

AI-Driven Deployable Water Quality Management: Benchmarking Effluent BOD Prediction at a European Pulp and Paper Mill

Published as part of ACS Engineering Au special issue "AI and Machine Learning in Chemical Engineering: Breakthroughs and Applications".

Fatemeh Naeijian, Ali Arefinia, Mohammed Tamim Zaki, Kevin D. Orner, Mohammadsoroush Tafazzoli, and Lewis S. Rowles*



Cite This: *ACS Eng. Au* 2026, 6, 170–184



Read Online

ACCESS |



Metrics & More



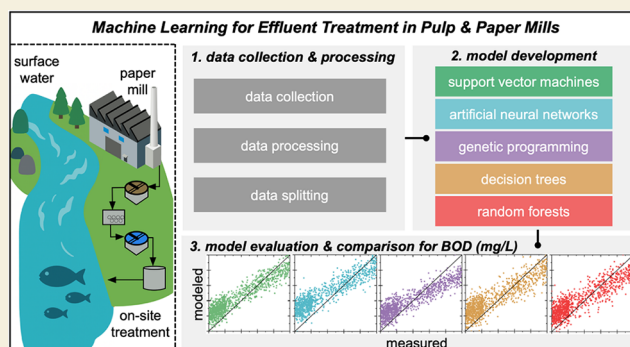
Article Recommendations



Supporting Information

ABSTRACT: The pulp and paper industry consumes 10–300 m³ of water per ton of product and generates substantial wastewater volumes. These facilities collect extensive operational data that remain underutilized for treatment optimization. However, systematic long-horizon comparisons of multiple machine learning algorithms for effluent prediction under full-scale operating conditions remain scarce. Using 1,033 daily records from a full-scale European mill (January 2020–October 2022), we evaluated five machine learning models to predict effluent biochemical oxygen demand (BOD) for environmental compliance monitoring. Input variables included flow rate, influent total suspended solids (TSS), influent BOD, temperature, aeration rate, hydraulic retention time, and chemical dosing. We benchmarked support vector machines (SVMs), artificial neural networks (ANNs), genetic programming (GP), decision trees (DTs), and random forests (RFs). SVM achieved the best performance ($R^2 = 0.85$, NSE = 0.82, RMSE = 5.5 mg/L), while ANN delivered competitive accuracy ($R^2 = 0.82$, RMSE = 6.8 mg/L) with the fastest runtime. GP produced interpretable mathematical equations, and DT/RF provided high operational transparency despite lower accuracy. These models can inform operational decisions related to aeration control, chemical dosing, and compliance management. By combining routinely monitored process variables with laboratory-measured influent BOD, our models demonstrated strong retrospective predictive performance across a three-year data set. This work establishes a benchmark for pulp and paper wastewater management and clarifies pathways for integrating real-time sensors to enable predictive process control.

KEYWORDS: biochemical oxygen demand (BOD), environmental compliance, Europe, industrial wastewater treatment, machine learning (ML), pulp and paper mills



1. INTRODUCTION

As the world's third largest water consumer, the pulp and paper industry uses 10–300 m³ of freshwater per ton of paper produced, depending on the type of process, raw materials, and the level of water reuse implemented.^{1–4} This high-water usage is closely linked to high energy consumption for extensive pumping and treatment.⁵ Consequently, the industry faces critical environmental challenges related to water and energy management particularly in wastewater treatment, where tightening discharge regulations impose increasingly strict limits on organic and nutrient loads.^{6,7} Insights into addressing these challenges may be found in the massive amount of real-time process and data generated by wastewater treatment operations, including continuous measurements of operation conditions, water quality parameters, and decision variables; thus, forming the foundation for data-driven monitoring,

optimization, and soft-sensing strategies widely discussed in recent wastewater digitalization studies.^{8–10} This wealth of data, combined with increasingly stringent environmental regulations, creates both an opportunity and an urgent need for advanced computational approaches to optimize treatment processes. Treatment facilities must ensure compliance with strict discharge limits for biochemical oxygen demand (BOD), chemical oxygen demand (COD), and sometimes nutrients

Received: October 9, 2025

Revised: January 2, 2026

Accepted: January 5, 2026

Published: January 17, 2026



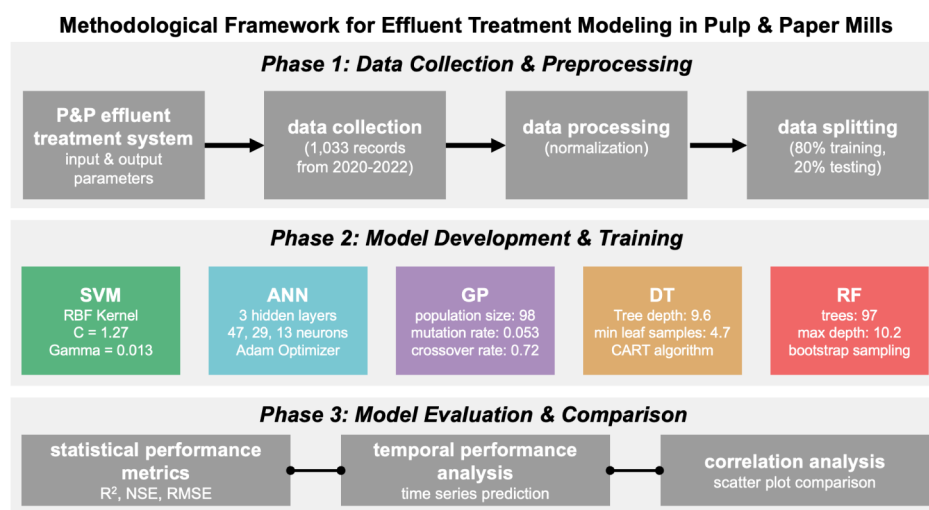


Figure 1. Methodological framework for applying machine learning models to wastewater treatment prediction in pulp and paper mills. The three-phase approach includes: (1) Data collection and preprocessing, involving acquisition of 1,033 operational records from the treatment facility (2020–2022), data normalization, and 80/20 train-test splitting; (2) model development and training, featuring the implementation of five machine learning models (SVM—green, ANN—blue, GP—purple, DT—orange, and RF—red) with their key hyperparameters; and (3) model evaluation and comparison using statistical performance metrics (R^2 , NSE, RMSE), temporal analysis, and correlation assessment. Input parameters include flow rate, TSS_{in}, BOD_{in}, temperature, aeration rate, retention time, and chemical addition, with BOD_{out} as the target output parameter.

that can severely impact receiving water bodies if not properly managed.^{11–15}

Traditional approaches to analyzing and optimizing wastewater treatment processes often rely on statistical methods or process models that can struggle to capture the complex, nonlinear relationships inherent in biological treatment systems.¹⁶ While these conventional methods have provided valuable insights, they frequently fall short when dealing with the volume and complexity of modern industrial process data. Environmental typologies, operational conditions, and biological processes interact in ways that create intricate patterns that simpler analytical approaches often miss.¹⁷ Additionally, the time-intensive nature of certain water quality measurements, such as the 5-day BOD test, creates a significant lag between process changes and performance feedback, highlighting the need for predictive capabilities. While mechanistic models can provide valuable insights into wastewater treatment processes, they often require detailed facility-specific parameters, extensive calibration, and long development times. In contrast, computational intelligence and machine learning (ML) approaches offer several advantages for pulp and paper industry applications. These models can be rapidly implemented using existing operational data, adapt to variability in wastewater composition, and scale across facilities with different treatment setups. While some ML models are considered “black box”, others (e.g., tree-based algorithms) include built-in feature importance metrics, offering transparency into variable contributions.^{18–20} This adaptability and interpretability make ML particularly useful in an industry where treatment processes and regulatory requirements vary widely.^{18,21–26}

ML techniques enable pattern discovery and prediction in complex wastewater processes^{27–31} and have been applied successfully in water and wastewater systems.^{30,32} The ability of ML methods to handle complex variable interactions makes them particularly valuable for optimizing wastewater treatment processes where multiple factors simultaneously influence treatment efficiency.^{33,34} Specifically, these methods are well

suited to multivariate, nonlinear, sensor-noisy processes in which complex variable interactions challenge purely mechanistic surrogates.^{19,34} Comparative evidence across different studies suggests characteristic trade-offs among different methods.^{35–39} Recent work also emphasizes explainability, uncertainty quantification, and hybridization to improve reliability and decision support at scale.^{18,36,40} Despite the promising applications of ML methods in wastewater treatment, significant knowledge gaps persist in comparing their predictive capabilities specifically in long-term benchmarking, cross-model comparison, and deployment-oriented validation for pulp and paper wastewater systems. Most studies examine single ML techniques using limited data sets.^{3,41} Additionally, many studies use short monitoring horizons or narrow plant-scale validation with sparse analyses of runtime, edge-case behavior, and deployment constraints.^{3,41}

The relative performance of different ML methods under varying operational conditions, and their ability to handle the unique characteristics of pulp-and-paper wastewater, has not been systematically evaluated. To address this gap, we benchmark five ML techniques using nearly three years of continuous full-scale operational data, offering broader insight than studies limited to single models or short-duration data sets. Beyond retrospective applications, these models can be framed as deployable soft sensors, i.e., virtual instruments that use readily available plant data (flow, influent TSS, temperature, aeration power) to estimate variables that are costly or delayed to measure (e.g., effluent BOD). While hard sensors (i.e., UV₂₅₄ or COD probes) have future potential, their uptake in pulp-and-paper facilities has been limited by cost, calibration demands, and maintenance requirements. Establishing robust ML baselines with conventional supervisory control and data acquisition (SCADA) and lab data is therefore a critical step before widespread hard-sensor deployment. From an operational perspective, these predictive capabilities directly support daily process control decisions where rapid feedback is otherwise unavailable due to delayed laboratory assays.

The overall goal of this study was to evaluate and compare different ML approaches for predicting BOD₅ in pulp and paper mill wastewater using a long-term plant-scale data set. Despite the growing body of ML-based studies in wastewater treatment, systematic long-horizon comparisons of multiple algorithms under real full-scale pulp-and-paper operating conditions remain scarce. In particular, the relative trade-offs among accuracy, robustness, and interpretability using continuous industrial data sets have not been rigorously benchmarked. Daily records collected from January 2020 to October 2022 ($n = 1,033$) at a full-scale facility were leveraged. Five advanced models, support vector machines (SVM), artificial neural networks (ANN), genetic programming (GP), decision trees (DT), and random forests (RF), were systematically benchmarked following a structured workflow including data preprocessing, model development, and performance evaluation. This work provides a comprehensive long-term benchmarking framework that quantifies performance trade-offs and offers practical guidance for compliance monitoring and water-reuse decision support. Overall, the findings establish a data-driven framework for translating routine process data into actionable insights for sustainable wastewater management in the pulp and paper industry.

2. METHODOLOGY

The methodology consisted of three main phases as illustrated in Figure 1: (1) data collection and preprocessing, including the acquisition of 1,033 operational records spanning 2020–2022, data normalization to address variable scaling differences, and partitioning into training (80%) and testing (20%) data sets; (2) development and training of five ML approaches (SVM, ANN, GP, DT, and RF), each with optimized model-specific hyperparameters; and (3) model evaluation and comparison using statistical performance metrics, temporal prediction analysis, and correlation assessment.

2.1. Case Study Wastewater Treatment System Facility, Data Set, and Parameters

The study utilized operational data from a large-scale pulp and paper facility in northern Europe with an annual production capacity of approximately 480,000 tonnes of paper and 330,000 tonnes of pulp. The specific facility has been anonymized at their request to protect proprietary operational data while still allowing the valuable findings from this research to benefit the broader pulp and paper industry. The mill operates a comprehensive wastewater treatment system designed to process 100,000 m³/d, integrating primary, secondary, and tertiary treatment processes. The primary treatment consists of screening and sedimentation for particle removal, followed by secondary treatment incorporating activated sludge processes to degrade organic matter. The tertiary treatment stage includes filtration and chemical treatments to ensure compliance with environmental regulations. The facility has also implemented recent modernization efforts aimed at enhancing the efficiency of its wastewater treatment and reducing emissions to both water and air, which has been a common trend across the industry.¹⁵

All data used in this study are provided in the [Supporting Information: Table S1](#) (Excel) contains the complete anonymized operational data set (1,033 daily records; seven input variables and the effluent BOD₅ target), and [Table S2](#) reports descriptive statistics (mean, standard deviation, minimum, and maximum) for each parameter; reproducible

MATLAB code is included in [SI Section S1](#) with extended methods in [Section S2](#). The data included critical input parameters such as flow rate (m³/h), influent total suspended solids (TSS in mg/L), influent BOD (mg/L), temperature (°C), aeration rate (kg O₂/h), retention time (hr), and supplemental chemical addition (kg). This complete data set is in the [Supporting Information \(SI\) Table S1](#) with descriptive statistics in [Table S2](#). The key output parameter examined in this study is effluent BOD (mg/L). The data set used in this paper consists of 1,033 records, representing daily samples spanning from January 1, 2020 to October 29, 2022. This data set enabled robust model training across the full spectrum of operational conditions, allowing for development of ML models applicable to both normal operation and extreme conditions that might challenge treatment efficiency. In practice, a deployable pathway involves replacing delayed laboratory BOD inputs with either lagged SCADA features or hard sensor surrogates. Our study focuses on the former to establish a retrospective benchmark, recognizing that the latter represents a natural extension when such instruments are available.

For preprocessing, missing values (<2% of records) were imputed using linear interpolation for continuous variables (e.g., flow, temperature, aeration rate) and last-observation-carried-forward for categorical or operational flags, preserving the consistency of the operational records. Features were normalized with min-max scaling fit on the training set only; the learned scalers were then applied to the test set to avoid data leakage. Effluent BOD measurements beyond three standard deviations from the mean were retained to represent operational extremes that are critical for industrial decision-making. As a robustness check, we also repeated the analysis with winsorized BOD (capped extremes); model rankings and conclusions were unchanged.

It is important to note that the effluent BOD₅ was obtained via the standard 5-day incubation assay and recorded in the facility's daily laboratory log. For benchmarking, influent and effluent BOD₅ were paired exactly as logged; consequently, models that include influent BOD₅ should be interpreted as retrospective performance benchmarks rather than deployable same-day soft sensors. All other process variables (flow, aeration rate, temperature, retention time, chemical addition) were collected online and used directly. We did not add explicit lag features or seasonality adjustments in this benchmarking study; extending the framework with lagged inputs and/or real-time surrogates (e.g., COD, TOC) is a natural direction for deployment-oriented work.

2.2. Development and Validation of the Machine Learning (ML) Models

We selected complementary model classes, i.e., kernel methods (SVM), neural networks (ANN), symbolic regression (GP), and tree-based learners (DT, RF), to balance accuracy, speed, interpretability, and uncertainty reporting for continuous BOD₅ prediction in industrial data sets. All data analyses and modeling were completed using MATLAB software. This code is in [SI Section S1](#). MATLAB was used in this study for its effective support of ML, neural networks, and optimization tasks relevant to engineering applications. All available parameters from the wastewater treatment facility were used without explicit feature selection, as this comprehensive inclusion of process variables aimed to provide insights into the complex interactions within biological treatment systems.

All records were randomly shuffled and normalized using min–max scaling and then divided into 80% for training and 20% for testing, treating each daily record as an independent sample rather than as a time-series.^{28,42,43} We adopted an 80/20 train-test split (826/207 daily samples) to retain an independent test set and to emulate retrospective deployment conditions in which models are trained on historical data and evaluated on unseen days. Our data set size ($n = 1,033$) is sufficiently large to provide adequate representation of operational variability in both training ($n = 826$) and testing ($n = 207$) sets, ensuring robust model development while maintaining computational efficiency. Additionally, this approach allows for evaluation of model performance on completely unseen data, which is particularly relevant for wastewater treatment applications where future prediction capabilities are essential for operational decision-making.³¹ ML models were calibrated using the training data to predict and evaluate the performance of the treatment processes, with their accuracy measured against the test data. At the model-tuning stage, we used grid search with 5-fold cross-validation; the complete hyperparameter search spaces and the selected optima for all algorithms are reported in Table S3 (Supporting Information). Table S4 lists the hyperparameter grid ranges tested for each model, including regularization and kernel parameters for SVM, hidden layer configurations and learning rates for ANN, evolutionary operators for GP, and tree structure constraints for DT and RF. Latency is not a limiting factor because predictions are generated at a daily cadence. Runtime is reported as a secondary, descriptive metric to indicate computational footprint for routine retraining and potential multisite deployment, whereas accuracy and robustness determine model selection. A brief description of the ML models used are discussed below. Because influent and effluent BOD are laboratory assays with inherent delays, models that use them provide retrospective predictive skill. Establishing these retrospective models is a necessary step to form the foundation for deployable soft sensors that can later be adapted to exclude lab inputs, add lag features, or incorporate online surrogates where available.

2.2.1. Support Vector Machines (SVM). SVMs are supervised learning method that can be formulated for both classification and regression. In this study we employ Support Vector Regression (SVR) to predict continuous effluent BOD₅ values. Rather than separating classes, SVR estimates a nonlinear regression function by mapping the inputs into a high-dimensional feature space using a kernel function and then fitting an ϵ -insensitive tube that balances predictive accuracy with model smoothness.^{44–47} This formulation makes SVM well-suited for a moderate data set with nonlinear relations among process variables, where it typically provides strong generalization and resistance to overfitting. Prediction uncertainty was estimated by fitting bootstrap-based SVR models and by training quantile-oriented variants to obtain interval estimates.³⁹

Hyperparameter optimization was performed using a grid search method to test a range of values for both C (0.1–10), the regularization parameter that controls error tolerance for deviations beyond the ϵ -tube and γ (0.001–0.1), the kernel coefficient of the radial basis function (RBF) controlling the locality of nonlinear influence in the feature space. These parameters influence model complexity and generalization. Detailed mathematical formulations, including kernel functions and optimization methods, are provided in the SI Section S2,

with Figure S1 illustrating the classification of two classes by SVM.

2.2.2. Artificial Neural Networks (ANNs). ANNs are machine-learning models inspired by biological neural systems and are widely used for recognizing nonlinear patterns in complex environmental data sets.^{48–52} Recent studies have demonstrated the strong performance of ANN architectures, including and hybrid ANN–SVM variants, for wastewater treatment process prediction. An ANN consists of interconnected processing units (neurons) arranged in an input layer, intermediate (hidden) layers, and output layer. The hidden layers transform input variables through nonlinear activation functions, while the output layer produces the model prediction.^{52,53}

The ANN architecture was optimized through systematic evaluation of 15 different configurations ranging from 1 to 5 hidden layers with varying neuron counts, showing approximately 7% improvement in R^2 over single-layer alternatives while maintaining computational efficiency.^{50,52} ReLU activation functions were implemented in hidden layers to accelerate convergence and prevent vanishing gradient problems, with a linear activation function was used in the output layer to allow unbounded BOD predictions. Learning rate was tuned within the range of 0.001 to 0.01, consistent with findings from recent water quality prediction studies.³¹ The comprehensive mathematical framework, including activation functions and the back-propagation algorithm, is detailed in SI Section S2, where Figure S2 presents a schematic representation of a simple ANN architecture.

2.2.3. Genetic Programming (GP). Genetic Programming GP is an evolutionary computation technique that evolves populations of symbolic mathematical expressions to model nonlinear relationships between inputs and output.^{54–56} Unlike traditional regression methods, GP does not require specifying a predefined functional form, instead, both the structure and parameters of the model simultaneously through genetic operators such as crossover, mutation, and reproduction. GP operates on tree structures representing mathematical expressions and iteratively improves them based on a fitness function. The resulting expressions are human-readable, allowing direct interpretation of the process variables that most strongly influence effluent quality.³⁷

The GP hyperparameters were optimized through a grid search approach covering 64 parameter combinations. Population size (tested range: 50–150), mutation rate (tested range: 0.01–0.1), crossover rate (tested range: 0.6–0.9), and tree depth (tested range: 3–7) were all evaluated to balance model stability, genetic diversity, and computational efficiency. Detailed implementation steps, genetic operations, and the evolutionary process are elaborated in SI Section S2, with Figure S3 illustrating the systematic workflow of the GP algorithm.

2.2.4. Decision Trees (DT). Decision Trees (DT) was implemented in regression mode to predict continuous BOD concentrations in the effluent. Regression trees recursively partition the input space by selecting splits that minimize impurity, typically measured using mean squared error, for continuous outcomes.⁵⁷ Each terminal (leaf) node stores a predicted numeric value, making DT models inherently interpretable for operators seeking rule-based decision structures.^{58,59} DTs can capture nonlinear interactions among process variables and handle mixed data types, which is advantageous in wastewater treatment data sets.^{42,60,61} To

address their natural instability, we apply pruning and minimum-leaf constraints and estimate uncertainty with bootstrap-based prediction intervals or quantile splits.

For DT, the “node size” parameter (i.e., determines the minimum number of observations required to split a node) was optimized using a trial-and-error approach, guided by predictive intervals at three quantile levels (2.5%, 50%, and 97.5%) to capture uncertainty in treatment outcomes.^{58,61} Hyperparameter optimization was conducted through a combination of 10-fold cross-validation and grid search techniques, evaluating 36 parameter combinations. The maximum tree depth was tested from 5 to 15 levels. Minimum leaf node samples (tested range: 2–10) were also evaluated based on the trade-off between granularity and robustness, particularly for extreme BOD values. A complete explanation of node splitting criteria, tree pruning techniques, and optimization procedures is provided in SI Section S2, and Figure S4 demonstrating the tree-like structure of the model.

2.2.5. Random Forests (RF). Random Forests (RF) was implemented as an ensemble of regression trees, where each tree predicts continuous BOD concentrations. Predictions from all trees are averaged to produce the final continuous output, which reduces variance and improves generalization compared with individual trees. This regression-based formulation is appropriate for continuous effluent BOD₅ prediction and avoids information loss associated with discretizing water-quality variables. RF represents an advanced ensemble learning method that extends the DT approach. RF combines multiple DT to create a more robust and accurate prediction model. While individual decision trees can suffer from high variance in their predictions, RF addresses this limitation by aggregating results from multiple trees.^{42,58} The RF method aggregates multiple DT, which enhances stability and predictive accuracy in various data modeling scenarios.^{59,60,62–65} Similar to DT, the “node size” parameter was optimized using a trial-and-error approach, guided by predictive intervals at three quantile levels (2.5%, 50%, and 97.5%) to capture uncertainty in treatment outcomes.^{58,61} Random forests stabilize tree predictions by averaging many trees and provide built-in cross-validation via out-of-bag error.³⁵ We quantify uncertainty from the spread of tree predictions and with quantile random forests; trade-offs include lower interpretability than a single tree and a tendency to smooth extremes.

Hyperparameters, including the number of trees (50–200) and maximum tree depth (5–15), were optimized using grid search with cross-validation. Prediction proceeds by processing each new input through all trees and averaging the resulting outputs. The mathematical foundation, bootstrap sampling methodology, and out-of-bag error estimation techniques are described in detail in SI Section S2, including the procedures for aggregating predictions across multiple trees.

2.3. Correlation Analysis and Model Evaluation

To understand the relationships between operational parameters and treatment efficacy, we employed correlation analysis to quantify the linear associations between input variables and effluent BOD. Correlation analysis was performed on the full data set to explore general relationships among variables and was not used for model training or feature selection. This statistical approach calculates correlation coefficients ranging from -1 to $+1$, where values closer to $|1|$ indicate stronger relationships. By characterizing pairwise linear associations

between parameters, the analysis supports targeted adjustments to reduce pollutant discharge while minimizing operational costs through evidence-based interventions. To evaluate and compare model performance on the independent test set, three key metrics were employed, three key metrics were employed: the correlation coefficient (R^2), Nash-Sutcliffe efficiency (NSE), and the root-mean-square error (RMSE) were used as performance indicators for each model. These metrics provide complementary insights into model accuracy, with R^2 and NSE values closer to 1 indicating a better fit between predictions and observations, while RMSE values closer to zero indicate better prediction accuracy.⁶⁶ The NSE, although more commonly used in hydrology, provides a robust measure of model efficiency by comparing the predictive accuracy of the models to the observed data.⁶⁷

These metrics together provide a comprehensive assessment of model performance, accounting for both absolute accuracy (i.e., RMSE) and relative performance (i.e., R^2 and NSE). The R^2 metric quantifies the proportion of variance explained by the model, NSE evaluates the model's predictive power relative to using the mean of observations, and RMSE provides a measure of prediction error in the original units of the data. Runtimes were reported as a secondary performance indicator to assess the computational efficiency and practical feasibility. In real-world applications, especially those involving large data sets or requiring real-time predictions, a model that performs well but takes an excessive amount of time to run may not be feasible. Therefore, reporting runtimes helps in evaluating the trade-off between accuracy and computational cost, ensuring that the model is not only accurate but also efficient and scalable for practical use. Simple baselines (e.g., mean-response and linear regression) were considered as reference points for interpreting model performance; however, quantitative baseline benchmarking was beyond the scope of this comparative ML-focused study.

3. RESULTS AND DISCUSSION

3.1. Correlation of Input Parameters with Effluent BOD

The correlation analysis revealed the relationships between operational parameters and effluent BOD, with correlation coefficients quantifying the strength and direction of these associations (Table 1). Most notably, influent BOD demonstrated the highest correlation with output (coefficient of 0.65), indicating that influent organic loading is the primary driver of treatment performance. This finding aligns with previous studies in pulp and paper mill treatment systems, which have similarly identified influent BOD as a critical control parameter for predicting treatment performance and effluent quality.^{12,68} Interestingly, retention time in the biological treatment process showed a strong negative correlation (-0.57) with effluent BOD. This inverse relationship is consistent with fundamental biological treatment principles and supports previous findings demonstrating that sufficient contact time is essential for effective organic matter removal.^{69,70} Aeration rate and TSS demonstrated moderate correlations (0.42 and 0.38, respectively), suggesting their role as secondary but important control parameters. The correlation with aeration rate aligns with research showing that oxygen transfer is often a rate-limiting factor in biological treatment systems.^{71,72} The moderate positive correlation of TSS with effluent BOD suggests that increased suspended solids can be associated with organic matter accumulation,

Table 1. Correlation Analysis of Input Parameters with Effluent BOD and Process Control Implications^{ab}

Input Feature	Correlation value	Implications for Process Control
Flow Rate (m ³ /h)	0.18	Higher flow rates reduce treatment effectiveness by limiting residence time.
TSS (mg/L)	0.38	Increased suspended solids burden treatment capacity.
Influent BOD (mg/L)	0.65	Primary driver of effluent quality, requiring careful load management.
Temperature (°C)	0.16	Moderate temperature dependence suggests robust biological activity.
Aeration Rate (kg O ₂ /h)	0.42	Critical for maintaining aerobic treatment processes.
Retention Time (h)	−0.57	Longer retention enables more complete treatment.
Supplemental Chemical Addition (kg)	−0.39	Nutrient dosing provides supplementary treatment support.

^aThis table quantifies the relationships between seven key operational parameters and final effluent BOD in the wastewater treatment system. ^bThe correlation values indicate the strength and direction of each relationship, while the right column summarizes the practical implications for process control and optimization.

potentially reflecting the complex interactions between particulate and dissolved components in biological treatment systems.¹² Supplemental nutrient addition showed a moderate negative correlation (−0.39), indicating that increased nutrient usage generally improves treatment efficiency. However, the relatively modest correlation suggests that nutrient addition plays a supporting rather than primary role in BOD removal, consistent with typical hierarchical treatment strategies in pulp and paper facilities. Flow rate and temperature exhibited weaker correlations (0.18 and 0.16, respectively), suggesting that the treatment system maintains relatively stable performance across normal operating ranges of these parameters. This operational stability is particularly valuable for pulp and paper facilities, which often experience variations in production schedules and seasonal temperature changes.

Overall, these statistical outcomes provide a clear foundation for deriving actionable operational insights. These findings provide operationally relevant insights that can inform treatment plant decision-making, particularly when used alongside detailed process knowledge and control constraints. Operators should prioritize careful management of influent BOD loads through production planning and pretreatment strategies, while also ensuring adequate retention times through proper flow control and basin management. Additionally, attention should be given to optimizing aeration systems while balancing energy consumption needs. The strategic use of chemical additives should be considered as a supplementary treatment measure to enhance overall system performance. These operational priorities align with established best practices in pulp and paper mill wastewater treatment and provide a framework for process optimization based on the relative importance of each parameter. These results also highlight the value of retrospective benchmarks as a reference point, i.e., without such baselines, it is difficult to judge the added benefit of lag-only configurations or future online surrogates.

3.2. Model Performance Comparison

In this comprehensive evaluation of ML approaches for predicting wastewater treatment processes in pulp and paper

mills, we systematically assessed five advanced models. The comparison aimed to identify the most effective predictive tools while considering both accuracy and practical implementation considerations (Table S4 and Table 2).

Table 2. Comparative Performance Metrics of Five Machine Learning Models Evaluated on the Independent Test Dataset^a

Model	R ²	NSE	RMSE (mg/L)	Average Runtime (s)
SVM	0.88	0.86	5.0	35
ANN	0.85	0.83	6.2	27
GP	0.82	0.81	7.1	40
DT	0.79	0.79	10.5	46
RF	0.76	0.75	14.4	49

^aThis table presents the evaluation metrics (R², NSE, RMSE) and computational runtime for each of the five models evaluated on the independent test dataset, providing a quantitative basis for comparing predictive accuracy and computational efficiency across different modeling approaches.

Among the evaluated models, the SVM with an RBF kernel demonstrated the strongest predictive performance on the independent test data set, achieving the highest R² (0.88), NSE (0.86), and the lowest RMSE (5 mg/L), with a runtime of 35 s (Table 2). This exceptional performance can be attributed to SVM's ability to handle complex nonlinear relationships inherent in wastewater treatment data. The superior performance of SVM aligns with previous studies in wastewater treatment modeling, which have similarly found SVM effective for capturing the complex interactions between operational parameters.^{73,74} Specifically, municipal wastewater studies report SVR performance around R² = 0.81.⁷⁵ With industrial, process-specific covariates, the present model attains R² = 0.88, indicating measurable gains from tailored inputs. In industrial settings, hybrid LSTM-XGBoost architectures show strong multi-indicator performance, while the corresponding standalone SVM baseline remains slightly below our accuracy.⁷⁶

The ANN model emerged as the second-best performing model, utilizing a multilayer perception (MLP) architecture with three hidden layers containing 47, 29, and 13 neurons, respectively. This architecture, optimized through extensive testing, employs ReLU activation functions for hidden layers and sigmoid functions for the output layer. The model achieved an R² = 0.85, NSE = 0.83, and RMSE = 6.2 mg/L. Notably, the ANN's execution time of 27 s was the fastest among all models, suggesting its potential for real-time applications. Reported ANN results for wastewater quality prediction often lie between R² = 0.80 and 0.84.⁷⁷ By incorporating process-specific covariates and a purpose-built MLP for pulp-and-paper conditions, the present study achieves R² = 0.85. This aligns with evidence that multilayer feed-forward networks are highly sensitive to input selection and domain-specific hyperparameterization.⁷⁸ Together, these comparisons emphasize that tailoring ANN structure and inputs to the operating environment enhances predictive accuracy and real-time applicability.

The GP model demonstrated robust but slightly lower performance metrics, with R² = 0.82, NSE = 0.81, and RMSE = 7.1 mg/L. This evolutionary approach utilized a population size of 98, mutation rate of 0.053, crossover rate of 0.72, and maximum tree depth of 5.3 levels. While GP required longer computational time (40 s), it offers the unique advantage of

producing interpretable mathematical expressions, which can provide insights into the underlying relationships in the data. GP-based effluent prediction in industrial systems is typically reported at $R^2 = 0.78$ – 0.80 ;⁷⁹ the present GP model attains $R^2 = 0.82$, likely reflecting tailored evolutionary settings and process-specific pulp-and-paper data. Work using multi-objective genetic algorithms highlights the interpretability and compute trade-off in wastewater control,⁸⁰ reinforcing GP's transparency and the gains from domain-informed tuning.

DT and RF models showed more modest performance levels. The DT model, configured with a tree depth limit of 9.6 and minimum leaf node samples of 4.7, achieved $R^2 = 0.79$, $NSE = 0.79$ and $RMSE = 10.5$ mg/L. Despite lower accuracy, the DT model offers high interpretability, making it valuable for process understanding. RF, utilizing 97 trees and a maximum depth of 10.2, demonstrated the lowest performance metrics ($R^2 = 0.76$, $NSE = 0.75$, and $RMSE = 14.4$ mg/L) but maintains value in its ability to handle outliers and missing data. Prior DT/RF studies on wastewater effluent prediction report MAEs 4.8 mg/L (DT, TSS) and 5.1 mg/L (RF, COD), and RF test-set R^2 around 0.72 and 0.58, i.e., below our RF $R^2 = 0.76$.^{42,81} Gains likely reflect process-specific features from pulp-and-paper operations and the relatively large data set, yielding better accuracy while preserving interpretability.

To move beyond aggregate scores, error distributions were compiled by absolute percent-error bins (see Table S5). Within $\pm 10\%$ error, SVM achieved 65.3% of predictions, compared with 59.1% for ANN, 59.2% for GP, 54.0% for DT, and 50.9% for RF. The share of $>20\%$ errors was lowest for SVM (7.9%) and higher for DT (15.9%) and RF (16.9%), highlighting differences in reliability relevant to operations. These comparative results suggest that while SVM provides the highest accuracy for effluent BOD prediction in pulp and paper mill wastewater treatment, each model offers distinct advantages that may be valuable in different operational contexts. The choice of model may ultimately depend on specific facility needs, balancing prediction accuracy with computational efficiency and interpretability requirements.

3.3. Model Testing and Validation

The testing phase provided critical insights into model generalization capabilities and reliability for real-world implementation. Testing on independent operational data provides a direct readout of generalization and reliability. Across all models, errors rose at the tails of the effluent BOD distribution (<5 mg/L and >25 mg/L) where observations comprised $<8\%$ of samples (Figure S5). This pattern is consistent with data imbalance during training and the difficulty of capturing abrupt upsets and rare operating regimes, yielding weaker performance in these edge regions. As shown in Table 3, the relative performance rankings observed across models are largely preserved on the independent test data set, though with some notable differences in the magnitude of performance metrics. The SVM model maintained its superior position during testing, achieving $R^2 = 0.85$, $NSE = 0.82$, and $RMSE = 5.5$ mg/L, with an execution time of 8.75 s. The limited reduction observed on the independent test data set indicates strong generalization capabilities. Overall, SVM's strong performance here likely reflects the data we used. First, the data set is moderate in size with a relatively high feature-to-sample ratio. i.e., conditions under which SVMs typically generalize well in small-to-medium sample regimes. Second, the operational variables

Table 3. Comparative Performance Metrics of Five Machine Learning Models Evaluated on the Independent 20% Testing Dataset^{ab}

Model	R^2	NSE	RMSE (mg/L)	Average Runtime (s)	Coverage (%)	Width (mg/L)
SVM	0.85	0.82	5.5	87	91.2	9.3
ANN	0.82	0.79	6.8	67	88.4	11.1
GP	0.79	0.76	7.5	102	87.1	12.5
DT	0.75	0.72	11.0	116	85.3	15.8
RF	0.72	0.68	15.0	122	83.7	18.2

^aThis table presents the evaluation metrics (R^2 , NSE, RMSE) and computational runtime, and uncertainty metrics (coverage and interval width for 90% prediction intervals). ^bThese metrics assess model generalization capabilities and performance when applied to previously unseen operational data.

show nonlinear relationships and differing scales; after normalization, an RBF kernel is well suited to capture these multidimensional patterns. Finally, noise levels are modest and predictors are only moderately correlated, allowing the maximum-margin formulation to learn a stable function without overfitting. These findings align with previous wastewater treatment modeling studies that have demonstrated SVM's robustness when dealing with new operational data.^{43,73,82} It is emphasized that these testing results reflect retrospective benchmarking performance, as influent BOD₅ is a laboratory measurement with an inherent analytical delay. For context, simple baseline approaches such as mean-value prediction or linear regression generally exhibit substantially lower predictive capability than the ML models evaluated here, reinforcing the value of nonlinear learning approaches for effluent BOD prediction.

The ANN model achieved $R^2 = 0.82$, $NSE = 0.79$, and $RMSE = 6.8$ mg/L, representing a 3.4% decrease in R^2 and a 3.5% decrease in NSE compared to the SVM model. While ANN models can overfit on smaller data sets, the strong generalization observed here can be attributed to the combination of optimal hyperparameters and the use of a robust data set with diverse operating conditions. The successful generalization can be attributed to the optimized three-layer architecture and careful selection of the Adam optimizer.

GP demonstrated interesting characteristics during testing, with $R^2 = 0.79$, $NSE = 0.76$, and $RMSE = 7.5$ mg/L. This reflects a 2.5% drop in R^2 , 2.6% drop in NSE, and a 5.3% increase in RMSE compared to the training phase. While this degradation is more than what was observed for SVM and ANN, it remained within acceptable limits.⁸³ The GP model's ability to maintain reasonable accuracy while providing interpretable mathematical expressions represents a valuable trade-off for process understanding and optimization.

DT and RF models showed the largest discrepancy between training and testing performance. The DT model achieved testing metrics of $R^2 = 0.75$, $NSE = 0.72$, and $RMSE = 11.0$ mg/L, while RF testing results showed $R^2 = 0.72$, $NSE = 0.68$, and $RMSE = 15.0$ mg/L. This more substantial degradation in performance (approximately 15–20% decline in R^2 values) suggests these models may be more susceptible to overfitting despite their inherent advantages for handling nonlinear relationships. Similar challenges with tree-based methods have been noted in other wastewater treatment studies.^{84,85}

A critical consideration in model validation is the consistency of performance across different operational ranges.

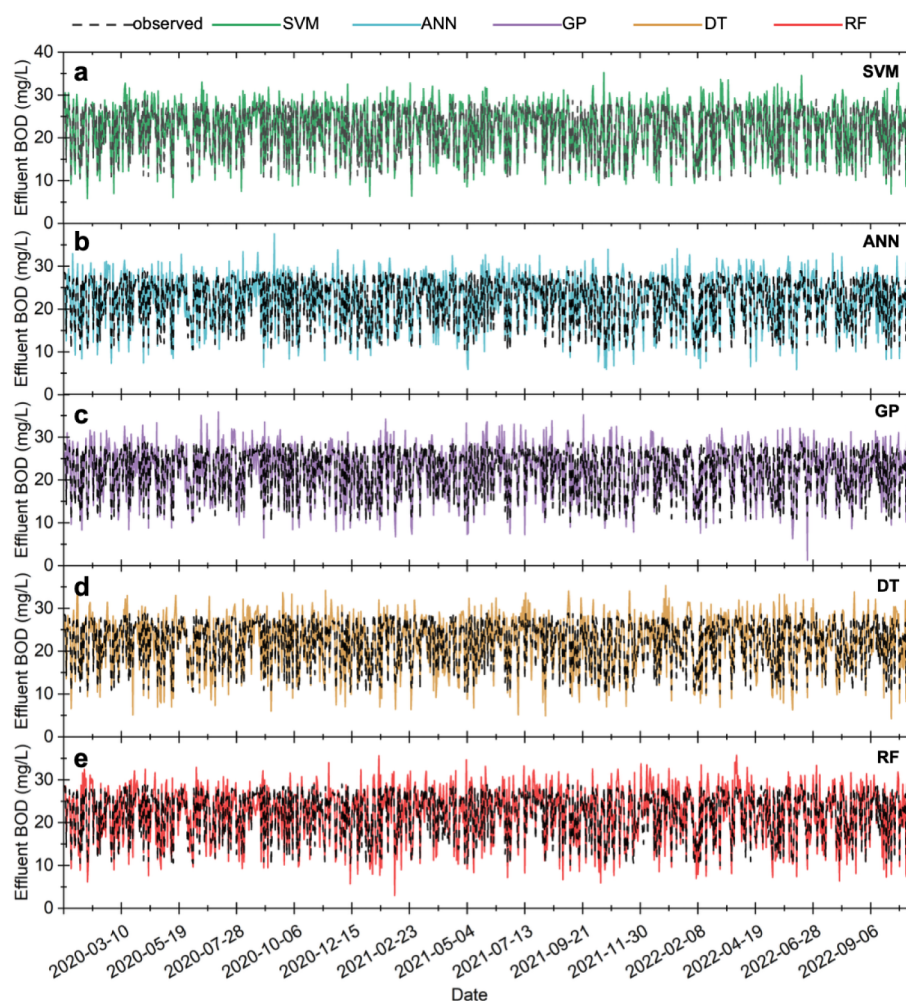


Figure 2. Temporal validation of machine learning models for effluent BOD prediction. Observed values (black dashed line) compared with predictions from (a) support vector machine (SVM), (b) artificial neural network (ANN), (c) genetic programming (GP), (d) decision tree (DT), and (e) random forest (RF). Data is shown across the testing period (January 2020–October 2022). Here, the temporal dimension is used strictly for visualization and comparative assessment of prediction behavior, rather than implying time-series model training or sequential dependence.

The testing results revealed that all models showed some degradation in prediction accuracy at extreme BOD values (specifically below 30 mg/L and above 250 mg/L). However, SVM and ANN maintained better reliability in these ranges, with less than a 10% reduction in R^2 compared to their performance in the midrange BOD values (70–200 mg/L). This behavior is particularly important for pulp and paper mill applications, where treatment systems must handle varying industrial loads and occasional extreme events. The computational efficiency during the testing phase also provides important insights for real-time applications. Although testing is typically performed on a smaller data set (20% of the original), the execution times still reflect the models' relative efficiency in generating predictions. For instance, the ANN model demonstrated notable computational speed with an average runtime of 6.75 s. This suggests that once trained, such models could support near-real-time decision-support once adapted to deployable input configurations. However, the trade-off between predictive accuracy and computational speed must be considered based on the specific operational demands of each facility. These retrospective models provide the essential baseline against which deployable variants (using SCADA variables or lag features) and future surrogate-based configurations can be evaluated.

3.4. Effluent BOD Prediction Performance

To provide a comprehensive evaluation of model performance, we employed both temporal and scatter plot analyses. Temporal analysis captured the models' ability to track BOD variations over time, while scatter plots visualized the correlation between observed and predicted values. This dual approach allowed for the assessment of both dynamic tracking capability and statistical agreement, operationally relevant performance characteristics. Temporal and scatter plot analyses were used to assess the predictive performance of the five ML models under real operational conditions. Our analysis begins with a full-period time-series evaluation (Figure 2), followed by a focused examination of a representative 100-day operational window and its corresponding correlation trends (Figure 3). A representative 100-day window was selected from the testing period to allow for a closer inspection of model performance under variable operational conditions. This window was chosen based on its inclusion of a broad range of influent BOD concentrations, flow variability, and broader operational variability over the monitoring period to ensure it captured typical and extreme scenarios encountered in pulp and paper mill wastewater treatment. Here, the temporal dimension is used strictly for visualization and comparative assessment of

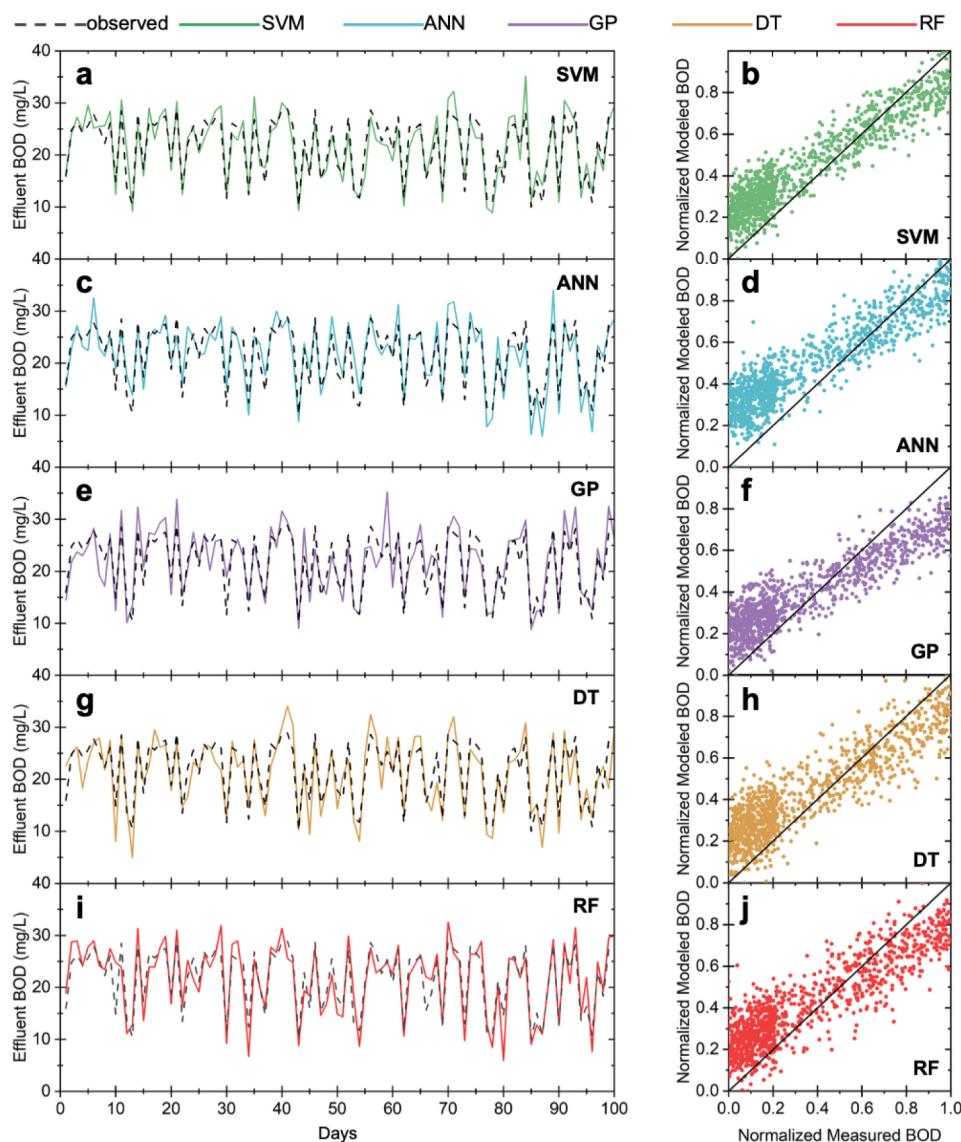


Figure 3. Comparison of observed and predicted effluent BOD in pulp and paper mill wastewater treatment using five machine learning models. Left panels (a, c, e, g, i) display temporal dynamics of effluent BOD (mg/L) over a 100-day testing period, comparing observed values (black dashed line) with predictions from (a) support vector machine (SVM, green), (c) artificial neural network (ANN, blue), (e) genetic programming (GP, purple), (g) decision tree (DT, orange), and (i) random forest (RF, red). Right panels (b, d, f, h, j) present corresponding scatter plots of normalized predicted versus measured BOD values with 1:1 reference lines for: (b) SVM, (d) ANN, (f) GP, (h) DT, and (j) RF. These scatter plot include the full data set. Here, the temporal dimension is used strictly for visualization and comparative assessment of prediction behavior, rather than implying time-series model training or sequential dependence.

prediction behavior, rather than implying time-series model training or sequential dependence.

Figure 2 illustrates the comparison between observed effluent BOD concentrations (black dashed line) and model predictions (colored lines) across the monitoring period from January 2020 to October 2022. In Figure 2a, the SVM model (green line) shows the closest overall tracking of the observed BOD pattern, with generally good alignment throughout the range of BOD values (10–30 mg/L). Figure 2b reveals that the ANN model (blue line) follows the general trend but exhibits more pronounced deviations at certain points, with occasional prediction spikes that extend beyond the observed values. The GP model in Figure 2c (purple line) demonstrates similar tracking characteristics to ANN but with its own distinct pattern of deviations from the observed data. Most notably, Figure 2d shows that the DT model (orange line) produces a

distinctly different prediction pattern with characteristic stepwise plateaus, where predicted values remain constant across ranges of time despite continuous changes in observed values. Similarly, Figure 2e displays the RF model predictions (red line), which also exhibit step-like behavior but with different plateau patterns than DT, generally capturing the central trend while showing reduced sensitivity to the full amplitude of the BOD fluctuations.

For more detailed analysis, Figure 3 provides a focused examination of model performance during a representative 100-day period (left panels) alongside correlation scatter plots (right panels). The left panels (a, c, e, g, i) display temporal dynamics over a selected 100-day window, while the right panels (b, d, f, h, j) show normalized scatter plots representing the statistical relationship between predicted and measured BOD values across the entire testing data set. The 100-day

period, from August 15 to November 22, 2021, was selected as a representative testing window. This period included moderate to high variability in influent BOD concentrations and spanned typical transitional seasonal shifts, capturing both operational stability and dynamic shifts. This allowed for clearer visual comparison between model predictions and observed patterns.

In Figure 3a, the SVM model (green line) tracks the observed values (black dashed line) with high fidelity, capturing both the magnitude and timing of BOD fluctuations throughout the 100-day period. The corresponding scatter plot (Figure 3b) shows the tightest clustering around the 1:1 reference line among all models, particularly in the midrange values (0.3–0.7 normalized). At extreme values, a slight pattern of underprediction at higher BOD levels and overprediction at lower levels is visible. The ANN model (Figure 3c) demonstrates good overall tracking but shows more pronounced deviations, particularly during rapid BOD transitions visible around days 10–20 and 65–75. Its scatter plot (Figure 3d) reveals a wider distribution of points compared to SVM, with a noticeable tendency toward overprediction at lower BOD values (<0.3 normalized) and underprediction at higher values (>0.7 normalized). Figure 3e shows the GP model's temporal predictions, which capture the general pattern but exhibit increased deviations during the sharp BOD fluctuations around days 40–50, where predicted values appear to lag behind observed fluctuations. The corresponding scatter plot (Figure 3f) displays more dispersion than SVM and ANN, with greater variance at both ends of the normalized BOD range. The DT model results (Figure 3g) exhibit a distinctive stepwise prediction pattern, characterized by horizontal plateaus where predicted values remain constant across ranges of time. This discretization is particularly evident around days 30–40 and 60–70. Its scatter plot (Figure 3h) confirms this behavior with distinct horizontal clustering patterns, creating a banded appearance that reflects the model's binary decision structure. Finally, the RF model (Figure 3i) shows similar stepwise characteristics to the DT model but with somewhat smoother transitions. The RF scatter plot (Figure 3j) displays the most pronounced horizontal banding among all models, indicating the model's tendency to produce categorical rather than continuous predictions across the entire range of observed values. These temporal and statistical patterns form the basis for evaluating how different model families behave under real-world variability, enabling direct practical comparisons.

Comparing all panels reveals that SVM and ANN models (panels a–d) provide smoother prediction profiles with higher accuracy. In contrast, tree-based models (panels g–j) generate more discretized outputs that may better serve threshold-based decision making and operational control systems in treatment facilities.⁵⁹ The GP model (panels e–f) represents a middle ground in both temporal tracking and statistical correlation, offering a balance between interpretability and prediction accuracy in water quality applications.⁵⁵ All models demonstrate better performance in the midrange of BOD values, with accuracy decreasing at the extremes of the measurement range, a common challenge in wastewater quality prediction attributed to both the relative scarcity of extreme events in training data sets and the inherent variability of biological treatment processes under nonstandard loading conditions.^{74,85}

To further assess model reliability beyond point-prediction accuracy, prediction interval performance was evaluated using coverage and width metrics. Coverage reflects the calibration and credibility of the prediction intervals, defined as the proportion of observed values falling within the nominal confidence bounds, while width represents predictive precision, with narrower intervals indicating lower uncertainty and greater operational usefulness.

In the context of water quality monitoring, where both reliability and actionable uncertainty bounds are essential for alarm setting and compliance decisions, joint consideration of these two metrics enables a more comprehensive evaluation. Analysis of the 90% prediction intervals showed that the SVM model achieved the most favorable uncertainty characterization, with a coverage of 91.2% (close to the nominal target) and an average interval width of 9.3 mg/L. The ANN model ranked second, with a coverage of 88.4% and a wider interval of 11.1 mg/L, indicating acceptable but reduced calibration. In contrast, the GP model exhibited moderate performance (coverage: 87.1%; width: 12.5 mg/L), while the DT and RF models demonstrated the weakest uncertainty estimates, with lower coverage (85.3% and 83.7%, respectively) and substantially wider intervals (15.8 and 18.2 mg/L). These wider intervals indicate reduced practical usefulness, as higher uncertainty is required to maintain nominal coverage.

The results indicate that models exhibiting stronger point-prediction accuracy (notably SVM and ANN) also provide more reliable and informative uncertainty estimates. This dual strength enhances their suitability for real-world decision-making in wastewater treatment operations, where both prediction accuracy and confidence bounds are critical for managing effluent quality and operational risk.

3.5. Industrial and Environmental Implications

This study investigated the application of ML models for optimizing wastewater treatment in pulp and paper mills, with the SVM model demonstrating superior performance on the independent test data set. This study addresses a key gap in the literature by providing a systematic long-horizon comparison of multiple ML algorithms under full-scale pulp-and-paper operating conditions, offering practical guidance for model selection based on accuracy, interpretability, and uncertainty quantification. While SVM achieved the highest accuracy with $R^2 = 0.85$ during testing, all models showed some limitations at extreme BOD values, suggesting opportunities for further refinement through additional variables and time-lag consideration.⁵⁸ Key limitations of this study include reliance on data from a single European facility, use of retrospective laboratory BOD measurements rather than real-time inputs, and the absence of k-fold cross-validation. Thus, future work should incorporate resampling-based validation to further verify robustness across splits and operating conditions. Taken together, these results demonstrate how the evaluated ML models, beyond predictive accuracy, can support practical, real-life decision-making in pulp and paper wastewater treatment by informing operational control, resource optimization, and long-term sustainability planning. Recent developments in computational approaches include hybrid models that combine mechanistic understanding with data-driven methods, addressing the limitations observed in purely black-box approaches. These hybrid frameworks integrate process knowledge with ML to improve prediction accuracy at extreme operational conditions while maintaining interpretability.^{21,31} Deep learn-

ing architectures, particularly long short-term memory (LSTM) networks, offer another promising direction for handling temporal dependencies in treatment processes, though with higher computational requirements than the models evaluated in this study.^{86,87}

The implementation of these ML models facilitates circular economy principles in pulp and paper manufacturing through several concrete applications. First, they enable predictive water quality management through advanced insight into effluent quality trends, which can support decision-making related to chemical dosing and aeration planning while maintaining effluent compliance.²⁸ Second, by providing predictive insight into effluent quality trends, these models can support water reuse decision framework.⁸⁸ Third, they facilitate precision resource recovery by predicting optimal conditions for extracting valuable components from wastewater streams, creating opportunities for materials recovery within the circular production cycle.¹² Overall, these predictive tools support the industry's progress toward zero-discharge systems by providing continuous monitoring and adaptive control of treatment processes, helping facilities meet increasingly stringent environmental regulations while maximizing water recycling throughout the manufacturing process.⁸⁹ By benchmarking multiple ML models against an extended 34-month, full-scale data set, this work adds to the limited body of industrial wastewater applications and demonstrates the value of long-term operational data for robust, comparative evaluation. These applications are discussed as decision-support use cases rather than implemented control actions, consistent with the retrospective benchmarking scope of the present study.

In practice, predictions can inform feed-forward chemical dosing, aeration set-point tuning, and reuse-routing when forecast BOD approaches permit thresholds. These predictive models provide a foundation for future deployable soft-sensor development, i.e., virtual instruments that infer effluent BOD from real-time operational data such as flow rate, TSS, temperature, and aeration (Figure S6). When adapted to SCADA-only or surrogate-enhanced inputs, these soft-sensor frameworks could enable feed-forward control strategies, allowing operators to proactively adjust aeration and chemical dosing based on predicted effluent quality. This approach reduces reliance on delayed lab measurements and enhances responsiveness to dynamic loading conditions. Similar soft sensor frameworks have been successfully applied in potable reuse systems to predict NDMA concentrations and optimize UV dosing,⁹⁰ in cooling tower operations to manage water chemistry,⁹¹ and in multimodal process monitoring to reduce sensor costs while maintaining accuracy.⁹² Integrating ML-based soft sensors into industrial water treatment systems offers a scalable pathway toward real-time compliance assurance and circular water management. For robust deployment, operators should use routine QA/QC and drift checks, periodic retraining on rolling windows, and report prediction intervals alongside statistical metrics. Key challenges (e.g., sensor gaps/drift, concept drift, and transferability across mills) can be mitigated via validation pipelines and site-specific fine-tuning. While SVM and ANN achieved the highest overall accuracy in this study, all models exhibited some degradation at extreme BOD values. To address this, future work could evaluate hybrid or ensemble strategies that combine complementary learners (e.g., a stacked SVM-ANN ensemble or mixture-of-experts that weights models by operating regime).

Prior reviews and applications in wastewater/process modeling indicate that hybridization and assembling can improve robustness and tail performance under variable operating conditions.^{21,31,59} For time-dependent effects, coupling with sequence models (e.g., LSTM or CNN-LSTM) has also been shown to reduce extreme-error rates in effluent prediction tasks.⁸⁶ Practical implementations could train an SVM tuned for midrange accuracy and an ANN regularized for rare/extreme regimes, then aggregate predictions via weighted averaging or a meta-learner; alternative designs include regime-switching ensembles triggered by influent load or retention time. Evaluation should incorporate quantile-aware losses and threshold-specific metrics to reflect compliance-relevant extremes.

Future research should explore model applications across diverse industrial sectors with complex wastewater characteristics, including textile manufacturing and chemical processing. The integration of ML models with real-time monitoring systems and Internet of Things technologies represents a promising direction for enhancing operational decision-making and circular economy.^{93,94} Additionally, further investigation into hybrid modeling approaches combining machine learning with physical or mechanistic models could address the prediction challenges at extreme operational conditions identified in this study.⁴³ Urgent technical gaps requiring attention include developing robust methods for handling extreme-event prediction, establishing transfer learning approaches for cross-facility deployment, and integrating online surrogate sensors to enable real-time prediction. By demonstrating strong retrospective performance with laboratory inputs, this study establishes a benchmark foundation upon which future deployable soft sensors can be developed. Future progress will come from adapting these frameworks to SCADA-only or surrogate-enhanced inputs, but such advances must be built on the baseline demonstrated here.

■ ASSOCIATED CONTENT

■ Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acseengineeringau.5c00088>.

Operational data set from the pulp and paper facility wastewater treatment system; descriptive statistics of wastewater treatment parameters; binned distribution of absolute percent error for test-set BOD predictions; hyperparameter grids and selected optima for each model; schematic structure of support vector machines (SVM); schematic of the components of a single-layer artificial neural networks (ANN); representation of mathematical relations in genetic programming (GP); schematic of the structure of a simple decision tree; signed percent error as a function of observed BOD; conceptual soft-sensor workflow for effluent BOD prediction and process control; MATLAB code for machine learning model development and validation; detailed methods on specific model development (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Lewis S. Rowles – Department of Civil Engineering and Construction, Georgia Southern University, Statesboro, Georgia 30458, United States; orcid.org/0000-0002-

3489-9179; Phone: (912) 478-0772; Email: lrowles@georgiasouthern.edu

Authors

Fatemeh Naeijian – Department of Civil Engineering and Construction, Georgia Southern University, Statesboro, Georgia 30458, United States

Ali Arefinia – Department of Irrigation and Reclamation Engineering College of Agriculture and Natural Resources, University of Tehran, Tehran 77871-31587, Iran

Mohammed Tamim Zaki – Energy Analysis and Environmental Impacts Division, Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States

Kevin D. Orner – Wadsworth Department of Civil and Environmental Engineering, West Virginia University, Morgantown, West Virginia 26506, United States;

orcid.org/0000-0001-8090-7603

Mohammadrosoush Tafazzoli – Department of Civil Engineering and Construction, Georgia Southern University, Statesboro, Georgia 30458, United States

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acsengineeringau.5c00088>

Author Contributions

CRedit: **Fatemeh Naeijian** conceptualization, formal analysis, investigation, writing - original draft; **Ali Arefinia** investigation, methodology, writing - original draft; **Mohammed Tamim Zaki** investigation, methodology, writing - review & editing; **Kevin D Orner** investigation, methodology, writing - review & editing; **Mohammadrosoush Tafazzoli** visualization, writing - review & editing; **Lewis S. Rowles** conceptualization, funding acquisition, investigation, project administration, supervision, visualization, writing - review & editing.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to Georgia Southern University for supporting this research through the Presidential Fellowship awarded to Fatemeh Naeijian. The authors also acknowledge the Department of Civil Engineering and Construction at Georgia Southern University for providing the necessary infrastructure and academic environment to conduct this research.

REFERENCES

- (1) Esmaeeli, A.; Sarrafzadeh, M.-H.; Zeighami, S.; Kalantar, M.; Bariki, S. G.; Fallahi, A.; Asgharnejad, H.; Ghaffari, S.-B. A Comprehensive Review on Pulp and Paper Industries Wastewater Treatment Advances. *Ind. Eng. Chem. Res.* **2023**, 62 (21), 8119–8145.
- (2) Kamali, M.; Khodaparast, Z. Review on Recent Developments on Pulp and Paper Mill Wastewater Treatment. *Ecotoxicol. Environ. Saf.* **2015**, 114, 326–342.
- (3) Singh, U. S.; Tripathi, Y. C. Characteristics and Treatment of Pulp and Paper Mill Effluents – A Review. *Int. J. Eng. Tech. Res.* **2020**, 10 (11), 454–4698.
- (4) Thompson, G.; Swain, J.; Kay, M.; Forster, C. F. The Treatment of Pulp and Paper Mill Effluent: A Review. *Bioresour. Technol.* **2001**, 77 (3), 275–286.
- (5) Bajpai, P. *Pulp and Paper Industry: Emerging Waste Water Treatment Technologies*; Elsevier: Amsterdam, Netherlands; Oxford, England; Cambridge, MA, 2017.

- (6) Borzooei, S.; Amerlinck, Y.; Panepinto, D.; Abolfathi, S.; Nopens, I.; Scibilia, G.; Meucci, L.; Zanetti, M. C. Energy Optimization of a Wastewater Treatment Plant Based on Energy Audit Data: Small Investment with High Return. *Environ. Sci. Pollut. Res.* **2020**, 27 (15), 17972–17985.
- (7) Borzooei, S.; Miranda, G. H. B.; Abolfathi, S.; Scibilia, G.; Meucci, L.; Zanetti, M. C. Application of Unsupervised Learning and Process Simulation for Energy Optimization of a WWTP under Various Weather Conditions. *Water Sci. Technol.* **2020**, 81 (8), 1541–1551.
- (8) Christenson, L.; Sims, R. Production and Harvesting of Microalgae for Wastewater Treatment, Biofuels, and Bioproducts. *Biotechnol. Adv.* **2011**, 29 (6), 686–702.
- (9) Borzooei, S.; Amerlinck, Y.; Abolfathi, S.; Panepinto, D.; Nopens, I.; Lorenzi, E.; Meucci, L.; Zanetti, M. C. Data Scarcity in Modelling and Simulation of a Large-Scale WWTP: Stop Sign or a Challenge. *J. Water Process Eng.* **2019**, 28, 10–20.
- (10) Borzooei, S.; Teegavarapu, R.; Abolfathi, S.; Amerlinck, Y.; Nopens, I.; Zanetti, M. C. Impact Evaluation of Wet-Weather Events on Influent Flow and Loadings of a Water Resource Recovery Facility. In *New Trends in Urban Drainage Modelling: Green Energy and Technology*, Mannina, G. Ed.; Springer International Publishing: Cham, 2019; pp. 706–711.
- (11) Achoka, J. D. The Efficiency of Oxidation Ponds at the Kraft Pulp and Paper Mill at Webuye in Kenya. *Water Res.* **2002**, 36 (5), 1203–1212.
- (12) Ashrafi, O.; Yerushalmi, L.; Haghighat, F. Wastewater Treatment in the Pulp-and-Paper Industry: A Review of Treatment Processes and the Associated Greenhouse Gas Emission. *J. Environ. Manage.* **2015**, 158, 146–157.
- (13) Borsetto, C.; Dykes, C.; Kockiri, B.; Song, L.; Wellington, E. M.; Abolfathi, S. Constructed Wetlands as Nature-Based Barriers: Mitigating Antimicrobial Resistance and Pathogen Dispersal in Riverine Systems. *J. Hazard. Mater.* **2025**, 495, 138855.
- (14) Dykes, C.; Pearson, J.; Bending, G. D.; Abolfathi, S. Hydraulic Efficiency and Mixing Dynamics in Surface Flow Constructed Wetlands: Influence of Design, Vegetation Phenology, and Climate Variability. *Water Res.* **2025**, 285, 124110.
- (15) Pokhrel, D.; Viraraghavan, T. Treatment of Pulp and Paper Mill Wastewater—a Review. *Sci. Total Environ.* **2004**, 333 (1), 37–58.
- (16) Sangamnere, R.; Misra, T.; Bherwani, H.; Kapley, A.; Kumar, R. A Critical Review of Conventional and Emerging Wastewater Treatment Technologies. *Sustain. Water Resour. Manag.* **2023**, 9 (2), 58.
- (17) Söderholm, P.; Bergquist, A.-K.; Söderholm, K. Environmental Regulation in the Pulp and Paper Industry: Impacts and Challenges. *Curr. For. Rep.* **2019**, 5 (4), 185–198.
- (18) Mohammadpour, A.; Gharehchahi, E.; Golaki, M.; Gharaghani, M. A.; Ahmadian, F.; Abolfathi, S.; Samaei, M. R.; Uddin, M. G.; Olbert, A. I.; Mousavi Khaneghah, A. Advanced Water Quality Assessment Using Machine Learning: Source Identification and Probabilistic Health Risk Analysis. *Res. Eng.* **2025**, 27, 105421.
- (19) Mohammadpour, A.; Gharehchahi, E.; Gharaghani, M. A.; Shahsavani, E.; Golaki, M.; Berndtsson, R.; Khaneghah, A. M.; Hashemi, H.; Abolfathi, S. Assessment of Drinking Water Quality and Identifying Pollution Sources in a Chromite Mining Region. *J. Hazard. Mater.* **2024**, 480, 136050.
- (20) Khosravi, K.; Farooque, A. A.; Karbasi, M.; Ali, M.; Heddam, S.; Faghfour, A.; Abolfathi, S. Enhanced Water Quality Prediction Model Using Advanced Hybridized Resampling Alternating Tree-Based and Deep Learning Algorithms. *Environ. Sci. Pollut. Res.* **2025**, 32 (11), 6405–6424.
- (21) Corominas, L.; Garrido-Baserba, M.; Villez, K.; Olsson, G.; Cortés, U.; Poch, M. Transforming Data into Knowledge for Improved Wastewater Treatment Operation: A Critical Review of Techniques. *Environ. Model. Softw.* **2018**, 106, 89–103.
- (22) Newhart, K. B.; Holloway, R. W.; Hering, A. S.; Cath, T. Y. Data-Driven Performance Analyses of Wastewater Treatment Plants: A Review. *Water Res.* **2019**, 157, 498–513.

- (23) Zaki, M. T.; Rowles, L. S.; Adjero, D. A.; Orner, K. D. A Critical Review of Data Science Applications in Resource Recovery and Carbon Capture from Organic Waste. *ACS EST Eng* **2023**, 3 (10), 1424–1467.
- (24) Zaki, M. T.; Rowles, L. S.; Hallowell, J.; Orner, K. D. A Data-Driven Framework to Inform Sustainable Management of Animal Manure in Rural Agricultural Regions Using Emerging Resource Recovery Technologies. *Clean. Environ. Syst* **2024**, 13, 100188.
- (25) Mozafari, Z.; Noori, R.; Bateni, S. M.; Jun, C.; Kim, D.; Saravani, M. J.; Naderian, D.; Siadatmousavi, S. M.; Afzalimehr, H.; Azizpour, J.; Sadrinasab, M.; Hosseinzadeh, M.; Kianmehr, P.; Abolfathi, S. Impact of Climatic Factors on Eutrophication in the World's Largest Lake. *Ecol. Indic* **2025**, 175, 113497.
- (26) Moghaddam, H. K.; Rahimzadeh Kivi, Z.; Abtahizadeh, E.; Abolfathi, S. Sustainable Water Allocation under Climate Change: Deep Learning Approaches to Predict Drinking Water Shortages. *J. Environ. Manage* **2025**, 385, 125600.
- (27) Abba, S. I.; Kilinc, H. C.; Tan, M. L.; Demir, V.; Ahmadianfar, I.; Halder, B.; Heddami, S.; Jawad, A. H.; Al-Areeq, A. M.; Yaseen, Z. M. Bio-Communal Wastewater Treatment Plant Real-Time Modeling Using an Intelligent Meta-Heuristic Approach: A Sustainable and Green Ecosystem. *J. Water Process Eng* **2023**, 53, 103731.
- (28) Elsayed, A.; Siam, A.; El-Dakhkhni, W. Machine Learning Classification Algorithms for Inadequate Wastewater Treatment Risk Mitigation. *Process Saf. Environ. Prot* **2022**, 159, 1224–1235.
- (29) Hong, Y.-S.; Bhamidimarri, R. Evolutionary Self-Organising Modelling of a Municipal Wastewater Treatment Plant. *Water Res* **2003**, 37 (6), 1199–1212.
- (30) Malviya, A.; Jaspal, D. Artificial Intelligence as an Upcoming Technology in Wastewater Treatment: A Comprehensive Review. *Environ. Technol. Rev* **2021**, 10 (1), 177–187.
- (31) Nagpal, M.; Siddique, M. A.; Sharma, K.; Sharma, N.; Mittal, A. Optimizing Wastewater Treatment through Artificial Intelligence: Recent Advances and Future Prospects. *Water Sci. Technol* **2024**, 90 (3), 731–757.
- (32) Faizollahzadeh Ardabili, S.; Najafi, B.; Shamshirband, S.; Minaei Bidgoli, B.; Deo, R. C.; Chau, K. Computational Intelligence Approach for Modeling Hydrogen Production: A Review. *Eng. Appl. Comput. Fluid Mech* **2018**, 12 (1), 438–458.
- (33) Arismendy, L.; Cárdenas, C.; Gómez, D.; Maturana, A.; Mejía, R.; Quintero, C. G. Intelligent System for the Predictive Analysis of an Industrial Wastewater Treatment Process. *Sustainability* **2020**, 12 (16), 6348.
- (34) Borzooei, S.; Teegavarapu, R.; Abolfathi, S.; Amerlinck, Y.; Nopens, I.; Zanetti, M. C. Data Mining Application in Assessment of Weather-Based Influent Scenarios for a WWTP: Getting the Most Out of Plant Historical Data. *Water. Air. Soil Pollut* **2019**, 230 (1), 5.
- (35) Mahdian, M.; Noori, R.; Salamattalab, M. M.; Heggy, E.; Bateni, S. M.; Nohegar, A.; Hosseinzadeh, M.; Siadatmousavi, S. M.; Fadaei, M. R.; Abolfathi, S. Anzali Wetland Crisis: Unraveling the Decline of Iran's Ecological Gem. *J. Geophys. Res.: Atmos.* **2024**, 129 (4), No. e2023JD039538.
- (36) Khosravi, K.; Attar, N.; Bateni, S. M.; Jun, C.; Kim, D.; Safari, M. J. S.; Heddami, S.; Farooque, A.; Abolfathi, S. Daily River Flow Simulation Using Ensemble Disjoint Aggregating M5-Prime Model. *Heliyon* **2024**, 10 (20), No. e37965.
- (37) Jafari-Asl, J.; Hashemi Monfared, S. A.; Abolfathi, S. Reducing Water Conveyance Footprint through an Advanced Optimization Framework. *Water* **2024**, 16 (6), 874.
- (38) Habib, M. A.; Abolfathi, S.; O'Sullivan, J. J.; Salaudinn, M. Efficient Data-Driven Machine Learning Models for Scour Depth Predictions at Sloping Sea Defences. *Front. Built Environ* **2024**, 10, 1343398.
- (39) Riazi, M.; Bateni, S. M.; Jun, C.; Farooque, A. A.; Khosravi, K.; Abolfathi, S. Enhancing Rainfall-Runoff Simulation in Data-Poor Watersheds: Integrating Remote Sensing and Hybrid Decomposition for Hydrologic Modelling. *Water Resour. Manage* **2025**, 39, 5529.
- (40) Zhu, J.-J.; Yang, M.; Ren, Z. J. Machine Learning in Environmental Research: Common Pitfalls and Best Practices. *Environ. Sci. Technol* **2023**, 57 (46), 17671–17689.
- (41) Wang, Y.; Cheng, Y.; Liu, H.; Guo, Q.; Dai, C.; Zhao, M.; Liu, D. A Review on Applications of Artificial Intelligence in Wastewater Treatment. *Sustainability* **2023**, 15 (18), 13557.
- (42) Zhou, P.; Li, Z.; Snowling, S.; Baetz, B. W.; Na, D.; Boyd, G. A Random Forest Model for Inflow Prediction at Wastewater Treatment Plants. *Stoch. Environ. Res. Risk Assess* **2019**, 33 (10), 1781–1792.
- (43) Zhu, J.; Jiang, Z.; Feng, L. Improved Neural Network with Least Square Support Vector Machine for Wastewater Treatment Process. *Chemosphere* **2022**, 308, 136116.
- (44) Mashford, J.; Marlow, D.; Tran, D.; May, R. Prediction of Sewer Condition Grade Using Support Vector Machines. *J. Comput. Civ. Eng* **2011**, 25 (4), 283–290.
- (45) Liu, Z.; Wan, J.; Ma, Y.; Wang, Y. Online Prediction of Effluent COD in the Anaerobic Wastewater Treatment System Based on PCA-LSSVM Algorithm. *Environ. Sci. Pollut. Res* **2019**, 26 (13), 12828–12841.
- (46) Singh, K. P.; Basant, N.; Gupta, S. Support Vector Machines in Water Quality Management. *Anal. Chim. Acta* **2011**, 703 (2), 152–162.
- (47) Yang, Y. H. Support Vector Machines for Environmental Informatics: Application to Modelling the Nitrogen Removal Processes in Wastewater Treatment Systems. *J. Environ. Inform* **2006**, 7 (1), 14–23.
- (48) Arefinia, A.; Bozorg-Haddad, O.; Chang, H. The Role of Data Mining in Water Resources Management. In *Essential Tools for Water Resources Analysis, Planning, and Management*, Bozorg-Haddad, O. Ed.; Springer: Singapore, 2021; pp. 85–99.
- (49) Hamed, M. M.; Khalafallah, M. G.; Hassanien, E. A. Prediction of Wastewater Treatment Plant Performance Using Artificial Neural Networks. *Environ. Model. Softw* **2004**, 19 (10), 919–928.
- (50) Jawad, J.; Hawari, A. H.; Javaid Zaidi, S. Artificial Neural Network Modeling of Wastewater Treatment and Desalination Using Membrane Processes: A Review. *Chem. Eng. J* **2021**, 419, 129540.
- (51) Nasr, M. S.; Moustafa, M. A. E.; Seif, H. A. E.; El Kobrosy, G. Application of Artificial Neural Network (ANN) for the Prediction of EL-AGAMY Wastewater Treatment Plant Performance-EGYPT. *Alex. Eng. J* **2012**, 51 (1), 37–43.
- (52) Warren-Vega, W. M.; Montes-Pena, K. D.; Romero-Cano, L. A.; Zarate-Guzman, A. I. Development of an Artificial Neural Network (ANN) for the Prediction of a Pilot Scale Mobile Wastewater Treatment Plant Performance. *J. Environ. Manage* **2024**, 366, 121612.
- (53) Zhang, S.; Omar, A. H.; Hashim, A. S.; Alam, T.; Khalifa, H. A. E.-W.; Elkotb, M. A. Enhancing Waste Management and Prediction of Water Quality in the Sustainable Urban Environment Using Optimized Algorithm of Least Square Support Vector Machine and Deep Learning Techniques. *Urban Clim* **2024**, 58, 102095.
- (54) Al-Obaidi, M. A.; Li, J.-P.; Kara-Zaitri, C.; Mujtaba, I. M. Optimisation of Reverse Osmosis Based Wastewater Treatment System for the Removal of Chlorophenol Using Genetic Algorithms. *Chem. Eng. J* **2017**, 316, 91–100.
- (55) Danandeh Mehr, A.; Nourani, V.; Kahya, E.; Hrnjica, B.; Sattar, A. M. A.; Yaseen, Z. M. Genetic Programming in Water Resources Engineering: A State-of-the-Art Review. *J. Hydrol* **2018**, 566, 643–667.
- (56) Hoseingholi, P.; Moeini, R. Pipe Failure Prediction of Wastewater Network Using Genetic Programming: Proposing Three Approaches. *Ain Shams Eng. J* **2023**, 14 (5), 101958.
- (57) Dalmau, M.; Rodriguez-Roda, I.; Ayesa, E.; Odriozola, J.; Sancho, L.; Comas, J. Development of a Decision Tree for the Integrated Operation of Nutrient Removal MBRs Based on Simulation Studies and Expert Knowledge. *Chem. Eng. J* **2013**, 217, 174–184.

- (58) Bahramian, M.; Dereli, R. K.; Zhao, W.; Giberti, M.; Casey, E. Data to Intelligence: The Role of Data-Driven Models in Wastewater Treatment. *Expert Syst. Appl.* **2023**, *217*, 119453.
- (59) Duarte, M. S.; Martins, G.; Oliveira, P.; Fernandes, B.; Ferreira, E. C.; Alves, M. M.; Lopes, F.; Pereira, M. A.; Novais, P. A Review of Computational Modeling in Wastewater Treatment Processes. *ACS EST Water* **2024**, *4* (3), 784–804.
- (60) Achite, M.; Samadianfard, S.; Elshaboury, N.; Sharafi, M. Modeling and Optimization of Coagulant Dosage in Water Treatment Plants Using Hybridized Random Forest Model with Genetic Algorithm Optimization. *Environ. Dev. Sustain.* **2023**, *25* (10), 11189–11207.
- (61) Crini, G.; Lichtfouse, E. Advantages and Disadvantages of Techniques Used for Wastewater Treatment. *Environ. Chem. Lett.* **2019**, *17* (1), 145–155.
- (62) Cheng, Q.; Chunhong, Z.; Qianglin, L. Development and Application of Random Forest Regression Soft Sensor Model for Treating Domestic Wastewater in a Sequencing Batch Reactor. *Sci. Rep.* **2023**, *13* (1), 9149.
- (63) Mateo Pérez, V.; Mesa Fernández, J. M.; Villanueva Balsera, J.; Alonso Álvarez, C. A Random Forest Model for the Prediction of FOG Content in Inlet Wastewater from Urban WWTPs. *Water* **2021**, *13* (9), 1237.
- (64) Mousavi, S. Z.; Momeni, S. A.; Dehdashti, B.; Dehdashti, D.; Fatehizadeh, A.; Amin, M. M. Advanced Treatment of Industrial Estate Effluent Using Ozone-Membrane Processes Based on Optimized Random Forest Model. *J. Water Process Eng.* **2022**, *48*, 102897.
- (65) Salem, M.; EL-Sayed Gabr, M.; Mossad, M.; Mahanna, H. Random Forest Modelling and Evaluation of the Performance of a Full-Scale Subsurface Constructed Wetland Plant in Egypt. *Ain Shams Eng. J.* **2022**, *13* (6), 101778.
- (66) Arefinia, A.; Bozorg-Haddad, O.; Ahmadaali, K.; Zolghadr-Asli, B.; Loáiciga, H. A. Cropping Patterns Based on Virtual Water Content Considering Water and Food Security under Climate Change Conditions. *Nat. Hazards* **2022**, *114* (2), 1709–1721.
- (67) Lamontagne, J. R.; Barber, C. A.; Vogel, R. M. Improved Estimators of Model Performance Efficiency for Skewed Hydrologic Data. *Water Resour. Res.* **2020**, *56* (9), No. e2020WR027101.
- (68) Mishra, D.; Khan, M. A.; Mudgal, M.; Padmakaran, P.; Chakradhar, B. Performance Evaluation of an Effluent Treatment Plant for a Pulp & Paper Mill. *Indian J. Chem. Technol.* **2009**, *16*, 79–83.
- (69) Al Kholif, M.; Subianto, A.; Sutrisno, J. Effect of Hydraulic Retention Time (HRT) in an Anaerobic Baffled Reactor (ABR) on the Reduction of BOD and COD in Slaughterhouse Industrial Wastewater. *Adv. Environ. Technol.* **2025**, *11* (1), 1–12.
- (70) Zulfikar, Z.; Nasrullah, N.; Kartini, K.; Aditama, W. Effect of Hydraulic Retention Time on the Levels of Biochemical Oxygen Demand and Total Suspended Solid with Simple Integrated Treatment as an Alternative to Meet the Household Needs for Clean Water. *Open Access Maced. J. Med. Sci.* **2022**, *10* (E), 6–11.
- (71) Garcia-Ochoa, F.; Gomez, E. Bioreactor Scale-up and Oxygen Transfer Rate in Microbial Processes: An Overview. *Biotechnol. Adv.* **2009**, *27* (2), 153–176.
- (72) Yang, H.-C.; Park, S.-K. Oxygen Transfer Characteristics of an Ejector Aeration System. *Int. J. Fluid Mach. Syst.* **2012**, *5* (1), 10–17.
- (73) Ahmad Yasmin, N. S.; Abdul Wahab, N.; Yusuf, Z. Modeling of Membrane Bioreactor of Wastewater Treatment Using Support Vector Machine. In *Modeling, Design and Simulation of Systems*; Mohamed Ali, M. S.; Wahid, H.; Mohd Subha, N. A.; Sahlan, S.; Yunus, M. A.; Wahap, A. R. Eds.; Springer: Singapore, 2017; pp 485–495.
- (74) Thurnhofer-Hemsi, K.; López-Rubio, E.; Molina-Cabello, M. A.; Najarian, K. Radial Basis Function Kernel Optimization for Support Vector Machine Classifiers. *arXiv* **2020**.
- (75) Sahu, P.; Londhe, S. N.; Kulkarni, P. S. Prediction of Water Quality Parameters Using Support Vector Regression. *Innov. Infrastruct. Solut.* **2023**, *8* (9), 229.
- (76) Xiong, Z.; Liu, X.; Igou, T.; Li, Z.; Chen, Y. Using Hybrid Machine Learning to Predict Wastewater Effluent Quality and Ensure Treatment Plant Stability. *Water* **2025**, *17* (13), 1851.
- (77) Jadhav, A. R.; Pathak, P. D.; Raut, R. Y. Water and Wastewater Quality Prediction: Current Trends and Challenges in the Implementation of Artificial Neural Network. *Environ. Monit. Assess.* **2023**, *195* (2), 321.
- (78) Mustafa, H. M.; Mustapha, A.; Hayder, G.; Salisu, A. Applications of IoT and Artificial Intelligence in Water Quality Monitoring and Prediction: A Review. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)*; IEEE: Coimbatore, India, 2021; pp 968–975.
- (79) Elsayed, A.; Ghaith, M.; Yosri, A.; Li, Z.; El-Dakhkhni, W. Genetic Programming Expressions for Effluent Quality Prediction: Towards AI-Driven Monitoring and Management of Wastewater Treatment Plants. *J. Environ. Manage.* **2024**, *356*, 120510.
- (80) Beraud, B.; Lemoine, C.; Steyer, J.-P. Multiobjective Genetic Algorithms for the Optimisation of Wastewater Treatment Processes. In *Computational Intelligence Techniques for Bioprocess Modelling, Supervision and Control*, Do Carmo Nicoletti, M.; Jain, L. C. Eds.; Springer: Berlin, Heidelberg, 2009; pp 163–195.
- (81) Rashidi-Khazaei, P.; Rezvantab, S.; Kheshti Monasebi, A. Predicting Wastewater Treatment Plant Effluent Quality Using Ensemble Learning. *Green Technol.* **2024**, *2* (1), 35–43.
- (82) Waqas, S.; Harun, N. Y.; Sambudi, N. S.; Arshad, U.; Nordin, N. A. H. M.; Bilal, M. R.; Saeed, A. A. H.; Malik, A. A. SVM and ANN Modelling Approach for the Optimization of Membrane Permeability of a Membrane Rotating Biological Contactor for Wastewater Treatment. *Membranes* **2022**, *12* (9), 821.
- (83) Ferreira, L. A.; Guimarães, F. G.; Silva, R. Applying Genetic Programming to Improve Interpretability in Machine Learning Models. *arXiv* **2020**.
- (84) Afan, H. A.; Melini Wan Mohtar, W. H.; Khaleel, F.; Kamel, A. H.; Mansoor, S. S.; Alsultani, R.; Ahmed, A. N.; Sherif, M.; El-Shafie, A. Data-Driven Water Quality Prediction for Wastewater Treatment Plants. *Heliyon* **2024**, *10* (18), No. e36940.
- (85) Hamada, M. S.; Zaqoot, H. A.; Sethar, W. A. Using a Supervised Machine Learning Approach to Predict Water Quality at the Gaza Wastewater Treatment Plant. *Environ. Sci. Adv.* **2024**, *3* (1), 132–144.
- (86) Li, Y.; Kong, B.; Yu, W.; Zhu, X. An Attention-Based CNN-LSTM Method for Effluent Wastewater Quality Prediction. *Appl. Sci.* **2023**, *13* (12), 7011.
- (87) Recio-Colmenares, R.; León Becerril, E.; Gurubel Tun, K. J.; Conchas, R. F. Design of a Soft Sensor Based on Long Short-Term Memory Artificial Neural Network (LSTM) for Wastewater Treatment Plants. *Sensors* **2023**, *23* (22), 9236.
- (88) Giannakis, S.; Lin, K.-Y. A.; Ghanbari, F. A Review of the Recent Advances on the Treatment of Industrial Wastewaters by Sulfate Radical-Based Advanced Oxidation Processes (SR-AOPs). *Chem. Eng. J.* **2021**, *406*, 127083.
- (89) Asghar, M. N.; Khan, S.; Mushtaq, S. Management of Treated Pulp and Paper Mill Effluent to Achieve Zero Discharge. *J. Environ. Manage.* **2008**, *88* (4), 1285–1299.
- (90) Newhart, K. B.; Thompson, K. A.; Branch, A.; Salvesson, A.; Safarik, J.; Plumlee, M. H. NDMA Soft-Sensors for Potable Reuse: A Model Development Study. *Water Res.* **2025**, *287*, 124451.
- (91) Zacherl, L.; Baumann, T. Digital Twin of an Open Cooling Tower: Experimental Studies and Numerical Validation. *ACS Eng. Au* **2025**, *5* (4), 347–358.
- (92) Das, A.; Prasad, V.; Srinivasan, R. Lazy Fusion of Multimodal Sensors for Cost-Effective Process Monitoring. *ACS Eng. Au* **2025**, *5* (4), 370–383.
- (93) Hacker, J.; Iqbal, F.; Osuna, M.; Perea, O.; Sánchez, K.; Page, C.; Torres, J.; Uddin, M. N.; Walden, H.; Westerbeke, M.; et al. Aquatic Quality Watch Informed by Communities (AQWIC) Facilitating the Adoption of Low-Cost Sensor Systems for Under-served Communities: A Review and Tutorial. *Environ. Sci. Adv.* **2025**, *4* (4), 512–529.

(94) Kumar, R.; Goyal, M. K. Role of AI&ML in Modernizing Water and Wastewater Treatment Processes. *Water. Air. Soil Pollut* **2025**, 236 (1), 31.