Machine Learning Algorithms for Dynamic System Identification in Wastewater Treatment Plant

Manuel Alejandro Ospina Alarcón[1*], Gabriel Elías Chanchí Golondrino[1], Liliana María Úsuga Manco[2]

[1]Faculty of Engineering, Systems Engineering Program, University of Cartagena, 30th, Consulate ave., 39 B-192, 130001 Cartagena, Colombia, mospinaa@unicartagena.edu.co, gchanchig@unicartagena.edu.co

[2]Faculty of Engineering, Chemical Engineering Program, University of Cartagena, 30th, Consulate ave., 39 B-192, 130001 Cartagena, Colombia, lusugam@unicartagena.edu.co

*Correspondence: mospinaa@unicartagena.edu.co

Abstract: A comprehensive study on the application of machine learning algorithms for dynamic system identification in wastewater treatment plants (WWTP) is presented. The research focuses on developing a flexible neural network model to predict the behavior of key variables in the aeration process of a pilot-scale water treatment plant. The methodology involves data collection from experimental trials, data preprocessing, neural network model development, validation, and implementation. The results demonstrate the effectiveness of the proposed approach in accurately predicting key variables such as dissolved oxygen, tank temperature, and tank level (mean squared error MSE=0.166 and coefficient of determination R2=0.967). The discussion highlights the importance of variable selection, data preprocessing techniques, model architecture design, and validation procedures. The conclusions emphasize the significance of machine learning techniques in optimizing wastewater treatment processes, improving energy efficiency, and facilitating real-time decision making. Recommendations for future research include scaling up the model to larger treatment plants, incorporating advanced deep learning techniques, and continuous validation and optimization of the model.

Keywords: wastewater treatment; machine learning; artificial neural networks; dynamic system identification; aeration process

1 Introduction of WWT Challenges

Wastewater treatment (WWT) stands as a critical endeavor in safeguarding environmental integrity and public health [1]. Aeration, a fundamental step within this process, plays a pivotal role by facilitating the removal of contaminants through the transfer of gases to liquids [2]. The efficacy of aeration profoundly influences treatment performance [3] and energy consumption within wastewater treatment plants (WWTP) [4].

Recently, several studies in Central Europe have also highlighted the growing interest in applying machine learning to wastewater treatment processes. For instance, [5] developed an ANN-based control system for aeration processes in Czech wastewater treatment plants, focusing on energy optimization. Similarly, [6] implemented real-time machine learning to improve nitrogen removal efficiency in Polish WWTPs. In Hungary, [7] explored deep learning techniques to enhance sludge treatment strategies. These regional efforts underscore the increasing relevance of data-driven approaches across different European contexts.

Traditional methods for modeling and optimizing WWT processes often rely on complex mathematical models, although they are capable of accurately capturing the dynamic behavior of the system, it is important to acknowledge that they can face significant limitations [8]. These limitations arise due to the inherent complexity of the models [3] and the large number of involved phenomena [9]. On one hand, the development of these models can be an extensive and laborious process, as it requires the inclusion and understanding of multiple variables and relationships between them [10]. This process can take a long time, meaning that obtaining complete and accurate models can be a daunting and costly task. Furthermore, once these models are obtained, their simulation can also be challenging. Running detailed and accurate simulations may require significant computational resources and considerable time. This can be problematic, especially in environments where quick and agile decisions are required, such as in modern industry. Consequently, despite the potential accuracy of these complex models, they may not be the most practical option for quick decision making in modern industry [11]. In such cases, more simplified approaches or agile methods may be preferred to ensure that decisions are made in a timely and effective manner, even if it means sacrificing some degree of precision in the system's behavior.

However, despite these contributions, few studies have concentrated specifically on dynamic system identification using flexible neural network architectures that integrate multiple input-output variables over time. Our study seeks to fill this gap by proposing a predictive ANN-based model tailored to capture the temporal behavior of key process variables in a pilot-scale WWTP.

In recent years, machine learning (ML) algorithms [12], particularly artificial neural networks (ANN) [13], have emerged as powerful tools for modeling complex systems [14] and predicting their behavior [15]. Despite the potential benefits of ML in WWTP, there is a lack of comprehensive studies that explore its application for dynamic system identification in treatment plants [16], [17]. Existing research often focuses on specific aspects of wastewater treatment [8] or employs simplistic models that do not fully capture the complexity of real-world systems [10].

This study aims to address this gap by proposing a novel approach for dynamic system identification in wastewater treatment plants using ANN algorithms. By integrating data from various sensors and process parameters, our proposed methodology seeks to develop accurate models capable of predicting system behavior, optimizing process performance, and facilitating proactive maintenance.

In this paper, we present the results of our investigation into the application of ML algorithms for dynamic system identification in a pilot-scale WWTP [3], [18]. Through rigorous experimentation and analysis, we demonstrate the effectiveness of our approach in improving the efficiency and reliability of WWT processes.

This research not only contributes to the existing body of literature but also presents a paradigm shift in WWT methodology, offering innovative solutions to complex challenges. By harnessing the power of ML, we aim to propel the field towards more efficient and sustainable practices, thereby advancing environmental stewardship and public health.

2 Methodology for Dynamic System Identification

This section outline the method developed for dynamic system identification in WWTPs using machine learning algorithms. Our approach consists of several key steps, including data collection, preprocessing, model selection, training, and validation (see Figure 1).

Two development methodologies were chosen: Kanban methodology and the Iterative Research Pattern (IRP) (see Figure 1). Kanban methodology focuses on continuously seeking improvement for the final product [19], while the IRP involves observations used to generate research questions addressed through a development cycle [20]. The IRP comprises four stages: observation, problem identification, solution development, and validation (see Figure 1). For observation, the aeration process of a WWTP was analyzed, requiring referencing previous studies that used mathematical models with dynamic balance equations containing necessary information [3], [18]. This yielded ample data by varying key parameters. Problem identification involved analyzing aeration process instructions and transferring the mathematical model to a neural network model,

which efficiently identified system dynamics and provided an appropriate solution based on mean squared error (MSE), the coefficient of determination (R^2), and the integral prediction index of absolute time error (ITAE). Subsequently, the ANN model was verified under different conditions to assess its performance and error.

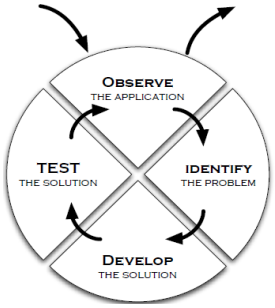


Figure 1
IRP methodology phases [20]

2.1 Observation of Process Variables

During the observation phase, an analysis of data collected from the pilot plant was conducted to determine which variables were essential for inclusion in the ANN model. Additionally, potential relationships between each variable were observed, along with validation of whether the obtained values corresponded to the pilot plant experimentation.

Data were collected from a pilot-scale wastewater treatment plant (see Figure 2), which was previously constructed as part of prior work [3], [18]. This pilot plant and all its instrumentation were detailed in these previous works, including the acquisition of data from the three fundamental variables used in this study: dissolved oxygen, temperature, and liquid level in the process tank.

The process parameters, as well as the plant parts, are thoroughly described in [3], [18]. Below, the input and output (measured) variables involved in the study are detailed. Table 1 delineates the dependent (measured) and independent variables. The input variables remain consistent since both the plant and the ANN are based on the same dataset. The objective is for the outputs to closely resemble each other, with the error tending towards zero.

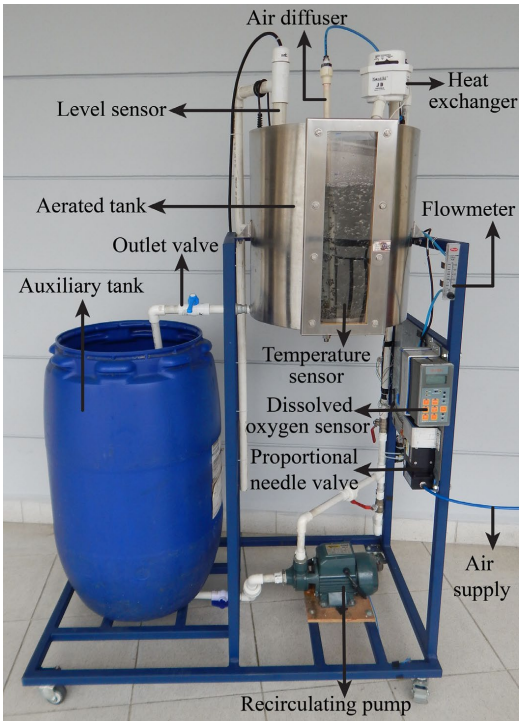


Figure 2
Aeration pilot plant used for data acquisition [3], [18]

Table 1
Dependent variables (measured) and independent variables (inputs)

Variable	Role	Description	Time Dependence
Volumetric flow rate of air	Input	% Bubble injection to aerator	Time dependent
Relay position	Input	Heat exchanger relay control (%)	Time dependent
Valve opening at the outlet stream	Input	Valve V-1 (%)	Time dependent
Fraction of pumping	Input	Pump percentage from auxiliary tank	Time dependent
Oxygen scavenger added	Input	Dosage of oxygen scavenger (g)	Time dependent
Temperature	Output	Aerator tank temperature (°C)	Time dependent
Dissolved oxygen	Output	DO concentration in aerator (mg/L)	Time dependent
Liquid level	Output	Water level in aerator tank (m)	Time dependent

Measurements of various process input variables such as relay position in the heat exchanger (% Relay), the fraction of volumetric flow rate of air (%Bubble), the valve opening at the outlet stream %V-1 , the fraction of pumping from auxiliary tank outlet (%Pump), and the amount of oxygen scavenger added to the reaction medium (g Scavenger) were obtained from sensors installed throughout the plant and recorded at regular intervals (see Figure 3).

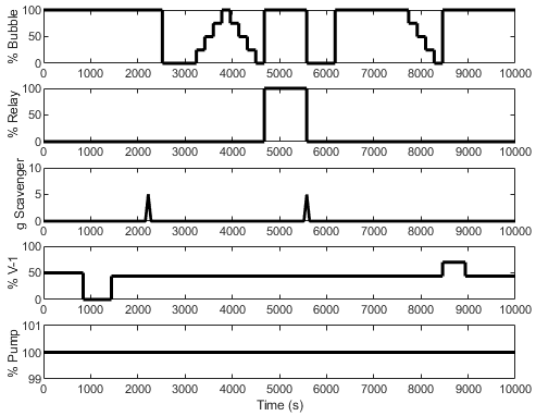


Figure 3

Applied inputs variables for dynamic system identification [3], [18]

From the input data depicted in Figure 3, measurements of the three key output variables, namely aerator tank temperature (T), liquid level in the aerator tank (L), and dissolved oxygen concentration (DO), were derived, thus completing the dataset for dynamic identification via ANN of the WWTP. Figure 4 illustrates the WWTP's response to changes in the input variables from Figure 3.

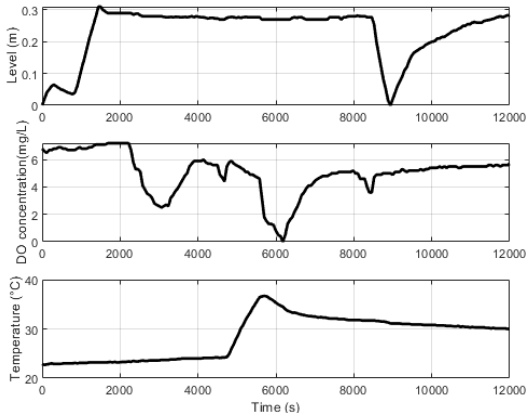
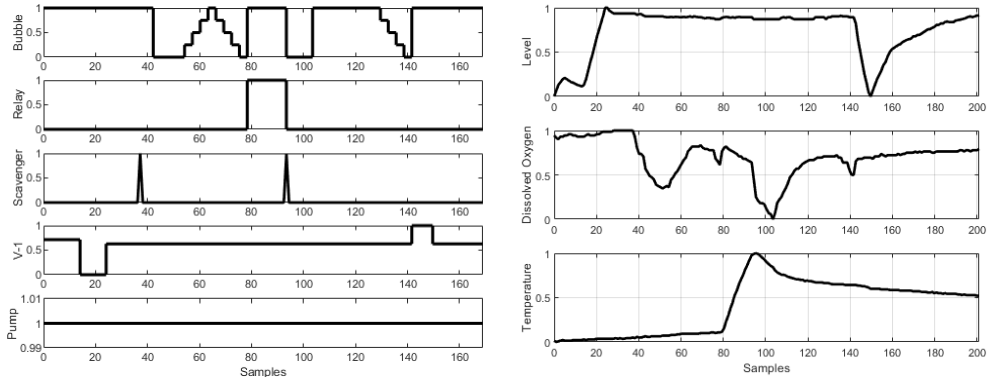


Figure 4

WWTP dynamic response [3], [15]

2.2 Data Cleaning, Selection, and Preprocessing

During this stage, it was determined within the dataset that certain variables provided by the WWTP were unnecessary for inclusion in the ANN model. Furthermore, a data cleaning process was conducted for the variables used in the model to prevent any contamination during model training. This involved handling empty data points, before training the machine learning models, the collected data underwent preprocessing to remove noise (refers to the preprocessing steps applied to improve data quality prior to model training), handle missing values, and normalize features (see Figure 5). Specifically, we used mean imputation to handle missing values, applied feature scaling (normalization) to bring all variables to a common range (0–1), and removed obvious outliers based on visual inspection and standard deviation thresholds. These procedures help reduce the effect of irregular or anomalous readings from sensors, thus improving the robustness of the ANN model.



a) b)

Figure 5
Normalized input (a) and output (b) data in the interval (0 - 1)

2.3 Development and Configuration of the ANN Architecture

In the early phase of model selection, we considered multiple machine learning algorithms including support vector machines (SVM), decision trees, and random forests. While these models offer advantages in interpretability and training speed, they are generally less effective in capturing the complex temporal dependencies and nonlinear relationships present in dynamic wastewater systems. Artificial Neural Networks (ANN) were ultimately selected due to their proven ability to model time-dependent, multivariate processes with high flexibility. Their architecture allows incorporating temporal delays and layered processing, which

are crucial for dynamic system identification tasks (see Figure 6). It was developed for dynamic system identification, leveraging the TensorFlow and Keras libraries in Python. The model architecture consisted of multiple layers of neurons incorporating suitable activation functions (RLu and hyperbolic tangent) and regularization techniques (gradient descent, regressors and adaptive moment estimation - ADAM) to effectively capture complex relationships within the data.

During this phase, the neural network programming took place, involving iterative testing to determine the optimal input data for the ANN. Through trial and error, parameters such as network architecture and error calculations were adjusted, ensuring the model trained equally well on both training and testing data. Additionally, meticulous review was conducted to confirm that the model performed consistently across both sets of data.

According to Figure 6, the utilization of regressors in data processing for prediction models is a method involving shifting data backward by a specified number of steps. This facilitates the creation of a dataset that incorporates historical and significant information for models reliant on the memory of previous events, as commonly encountered in time series analysis.

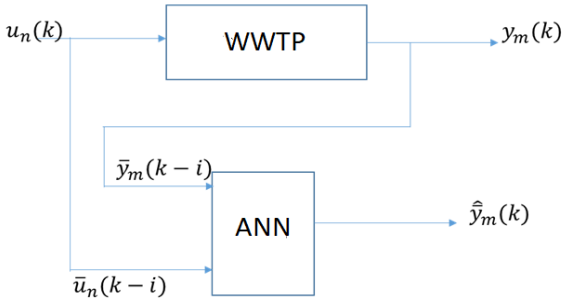


Figure 6
ANN model implemented

Incorporating temporal delays in neural networks is pivotal for analyzing sequences and time-varying data, as it enables the network to consider historical information in predictions and decision making. By manually deciding on the delay configuration in variables, rather than opting for TDNN (Time-Delay Neural Network) architectures, greater flexibility is achieved in tailoring the data to the model's requirements [21].

Compared to recent machine learning-based solutions in Central Europe, such as the ANN controller for aeration in Czech WWTPs [5], or the ML-driven nitrogen removal monitoring system in Poland [6], our model presents a novel approach by focusing on dynamic system identification using historical regressors and ANN time-delay configuration. Additionally, unlike the deep learning application for sludge treatment proposed in Hungary [7], our model captures a broader process scope, integrating variables like DO concentration, tank level, and temperature

within a single predictive framework. This distinction underlines the originality of our approach and reinforces its potential for enhancing decision making and predictive maintenance in wastewater treatment operations.

2.4 Model Testing and Evaluation Strategy

During the test phase, a thorough examination was conducted to ensure that the model did not suffer from overfitting, and that the graphical representation of the ANN output aligned with expectations. This entailed scrutinizing whether the predicted values closely approximated the actual values of the output variables (measured variables, L, T, and DO), thereby confirming the model's accuracy. Additionally, measures were taken to prevent the model from becoming overly tailored to the training data, which could compromise its ability to generalize to unseen data. The validation process aimed to validate the model's performance and verify its reliability in producing predictions consistent with real-world outcomes.

2.4.1 ANN Training Phase: Data Splitting and Optimization Techniques

The selected ANN model was trained using a subset of the collected data (60%, with the remaining data reserved for validation (20%) and testing (20%). During training, techniques such as gradient descent optimization, batch normalization, and early stopping to prevent overfitting and improve convergence were employed.

2.4.2 Validation Metrics and Testing Methodology

The trained model was validated using the reserved validation dataset to assess its predictive accuracy and generalization ability. We evaluated performance metrics such as mean squared error (MSE) expressed as in (1), the coefficient of determination (R2) as in (2), and the integral prediction index of absolute time error (ITAE) expressed as in (3) to quantify the model's performance in terms of the three available data sources: level, temperature, and dissolved oxygen concentration.

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

The MSE presented in (1) is a commonly used metric for assessing the accuracy of a model's predictions as ANN. It is calculated as the average of the squared differences between the n predicted values $\hat{y}_i$ and the n measurement values y_i. This measure is particularly useful in dynamic systems identification problems, where the goal is to minimize the discrepancy between the model's predictions and the observed data, providing a quantitative way to evaluate its performance. A lower MSE value indicates higher accuracy in the model's predictions.

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

The R^2 coefficient as in (2), is a statistical measure that represents the proportion of the variance in the measurement variables y_i that are predictable from the estimated output variables $\hat{y}_i$ in an ANN model. It ranges from 0 to 1, where a value closer to 1 indicates a better fit of the model to the data.

The prediction index ITAE in (3) is a performance metric used in dynamic systems identification to evaluate the transient response of a process. Unlike traditional error measures, ITAE places more emphasis on reducing the integral of the absolute time-weighted error over a specified time interval. It is particularly useful for systems where minimizing the duration and magnitude of error is crucial [3].

$$ITAE = \int_{t_0}^{t_f} t(y_i - \hat{y}_i)\,dt \quad (3)$$

where t is the time, t_0 and t_f are the initial and final times, respectively, y_i is the i-th measurement, and $\hat{y}_i$ is the model prediction for that measurement.

In the case of the WWTP, these three metrics (1), (2), and (3) must be calculated for the three measured variables: L, T, and DO in the aerated tank. These metrics are essential for assessing the performance and efficiency of the treatment processes in wastewater management systems.

3 Results and Discussion: Accuracy and Robustness

A preliminary ANN scheme was designed (see Figure 7), which would later be implemented in code. Phyton and Google Cola were chosen as the development and training environment for its collaborative features and access to GPU resources. This facilitated the seamless transition from conceptualization to practical implementation, allowing for efficient development and training of the neural network model.

This choice presented in Figure 7 allows for better portability of the model, eliminating the need for specialized hardware for training the ANN. Google Colab's provision of GPU-equipped servers enables model training without consuming local or remote resources. Constructing the model involved selecting existing data analysis and machine learning tools. Python was chosen as the primary programming language due to its extensive libraries. Pandas facilitated dataset manipulation from Excel files, while NumPy handled matrix operations required in machine learning algorithms. Scikit-Learn, specialized in machine learning algorithms, validated the model's test data. Matplotlib facilitated result visualization by plotting the ANN outputs. Tensorflow libraries were used for

ANN training, with Keras serving as a high-level API tailored for ANN construction. This comprehensive toolset formed the foundation of the project's development framework.

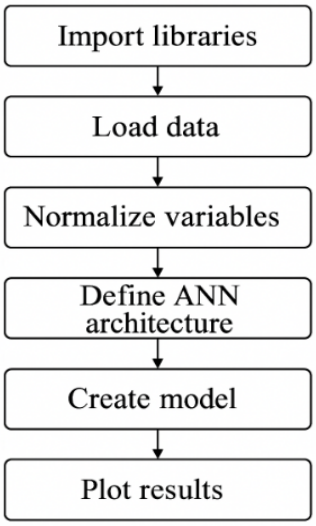


Figure 7
Sequence diagram of ANN construction

Starting from Figure 7, the model development follows a logical sequence, initiating with the importation of necessary libraries and the creation of the regressors function. Subsequently, data loading occurs, and input and output (measured) variables are defined, followed by their conversion into normalized arrays. The ANN architecture is then designed using Keras. Once the structure is defined, K-fold cross-validation is performed to assess the model's prediction fit. Finally, using libraries, the predicted variables are visualized and compared with the measured variables through respective graphs, marking the culmination of the model process.

3.1 Regressor Function and Time Dependency Handling

As shown in Figure 6, The regressor function was implemented to adjust the temporal alignment of variables (input and output) consisting of historical data, indicating their time-dependence. While time itself isn't used in model training, it governs the sequence and order of the data. The regressor function was necessary to shift the values of each variable array backward in time by a defined number of positions, denoted by the variable 'num_delay', which in this case is set to two time lags. The regressor function was implemented using Python and NumPy to shift time series data, allowing the ANN to incorporate temporal dependencies.

The function introduces a delay in the input and output variables, which is critical in dynamic system identification tasks.

Applying this adjustment fixes the initial values to zero, with subsequent data points reassigned to maintain the correct sequence. Figure 8 shows the second regressor of the data along with their respective original value for the sum input (Bubble, Relay, and Scavenger) variables, and the first regressor to all output (Level, DO, and Temperature) variables presented in Figures 3, 4, and 5. These regressors are adjusted based on the data's temporal behavior as depicted in Figure 8.

Incorporating information on data behavior over time enhances the predictive model's accuracy, particularly evident in measured variables such as dissolved oxygen, tank level, and tank temperature. The effectiveness of the regressors function is demonstrated by the alignment between original (measured) variables and estimated regressors values (see Figure 8).

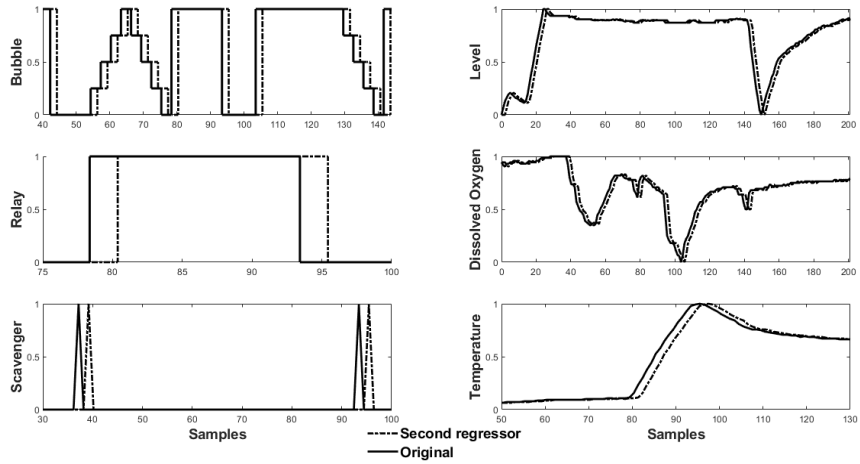


Figure 8
Variables with a regressor function

3.2 ANN Model Configuration Using Keras Framework

The ANN model was built using TensorFlow and Keras. It consists of four layers: three hidden layers with ReLU activation and one output layer using a hyperbolic tangent function. The Adam optimizer was applied for training, and mean squared error (MSE) was used as the loss function. This architecture was selected to ensure high performance in modeling nonlinear and time-dependent relationships. The consistency of the network with multiple layers enables the model to learn patterns and relationships among the data, thereby enhancing prediction accuracy.

ReLU activation function was applied to the hidden layers, chosen for its precision in regression problems and its ability to handle and mitigate the issue gradient decrease vanishing in the network.

As the optimization method, the Adam algorithm was chosen over the traditional gradient descent due to its ability to automatically adjust the learning rate for each parameter, thereby enhancing the effectiveness of the process. Adam was used in conjunction with the mean squared error (MSE) as the loss function, as it imposes a more severe penalty on erroneous predictions compared to MSE by assigning different weights to errors. This results in a more reliable and accurate model.

3.3 Validation through Visual and Quantitative Methods

During the validation phase, the model's robustness was assessed by analyzing its predictive capacity and determining whether it optimally fit the data or exhibited overfitting. Iterations and modifications were made during the development stage to enhance the model's accuracy. This was achieved by using graphs that compared ANN predictions with measured variables and by observing key metrics such as (1), (2), and (3). These evaluations provided insights into the model's performance and guided adjustments to optimize its predictive capability.

Detecting overfitting through validation with test data was necessary to ascertain the model's ability to accurately predict new data and verify its adequacy in prediction. The K-fold method, facilitated by Scikit-Learn, was chosen for this purpose. It partitions the dataset into subsets, iteratively training and evaluating on one subset as the test set and the remaining subsets as the training set. This process is repeated for each subset, allowing for the comparison and averaging of errors across subsets. Low error indicates that the model does not suffer from overfitting.

3.3.1 Graphical Comparison of Measured and Predicted Variables

Once the model was trained, the accuracy of the ANN's predictions regarding the actual data of the measured variables was examined. Figures 9, 10, and 11 illustrate the comparison between measured values and predictec values by the ANN model for three distinct variables: dissolved oxygen (DO) (see Figure 9), tank temperature (see Figure 10), and tank level (see Figure 11). In each graph, the lines depict the time series of measured data (in black) versus the model estimates (in red). The common feature across these graphs is the notable consistency between predictions and real values, indicating the model's high accuracy in its estimations for different variable types. This level of precision suggests that the model adequately captures the underlying dynamics of the WWTP.

From Figures 9, 10 and 11, the model achieved high prediction accuracy for variables such as dissolved oxygen (DO) concentration, tank temperature, and

water level, which are critical for optimizing treatment processes. The model's ability to accurately forecast these variables contributes to improve operational efficiency and resource utilization in WWTP.

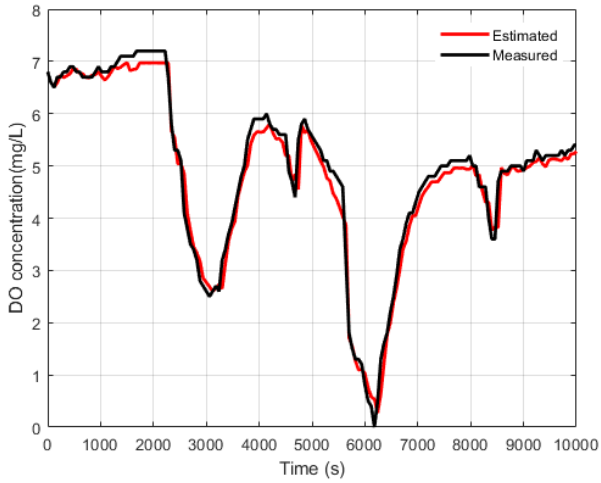


Figure 9
Comparison of measured and estimated DO

Through dynamic system identification (see Figures 9, 10, and 11), the model successfully captured the complex temporal dynamics inherent in wastewater treatment processes. By analyzing time-series data from sensors installed throughout the plant, can be identified patterns and trends that influence system behavior, enabling better understanding and control of the treatment process.

While the performance of the ANN approach was thoroughly evaluated, direct comparisons with traditional baseline methods like linear regression and time-series analysis were not conducted in this research. Instead, the focus remained on assessing the predictive capability and robustness of the machine learning model independently. The results obtained from Figures 9, 10, and 11, the qualitative comparison in the figures underscored the effectiveness of the machine learning model in enhancing WWTP operations, showcasing its potential for real-world applications.

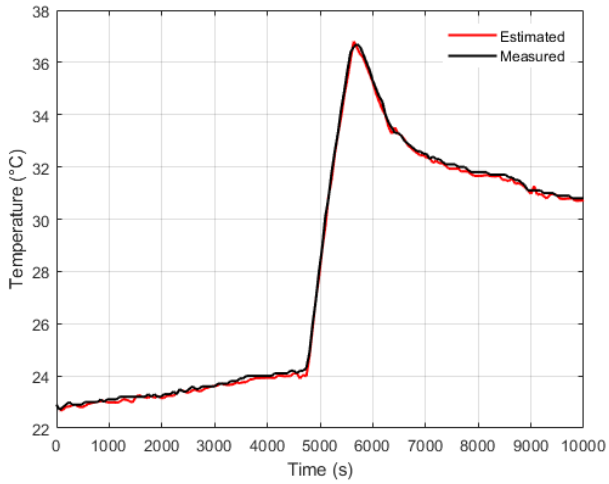


Figure 10
Comparison of measured and estimated temperature

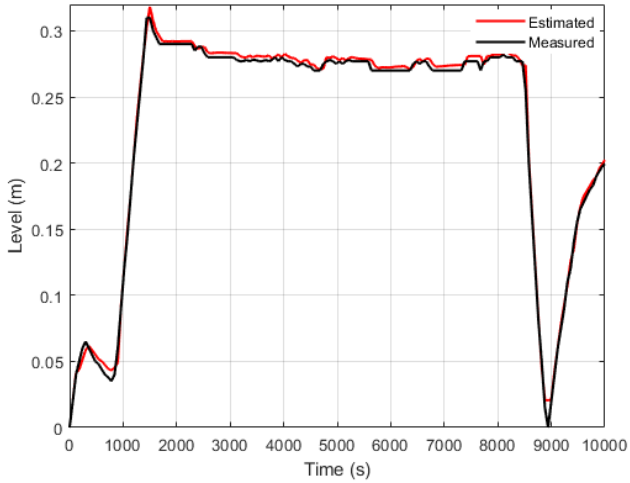


Figure 11
Comparison of measured and estimated level

3.3.2 Quantitative Evaluation Using MSE and R^2 Metrics

To evaluate model performance, we used standard functions from the Scikit-learn library to compute the Mean Squared Error (MSE) (1) and the coefficient of determination (R^2) (2) between predicted and observed values. These metrics provided a quantitative assessment of model accuracy and fit. The MSE quantifies

the discrepancy between actual values and those predicted by the model, with lower values indicating higher prediction accuracy. The R^2 denotes the proportion of variability in the data values explained by the model, with a higher value (approaching 1) on a scale of 0 to 1 indicating a greater likelihood of explaining data variability and better model fit. For the implemented ANN, satisfactory values were obtained for both MSE and R2 (see Table 2).

Table 2 summarizes the model's performance in terms of accuracies for the measured variables. The values of MSE and R^2 indicate a high-level of accuracy. For the dissolved oxygen variable, as well as for tank temperature and level, the R^2 values approach unity, signifying high robustness in the predictability and adaptability of the model.

Table 2
Comparison of EMC and R2 calculation for the three measured variables

Measured Variable	MSE	R^2
DO concentration	0.109	0.948
Temperature	0.043	0.997
Level	0.261	0.996

Based on (1) and (2), the calculation function is capable of individually computing each prediction of the measured output variables, as well as performing the combined calculation of all three variables. This feature allows for a granular analysis of each predicted output variable's performance while also providing a comprehensive assessment of the model's overall predictive accuracy. By offering the flexibility to assess individual predictions and their collective impact, this capability enhances the model's interpretability and facilitates targeted adjustments or interventions as needed. Additionally, the ability to aggregate predictions enables a holistic evaluation of the model's efficacy in capturing system dynamics and predicting outcomes across multiple variables simultaneously. This comprehensive functionality underscores the versatility and robustness of the calculation function in supporting informed decision making and optimizing system performance.

The developed ANN model demonstrated robust performance in predicting key process variables within the WWTP. Evaluation metrics such as mean squared error (MSE) and R-squared coefficient indicated strong agreement between predicted and observed values, validating the model's effectiveness.

3.3.3 ITAE-Based Assessment and Comparison with Literature Models

The ITAE metric (3) was computed using the trapezoidal rule (via the trapz function from SciPy) to assess the model's transient error response. This helped compare the ANN model against existing literature, particularly for dynamic behavior in dissolved oxygen concentration [3], [18]. This approach was adopted

to validate the data and theories underlying semi-empirical models of phenomenological bases.

The ITAE calculated from the ANN model's predictions aligns with the data reported in the literature (refer to Table 3). It's important to note that while the ANN model doesn't fully describe or interpret the phenomenon involved in WWTP, it can serve for quick decision making and the implementation of efficient intelligent control strategies in such processes.

Table 3
ITAE predictions with literature

Measured Variable	ITAE	Best ITAE in [3]	Best ITAE in [18]
DO concentration	$1.432x10^4$	1.50×10^4	$1;41x10^4$
Temperature	$7.130x10^4$	not reported	not reported
Level	$2.653x10^3$	not reported	not reported

The comparison of ITAE values between the ANN model and phenomenological based models [3] and [18] in Table 3 provides valuable insights into the predictive performance of the model. Despite its limitations in fully elucidating the complex dynamics of WWTP, the neural network model demonstrates its utility in the deployment of effective control strategies. By leveraging historical data and leveraging its predictive capabilities, the model offers a practical tool for enhancing operational efficiency and optimizing resource allocation within WWTPs. By examining the importance of model characteristics and coefficients, valuable information was obtained about the underlying mechanisms driving the dynamics of the process.

Despite its effectiveness, the model has certain limitations, including the need for extensive data preprocessing and computational resources for training. Additionally, future research could explore the integration of advanced machine learning techniques, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), to further improve predictive performance and accommodate nonlinear relationships.

The research results highlight the potential of machine learning algorithms in particular the ANN for dynamic system identification in WWTP. By leveraging advanced modeling techniques and real-time data analysis, operators can optimize plant performance, minimize energy consumption, and improve environmental sustainability.

Conclusions

This study applied artificial neural networks (ANN) for dynamic system identification in a pilot-scale wastewater treatment plant (WWTP). The proposed model demonstrated high accuracy in predicting key process variables—dissolved oxygen concentration, tank temperature, and liquid level—based on time series data collected from sensor measurements.

By leveraging historical input-output relationships and incorporating temporal delays, the ANN model captured nonlinear dynamics of the aeration process with promising performance metrics (MSE = 0.166; R^2 = 0.967). These results support the viability of machine learning as a tool for improving process understanding, operational efficiency, and decision-making in WWTPs.

Compared to traditional modeling approaches, the ANN offers advantages in flexibility, adaptability, and reduced need for extensive physical modeling. However, the model requires significant data preprocessing and training, which presents some limitations in real-time deployment scenarios.

Future research may explore the integration of recurrent neural networks (RNN), reinforcement learning for control applications, and model scalability to full-scale plants. Additionally, incorporating more process variables and testing under varying operational conditions could enhance model generalizability.

Acknowledgements

The authors wish to express their sincere gratitude to the University of Cartagena-Colombia for their invaluable support throughout the development of this research. Additionally, we extend our appreciation to Christian Zuluaga, Maribel Ruiz, and José García for generously providing the data necessary for conducting the research and facilitating comparisons with previous studies. Their assistance was instrumental in the successful completion of this study.

References

[1] J. K. Bwapwa, N. Mkhize, and M. Seyam, "Evaluation of operational efficiency and performance for a water treatment plant," *South African J. Chem. Eng.*, 2024, doi: 10.1016/j.sajce.2024.04.003

[2] J. Li, Z. Wang, and Y. Wang, "Integrating membrane aerated biofilm reactors with biological nitrogen removal processes: A new paradigm for achieving sustainable wastewater treatment plants," *Chem. Eng. J.*, Vol. 475, p. 146025, 2023, doi: 10.1016/j.cej.2023.146025

[3] C. Zuluaga-Bedoya, M. Ruiz-Botero, M. A. Ospina-Alarcón, and J. Garcia-Tirado, "A dynamical model of an aeration plant for wastewater treatment using a phenomenological based semi-physical modeling methodology," *Comput. Chem. Eng.*, Vol. 117, pp. 420-432, 2018, doi: 10.1016/j.compchemeng.2018.07.008

[4] Y. Q. Wang *et al.*, "Machine learning framework for intelligent aeration control in wastewater treatment plants: Automatic feature engineering based on variation sliding layer," *Water Res.*, Vol. 246, p. 120676, 2023, doi: 10.1016/j.watres.2023.120676

[5] M. Andreides, P. Dolejš, and J. Bartáček, "The prediction of WWTP influent characteristics: Good practices and challenges," *J. Water Process Eng.*, Vol. 49, p. 103009, Oct. 2022, doi: 10.1016/j.jwpe.2022.103009

[6] J. Drewnowski, B. Szeląg, F. Sabba, M. Piłat-Rożek, A. Piotrowicz, and G. Łagód, "Innovations in Wastewater Treatment – Harnessing Mathematical Modeling and Computer Simulations with Cutting-Edge Technologies and Advanced Control Syste," *J. Ecol. Eng.*, Vol. 24, No. 12, pp. 208-222, 2023, doi: 10.12911/22998993/173076

[7] K. Szatmári, T. Chován, S. Németh, and A. Kummer, "How to support decision making with reinforcement learning in hierarchical chemical process control?," *Chem. Eng. J. Adv.*, Vol. 22, p. 100753, May 2025, doi: 10.1016/j.ceja.2025.100753

[8] Z. Shang *et al.*, "Direct and indirect monitoring methods for nitrous oxide emissions in full-scale wastewater treatment plants: A critical review," *J. Environ. Manage.*, Vol. 358, p. 120842, 2024, doi: 10.1016/j.jenvman.2024.120842

[9] S. Kumari, J. Chowdhry, and M. Chandra Garg, "AI-enhanced adsorption modeling: Challenges, applications, and bibliographic analysis," *J. Environ. Manage.*, Vol. 351, p. 119968, 2024, doi: 10.1016/j.jenvman.2023.119968

[10] P. Sanchis-Perucho, J. Harmand, A. Feddaoui-papin, D. Aguado, and Á. Robles, "Building a simple multivariable filtration model to predict irreversible fouling when directly filtering municipal wastewater," *J. Environ. Chem. Eng.*, Vol. 12, No. 3, p. 112653, 2024, doi: 10.1016/j.jece.2024.112653

[11] S. S. Gill *et al.*, "Modern computing: Vision and challenges," *Telemat. Informatics Reports*, Vol. 13, p. 100116, 2024, doi: 10.1016/j.teler.2024.100116

[12] S. Sahu, A. Kaur, G. Singh, and S. Kumar Arya, "Harnessing the potential of microalgae-bacteria interaction for eco-friendly wastewater treatment: A review on new strategies involving machine learning and artificial intelligence," *J. Environ. Manage.*, Vol. 346, p. 119004, 2023, doi: 10.1016/j.jenvman.2023.119004

[13] J. Luo *et al.*, "Prediction of biological nutrients removal in full-scale wastewater treatment plants using H2O automated machine learning and back propagation artificial neural network model: Optimization and comparison," *Bioresour. Technol.*, Vol. 390, p. 129842, 2023, doi: 10.1016/j.biortech.2023.129842

[14] X. Wang, M. Daryapour, A. Shahrabadi, S. Pirasteh, and F. Razavirad, "Artificial neural networks in predicting of the gas molecular diffusion coefficient," *Chem. Eng. Res. Des.*, Vol. 200, pp. 407-418, 2023, doi: 10.1016/j.cherd.2023.10.035

[15] B. M. Maurya *et al.*, "Artificial intelligence and machine learning algorithms in the detection of heavy metals in water and wastewater: Methodological and ethical challenges," *Chemosphere*, Vol. 353, p. 141474, 2024, doi: 10.1016/j.chemosphere.2024.141474

[16] M. Khalil, A. AlSayed, Y. Liu, and P. A. Vanrolleghem, "Machine learning for modeling N2O emissions from wastewater treatment plants: Aligning model performance, complexity, and interpretability," *Water Res.*, Vol. 245, p. 120667, 2023, doi: 10.1016/j.watres.2023.120667

[17] M. Alvi *et al.*, "Deep learning in wastewater treatment: a critical review," *Water Res.*, Vol. 245, p. 120518, 2023, doi: 10.1016/j.watres.2023.120518

[18] M. Ruiz Botero, "Modelo semifísico de base fenomenológica de la transferencia de oxígeno para el tratamiento de aguas residuales en una planta piloto aireada por difusión," Instituto Tecnológico Metropolitano, 2019 [Online] Available: https://repositorio.itm.edu.co/handle/20.500.12622/1400

[19] K. Ahmad, "Leadership and work motivation from the cross cultural perspective," *Int. J. Commer. Manag.*, Vol. 19, No. 1, pp. 72-84, Mar. 2009, doi: 10.1108/10569210910939681

[20] K. S. Pratt, "Design Patterns for Research Methods: Iterative Field Research," in *AAAI Spring Symposium: Experimental Design for Real*, 2009, pp. 1-7

[21] M. A. Ospina-Alarcón, L. M. Úsuga-Manco, and G. E. Chanchí-Golondrino, "Particle Motion in Jigs Using Linear and Nonlinear Empirical Models," *ARPN J. Eng. Appl. Sci.*, Vol. 17, No. 12, pp. 1280-1287, 2022