang, Y., 2019. Online prediction of effluent COD in the anaerobic wastewater treatment system based on PCA-LSSVM algorithm, Environmental Science and Pollution Research, 26, 12828–12841. http://dx.doi.org/10.1007/s11356-019-04671-8
[23] Pörhö, H., Tomperi, J., Sorsa, A., Juuso, E., Ruuska, J., Ruusunen, M., 2023. Data-based modelling of chemical oxygen demand for industrial wastewater treatment, Applied Sciences, 13, 7848. http://dx.doi.org/10.3390/app13137848
[24] Balogun, S., Ogwueleka, T.C., Arudi, I.S., Olu, A.L., 2026. Machine learning-based modelling of effluent quality of Wupa municipal wastewater treatment plant, Abuja, Nigeria, Cleaner Water, 6, 100238. http://dx.doi.org/10.1016/j.clwat.2026.100238
[25] Ching, P.M.L., Zou, X., Wu, D., So, R.H.Y., Chen, G.H., 2022. Development of a wide-range soft sensor for predicting wastewater BOD5 using an eXtreme gradient boosting (XGBoost) machine, Environmental Research, 210, 112953. http://dx.doi.org/10.1016/j.envres.2022.112953
[26] Yang, S., Kim, J., Eom, J., Kim, M., Lee, M., Lee, K.H., 2025. Machine learning based prediction of waste activated sludge generation for optimization of WWTP operational efficiency, Environmental Research, 286, 122990. http://dx.doi.org/10.1016/j.envres.2025.122990
[27] Wang, Z., Man, Y., Hu, Y., Li, J., Hong, M., Cui, P., 2019. A deep learning based dynamic COD prediction model for urban sewage, Environmental Science: Water Research & Technology, 5, 2210–2218. http://dx.doi.org/10.1039/C9EW00505F
[28] Xu, B., Pooi, C.K., Tan, K.M., Huang, S., Shi, X., Ng, H.Y., 2023. A novel long short-term memory artificial neural network (LSTM)-based soft-sensor to monitor and forecast wastewater treatment performance, Journal of Water Process Engineering, 54, 104041. http://dx.doi.org/10.1016/j.jwpe.2023.104041
[29] Li, J., Lin, S., Zhang, L., Zhong, L., Ding, L., 2025. Innovative multistep and synchronous soft sensing prediction of COD and NH3 in WWTPs via multimodal data and multiple attention mechanisms, Water Research, 278, 123405. http://dx.doi.org/10.1016/j.watres.2025.123405

[30] Samadianfard, S., Aras, E., Yılmaz Çelik, D., Sattari, M.T., Gündüz, O., 2026. Hybrid CEEMDAN-deep learning models for COD prediction in wastewater treatment plants, Environmental Science: Water Research & Technology, 12, 1–15. http://dx.doi.org/10.1039/D6EW00014B
[31] Qiu, C., Huang, F.-Q., Zhong, Y.-J., Wu, J.-Z., Li, Q.-L., Zhan, C.-H., 2025. Comparative analysis and application of soft sensor models in domestic wastewater treatment for advancing sustainability, Environmental Technology, 46, 1959–1980. http://dx.doi.org/10.1080/09593330.2024.2415722
[32] Hassnain, M., Lee, S.M.W., Azhar, M.R., 2025. Soft sensing of biological oxygen demand in industrial wastewater using machine learning models, Journal of Water Process Engineering, 78, 108699. http://dx.doi.org/10.1016/j.jwpe.2025.108699
[33] Lv, J., Wang, H., Chen, Y., Zhang, X., 2024. Enhancing effluent quality prediction in wastewater treatment plants through the integration of factor analysis and machine learning, Bioresource Technology, 393, 130008. http://dx.doi.org/10.1016/j.biortech.2023.130008
[34] Yang, X., Wei, J., Ye, G., Zhao, Y., Li, Z., Qiu, G., Li, F., Wei, C., 2020. The correlations among wastewater internal energy, energy consumption and energy recovery/production potentials in wastewater treatment plant, Science of the Total Environment, 714, 136655. http://dx.doi.org/10.1016/j.scitotenv.2020.136655
[35] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. LightGBM: a highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems, 30, 3146–3154.
[36] Melbourne Water, 2021. Melbourne Water. https://www.melbournewater.com.au/about/strategies-and-reports/annual-report
[37] Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, 785–794. http://dx.doi.org/10.1145/2939672.2939785
[38] Yu, Y., Zou, Z., Wang, S., 2019. Statistical regression modeling for energy consumption in wastewater treatment, Journal of Environmental Sciences, 75, 201–208. http://dx.doi.org/10.1016/j.jes.2018.03.023