Revealing key parameters and mechanisms governing membrane selectivity for dye/NaCl mixtures by explainable machine learning

Citation for the original published paper (version of record):
Xue, Y., Chew, J. (2026). Revealing key parameters and mechanisms governing membrane selectivity for dye/NaCl mixtures by
explainable machine learning. Journal of Membrane Science, 757.
http://dx.doi.org/10.1016/j.memsci.2026.125908

Journal of Membrane Science 757 (2026) 125908

Contents lists available at ScienceDirect

Journal of Membrane Science

journal homepage: www.elsevier.com/locate/memsci

Revealing key parameters and mechanisms governing membrane selectivity for dye/NaCl mixtures by explainable machine learning

Yongtao Xue, Jia Wei Chew *

Department of Chemistry and Chemical Engineering, Chalmers University of Technology, Gothenburg, 412 96, Sweden

ARTICLE INFO

Keywords:
Membrane separation
Dye/NaCl mixtures
Machine learning
Feature importance
Interpretable model

ABSTRACT

Membrane selectivity is crucial for the separation of dye/NaCl mixtures. However, because separation performance is governed by the interplay of multiple factors, including membrane properties, operating conditions, and feed characteristics , identification of the key parameters and underlying mechanisms remains challenging. To address this, this study employed several machine learning methods to predict membrane selectivity and investigatethe mechanisms governing separation. Data were extracted from 157 peer-reviewed papers. Among the four models employed (namely, Random Forest (RF), Support Vector Regression (SVR), Artificial Neural Network (ANN), and Extreme Gradient Boosting (XGB)), XGB gave the best predictive performance with R^2 values of approximately 0.7. SHapley Additive exPlanations (SHAP) analysis was applied to evaluate the effect of dye and membrane properties as well as operation-related parameters. The maximum projection radius of the dye, membrane pore radius, membrane water contact angle, and effective membrane area were identified as the four predominant features influencing the separation factors of dye/NaCl mixtures. Model results indicated that high separation factors were achieved under specific conditions, namely, membrane pore radius below 4 nm, membrane water contact angle below 25°, maximum dye projection radius above 1.2 nm, and effective membrane area below 20 cm^2. Overall, this study demonstrates the potential of machine learning for predicting membrane selectivity, and providing new insights for the optimization of membrane properties and operating conditions for the treatment of textile wastewater.

1. Introduction

Textile industries produce large quantities of high-salinity wastewater, mainly containing various dyes and inorganic salts (mainly NaCl), which can cause severe ecological and environmental impacts if directly discharged without any post-treatment [1,2]. However, it is challenging to treat the high-salinity wastewater through conventional biological, physical, and chemical methods [3]. In addition, the high concentration of inorganic salts in textile wastewater is a valuable product that can be recovered and reused [4]. Therefore, developing a method that can efficiently remove dyes and recovery of salts from textile wastewater is crucial for sustainable wastewater treatment. To address this, membrane-based separation has emerged as a promising technique for the treatment of textile wastewater due to its high removal effectiveness, high selectivity, as well as energy efficiency [5,6].

Recently, remarkable progress has been made in membrane customization for efficient separation of dye-salt mixtures. However, the properties of dyes vary greatly, resulting in significant differences in membrane selectivity for different types of wastewaters. For instance, Li et al. [7] reported that the prepared MXene/sodium alginate composite membrane demonstrated an excellent separation factor (783.1) for a mixture of Evans blue-NaCl, but also exhibiting a very low separation factor (3.53) for a mixture of Crystal violet-NaCl under the same conditions. In addition, membrane properties and operating conditions can also affect the separation efficiency of mixtures of dye/NaCl. Liu et al. [8] observed that the prepared composited membrane exhibited two different separation factors of 219.4 and 122 for mixtures of brilliant blue G-NaCl at the dye concentrations of 10 ppm and 100 ppm, respectively. In summary, membrane properties (e.g., pore radius, hydrophilicity, surface roughness), dye properties (molecular size, charge), and operating conditions (e.g., concentrations, pressure) are important in dictating the efficient separation of dye and NaCl [9,10]. Unfortunately, experimental methods alone are limited in optimizing membrane selectivity for mixtures of dyes-NaCl due to the large number of parameters and complex underlying separation mechanisms, which require a huge amount of experiments and thus long development

* Corresponding author.
E-mail address: jia.chew@chalmers.se (J.W. Chew).

https://doi.org/10.1016/j.memsci.2026.125908
Received 20 May 2026; Received in revised form 4 July 2026; Accepted 9 July 2026
Available online 9 July 2026

cycles.

The advances of data-driven methods provide promising tools for elucidating the relationships among the myriad of variables and outcomes, as well as the underlying mechanisms [11,12]. Recently, machine learning has made significant progress in the field of environmental remediation research. For instance, Hu et al. [13] established a robust Extreme Gradient Boosting (XGBoost) model to predict the rejection efficiency of organic micropollutants by nanofiltration and reverse osmosis membranes. Ji et al. [14] achieved excellent lithium permeation percentage (90%) and Li^+/Mg^{2+} selectivity (158) by designing and optimizing membranes using machine learning. Moreover, machine learning has been applied to predict the separation efficiency of membranes for dyes and salts, as well as to identify the key parameters governing separation efficiency. For instance, Liu et al. [15] developed density functional theory (DFT)-assisted machine learning models to predict membrane separation performance for dye/salt and dye/dye systems. Similarly, Zhang et al. [16] used Shapley Additive exPlanations (SHAP) to demonstrate that membrane properties played a dominant role in dye/salt separation based on systems involving four salts (namely, NaCl, Na_2SO_4, $MgCl_2$, $MgSO_4$) and two dyes (namely, Congo red, Direct red 23).

Although previous studies have demonstrated the feasibility of employing machine learning to predict dye/salt separation performance, the mechanistic understanding of membrane selectivity in dye/NaCl mixtures remains incomplete. Existing studies have generally focused on specific dye/salt systems, multiple salt types, or selected membrane materials, providing valuable data-driven insights. However, the simultaneous consideration of different salts, or the use of limited dye systems, makes it difficult to isolate the role of dye molecular properties in governing membrane selectivity. In addition, most reported models have primarily emphasized performance prediction, whereas the underlying selectivity mechanisms and the intertwined effects of dye characteristics, membrane properties, and operating conditions remain insufficiently clarified. In practical textile wastewater, NaCl is one of the most representative salts; therefore, the selective separation of diverse dyes from NaCl is crucial for dye recovery and salt reuse. However, dye/NaCl separation is highly sensitive to variations in dye molecular structures, membrane properties, and experimental conditions. Accordingly, the value of machine-learning models in this field should not be assessed solely by numerical prediction accuracy, but also by their ability to capture general separation trends, identify key governing parameters, and provide interpretable insights into membrane selectivity. Based on this background, this study develops a systematic and interpretable machine-learning framework to analyze dye/NaCl separation performance, targeted at advancing the understanding of the relationships among membrane performance, operating conditions, and the selective permeation behavior of dyes and NaCl.

Herein, a comprehensive dataset of membrane selectivity for a mixture of dyes-NaCl was curated, including separation factors, membrane properties (e.g., pore radius, hydrophilicity, surface roughness), dye properties (molecular size, charge), and operating conditions (e.g., concentrations, pressure). Subsequently, four different models, including Random Forest (RF), Support Vector Regression (SVR), Artificial Neural Network (ANN), and Extreme Gradient Boosting (XGB), were employed to predict the separation efficiency of a mixture of dyes-NaCl, with coefficient of determination (R^2) and root mean square error (RMSE) calculated to evaluate the performance of the models. Furthermore, SHAP, partial dependence plots, and individual conditional expectation (ICE) were performed to investigate the key parameters on membrane selectivity and the underlying mechanisms during the separation. This study is expected to provide new insights into optimizing membrane performance for dye/NaCl mixtures, thereby advancing sustainable wastewater treatment.

2. Materials and methods

2.1. Data collection and preprocessing

The workflow of this study is illustrated in Fig. 1. Firstly, a comprehensive literature survey on the membrane separation studies for dyes/NaCl mixture was conducted through the utilization of various databases (e.g., Web of Science, Google Scholar, and Scopus) with numerous keywords (e.g., membrane, separation, dye and salt mixtures). To restrict the scope for training the models, peripheral studies, such as electrochemical membranes, solar-driven membrane-separation, membrane distillation, effect of temperature and pH, were not considered. The separation factor (SF) was calculated by Eq. (1).

$$SF = \frac{100 - R_{NaCl}}{100 - R_{dye}} \tag{1}$$

where R_{NaCl} and R_{dye} are the rejection percentages for NaCl and dyes, respectively.

A total of 157 peer-reviewed literature studies were collected (Table S1), resulting in the collation of 527 data rows (total of 4712 data points). Most of the data points were directly extracted from experimental tables or charts. For the data presented in figures, the WebPlotDigitizer web platform (https://apps.automeris.io/wpd4/) was used to extract the data points. All extracted data were directly used. The datasets mainly include membrane properties (e.g., pore radius, zeta potential, water contact, surface roughness) and operating conditions (e.g., the type of reactor, NaCl concentration, dye concentration, pressure, and effective membrane area), which are well-acknowledged to significantly affect the separation factors for dye/NaCl mixtures. For articles that provide MWCO value without pore radius, the pore radius (r_p, nm) was calculated using Eq. (2) [13].

$$r_p = \left(\frac{MWCO}{1163.7}\right)^{\frac{1}{1.58}} \tag{2}$$

In addition, a total of 44 different kinds of dyes are accounted for in the dataset, with surface charge (charge value at pH = 7), hydrophilicity (logD at pH = 7, and D denotes the octanol-water distribution coefficient), and molecular size (maximum projection radius) properties obtained from Pubchem and Chemicalize, as summarized in Table S2. Missing values in the dataset were bridged using the multiple imputations by chained equations (MICE) algorithm, and extreme randomized trees (ETR) were used as a nonlinear model to predict the missing values. To evaluate the accuracy and effectiveness of the method, 10% of the known data were randomly masked, and then the masked data were imputed again using the same method, and the values of the coefficient of determination (R^2) were calculated (all R^2 >85% in this study). Since the MICE-generated data may be disagree with the curated data, manual checking was performed to ensure consistency.

2.2. Feature engineering and model construction

Feature engineering was applied to increase the credibility and performance of machine learning models through the transformation and selection of related features. In this study, the type of membrane operation (i.e., dead-end or cross-flow) was coded by categories for subsequent model development. In addition, Pearson correlation coefficient analysis was conducted to identify the correlations between features. The Pearson correlation coefficient (R_p) was calculated using Eq. (3) [17].

$$R_p = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\sum_{i=1}^{n}(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{3}$$

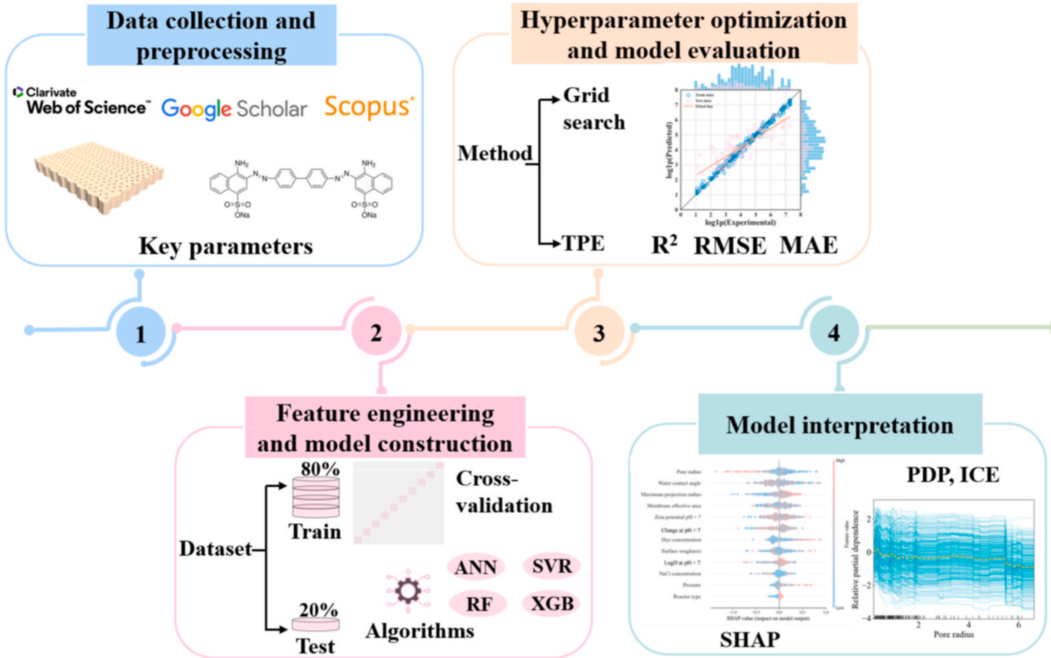

Fig. 1. The machine-learning workflow includes data collection and preprocessing, feature engineering and model construction, hyperparameter optimization and model evaluation, and model interpretation.

where $\overline{x}$ and $\overline{y}$ represent the average values of the input variables, and n refers to the number of data points.

Furthermore, the dataset was randomly split into a training set and a test set at a ratio of 80% to 20%. Subsequently, the 10-fold cross-validation was applied to the training set to prevent the overfitting of the model. Specifically, the training data was divided into ten subsets, with nine subsets used for training and one subset reserved for validation in each iteration, thereby ensuring comprehensive training and evaluation of the model. To reduce the randomness of dataset splitting and increase model accuracy, each calculation was repeated 12 times.

In this study, four machine learning algorithms, including Random Forest (RF), Support Vector Regression (SVR), Artificial Neural Network (ANN), and Extreme Gradient Boosting (XGB), were employed and the corresponding results compared. The RF model is a supervised learning algorithm based on ensemble learning through multiple decision trees. It has some advantages, including handling complex multivariable data, managing non-linear relationships, and preventing model overfitting [18]. The SVR model is widely applied for the investigation of nonlinear relationships through kernel-based functions to map data into high-dimensional feature spaces, offering advantages such as low computational complexity and strong generalization capabilities [19]. ANN is a computational algorithm inspired by the structural and functional principles of the human nervous system. By iteratively adjusting weights through backpropagation and gradient descent, the ANN minimizes the deviation between predicted and observed outcomes [20]. XGB is another popular machine learning algorithm based on the gradient boosting framework, which combines multiple weak predictive models (e.g., decision trees) into a strong predictive model to improve the prediction performance [21]. The performance of each model was evaluated based on the coefficient of determination (R^2) and root mean square error (RMSE). The R^2 and RMSE were calculated using Eq. (4) and Eq. (5), respectively [22].

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y}_i)^2} \tag{4}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2} \tag{5}$$

where y_i represents the experimental value, $\overline{y}_i$ is the average value of experimental values, and $\widehat{y}_i$ is the predicted values obtained from machine learning models. The grid search method was employed for hyperparameter tuning. The parameter ranges for different models when employing grid search are summarized in Table S3.

2.3. Model explanation

In this study, SHAP were employed to identify the key driving factors for the model predictions. The SHAP values explain how the input features influence the target variable, facilitating the identification of the most important features on the model prediction, as well as quantification of the relative contributions of the input features on the model output [23]. In addition, the negative or positive effects of the features on the target variable can also be observed. Therefore, SHAP provides valuable insights into how various features influence separation factors, offering understanding on the importance of each feature. In addition, the individual conditional expectation (ICE) plots were applied to analyze the relationship between the prediction value and a particular feature while keeping all other features unchanged. The ICE plots use each line to represent a sample, showing how the prediction depends on the features of each sample [24]. While partial dependence plots (PDP) also show the impacts of features on model predictions, they show average effects between input features and the target variable [25]. Furthermore, bivariate partial dependence plots highlight the interactions between two features and the outcome.

3. Results and discussion

3.1. Statistical analysis of variables

Fig. 2a–l presents the box and scatter plots of twelve variables, including membrane properties (pore radius, zeta potential, water contact angle, surface roughness), operating conditions (NaCl

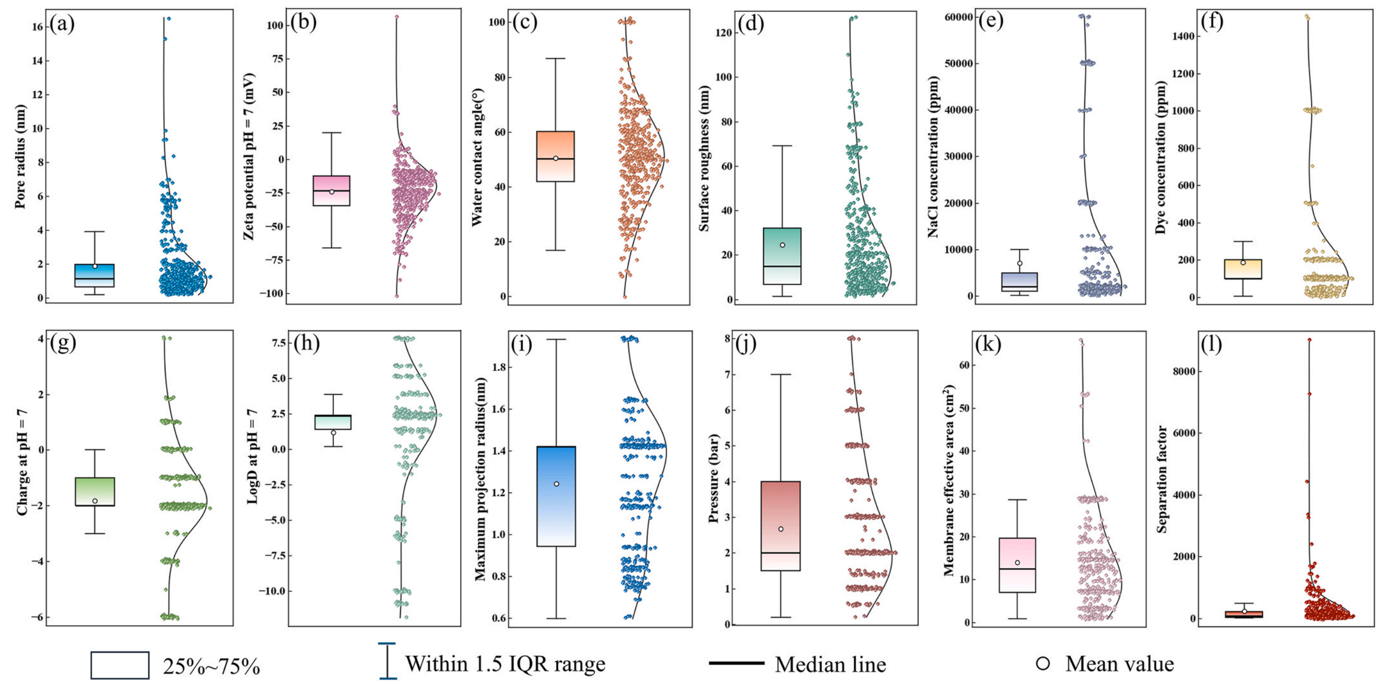

Fig. 2. Box and scatter plots of the reported membrane separation process variables: pore radius (a), zeta potential (b), water contact angle (c), surface roughness (d), NaCl concentration (e), dye concentration (f), charge values at pH = 7 of dyes (g), logD values at pH = 7 of dyes (h), maximum projection radii of dyes (i), pressure (j), effective membrane area (k), and separation factor (l).

concentration, dye concentration, pressure, effective membrane area), dye properties (charge at pH = 7, logD at pH = 7, maximum projection radius), and separation factors. The pore radii of most membranes are concentrated between 0.63 nm and 1.96 nm, and the average pore radius is 1.85 nm (Fig. 2a), indicating that most of the membranes in the dataset are nanofiltration membranes. The range of the zeta potential of the membranes is extensive, spanning from negative charges (−101.6 mV) to positive charges (106.2 mV), with the charges primarily negative and with an average value of −24.2 mV (Fig. 2b). The water contact angle exhibits a uniform distribution, with most values between 42° and 60° (Fig. 2c), suggesting that the studied membranes are largely hydrophilic. In Fig. 2d, most membranes exhibit surface roughness of between 6.8 nm and 32.1 nm, which is characteristic of relatively smooth membrane surfaces. In summary, the membranes in the dataset for the separation mixture of dye and salt are predominantly negatively charged, hydrophilic, and smooth nanofiltration membranes. In addition, most NaCl concentrations are below 10,000 ppm (Fig. 2e), while most dye concentrations are below 200 ppm (Fig. 2f), which aligns with the conditions found in actual textile wastewater. Fig. 2g–i demonstrates a uniform distribution of dye properties, including charge at pH = 7, logD at pH = 7, and maximum projection radius. Fig. 2j shows that these separation experiments were usually conducted at relatively low pressures, typically between 1.5 bar and 4 bar. Additionally, almost all the membranes' effective areas were below 30 cm^2 (Fig. 2k), which indicates that these experiments were mainly conducted at the laboratory scale. The separation factors are mostly below 2000, with some larger values also reported, giving the mean value of the separation factor as 232. The most frequently studied dye is Congo red, accounting for approximately 40%, followed by methyl blue (9.5%) and rhodamine B (4.5%). Moreover, approximately 64% of the separation experiments in the dataset were conducted using cross-flow systems, which has some advantages, such as reduced fouling, continuous operation, and higher flux stability compared to dead-end systems. To further explore the combined effects of variables on the separation, multivariate scatter plots are presented in Fig. S1. The separation factor shows considerable variation across the examined paired-variable combinations, rather than exhibiting a clear monotonic trend with any specific two-variable combination. This suggests that dye/NaCl separation performance cannot be fully explained by a limited number of isolated variables. Instead, it is likely governed by the interplay of operating conditions, membrane properties, and dye characteristics. Therefore, explainable machine-learning analysis was further employed to provide a more comprehensive interpretation of feature contributions and nonlinear dependencies in dye/NaCl separation.

Prior to model construction, Pearson correlation coefficient analysis was performed for the input features to assess potential multicollinearity among all variables, as high correlation between two variables can compromise the stability and interpretability of the model. In Fig. 3, Pearson correlation heatmaps use both color intensity and bubble size to visualize correlations, with red indicating positive correlation and blue indicating negative correlation, where greater color intensity and larger bubbles represent stronger correlations. The Pearson correlation coefficient ranges from −1 to 1, and the values greater than 0.8 or less than −0.8 are generally considered to indicate a high correlation between the two variables [26]. For instance, the heatmaps demonstrate a negative correlation (−0.71) between charge values at pH = 7 and maximum projection radius of the dyes, whilst exhibiting a positive correlation (0.65) between charge values at pH = 7 and log D values at pH = 7 of the dyes. In Fig. 3, all the absolute values of Pearson correlation coefficients are below 0.8, indicating strong independence of the features. In addition, the correlations between the dye/NaCl separation factor and other variables are relatively low, which suggest weak linear relationships.

3.2. Model development and evaluation

Considering the broad distribution and the existence of a few large values for the separation factor (Fig. 2l), two common methods, including log-transformation and winsorization, were applied in the model development. Applying the log1p transformation to standardize the separating factor values can make the data distribution more concentrated and improve the stability of the model [27]. In addition, winsorization is an efficient process that can replace outliers and

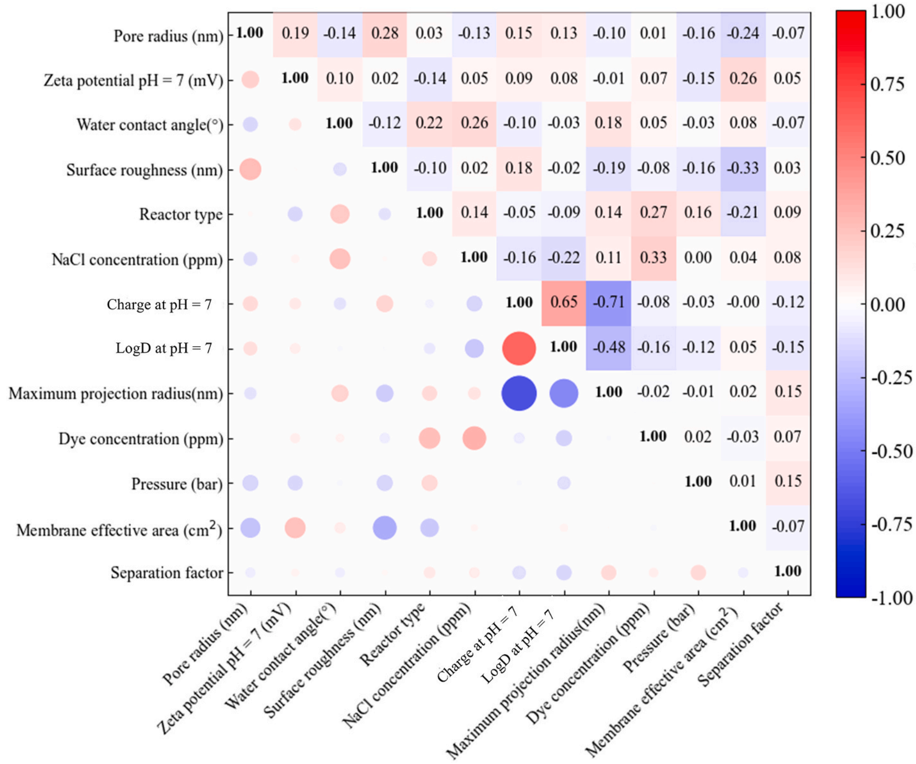

Fig. 3. The Pearson correlation matrix of input features and separation factor.

extreme values with specific values, thereby decreasing the effects of outliers on the models [28]. As illustrated in Fig. 4a–d and Fig. S2, log-transformed outcomes were applied for the prediction of separation factors using four different models, namely, RF, SVR, ANN, and XGB. The y = x line indicates perfect agreement between predicted and experimental values, whereas the histogram reflects the distribution of the variable. Clearly, the XGB model results in the highest average R^2 and lowest RMSE (0.5557, 0.9607), compared with RF (0.5308, 0.9906), SVR (0.4340, 1.0857), and ANN (0.2670, 1.2379). With regards to winsorization, Fig. S3–S4 show that the RF model displays the highest

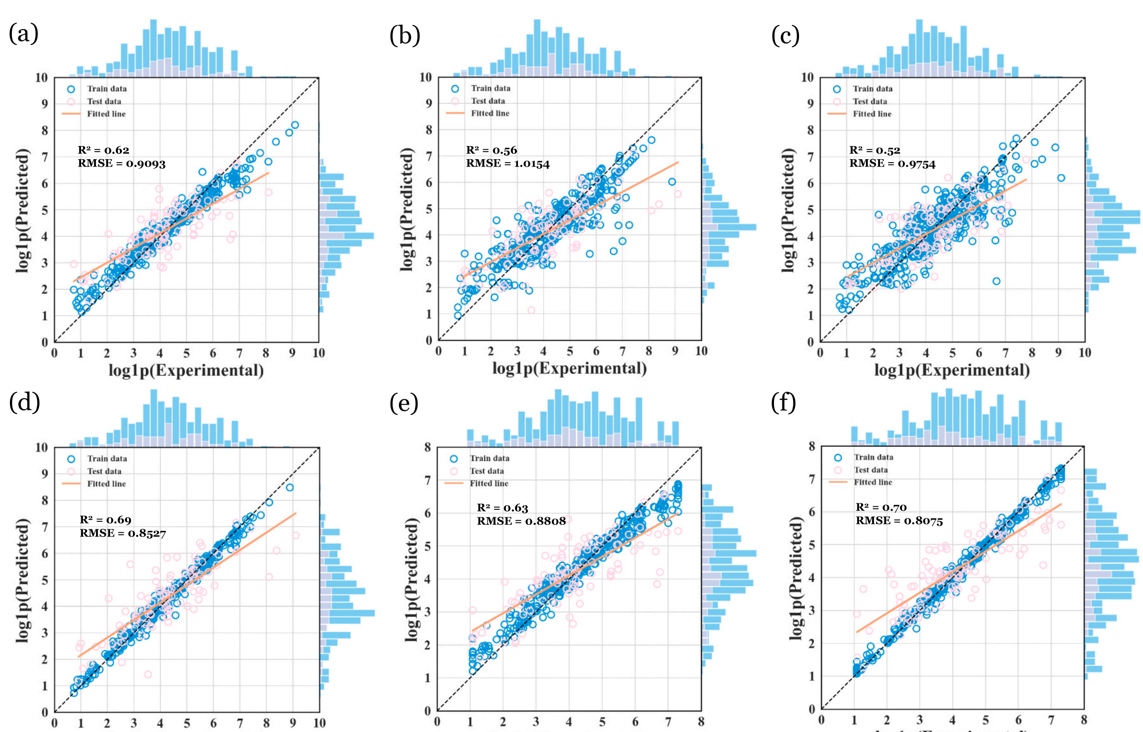

Fig. 4. Comparison of predicted vs. experimental values for different models with log-transformed outcomes: RF (a), SVR(b), ANN(c), XGB(d), RF with winsorization (e), and XGB with winsorization (f).

average R^2 and lowest RMSE (0.3366, 237.35) compared with XGB (0.3262, 238.72), SVR (0.2079, 259.38), and ANN (0.0919, 278.31) models. Therefore, the results indicate that both XGB and RF are effective for predicting the separation factors. Furthermore, log1p transformation outperforms winsorization. To further improve the accuracy of the models, the collected separation factors were first winsorized and then log1p transformed. As shown in Fig. 4e–f and Fig. S5, the performance of RF (R^2: 0.5497, RMSE: 0.9475) and XGB (R^2: 0.5748, RMSE: 0.9188) for the prediction of separation factors was slightly enhanced. Therefore, the RF and XGB models with log-transformation and winsorization were employed for subsequent analysis.

3.3. Model interpretation

Model interpretation is important for understanding the relative feature contributions to the target variable, as well as the underlying mechanisms [29], which is crucial for ensuring that model predictions align with and further enhance mechanistic understandings. In this study, the SHAP method was applied to investigate the relationship between input features and the target variable based on the RF and XGB models. In Fig. 5a, each dot represents a feature in the dataset, and its color indicates the magnitude of the feature value. If the high feature values lead to positive SHAP values, it indicates that the feature exhibits a positive effect on the target variable. On the contrary, if the high feature values lead to negative SHAP values, the feature exhibits a negative effect on the target variable [30]. For example, the high value of the maximum projection radius of dyes displays positive SHAP values, while the large pore radius of the membrane shows negative SHAP values, indicating that large projection radius of the dyes and small pore radius of the membranes are beneficial for the separation of dye and NaCl mixtures (Fig. S6a–b). According to the calculation method for separation factors (Eq. (1)), high rejection of dyes and low rejection of NaCl result in high separation factors. Therefore, consistent with the size exclusion theory and experimental results, large dye molecules and small pore radii of membranes enhance the separation of dye and NaCl mixtures. In addition, membrane water contact angle exhibits a negative effect on the separation factors (water contact angle bellow 45°, Fig. S6c), which aligns with the well-acknowledged phenomena of hydrophilic membranes typically providing superior filtration of dyes [31]. Similarly, consistent with expectations, the effective membrane area (Fig. S6d), surface roughness [6], and zeta potential of the membrane [32] exhibit weakly positive effects on the separation factors. In addition, the effective membrane area represents a process design-related parameter associated with the actual filtration process. Variations in effective membrane area can influence permeate flux distribution, local hydraulic conditions, residence time, concentration polarization, and mass-transfer resistance. These factors may further affect dye accumulation near the membrane surface and thereby dye/NaCl selectivity. Therefore, the significant contribution of effective membrane area suggests that dye/NaCl separation is governed not only by dye molecular structure and intrinsic membrane properties, but also by process-level mass-transfer behavior. To further explore the impact of input features on model prediction, the mean absolute SHAP values and their importance distribution were calculated. As illustrated in Fig. 5b, the maximum projection radius of dyes, pore radius of membranes, water contact angle of membranes, and effective membrane area are the four most important features for the prediction of separation factors, accounting for 13.89%, 11.78%, 11.17%, and 10.86%, respectively. Additionally, the applied pressure, NaCl concentration, log D at pH = 7, and the type of reactor have little impact on separation factors,

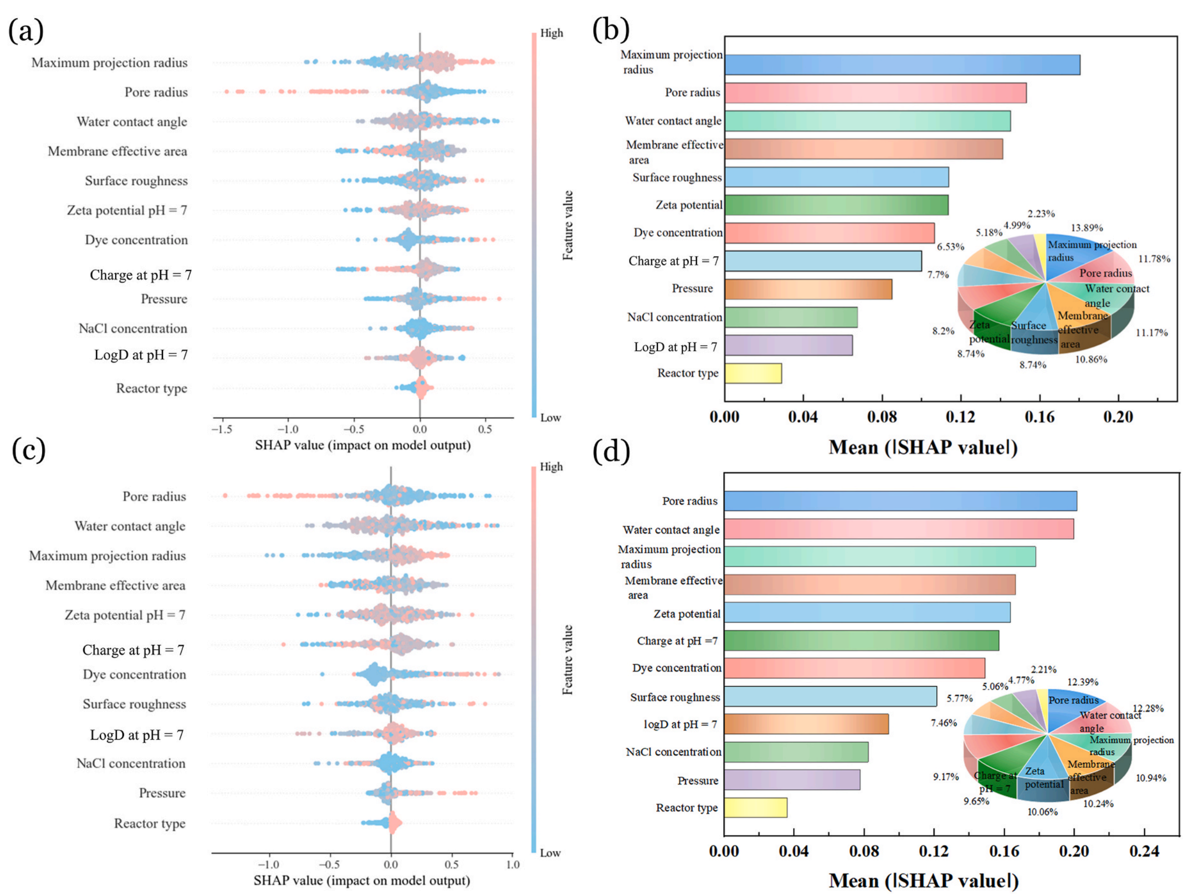

Fig. 5. SHAP values analysis of different variables for separation factors by using the RF (a) and XGB (c) models, the impact of input variables on model prediction, and their importance distribution by using the RF (b) and XGB (d) models.

accounting for 6.53%, 5.18%, 4.99%, and 2.23%, respectively. It should be noted that there is no significant difference between the most important feature of the maximum projection radius of dyes (13.89%) and the least important feature of the type of reactor (2.23%), revealing that all the features contribute varying extents to the separation factors. Furthermore, SHAP analysis was conducted based on the XGB model. As shown in Fig. 5c–d and Fig. S7, the influences of all features on the separation factor are similar to that resulting from the RF model. The four most predominant features for the prediction of separation factors are maximum projection radius of dyes, pore radius of membrane, water contact angle of membrane, and effective membrane area, while logD at pH = 7 of dyes, NaCl concentration, applied pressure, and type of reactor play minor roles. For the SHAP analysis based on the RF and XGB models, the four most important features are identical, and the four least important features are also identical, indicating consistently the physicochemical correlation between the maximum projected radius of the dye, the pore size of the membrane, the water contact angle of the membrane, the effective membrane area, and the separation factor. However, the contribution rates of the four most important features to the model output remain relatively low, ranging from 10% to 14%. This suggests that dye/NaCl separation is not governed by a single dominant variable, but rather by the combined effects of multiple relevant parameters. For instance, the maximum projection radius of dye molecules and the membrane pore radius jointly determine the size-sieving effect. In addition, the water contact angle affects membrane wettability and water transport behavior, whereas the effective membrane area may influence local flux distribution, concentration polarization, and mass-transfer resistance. Therefore, the relatively dispersed feature contributions indicate significant multi-factor coupling effects. This interpretation is further supported by the subsequent bivariate partial dependence analysis, which reveals interactions between key variables and confirms that dye/NaCl separation selectivity arises from the combined, rather than isolated, contributions of these features. Furthermore, the XGB model was applied for further analysis due to the slightly better performance and similar results relative to the RF model.

To further improve the prediction accuracy of the model, deleting some less important features or adding new features was considered. In addition, a versatile hyperparameter tuning method, namely, Tree-structured Parzen Estimator (TPE), was applied. In one case, two less important features, namely, pressure and reactor type, were excluded in the model construction. Fig. S8 shows that the mean R^2 value for the XGB model slightly decreases, and the mean RMSE value slightly increases, consistent with the earlier results that all parameters exert some influence on the model's predictions. In another case, since size exclusion effects and electrostatic interactions are generally considered key mechanisms in membrane separation [33], size effects (the ratio of dye's maximum projection radius and pore radius of membrane) and charge effects (the product of zeta potential of membrane and charge of dye) were included in the model. Fig. S9 shows that the Pearson correlation coefficients for all features are below 0.8, indicating that the introduction of size effects and charge effects do not significantly affect the subsequent model development and interpretation. As illustrated in Fig. S10, the integration of size effects and charge effects into the XGB model do not markedly improve the model's performance, suggesting that too many features may hinder the improvement of the model's predictive capability. Therefore, the original twelve features (i.e., without the deletion and addition of features) were used in subsequent model interpretations.

The specific influence of important features on the separation factor was investigated by PDP and ICE. As illustrated in Fig. 6 and Fig. S11, pore radius and membrane water contact angle display greater impact on their partial dependence compared to other features, suggesting their major roles in the separation factors of dyes and NaCl. Fig. 6a shows that, when the membrane pore radius is below approximately 3.3 nm, the separation factor slightly increases as the membrane pore radius increases. This is because small pore radius facilitates effective rejection

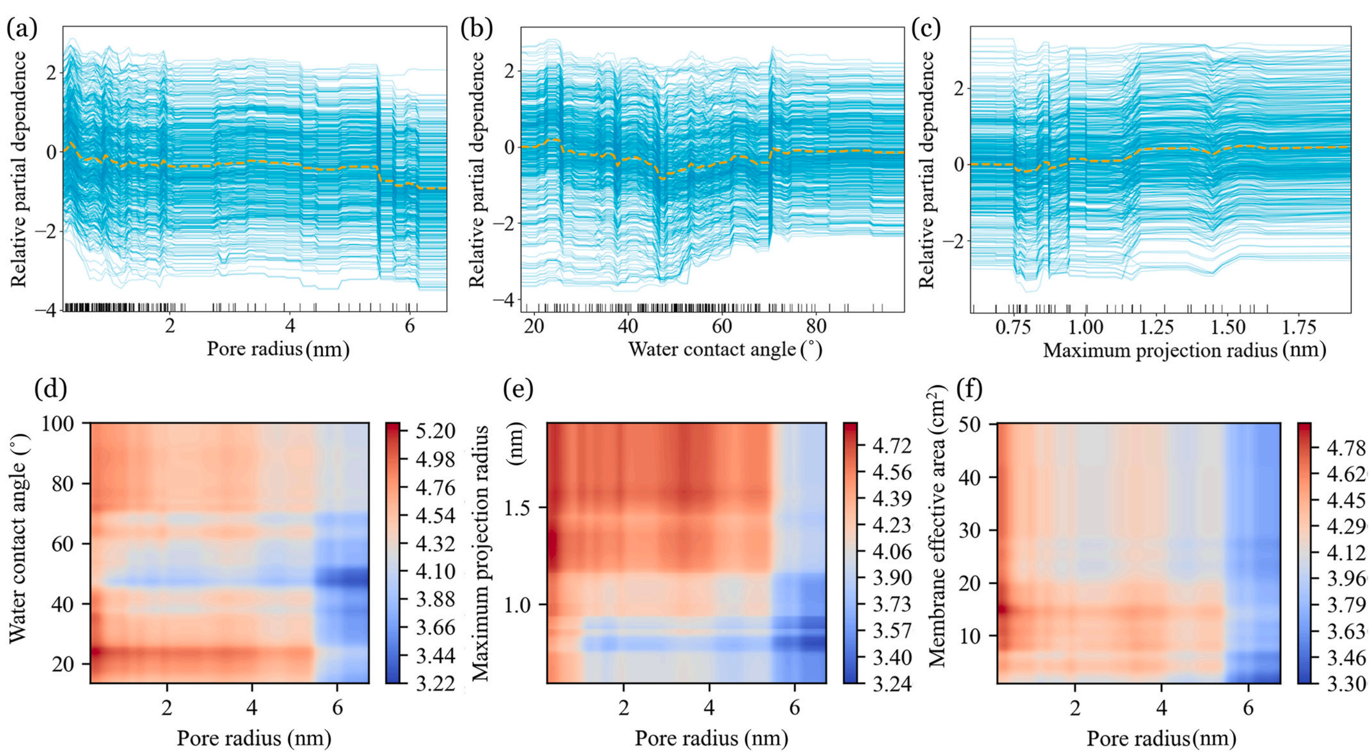

Fig. 6. ICE (cyan solid line) and PDP (orange dotted line) showing the relationships among pore radius (a), water contact angle (b), and maximum projection radius (c) for separation factors. Bivariate partial dependence plots of different variables for the separation factors, pore radius and water contact angle (d), pore radius and maximum projection radius (e), and pore radius and effective membrane area (f). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

of dyes while allowing NaCl to pass through, resulting in high separation factors for dyes/NaCl. When the pore radius exceeds 3.3 nm, the separation factor decreases as the pore radius increases, which is tied to the rejection efficiency for both dyes and NaCl decreasing due to the weak size exclusion effects. When the pore radius exceeds 5.4 nm, the separation factors drop significantly, as the membrane cannot efficiently reject dyes and NaCl, leading to all the dyes and NaCl readily permeating through the membrane. Fig. 6b shows a non-monotonic relationship between the separation factor and the membrane water contact angle. The PDP curve indicates that the model-predicted separation factor decreases as the water contact angle increases to approximately 47°, followed by an increase at higher water contact angles. This suggests that membrane wettability provides important interfacial information for predicting dye/NaCl separation, as the water contact angle reflects the affinity between the membrane surface and the aqueous solution. Its influence may be associated with changes in surface hydration and dye-membrane interactions [34]. In the ICE analysis, the effect of membrane water contact angle is assessed by varying the water contact angle of each sample while keeping the other variables fixed, thereby reflecting the response of the model-predicted separation factor to changes in membrane wettability. Therefore, the observed trend should be interpreted as a data-driven indication of correlation associated with membrane surface properties, rather than direct evidence of a specific physical mechanism. Further experimental studies are needed to verify the underlying mechanisms. Additionally, the maximum projection radius (Fig. 6c), effective membrane area (Fig. S11a), zeta potential (Fig. S11b), and dye concentration (Fig. S11c) exhibit minimal effect on their partial dependence, suggesting that these features have limited impact on the separation factors of dyes and NaCl.

Fig. 6d–f and Fig. S12 display the synergistic effects of two features on separation factors. Fig. 6d shows that the separation factor decreases significantly when the pore radius increases to around 6 nm, while the separation factor firstly decreases and then increases as the water contact angle increases, suggesting that both pore radius and water contact angle of the membrane are important influences of the separation factor. In addition, the highest partial dependence values are achieved when the pore radius is smaller than 4 nm and the water contact angle is lower than 25°, indicating that this combination of pore radius and water contact angle range is associated with higher predicted separation factors in the dataset. However, it should be noted that this result reflects a favorable feature combination identified from the bivariate partial dependence plots, rather than a monotonic dependence of the separation factor on pore size alone. In Fig. 6e, the optimal separation factors are when the pore radius is smaller than 4 nm and maximum projection radius greater than 1.2 nm. This is mainly due to the large size ratio (molecular size of dyes to pore radius of membrane) benefitting the rejection efficiency of dyes, affirming that size exclusion effects are crucial for separating dye and NaCl. Fig. 6f demonstrates that when the pore radius is below 4 nm, the separation factor is optimal when the effective membrane area is smaller than 20 cm^2. Fig. S12 also confirms that higher maximum projection radius, lower water contact angle, and smaller effective membrane areas are favorable to separation factors. This analysis collectively confirms that the pore radius and water contact angle of the membrane, as well as the maximum projection radius of dyes, play key roles in the separation of mixtures of dyes and NaCl.

3.4. Implications and outlook

In view of the complex interplays of various influences and the absence of physical equations, machine learning provides an efficient strategy for predicting the separation factors of mixtures of dyes and NaCl. This can significantly save time, cost, and effort relative to traditional experimental methods. In this study, the influence of commonly reported variables on the separation factors is investigated through SHAP, ICE, and PDP, and the underlying mechanisms are elucidated by investigating the relationships between input features and outcomes. This study provides new insights for the optimization of separation factors to enhance the separation of mixtures of dyes and NaCl. Notably, this study has some limitations listed as follows.

(1) The accuracy of models needs to be improved. Although four well-reported models are applied for the prediction of separation factors, the best model of XGB gives a modest R^2 value of 0.7, which can be attributable to the following aspects. The used dataset contains 527 data rows (total of 4712 data points); since data-driven black-box models are inherently data-costly [35], more data may improve the R^2 values. In addition, because the data were obtained from experiments conducted under different conditions (e.g., different laboratory settings, different membrane module designs, different equipment, different hydrodynamics, etc.), data consistency may expectedly be inadequate. Furthermore, several specific sources of error affecting model accuracy should be considered. First, the structural diversity of dye molecules, including differences in molecular size, charge, functional groups, hydrophobicity, and aggregation behavior, can influence membrane rejection and dye-membrane interactions. Second, variations among membrane types, such as pore size distribution, surface charge, hydrophilicity, and chemical composition, can result in different separation behaviors and thereby prediction errors. Third, dye/NaCl separation is a complex process governed by multiple mechanisms, including size exclusion, electrostatic repulsion, Donan exclusion, and adsorption, making it difficult for models to fully capture all separation behaviors. Therefore, future work should not only aim to expand the dataset to improve the R^2 value, but also seek to reduce these sources of error by incorporating more detailed dye and membrane descriptors, as well as mechanism-related parameters.
(2) Some factors are not consistently reported in literature, and thus are not collated in the dataset and thereby not employed in the development of the models. Some examples include solution pH, temperature, and presence of ions and pollutants. Therefore, the models may be limited in predicting separation factors in real dye wastewater. In addition, the key operational feasibility considerations, like lifecycle analysis and technoeconomic analysis, are not factored into the models, which further restricts the direct applicability of the models in practice.

In future, more parameters (e.g., solution pH, temperature, stability, CapEx (Capital Expenditures) and OpEx (Operating Expenses)) should be incorporated into the machine learning models to enhance the quality of the dataset. In addition, customizing suitable models based on the dataset, adjusting input parameters, optimizing feature engineering, and enriching output variables can improve the performance of such models and further enhance the practical application of the results. Furthermore, although machine-learning interpretations provide useful insights into the potential roles of dye characteristics, membrane properties, and operating conditions in dye/NaCl separation, these relationships should be regarded as data-driven mechanistic implications rather than direct evidence of causal mechanisms. In particular, SHAP, PDP, and ICE analyses can identify important variables and trends, thus offering data-driven guidance for understanding key factors and designing future experimental validation studies. To further exploit the benefits of membrane-based dye/NaCl separation, future research should integrate controlled separation experiments, membrane surface characterization, and pore structure analysis to validate the proposed mechanisms and further improve the physical interpretability of the model.

4. Conclusion

In this study, machine learning was employed to elucidate the important influences and underlying mechanisms of membrane selectivity for dye/NaCl mixtures. The commonly reported parameters (including membrane pore radius and water contact angle , dye and NaCl concentrations, effective membrane area, and separation factors for dye/NaCl mixtures) were extracted from 157 publications, and the MICE method was applied for handling missing values in the dataset. Subsequently, four models (namely, RF, SVR, ANN, XGB) were employed to predict the separation factors, with the XGB model demonstrating the best performance with an R^2 value of 0.7. SHAP analysis confirmed that the maximum projection radius of dyes, membrane pore radius, membrane water contact angle , and effective membrane area were the four most influential features in the prediction of separation factors. The PDP and ICE results indicated that size exclusion was the primary mechanism governing the selective separation of dye/NaCl mixtures, while hydrophilicity-related effects and charge-associated electrostatic interactions further regulated dye retention and NaCl transport. Finally, the limitations of the study and recommendations for future research were identified. Overall, this research provides a practical framework for understanding and predicting the separation factors of dye/NaCl mixtures, thereby advancing the broader application of membrane-based separation in textile wastewater remediation.

CRediT authorship contribution statement

Yongtao Xue: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Jia Wei Chew:** Funding acquisition, Methodology, Project administration, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful for funding from the Swedish Research Council (VR; 2025-07412), Chalmers Gender Initiative for Excellence (Genie), and Chalmers Academic Excellence Award.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.memsci.2026.125908.

Data availability

Data will be made available on request.

References

[1] T. Li, R. Liu, S. Liu, J. Xu, J. Yu, L. Guo, F. Li, Highly permeable hybrid membranes constructed from MXene and TA-APTES nanoaggregates for effective dye/salt fractionation, J. Membr. Sci. 734 (2025) 124470, https://doi.org/10.1016/j.memsci.2025.124470.
[2] H. Zhang, X. Lai, J. You, W. Wang, M. Wu, L. Liu, H. Chen, Z. Luo, H. Zhu, Y. Wang, Y. Lai, W. Cai, Facile construction of novel poly(ether sulfone ether ketone ketone) loose nanofiltration membrane for efficient dye/salt separation, Desalination 601 (2025) 118573, https://doi.org/10.1016/j.desal.2025.118573.
[3] H. Zhou, X. Li, Y. Li, R. Dai, Z. Wang, Tuning of nanofiltration membrane by multifunctionalized nanovesicles to enable an ultrahigh dye/salt separation at high salinity, J. Membr. Sci. 644 (2022) 120094, https://doi.org/10.1016/j.memsci.2021.120094.
[4] W. Xie, K. Zhao, L. Xu, N. Gao, H. Zhao, Z. Gong, L. Yu, J. Jiang, Oxalic acid cross-linked sodium alginate and carboxymethyl chitosan hydrogel membrane for separation of dye/NaCl at high NaCl concentration, Chin. Chem. Lett. 33 (2022) 1951–1955, https://doi.org/10.1016/j.cclet.2021.11.058.
[5] J. Li, X. Bai, X. Ma, Z. Wang, Fabrication of high permselectivity nanofiltration membranes based on 3,5-dihydroxybenzoic acid for separation of dye/salts, Desalination 607 (2025) 118796, https://doi.org/10.1016/j.desal.2025.118796.
[6] Y. Xue, J. Shao, H.X. Le, J.W. Chew, Facile fabrication of hierarchical PVDF/Mg–Al LDH/Chitosan membrane for textile wastewater remediation: role of LDH in anchoring chitosan, J. Membr. Sci. 739 (2026) 124914, https://doi.org/10.1016/j.memsci.2025.124914.
[7] H. Li, H. Lin, S. Raza, Z. Zhao, S. Chen, Y. Wang, Q. Zeng, C. Chen, W. Yu, L. Shen, Molecular insights in outstanding performance of Ca2+ bridging MXene/Sodium alginate composite membranes, Water Res. 286 (2025) 124296, https://doi.org/10.1016/j.watres.2025.124296.
[8] P. Liu, Q. Xing, R. Huang, P. Bai, J. Lyu, Incorporating hydrophilic amino-modified UiO-66 nanoparticles into polyphenylsulfone membrane for improved dye/salt separation, Chem. Eng. J. 499 (2024) 156096, https://doi.org/10.1016/j.cej.2024.156096.
[9] U.S. Joshi, S. Samanta, S.K. Jewrajka, Low fouling polyelectrolyte layer-by-layer self-assembled membrane for high performance dye/salt fractionation: sequence effect of self-assembly, ACS Appl. Mater. Interfaces 16 (2024) 32748–32761, https://doi.org/10.1021/acsami.4c06169.
[10] H. Chen, F. Zhao, Z. Li, J. Ge, C. Wang, L. Yang, A novel loose nanofiltration membrane constructed by 1, 2-bis (N-aminoethyl imidazoline) ethane as an aqueous monomer for precise dye/salt separation, Desalination 587 (2024) 117955, https://doi.org/10.1016/j.desal.2024.117955.
[11] X. Song, J. Zhao, M. Jia, R. Yin, H. Ren, Elucidating the role of reactive nitrogen species in micropollutant degradation using machine learning, Water Res. 292 (2026) 125345, https://doi.org/10.1016/j.watres.2026.125345.
[12] M. Liu, Q. Tang, E. Leng, Y. Deng, D. Lv, J. E, Multi-staged machine learning-driven investigation of biomass-derived hard carbon anodes for sodium-ion batteries: connecting precursor, structure, and performance, Chem. Eng. J. 528 (2026) 172350, https://doi.org/10.1016/j.cej.2025.172350.
[13] A. Hu, Y. Liu, X. Wang, S. Xia, B. Van der Bruggen, A machine learning based framework to tailor properties of nanofiltration and reverse osmosis membranes for targeted removal of organic micropollutants, Water Res. 268 (2025) 122677, https://doi.org/10.1016/j.watres.2024.122677.
[14] Z. Ji, H. Guan, Y. Dong, Performance and mechanism of lithium extraction from water via machine learning-powered nanofiltration, J. Membr. Sci. 733 (2025) 124344, https://doi.org/10.1016/j.memsci.2025.124344.
[15] P. Liu, H. Xu, P. Jin, X. Zhu, J. Zheng, Y. Liu, J. Yang, D. Xu, H. Liang, DFT-assisted machine learning for polyester membrane design in textile wastewater recovery applications, Water Res. 279 (2025) 123438, https://doi.org/10.1016/j.watres.2025.123438.
[16] Y. Zhang, B. Han, Q. Song, S. Jin, Z. Lu, S. Lu, X. Zhang, Y. Zhang, D. Li, Revealing key structural and operating parameters on salt/dye separation of loose nanofiltration membrane by ensemble machine learning, J. Membr. Sci. 732 (2025) 124274, https://doi.org/10.1016/j.memsci.2025.124274.
[17] Z. Chen, F. Gong, J. Song, K. Zhang, Interpretable prediction of biomass-derived biochar characteristics: combining machine learning with shapley additive explanation, Energy 335 (2025) 137877, https://doi.org/10.1016/j.energy.2025.137877.
[18] X. He, H. Wang, Accurate prediction of trace elements in aluminum alloys using random forest regression and femtosecond laser-ablation spark-induced breakdown spectroscopy, Microchem. J. 216 (2025) 114813, https://doi.org/10.1016/j.microc.2025.114813.
[19] D. Xu, J. Yin, An enhanced hybrid scheme for ship roll prediction using support vector regression and TVF-EMD, Ocean. Eng. 307 (2024) 117951, https://doi.org/10.1016/j.oceaneng.2024.117951.
[20] R. Jradi, C. Marvillet, M.R. Jeday, Multi-objective optimization and performance assessment of response surface methodology (RSM), artificial neural network (ANN) and adaptive neuro-fuzzy interfence system (ANFIS) for estimation of fouling in phosphoric acid/steam heat exchanger, Appl. Therm. Eng. 248 (2024) 123255, https://doi.org/10.1016/j.applthermaleng.2024.123255.
[21] C.H. Brillianto Apribowo, S.P. Hadi, F.D. Wijaya, M.I. Bambang Setyonegoro, Sarjiya, early prediction of battery degradation in grid-scale battery energy storage system using extreme gradient boosting algorithm, Results Eng. 21 (2024) 101709, https://doi.org/10.1016/j.rineng.2023.101709.
[22] P. Pascacio, D.J. Vicente, F. Salazar, S. Guerra-Rodríguez, J. Rodríguez-Chueca, Predictive modeling of Enterococcus sp. removal with limited data from different advanced oxidation processes: a machine learning approach, J. Environ. Chem. Eng. 12 (2024) 112530, https://doi.org/10.1016/j.jece.2024.112530.
[23] S.M.K. Alam, P. Li, M. Rahman, M. Fida, V. Elumalai, Key factors affecting groundwater nitrate levels in the Yinchuan Region, Northwest China: research using the eXtreme Gradient Boosting (XGBoost) model with the SHapley Additive exPlanations (SHAP) method, Environ. Pollut. 364 (2025) 125336, https://doi.org/10.1016/j.envpol.2024.125336.
[24] L.S. Ho, V.Q. Tran, Machine learning approach for predicting and evaluating California bearing ratio of stabilized soil containing industrial waste, J. Clean. Prod. 370 (2022) 133587, https://doi.org/10.1016/j.jclepro.2022.133587.
[25] G. Xie, C. Zhu, C. Li, Z. Fan, B. Wang, Predicting the adsorption of ammonia nitrogen by biochar in water bodies using machine learning strategies: model optimization and analysis of key characteristic variables, Environ. Res. 267 (2025) 120618, https://doi.org/10.1016/j.envres.2024.120618.
[26] MdM. Hossain, R. Sikder, G. Hua, T. Ye, From model development to mitigation: machine learning for predicting and minimizing iodinated trihalomethanes in water treatment, Environ. Sci. Technol. 59 (2025) 11638–11652, https://doi.org/10.1021/acs.est.5c05409.

[27] K.P. Fourkiotis, A. Tsadiras, Comparing machine learning techniques for house price prediction, in: I. Maglogiannis, L. Iliadis, J. MacIntyre, M. Dominguez (Eds.), Artificial Intelligence Applications and Innovations, Springer Nature Switzerland, Cham, 2023, pp. 292–303.
[28] J. Chung, K. Neo, M. Khan, E. Cruz, Z. Zheng, T. Munasinghe, J.C. Wei, Exploring flooded fraction prediction through machine learning models focusing on medical infrastructure in the Southeast U.S. coastal areas, in: 2025 IEEE International Conference on Big Data and Smart Computing (BigComp), 2025, pp. 49–56, https://doi.org/10.1109/BigComp64353.2025.00017.
[29] C.L. Loi, C.-C. Wu, Y.-C. Liang, Prediction of tropical cyclogenesis based on machine learning methods and its SHAP interpretation, J. Adv. Model. Earth Syst. 16 (2024), https://doi.org/10.1029/2023MS003637.
[30] J. Liu, Y. Wu, A. Zhu, Z. Ji, Data-driven modeling and explainable optimization of organic micropollutant removal in forward osmosis membranes, J. Environ. Chem. Eng. 14 (2026) 120805, https://doi.org/10.1016/j.jece.2025.120805.
[31] H. Nawaz, M. Umar, I. Nawaz, A. Ullah, M. Tauseef Khawar, M. Nikiel, H. Razzaq, M. Siddiq, X. Liu, Hybrid PVDF/PANI membrane for removal of dyes from textile wastewater, Adv. Eng. Mater. 24 (2022) 2100719, https://doi.org/10.1002/adem.202100719.
[32] N. Basel, Q. Liu, L. Fan, Q. Wang, N. Xu, Y. Wan, Q. Dong, Z. Huang, T. Guo, Surface charge enhanced synthesis of TpEB-based covalent organic framework (COF) membrane for dye separation with three typical charge properties, Sep. Purif. Technol. 303 (2022) 122243, https://doi.org/10.1016/j.seppur.2022.122243.
[33] J. Guo, Y. Zhang, F. Yang, B.B. Mamba, J. Ma, L. Shao, S. Liu, Ultra-Permeable dual-mechanism-driven graphene oxide framework membranes for precision ion separations, Angew. Chem. Int. Ed. 62 (2023) e202302931, https://doi.org/10.1002/anie.202302931.
[34] Y. Liu, J. Du, H. Wu, C. Cong, H. Zhang, J. Wang, Z. Wang, Antifouling streptomycin-based nanofiltration membrane with high permselectivity for dye/salt separation, Sep. Purif. Technol. 297 (2022) 121443, https://doi.org/10.1016/j.seppur.2022.121443.
[35] J.W. Chew, T. Bierweiler, P. Ryttestal, R. Andersson, T. Wik, AI in chemical engineering: from promise to practice, AIChE J. (2026) e70358, https://doi.org/10.1002/aic.70358. (Accessed 17 May 2026).