A Comprehensive Review of Machine Learning Applications in Wastewater Treatment: Current State, Comparative Analysis, and Future Directions

Mir Junaid Rasool[1*], J. Somashekar[2]

[1]Research Scholar, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jain (Deemed-to-be University), Bangalore, Karnataka, India
[2]Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jain (Deemed-to-be University), Bangalore, Karnataka, India

Email: mirjunaidmir@gmail.com[1*], jsomasekar@gmail.com[2]

Abstract

As the world's need for clean water keeps rising and pollution continues to worsen, there is a growing push for better wastewater treatment systems. Treatment plants (WWTPs) are essential not only for protecting public health but also for keeping the environment safe. Still, running these plants is not easy because the quality of incoming water often changes, the biological processes are complex, and regulations are very strict. Traditional methods usually fall short, being slow and inefficient. Newer approaches, like machine learning (ML) and artificial intelligence (AI), bring fresh opportunities by making it possible to predict issues in real time, spot irregularities, improve processes, and support better decision-making. This literature review brings together findings from five key research papers and over 40 additional studies published between 2018 and 2025. The review highlights a significant shift towards advanced deep learning (e.g., LSTM, GRU) and ensemble models, demonstrating superior performance in capturing complex, time-dependent data. Key trends include multi-source data fusion, expanding focus on effluent quality prediction for regulatory compliance, nutrient removal, energy optimization, and predictive maintenance. Despite these advancements, persistent challenges include data quality and availability, model interpretability ("black box" nature), generalizability across diverse WWTPs, and integration with existing infrastructure. Future research directions emphasize hybrid and physics-informed models, Explainable AI (XAI), Digital Twins, Reinforcement Learning for optimal control, and fostering interdisciplinary collaboration. Ultimately, ML/AI holds immense potential to revolutionize wastewater management, transitioning from reactive to proactive strategies, contingent on addressing these critical limitations for widespread and sustainable adoption.

Keywords

Wastewater Treatment Plants (WWTPs), Machine Learning, Artificial Intelligence, Effluent Quality Prediction, Process Optimization, Predictive Maintenance

Submission: 30 October 2025; **Acceptance:** 27 November 2025; **Available online:** December 2025

Introduction

Efficient and sustainable use of water resources has become one of the major global issues of this century. With rapid urbanization, rising population, and industrial development, the demand for reliable wastewater treatment is greater than ever(Malviya & Jaspal, 2021; Wang et al., 2024). Wastewater treatment plants (WWTPs) serve as critical infrastructure to protect both human health and ecosystems by eliminating pollutants from municipal and industrial wastewater before it is released or reused. However, operating these plants is highly complex because the processes are dynamic, non-linear, and affected by numerous biological, chemical, and physical factors, which can vary daily, seasonally, and annually(Inbar & Avisar, 2024; Malviya & Jaspal, 2021; Zamfir et al., 2025).

Traditionally, WWTPs depended on empirical methods and offline laboratory testing. This approach is typically slow, resource-heavy, and costly, since important water quality indicators like biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), and total suspended solids (TSS) often take several days to analyze. The delay in results prevents operators from making timely, data-informed decisions. As a consequence, plants may face higher operational expenses, greater energy use, and in some cases, difficulties in meeting strict environmental standards(Inbar & Avisar, 2024; Zamfir et al., 2025). Conventional systems frequently fail to keep pace with modern pollutants and variable influent quality. To overcome these limitations, advanced technologies such as artificial intelligence (AI), machine learning (ML), and deep learning (DL) are increasingly being recognized as game-changing tools for the wastewater treatment sector(Malviya & Jaspal, 2021; Yang et al., 2024). These data-driven methods, often described as “soft sensors” or “virtual instruments,” make use of large collections of past and live information from online sensors to create predictive models of key effluent parameters. In contrast to traditional mechanistic approaches built on first-principle equations and often restricted by their assumptions and complexity, AI models can reveal complex, non-linear connections among different process variables without the need for a detailed grasp of the underlying physics. This capability supports continuous monitoring, process optimization, and early alerts, helping operators shift from reactive control to a proactive management strategy(Chen & Kao, 2025; Malviya & Jaspal, 2021; Yang et al., 2024). The purpose of this paper is to provide a comprehensive and expert-level review of AI applications in WWTPs. It synthesizes recent advancements in predictive modeling methodologies, critically analyzes the performance of various algorithms, and discusses key strategies for overcoming persistent implementation challenges. The report will explore emerging paradigms such as Digital Twins and Explainable AI, and finally, present a roadmap for future research directions that can accelerate the integration of intelligent systems into sustainable wastewater management practices.

Literature Review: A Synthesis of AI/ML/DL Methodologies and Applications in WWTPs:

Key Wastewater Parameters:

To comprehend the application of AI in wastewater treatment, it is first necessary to understand the fundamental parameters that govern the process. These parameters serve as both the inputs and the targets for AI models, and their accurate measurement and prediction are essential for effective plant operation and regulatory compliance.

A diverse range of contaminants and quality indicators are monitored throughout the wastewater treatment process. These include:

Total Suspended Solids (TSS): This refers to the concentration of particulate matter present in wastewater. Excess solids can harm aquatic life and increase treatment costs. AI-driven models help forecast TSS levels so that treatment plants can stay within environmental standards(Inbar & Avisar, 2024).

Biochemical Oxygen Demand (BOD5): This parameter reflects how much oxygen microorganisms need to break down organic material during a five-day test. It is essential for wastewater plant design, evaluating operational performance, and calculating aeration needs, which account for a major share of a facility's energy usage(Inbar & Avisar, 2024).

Chemical Oxygen Demand (COD): COD indicates the overall oxygen required to chemically oxidize both organic and inorganic substances in wastewater. Because it is faster to measure than BOD5, it serves as a practical substitute for estimating total pollution. Comparing COD with BOD5 also helps determine the biodegradability of the influent and detect potential industrial discharges(Inbar & Avisar, 2024).

Total Nitrogen (TN) and Ammonia (NH3-N): Nitrogen compounds, especially ammonia, are primary contributors to eutrophication and oxygen depletion in natural waters. Their removal through nitrification and denitrification is a key treatment step in many plants. AI approaches are commonly applied to forecast how efficiently these nitrogen species can be reduced and to monitor effluent concentrations(Inbar & Avisar, 2024).

Total Phosphorus (TP): Like nitrogen, phosphorus is a major nutrient that contributes to eutrophication. Its removal is often a complex biological process, and TP is a parameter that frequently exceeds regulatory thresholds. Predictive models for TP are therefore crucial for ensuring compliance(Inbar & Avisar, 2024).

Core AI Paradigms

The AI methodologies applied to these parameters can be broadly categorized based on their learning objectives. Regression models are utilized to predict continuous, numerical values such as the concentration of a pollutant in mg/L. Conversely, classification models are designed to predict discrete outcomes, for example, whether a specific parameter will be above or below a regulatory threshold (a binary classification task) or to identify which operational state a plant is in (a multi-class task). Finally, time-series forecasting models are a specialized category, particularly effective for sequential data like daily or hourly measurements, to predict future values based on past trends and patterns(Inbar & Avisar, 2024; Malviya & Jaspal, 2021).

Predictive Modeling of Effluent Quality:

The development of accurate predictive models for effluent quality is a cornerstone of AI application in wastewater treatment. The literature reveals a wide range of approaches, from traditional statistical and machine learning models to complex deep learning architectures and hybrid systems.

Traditional and Ensemble Machine Learning Models

A range of well-known machine learning techniques has been applied effectively to forecast wastewater parameters. Selecting an appropriate algorithm typically depends on how complex the problem is and the trade-off between accuracy and interpretability.

- **Artificial Neural Networks (ANN) and Multilayer Perceptron (MLP):** Inspired by the architecture of the human brain, these approaches excel at capturing intricate, non-linear patterns(Omarova et al., 2023). Research has demonstrated their strength in estimating key indicators. For example, one comparative analysis identified an ANN-MLP configuration as the best performer for TSS estimation, reporting an R^2 of 0.8 on the test data. Other studies have achieved an R^2 of 0.97 for COD prediction(Inbar & Avisar, 2024). Despite their strong predictive capability, ANNs are sometimes considered "black boxes," making them harder for operators to interpret and trust(Masood et al., 2024).
- **Tree-Based Ensemble Models:** These methods merge many individual decision trees into a single, more generalized model, which typically improves accuracy and adaptability.
- **Random Forest (RF):** RF constructs numerous decision trees and combines their outputs. It handles high-dimensional inputs effectively and has proven superior for estimating COD and TSS in some cases, reaching about 91 % and 95 % accuracy, respectively(Mahanna et al., 2024).
- **AdaBoost (Adaptive Boosting):** AdaBoost trains a sequence of models, each one correcting the mistakes of its predecessor. In one study on TSS prediction, it slightly outperformed other approaches, yielding a test-set R^2 of 0.77(Inbar & Avisar, 2024).
- **XGBoost (Extreme Gradient Boosting):** This enhanced form of gradient boosting consistently delivers strong accuracy and reliability. In an Effluent Quality Index study, XGBoost achieved the smallest Mean Absolute Percentage Error (MAPE) and an R^2 of 0.813, surpassing alternatives such as AdaBoost and Support Vector Regression(Bo-Qi et al., 2025).
- **Gene Expression Programming (GEP):** A unique type of evolutionary algorithm, GEP is notable for its ability to produce explicit mathematical expressions that link input variables to the target output. This provides a level of model transparency and interpretability that is often missing from other ML models. A study on predicting WWTP influent parameters found that GEP models were the most accurate for BOD5 and COD, with R2 values of 0.784 and 0.861, respectively(Inbar & Avisar, 2024).The derived equations provide a clear, functional relationship that can be easily understood and implemented by operators. The model for BOD5, for example, involved terms related to TSS, OrgN, and OrgP, while excluding ammonia and inorganic phosphorus, which aligns with the biochemical principles of BOD5 measurement(Inbar & Avisar, 2024).
- **Support Vector Regression (SVR):** Based on the principle of finding an optimal hyperplane to separate data, SVR models have also shown strong performance in predicting wastewater parameters. One study found that SVR excelled in fitting accuracy for an effluent quality index (EQI), achieving the highest R2 of 0.826, though it exhibited less stability in its predictions compared to XGBoost(Bo-Qi et al., 2025).

It is apparent that the performance of a given algorithm is not universal but is highly dependent on the specific wastewater parameter being predicted and the characteristics of the dataset. For instance, while ANN-MLP performed best for TSS in one study, GEP was superior for BOD5 and COD in another, and XGBoost was the top performer for a composite EQI(Malviya

& Jaspal, 2021; Pisa et al., 2021). This highlights a crucial consideration: the selection of a model is not a one-size-fits-all problem. It is a process of balancing multiple objectives—including predictive accuracy, computational efficiency, data requirements, and model interpretability—based on the unique context of the application. The transparency offered by a model like GEP, for example, may be more valuable to a plant manager than a marginal increase in R2 from a more complex, opaque model.

Advanced Deep Learning Architectures

Deep learning, a branch of machine learning, employs multiple neural layers capable of automatically extracting complex features from data. These techniques excel when working with large and intricate datasets and are especially valuable for recognizing long-range patterns in time-series information, an important aspect of monitoring wastewater treatment processes.

- **Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU):** LSTM and GRU are specialized forms of recurrent neural networks designed to handle sequential data. They effectively address issues such as vanishing or exploding gradients that often affect standard RNNs. Their built-in gating structures allow them to retain or discard information as needed, making them well suited for capturing the temporal behavior of WWTP operations(Inbar & Avisar, 2024). One study that compared LSTM, GRU, and Transformer models on a simulated WWTP dataset (BSM2) found that GRU consistently demonstrated the most robust performance during dynamic conditions like rain and storm events. It effectively balanced predictive responsiveness and stability, whereas LSTM struggled with rapid fluctuations. The Transformer model, while excelling in stable dry weather, was more sensitive to sudden changes(Voipan et al., 2025). This indicates that GRU's simplified architecture and efficient information flow make it particularly well-suited for the unpredictable and dynamic nature of wastewater data(Voipan et al., 2025).
- **Transformers:** Emerging from natural language processing, Transformer models use self-attention mechanisms to weigh the importance of different data points in a sequence. A comparative study noted that the Transformer network delivered the best performance under stable dry weather conditions and showed a slight advantage in capturing complex rebound effects after rainfall. However, its overall performance during storms was less stable than that of the GRU model(Voipan et al., 2025). This suggests that while Transformer models are powerful, their application in WWTPs may be better suited for systems with relatively stable inputs or as a component of a hybrid model.

The performance differences between these deep learning models during stable vs. dynamic conditions is a key finding in literature. While all models, even traditional ones, can perform well under stable dry weather, they all experience increased prediction errors at the onset of rain or storm events due to sudden changes in flow and pollutant loads(Voipan et al., 2025). The ability of models like GRU to quickly adapt to these sudden changes and maintain stability during the critical post-event stabilization phase makes them particularly valuable for building real-time monitoring and early warning systems. This highlights the need to evaluate models not just on their average performance, but on their robustness during the most challenging operational periods.

Hybrid and Multi-Output Models

An emerging direction in this field is the creation of hybrid approaches that integrate the advantages of multiple model types. This approach helps address the weaknesses of individual models and enhances overall predictive accuracy and reliability.

- **Hybrid Architectures:** These methods merge different learning frameworks to build a stronger predictive system. One illustration is a deep learning framework that links a Temporal Convolutional Network (TCN) with a Long Short-Term Memory (LSTM) network (TCN-LSTM). This combined model was designed to forecast hourly total nitrogen (TN) levels in wastewater treatment plants and achieved 33.1% greater accuracy than using TCN or LSTM alone, and 63.6% better than a conventional feedforward neural network (FFNN)(Y. Xie et al., 2024). The result highlights how combining complementary deep learning techniques—using TCN for local pattern detection and LSTM for long-term sequence modeling—can significantly boost performance.
- **Clustering-Based Hybrid Models:** Another innovative method employs a two-step prediction strategy. It first applies clustering to divide time-series data into segments that reflect different operating states (such as normal or rainy conditions) and then uses the best-fit model for each cluster. A study applying this technique on the BSM2 platform utilized a Partial Least Squares Random Weight Neural Network (PLS-RWNN) for large-sample, stable conditions, and a Multi-output Correlation Vector Machine (MRVM) for small-sample, high-variability situations. This combined system improved the root mean square successive difference (RMSSD) by 42.17% compared to using a single model on unclustered data(Inbar & Avisar, 2024).
- **Multi-Source Data Fusion:** Modern WWTPs generate a wealth of data from various sources, including water quantity, process variables, energy consumption, and traditional water quality measurements. Advanced models are now designed to fuse this multi-source data for a more comprehensive and accurate prediction. A deep learning framework using LSTM and GRU at an Industrial Effluent Treatment Plant (IETP) in China found that fusing these data sources enabled the deep learning models to significantly outperform traditional machine learning models in predicting effluent quality(Inbar & Avisar, 2024). This approach validates the existence of complex relationships between seemingly disparate variables, such as energy consumption and water quality(Inbar & Avisar, 2024).

This movement towards hybridization and multi-source data fusion represents a key development in the field. These approaches are not simply about achieving incremental performance gains; they are a strategic response to the inherent weaknesses of single models and single data streams. Mechanistic models provide interpretability but struggle with real-world data variability, while data-driven models are powerful but can be "black boxes" that require massive datasets. Hybrid and data fusion models, by integrating different methodologies and data sources, aim to create more resilient, accurate, and comprehensive solutions that are better suited for the complexities of real-world WWTPs.

Data-Centric Strategies for Model Enhancement:

The performance of any data-driven model is fundamentally tied to the quality and relevance of the data it is trained on. Consequently, advanced data-centric strategies, such as feature selection and data management, are as crucial as the model architecture itself.
Feature selection involves pinpointing and choosing the most important input variables for a predictive model. This key step simplifies computations, lowers the chance of overfitting, and enhances both the accuracy and the interpretability of the model(Inbar & Avisar, 2024).

Methodologies: Research identifies several main types of feature-selection techniques.

- **Filter methods** (e.g., Correlation, Mutual Information): These approaches pick variables according to how strongly they relate to the target outcome, independent of the specific machine-learning algorithm(Malviya & Jaspal, 2021).
- **Wrapper methods** (e.g., Sequential Backward Selection, SBS): Here, subsets of variables are chosen by repeatedly training a model and checking how well it performs, effectively integrating the learning algorithm into the selection cycle(Inbar & Avisar, 2024).
- **Embedded methods** (e.g., Least Absolute Shrinkage and Selection Operator, LASSO): These techniques carry out variable selection as part of the model-training routine itself(Inbar & Avisar, 2024).

Importance and Insights:

Studies using these methods consistently find that a small subset of features often account for the majority of a model's predictive power. For example, a study on TSS prediction found that a scenario using only four features (CODe, BOD5e, BOD5i, TN) selected by the SBS method was the most efficient and achieved the highest R2 value(Inbar & Avisar, 2024). Similarly, a comparative study using Monte Carlo Simulation (MCS) and GEP found that TSS was the most influential parameter for both BOD5 and COD estimation, with a 10% increase in TSS leading to approximately a 7.9% increase in both target parameters(Inbar & Avisar, 2024).The ability of these techniques to identify key, non-redundant parameters not only improves model performance but also provides valuable operational insights. For instance, the high importance of TSS in predicting BOD and COD confirms that a major portion of biodegradable material is in particulate form. However, a crucial distinction must be made between correlation and causation. A study on GEP models for COD prediction noted that while ammonia (NH3) was an influential factor, this may be due to a hidden correlation with other toxic compounds rather than a direct chemical link, demonstrating the need for domain expertise to correctly interpret model outputs and avoid erroneous assumptions(Inbar & Avisar, 2024).

Addressing Data Challenges

Data-driven models are heavily reliant on large, clean, and representative datasets. However, real-world WWTP data is often characterized by scarcity, noise, and non-Gaussian distributions, posing significant challenges for modeling(Yang et al., 2024).

Data Scarcity and Quality: The high cost and time-intensive nature of manual laboratory tests and the susceptibility of online sensors to errors and malfunctions often result in incomplete or low-quality datasets(Shahab et al., n.d.). This is a major limitation for many deep learning models

that require vast amounts of data to train effectively and generalize to unseen conditions(Shahab et al., n.d.).

Advanced Techniques for Data Limitations:

- **Transfer Learning (TL):** TL is an emerging technique that addresses data scarcity by transferring knowledge from a model trained on a large dataset (the "source domain") to a new, data-poor task (the "target domain")(Pisa et al., 2021). This can involve using a well-established simulation model like BSM2 as a source and transferring its learned knowledge to a real-world plant with limited data, thereby improving the target model's performance without extensive retraining(Koksal & Aydin, 2024). A key finding is that TL-based controllers can improve performance by 40-99% compared to conventional methods(Pisa et al., 2021).
- **Semi-Supervised Learning (SSL):** This approach combines the use of a small amount of labeled data with a large amount of unlabeled data to train a model(Jia et al., 2025). This is particularly useful in contexts where obtaining labeled data is a significant challenge. For example, an SSL method for identifying microparticles in wastewater was shown to significantly improve detection accuracy with a limited number of labeled images, highlighting its potential for long-term monitoring where manual labeling is unfeasible(Jia et al., 2025).

Discussion: Critical Insights, Challenges, and a Roadmap for the Future

The synthesis of recent literature on AI in wastewater treatment reveals a field of rapid innovation and significant promise. However, it also highlights a number of persistent challenges that must be addressed to enable widespread, real-world adoption. This discussion will provide a critical analysis of the current state, connecting the various findings to form a holistic view of the field's trajectory and outlining a concrete roadmap for future research.

Overarching Trends and Key Findings

The extensive body of work on predictive modeling in WWTPs has consistently shown that no single AI algorithm is universally superior. Instead, the efficacy of a model is determined by a complex interplay of the problem type, data characteristics, and operational objectives.

A clear trend is the shift from traditional machine learning to more sophisticated models. While classical algorithms like ANN, XGBoost, and RF are still widely and effectively used, there is a growing consensus on the superiority of deep learning (DL) architectures, particularly LSTM and GRU, for handling time-series data and capturing the complex, non-linear dynamics of WWTPs(Voipan et al., 2025). These models demonstrate exceptional performance in predicting effluent quality under both stable and dynamic conditions, with GRU often showing better stability during disruptive events like rain or storms(Voipan et al., 2025).

Furthermore, research has moved beyond simple predictions to more integrated and advanced systems. The rise of hybrid and multi-source data fusion models, which combine the strengths of different algorithms or data types, is a notable trend(Y. Xie et al., 2024). This is a direct response to the limitations of single models when faced with a wide range of operational conditions and data sources. Emerging paradigms such as Digital Twins (DTs)(Wang et al., 2024), Explainable AI (XAI) (Sheik et al., 2025), and Reinforcement Learning (RL) (Zhu et al., 2025)are also gaining traction, moving the field towards autonomous control, transparent decision-making, and holistic plant management. This represents a fundamental shift from merely anticipating a problem to learning the optimal solution and implementing it autonomously.

Comparative Analysis Across Studies: Table 1 provides a concise comparison of selected studies, outlining their objectives, methodologies, datasets, main findings, and notable strengths and limitations in wastewater process modelling.

Table 1. Comparative Analysis Across Studies

Paper	Objective	Methods / Algorithms	Dataset	Key Findings	Strengths	Limitations
(Gholizadeh et al., 2024)	Predict effluent Total Suspended Solids (TSS) and evaluate the effect of feature-selection methods.	Artificial Neural Network – Multi-Layer Perceptron (ANN-MLP), k-Nearest Neighbours (KNN), AdaBoost; feature-selection methods (Correlation, Mutual Information, Sequential Backward Selection, LASSO, Tree-Based, Variance Threshold).	Tehran Municipal WWTP, Iran; daily data 2016–2020 (654 samples).	ANN-MLP with Sequential Backward Selection achieved the highest performance (R^2 = 0.80); appropriate feature selection improved accuracy by ~6%.	Demonstrates the importance of feature selection; rigorous validation using k-fold and grid-search cross-validation.	Single WWTP dataset; limited range of algorithms.
(X. Xie et al., 2025)	Develop a multi-output hybrid model for Total Nitrogen (TN), Soluble Nitrate (SNO) and Soluble Oxygen (SO);	Gaussian Mixture Model (GMM) clustering; Hybrid Partial Least Squares – Random Wavelet Neural Network / Multi-	Benchmark Simulation Model No. 2 (BSM2) platform; 364 days, 34 944 sets (15-min intervals).	Clustering-based hybrid model improved performance by 42.17% (RMSSD = 0.6189).	Tackles data fluctuations; combines complementary models; comprehensive validation.	Simulation-only (no real-world data); MRVM is computationally slow on large datasets.

	mitigate single-model degradation via clustering.	Relevance Vector Machine (PLS-RWNN / MRVM).				
(Aghdam et al., 2023)	Predict influent Biochemical Oxygen Demand (BOD_5) and Chemical Oxygen Demand (COD); derive mathematical expressions and identify influential parameters.	Gene Expression Programming (GEP), MLP Neural Network, KNN, Gradient Boosting, Regression Trees, Random Forest; Monte Carlo Simulation.	Seven Hong Kong municipal WWTPs; monthly data 2018–2020.	GEP produced the most accurate results (BOD_5 R^2 = 0.727; COD R^2 = 0.861) and explicit equations.	Provides interpretable mathematical expressions; highlights influential parameters such as TSS.	Monthly (not real-time) data; only municipal plants; potential BOD_5 measurement errors.
(Inbar & Avisar, 2024)	Predict effluent Total Phosphorus (TP) compliance (binary classification); analyse nutrient removal efficiency.	XGBoost, Random Forest, Support Vector Machine, ANN, Long Short-Term Memory (LSTM).	Kfar Saba – Hod Hasharon WWTP, Israel; 11-year daily dataset (1 624 samples).	XGBoost achieved 87% accuracy and 85% precision; Random Forest had highest recall (90%).	Long-term, real-world dataset; explicit focus on regulatory compliance; precision–recall trade-off analysis.	Single WWTP; limited to TP; uncertain generalizability to other climates.
(Zhang et al., 2025)	Predict effluent COD, Ammonia-Nitrogen (NH_3-N), TN and TP; integrate multi-source data; compare deep	Random Forest, MLP, LSTM, Gated Recurrent Unit (GRU); RReliefF for feature importance.	Industrial Effluent Treatment Plant, Anhui Province, China; one-year hourly data (8 689 sets).	Deep learning (LSTM, GRU) outperformed traditional ML; GRU slightly superior for COD/NH_3-N.	Demonstrates multi-source data fusion; shows DL advantages on complex data; includes feature-importance analysis.	Single-site dataset; COD prediction still challenging.

	learning with traditional ML.					
(Pisa et al., 2021)	Apply Transfer Learning (TL) to design WWTP control loops without extensive process knowledge.	LSTM-based controllers with transfer-learning strategies.	Experimental control-loop data (details not fully specified).	TL controllers reduced oscillations and improved Integral Absolute Error by 40–94% and Integral Squared Error by 34–99%.	Dramatically cuts model-design and training time by reusing knowledge; marked improvement over conventional controllers.	Focuses on control loops rather than effluent-quality prediction; future research directions not clearly stated.
(Bøhn et al., 2025)	Identify foundational requirements for data-driven modelling in WWTPs.	Linear models (ElasticNet) and non-linear models (LSTM, Temporal Convolutional Network).	Pilot denitrification reactor, Veas facility, Norway.	Non-linear models fit training data best but linear models generalised better; temperature shifts strongly affected performance.	Emphasises practical issues; publicly shares code and data for reproducibility.	Single pilot reactor; unmeasured factors (e.g. biofilm carrier loss) remain challenging.
(Alvi, 2024)	Review deep-learning methods and applications in wastewater process modelling.	Narrative review of DL models (LSTM, GRU) and mechanistic models.	None (review article).	DL emerging as an alternative to semi-mechanistic models; highlights limited cross-community understanding.	Bridges knowledge gap between two research communities; identifies open research problems.	No original data or model results.

Conclusion

The integration of artificial intelligence and machine learning into wastewater treatment has transitioned from a theoretical concept to a practical and transformative reality. AI models, particularly advanced deep learning architectures like GRU and hybrid systems, have demonstrated a superior capability to predict key effluent parameters and handle the complex, non-linear dynamics of WWTPs. These models act as powerful soft sensors, providing operators with the real-time insights necessary for proactive management, optimized energy consumption, and ensured regulatory compliance.

The field has evolved beyond simply predicting future states. Emerging paradigms such as Digital Twins and Reinforcement Learning are enabling the development of sophisticated prescriptive control systems that can autonomously optimize complex, multi-objective functions. Furthermore, advancements in data-centric strategies like feature selection and techniques like Explainable AI are addressing the critical barriers of data quality and model interpretability, which are essential for building trust and facilitating real-world adoption.

While challenges remaining, including data scarcity, a lack of standardization, and the need for greater interdisciplinary collaboration, the future of AI in wastewater treatment is profoundly promising. Continued research into hybrid and physics-informed models, coupled with an expanded focus on multi-objective optimization and novel applications like resource recovery and predictive maintenance, will pave the way for a new era of intelligent, sustainable, and resilient wastewater management. The transition to a data-driven paradigm is not just about technological advancement; it is about securing a cleaner, more sustainable future for our water resources.

Acknowledgement

I sincerely thank my guide and co-author for their constant support and valuable guidance throughout this work. I am grateful to Jain (Deemed-to-be University) for providing the resources and research environment.

References:

Aghdam, E., Mohandes, S. R., Manu, P., Cheung, C., Yunusa-Kaltungo, A., & Zayed, T. (2023). Predicting quality parameters of wastewater treatment plants using artificial intelligence techniques. *Journal of Cleaner Production*, *405*. https://doi.org/10.1016/j.jclepro.2023.137019

Alvi, M. (2024). *Deep learning for sensing in wastewater treatment plants* [The University of Western Australia]. https://doi.org/10.26182/0n17-k211

Bøhn, E., Eidnes, S., & Jonassen, K. R. (2025). *Machine learning in wastewater treatment: insights from modelling a pilot denitrification reactor*. http://arxiv.org/abs/2412.14030

Bo-Qi, L., Ding-Jie, Z., Yang, Z., & Long-Yu, S. (2025). Comparative analysis of supervised learning models for effluent quality prediction in wastewater treatment plants. *PLOS ONE*, *20*(6 June). https://doi.org/10.1371/journal.pone.0325234

Chen, K.-J., & Kao, C.-M. (2025). Optimization of Municipal Wastewater Treatment Plants Management through Digital Twin Modeling. *Journal of Environmental Engineering*, *151*(4). https://doi.org/10.1061/joeedu.eeeng-8046

Gholizadeh, M., Saeedi, R., Bagheri, A., & Paeezi, M. (2024). Machine learning-based prediction of effluent total suspended solids in a wastewater treatment plant using different feature selection approaches: A comparative study. *Environmental Research*, *246*. https://doi.org/10.1016/j.envres.2024.118146

Inbar, O., & Avisar, D. (2024). Enhancing wastewater treatment through artificial intelligence: A comprehensive study on nutrient removal and effluent quality prediction. *Journal of Water Process Engineering*, *61*. https://doi.org/10.1016/j.jwpe.2024.105212

Jia, T., Yu, J., Sun, A., Wu, Y., Zhang, S., & Peng, Z. (2025). *Semi-supervised learning-based identification of the attachment between sludge and microparticles in wastewater treatment*. https://doi.org/10.1016/j.jenvman.2025.124268

Koksal, E. S., & Aydin, E. (2024). *A Hybrid Approach of Transfer Learning and Physics-Informed Modeling: Improving Dissolved Oxygen Concentration Prediction in an Industrial Wastewater Treatment Plant*. https://arxiv.org/abs/2401.11217

Mahanna, H., ELRahsidy, N., Kaloop, M. R., El-Sapakh, S., Alluqmani, A., & Hassan, R. (2024). Prediction of wastewater treatment plant performance through machine learning techniques. *Desalination and Water Treatment*, *319*. https://doi.org/10.1016/j.dwt.2024.100524

Malviya, A., & Jaspal, D. (2021). Artificial intelligence as an upcoming technology in wastewater treatment: a comprehensive review. *Environmental Technology Reviews*, *10*(1), 177–187. https://doi.org/10.1080/21622515.2021.1913242

Masood, A., Srivastava, A., Hameed, M. M., Saquib Tanweer, M., Ahmad, K., & Razali, S. F. M. (2024). Machine Learning for Water Quality Soft Sensing in Wastewater Treatment: State of the Art and Future Directions. In *Innovative and Hybrid Technologies for Wastewater Treatment and Recycling* (pp. 364–380). CRC Press. https://doi.org/10.1201/9781003454199-17

Omarova, P., Amirgaliyev, Y., Kozbakova, A., & Ataniyazova, A. (2023). Application of Physics-Informed Neural Networks to River Silting Simulation. *Applied Sciences (Switzerland)*, *13*(21). https://doi.org/10.3390/app132111983

Pisa, I., Morell, A., Vilanova, R., & Vicario, J. L. (2021). Transfer learning in wastewater treatment plant control design: From conventional to long short-term memory-based controllers. *Sensors*, *21*(18). https://doi.org/10.3390/s21186315

Shahab, S., Anjum, M., & Sarosh Umar, M. (n.d.). Deep Learning Applications in Solid Waste Management: A Deep Literature Review. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 13, Issue 3). https://dx.doi.org/10.14569/IJACSA.2022.0130347

Sheik, A. G., Kumar, A., Srungavarapu, C. S., Azari, M., Ambati, S. R., Bux, F., & Patan, A. K. (2025). Insights into the application of explainable artificial intelligence for biological wastewater treatment plants: Updates and perspectives. *Engineering Applications of Artificial Intelligence*, *144*. https://doi.org/10.1016/j.engappai.2025.110132

Voipan, D., Voipan, A. E., & Barbu, M. (2025). Evaluating Machine Learning-Based Soft Sensors for Effluent Quality Prediction in Wastewater Treatment Under Variable Weather Conditions. *Sensors*, *25*(6). https://doi.org/10.3390/s25061692

Wang, A. J., Li, H., He, Z., Tao, Y., Wang, H., Yang, M., Savic, D., Daigger, G. T., & Ren, N. (2024). Digital Twins for Wastewater Treatment: A Technical Review. In *Engineering* (Vol. 36, pp. 21–35). Elsevier Ltd. https://doi.org/10.1016/j.eng.2024.04.012

Xie, X., Deng, X., Huang, L., & Ning, Q. (2025). A multi-output hybrid prediction model for key indicators of wastewater treatment processes. *Chemometrics and Intelligent Laboratory Systems*, *258*. https://doi.org/10.1016/j.chemolab.2025.105316

Xie, Y., Chen, Y., Wei, Q., & Yin, H. (2024). A hybrid deep learning approach to improve real-time effluent quality prediction in wastewater treatment plant. *Water Research*, *250*, 121092. https://doi.org/10.1016/j.watres.2023.121092

Yang, C., Guo, Z., Geng, Y., Zhang, F., Wei, W., & Liu, H. (2024). Optimized deep learning models for effluent prediction in wastewater treatment processes. *Environ. Sci.: Water Res. Technol.*, *10*(5), 1208–1218. https://doi.org/10.1039/D3EW00875D

Zamfir, F.-S., Carbureanu, M., & Mihalache, S. F. (2025). Application of Machine Learning Models in Optimizing Wastewater Treatment Processes: A Review. *Applied Sciences*, *15*(15), 8360. https://doi.org/10.3390/app15158360

Zhang, S., Cao, J., Gao, Y., Sun, F., & Yang, Y. (2025). A Deep Learning Algorithm for Multi-Source Data Fusion to Predict Effluent Quality of Wastewater Treatment Plant. *Toxics*, *13*(5). https://doi.org/10.3390/toxics13050349

Zhu, Z., Dong, S., Zhang, H., Parker, W., Yin, R., Bai, X., Yu, Z., Wang, J., Gao, Y., & Ren, H. (2025). Bayesian Optimization-Enhanced Reinforcement learning for Self-adaptive and multi-objective control of wastewater treatment. *Bioresource Technology*, *421*, 132210. https://doi.org/10.1016/j.biortech.2025.132210