Contents lists available at ScienceDirect

Desalination

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

Predicting the boron removal of reverse osmosis membranes using machine learning

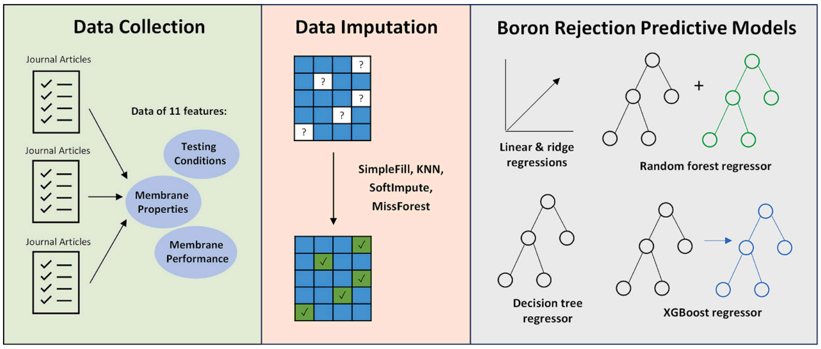
Sukarno [a,b], Jeng Yi Chong [c,d,*], Gao Cong [a]

[a] Division of Data Science, College of Computing and Data Science, Nanyang Technological University, Singapore
[b] Earth Observatory of Singapore, Nanyang Technological University, Singapore
[c] Singapore Membrane Technology Centre, Nanyang Environment and Water Research Institute, Nanyang Technological University, 1 Cleantech Loop, 637141, Singapore
[d] School of Chemistry, University of Southampton, Southampton SO17 1BJ, UK

HIGHLIGHTS

- Predictive models for boron rejection of RO membrane were developed using ML.
- Tree-based models such as XGBoost regressor showed outstanding performance.
- For SWRO membranes, NaCl rejection >99.6 % will yield high boron rejection.
- BWRO membranes with a looser structure will still perform well at pH >9.
- Membrane surface properties showed minimal effects on boron rejection.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Boron rejection
Machine learning
Artificial intelligence
Reverse osmosis
Polyamide TFC membranes

ABSTRACT

Reverse osmosis (RO) is a key technology for seawater desalination, but boron removal remains challenging due to the relatively low and varying boron rejection of RO membranes. This study explored the use of machine learning (ML) to develop predictive models for boron removal of RO membranes. Data of 11 features encompassing membrane properties, testing conditions and membrane performance were collected from journal articles. Missing data were recovered using data imputation algorithms. The predictive models were developed using five regression algorithms: linear, ridge, decision tree, random forest and XGBoost regressors, and the tree-based XGBoost regressor performed the best ($R^2 = 0.84$). Feature importance analysis and tree diagrams revealed that membrane type, feed pH and NaCl rejection as key factors in influencing boron rejection, while membrane surface properties showed minimal impact. Partial dependence plots were generated to further analyze each feature. High NaCl rejection of >99.6 % is highly desirable for SWRO membranes to achieve high boron rejection. For BWRO membranes at pH >9, a looser structure with a NaCl rejection >95 % could be applied. The study successfully applied ML to a dataset with large portion of missing values, and the results provide valuable insights for future membrane design and boron removal processes.

[*] Corresponding author at: School of Chemistry, University of Southampton, Southampton SO17 1BJ, UK.
E-mail address: j.y.chong@soton.ac.uk (J.Y. Chong).

https://doi.org/10.1016/j.desal.2024.117854
Received 5 April 2024; Received in revised form 6 June 2024; Accepted 16 June 2024
Available online 22 June 2024

1. Introduction

Water scarcity is a pressing global challenge exacerbated by population growth and industrialization. Seawater desalination has emerged as a critical source of clean water to meet the increasing water demand [1,2]. Membrane separation has gained prominence as a transformative technology for seawater desalination. In the reverse osmosis (RO) process, semipermeable membranes can effectively retain salts and impurities in seawater, producing high-purity clean water. Operating at ambient temperature, this pressure-driven process offers a more energy-efficient alternative to thermo-driven desalination processes like multistage flash distillation [3,4]. Polyamide thin-film composite (TFC) membranes have been widely used in the seawater RO (SWRO) desalination process as they have a high salt rejection and excellent water permeability. In recent years, large-scale SWRO desalination plants have been constructed in countries with serious water scarcity, for example the Taweelah SWRO plant (909,200 m^3/day) in the United Arab Emirates and the Sorek (624,000 m^3/day) SWRO plants in Israel [5].

One of the key challenges of SWRO desalination is the boron content in the water produced [6,7]. Seawater contains boron at a concentration of about 5 mg/L. Though boron is an essential micronutrient for most life, the amount needed is generally very small. At concentration >1 mg/L, boron can be harmful for most plants, and high boron level can be toxic to human [8]. The World Health Organisation requires the boron concentration to be <0.5 mg/L in drinking water [9]. To meet the boron concentration requirement, it is essential for the SWRO process to have a high boron removal rate, preferably >90 %. However, boron is a small element and form a non-ionized boric acid (H_3BO_3) in seawater. With an ionic radius of 0.244–0.261 nm, boric acid can diffuse through the RO membranes in a non-ionic manner, resulting in a low boron rejection [10]. Many studies have been conducted to improve the boron rejection of RO membranes. One of the main approaches is through sealing the "defects" or plugging the network pores, so that boron can be excluded from passing through the membrane [11,12]. Some studies also modified the surface change of the membranes so that the stronger electrostatic charge can enhance the boron rejection [13,14].

While most SWRO membranes have a consistently high NaCl rejection >99 %, their boron rejection can vary significantly from 70 to 99 % [5,15]. To achieve a low boron concentration in the water produced, many SWRO plants have adopted a two-pass RO process where portion of the permeate from the first pass is fed to a second RO membrane unit [16,17]. Brackish water RO (BWRO) membranes with a higher water permeability are typically used in the second pass, and a lower pressure will be applied due to the low salt concentration. Though the second pass RO unit can further remove the boron content to meet the requirement, the extra step will require additional energy and cost. Alternatively, the boron rejection in the RO process can be improved by increasing the pH of the feed. As pH increases, boron will be transformed into negatively charged borate ions ($B(OH)_4^-$), could be rejected more effectively by the RO membranes through the Donnan's effect [16,18]. This procedure however requires large amount of chemical dosing to change the pH of the feed, and is more commonly used in the second pass RO. Therefore, SWRO membranes with a high boron rejection is highly desired so that boron removal can be done efficiently in the single-pass RO process. Despite ongoing efforts, the progress in developing membranes with desirable boron rejection performance was still slow.

The use of machine leaning (ML) techniques can be beneficial in accelerating membrane development and facilitating the study of membrane processes. ML has gained popularity in many areas and its applications in chemical and environmental engineering, and material science have becoming prevalent in recent years [19–21]. Most studies on membrane development involved laborious experimental work and optimizing the membrane performance can be challenging with various parameters. One major advantage of ML is its capacity to handle and analyze multi-dimensional data and this will help researchers to search for useful materials and synthesis techniques more efficiently. Some recent studies have demonstrated the potential of ML techniques in membrane research. Yeo et al. employed ML technique gradient boosting tree model to predict the salt pass rate of thin-film nanocomposite (TFN) RO membranes [22]. The contributions of different parameters such as the loading, pore size and shape of nanoparticles on the membrane performance could be analysed effectively using data collected from the literature. In another study by Hu et al., ML techniques were used to predict the performance of organic solvent nanofiltration (OSN) membranes [23]. ML models such as artificial neural network, support vector machine and random forest were used as the alternative to traditional mathematical equations. Meanwhile, Zhu et al. [24] predicted the organic contaminants rejection of nanofiltration (NF) and RO membranes using traditional algorithm (e.g., multiple linear regression) and ensemble models (e.g., random forest and gradient boosting) using data collected from the journal articles. Their study showed that the ensemble models outperformed traditional models in predicting the removal efficiency [24].

Data is an essential part of ML and experimental data published in the literature is a valuable source of data. In recent years, databases of membrane research have been established, for example: the Open Membrane Database (OMD) for reverse osmosis and nanofiltration membranes, and the Membrane Database for polymer gas separation membranes [25,26]. Though the OMD has an extensive database on RO membranes, the focus is mainly on the water and salt permeabilities, and boron rejection data was not collected. A recent study by Ajali-Hernández et al. applied ML models to study boron permeability in SWRO, but the data used originated from a single desalination plant [27]. Therefore, one key challenge of utilizing ML techniques to study the boron rejection performance of RO membranes is collecting data from the literature. Another common challenge encountered when applying ML in membrane research is the missing data. Though most studies reported key membrane performances such as water permeability and salt rejection, the comprehensiveness of other parameters such as detailed testing conditions and membrane properties can vary. Missing data will affect the modeling of the data as typical ML models cannot use data entries containing missing values. Therefore, data imputation is an important step to fill up the missing data so that all the collected data can be utilized for ML modeling, which is particularly important for problems without a large amount of data. However, many previous studies using ML for membrane research only employed basic strategies such as the median of the variables to fill up missing data [22]. Yuan et al. have previously used data imputation techniques to predict missing gas separation performance of membranes [28]. While Gao et al. have applied XGBoost and CatBoost to recover missing data in their study on NF membranes [29]. Advanced data imputation techniques can be further explored to study datasets in membrane research.

This paper focused on the study of boron rejection performance in RO process using ML techniques. We first collected the experimental data from journal articles, with the focus on polyamide-based SWRO, BWRO and NF membranes. Data from three main categories were collected: membrane properties, testing conditions and membrane performance. Descriptive analytics were performed on the compiled dataset to study the trend and distribution of the parameters. Afterwards, missing entries in the dataset were recovered using data imputation algorithms. Several data imputation methods: SimpleFill, k-nearest neighbours, SoftImpute and MissForest were applied. Finally, regression algorithms were performed to the predict boron rejection based on the parameters. Five regression models: linear, ridge, decision tree, random forest and XGBoost regressions were applied in this study. The rationale for using these methods is that they mostly have great explainability, which is important for understanding the factors influencing boron rejection. To further improve the accuracy of the regression models, data classification based on the membrane type was also conducted. The feature importance from the regression models and the tree diagram were analysed to understand parameters that will affect the boron rejection performance of the membranes significantly. Furthermore, the trained regression algorithms were used to generate boron rejection performance data under various conditions, enabling partial dependency analysis on the features. The interaction between parameters and optimum conditions for excellent boron rejection performance could be determined. The insight obtained from this study can serve as a valuable reference for researchers during membrane development and process design, and such information could be otherwise challenging to discover using non-ML methods.

2. Method

2.1. Data collection

In this work, experimental data were collected from journal articles published from year 2005 to 2022 (full list in Table S1 Supplementary material). We focused on collecting data on polyamide TFC membranes as they are the most commonly used and studied membranes for seawater desalination and boron removal. Data points from graphs and charts were extracted using Plot Digitizer (http://plotdigitizer.sourceforge.net/). Boron rejection was set as our target of prediction, and we collected data of 11 other parameters as features of the predictive models.

2.2. Data imputation

The collected dataset contained some missing data because not all the parameters were reported in each journal article. Data imputation was required to recover the missing entries before the dataset could be further analysed and modelled. The entire dataset was used for data imputation without splitting into training or testing dataset. Several imputation algorithms were studied, and their brief descriptions are as followed. Details on how the algorithm work, source codes and installation guides can be found in the links provided in the Supplementary material.

SimpleFill The missing entries in each parameter are filled with the mean of respective parameter.

K-nearest neighbours (KNN) The algorithm finds the k nearest neighbours with the highest similarity score and subsequently uses the weighted sum of these k rows to recover missing entries [30].

SoftImpute Data imputation based on matrix completion method, according to the iterative soft thresholding of singular value decomposition (SVD) [31].

MissForest MissForest selects a parameter with the lowest missing entries and recover the missing entries in this parameter using random forest algorithm. Non-selected parameters are filled with mean of the respective parameter. The same procedure is repeated to next parameter with the lowest missing entries [32].

Root mean square error (RMSE) was used to evaluate the accuracy of data imputation, which is defined as follow:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left|y_i^{real} - y_i^{pred}\right|^2}{n}} \tag{1}$$

where y_i^{real}: collected data, y_i^{pred}: predicted data by model, n: number of sample points.

Before imputation algorithm is run, 10 available entries are hidden. These 10 hidden entries are then assessed against entries recovered from imputation algorithms. The errors are subsequently quantified using RMSE. This step is repeated until all available entries in each parameter are assessed. Since there will be multiple RMSE in each parameter, an average of RMSE is obtained as the representative of each parameter.

2.3. Boron rejection regression models

Regression algorithms were used to develop predictive models to predict boron rejection performance of membranes using the imputed dataset. Five regression algorithms were studied and their brief description are as below. More detailed information on the algorithms and the source codes can be found in the references indicated in the Supplementary material.

Linear regression Linear regression is used as our baseline model since this is one of the widely used algorithms before other ML techniques start to gain popularity.

Ridge regression Ridge regression is the extension of linear regression, with an additional feature to penalise coefficients of highly correlated parameters [33].

Decision tree regressor In decision tree regressor, the dataset is broken into smaller and smaller subsets based on the best split [7,34]. The quality of the split is determined using squared error (SE) as defined in Eq. (2). In our experiment, the maximum depth of the full tree is set to 9 to prevent overfitting. SE is defined as follow:

$$SE = \sum_{i=1}^{n}\left(y_i^{real} - y_i^{pred}\right)^2 \tag{2}$$

Random forest regressor Similar to decision tree regressor, random forest regressor however builds a group of trees using different subset of data [35]. Predicted values obtained from different trees are being aggregated as the final predicted value. In comparison to decision tree where only one tree is used, using a group of trees will reduce over-fitting. When growing the trees in random forest regressor, SE is used to determine the quality of split.

XGBoost regressor Similar to random forest regressor, XGBoost regressor builds a group of progressive trees [36]. After the first tree is built, the subsequent tree is built based on the previous tree, which is to reduce the errors committed by previous tree.

2.4. Interpretation of boron rejection regression models

Besides analysing the feature importance of the predictive models, two other methods were used to further analyze and interpret the models. In the first method, a tree diagram of the decision tree regressor was generated using export_graphviz. The tree branches were analysed to gain more understand on the effects of the features. The second method was to construct the univariate and bivariate partial dependence plots (PDP) using the predictive models developed. The PDP plots enabled the study of the effects of variables under a wider range of conditions. The univariate PDP was generated by varying the condition of a single parameter, while keeping the remaining parameters constant. Boron rejections of each single data point was predicted and an average was calculated and shown in the graphs. For the bivariate PDP, two parameters were varied at the same time while keeping remaining parameters constant.

3. Results and discussion

3.1. Descriptive analytics of data collected from literature

Fig. 1a illustrated the number of publications on the study of boron rejection of RO membranes from 2005 to 2023, with a total of 77 papers. There has been a consistent interest in this area over the years, with studies mainly focusing on the development of RO membranes with an enhanced boron rejection performance and the understanding of boron rejection in RO process under different conditions. From the journal articles, we successfully aggregated 534 rows of data and Fig. 1b shows the boron rejection distribution of the membranes. Boron rejection >70 % were typically observed in SWRO membranes, while it spread between 45 % to 99 % for BWRO membranes. SWRO membranes with a tighter pore network structure overall had a higher boron rejection compared to BWRO membranes. We also included NF membranes in this study as similar polyamide TFC membranes were investigated for boron removal application in some studies. As expected, NF membranes with a looser pore structure had the lowest boron rejection, typically below 65 %. Of all these data points, there was similar number of data points for SWRO and BWRO, each accounted for about 44 %, while only 12 % for NF, as shown in Fig. 1c.

We collected data of 11 parameters as the features for the ML modeling, as shown in Fig. 1d. These parameters are the operating or testing conditions (temperature, pressure, pH, cross-flow velocity, boron concentration and NaCl concentration), membrane properties (contact angle, surface charge and surface roughness), and membrane performance (water permeability and NaCl rejection). These variables are commonly reported in the literature and many of them could affect the boron rejection performance of the membranes. Most studies included the operating conditions in the publication, therefore these variables had fewer missing data in our data collection. However, membrane properties were not reported in many studies especially those focused on membrane processes, leading to a high percentage of missing data (>50 %) in these variables. In our data collection, we included TDS rejection and salt rejection under the NaCl rejection variable for studies where mixed salt solution was tested. TDS and salt rejections were normally reported when real feeds like seawater were tested, and NaCl was still the main salt component in the solution [37,38]. Including the TDS and salt rejection data will reduce the missing data in the NaCl rejection feature, which can subsequently improve the data quality and ML modeling, as discussed later in Section 3.3.

Fig. 2a to f show the data distribution of the features and the membrane type was also indicated in the distributions. The boron rejection performance have been studied at a wide range of pressure from 2 to 62 bar. For the SWRO membranes, a higher pressure was normally applied, with most data points >40 bar. Feeds with a high NaCl concentration such as seawater were normally tested with SWRO membranes, and a higher pressure was required to overcome the osmotic pressure. On the other hand, the pressure applied was lower for BWRO membranes

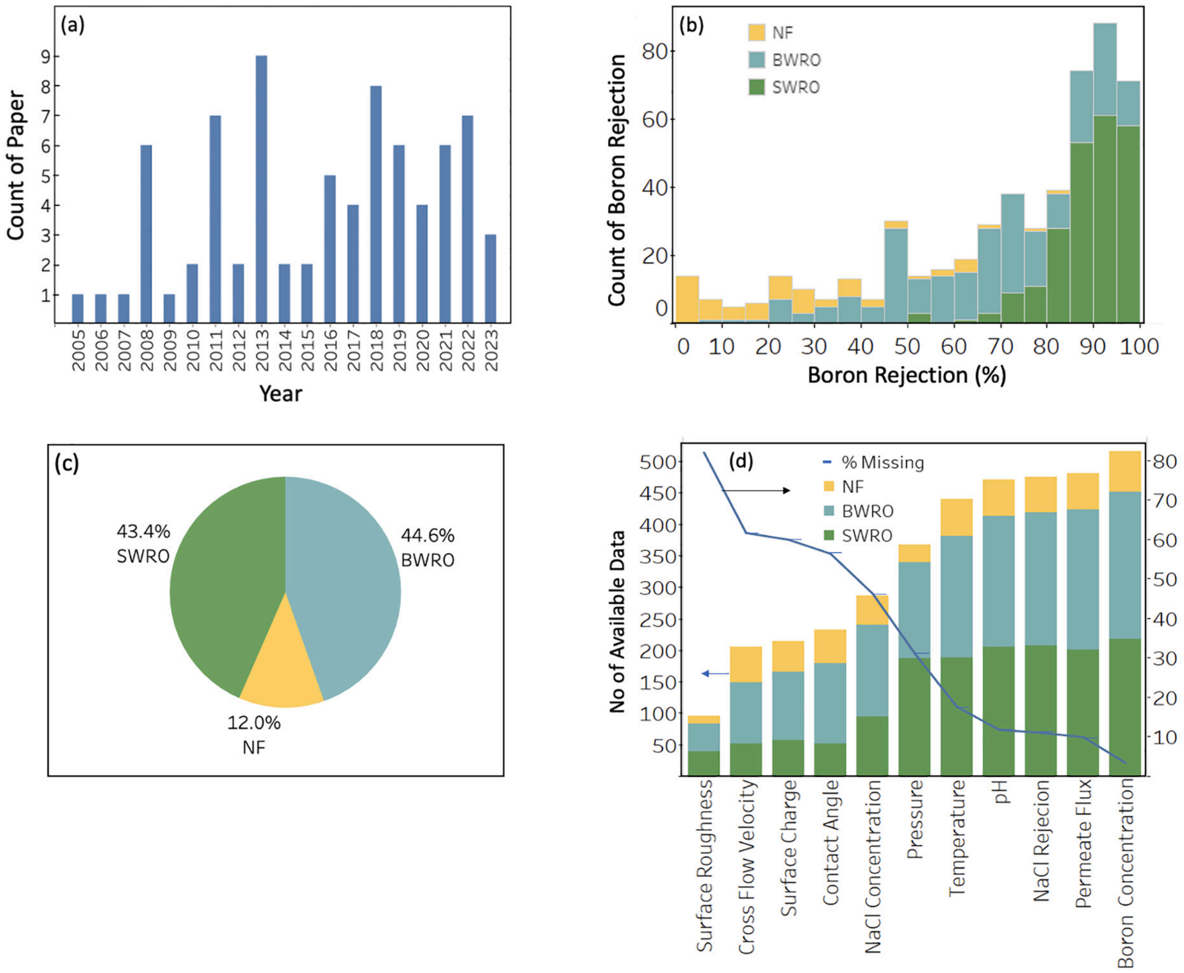


Fig. 1. (a) Number of publications by year, (b) distribution of boron rejection by membrane type, (c) distribution of membrane types, (d) dataset parameters and missing data percentage.

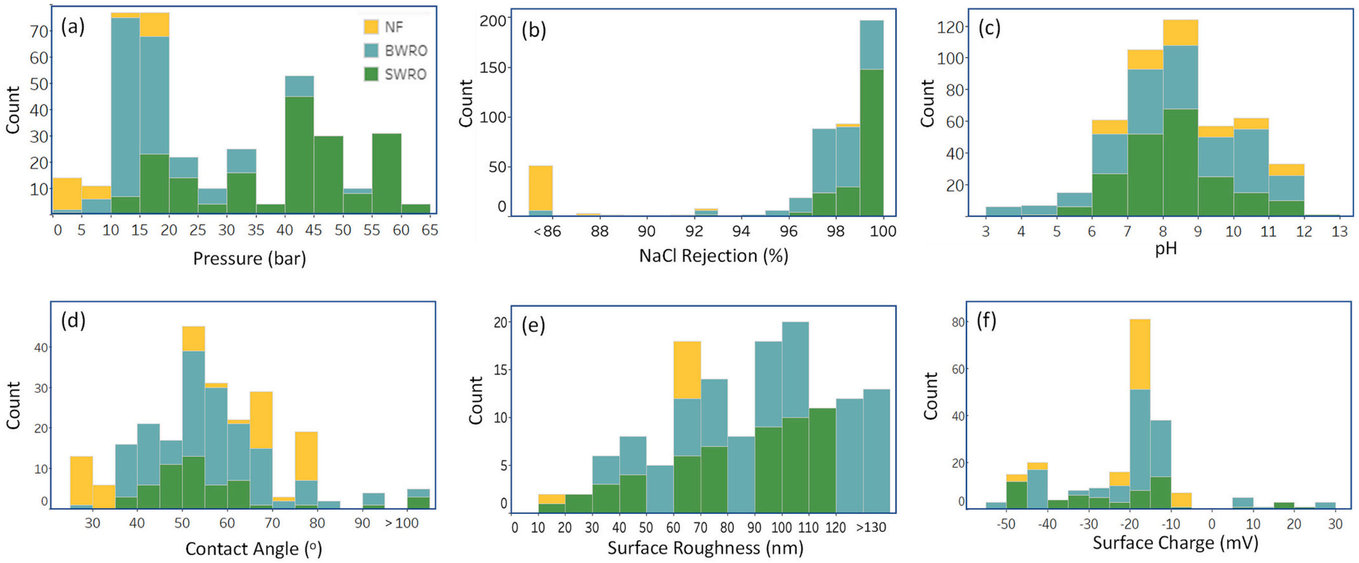


Fig. 2. Distribution of data collected: (a) pressure, (b) NaCl rejection, (c) pH, (d) water contact angle, (e) surface roughness, and (f) surface charge.

(typically between 5 and 35 bar), and NF membranes (below 20 bar), as feeds with a lower salt concentration were normally tested. Besides that, SWRO membranes had the highest NaCl rejection >99 %. The NaCl rejection was lower for BWRO membranes (97–99 %), and <86 % for NF membranes. The common pH used to study the boron rejection performance was between 6 and 10. However, it was observed that BWRO membranes were also frequently tested at pH >10 to obtain better boron rejection performance [39].

For the membrane properties, the water contact angle of the membranes mostly fell between 25 and 80° as the polyamide TFC membranes are generally hydrophilic. While for the membrane surface roughness, the data ranged between 10 and 130 nm, but with majority of them above 60 nm. The rough membrane surface was contributed by the common ridge-and-valley structure of polyamide thin films [40]. For the membrane surface charge, the polyamide TFC membranes generally had a negatively charged surface with most of them with the surface charge between −10 and −25 mV. Some studies also reported highly negatively charged membrane surface with a zeta potential below −40 mV.

3.2. Data imputation of missing entries

The missing entries in our dataset were recovered using 4 data imputation algorithms: SimpleFill, KNN, SoftImpute and MissForest. Randomisation of the dataset was required for SoftImpute and MissForest, and 3 different random states were used during the imputation. Results of different random states in Table S2 in the Supplementary material showed consistent RMSE values, and an average RMSE value was calculated. Table 1 presents the results of the data imputation where a smaller RMSE value indicates a higher accuracy of the imputation. MissForest outperformed the other methods, having smaller RMSE values in most parameters. The magnitude of the RMSE values varied among the parameters as it depends on the absolute value of the data. Our study demonstrated that the MissForest method can further improve the data imputation as compared to the most used SimpleFill method. Data imputation step is an important step when applying ML techniques in experimental data as missing data is commonly encountered. MissForest was therefore used to fill up the missing entries to obtain the final dataset for the predictive models.

Fig. 3a–c show the collected and imputed data points for the features: pressure, NaCl rejection and pH. The imputed values were within the reasonable range of data collected from the articles. Fig. 3d–f show the data distribution with the imputed data included. The distributions were similar to those before the imputation (Fig. 2a–c).

Table 1
Data imputation performance. The figures in bold represent the lowest RMSE for each parameter.

Parameter	RMSE			
	SimpleFill	KNN	SoftImpute	MissForest
Surface charge	12.9	14.9	12.3	**9.9**
Surface roughness	81.5	74.7	**72.1**	99.0
Contact angle	23.8	18.8	14.9	**13.6**
NaCl rejection	35.4	**4.1**	12.4	6.0
Permeate flux	22.1	22.1	19.1	**17.6**
Cross flow velocity	0.2	0.2	0.2	**0.1**
Pressure	12.7	13.2	16.0	**7.8**
Temperature	5.3	3.3	3.3	**2.5**
pH	2.0	2.2	**1.5**	1.6
NaCl concentration	6902.2	6363.3	7931.9	**1559.8**
Boron concentration	**110.0**	431.8	158.7	174.1

3.3. Boron rejection prediction models

After the data imputation, the complete dataset was used to develop predictive models for boron rejection of RO membranes using 5 different regression models. Without data imputation, the data entries containing missing values could not be used effectively by the regression methods to build the predictive models. All the variables listed in Table 1 were used as features of the modeling, except temperature as most studies only conducted the experiments at room temperature. A 5-fold cross validation was used for the regression assessments, and Table 2 reports the performance of each model in terms of their R^2, RMSE and MAE values. We first conducted the modeling without categorizing the dataset based on the membrane type. As shown in Table 2(a), the XGBoost regressor performed the best with the highest R^2 of 0.78. Compared to the tree-based models, both linear and ridge regressions did not give a good prediction of the boron rejection, with a low R^2 of 0.31. Some parameters in the dataset, such as pH and NaCl concentration behave like categorical variables. For example, as the pH was higher 9, the boron rejection increased significantly. This created a scenario where boron rejection performance is bimodal between pH below and above 9. In this scenario, algorithms capable of handling categorical data such as decision tree, random forest and XGBoost regressor would perform better. Our results also demonstrated the superiority of the ML

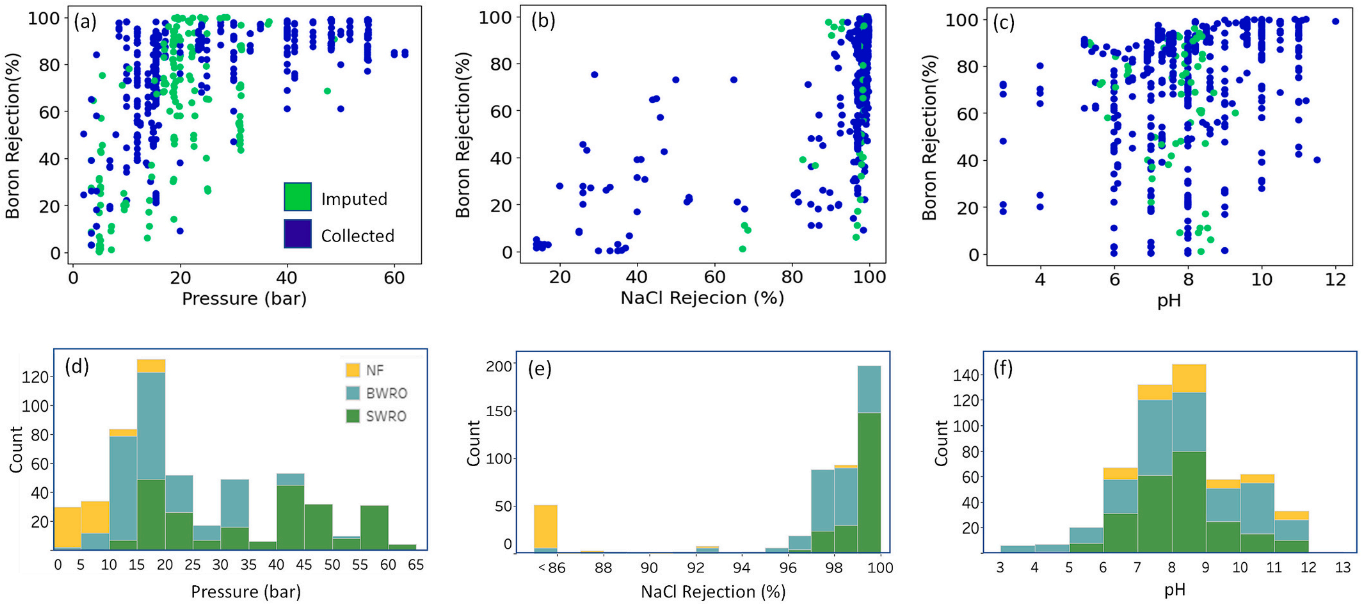


Fig. 3. Visualisation of recovered missing entries for (a) pressure, (b) NaCl rejection, and (c) pH. Data distribution after imputation for (d) pressure, (e) NaCl rejection, and (f) pH.

Table 2
Boron rejection prediction model performance: (a) without membrane type feature, (b) with membrane type feature, and (c) with membrane type feature and data of TDS/salt rejection. (σ is the standard deviation).

	R^2	RMSE	MAE
(a)			
Linear regression	0.31 (σ = 0.10)	21.44 (σ = 1.21)	16.38 (σ = 1.19)
Ridge regression	0.31 (σ = 0.10)	21.43 (σ = 1.20)	16.36 (σ = 1.17)
Decision tree regressor	0.60 (σ = 0.09)	16.28 (σ = 1.44)	11.30 (σ = 1.04)
Random forest regressor	0.62 (σ = 0.09)	15.86 (σ = 1.41)	11.35 (σ = 0.59)
XGBoost regressor	0.78 (σ = 0.05)	12.10 (σ = 1.25)	7.59 (σ = 0.59)
(b)			
Linear regression	0.66 (σ = 0.07)	14.84 (σ = 0.74)	10.99 (σ = 0.39)
Ridge regression	0.67 (σ = 0.07)	14.83 (σ = 0.73)	10.98 (σ = 0.39)
Decision tree regressor	0.76 (σ = 0.06)	12.47 (σ = 0.99)	8.64 (σ = 0.60)
Random forest regressor	0.76 (σ = 0.07)	12.42 (σ = 1.15)	8.60 (σ = 0.48)
XGBoost regressor	0.84 (σ = 0.03)	10.45 (σ = 1.12)	6.49 (σ = 0.65)
(c)			
Linear regression	0.70 (σ = 0.06)	14.09 (σ = 0.61)	10.19 (σ = 0.42)
Ridge regression	0.70 (σ = 0.06)	14.08 (σ = 0.60)	10.19 (σ = 0.43)
Decision tree regressor	0.74 (σ = 0.08)	12.94 (σ = 1.56)	8.72 (σ = 1.04)
Random forest regressor	0.79 (σ = 0.05)	11.85 (σ = 0.96)	8.03 (σ = 0.80)
XGBoost regressor	0.84 (σ = 0.03)	10.09 (σ = 0.52)	6.08 (σ = 0.31)

models in handling the dataset from the membrane research area.

It was shown earlier in Fig. 2 that the data distributions varied for different membrane types. The testing conditions and membrane performance for SWRO, BWRO and NF membranes could be quite different. To enable the models to learn better using the dataset, we introduced a new feature: membrane type. SWRO, BWRO and NF membranes were encoded into 1, 2 and 3, respectively, for the membrane type feature. The regression results with the addition feature are shown in Table 2(b). The performance of all the models improved significantly, with a higher R^2 value and lower RMSE. The R^2 value of the best model (XGBoost regressor) improved from 0.78 to 0.84, while the RMSE decreased from 12.10 to 10.45. The additional feature also improved the non-tree regression methods, linear and ridge regressions. Our study demonstrated that categorizing the dataset can be an effective way to improve the learning process, especially for data with multiple distributions.

From the data collection, we observed that not all the articles reported the NaCl rejection of the membranes, resulting in a high percentage of missing data in this feature. For studies testing mixed solute solutions or real feeds like seawater, the TDS or salt rejection was more commonly reported. Introducing a separate feature for TDS or salt rejection would cause more missing data as most studies did not report this information. As NaCl was still the main component in the mixed solute solutions, the TDS/salt rejection would be reasonably close to the NaCl rejection of the membranes. Therefore, we performed an assessment by including TDS/salt rejection data into the NaCl rejection feature for the regression fitting, and the results are shown in Table 2(c). The R^2 values further increased and the RMSE decreased for nearly all regression methods. The XGBoost regression still performed the best and had an R^2 value of 0.84. The utilization of TDS/salt rejection reduced the missing data in the NaCl rejection feature, which could improve the data quality for modeling. For the subsequent analysis, we have included the TDS/salt rejection data.

To further test the performance of the regression fittings, the dataset was divided into 80 % for training and 20 % for testing. The training dataset was first used to train the regression model, which was afterwards used to predict the boron rejection of both the training and testing datasets. The predicted data were plotted against collected data and the R^2 score was calculated, as shown in Fig. 4. The tree-based ML algorithms: decision tree, random forest and XGB regressors had R^2 values of 0.64, 0.75 and 0.79, respectively, which outperformed the linear and ridge regressions, both with an R^2 of 0.62. There was a good agreement between predicted and collected data, especially in the high boron rejection region. This can be attributed to more data points being collected for the high boron rejection region, as shown in Fig. 1b.

The relative contributions of each parameter to the boron rejection performance of the membranes in different predictive models are plotted in Fig. 5. The contribution of pH was consistently high across these models; it was the most important parameter in the decision tree and XGBoost regressors and the second most important parameter in the random forest regressor. NaCl rejection was also shown as an important parameter across different models. Membranes with the capability to reject NaCl and salt efficiently are more likely to have a more selective and defect-free separating layer that could have a better boron rejection [15]. Membrane properties such as surface charge, surface roughness and contact angle did not exhibit strong effects on the boron rejection

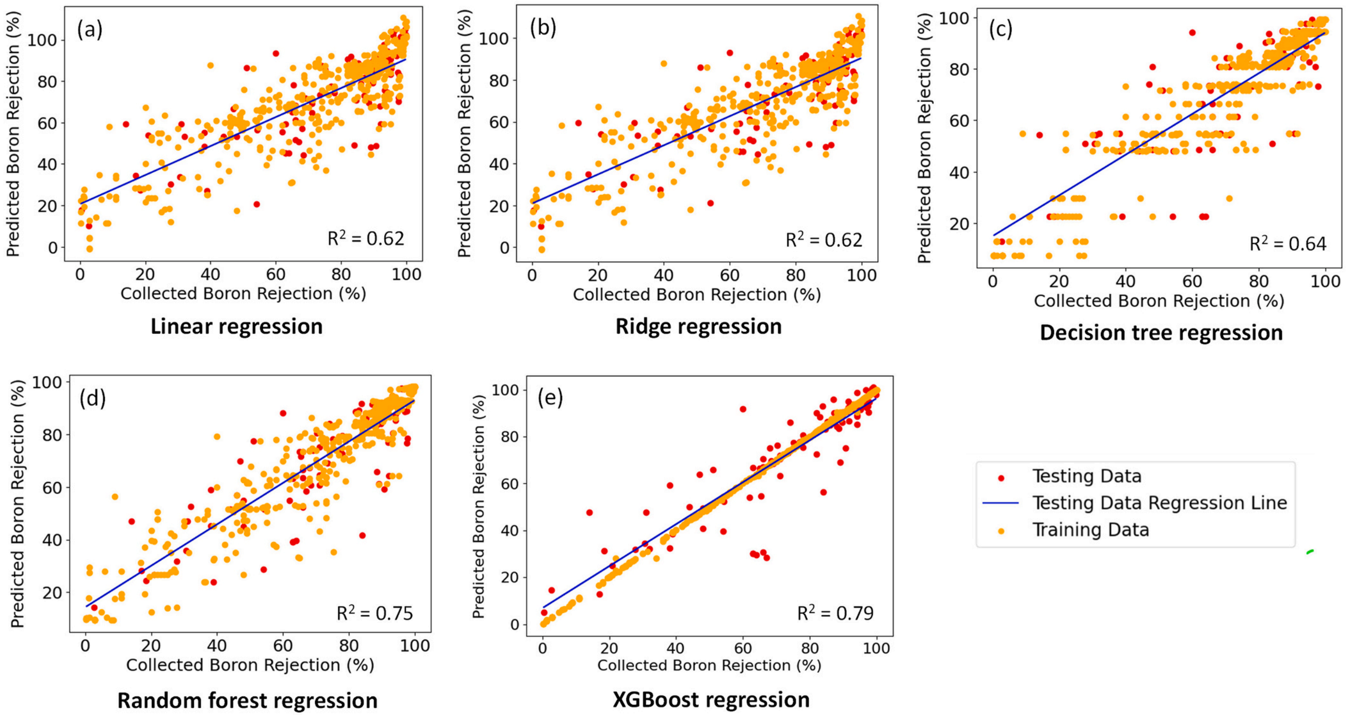


Fig. 4. Boron rejection prediction model performance: (a) linear regression, (b) ridge regression, (c) decision tree regression, (d) random forest regression, and (e) XGBoost regression.

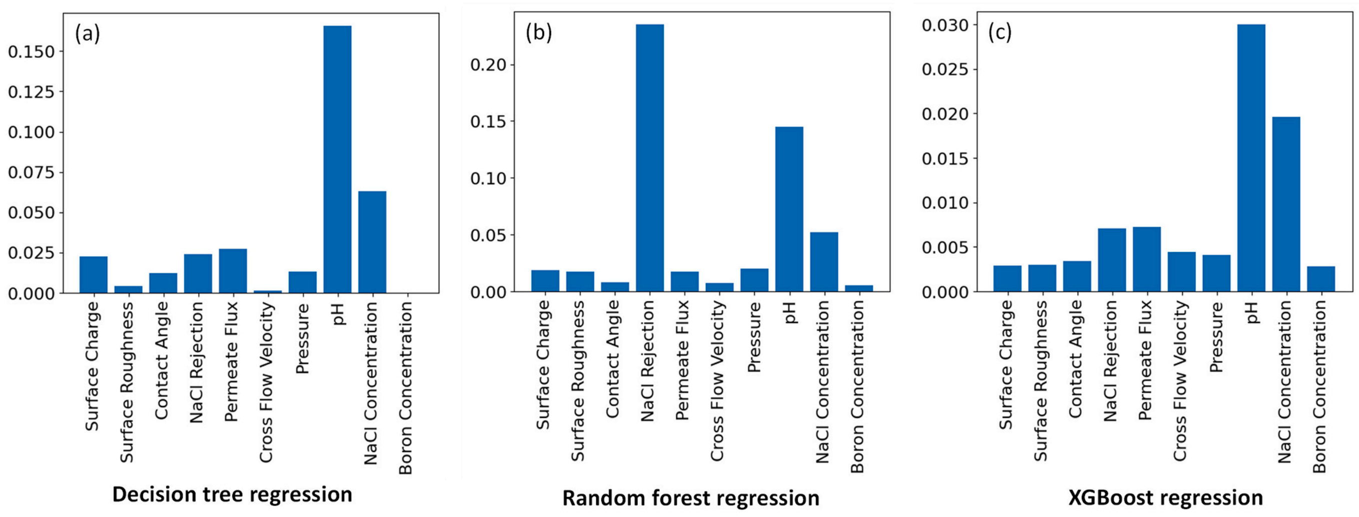


Fig. 5. Feature Importance generated by (a) decision tree regression, (b) random forest regression, and (c) XGBoost regression.

performance.

The importance of the features was also examined using a tree diagram and Fig. 6 illustrates the diagram generated by the decision tree regressor. The tree first segregated the data points into 3 clusters according to the membrane type. The tree model identified that SWRO, BWRO and NF membranes had distinct boron removal performance, and thus classified them in the first node of the tree. Within each membrane type cluster, the tree subsequently divided the data points based on the pH, with pH ~9 as the cut-off. The acid dissociation constant of boric acid is 9.2, where charged boron species becomes dominant above this pH. The boron rejection will increase significantly at pH >9 as the membranes can retain the charged species more effectively through surface charge effect [8]. The two most important and influential features, membrane types and pH, were captured at the upper nodes of the tree. The decision tree is capable to generalise and learn from the dataset intuitively, producing outcome that aligns with current knowledge.

The tree diagram generated not only shows the important features but also the conditions when a high boron rejection can be achieved. For SWRO membranes tested at pH <8.8, a higher pressure >15.8 bar could contribute to a higher boron rejection. The next level of the branch showed that NaCl rejection >99.6 % would yield a high boron rejection of 92.8 %. Though SWRO membranes generally have a high NaCl rejection >99 %, the ML model suggested that a higher NaCl rejection is likely needed to ensure a high boron rejection. This aligns with the understanding that a defect-free polyamide separating layer with a high crosslinking density is needed to achieve high NaCl rejection [15]. Such highly selective membranes will also have excellent solute sieving properties, which will be beneficial for boron rejection. However, the

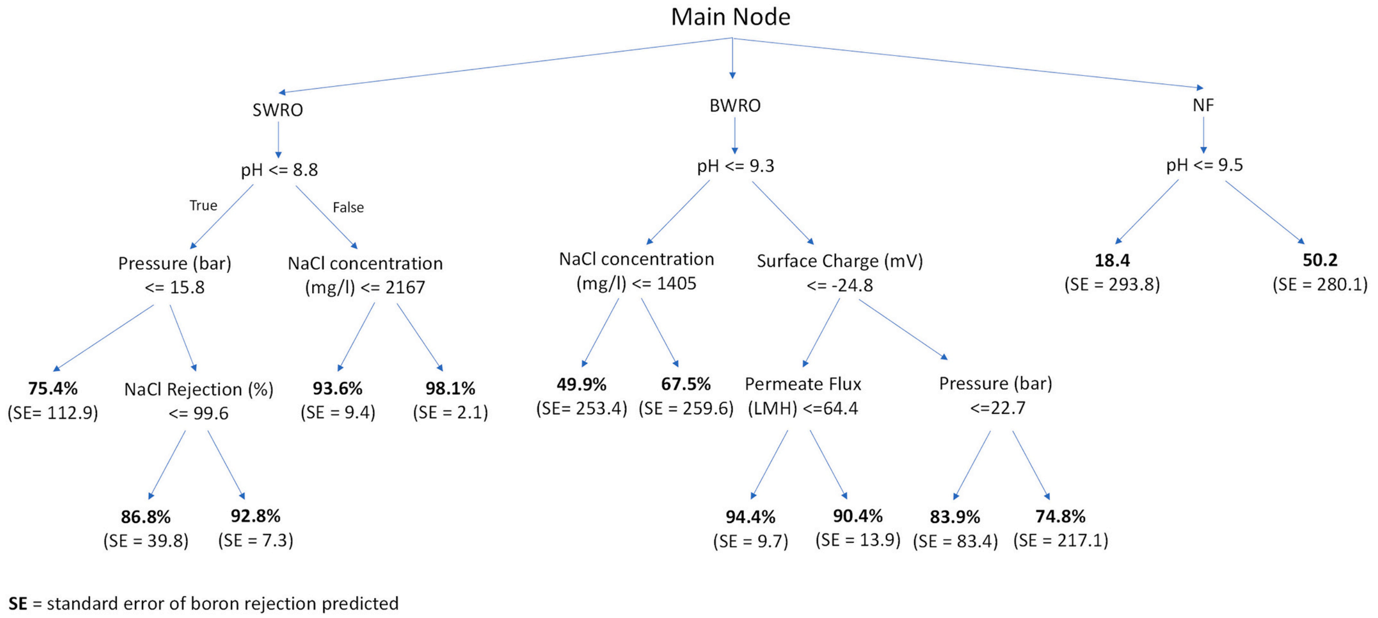


Fig. 6. A snip of tree diagram generated from the decision tree regressor.

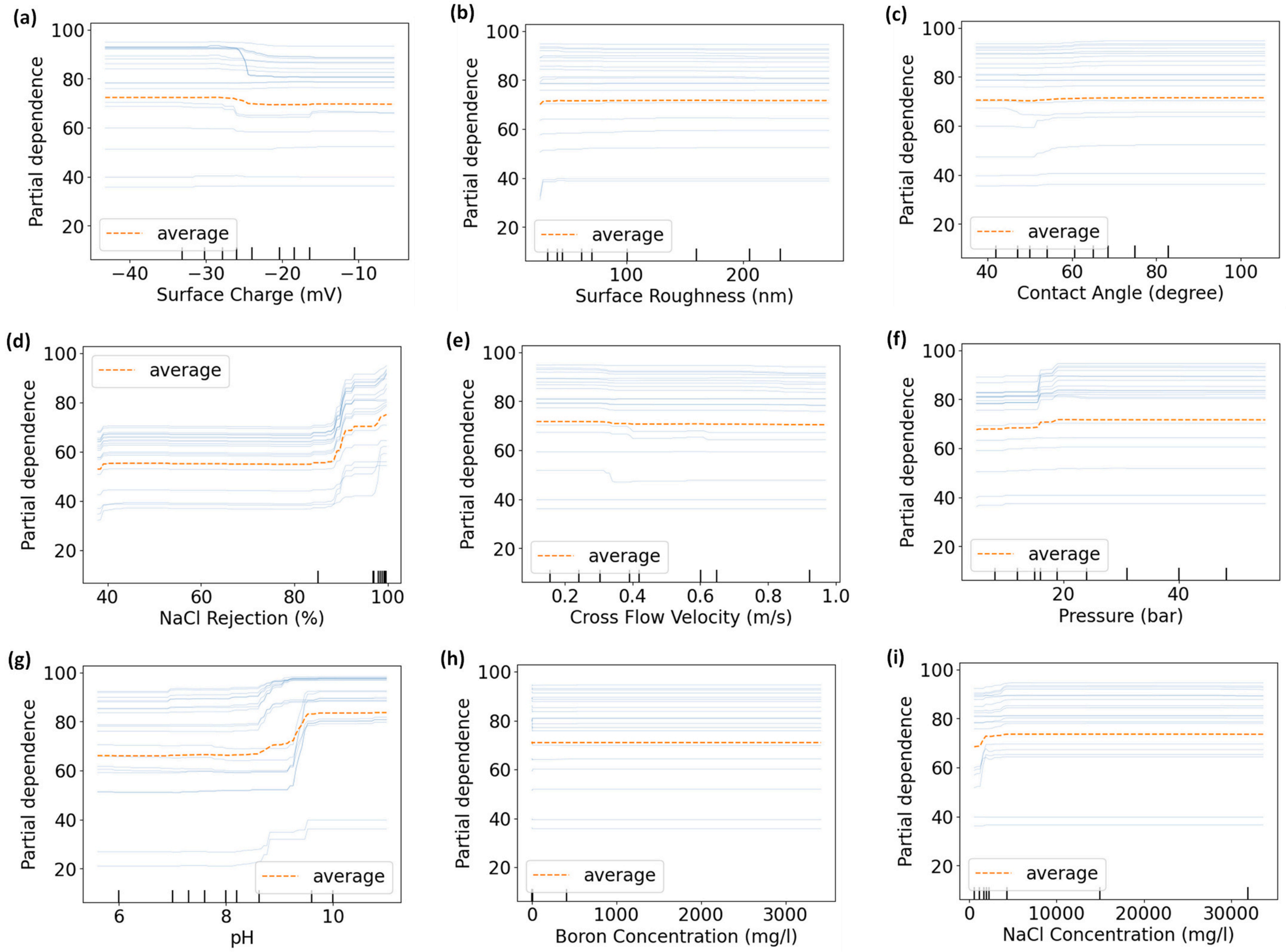


Fig. 7. Univariate PDP for (a) surface charge, (b) surface roughness, (c) contact angle, (d) NaCl rejection, (e) cross flow velocity, (f) pressure, (g) pH, (h) boron concentration, and (i) NaCl concentration.

higher NaCl rejection may compromise the water permeability of the membranes. Breakthroughs in membrane material synthesis are required to overcome the permeability-selectivity trade-off relationship. On the other hand, BWRO membranes generally had a lower boron rejection, and a higher pH > 9.3 was required in order to achieve a high boron rejection. The membrane surface charge was shown to be an important property to have a high boron rejection at high pH, and a highly negatively charged surface with zeta potential lower than −24.8 eV is favorable. For NF membranes, the boron rejection unfortunately was very low even when the pH was high. Though NF membranes have a high water permeability, the looser structure of the separating layer was insufficient for boron removal applications. The full tree generated in our experiment is included in Fig. S1 in the Supplementary material.

3.4. Univariate and bivariate partial dependence analyses

To further assess the effects of each feature on the boron rejection performance, partial dependence analysis was conducted. The univariate partial dependence plots (PDP) were obtained by varying one parameter at a time, and the boron rejection was predicted with the conditions of other parameters remained unchanged. An average was obtained from all the data points and depicted using orange dotted line in Fig. 7. The predictive model developed from the random forest regressor was used in this analysis. Though the XGBoost regressor gave slightly higher accuracy (in Section 3.3), the PDP produced had more noise. This could be due to the small dataset and overfitting, making result interpretation challenging. The PDP generated from the XGBoost regressor is included in the Supplementary material for reference. Among all the features, only NaCl rejection and pH showed significant effects on the boron rejection. As the NaCl rejection increased above 90 %, the boron rejection increased by about 10 %, and it further increased when the NaCl rejection was above 99 %. This result supports the observation in the tree diagram (Fig. 6) that NF membranes with a low NaCl rejection <90 % may not be suitable boron removal even in the second pass of a two-pass RO process. RO membranes with a very high NaCl rejection will enhance the boron removal performance, and this observation is consistent with the "defect-plugging" approach taken by many researchers when developing RO membranes [15]. Meanwhile, as the pH was increased above 9, the boron rejection increased by about 15 %. The result shows that pH around 10 could be sufficient to obtain the optimal effect of pH on boron removal. For other operating conditions, higher pressure showed a slight positive effect on the boron rejection, but the cross-flow velocity did not have much effect. The effects of membrane properties, surface charge, surface roughness and contact angle was also minimum.

Bivariate dependence analysis was conducted to further understand the interaction between two features and their mutual effects. Likewise, the random forest regressor was used to produce the bivariate PDP in Fig. 8. From the earlier finding, pH consistently showed as an prominent feature and therefore was chosen as the primary parameter to study with another parameter. Membrane surface roughness, contact angle and crossflow velocity showed minimum effect on the boron rejection across different pH range. A slight effect was observed for surface charge at pH <8, but the effect of pH remained dominant. Surface modification of RO membranes to improve their surface charge may not yield substantial improvement especially when most RO membranes are already highly negatively charged. However, surface charge lower than −24.8 eV is still vastly desirable for BWRO membranes applied at high pH, as discussed in Section 3.3. On the other hand, NaCl rejection and pressure showed some influence on boron rejection when pH is below 9. The values of boron rejection shown in the figures were relatively low as all three types of membranes were included in this analysis.

We also generated the bivariate PDP for SWRO and BWRO membranes separately, focusing on key features, as illustrated in Fig. 9. In seawater desalination, seawater is usually fed directly to the SWRO membrane unit. Changing the pH of the feed is usually impractical due to the large amount of water being processed. Furthermore, the high salt concentration of seawater also makes it difficult to control the pH. The

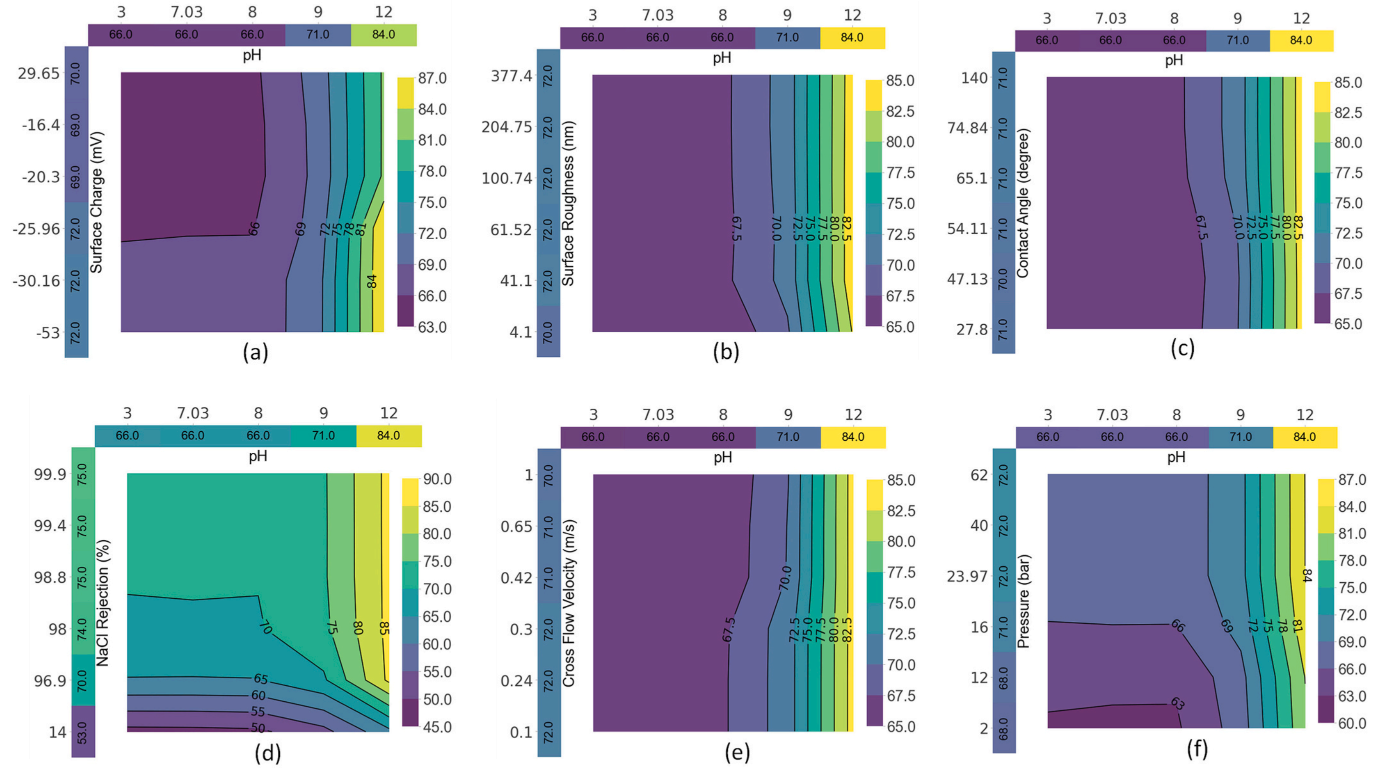


Fig. 8. Bivariate PDP between pH and (a) surface charge, (b) surface roughness, (c) contact angle, (d) NaCl rejection, (e) cross-flow velocity, and (f) pressure.

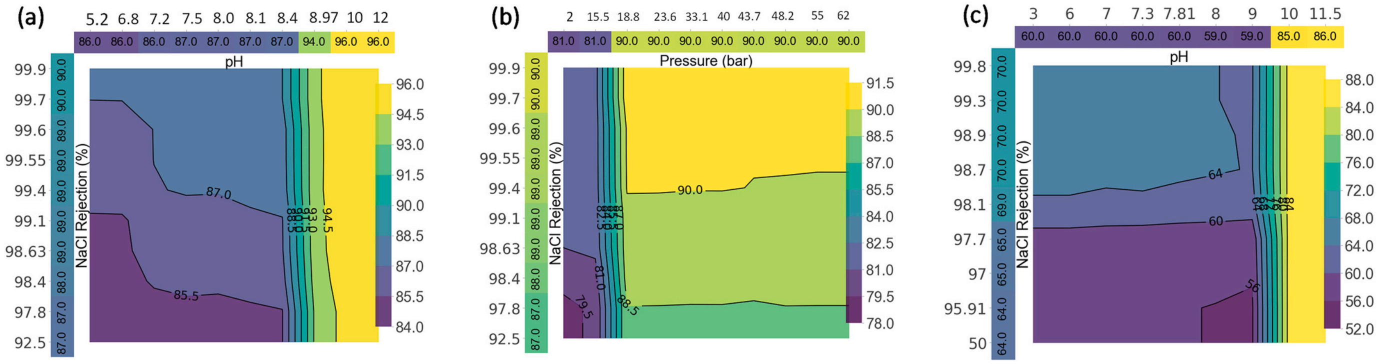


Fig. 9. Bivariate PDP of (a, b) SWRO membranes and (c) BWRO membranes.

pH of seawater is about 8, and it can be observed from Fig. 9(a) that NaCl rejection >99.4 % could further improve the boron rejection of SWRO membranes. Similar observation was shown in the NaCl rejection – pressure PDP in Fig. 9(b). High operating pressure is essential for achieving high boron rejection, and higher NaCl rejection could further enhance the boron retention. On the other hand, BWRO membranes are more likely used in the second pass of a two-pass RO process. High pH is needed for high boron rejection, but a very high NaCl rejection may not further improve the boron rejection. Therefore, relatively loose BWRO membranes can be used if a high pH is applied. The higher water permeability of looser BWRO membranes can potentially reduce the energy consumption of the process.

4. Conclusions

In this study, we successfully developed predictive models for boron removal performance of RO membranes using ML techniques based on published experimental data. Data of 11 features on membrane properties, operating conditions and membrane performance was first collected from the literatures. The missing data entries was then recovered through data imputation and several techniques including SimpleFill, KNN, SoftImpute and MissForest were examined. Our results demonstrated that ML-based MissForest algorithm could best impute the missing experimental data. The predictive models for boron removal performance were subsequently developed by training the regression models using the imputed dataset. Five regression models were studied and the tree-based algorithms (decision tree, random forest and XGBoost regressors) outperformed the linear and ridge regressions. The XGBoost regressor demonstrated the best performance, with an R^2 of 0.84, which could be attributed to its ability to handle categorical parameters. Our study also showed two approaches to improve the training of the ML models. Firstly, the appropriate classification of the dataset enhanced the trainings, and the accuracy of the models improved significantly with the additional membrane type feature. Secondly, including the salt/TDS rejection data in the NaCl rejection feature further improved the training by reducing the missing entries in the dataset. To further improve the accuracy of the ML models, additional experiments can be designed in future work to obtain more data for model trainings.

The effects of the features on the boron removal performance were studied by analysing the feature importance in the predictive models and the tree diagram. The membrane type, pH and NaCl rejection appeared to be the three most important parameters in affecting the boron rejection performance. The tree diagram showed that pH >9 was a key cut-off point for high boron removal rate, and NaCl rejection >99.6 % was required to achieve a high boron rejection for SWRO membranes at pH < 9. Membrane properties such as surface charge, surface roughness and contact angles showed minimum effects on the boron rejection. However, membrane surface charge of <−24.8 eV is still desirable for high boron rejection in BWRO at pH > 9. Furthermore, the predictive models were used to generate univariate and bivariate PDP, allowing further analysis of the variables under a wider range of conditions. pH and NaCl rejection continued to show as the most prominent parameters in determining high boron rejection. For SWRO, defect-free membranes with a high NaCl rejection >99.6 % are highly desirable for high boron rejection. However, for the application of BWRO in a second pass RO process, a looser structure with a NaCl rejection of >95 % could perform well at pH >9. Our results demonstrated that ML could effectively learn from the dataset, producing meaningful outcomes that align with current knowledge. The models also identified conditions for further improving the boron removal performance. Since this study relied on data from published papers, the explored features were limited. Further experiments are necessary to obtain sufficient relevant data to explore the effects of other features. The predictive models developed can serve as a useful tool for researchers and membrane users to predict the membrane performance. The detailed analysis also provides researchers with more systematic guidelines for developing membranes and designing membrane processes.

CRediT authorship contribution statement

Sukarno: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. **Jeng Yi Chong:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Gao Cong:** Writing – review & editing, Validation, Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

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.

Data availability

Data will be made available on request.

Acknowledgements

The authors would like to acknowledge Singapore Telecommunications Limited (Singtel) for its support, through Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU). We also acknowledge Mr. Shuai Liu for his assistance and valuable suggestions on the project. Jeng Yi Chong would also like to acknowledge Nanyang Technological University for the Presidential Postdoctoral Fellowship.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.desal.2024.117854.

References

[1] M. Ayaz, M.A. Namazi, M.A. ud Din, M.I.M. Ershath, A. Mansour, el H.M. Aggoune, Sustainable seawater desalination: current status, environmental implications and future expectations, Desalination 540 (2022) 116022, https://doi.org/10.1016/J.DESAL.2022.116022.

[2] M. Elimelech, W.A. Phillip, The future of seawater desalination: energy, technology, and the environment, Science 333 (2011) (1979) 712–717, https://doi.org/10.1126/SCIENCE.1200488/SUPPL_FILE/ELIMELECH.SOM.PDF.

[3] H. Nassrullah, S.F. Anis, R. Hashaikeh, N. Hilal, Energy for desalination: a state-of-the-art review, Desalination 491 (2020) 114569, https://doi.org/10.1016/J.DESAL.2020.114569.

[4] S. Lin, H. Zhao, L. Zhu, T. He, S. Chen, C. Gao, L. Zhang, Seawater desalination technology and engineering in China: a review, Desalination 498 (2021) 114728, https://doi.org/10.1016/J.DESAL.2020.114728.

[5] Y.J. Lim, K. Goh, M. Kurihara, R. Wang, Seawater desalination by reverse osmosis: current development and future challenges in membrane fabrication – a review, J. Membr. Sci. 629 (2021) 119292, https://doi.org/10.1016/J.MEMSCI.2021.119292.

[6] N. Najid, S. Kouzbour, A. Ruiz-Garcia, S. Fellaou, B. Gourich, Y. Stiriba, Comparison analysis of different technologies for the removal of boron from seawater: a review, J. Environ. Chem. Eng. 9 (2021) 105133, https://doi.org/10.1016/J.JECE.2021.105133.

[7] E. Güler, C. Kaya, N. Kabay, M. Arda, Boron removal from seawater: state-of-the-art review, Desalination 356 (2015) 85–93, https://doi.org/10.1016/J.DESAL.2014.10.009.

[8] H. Koseoglu, N. Kabay, M. Yüksel, S. Sarp, Ö. Arar, M. Kitis, Boron removal from seawater using high rejection SWRO membranes - impact of pH, feed concentration, pressure, and cross-flow velocity, Desalination 227 (2008), https://doi.org/10.1016/j.desal.2007.06.029.

[9] S. Bolan, H. Wijesekara, D. Amarasiri, T. Zhang, P. Ragályi, M. Brdar-Jokanović, M. Rékási, J.Y. Lin, L.P. Padhye, H. Zhao, L. Wang, J. Rinklebe, H. Wang, K.H. M. Siddique, M.B. Kirkham, N. Bolan, Boron contamination and its risk management in terrestrial and aquatic environmental settings, Sci. Total Environ. 894 (2023) 164744, https://doi.org/10.1016/J.SCITOTENV.2023.164744.

[10] L.J. Banasiak, A.I. Schäfer, Removal of boron, fluoride and nitrate by electrodialysis in the presence of organic matter, J. Membr. Sci. 334 (2009) 101–109, https://doi.org/10.1016/J.MEMSCI.2009.02.020.

[11] S. Shultz, M. Bass, R. Semiat, V. Freger, Modification of polyamide membranes by hydrophobic molecular plugs for improved boron rejection, J. Membr. Sci. 546 (2018), https://doi.org/10.1016/j.memsci.2017.10.003.

[12] Y. Li, S. Wang, X. Song, Y. Zhou, H. Shen, X. Cao, P. Zhang, C. Gao, High boron removal polyamide reverse osmosis membranes by swelling induced embedding of a sulfonyl molecular plug, J. Membr. Sci. 597 (2020) 117716, https://doi.org/10.1016/J.MEMSCI.2019.117716.

[13] H. Raval, V. Sundarkumar, Low-energy reverse osmosis membrane with high boron rejection by surface modification with a polysaccharide, Can. J. Chem. Eng. 97 (2019), https://doi.org/10.1002/cjce.23375.

[14] X. Zhai, J. Meng, R. Li, L. Ni, Y. Zhang, Hypochlorite treatment on thin film composite RO membrane to improve boron removal performance, Desalination 274 (2011) 136–143, https://doi.org/10.1016/J.DESAL.2011.02.001.

[15] Z. Ali, Y. Al Sunbul, F. Pacheco, W. Ogieglo, Y. Wang, G. Genduso, I. Pinnau, Defect-free highly selective polyamide thin-film composite membranes for desalination and boron removal, J. Membr. Sci. 578 (2019) 85–94, https://doi.org/10.1016/J.MEMSCI.2019.02.032.

[16] X. Liu, C. Xu, P. Chen, K. Li, Q. Zhou, M. Ye, L. Zhang, Y. Lu, Advances in technologies for boron removal from water: a comprehensive review, Int. J. Environ. Res. Public Health 19 (2022), https://doi.org/10.3390/ijerph191710671.

[17] Y. Du, L. Xie, Y. Liu, S. Zhang, Y. Xu, Optimization of reverse osmosis networks with split partial second pass design, Desalination 365 (2015) 365–380, https://doi.org/10.1016/J.DESAL.2015.03.019.

[18] B. Chen, F. Li, X. Zhao, Boron removal with modified polyamide RO modules by cross-linked glutaric dialdehyde grafting, J. Chem. Technol. Biotechnol. 96 (2021), https://doi.org/10.1002/jctb.6561.

[19] A.M. Schweidtmann, E. Esche, A. Fischer, M. Kloft, J.U. Repke, S. Sager, A. Mitsos, Machine learning in chemical engineering: a perspective, Chem. Ing. Tech. 93 (2021), https://doi.org/10.1002/cite.202100083.

[20] H. Yin, M. Xu, Z. Luo, X. Bi, J. Li, S. Zhang, X. Wang, Machine learning for membrane design and discovery, Green Energy Environ. 9 (2024) 54–70, https://doi.org/10.1016/J.GEE.2022.12.001.

[21] G. Ignacz, N. Alqadhi, G. Szekely, Explainable machine learning for unraveling solvent effects in polyimide organic solvent nanofiltration membranes, Adv. Membr. 3 (2023) 100061, https://doi.org/10.1016/J.ADVMEM.2023.100061.

[22] C.S.H. Yeo, Q. Xie, X. Wang, S. Zhang, Understanding and optimization of thin film nanocomposite membranes for reverse osmosis with machine learning, J. Membr. Sci. 606 (2020), https://doi.org/10.1016/j.memsci.2020.118135.

[23] J. Hu, C. Kim, P. Halasz, J.F. Kim, J. Kim, G. Szekely, Artificial intelligence for performance prediction of organic solvent nanofiltration membranes, J. Membr. Sci. 619 (2021), https://doi.org/10.1016/j.memsci.2020.118513.

[24] T. Zhu, Y. Zhang, C. Tao, W. Chen, H. Cheng, Prediction of organic contaminant rejection by nanofiltration and reverse osmosis membranes using interpretable machine learning models, Sci. Total Environ. 857 (2023), https://doi.org/10.1016/j.scitotenv.2022.159348.

[25] C.L. Ritt, T. Stassin, D.M. Davenport, R.M. DuChanois, I. Nulens, Z. Yang, A. Ben-Zvi, N. Segev-Mark, M. Elimelech, C.Y. Tang, G.Z. Ramon, I.F.J. Vankelecom, R. Verbeke, The open membrane database: synthesis–structure–performance relationships of reverse osmosis membranes, J. Membr. Sci. 641 (2022), https://doi.org/10.1016/j.memsci.2021.119927.

[26] A. Thornton, B. Freeman, L. Robeson, Polymer Gas Separation Membrane Database. https://membrane-australasia.org/polymer-gas-separation-membrane-database/, 2012.

[27] N.I. Ajali-Hernández, A. Ruiz-Garćıa, C.M. Travieso-González, ANN based-model for estimating the boron permeability coefficient as boric acid in SWRO desalination plants using ensemble-based machine learning, Desalination 573 (2024) 117180, https://doi.org/10.1016/J.DESAL.2023.117180.

[28] Q. Yuan, M. Longo, A.W. Thornton, N.B. McKeown, B. Comesaña-Gándara, J. C. Jansen, K.E. Jelfs, Imputation of missing gas permeability data for polymer membranes using machine learning, J. Membr. Sci. 627 (2021), https://doi.org/10.1016/j.memsci.2021.119207.

[29] H. Gao, S. Zhong, W. Zhang, T. Igou, E. Berger, E. Reid, Y. Zhao, D. Lambeth, L. Gan, M.A. Afolabi, Z. Tong, G. Lan, Y. Chen, Revolutionizing Membrane Design Using Machine Learning-Bayesian Optimization, in: Cite This: Environ. Sci. Technol, 2022, p. 2572, https://doi.org/10.1021/acs.est.1c04373.

[30] S. Liu, X. Li, G. Cong, Y. Chen, Y. Jiang, Multivariate Time-Series Imputation With Disentangled Temporal Representations. The Eleventh International Conference on Learning Representations, 2023. https://openreview.net/forum?id=rdjeCNUS6TG.

[31] R. Mazumder, T. Hastie, R. Tibshirani, Spectral regularization algorithms for learning large incomplete matrices, J. Mach. Learn. Res. 11 (2010).

[32] D.J. Stekhoven, P. Bühlmann, Missforest-non-parametric missing value imputation for mixed-type data, Bioinformatics 28 (2012), https://doi.org/10.1093/bioinformatics/btr597.

[33] Y. Wu, Can't ridge regression perform variable selection? Technometrics 63 (2021) https://doi.org/10.1080/00401706.2020.1791254.

[34] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, 2017, https://doi.org/10.1201/9781315139470.

[35] L. Breiman, Random forests, Mach. Learn. 45 (2001), https://doi.org/10.1023/A:1010933404324.

[36] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016), https://doi.org/10.1145/2939672.2939785.

[37] E. Güler, N. Kabay, M. Yüksel, E. Yavuz, Ü. Yüksel, A comparative study for boron removal from seawater by two types of polyamide thin film composite SWRO membranes, Desalination 273 (2011) 81–84, https://doi.org/10.1016/J.DESAL.2010.10.045.

[38] P.P. Mane, P.K. Park, H. Hyung, J.C. Brown, J.H. Kim, Modeling boron rejection in pilot- and full-scale reverse osmosis desalination processes, J. Membr. Sci. 338 (2009) 119–127, https://doi.org/10.1016/J.MEMSCI.2009.04.014.

[39] E. Yavuz, Ö. Arar, M. Yüksel, Ü. Yüksel, N. Kabay, Removal of boron from geothermal water by RO system-II-effect of pH, Desalination 310 (2013) 135–139, https://doi.org/10.1016/J.DESAL.2012.07.044.

[40] X. Ma, Z. Yang, Z. Yao, H. Guo, Z. Xu, C.Y. Tang, Tuning roughness features of thin film composite polyamide membranes for simultaneously enhanced permeability, selectivity and anti-fouling performance, J. Colloid Interface Sci. 540 (2019), https://doi.org/10.1016/j.jcis.2019.01.033.