applied sciences

MDPI

Article

Machine Learning Methods for the Prediction of Wastewater Treatment Efficiency and Anomaly Classification with Lack of Historical Data

Igor Gulshin [1,*] and Olga Kuzina [2]

[1] Research and Education Centre "Water Supply and Wastewater Treatment", Moscow State University of Civil Engineering, 26, Yaroslaskoye Highway, Moscow 129337, Russia
[2] Department of Information Systems, Technologies and Automation in Construction, Moscow State University of Civil Engineering, 26, Yaroslaskoye Highway, Moscow 129337, Russia
* Correspondence: gulshinia@mgsu.ru

Abstract: This study examines an algorithm for collecting and analyzing data from wastewater treatment facilities, aimed at addressing regression tasks for predicting the quality of treated wastewater and classification tasks for preventing emergency situations, specifically filamentous bulking of activated sludge. The feasibility of using data obtained under laboratory conditions and simulating the technological process as a training dataset is explored. A small dataset collected from actual wastewater treatment plants is considered as the test dataset. For both regression and classification tasks, the best results were achieved using gradient-boosting models from the CatBoost family, yielding metrics of SMAPE = 9.1 and ROC-AUC = 1.0. A set of the most important predictors for modeling was selected for each of the target features.

Keywords: machine learning algorithms; wastewater treatment; effluent quality; soft sensors

check for updates

Citation: Gulshin, I.; Kuzina, O. Machine Learning Methods for the Prediction of Wastewater Treatment Efficiency and Anomaly Classification with Lack of Historical Data. *Appl. Sci.* **2024**, *14*, 10689. https://doi.org/10.3390/app142210689

Academic Editor: George Besseris

Received: 15 October 2024
Revised: 11 November 2024
Accepted: 13 November 2024
Published: 19 November 2024

1. Introduction

Wastewater treatment is a critical human activity, directly influencing the sustainable development of a comfortable living environment. It is essential to all types of industrial processes and is a key infrastructure component in communities of any size.

Despite significant capital investments, consistent improvements in wastewater treatment quality are not always achieved. Recent studies on existing treatment facilities reveal that inadequate performance is often linked to operational errors. Modern wastewater treatment plants are equipped with sufficient automation systems, enabling the control of complex processes under normal operating conditions. However, because wastewater treatment relies primarily on biological treatment facilities, which function as finely tuned bioreactors, operating under quasi-stationary conditions, even minor deviations in one of the many technological parameters can lead to emergencies. The shortage of qualified personnel in the industry often results in delayed responses to emerging technological issues, causing instability in the biological system, a decline in treatment quality, and potential system failure. Such situations pose direct risks of technological, environmental, and sanitary disasters, leading to substantial financial losses for both the state and operating organizations.

In recent years, research on the predictive modeling of wastewater treatment plant operations has gained considerable attention worldwide. These studies aim to forecast treated wastewater quality and to predict abnormal and emergency situations at treatment facilities. Various model classes are being explored to achieve these objectives. Regression models focus on implementing "soft sensors"—models designed to replace physical sensors at treatment plants to reduce costs. Zhang [1] used machine learning models to refine rates of nitrogen and phosphorus removal from wastewater, using activated sludge species composition as input parameters. The XGBoost model demonstrated a determination coefficient (DC) of over 0.8. For the dataset, the authors suggest using 16S rRNA sequencing. This study is notable for its approach to predictor selection, as similar studies have not been conducted previously due to insufficient data. However, a drawback of this approach is its technical complexity, as very few wastewater treatment plants are equipped to perform regular PCR tests for detecting indicator microorganisms. For broader practical relevance, it is recommended to consider more accessible characteristics as predictors.

Wang [2] investigated the impact of operational characteristics at the Umeå wastewater treatment plant (Sweden) on the efficiency of suspended solids and orthophosphate removal. The dataset included 105,763 entries across 32 predictors. The primary outcome of the study was a Variable Importance Measure (VIM) analysis, which allowed the authors to form several hypotheses regarding technological dependencies at the plant. A limitation of this approach is that the model, based on such assumptions, functions as a "black box", presenting certain challenges during the technological development phase. Nevertheless, the authors address the critical issue of time lags associated with treatment processes and their effect on model performance. Approaches to handling wastewater treatment duration at various stages should be adopted in future research. For example, a similar approach is used in the work of Xu [3], which focuses on developing a new Long Short-Term Memory (LSTM) model. In this study, a model was developed to predict wastewater treatment quality using a "soft sensor", based on data from inexpensive and limited physical sensors, while accounting for the time lag in wastewater treatment duration. The Mean Absolute Percentage Error (MAPE) for total nitrogen removal in this model was 2.3%, significantly lower than the results from the traditional multivariate models tested on the same data by the authors. However, it should be noted that the model's performance is highly dependent on the selection and quality of the predictors [4]. Regarding LSTM, this recurrent neural network has proven effective in time series forecasting, as demonstrated by Recio-Colmenares [5], who applied it to predict two substrate concentrations in the synthetic ASM1 model, achieving a MAPE of 1.31%. Nevertheless, LSTM models have several limitations; they are prone to overfitting with small datasets and are sensitive to data outliers. Despite these challenges, many researchers currently approach the task of predicting wastewater treatment quality as a time series forecasting problem. El-Rawy [6] applied the Deep Learning Time Series Forecasting (DLTSF) approach to predict key parameters of treated wastewater at the El-Berka treatment plant in Egypt. The Root Mean Squared Error (RMSE) for total nitrogen was 1.92, which is considered a fairly accurate prediction. Singh [7] presented a comprehensive model comparing the performance of various regression models—Artificial Neural Network (ANN), Fuzzy Logic (FL) algorithms, Random Forest (RF), and LSTM—for predicting treated wastewater parameters. The models were evaluated using metrics such as RMSE, MAPE, Mean Squared Error (MSE), and the Determination Coefficient (DC). The controlled parameters included organic pollutants, nutrients, suspended solids, and heavy metals.

According to the reviewed sources, current models can predict treated wastewater quality with a MAPE of around 1.0%, which displays a remarkably high accuracy for machine learning models. The most accurate model identified was the Outlier Robust Extreme Learning Machine (ORELM), a modification of the standard Extreme Learning Machine (ELM), specifically adapted to handle significant data outliers. The study confirms the presence of numerous data outliers in wastewater treatment studies, which is associated with the high variability of the processes. When developing models, it is essential to implement measures that mitigate the impact of outliers on modeling results. ORELM has been increasingly used for optimizing drinking and wastewater treatment processes, likely due to the unique distribution characteristics of the input data [8–11].

Zaghloul and Achari [12] applied an ensemble method to predict 15 operational parameters of wastewater treatment plants, including biomass properties. The models were trained using data collected over a 10-year period. The authors proposed a six-stage framework in which subsequent parameters were predicted based on previously forecasted ones. According to the authors, the symmetric mean absolute percentage error (SMAPE) using this approach was 7.5%. Another advantage of the multi-stage approach is the reduction in the "black box" effect, which is common in other data-driven models. The models used in this study included Artificial Neural Networks (ANN), Adaptive Neuro-Fuzzy Inference System (ANFIS), and Support Vector Regression (SVR). The authors raised an important issue regarding the development of feature generation principles in wastewater quality models. In this study, intermediate features related to the sludge dose (TSS) in various parts of the bioreactor were generated at different stages, though this alone was insufficient to significantly improve model quality. Simple sequential transformations of predictors increase the risk of model noise (or data leakage), which greatly reduces model performance using test data.

However, using forecasted values for intermediate calculations can enhance model accuracy. In such cases, selecting the appropriate calculation model is crucial. For example, Xu [13] integrated machine learning approaches into the calculations of the Activated Sludge Model #3 (ASM3). The authors aimed to address the limitations of the mechanistic model, particularly those associated with the high variability of incoming wastewater. A machine learning model was employed to predict the biodegradability of wastewater during the treatment of petrochemical effluents. As a result of the study, two models were developed, MLR-ASM3 (based on multiple linear regression) and DF-ASM3 (based on decision forests). The hybrid model achieved MAPE values of less than 25%, which represents a relatively high performance compared to classical ASM3 modeling. This hybrid modeling serves as a practical example of combining "white-box" approaches with high-performance, data-driven computations. An important aspect of regression tasks in wastewater treatment is the optimization of energy consumption. Alali [14] conducted a comparative analysis of 23 models for predicting energy consumption at wastewater treatment facilities in Melbourne. Optimizing energy consumption is a crucial task, as estimates suggest that approximately 7% of global electricity consumption is attributable to the pumping and treatment of wastewater [15]. The time lag used in the model was set to 1 day. The most effective model was the gradient-boosting model XGBoost, which achieved an RMSE metric of approximately 12%.

It is noteworthy that the study was conducted using open data, thus the issues of data collection and preparation were not addressed. However, the challenge of obtaining a representative sample for training and validating the model remains one of the primary issues that must be addressed when employing machine learning approaches in the modeling of wastewater treatment processes. The issue of high-frequency data collection of wastewater characteristics is addressed in the work of Asadi [16]. This study focuses on machine learning algorithms for optimizing the operation of aeration equipment. A total of 35 input parameters were utilized, of which only a few (such as dissolved oxygen concentration) could be collected in real-time. Nonetheless, the optimization of equipment operation resulted in a 31% reduction in energy consumption, which is a significant achievement. Increasing the sample size across all predictors could further enhance this outcome.

This issue is also discussed in the study by Asami [17]. Various models, including Artificial Neural Networks (ANN) and the M5 model tree, were considered, yielding satisfactory metrics. However, despite attempts to fine-tune the models, the primary limitation for improving predictive quality is the insufficient amount of representative data for training. This lack of data is primarily associated with inadequate funding for the operation of treatment facilities. In many cases, automatic analytical sensors are absent at the plants, and the quality control of treated wastewater is performed manually.

The issue of data collection in the absence of sufficient technical resources is addressed in the study by Safder [18], who trained over 20 models primarily based on Artificial Neural Network (ANN) architecture. The achieved metric values were relatively high; however, the data collection and forecasting structure rely on time series analysis, which necessitates a large number of continuously operating sensors. Consequently, models that reduce noise were examined in conditions of unstructured input data and a significant number of missing values. The best results were obtained using the Multihead-Attention-Based Gated Recurrent Unit (MAGRU) model. The modeling of total nitrogen was conducted with a time lag of 3 h, corresponding to the adopted technological scheme for wastewater treatment. This model has been referenced in several studies, demonstrating a generally satisfactory forecasting quality [19–21].

The second prevalent task in the field of wastewater treatment is classification. In most cases, this involves anomaly detection and the prevention of emergency situations at treatment facilities. For instance, researchers led by Bellamoli [22] developed an approach for classifying anomalies in wastewater treatment plants utilizing Sequencing Batch Reactor (SBR) systems. They identified the most representative predictors and time cycles for modeling. The most accurate results were achieved with ensemble models based on decision trees, specifically XGBoost and LightGBM. The recall metric reached values of 0.83 at various stages of modeling. The authors conclude that one direction for advancing classification modeling approaches in wastewater treatment plants is to focus on domain distribution characteristics.

Elsayed [23] conducted an analysis of 23 models for the effectiveness of classifying wastewater treatment quality. It is worth noting that the quality of treatment was classified based on individual parameters, resulting in a binary classification of “High” and “Low”. This approach appears somewhat unusual, as a regression model could have been employed for the same purposes. Nonetheless, the available predictors were tested using classical classification models, including k-nearest neighbor (KNN), support vector machine (SVM), and decision trees (DT). The classification accuracy reached 88%; however, given that the classes were unbalanced, it would be prudent to assess additional classification metrics. A similar study was conducted by Nasir [24], who also addressed the classification task to predict the quality of treated wastewater. Interestingly, the training dataset consisted of data from various treatment plants. One of the models used, CatBoost, demonstrated the highest effectiveness, achieving a classification accuracy of 94.51%. CatBoost is a multi-platform machine learning library developed by Yandex (Yandex, Moscow, Russia), designed for implementing gradient-boosting methods. Currently, it is one of the most popular models in data analysis across various scientific fields. Interestingly, in the articles reviewed in recent years concerning wastewater treatment, CatBoost models are rarely encountered. Nevertheless, among the advantages of these models are enhanced support for categorical features with automatic encoding, resilience to overfitting, rapid training times, versatility (the ability to address regression, classification, and ranking tasks), interpretability, and support for various platforms. CatBoost demonstrates high speed and accuracy, which contribute to its growing popularity in machine learning project applications [25–28].

Many modern approaches to wastewater parameter modeling rely on deep neural networks. These include Recurrent Neural Networks (RNN) and their variants, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), which are widely used for time series forecasting and sequential data, making them especially useful for monitoring wastewater parameters (e.g., BOD5, COD, TSS) over time. These networks can capture temporal dependencies, which is critical when analyzing seasonal or short-term fluctuations in wastewater characteristics. However, they may encounter difficulties in handling long-term dependencies [29]. Transformers, such as Temporal Fusion Transformers (TFT), are well-suited for modeling complex temporal dependencies and can capture relationships among various wastewater characteristics. They are particularly useful for multivariate data when several wastewater indicators (e.g., BOD, COD, pH, TSS) need to be considered simultaneously in forecasting. These models support variable and temporal dependency analysis in complex, highly variable data [30,31]. Neural networks based on N-BEATS (Neural Basis Expansion Analysis for Time Series Forecasting) represent a relatively new architecture for time series forecasting, enabling the effective modeling of long-term dependencies without using recurrent blocks. This architecture is suitable for predicting complex trends and analyzing temporal data, especially when seasonality and trends are less pronounced, as can occur in wastewater [32,33]. Combinations of CNN and RNN (e.g., CNN-LSTM) or GRU with autoencoders, such as top-sparse autoencoders (TSA), are used for processing and forecasting high-dimensional wastewater data. These architectures help extract spatial and temporal dependencies, enhancing model accuracy and adaptability under variable loads at treatment facilities [34,35].

However, working with time series data requires high-quality preprocessing and the regular measurement of parameters, which in many cases is challenging to achieve. This study addresses the issue of accumulating training datasets in modeling not based on the principle of time series. This is particularly relevant in cases where the technical infrastructure of wastewater treatment plants does not allow for the regular measurements of the characteristics necessary for constructing a time series. One of the primary objectives of this work is to develop principles for obtaining training datasets for model training, which can subsequently be applied to operational treatment facilities.

The feasibility of using laboratory sample data as a training set, followed by validation against data from real treatment plants, is examined. An essential stage of the research involved is analyzing the significance of predictors for establishing guidelines regarding the placement of analytical equipment at wastewater treatment stations [36].

In this work, models compatible with Shapley's value-based analysis tools (using the SHAP library) are employed, which enabled the compilation of a list of predictors for each specific task. Thus, the following research objectives have been formulated:

- To establish a system that facilitates the collection of the necessary amount of data for training supervised models based on a laboratory setup.
- To test the hypothesis regarding the equality of the ratios of input and output parameters of wastewater in the populations of the laboratory setup and the operational treatment plants under investigation.
- To develop a strategy for data preprocessing.
- To implement the addition and selection of predictors for model training.
- To identify the most effective models with optimal hyperparameters for classification and regression tasks using cross-validation tools.
- To evaluate the models on a test dataset using selected quality metrics.

To achieve these objectives, an experiment was conducted using a laboratory-scale bioreactor simulating the wastewater treatment process. Statistical tests were carried out to confirm the applicability of the data collected in laboratory conditions as a training dataset. The test dataset comprised data from full-scale wastewater treatment plants (WWTPs). This approach assesses the feasibility of training a machine learning model under conditions of limited historical data from real wastewater treatment plants. Well-established regression and classification machine learning models were employed for training, with performance evaluated across multiple metrics, significant parameters identified through SHAP analysis, and rigorous quality assessment of the model predictions.

2. Materials and Methods

2.1. Data Description

The primary objective of the study was to assess the feasibility of using training and test datasets from different sources for modeling the quality of treated wastewater and detecting anomalies. The test dataset comprised the results of technological monitoring from wastewater treatment plants located in the Moscow region, Russia. The monitoring was conducted during 2022–2023, resulting in a sample size of 140 data entries across 15 parameters. This dataset was insufficient for forming a robust training dataset; therefore, the missing information was obtained using a laboratory bioreactor whose operational scheme allowed for the generation of parameter distributions corresponding to the data from actual treatment plants. The equipment used is described in Section 2.2.

During the laboratory experiment, a dataset consisting of 4303 data entries across 15 parameters was collected. The quality of incoming and treated wastewater (only at the biological stage) was measured based on the following characteristics: biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), ammonium nitrogen (NH_4), orthophosphates (PO_4), nitrates, and nitrites (NO_2, NO_3, only in treated water). The technological parameters measured included wastewater temperature (t), dissolved oxygen concentration in the aerobic zone of the bioreactor (DO), and the proportion of carbon dioxide in the exhaust air (CO_2).

Categorical features included the type of external carbon source (acetate or a solution of synthesized volatile fatty acids after acidification (VFA)), the point of introduction of the external carbon source (into the bioreactor (In) or into the mixer before the bioreactor (Out)), and the presence of filamentous bulking of activated sludge. Among all the features, the characteristics of treated wastewater and the filamentous bulking were the primary targets for modeling. Thus, the dataset enabled the resolution of two tasks, regression and classification (detection of abnormal situations related to filamentous bulking of activated sludge). The collected parameters and their average values are presented in Table 1.

Table 1. The average values of the parameters in the datasets.

Parameter	Lab-Scale	Full-Scale
BOD_5 (influent) [mgO_2/L]	107	108
COD (influent) [mgO/L]	135	136
Ammonium (influent) (NH_4) [mg/L]	35.0	34.6
Phosphorus (influent) (PO_4) [mg/L]	8.0	7.6
DO [mg/L]	2.51	2.53
CO_2 [%]	3.3	3.0
Temperature (t) [°C]	20.0	19.8
External carbon source (substrate)	Acetate/VFA	
Inlet point	In/Out	
BOD_5 (effluent) [mgO_2/L]	3.4	3.5
COD (effluent) [mgO_2/L]	4.2	4.4
Ammonium (effluent) [mg/L]	1.0	1.0
Phosphorus (effluent) (PO_4) [mg/L]	0.35	0.17
Nitrate (effluent) (NO_3) [mg/L]	5.7	5.4
Nitrite (effluent) (NO_2) [mg/L]	0.03	0.03
Filamentous bulking	Yes/No	

Preliminary data preparation involved operations such as data import, type checking, and correction of missing values and duplicates. All numerical features incorporated into the dataframe represent independent results from analytical measurements of water quality. The presence of missing values may indicate either that measurements were not conducted or that a technical malfunction occurred during the data transmission or report compilation phases. This type of missing data, according to Donald Rubin's classification, falls under the category of Missing at Random (MAR), meaning that the missingness depends on an observable variable. In this case, the observable variable could be the measurement interval, which can be incorporated into the dataset as a separate feature. Removing records with missing values in "long" analyses may reduce the dataset's informativeness and render the remaining data points meaningless. However, after establishing the relationships among the features, it is possible to determine the sets of information necessary for modeling a specific target variable. If, for a particular model, the "long" analyses do not carry weight, they can be excluded, while a more comprehensive dataset using other parameters can enhance the modeling quality.

Thus, the recommended approach to address missing values in this case is Pairwise Deletion, which involves retaining the missing values and filling in existing gaps with placeholders. The placeholders may represent feature values outside the feasible measurement range. These placeholders can be temporarily disabled during analysis through logical data slicing. In instances where the data are time-distributed (i.e., the time of the experiment is recorded), the method of Missing Values Imputation can be applied, which consists of filling in the missing values. This method can also be automated within the pipeline using the SimpleImputer tool from the sklearn library.

Regarding categorical features, the presence of a missing value in one of them may be related to a specific experimental design. For example, an absence of values in the substrate feature might indicate that the substrate was not supplied at all, and the data collection system did not transmit any value. In such cases, it is advisable to use the most frequent value for the features. The primary objective of the statistical analysis of data is to identify the convergence of sample means between the training and test datasets. To achieve this, it is essential to establish the distribution characteristics of the indicator variables. This can be accomplished using both graphical methods, such as histograms and box plots, and statistical analysis methods. For the statistical analysis, the Shapiro–Wilk test is proposed. The null hypothesis H_0 of the Shapiro–Wilk test posits that the random variable, for which the sample is known, follows a normal distribution. Conversely, the alternative hypothesis H_1 states that the distribution is not normal.

In the case of normally distributed data, standard tools for parametric statistical analysis can be employed. For instance, a two-tailed Student's *t*-test can be utilized, accompanied by the corresponding null and alternative hypotheses and the significance level criterion. In cases where the sample distributions are not normal—an expected occurrence for water treatment efficiency, which often exhibits an exponential distribution—it is advisable to utilize bootstrap analysis methods. In this approach, 100,000 random samples, each comprising 10% of the total dataset, are generated for each indicator variable, followed by the assessment of the *p*-value with a significance level of 0.05.

The null hypothesis posits that the sample means are equal, while the alternative hypothesis asserts that they differ. The allowable range for differences is limited to the variability of the absolute means of the samples; for instance, in the case of biochemical oxygen demand (BOD) efficiency, this variability was found to be 0.0007. In instances of significant discrepancies between the sample means, it is recommended to implement changes to the operational technology regime of the laboratory setup. Conversely, if convergence is observed, a correlation analysis may be conducted.

Correlation analysis is conducted using both graphical and computational methods. Given that the dataset includes categorical variables and potentially discrete features, the standard calculation of correlation coefficients is not applicable. Instead, it is recommended to employ correlation calculation tools based on chi-squared statistics, such as the construction of a correlation matrix using the phik_matrix tool from the PhiK library. Calculating correlation through chi-squared statistics allows for the identification of relationships among all types of data—categorical, continuous, and discrete—which is not feasible with conventional linear correlation methods. Additionally, an assessment of the multicollinearity among the features was performed in preparation for the selection of linear models.

2.2. Lab-Scale Equipment

The laboratory component of the study was conducted using an automated bioreactor-fermentor equipped with a set of auxiliary sensors. The setup is based on the Yocell YC-JG reactor (Yocell Biotechnology, Qingdao, China) and includes a reactor vessel with a working volume of 11 L, made of borosilicate glass. The vessel is equipped with a controlled electromechanical stirrer featuring adjustable rotational speed, a pneumatic aeration system with adjustable aeration intensity, and a heating and cooling system through an external circuit. The setup was further equipped with a set of analytical sensors that correspond to those used in actual wastewater treatment facilities. These include the Hamilton VisiFerm DO sensor for dissolved oxygen (Hamilton Company, Reno, NV, USA), the Hamilton Polilyte Plus pH ARC sensor for pH measurement (Hamilton Company, Reno, NV, USA), the HACH A-ISE system for nitrogen compound analysis (Hach, Loveland, CO, USA), the BlueSens BlueVary system for analyzing carbon dioxide content in the outgoing air stream (BlueSens Gas Sensor GmbH, Herten, Germany), and the CarboVis 701/705 IQ system for analyzing COD and BOD (Xylem Analytics, San Diego, CA, USA).

Control and calibration measurements were conducted using the HACH Lange DR6000 UV-Vis spectrophotometer (Hach, Loveland, CO, USA), the WTW OxiTOP-IDS system (Xylem Analytics, San Diego, CA, USA), the WTW Oxi3310 equipped with the CellOX 325 sensor (Xylem Analytics, San Diego, CA, USA), and the WTW pH 3310 analyzer (Xylem Analytics, San Diego, CA, USA). The control tests for suspended solids were performed using standard methods. Figure 1 shows a photograph of the bioreactor.

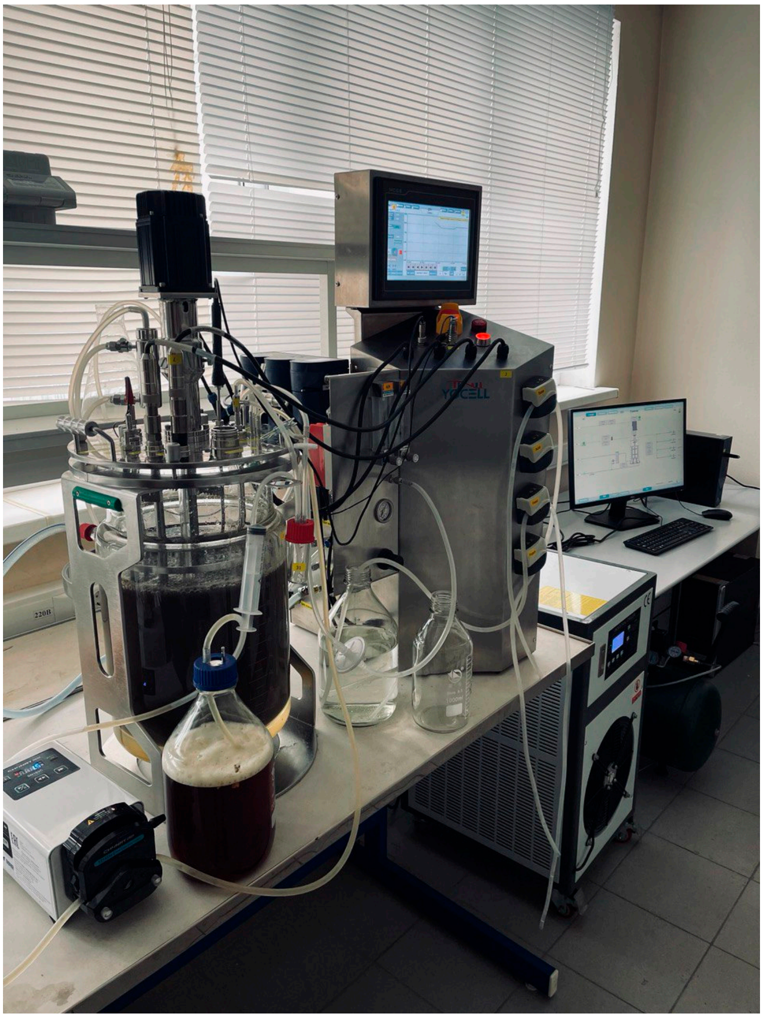


Figure 1. The lab-scale bioreactor-fermentor.

The operation of the system was managed through an integrated controller, as well as via a personal computer using Siemens Simatic software (SIMATIC WinCC V7.5 SP2 Upd15). In addition to automating the control of the main components of the setup, the control program was configured to facilitate data collection from all analyzers. The collected data were exported to a .csv file for further analysis. For the operation of the system, wastewater from active treatment facilities was used as the substrate, along with additional substrates consisting of solutions of volatile fatty acids obtained from the acidification of the sludge and acetic acid (CH_3COOH, AnalaR NORMAPUR, 99%). The volatile fatty acids were produced through the fermentation of sludge from the same treatment facilities.

The inoculum biomass was sourced from the aeration tank of the studied treatment facilities. The operational regime of the system was aligned with the technological processes of the existing facilities.

The bioreactor's operating conditions were optimized to achieve comparable performance characteristics of the system. The modeling was conducted in accordance with the computational principles outlined in a previous study [37]. Thus, the actual technological process was replicated under controlled laboratory conditions. The data collection scheme for the laboratory setup is illustrated in Figure 2.

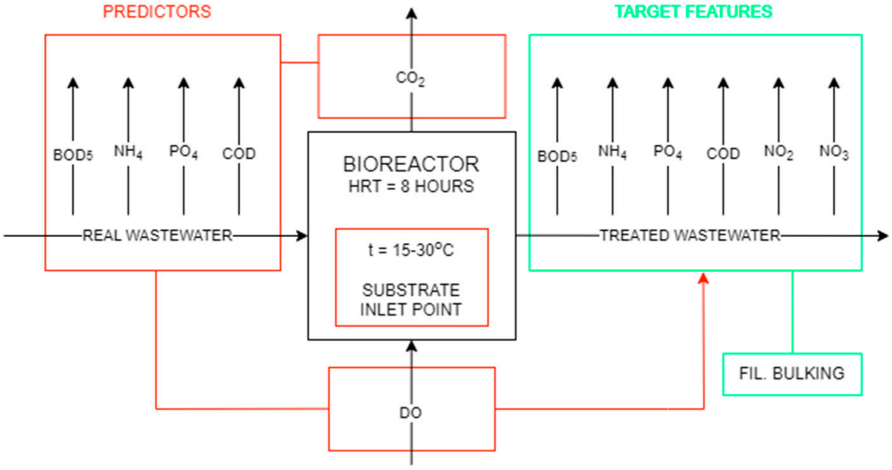


Figure 2. The data collection scheme for the laboratory setup.

The laboratory bioreactor operated in sequencing batch reactor (SBR) mode. The operational parameters of the setup are listed in Table 2.

Table 2. Lab-scale bioreactor operation parameters.

Parameter	Values
Fill time [min]	15
Idle time [min]	5
Aeration time [min]	210
Mixing time [min]	210
Settling time [min]	30
Decanting time [min]	10
Hydraulic retention time (HRT) [h]	8
Mixed liquor volatile suspended solids (MLVSS) [g/L]	4.5
Dissolved oxygen (aeration stage) [mg/L]	3.5
Dissolved oxygen (anoxic stage) [mg/L]	0.2
External carbon source (BOD_5 eq.) [mgO_2/L]	150
Temperature (t) [°C]	15–30

2.3. Methodology

The primary objective of this work is to develop algorithms for solving regression and classification tasks. The regression task focuses on predicting numerical target features that serve as indicators of water quality after treatment. A separate model is trained for each feature. The pipeline for model selection and training consists of the following components:

- Data processing (encoding and scaling of features).
- Randomized cross-validation of selected models to find the optimal value of the chosen metric.
- Model evaluation, including adequacy testing, assessment on a test set using a set of metrics, and evaluation of feature importance using SHAP values. The Shapley values are calculated using Formula (1), as follows:

$$\Phi_i(p) = \sum_{S \in N/\{i\}} \frac{|S|!(n-|S|-1)!}{n!}(p(S \cup \{i\}) - p(S)), \quad (1)$$

where n represents the number of features in the model, S is a subset of features that does not include feature i, $|S|$—denotes the number of features in subset S, $p(S)$ is the value obtained from subset S, and $(p(S \cup \{i\})$ is the value obtained from subset S that includes feature i.

In addressing the regression task, the following machine learning models were employed: DecisionTreeRegressor, LinearRegression, ElasticNet, Lasso Regression, Ridge Regression, XGBoost, LightGBM, and CatBoost. The defining metric for cross-validation was the Symmetric Mean Absolute Percentage Error (SMAPE), which is determined using Formula (2).

$$SMAPE = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}, \tag{2}$$

where n is the number of observations, F_t is the predicted value, and A_t is the actual value.

After selecting significant features, re-cross-validation is performed. When the values of the target metric increase, a new model with the selected hyperparameters is accepted. Subsequently, the model metrics are evaluated on the test set. For the regression task, the metrics R^2, RMSE, and MAE are assessed (as defined in Formulas (3)–(5), respectively).

$$R^2 = 1 - \frac{D[y|x]}{D[y]} = 1 - \frac{\sigma^2}{\sigma_y^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (\overline{y_i} - y_i)^2}, \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}, \tag{4}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|, \tag{5}$$

where $D[y]$ = σ_y^2 represents the variance of the random variable, $D[y|x]$ = σ^2 denotes the variance of the model error, n is the number of observations, y_i is the target value, $\hat{y}_i$ is the predicted value, and $\overline{y_i}$ is the mean value.

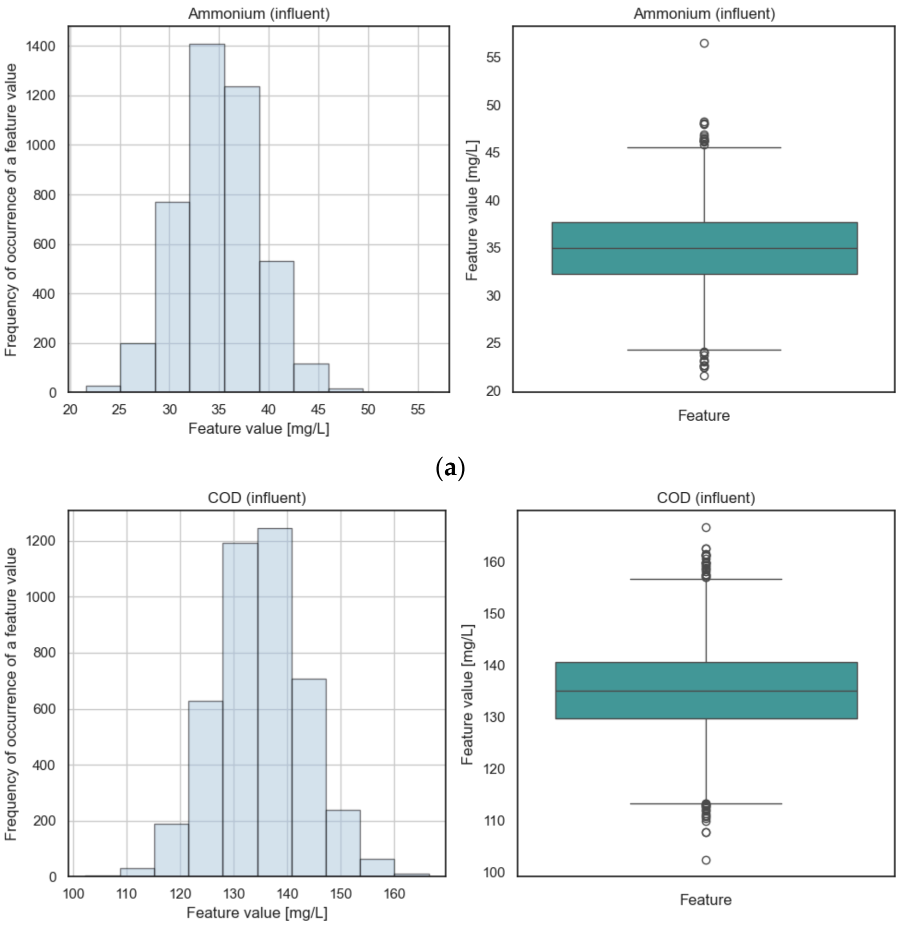
The adequacy of the model was assessed by comparing the target metric of the model with the metric of the constant model, DummyRegressor. In addressing the classification task, the following models were considered: DecisionTreeClassifier, LogisticRegression, KNeighborsClassifier, SVC (Support Vector Classification), as well as XGBoost, LightGBM, and CatBoost. The metric employed for the evaluation was the ROC-AUC (Receiver Operating Characteristic—Area Under the Curve), which is a metric used to assess the performance of binary classifiers. This metric is based on the analysis of a curve constructed from the values of true positive and false positive rates at various classification thresholds. An AUC greater than 0.5 indicates that the trained model performs better than random guessing.

Considering the research objectives, data from actual wastewater treatment facilities were used as the test set for model evaluation (preceded by a statistical analysis of both laboratory and real-world data). For hyperparameter tuning, the cross-validation tool RandomizedSearchCV (from Scikit-learn V1.5.2) was employed, eliminating the need for a separate validation dataset within the training dataset.

3. Results

Prior to model construction, a comprehensive statistical analysis was conducted on the datasets obtained from both operational wastewater treatment facilities and the laboratory experiment. The results of the statistical evaluation indicated that the numerical features exhibit varying measurement scales, necessitating data scaling during modeling to ensure uniformity. Additionally, no anomalous values were detected within the datasets, confirming their suitability for modeling purposes. Information from both dataframes was examined to facilitate a comparative analysis. It was observed that the nature of the feature values corresponds between datasets, thereby enabling further progress on the project. To test the hypotheses regarding the equality of sample values of feature ratios between the populations of the laboratory experiment and the operational facilities, additional features were introduced. These include the ratio of easily oxidizable organic matter (BOD_5) to ammonium nitrogen (NH_4) in the incoming wastewater (C/N ratio) and the efficiency of removal of key pollutants.

Overall, all numerical features, both input and target variables, exhibited distributions approaching normality, attributed to the controlled conditions of the experiment. Outliers were present for all features, defined as values lying outside the 1.5 interquartile range. The presence of outliers aligns with the inherent variability in wastewater treatment processes. The measured values fell within the analytical equipment's range and corresponded to the technological scheme and operational conditions of the installations. Figure 3 presents the distribution diagrams of ammonium nitrogen and COD values in the incoming wastewater, illustrating the variability and distribution characteristics of these key parameters.



(**a**)

(**b**)

Figure 3. Distribution of predictor values: (**a**) ammonium (influent); (**b**) COD (influent).

To prevent negative impacts on model training quality due to class imbalance, oversampling of the target categorical feature (filamentous bulking occurrence) was performed as a stratification method. Oversampling was conducted using the tools available in the imbalanced-learn library, specifically the "RandomOverSampler" function, to ensure adequate representation of the minority class during model training. The distribution of class values concerning the presence of filamentous bulking after oversampling is illustrated in Figure 4.

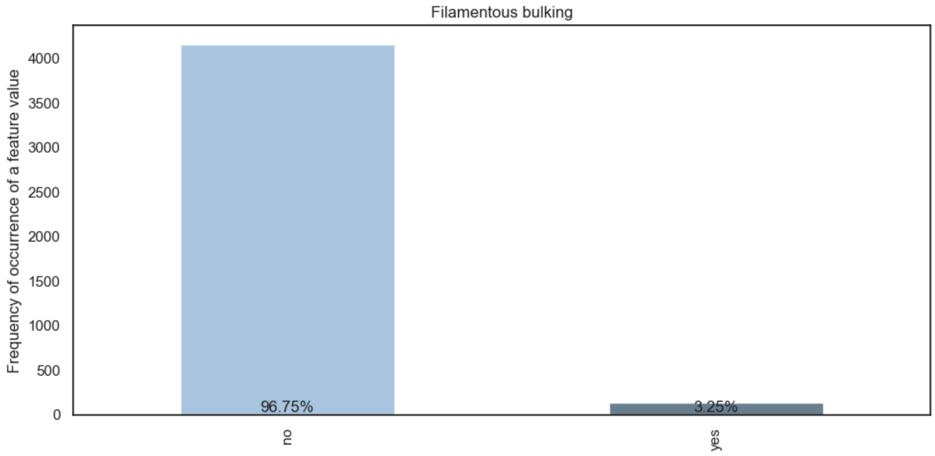
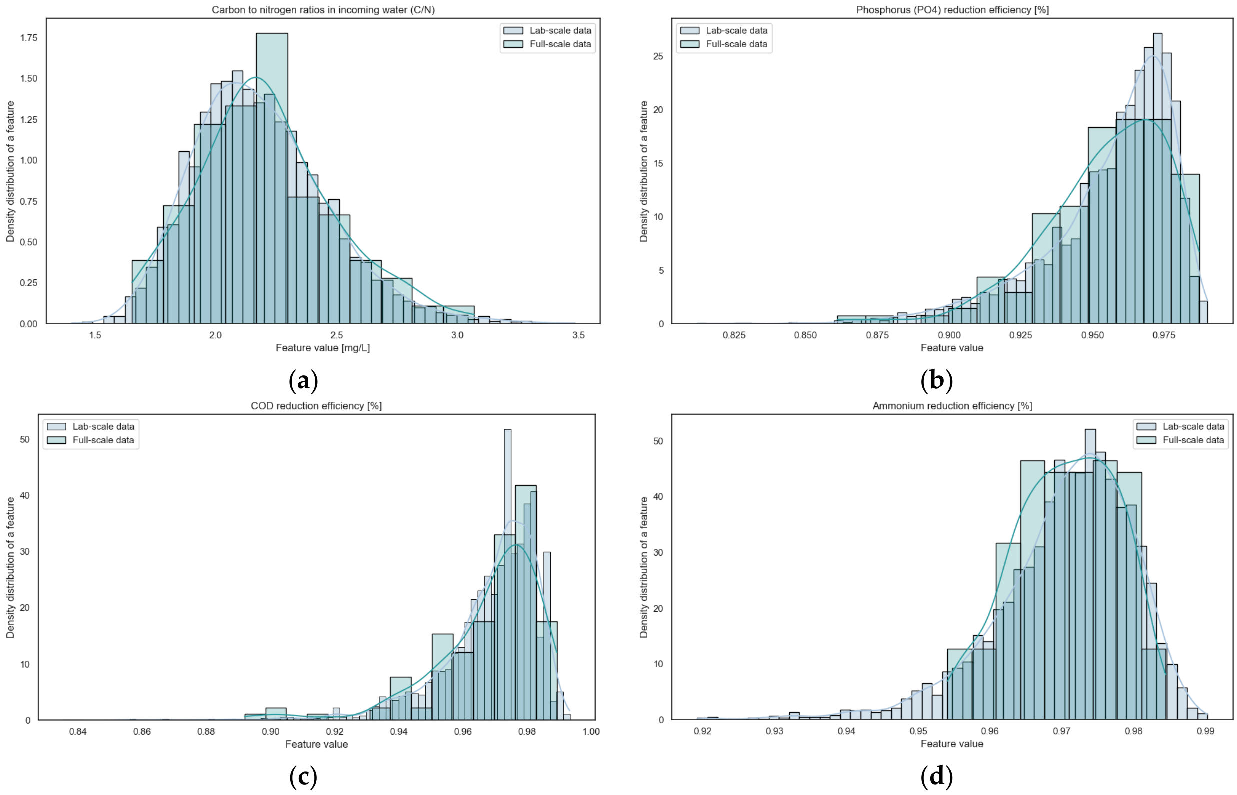


Figure 4. Distribution of the target feature of the presence of filamentous bulking in the bioreactor.

According to the Shapiro–Wilk statistical test, the distributions of indicator variables—primarily the efficiency of wastewater treatment—are generally not normal ($p < 0.05$). A visual comparison of the samples was conducted using histograms with a highlighted density. The comparative histograms are presented in Figure 5. Despite deviations from normality, the histograms indicate that many distributions approach normality, and the variances of the samples are nearly equal.



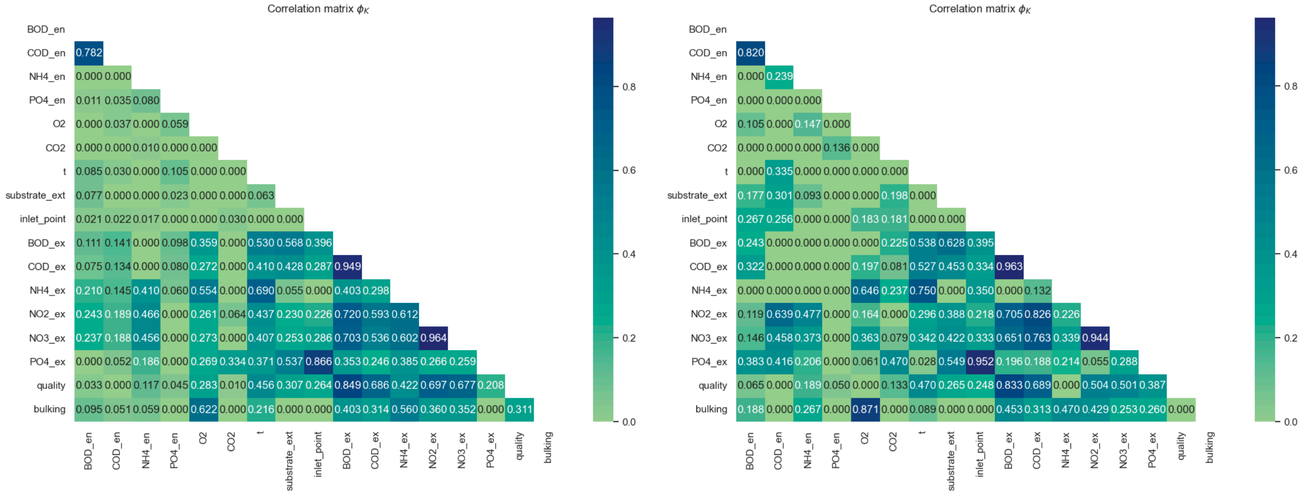
(**a**) (**b**)

(**c**) (**d**)

Figure 5. Comparative histograms of the bioreactors (Lab-scale and Full-scale) operation: (**a**) C/N ratio; (**b**) orthophosphates reduction efficiency [%]; (**c**) COD reduction efficiency [%]; (**d**) ammonium reduction efficiency [%].

Applying Student's *t*-test to non-normalized samples yields results with significant assumptions; therefore, a bootstrap analysis with 1000 resamples was conducted. The results supported all null hypotheses regarding the convergence of sample means for all indicator variables. According to the statistical tests and bootstrap analysis, the sample means of the populations for the majority of indicator variables are equal ($p > 0.05$). Consequently, we assumed that the results of laboratory modeling can be applied to the operation of existing wastewater treatment facilities.

Since the statistical tests indicated that the training and test samples can be used in their current form, we performed a correlation analysis of the features from both samples to verify the convergence of their interrelationships. It was essential to confirm that there are no significant differences in the feature distributions between the training and test samples and that the model evaluation of the test sample would be valid. Furthermore, the analysis assessed the multicollinearity of the input features. A correlation matrix was constructed, incorporating the calculation of Pearson correlation coefficients for continuous variables and transformed chi-squared statistics for categorical variables. The correlation matrix for the training sample is presented in Figure 6.

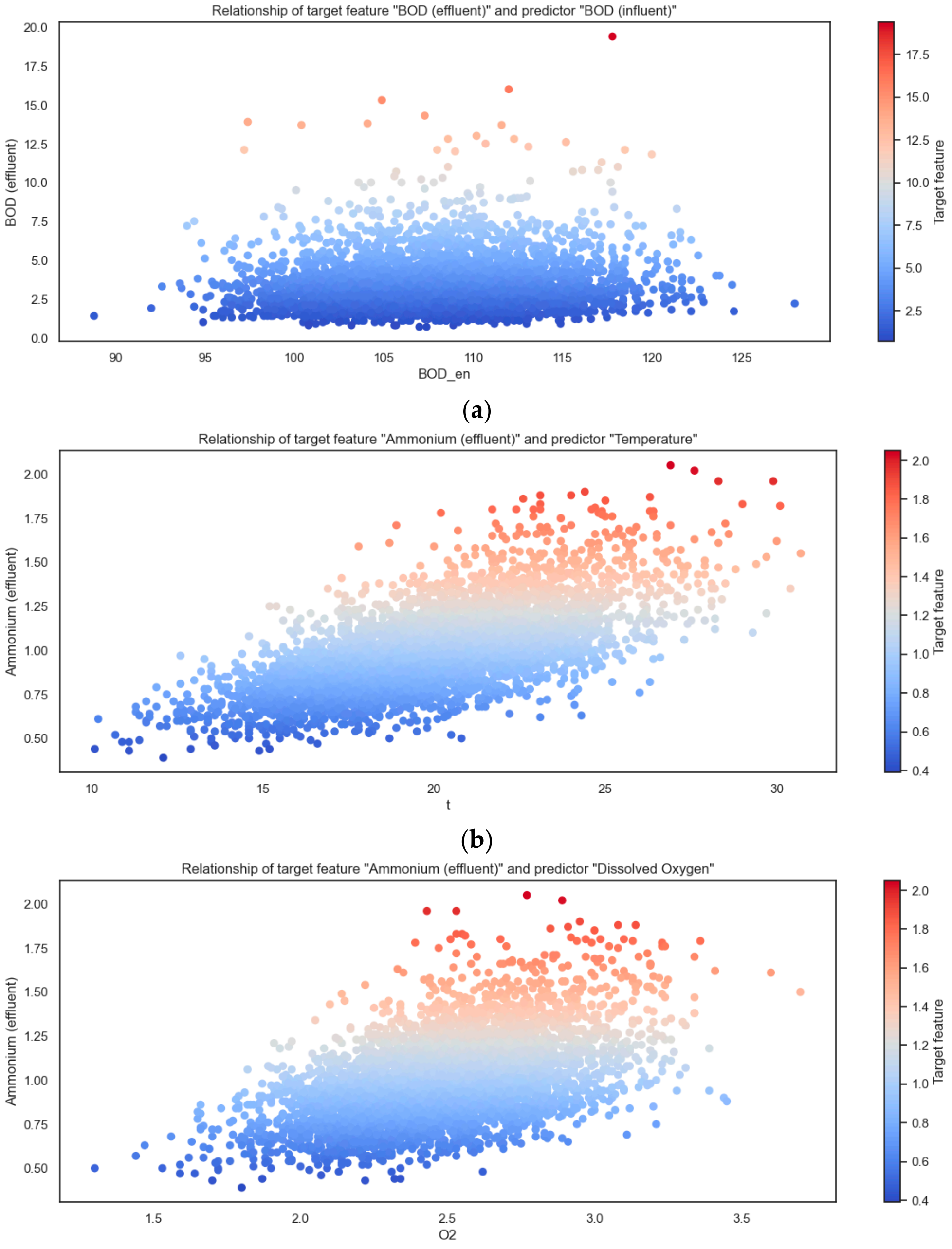


(**a**) (**b**)

Figure 6. Correlation matrix PhiK: (**a**) lab-scale data; (**b**) full-scale data.

As observed, the overall nature of the relationships between the features in both samples is consistent, allowing for the adequate evaluation of the models using the test sample. No significant multicollinearity (correlation coefficient > 0.8) was detected among the input features, suggesting that each predictor contributes unique information to the model. However, a strong correlation was identified between certain target parameters, such as COD/BOD_5, substrate input point/phosphates, and nitrates/nitrites. To assess the presence of linear dependencies among the features, scatter plots and box plots were additionally constructed for each pair of predictors and target variables. Some of these plots are presented in Figure 7.

Model validation was conducted using cross-validation with randomized optimization of the loss function, employing the RandomizedSearchCV tool from scikit-learn. For the regression task, the validation metric employed was SMAPE (Symmetric Mean Absolute Percentage Error), providing a scale-independent measure of predictive accuracy. For the classification task, the area under the receiver operating characteristic curve (ROC-AUC) was used as the validation metric, assessing the model's ability to discriminate between classes.



(**a**)

(**b**)

(**c**)

Figure 7. Feature span diagrams: (**a**) BOD effluent/BOD influent; (**b**) ammonium effluent/temperature; (**c**) ammonium effluent/dissolved oxygen.

The analysis of SHAP (SHapley Additive exPlanations) values facilitated the identification of the most significant predictors for each target variable, enhancing model interpretability. Figure 8 presents the distribution of SHAP values for the predictors associated with the target variables, COD effluent and orthophosphates effluent. The most influential features included the type of external substrate, its point of introduction, dissolved oxygen concentration, and temperature.

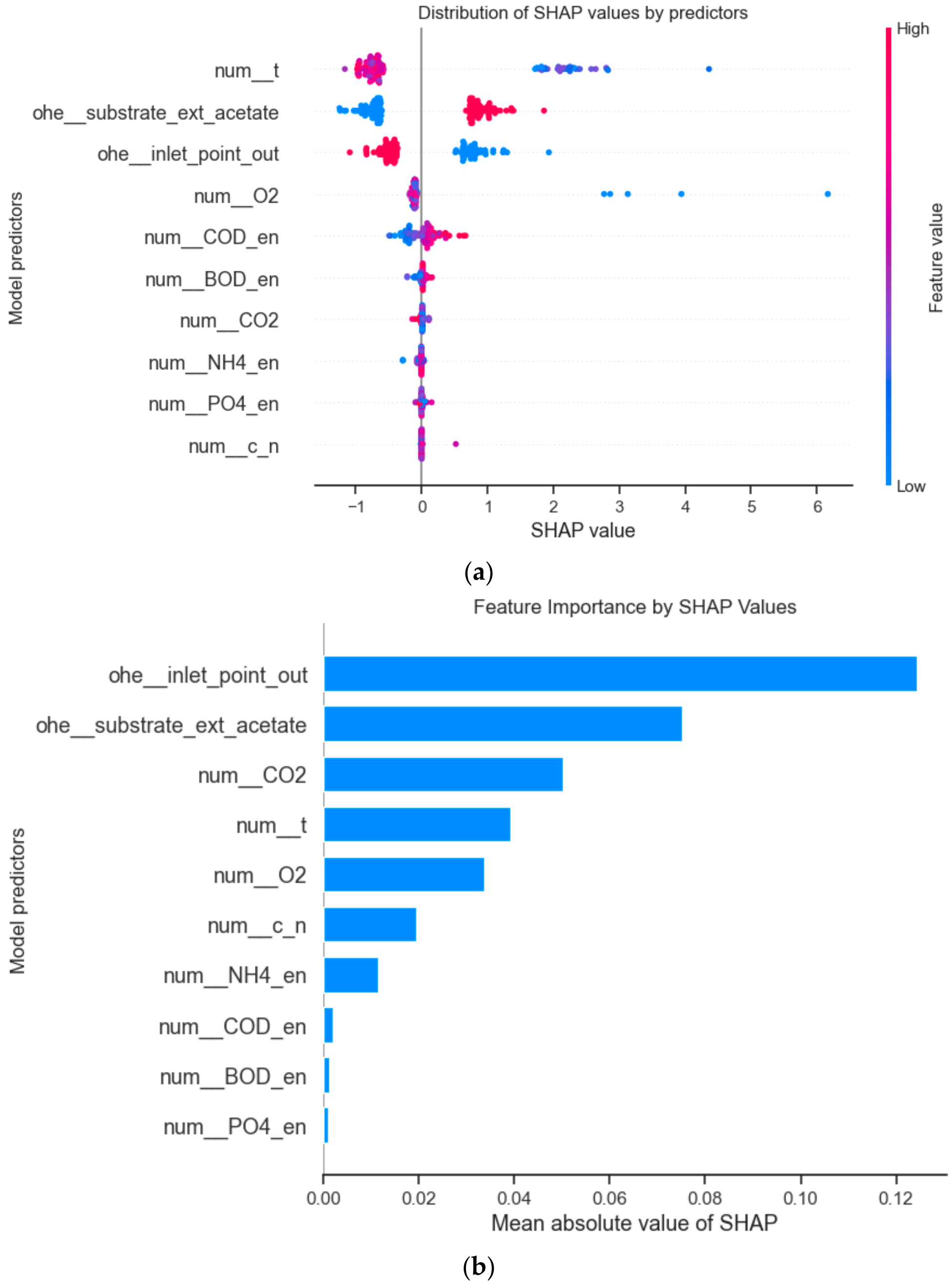


Figure 8. The most significant predictors for modeling COD of treated water: (**a**) SHAP beeswarm plot; (**b**) SHAP summary bar plot.

The results of the modeling are presented in Table 3. The CatBoost model demonstrated superior performance in both regression and classification tasks, with SMAPE values below 10% for key effluent parameters and a ROC-AUC score of 0.95 for the classification of filamentous bulking events.

As observed, modeling involves limited sets of features, which can generally be explained in the context of the relevant mechanistic models [13]. The classification task necessitates the exclusion of false negative predictions from the model. A false negative prediction (or missed detection of an emergency) means that the model failed to recognize a potential emergency, predicting a normal state when the situation is actually hazardous. The model's operation should be aimed at preventing emergencies at the station, even at the expense of overall accuracy (i.e., false positive predictions are acceptable). Figure 9 presents the confusion matrix, indicating that no changes to the prediction thresholds are required.

Table 3. Summary table of modeling results within the framework of the tasks set.

Target Feature	Best Model	Hyperparameters	Important Features	Metrics Values
BOD_5 (effluent)	CatBoostRegressor	learning_rate 0.05 iterations 150 depth 7 random_strength 2	Temperature Substrate Inlet point DO	SMAPE 19.2 RMSE 0.962 MAE 0.670 R^2 0.78
COD (effluent)		learning_rate 0.35 iterations 180 depth 6 random_strength 2	Temperature Substrate Inlet point DO BOD influent COD influent	SMAPE 19.1 RMSE 1.211 MAE 0.854 R^2 0.85
Ammonium (effluent)		learning_rate 0.25 iterations 210 depth 6 random_strength 2	DO Temperature C/N ratio	SMAPE 10.6 RMSE 0.145 MAE 0.111 R^2 0.79
Phosphorus (effluent)		learning_rate 0.12 iterations 160 depth 8 random_strength 2	DO Temperature NH_4 influent C/N ratio Substrate Inlet point	SMAPE 9.1 RMSE 0.054 MAE 0.036 R^2 0.81
Nitrate (effluent)		learning_rate 0.08 iterations 130 depth 7 random_strength 2	DO Temperature NH_4 influent C/N ratio Substrate Inlet point	SMAPE 9.1 RMSE 0.765 MAE 0.484 R^2 0.82
Nitrite (effluent)		learning_rate 0.05 iterations 140 depth 7 random_strength 2	DO Temperature NH_4 influent	SMAPE 12.4 RMSE 0.006 MAE 0.003 R^2 0.88
Filamentous bulking	CatBoostClassifier	learning_rate 0.8 iterations 150 depth 6 random_strength 2	DO Temperature NH_4 influent C/N ratio	ROC-AUC 1.0

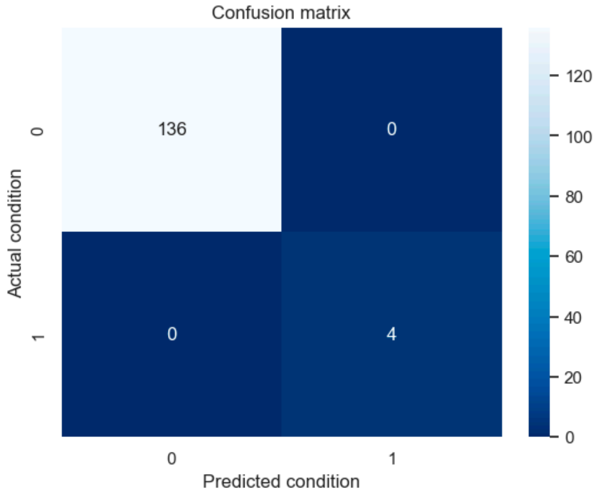
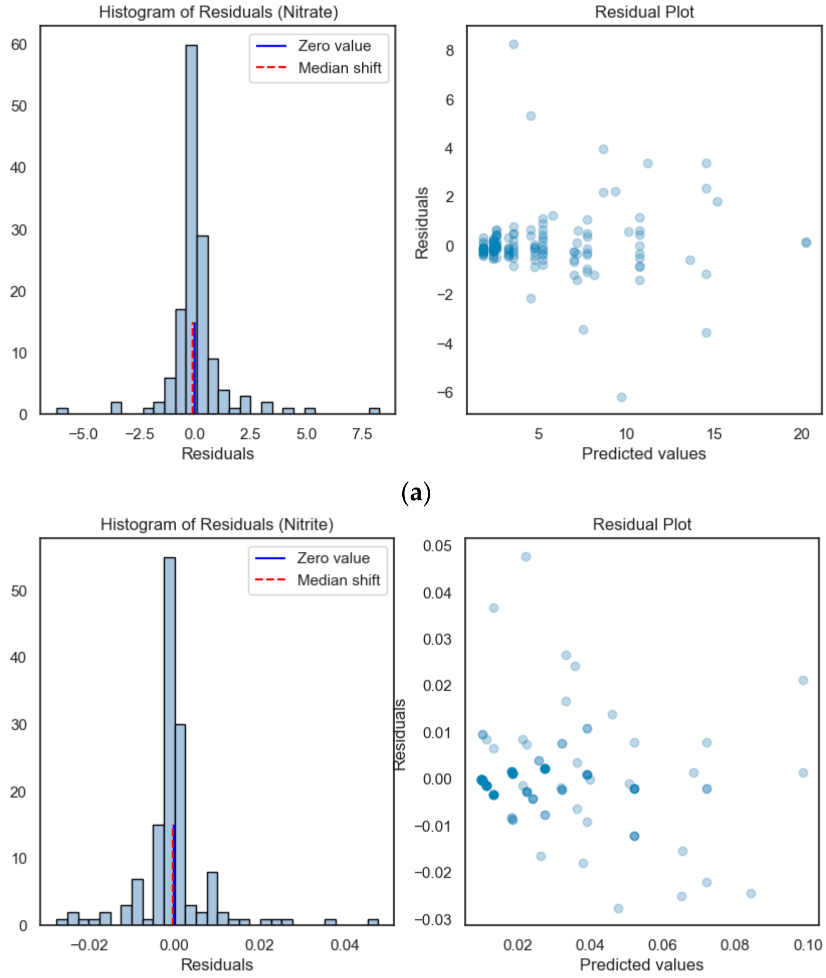


Figure 9. Confusion matrix for predicting filamentary bulking of activated sludge using test data.

For the regression task results, residual analysis was performed for each target feature. A graphical representation of the analysis for some of the target features is shown in Figure 10.



(**a**)

(**b**)

Figure 10. Residual analysis: (**a**) nitrate; (**b**) nitrite.

The residual analysis demonstrated favorable results. Residuals were randomly distributed around zero, indicating no systematic bias in the model predictions. Slight biases in the median were observed, which can be addressed by incorporating a larger dataset of historical data [15].

4. Discussion

Filamentous bulking of activated sludge is classified as an emergency situation within wastewater treatment systems, often resulting in extremely negative outcomes for process development. Essentially, filamentous bulking frequently leads to the complete shutdown of the entire system due to the loss of sludge settleability, causing effluent quality deterioration. The probability of its occurrence in conventional treatment plants operating under normal conditions is low, estimated at approximately 3% in the laboratory experiment. When training the model, stratification of the data or oversampling is necessary to mitigate the adverse effects of class imbalance in the target variable. The application of oversampling techniques ensures that rare but critical events, such as filamentous bulking, are adequately represented in the training process, improving the model's ability to predict such events.

The histograms indicate that many distributions approach normality, and the variances of the samples are nearly equal. However, the issue of non-normality is likely related to the absence of ultra-low values for the indicators due to the detection limits of the equipment. This limitation affects statistical analyses that assume normal distribution, emphasizing the need for robust statistical methods or data transformations.

Applying Student's *t*-test to non-normalized samples introduces significant assumptions. Thus, the use of bootstrap analysis provided a more reliable assessment of sample means without relying on normality assumptions. Confirming that laboratory data can be applied to operational facilities is significant, as it validates the relevance of the experimental findings to real-world applications. The constructed scatter plots revealed that a linear dependence on the target numerical variables is present only for ammonium, dissolved oxygen, and temperature. This is primarily related to the nature of biochemical processes, where reaction rates depend largely on ambient temperature and the availability of dissolved oxygen to the biomass. Additionally, the availability of organic substrates relative to the amount of incoming nitrogen (C/N ratio) plays a significant role in microbial activity and nutrient removal efficiency. Consequently, linear regression models are unlikely to yield high performance metrics in assessing treated wastewater quality due to the complex, non-linear relationships between process variables. Similar findings regarding the absence of clear linear dependencies between predictors and target variables have been reported in other studies [12,14,17]. This suggests that advanced modeling techniques capable of capturing non-linear interactions, such as gradient-boosting algorithms, are more suitable for this application.

Notably, the results of the significance analysis differ from those of previous studies [17,20,22]. This discrepancy arises because, in this work, the most influential features include those not utilized in other studies, specifically the type of external substrate and its point of introduction. The inclusion of these operational parameters provides a more comprehensive understanding of the factors influencing effluent quality. Furthermore, the best-performing model identified was CatBoost, a gradient-boosting algorithm that effectively handles categorical features. A comparison with studies that also evaluated the CatBoost model indicates that the significance of common features is similar [24–28], reinforcing the validity of our findings.

The modeling involves limited sets of features, which can generally be explained in the context of relevant mechanistic models [13]. Using a limited but significant set of predictors enhances model interpretability and reduces the risk of overfitting. The classification task necessitates the exclusion of false negative predictions from the model. A false negative prediction (or missed detection of an emergency) means that the model failed to recognize a potential emergency, predicting a normal state when the situation is actually hazardous. The model's operation should be aimed at preventing emergencies at the station, even at the expense of overall accuracy (i.e., false positive predictions are acceptable). This approach prioritizes safety and environmental protection, ensuring that potential risks are not overlooked. The confusion matrix presented in Figure 9 indicates that no changes to the prediction thresholds are required, as the model successfully identifies all instances of filamentous bulking without producing false negatives. This result underscores the effectiveness of the model in emergency prediction. The residual analysis demonstrates favorable results; however, the presence of slight biases in the median suggests that incorporating a larger dataset of historical data could further enhance model performance [15]. Expanding the dataset would improve the model's generalizability and robustness, capturing a wider range of operating conditions and the variability inherent in wastewater treatment processes.

Overall, this study demonstrates the feasibility and effectiveness of using advanced machine learning models, specifically CatBoost, for predicting effluent quality and potential emergency situations in wastewater treatment plants. The inclusion of operational parameters unique to this study contributes to a deeper understanding of process dynamics. Future research should explore the impact of external substrates and their introduction points on treatment efficiency, as well as the integration of larger, more diverse datasets to enhance model reliability.

5. Conclusions

It is important to note that the primary goal of this project was not the development of specific models but rather the formulation of approaches in the form of recommendations for data acquisition, preparation for modeling, and subsequent model selection. Each modification of the technological scheme parameters (for instance, changes in treatment duration) necessitates retraining the model. The recommendations can be articulated as follows:

1. It is essential to ensure complete adherence to the technological parameters of the investigated treatment facilities during laboratory modeling. First and foremost, this includes aligning the average values of incoming pollutant concentrations, dissolved oxygen concentrations, and temperature. The proportion of carbon dioxide is not regulated; therefore, in cases of significant discrepancies in values, it is recommended to exclude this parameter from the modeling process. It is advisable to use analytical equipment that is similar to that used at the facility. Additionally, it is recommended to automate the collection of sensor readings, focusing on the longest measurement intervals.
2. It is recommended to incorporate a new feature into the model, the carbon-to-nitrogen (C/N) ratio. This feature holds significant importance for nearly all target variables.
3. It is recommended to remove all complete duplicates from the training dataset to prevent data leakage during cross-validation. Missing values in quantitative features should be handled using the Pairwise Deletion strategy. In cases where time series data are present in the dataset, missing values can be filled with the average of the two adjacent values. For categorical features, it is advisable to impute missing values using the mode.
4. It is recommended not to remove outliers that fall outside the range of the interquartile range, as outliers are significant indicators for the occurrence of emergency situations. Before training models on the training dataset, statistical tests for the comparison of sample means should be conducted. Student's *t*-test can be applied when there are more than 50 observations in the sample. Additionally, a correlation analysis should be performed primarily to identify multicollinearity among predictors. If the values in the samples are not suitable for statistical analyses (for example, according to Student's *t*-test), it is advisable to conduct a bootstrap analysis, as water treatment efficiency indicators often exhibit exponential distribution.
5. It is recommended to apply oversampling methods or stratification of target categorical features. In the present project, high-quality predictions were achieved despite class imbalance; however, verification is necessary for each specific case. Feature scaling should be performed using robust methods, such as the RobustScaler.
6. For the examined set of features, gradient-boosting methods are the most preferred models for both regression and classification tasks, with CatBoost identified as the optimal model based on the results of the study. Each target feature has its own optimal set of hyperparameters; therefore, it is recommended to utilize cross-validation procedures for hyperparameter tuning during modeling.
7. The most objective metrics for model evaluation in this project are proposed to be SMAPE for regression and ROC-AUC for classification. It is recommended to assess model adequacy by comparing these metrics against a constant model during the model-building process.
8. The necessary sets of features for use as predictors (minimum quantity) are presented in Table 2 for each of the target metrics. These features were selected based on the evaluation of SHAP values and are intended to enhance the models by eliminating noise.
9. In addressing classification tasks related to system failure detection, it is essential to prevent the occurrence of false negative predictions. This means excluding the oversight of forecasting emergency situations, which are encoded as 0 by the LabelEncoder. To achieve this, it is necessary to conduct an evaluation using the confusion matrix

and, if needed, to re-calibrate the thresholds of the models based on the metrics of Recall and Precision.

10. The results obtained from the modeling can be utilized to predict both the efficiency of wastewater treatment and the likelihood of emergency situations, with a time lag equivalent to the Hydraulic Retention Time (HRT). The HRT is incorporated into the design calculations of wastewater treatment facilities and is adjusted based on the hydraulic load on the bioreactor. This factor must be taken into consideration in each specific case.

Author Contributions: Conceptualization, O.K. and I.G.; methodology, O.K.; software, I.G.; validation, I.G. and O.K.; formal analysis, I.G.; investigation, I.G.; resources, O.K.; data curation, I.G.; writing—original draft preparation using I.G.; writing—review and editing, I.G.; visualization, I.G.; supervision, O.K.; project administration, I.G.; funding acquisition, I.G. All authors have read and agreed to the published version of the manuscript.

Funding: The research was funded by the National Research Moscow State University of Civil Engineering (grant for fundamental and applied scientific research, project No. 27-392/130).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, Y.; Wu, H.; Xu, R.; Wang, Y.; Chen, L.; Wei, C. Machine learning modeling for the prediction of phosphorus and nitrogen removal efficiency and screening of crucial microorganisms in wastewater treatment plants. *Sci. Total Environ.* **2024**, *907*, 167730. [CrossRef] [PubMed]
2. Wang, D.; Thunéll, S.; Lindberg, U.; Jiang, L.; Trygg, J.; Tysklind, M.; Souihi, N. A machine learning framework to improve effluent quality control in wastewater treatment plants. *Sci. Total Environ.* **2021**, *784*, 147138. [CrossRef]
3. Xu, B.; Pooi, C.K.; Tan, K.M.; Huang, S.; Shi, X.; Ng, H.Y. A novel long short-term memory artificial neural network (LSTM)-based soft-sensor to monitor and forecast wastewater treatment performance. *J. Water Process Eng.* **2023**, *54*, 104041. [CrossRef]
4. Abouzari, M.; Pahlavani, P.; Izaditame, F.; Bigdeli, B. Estimating the chemical oxygen demand of petrochemical wastewater treatment plants using linear and nonlinear statistical models—A case study. *Chemosphere* **2021**, *270*, 129465. [CrossRef] [PubMed]
5. Recio-Colmenares, R.; León Becerril, E.; Gurubel Tun, K.J.; Conchas, R.F. Design of a Soft Sensor Based on Long Short-Term Memory Artificial Neural Network (LSTM) for Wastewater Treatment Plants. *Sensors* **2023**, *23*, 9236. [CrossRef]
6. El-Rawy, M.; Abd-Ellah, M.K.; Fathi, H.; Ahmed, A.K.A. Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques. *J. Water Process Eng.* **2021**, *44*, 102380. [CrossRef]
7. Singh, N.K.; Yadav, M.; Singh, V.; Padhiyar, H.; Kumar, V.; Bhatia, S.K.; Show, P.L. Artificial intelligence and machine learning-based monitoring and design of biological wastewater treatment systems. *Bioresour. Technol.* **2023**, *369*, 128486. [CrossRef]
8. Boumezbeur, H.; Laouacheria, F.; Heddam, S.; Djemili, L. Modelling coagulant dosage in drinking water treatment plant using advance machine learning model: Hybrid extreme learning machine optimized by Bat algorithm. *Environ. Sci. Pollut. Res.* **2023**, *30*, 72463–72483. [CrossRef]
9. Xiong, J.; Tao, Z.; Hua, L.; Qiao, X.; Peng, T.; Nazir, M.S.; Zhang, C. An evolutionary robust soft measurement technique via enhanced atom search optimization and outlier robust extreme learning machine for wastewater treatment process. *J. Water Process Eng.* **2023**, *55*, 104102. [CrossRef]
10. Razmi, M.; Saneie, M.; Basirat, S. Estimating discharge coefficient of side weirs in trapezoidal and rectangular flumes using outlier robust extreme learning machine. *Appl. Water Sci.* **2022**, *12*, 176. [CrossRef]
11. Hasani, F.; Shabanlou, S. Outlier robust extreme learning machine to simulate discharge coefficient of side slots. *Appl. Water Sci.* **2022**, *12*, 170. [CrossRef]
12. Zaghloul, M.S.; Achari, G. Application of machine learning techniques to model a full-scale wastewater treatment plant with biological nutrient removal. *J. Environ. Chem. Eng.* **2022**, *10*, 107430. [CrossRef]
13. Xu, B.; Pooi, C.K.; Yeap, T.S.; Leong, K.Y.; Soh, X.Y.; Huang, S.; Shi, X.; Mannina, G.; Ng, H.Y. Hybrid model composed of machine learning and ASM3 predicts performance of industrial wastewater treatment. *J. Water Process Eng.* **2024**, *65*, 105888. [CrossRef]
14. Alali, Y.; Harrou, F.; Sun, Y. Unlocking the potential of wastewater treatment: Machine learning based energy consumption prediction. *Water* **2023**, *15*, 2349. [CrossRef]

15. Plappally, A.; Lienhard V, J.H. Energy requirements for water production, treatment, end use, reclamation, and disposal. *Renew. Sustain. Energy Rev.* **2012**, *16*, 4818–4848. [CrossRef]
16. Asadi, A.; Verma, A.; Yang, K.; Mejabi, B. Wastewater treatment aeration process optimization: A data mining approach. *J. Environ. Manag.* **2017**, *203*, 630–639. [CrossRef]
17. Asami, H.; Golabi, M.; Albaji, M. Simulation of the biochemical and chemical oxygen demand and total suspended solids in wastewater treatment plants: Data-mining approach. *J. Clean. Prod.* **2021**, *296*, 126533. [CrossRef]
18. Safder, U.; Kim, J.; Pak, G.; Rhee, G.; You, K. Investigating machine learning applications for effective real-time water quality parameter monitoring in full-scale wastewater treatment plants. *Water* **2022**, *14*, 3147. [CrossRef]
19. Li, J.; Dong, J.; Chen, Z.; Li, X.; Yi, X.; Niu, G.; He, J.; Lu, S.; Ke, Y.; Huang, M. Free nitrous acid prediction in ANAMMOX process using hybrid deep neural network model. *J. Environ. Manag.* **2023**, *345*, 118566. [CrossRef]
20. Xie, Y.; Mai, W.; Ke, S.; Zhang, C.; Chen, Z.; Wang, X.; Li, Y.; Dionysiou, D.D.; Huang, M. Artificial intelligence-implemented prediction and cost-effective optimization of micropollutant photodegradation using g-C_3N_4/Bi_2O_3 heterojunction. *Chem. Eng. J.* **2024**, *499*, 156029. [CrossRef]
21. Wan, X.; Li, X.; Wang, X.; Yi, X.; Zhao, Y.; He, X.; Wu, R.; Huang, M. Water quality prediction model using Gaussian process regression based on deep learning for carbon neutrality in papermaking wastewater treatment system. *Environ. Res.* **2022**, *211*, 112942. [CrossRef] [PubMed]
22. Bellamoli, F.; Di Iorio, M.; Vian, M.; Melgani, F. Machine learning methods for anomaly classification in wastewater treatment plants. *J. Environ. Manag.* **2023**, *344*, 118594. [CrossRef] [PubMed]
23. Elsayed, A.; Siam, A.; El-Dakhakhni, W. Machine learning classification algorithms for inadequate wastewater treatment risk mitigation. *Process Saf. Environ. Prot.* **2022**, *159*, 1224–1235. [CrossRef]
24. Nasir, N.; Kansal, A.; Alshaltone, O.; Barneih, F.; Sameer, M.; Shanableh, A.; Al-Shamma'a, A. Water quality classification using machine learning algorithms. *J. Water Process Eng.* **2022**, *48*, 102920. [CrossRef]
25. Jiang, J.; Xiang, X.; Zhou, Q.; Zhou, L.; Bi, X.; Khanal, S.K.; Wang, Z.; Chen, G.; Guo, G. Optimization of a Novel Engineered Ecosystem Integrating Carbon, Nitrogen, Phosphorus, and Sulfur Biotransformation for Saline Wastewater Treatment Using an Interpretable Machine Learning Approach. *Environ. Sci. Technol.* **2024**, *58*, 12989–12999. [CrossRef]
26. Al Nuaimi, H.; Abdelmagid, M.; Bouabid, A.; Chrysikopoulos, C.V.; Maalouf, M. Classification of WatSan Technologies using machine learning techniques. *Water* **2023**, *15*, 2829. [CrossRef]
27. Wang, Q.; Li, Z.; Cai, J.; Zhang, M.; Liu, Z.; Xu, Y.; Li, R. Spatially adaptive machine learning models for predicting water quality in Hong Kong. *J. Hydrol.* **2023**, *622*, 129649. [CrossRef]
28. Halalsheh, N.; Alshboul, O.; Shehadeh, A.; Al Mamlook, R.E.; Al-Othman, A.; Tawalbeh, M.; Almuflih, A.S.; Papelis, C. Breakthrough curves prediction of selenite adsorption on chemically modified zeolite using boosted decision tree algorithms for water treatment applications. *Water* **2022**, *14*, 2519. [CrossRef]
29. Farhi, N.; Kohen, E.; Mamane, H.; Shavitt, Y. Prediction of wastewater treatment quality using LSTM neural network. *Environ. Technol. Innov.* **2021**, *23*, 101632. [CrossRef]
30. Burrichter, B.; Koltermann da Silva, J.; Niemann, A.; Quirmbach, M. A Temporal Fusion Transformer Model to Forecast Overflow from Sewer Manholes during Pluvial Flash Flood Events. *Hydrology* **2024**, *11*, 41. [CrossRef]
31. Sun, X.; Zhang, L.; Wang, C.; Yang, Y.; Wang, H. Dynamic Real-Time Prediction of Reclaimed Water Volumes Using the Improved Transformer Model and Decomposition Integration Technology. *Sustainability* **2024**, *16*, 6598. [CrossRef]
32. Zhang, Y.; Suzuki, G.; Shioya, H. Prediction and Detection of Sewage Treatment Process Using N-BEATS Autoencoder Network. *IEEE Access* **2022**, *10*, 112594–112608. [CrossRef]
33. Hao, Z. A dissolved oxygen prediction model based on GRU–N-Beats. *Front. Mar. Sci.* **2024**, *11*, 1365047. [CrossRef]
34. Chen, Z.; Hu, J.; Min, G.; Zomaya, A.Y.; El-Ghazawi, T. Towards Accurate Prediction for High-Dimensional and Highly-Variable Cloud Workloads with Deep Learning. *IEEE Trans. Parallel Distrib. Syst.* **2020**, *31*, 923–934. [CrossRef]
35. Yang, Y.; Kim, K.R.; Kou, R.; Li, Y.; Fu, J.; Zhao, L.; Liu, H. Prediction of effluent quality in a wastewater treatment plant by dynamic neural network modeling. *Process Saf. Environ. Prot.* **2022**, *158*, 515–524. [CrossRef]
36. Wang, D.; Thunéll, S.; Lindberg, U.; Jiang, L.; Trygg, J.; Tysklind, M. Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. *J. Environ. Manag.* **2022**, *301*, 113941. [CrossRef]
37. Gogina, E.; Gulshin, I. Characteristics of low-oxygen oxidation ditch with improved nitrogen removal. *Water* **2021**, *13*, 3603. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.