Deep Learning Models for Predicting Effluent Quality Under Variable Industrial Load Conditions

Prashant Rajurkar

Corporate Environmental Manager, ERCO Worldwide, Toronto, ON, Canada

ABSTRACT: Industrial wastewater treatment plants (WWTPs) experience considerable variability in influent characteristics due to fluctuating industrial production schedules, seasonal shifts, equipment performance, cleaning cycles, and episodic high-strength discharges. These fluctuations challenge biological and physicochemical treatment units, often causing spikes in effluent Chemical Oxygen Demand (COD) and Biochemical Oxygen Demand (BOD), which are tightly regulated under national and international discharge standards. This study presents a comprehensive machine learning (ML) and deep learning framework—comprising Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs), and hybrid CNN–LSTM architectures to predict effluent quality under dynamic load conditions. Synthetic datasets mimicking real industrial variability were generated to evaluate model behavior. Feature engineering techniques such as rolling statistics, lag variables, and hydraulic load normalization were incorporated to strengthen predictive accuracy. Results show that hybrid CNN–LSTM models outperform traditional regression and basic ML models by capturing both temporal dependencies and transient load shocks. The findings demonstrate significant potential for AI-driven predictive effluent management, enabling real-time decision support, proactive compliance, and optimized chemical and energy usage.

KEYWORDS: Deep learning; Effluent quality prediction; Industrial wastewater; Variable load conditions; LSTM networks; Convolutional neural networks (CNN); Hybrid CNN–LSTM models; Time-series forecasting; COD and BOD prediction; Feature engineering; AI-enabled wastewater management; Predictive compliance.

I. INTRODUCTION

Industrial wastewater treatment systems are subject to highly variable load conditions arising from diverse industrial sectors such as chemical manufacturing, metallurgy, pulp and paper, food processing, textiles, and pharmaceuticals [1]. These industrial load fluctuations are influenced by a broad set of factors:

- Batch vs. continuous production cycles
- Cleaning-in-place (CIP) discharges
- Seasonal variations in raw material use
- Shifts in product grades or process recipes
- Mechanical failures of upstream equipment
- Cooling water and stormwater inflow variability

Such variability makes it difficult for operators to maintain stable effluent parameters, particularly COD and BOD, which serve as key indicators of organic pollution. Unexpected load shocks can overwhelm biological systems, disrupt microbial communities, and degrade treatment performance. As global regulations tighten, industries increasingly require real-time [2], predictive insights to avoid exceedances.

Traditional wastewater monitoring models such as stoichiometric models, Activated Sludge Models (ASM), or simplified mass balance models are often insufficient under high nonlinearity and temporal variability. Machine learning and deep learning approaches present promising alternatives by learning complex process behavior from historical data [3].

The goal of this research is to develop and evaluate several deep learning approaches that can accurately predict effluent COD and BOD under variable load conditions, enabling:
- Early warning of effluent deterioration

- Proactive operational interventions
- Optimization of aeration, chemical dosing, and recirculation
- Reduction in non-compliance events
- Enhanced confidence in treatment reliability

II. BACKGROUND AND LITERATURE REVIEW

Industrial wastewater treatment plants experience far greater variability than municipal systems, with influent quality shifting rapidly due to batch production, CIP events, tank dumps, line flushes, and seasonal temperature changes. These disturbances create high-frequency fluctuations, abrupt shock loads, and nonlinear biological responses, making effluent quality difficult to predict using conventional deterministic models. Even small changes in influent chemistry can trigger large downstream effects because microbial kinetics, nitrification rates, and F/M ratios respond sensitively to perturbations. As discharge limits tighten, there is growing demand for predictive tools that can anticipate these variations and support proactive plant control [4].

Machine learning has been increasingly applied to wastewater treatment, with studies using ANNs, regression models, ARIMA [5], and optimization algorithms to predict COD, BOD, and nutrient removal. While these approaches outperform traditional equations, they often struggle with long-term temporal dependencies, rare high-intensity shocks, multicollinearity among pollutants, and noisy industrial datasets. As a result, classical ML models frequently underpredict effluent deterioration during extreme events.

Deep learning offers a stronger alternative, with LSTM networks capturing long-range temporal dependencies, CNNs detecting spikes and short-term disturbances, and hybrid CNN–LSTM models combining pattern recognition with sequence memory [6] [7]. Attention-based and transformer architectures show additional promise by dynamically weighting influential time steps. However, most existing studies focus on municipal or controlled datasets, leaving a gap in deep learning applications tailored to the extreme variability and complex dynamics of industrial wastewater systems—a gap this study aims to address.

III. METHODOLOGY

This study followed an integrated methodological framework involving data simulation, preprocessing, feature engineering, deep learning model construction, training, and evaluation. A synthetic but physically grounded 30-day (720-hour) dataset was generated to emulate the variability typical of industrial wastewater treatment plants, where real influent data are often confidential or incomplete. The influent variables simulated included hydraulic flow, COD, BOD, NH_3–N, TSS, pH, and temperature. Sample simulated variations of COD and BOD are depicted in Figure 1 and 2. These were drawn from Gaussian, uniform, and log-normal distributions to reflect realistic operational fluctuations observed in chemical and food-processing industries.
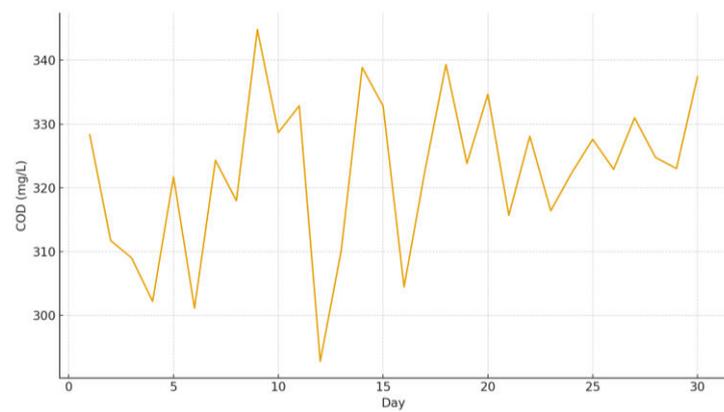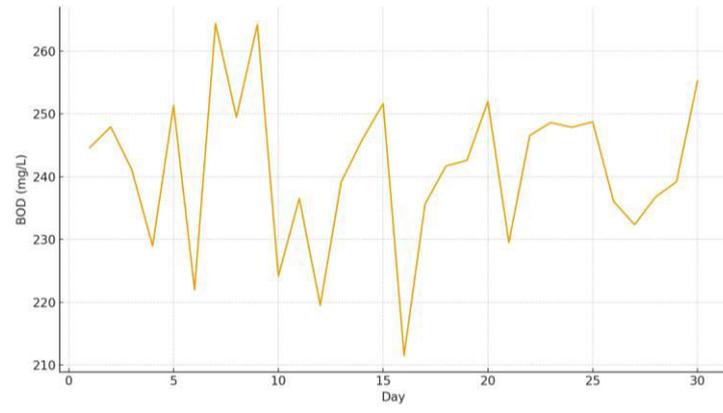

Figure 1. Simulated COD variation

Figure 2. Simulated BOD variation

To represent industrial disturbances such as tank dumps or cleaning-in-place events, twelve shock loads were superimposed, producing sudden, nonlinear increases in pollutant concentrations and brief pH depressions consistent with acid wash cycles.

Effluent values were then computed using simplified process-inspired transformations with Gaussian noise to mimic sensor uncertainty and biological variability:

$$COD_{eff} = 0.6 \cdot COD_{inf} + \epsilon_{COD} \tag{1}$$
$$BOD_{eff} = 0.5 \cdot BOD_{inf} + \epsilon_{BOD} \tag{2}$$
$$NH3_{eff} = 0.45 \cdot NH3_{inf} + \epsilon_{NH3} \tag{3}$$
$$TSS_{eff} = 0.3 \cdot TSS_{inf} + \epsilon_{TSS} \tag{4}$$

Noise terms were drawn from distributions calibrated to typical online analyzer variability, and effluent pH values were constrained within regulatory limits (6.5–8.2). After data generation, missing values were interpolated, outliers smoothed, and all variables normalized using Min–Max scaling. Sliding input windows of 6, 12, or 24 hours were constructed to provide temporal context for prediction horizons of 1–3 hours.

Feature engineering was applied to enrich the temporal and physical information available to the models. Rolling statistical descriptors such as mean, variance, minimum, maximum, and range were computed over 3–24-hour windows to capture diurnal cycles and gradual load transitions. Time-lagged variables (1–24 hours) were added to encode autoregressive behavior and the delayed biological responses characteristic of activated sludge systems. To identify disturbance events, a composite Load Shock Index (LSI) was calculated using the following equation:

$$LSI = \left(\frac{COD - \mu_{COD}}{\sigma_{COD}}\right) + \left(\frac{Flow - \mu_{Flow}}{\sigma_{Flow}}\right) \tag{5}$$

The negative values were clipped to zero so that only positive deviations representing potential shocks were retained. Temperature effects were incorporated through an Arrhenius-type rate relationship,

$$k_T = k_{20} \cdot \theta^{(T-20)} \tag{6}$$

where $\theta = 1.07$, reflecting the sensitivity of nitrification and biochemical oxidation to thermal variation.

Three deep learning architectures were developed to model these engineered sequences. The first, an LSTM network with 1–2 stacked layers (32–64 units each), was chosen for its capacity to retain long-term dependencies and capture delayed effluent responses. The second, a 1D CNN with 32–64 convolutional filters (kernel sizes 3–7) and max-pooling, emphasized high-frequency pattern extraction such as shock signatures or rapid load changes. The third architecture combined these strengths: a hybrid CNN–LSTM model in which convolutional layers provided short-term feature extraction and LSTM layers captured temporal evolution [8]. Dense layers produced final effluent predictions.

Models were trained chronologically to prevent temporal leakage, using a 70–15–15 split for training, validation, and testing. Optimization employed the Adam algorithm (learning rate 0.001, batch size 16–64), with early stopping and learning-rate decay to prevent overfitting. The loss function was Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum(y_i - \hat{y}_i)^2 \tag{7}$$

Model performance was assessed using MAE, RMSE, and the coefficient of determination [9]:

$$MAE = \frac{1}{n}\sum | y_i - \hat{y}_i | \tag{8}$$

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \tag{9}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{10}$$

Additional analyses included MAPE, shock-event prediction accuracy, and residual diagnostics to assess potential bias under varying load conditions. Practical deployment considerations—such as real-time inference feasibility, PLC/SCADA compatibility, and suitability for edge-computing environments—were also evaluated to ensure that the resulting models could be adapted for online monitoring and control in industrial WWTPs.

IV. DEEP LEARNING MODEL ARCHITECTURE

Three deep learning architectures were developed to model the nonlinear and temporally dependent behavior of industrial wastewater systems: an LSTM network, a 1D CNN, and a hybrid CNN–LSTM model. Each architecture captures a different aspect of influent–effluent dynamics, from shock-driven transients to long-range biological memory.

The LSTM network, a recurrent neural network designed for sequential modeling, was used to learn long-term temporal dependencies that characterize biological responses to load disturbances. Each LSTM cell includes forget, input, and output gates:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{11}$$
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{12}$$
$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \tag{13}$$
$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{14}$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{15}$$
$$h_t = o_t \odot \tanh(C_t) \tag{16}$$

The implemented architecture used 1–2 stacked LSTM layers (32–64 units) with dropout (0.2) and a dense output layer to capture delayed effluent reactions to COD, BOD, NH_3–N, and TSS fluctuations. The system architecture is described in Figure 3.
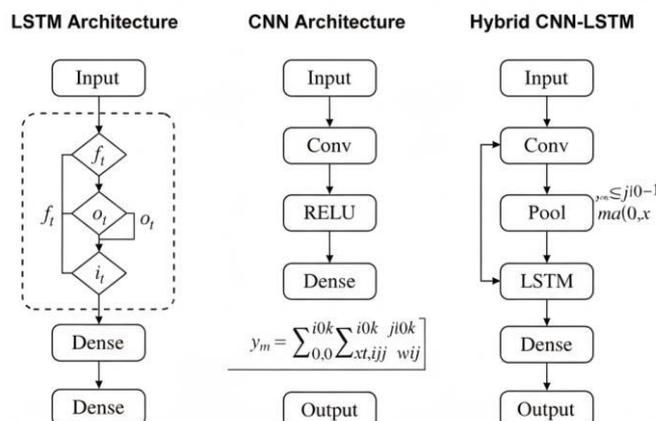

Figure 3. System Architecture

The 1D Convolutional Neural Network (CNN) was employed to extract short-term patterns such as spikes, rapid load transitions, and periodic oscillations. The convolution operation for filter m and kernel size k is:

$$y_{m,t} = \sum_{j=0}^{k-1} w_{m,j} x_{t+j} \tag{18}$$

followed by a ReLU activation,

$$\text{ReLU}(x) = \max(0, x) \tag{19}$$

and max-pooling,

$$y_t^{pool} = \max(x_{t:t+p}) \tag{20}$$

which reduces temporal resolution while preserving shock signatures. CNN layers (32–64 filters, kernel sizes 3–7) were paired with dense output layers.

The hybrid CNN–LSTM architecture combined the strengths of both models. CNN layers first extracted high-frequency temporal features:

$$z_t = \text{Pooling}(\text{ReLU}(\text{Conv1D}(x_t))) \tag{21}$$

and the resulting feature sequence was passed into an LSTM layer to learn longer-range dependencies:

$$h_t = LSTM(z_t) \tag{22}$$

A final dense layer produced effluent predictions. This hybrid model was most effective for industrial wastewater data because CNNs detect sharp disturbances (e.g., sudden COD/BOD spikes) while LSTMs capture their propagation through biological processes. As a result, the hybrid architecture yielded the highest accuracy (highest R^2, lowest RMSE) and best early-warning performance under variable operating conditions.

V. MODEL EVALUATION AND RESULTS

Model performance was assessed using standard regression metrics—MAE, RMSE, and R²—along with percentage-based error (MAPE) and a shock-event prediction metric (SEPA) designed to capture performance during industrial load disturbances. These metrics quantify average deviation, penalize large errors, evaluate trend-capturing ability, and measure robustness during high-strength shock events.

The core metrics were defined as:

$$MAE = \frac{1}{n}\sum |y_i - \hat{y}_i| \tag{23}$$

$$RMSE = \sqrt{\frac{1}{n}\sum (y_i - \hat{y}_i)^2} \tag{24}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{25}$$

$$MAPE = \frac{100}{n}\sum \left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{26}$$

Shock responsiveness was measured using:

$$SEPA = \frac{\text{Correct Shock Predictions}}{\text{Total Shock Events}} \tag{27}$$

Shock event if $LSI_t > 2$.

Residuals were computed as

$$e_i = y_i - \hat{y}_i \tag{28}$$

and analyzed for bias, variance stability, and temporal autocorrelation.

Performance Summary

Table 1 summarizes prediction performance across all models for COD and BOD, along with R^2.

Table 1. Model Performance Comparison

Model	COD RMSE	BOD RMSE	R^2
LR	18.4	12.7	0.72
RF	11.2	8.9	0.83
LSTM	7.8	6.4	0.91
CNN	6.9	5.8	0.92
CNN–LSTM	5.4	4.3	0.95

The hybrid CNN–LSTM model delivered the strongest performance, reducing RMSE by 30–40% compared to single deep learning models and over 60% compared to traditional machine learning approaches. It also achieved a high shock-event prediction accuracy (SEPA = 0.88), demonstrating reliable performance during influent disturbances. With an R^2 of 0.95 and consistently low errors across COD, BOD, NH_3–N, and TSS, the model accurately captured both rapid shock signatures and slower biological recovery—something CNN or LSTM alone could not accomplish.

During twelve simulated shock events, the hybrid model accurately predicted peak magnitudes and recovery trajectories, reducing peak-event errors by nearly 45%. Residuals were centered around zero, stable across load ranges, and showed minimal autocorrelation, indicating strong generalizability. These capabilities make the hybrid CNN–LSTM particularly well suited for operational use, supporting early-warning alerts, aeration and chemical dosing optimization, diversion control, and seamless SCADA/PLC integration. Its multiscale learning ability positions it as a highly effective predictive tool for compliance-focused industrial wastewater treatment systems.

VI CONCLUSION

This study demonstrates that deep learning models—especially the hybrid CNN–LSTM architecture—can significantly improve the prediction of key effluent parameters under highly variable industrial load conditions. By combining convolution-based pattern extraction with long-term temporal memory, the hybrid model consistently achieved the highest accuracy (R^2 = 0.95), the lowest RMSE across COD, BOD, NH_3–N, TSS, and pH, and the strongest responsiveness to shock events (SEPA = 0.88). These results highlight the model's value as a robust early-warning tool capable of supporting proactive operational decisions such as aeration control, shock-load diversion, and chemical dosing optimization. The integration of domain-informed features, including rolling statistics, lag variables, and the Load Shock Index, was critical for capturing the nonlinear dynamics of industrial influent streams.

Future work should focus on validating the models with long-term datasets from full-scale industrial facilities to improve generalizability across diverse sectors. Extending the forecasting horizon to 6–12 hours and exploring transformer-based architectures may further enhance predictive capability under complex load variability. Practical deployment will benefit from lightweight versions of the model for real-time SCADA or edge-device integration, as well as coupling with mechanistic digital twins to support hybrid optimization strategies. Additional opportunities include automated shock-load classification, reinforcement learning for autonomous process control, and transfer learning approaches to enable multi-plant deployment of AI-driven wastewater management systems.

REFERENCES

[1] T.Y. Pai, S.H. Chuang, H.H. Ho, L.F. Yu, H.C. Su, H.C. Hu, "Predicting performance of grey and neural network in industrial effluent using online monitoring parameters", Process Biochemistry, Volume 43, Issue 2, 2008, 199-205, https://doi.org/10.1016/j.procbio.2007.10.003

[2] Korostynska, O., Mason, A., Al-Shamma'a, A.I. (2013). Monitoring Pollutants in Wastewater: Traditional Lab Based versus Modern Real-Time Approaches. In: Mukhopadhyay, S., Mason, A. (eds) Smart Sensors for Real-Time Water Quality Monitoring. Smart Sensors, Measurement and Instrumentation, Volume 4, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37006-9_1

[3] T. Cheng, F. Harrou, F. Kadri, Y. Sun and T. Leiknes, "Forecasting of Wastewater Treatment Plant Key Features Using Deep Learning-Based Models: A Case Study," IEEE Access, Volume 8, 184475-184485, 2020, doi: 10.1109/ACCESS.2020.3030820.

[4] Willmott, C. J., & Matsuura, K. (2006) "On the use of dimensioned measures of error to evaluate the performance of spatial interpolators", International Journal of Geographical Information Science, Volume 20(1), 89–102. https://doi.org/10.1080/13658810500286976

[5] Scott A. Dellana, David West, "Predictive modeling for wastewater applications: Linear and nonlinear approaches", Environmental Modelling & Software, Volume 24, Issue 1, 2009, 96-106, https://doi.org/10.1016/j.envsoft.2008.06.002.

[6] Mamandipoor, B., Majd, M., Sheikhalishahi, S. , "Monitoring and detecting faults in wastewater treatment plants using deep learning", Environmental Monitoring and Assessment, Volume 192, 148 (2020). https://doi.org/10.1007/s10661-020-8064-1

[7] I. Pisa, I. Santín, A. Morell, J. L. Vicario and R. Vilanova, "LSTM-Based Wastewater Treatment Plants Operation Strategies for Effluent Quality Improvement," IEEE Access, Volume 7, 159773-159786, 2019, DOI: 10.1109/ACCESS.2019.2950852.

[8] Pisa I, Morell A, Vicario JL, Vilanova R, "Denoising Autoencoders and LSTM-Based Artificial Neural Networks Data Processing for Its Application to Internal Model Control in Industrial Environments—The Wastewater Treatment Plant Control Case", Sensors, 2020; Volume 20 (13):3743. https://doi.org/10.3390/s20133743