Advances in Science and Technology Research Journal

Advances in Science and Technology Research Journal, 2026, 20(10), 351–363
https://doi.org/10.12913/22998624/225615
ISSN 2299-8624, License CC-BY 4.0

Received: 2026.06.25
Accepted: 2026.07.28
Published: 2026.09.01

Machine learning-based prediction of chlorophenol removal from wastewater using reverse osmosis

Emad Majeed Hameed[1*], Ahmed Abdul Azeez Ismael[1], Ali Mahmood Khalaf[2]

[1] Technical Institute of Baquba, Middle Technical University, Diyala, Iraq
[2] Tikrit University, Tikrit, Iraq
* Corresponding author's email: emadmajeed@mtu.edu.iq

ABSTRACT

One of the most popular methods for reducing the presence of extremely toxic compounds in wastewater is the reverse osmosis (RO) process. The purpose of this research was to examine the prosperity of different machine learning (ML) techniques based optimisation for the removal of chlorophenol from wastewater using a single module of RO process. The main intention was to predict the optimal operating conditions that would introduce a maximum chlorophenol rejection using different ML techniques (the Linear Regression, Ridge, Lasso, Decision Tree, Random Forest, Gradient Boosting, SVR, XGBoost, and Neural Networks) based on available experimental data. The performance of the developed models was evaluated using root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and cross-validation. The results demonstrated that the Gradient Boosting algorithm can achieve the best performance in predicting chlorophenol removal, followed by the XGBoost algorithm. However, neural networks performed the worst. The results show that feed concentration and pressure are the essential operating conditions that can dominate the rejection rate of chlorophenol. More importantly, it was concluded that feed pressure of 12.76 atm, feed concentration of 5079.05 kmol/m^3, wastewater temperature of 31.11 °C, and feed flowrate of 2.36×10^{-4} m^3/s are the optimal conditions to attain the maximum chlorophenol rejection of 82.37%.

Keywords: wastewater, chlorophenol rejection, reverse osmosis, machine learning techniques.

INTRODUCTION

A wide range of chemical industry effluents contain micro-pollutants, including common ones like phenolic compounds, particularly the highly toxic chlorophenol. These compounds are present in wastes from pharmaceutical, refinery, and petrochemical operations [1,2]. Unfortunately, its negative effects on the environment in general and on human health in particular, are undeniable, even at very low concentrations [3]. Consequently, numerous regulatory bodies have strictly limited the release of phenolic wastewater without adequate treatment. As an example, the permissible level for phenol in surface water was set at 0.5 ppm, according to the standards of Japan Environmental Governing [4]. Also, the recycling of wastewater for reuse should not have any phenolic substances. To successfully attain these restrictions, various treatment techniques were created to reduce these compounds across numerous industrial sectors. For instance, membrane and extraction processes, distillation, adsorption, oxidation are primitive and famous methods utilised to eliminate phenolic substances from wastewater [5]. Membrane filtration, particularly the RO method, has been widely employed for removing a broad spectrum of organic contaminants, with notable effectiveness against phenolic compounds [6].

RO process is basically a semi-permeable membrane designed to reject dissolved organic contaminants such as 2,4-dichlorophenol and pentachlorophenol. Numerous laboratory

investigations were conducted to eliminate phenol, chlorophenol, and dimethylphenol from simulated industrial effluent, which verified the viability of a pilot-scale RO system [7–9].

The prediction of organic compounds removal from wastewater using ML models is a rapidly evolving field where developed models demonstrated significant potential. These models can analyse complex, non-linear relationships between operational parameters (like pressure, pH, temperature, and feed concentration) and the removal efficiency. Common methods involve Artificial Neural Networks, Support Vector Machines, and Random Forest, they are often employed to forecast system performance with high accuracy, optimising the process and reducing experimental costs [10,11].

Accurate prediction of the effluent characteristics, released from wastewater treatment plants (WWTPs), is essential to lowering labour, expenses, sample needs, and environmental contamination. ML methods can be efficient in achieving this goal. Gholizadeh et al. [7] aimed to investigate the impact of six feature selection (FS) approaches on the performance of three ML techniques for predicting total suspended solids (TSS). The methods used included defining five scenarios based on FS results and applying ANN-MLP, KNN, and AdaBoost algorithms. Using K-fold cross-validation approach, the dataset was divided into training and testing sets. The results showed the most effective case was IV case using Sequential Backward Selection. The ANN-MLP algorithm proved the most effective with high R^2 values (0.78 training, 0.8 testing). Warren-Vega et al. [8] attempted to treat piggery wastewater using Fenton and solar photo-Fenton processes. The methods used involved a laboratory-scale Taguchi experimental design to determine optimal parameters of (pH, time, $[H_2O_2]$, $[Fe^{2+}]$), followed by a cascade forward neural network to model the data. The results showed high removal efficiencies (97.89% colour) under optimal conditions. The neural network achieved an excellent R^2 of 0.99, enabling accurate prediction for scaling up the process.

Maqsood et al. [9] explored the growing use of advanced computational models. The aim was to review their application. Methods used include decision trees, artificial neural networks, long short-term memory, fuzzy logic, internet of things, and recurrent neural networks. These methods were studied for optimising key indicators like biochemical oxygen demand. The results showed that these models can improve efficiency and monitoring. However, challenges like data access and reproducibility still remain. Hybrid models combining approaches showed better performance.

Wang et al. [12] intended to develop a mathematical model frame for optimising RO performance in brackish waters applying response surface methodology RSM and ANN methods. The used methods were RSM based on 15 experimental datasets and a Genetic Algorithm optimised ANN with a 3–10-4 structure trained on 70% training data. The results showed that the accuracy of ANN technique is higher than RSM with R^2 of 0.9936 for performance index and 0.9932 for specific energy consumption and a low RMSE. A two-step optimisation method combining RSM and ANN was used to propose the best process conditions for particular feed TDS scenarios.

Ahmad et al. [13] reviewed the application of ML in membrane desalination for optimisation. The methods used include analysing techniques like ANNs, SVMs, and RF applied to systems such as RO and forward osmosis (FO). The results demonstrate that ML models, particularly ANNs, can enhance the performance prediction and fouling mitigation by correlating operational parameters with efficiency.

Al-Obaidi et al. [14] focused on enhancing the efficiency of RO process for the removal of dimethylphenol from wastewater using response surface methodology (RSM). The researchers developed a vigorous forecasting model based on actual experimental data. The model inspected main operational parameters of feed pressure, concentration temperature, and feed flow rate to evaluate their effect on the effectiveness of dimethylphenol removal. Particularly, the results introduced that feed pressure and concentration can positively affect dimethylphenol removal rate, in contrast to a detrimental effect on feed flow rate and water temperature. Statistically, the analysis of variance emphasised by an impressive F-value of 240.38, highlighted the noteworthy role of feed pressure and concentration in the dimethylphenol rejection. Also, the model revealed a high-predictive precision, verified by an R^2 value of 0.9772, ascertaining the reliability of the developed RSM model.

The current research intended to investigate optimal conditions of a single module of RO process for efficient removal of chlorophenol from

wastewater by deploying various ML techniques relied on collected experimental data from the previous studies. The research aimed at identifying the optimal operational conditions to maximise chlorophenol rejection by assessing and comparing the performance of various ML models like Linear Regression, Ridge, Lasso, Decision Tree, Random Forest, Gradient Boosting, Support Vector Regression (SVR), XGBoost, and Neural Networks. The importance of this research lies in its potential industrial implications. First, improving the efficiency of wastewater treatment processes is vital as toxic compounds such as chlorophenol pose serious environmental and health risks. Second, the gained visions from this research can help decision-makers to adopt superior data-driven techniques for process optimisation, leading to better agreement with environmental guidelines besides developing more sustainable wastewater management solutions. Third, this research not only contribute to the progression of greener technologies in the chemical and processing industries but also resolve public health concerns due to toxic pollutants.

MATERIAL AND METHOD

The dataset

A laboratory standard pilot scale cross flow RO treatment unit was employed by Sundaramoorthy et al. [15] for the removal of chlorophenol from wastewater at a set of variable inlet conditions. The RO setup utilised a commercial thin film composite polyamide membrane with an effective area of 7.845 m², which was housed inside a spiral wound membrane module. The module's specifications are provided in Table 1.

Figure 1 displays a schematic representation of the RO process. A feed tank containing wastewater supplies the RO module via a high-pressure pump, which can achieve pressures up to 20 atm. Chlorophenol solutions were produced at variable concentrations, ranging between 0.778×10^{-3} to 6.226×10^{-3} kmol/m³. Additional operating parameters, such as flow rate and pressure were adjusted across ranges of 2.166×10^{-4} to 2.583×10^{-4} m³/s and 5.83 to 13.58 atm, respectively, for each particular rate of feed flow. The treatment occurred at operating temperatures between 29.5 °C and 32.5 °C. Sundaramoorthy et al. [15] utilised 70 sets of comprehensive experimental data as provided in Table 2. The data comprise the feed flowrate (Fb), feed pressure (Pb), inlet water temperature (Tb), feed chlorophenol concentration (Cb), and overall chlorophenol rejection (Rej).

Methodology

The objective of this research was to investigate optimum chlorophenol removal from wastewater based on an integrated unit of RO process via machine learning. The target was to predict the rejection rate and identify optimal associated operating conditions. The methodology follows a comprehensive machine learning pipeline:

Data loading and exploration

This step loads experimental data with features like feed pressure, temperature, concentration, and flow rate. Exploratory data analysis was performed, including correlation matrices,

Table 1. Specification of RO module [15]

Property	Value
Maker and configuration	Ion Exchange, India Ltd., TFC Polyamide, spiral wound
Feed (t_f) and permeate (tp) channel thickness	0.8 (mm) and 0.5 (mm)
Membrane length (L) and width (W)	93.4 (cm) and 840 (cm)
Membrane volume	6.2764×10^{-3} m³
Effective membrane area (A)	7.845 m²
Module diameter	8.25 (cm)
Maximum feed temperature (°C)	40
Maximum feed pressure (atm)	24.771
Maximum pressure drop per element (atm)	1.381
Maximum and minimum feed flow rate (m³/s)	$1 \times 10^{-4} – 1 \times 10^{-3}$
Permeate pressure (atm)	1

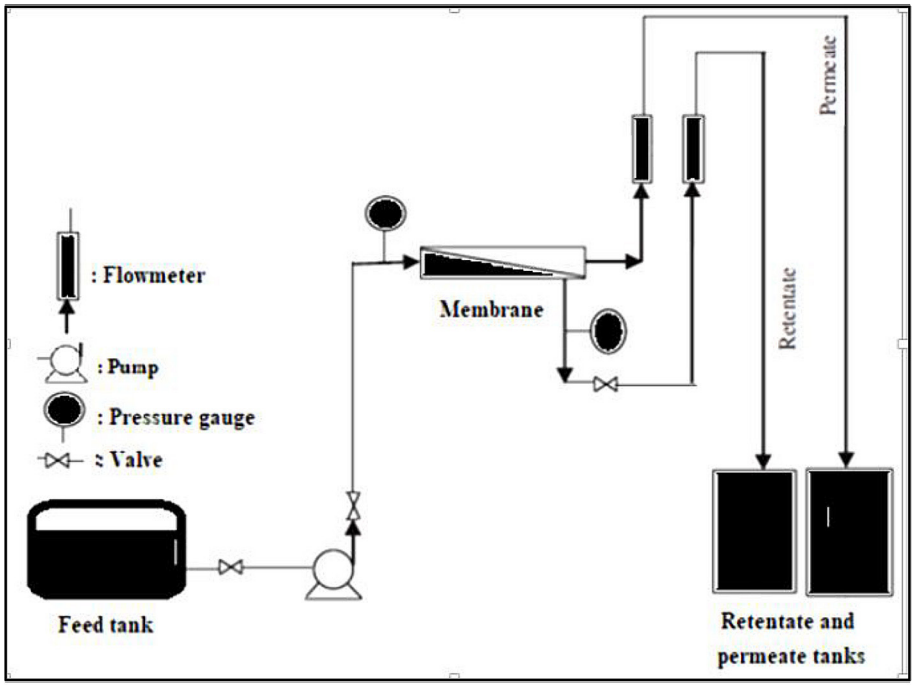

Figure 1. An illustration of RO process (taken from [15])

distribution plots. Table 3 shows samples of the used dataset, while Table 4 introduces the statistical information of dataset.

Figure 2 depicts the correlation matrix of inlet variables and the performance indicator of chlorophenol rejection. The correlation matrix is characterised in a colour-coded format where colour intensity specifies the strength of the correlation. Dark red values indicate a robust positive correlation (close to +1), while light colours or blue shades specify weak or negative correlations (close to 0 or negative values). The correlation matrix in Figure 2 shows relationships between operating variables. P (atm) has a moderate positive correlation with Rej (0.38) and negligible correlations with all other variables. Tb (°C) has moderate positive correlations with Cb (0.28) and Rej (0.37) and a weak negative correlation with Fb (-0.25). Cb (kmol/m³) has a very strong positive correlation with Rej (0.86) and moderate positive correlations with Tb (0.28). Fb (m³/s x100l) has a negligible correlation with the other variables. Rej (%) has a very strong positive correlation with Cb (0.86) and moderate positive correlations with P (0.38), and Tb (0.37).

The data distribution of each feature in the dataset is shown in Figure 3. The distribution of each feature against the target feature "chlorophenol rejection (Rej)." is shown in Figure 4.

Data preprocessing

Data preprocessing involved splitting the data into training (56 samples) and test sets (14 samples). The processes of normalisation and standardisation utilise techniques to scale numerical data to a uniform range. Normalisation transforms values into a 0 to 1 interval, while standardisation adjusts them to possess a mean of 0 and a standard deviation of 1. These methods enhance the convergence of optimisation models and ensure feature comparability. Converting Categorical Data: Since most machine learning models need numerical data, categorical variables must be converted. Approaches like one-hot encoding, label encoding, and target encoding are employed to change categorical information into a numerical form suitable for algorithms [12]. This study applied standard scaling to the features.

Model training and evaluation

This step trains multiple regression models including Linear Regression, Ridge, Lasso, Decision Tree, Random Forest, Gradient Boosting, SVR, XGBoost, and Neural Networks. The models were evaluated using RMSE, MAE, R^2, and cross-validation. Linear Regression models the relationship between features and a target variable using a linear approach. Ridge Regression is a regularised linear model that prevents overfitting

Table 2. Experimental data collection [15]

Exp. No.	Pb (atm)	Tb (°C)	Cb (kmol/m³ x10³)	Fb (m³/s x10⁴)	Rej (%)	Exp. No.	Pb (atm)	Tb (°C)	Cb (kmol/m³ x10³)	Fb (m³/s x10⁴)	Rej (%)
1	5.83	30	0.778	2.166	56.7	36	5.83	31	2.335	2.33	68.7
2	7.77	30	0.778	2.166	59.3	37	7.77	31	2.335	2.33	70
3	9.71	30	0.778	2.166	61.4	38	9.71	31	2.335	2.33	71.4
4	11.64	30	0.778	2.166	63.8	39	11.64	31	2.335	2.33	72.9
5	13.58	30	0.778	2.166	66.2	40	13.58	31	2.335	2.33	74.4
6	5.83	32	1.556	2.166	61.9	41	5.83	31	6.226	2.33	74.5
7	7.77	32	1.556	2.166	63.9	42	7.77	31	6.226	2.33	76.6
8	9.71	32	1.556	2.166	65.9	43	9.71	31	6.226	2.33	79.8
9	11.64	32	1.556	2.166	67.9	44	11.64	31	6.226	2.33	81.2
10	13.58	32	1556	2.166	70	45	13.58	31	6.226	2.33	82.2
11	5.83	32	2.335	2.166	65.6	46	5.83	29.5	0.778	2.58	57.8
12	7.77	32	2.335	2.166	66.8	47	7.77	29.5	0.778	2.58	60.6
13	9.71	32	2.335	2.166	68.4	48	9.71	29.5	0.778	2.58	62.8
14	11.64	32	2.335	2.166	69.6	49	11.64	29.5	0.778	2.58	64.3
15	13.58	32	2.335	2.166	71.1	50	13.58	29.5	0.778	2.58	66.3
16	5.83	32	3.891	2.166	70.7	51	5.83	31	1.556	2.58	65.2
17	7.77	32	3.891	2.166	72.3	52	7.77	31	1.556	2.58	67.5
18	9.71	32	3.891	2.166	71.7	53	9.71	31	1.556	2.58	69.7
19	11.64	32	3.891	2.166	74.8	54	11.64	31	1.556	2.58	71.2
20	13.58	32	3.891	2.166	76.4	55	13.58	31	1.556	2.58	72.5
21	5.83	31	6.226	2.166	75.5	56	5.83	31	2.335	2.58	69.9
22	7.77	31	6.226	2.166	76.7	57	7.77	31	2.335	2.58	71.7
23	9.71	31	6.226	2.166	79.8	58	9.71	31	2.335	2.58	72.8
24	11.64	31	6.226	2.166	81	59	11.64	31	2.335	2.58	74.2
25	13.58	31	6.226	2.166	81.9	60	13.58	31	2.335	2.58	75.7
26	5.83	30	0.778	2.33	56.2	61	5.83	32	3.891	2.58	71.7
27	7.77	30	0.778	2.33	58.1	62	7.77	32	3.891	2.58	72.8
28	9.71	30	0.778	2.33	60.3	63	9.71	32	3.891	2.58	75.0
29	11.64	30	0.778	2.33	62.4	64	11.64	32	3.891	2.58	76.1
30	13.58	30	0.778	2.33	64.5	65	13.58	32	3.891	2.58	77.4
31	5.83	32	1.556	2.33	62.9	66	5.83	31	6.226	2.58	72.6
32	7.77	32	1.556	2.33	64.6	67	7.77	31	6.226	2.58	77.8
33	9.71	32	1.556	2.33	66.4	68	9.71	31	6.226	2.58	79.4
34	11.64	32	1.556	2.33	68.6	69	11.64	31	6.226	2.58	81.5
35	13.58	32	1.556	2.33	70.4	70	13.58	31	6.226	2.58	**83.0**

by adding L2 penalty. Lasso Regression uses L1 regularisation for feature selection and can yield sparse models. Decision Tree splits data into subsets based on feature values to make predictions. Random Forest is an ensemble of decision trees that improves accuracy and reduces overfitting. Gradient Boosting builds models sequentially where each tree corrects errors of the previous one. SVR finds a hyperplane to predict continuous values with maximum margin. XGBoost is an optimised gradient boosting system known for its speed and performance. Neural Networks use interconnected layers of neurons to learn complex non-linear patterns [13–16].

RMSE measures the square root of the average squared differences between predicted and actual values. It is given by Equation 1. RMSE is sensitive to large errors, because it squares the differences before averaging. A lower RMSE indicates better model accuracy [17].

Table 3. Dataset samples

Exp. No.	Pb (atm)	Tb (°C)	Cb (kmol/m³ x10³)	Fb (m³/s x10⁴)	Rej (%)
1	5.83	30	0.778	2.166	56.7
2	7.77	30	0.778	2.166	59.3
3	9.71	30	0.778	2.166	61.4
4	11.64	30	0.778	2.166	63.8
5	13.58	30	0.778	2.166	66.2

Table 4. Statistical information of dataset

Parameter	Pb (atm)	Tb (°C)	Cb (kmol/m³ x10³)	Fb (m³/s x10⁴)	Rej (%)
count	70	70	0.70	0.70	70
mean	9.706	31.1071	2.8905	2.3607	70.2128
std	2.7591	0.8115	0.9970	0.1772	6.7796
min	5.83	29.5	0.778	2.166	56.2
25%	7.77	31.0	1.556	2.166	65.3
50%	9.71	31.0	2.335	2.33	70.55
75%	11.64	32.0	3.891	2.58	74.95
max	13.58	32.0	6.226	2.58	3.0

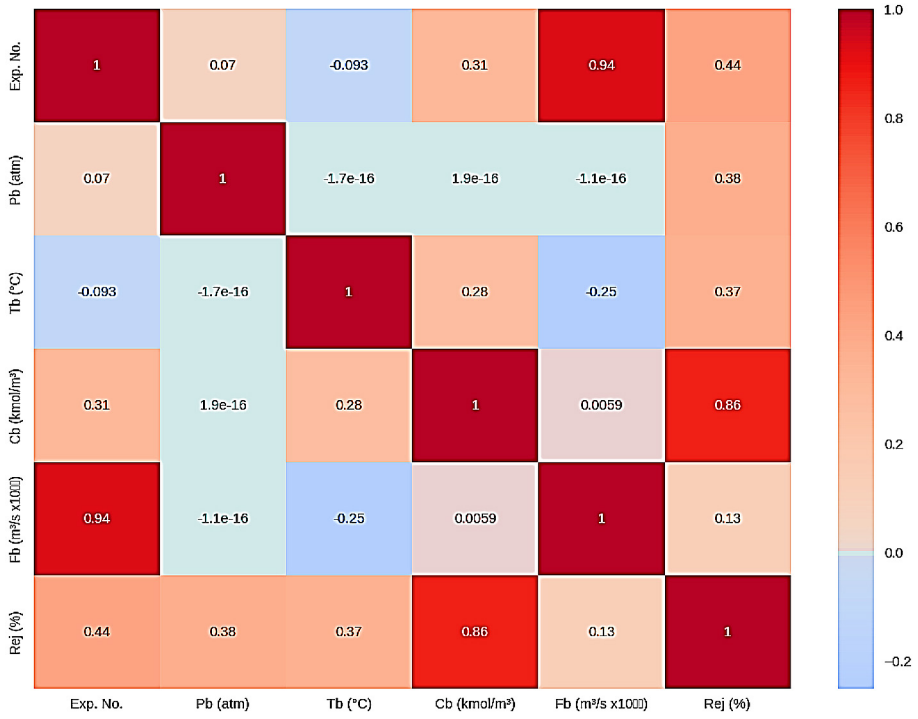

Figure 2. The correlation matrix of variables

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (1)$$

MAE measures the average absolute differences between predicted and actual values. It is calculated by Equation 2. MAE is more robust to outliers than RMSE, because it does not square the errors. A lower MAE signifies better model performance.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (2)$$

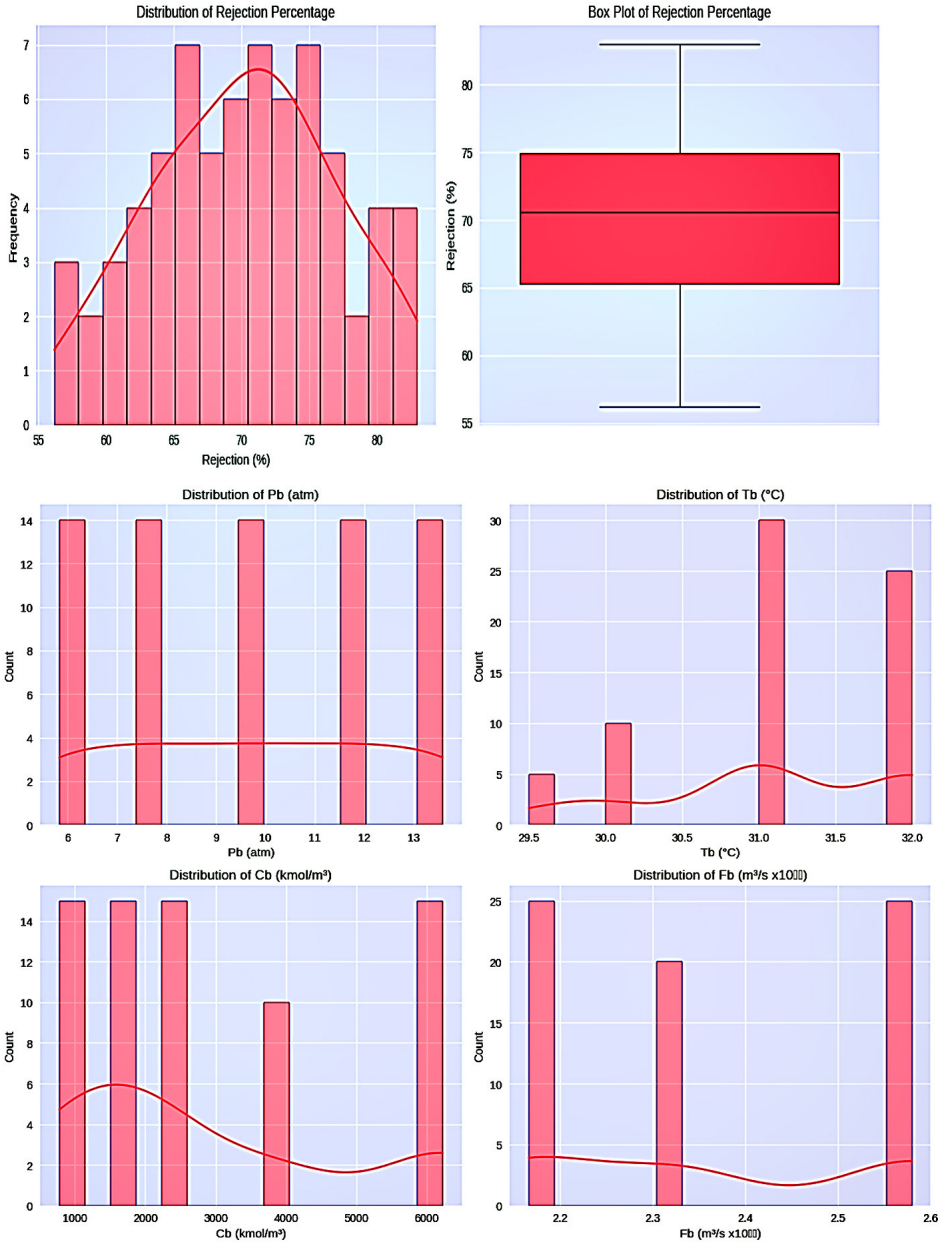

Figure 3. Data distribution of dataset features

R^2 measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It is defined by Equation 3. R^2 is between 0 to 1, where 1 denotes perfect prediction. A higher R^2 value indicates a better fit.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{3}$$

where: y_i is actual value of the ith sample,$\hat{y}_i$ is predicted value of the ith sample generated by the forecasting model, $\bar{y}_l$ is the mean of all actual values, and N is Total number of samples.

Hyper-parameter tuning

Hyper-parameter tuning is a crucial step in machine learning model development that involves finding the optimal set of hyper-parameters for a given algorithm GridSearchCV is a popular technique for this purpose. It works by exhaustively searching through a specified subset of hyper-parameters and evaluating each combination using cross-validation.

The process involves defining a parameter grid which is a dictionary where keys are the hyper-parameter names and values are lists of settings to try. For example for a Random Forest the grid might include n_estimators [50 100 200] max_depth

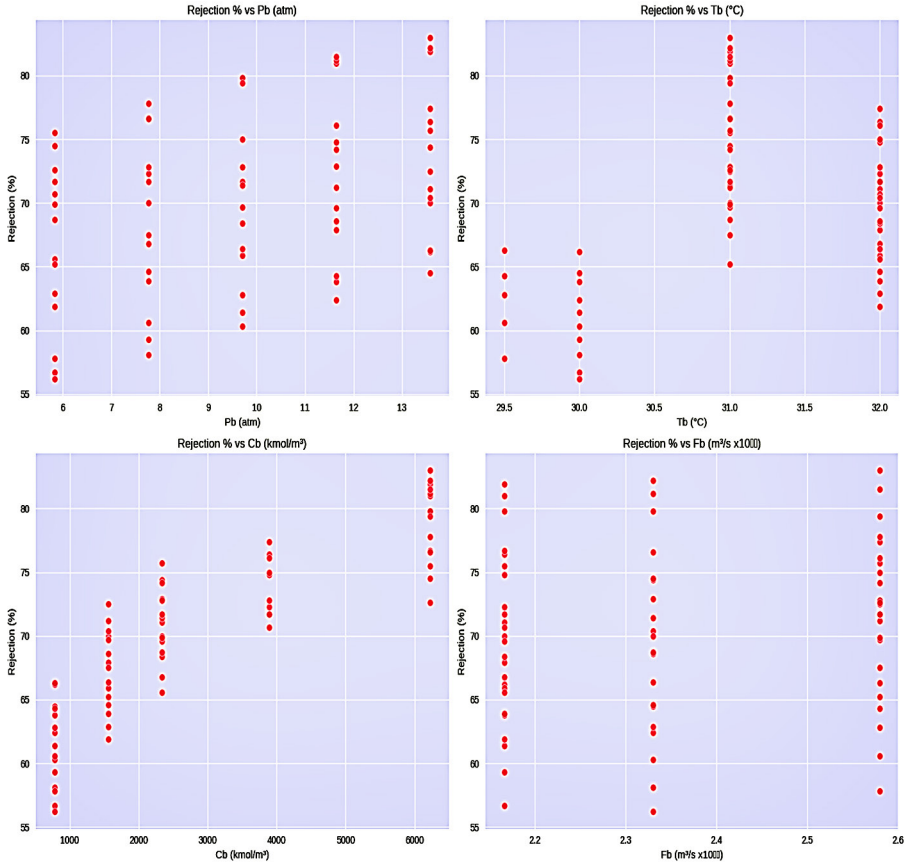

Figure 4. Distribution of each feature against the target feature

[None 10 20] and min_samples_split [2 5 10] GridSearchCV then trains and evaluates a model for every possible combination of these values.

The main advantage of GridSearchCV is its thoroughness it is guaranteed to find the best combination within the provided grid. However it can be computationally expensive, especially with a large parameter space The best parameters and the best cross-validation score can be accessed after fitting via the best_params_ and best_score_ attributes, respectively [18,19].

Feature importance analysis

This step identifies the most influential features on the rejection percentage. The process of eliminating the least relevant attributes from a dataset before applying machine learning methods to improve model efficiency is called feature selection. Excessive irrelevant features can lengthen training duration and raise the likelihood of overfitting. Feature selection picks smaller groups of attributes without compromising predictive performance. It aims to reduce dimensionality by discarding redundant and non-informative variables. Its objective is to remove unrelated characteristics to increase precision. Typically, feature selection operates by computing a metric for every attribute and selecting those with the best ratings. A minimum score can be established as a cutoff. The techniques for feature selection are commonly grouped into three types: filter methods based on statistical scores, wrapper methods that perform feature subset searches, and embedded methods that integrate feature choice during model training [20–22].

Finally, creating an optimisation surface to find the best operating conditions for maximum rejection and conducting a sensitivity analysis to understand the impact of each parameter. Then, saving the best trained model for future use.

RESULTS AND DISCUSSION

The Sundaramoorthy dataset incorporated various input parameters of the *Fb*, Pb, Tb, Cb, and response parameter of chlorophenol Rej. Seventy data points from this dataset were used by various machine learning techniques and optimisation methods. In this dataset, 80% of the

input data is used for training, and the remaining 20% is used for testing.

The results shown in Table 5 and Figure 5 reveal that the Gradient Boosting algorithm achieved the best performance in predicting chlorophenol removal, followed by the XGBoost algorithm. Neural networks, on the other hand, recorded the worst performance, with a negative R^2 value indicating a significant unsuitability between the model and the data. It is noteworthy that machine learning models, such as Gradient Boosting and XGBoost, demonstrated high prediction accuracy based on input factors (temperature, concentration, pressure, and flow rate), as evidenced by the low RMSE and MAE values and R^2 close to one. The table also shows that the performance of these algorithms remained stable during cross-sectional verification (CV), as evidenced by the CV R^2 values with a relatively low standard deviation, confirming the robustness of these models.

When assessing the importance of features on prediction performance, the results demonstrated as shown in Table 6 and Figure 6 that the most important feature is Cb with 0.773, Pb is the second most important feature, while temperature and flow rate are less influential. The associated results show concentration dominates the model's predictions significantly.

Figure 7 (the left part) demonstrates the actual and predicted rejection percentage. It is observable from this figure that points seem to closely cluster along a diagonal line, indicating a high degree of agreement between the data and the model's predictions. This suggests that, for this dataset, the model has good predictive accuracy.

Table 5. The performance metrics of ML techniques

Model	RMSE	MAE	R^2	CV R^2 Mean	CV R^2 Std
Gradient boosting	0.805690	0.657650	0.985633	0.969554	0.015233
XGBoost	0.960817	0.807506	0.979569	0.946920	0.014165
Ridge regression	1.250577	1.092827	0.965387	0.906605	0.031382
Linear regression	1.254757	1.064630	0.965155	0.910361	0.021730
Random forest	1.476845	1.248643	0.951729	0.938224	0.021532
Lasso regression	1.920516	1.301985	0.918370	0.848812	0.043266
Decision tree	2.030130	1.557143	0.908785	0.890561	0.036227
Support vector regression	4.411005	3.211346	0.569383	0.628199	0.047602
Neural network	9.027496	6.636322	-0.803645	-0.984923	0.696084

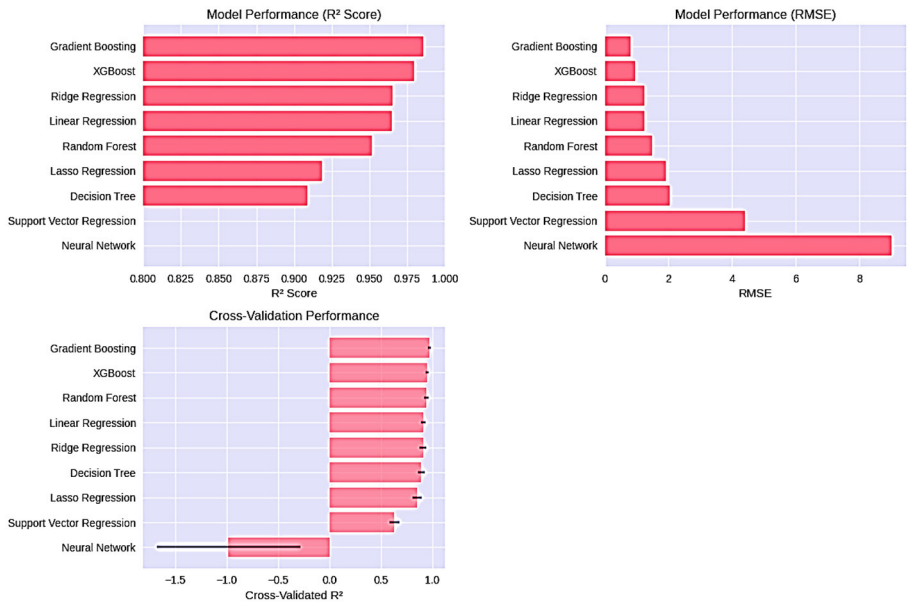

Figure 5. Performance metrics of ML techniques

Table 6. Feature importance

Cb (kmol/m³)	0.773104
Pb (atm)	0.130159
Tb (°C)	0.077735
Fb (m³/s x10^{-4})	0.019002

On the other hand, the right part of Figure 7 is residual plot which shows the prediction errors. The residuals are randomly scattered around zero without forming any clear patterns. This random scatter is ideal and suggests that the model's assumptions are likely met, meaning it is equally accurate across the range of predicted values and that there is no systematic error.

The results of Figure 8 show that the optimal conditions to achieve the maximum predicted rejection of 82.37%. These ideal parameters are a pressure of 12.76 atm, a very high concentration of 5079.05 kmol/m³, a moderate temperature of 31.11 °C, and a low flowrate of 2.36 × 10^{-4} m³/s.

Performing a one-at-a-time sensitivity analysis involves systematically varying each input feature (Pressure, Concentration, Temperature, Flow Rate) across its realistic range while holding the others constant at baseline values, then recording the resulting change in the target output, rejection percentage. This quantifies the individual influence of each parameter and identifies which feature exerts the strongest control over the process performance. Figure 9 with Table 7 show the sensitivity analysis of each feature and the change in the target feature. It is noticeable from Table 7 that the feature pressure increase from 12.76 to 13.8 atm caused a -7.57% change in rejection. This shows high sensitivity and pressure is an effective performance factor. The concentration feature had a moderate sensitivity when increased from 5079.05 to 6500 kmol/m³ resulting change from baseline of -3.47%. Both the features temperature and flow rate show high sensitivity caused a -10.87% and -8.87%, respectively change in rejection.

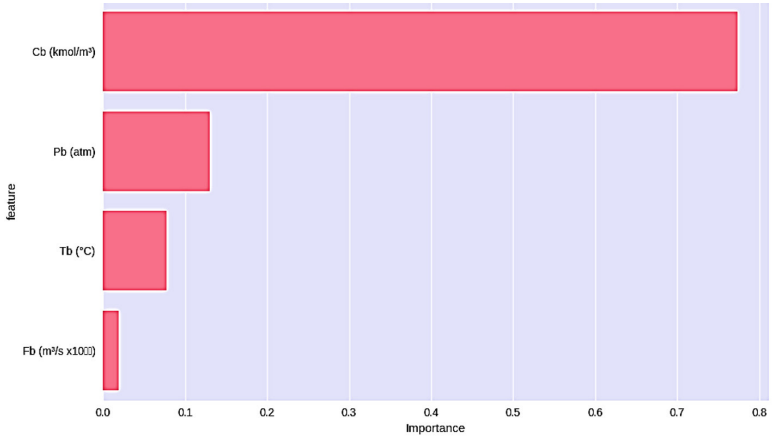

Figure 6. Feature importance

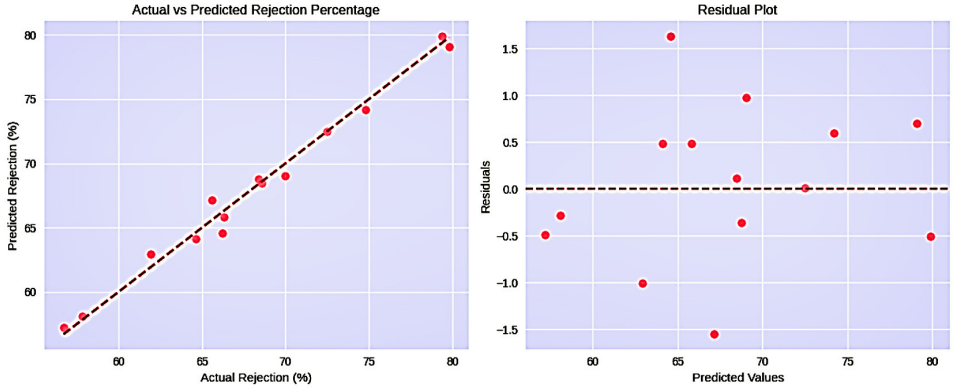

Figure 7. Visual analysis of model performance

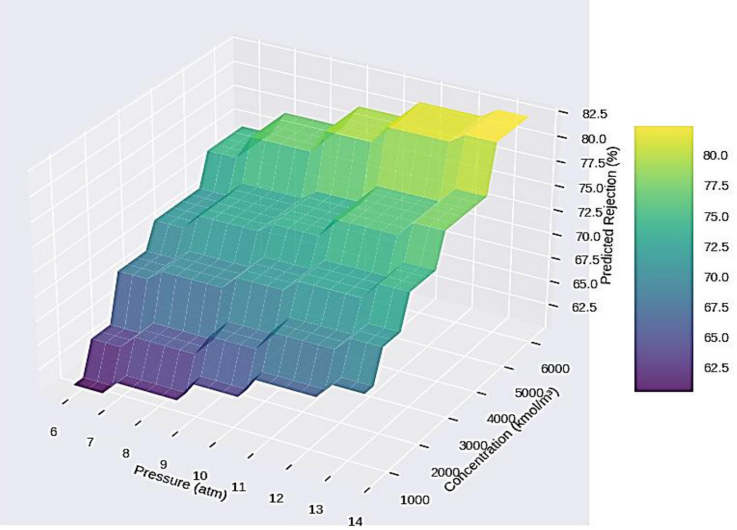

Figure 8. Optimal conditions for maximum rejection

Table 7. Sensitivity analysis

Analysis step	Variable changed	Baseline values	New value	Resulting rejection	Change from baseline	Sensitivity
Baseline	(All at baseline)		-	82.37%	-	-
1	Pressure	12.76 atm	Increased to 13.8 atm	74.8%	-7.57%	High
2	Concentration	5079.05 kmol/m³	Increased to 6500 kmol/m³	78.9%	-3.47%	Moderate
3	Temperature	31.11 °C	Increased to 31.25 °C	71.5%	-10.87%	High
4	Flow rate	2.36 × 10^{-4} m³/s	Increased to 2.58 x10^{-4} m³/s	73.5%	-8.87%	High

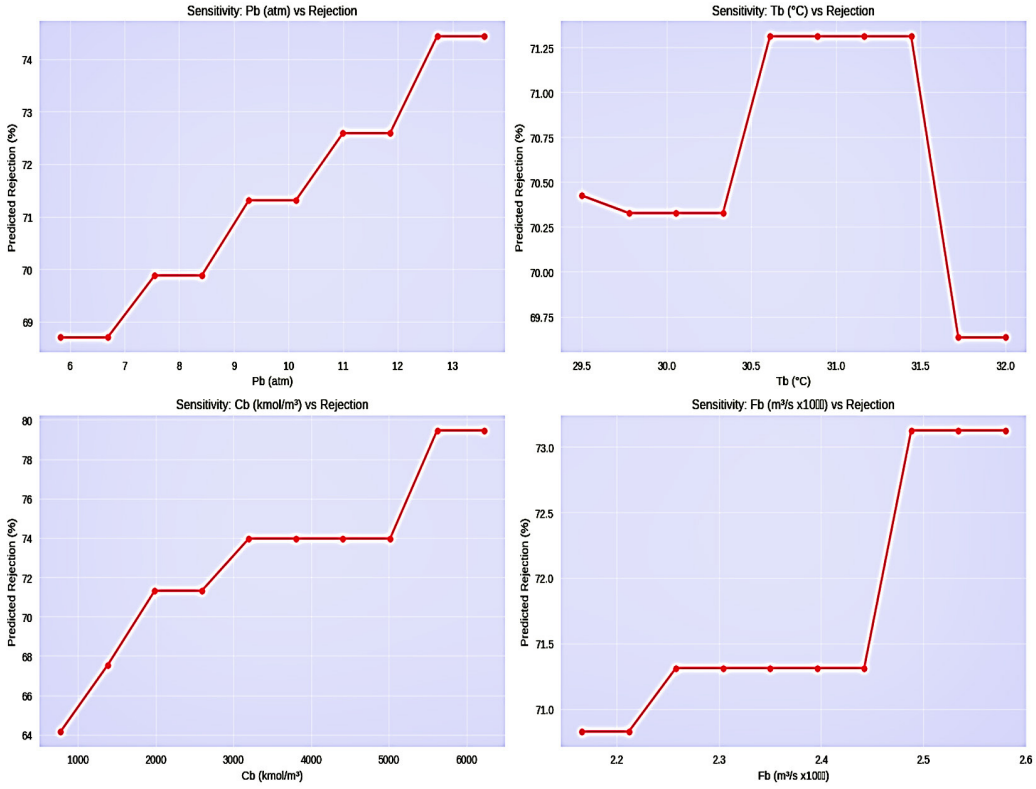

Figure 9. Sensitivity analysis

CONCLUSIONS

ML models show great promise in the quickly developing field of predicting the removal of chlorophenol from wastewater by reverse osmosis. This study examined optimisation of machine learning techniques for the removal of chlorophenol from wastewater by reverse osmosis. The goal was to determine the ideal operating conditions and predict the rejection percentage. The results showed that random forest and gradient boosting provided excellent performance in predicting chlorophenol rejection percentages. The findings also demonstrated that feed concentration and pressure are the most significant factors, while temperature and flow rate have moderate influence. The optimal conditions identified in this research can guide process optimisation, higher pressure generally leads to better rejection and moderate to high concentrations can be effectively treated. The models in this paper can be used for process optimisation, helps in determining the most efficient operating conditions, and reduces the need for extensive experimental trials.

REFERENCES

1. Al-Obaidi M.A., Kara-Zaitri C., Mujtaba I.M. Simulation and optimisation of a two-stage/two-pass reverse osmosis system for improved removal of chlorophenol from wastewater. J. Water Process Eng. 2018;22:131–137. https://doi.org/10.1016/j.jwpe.2018.01.012
2. Mohammad A.T., Al-Obaidi M.A., Hameed E.M., Basheer B.N., Mujtaba I.M. Modelling the chlorophenol removal from wastewater via reverse osmosis process using a multilayer artificial neural network with genetic algorithm. J. Water Process Eng. 2020;33:100993. https://doi.org/10.1016/j.jwpe.2019.100993
3. Abd Gami A., Shukor M.Y., Khalil K.A., Dahalan F.A., Khalid A., Ahmad S.A. Phenol and its toxicity. J. Environ. Microbiol. Toxicol. 2014;2:11–23.
4. Japan Environmental Governing Standard. Department of Defence, Chapter 4: Wastewater. Headquarters, US Forces Japan, Japan; 2016.
5. Mohammed A.E., Jarullah A.T., Gheni S.A., Mujtaba I.M. Optimal design and operation of an industrial three phase reactor for the oxidation of phenol. Comput. Chem. Eng. 2016;94:257–271. https://doi.org/10.1016/j.compchemeng.2016.07.018
6. Al-Obaidi M.A., Kara-Zaitri C., Mujtaba I.M. Evaluation of chlorophenol removal from wastewater using multi-stage spiral-wound reverse osmosis process via simulation. Comput. Chem. Eng. 2019;130:106522. https://doi.org/10.1016/j.compchemeng.2019.106522
7. Gholizadeh M., Saeedi R., Bagheri A., Paeezi M. Machine learning-based prediction of effluent total suspended solids in a wastewater treatment plant using different feature selection approaches: A comparative study. Environ. Res. 2024;246:118146. https://doi.org/10.1016/j.envres.2024.118146
8. Warren-Vega W.M., Montes-Pena K.D., Romero-Cano L.A., Zarate-Guzman A.I. Development of an artificial neural network (ANN) for the prediction of a pilot scale mobile wastewater treatment plant performance. J. Environ. Manag. 2024;366:121612. https://doi.org/10.1016/j.jenvman.2024.121612
9. Maqsood Q., Fatima R., Rafaqat F., Mehmood T., Ali S.W., Hussain M. Revolutionizing water and wastewater treatment: AI and machine learning-driven approaches for advanced solutions. Desalin. Water Treat. 2025:101432. https://doi.org/10.1016/j.dwt.2025.101432
10. Caglar Gencosman B., Eker Sanli G. Prediction of polycyclic aromatic hydrocarbons (PAHs) removal from wastewater treatment sludge using machine learning methods. Water Air Soil Pollut. 2021;232:87. https://doi.org/10.1007/s11270-021-05049-8
11. Zamfir F.S., Carbureanu M., Mihalache S.F. Application of machine learning models in optimizing wastewater treatment processes: A review. Appl. Sci. 2025;15:8360. https://doi.org/10.3390/app15158360
12. Wang X., Sun X., Wu Y., Gao F., Yang Y. Optimizing reverse osmosis desalination from brackish waters: predictive approach employing response surface methodology and artificial neural network models. J. Membr. Sci. 2024;704:122883. https://doi.org/10.1016/j.memsci.2024.122883
13. Ahmad U., Abdala A., Ng K.C., Akhtar F.H. Machine learning-driven design of membranes for saline and produced water treatment across scales. Environ. Sci.: Water Res. Technol. 2025;11:2080–2099. https://doi.org/10.1039/d5ew00417a
14. Al-Obaidi M., Al-Nedawe B., Mohammad A., Mujtaba I. Response surface methodology for predicting the dimethylphenol removal from wastewater via reverse osmosis process. Chem. Prod. Process Model. 2021;16:193–203. https://doi.org/10.1515/cppm-2020-0025
15. Sundaramoorthy S., Srinivasan G., Murthy D.V.R. Reprint of: An analytical model for spiral wound reverse osmosis membrane modules: Part II—experimental validation. Desalination 2011;280:432–439. https://doi.org/10.1016/j.desal.2011.03.047

16. Hameed E.M., Joshi H. Performance comparison of machine learning techniques in prediction of diabetes risk. AIP Conf. Proc. 2024;3051(1). https://doi.org/10.1063/5.0191611
17. James G., Witten D., Hastie T., Tibshirani R., Taylor J. Linear Regression. In: An Introduction to Statistical Learning. Springer Texts in Statistics. Springer International Publishing, Cham; 2023. p. 69–134. https://doi.org/10.1007/978-3-031-38747-0_3
18. Hameed E.M., Joshi H., Ismael A.A.A. The effect of combining datasets in diabetes prediction using ensemble learning techniques. CommIT J. 2025;19:129–140. https://doi.org/10.21512/commit.v19i1.12064
19. Barahi S., Azizi A., Taky M., Belhamidi S. Machine learning algorithms in wastewater technology: Predicting treatment quality and efficiency. J. Water Land Dev. 2025:137–151. https://doi.org/10.24425/jwld.2025.155310
20. Feurer M., Hutter F. Hyperparameter Optimization. In: Automated Machine Learning: Methods, Systems, Challenges. Springer International Publishing, Cham; 2019. p. 3–33. https://library.oapen.org/bitstream/handle/20.500.12657/23012/1/1007149.pdf
21. Acikkar M. Fast grid search: A grid search-inspired algorithm for optimizing hyperparameters of support vector regression. Turk. J. Electr. Eng. Comput. Sci. 2024;32:68–92. https://doi.org/10.55730/1300-0632.4056
22. Alexandropoulos S.A.N., Kotsiantis S.B., Vrahatis M.N. Data preprocessing in predictive data mining. Knowl. Eng. Rev. 2019;34:e1. https://doi.org/10.1017/S026988891800036X