Machine Learning-Based Modelling and Analysis of Water/Salt Selectivity of Polyamide Nanofiltration Membranes

Palvi Soni[1], Gajendra Tandan[2]

[1]Assistant Professor, Department of CS & IT, Kalinga University, Raipur, India.
[2]Research Scholar, Department of CS & IT, Kalinga University, Raipur, India.

Nanofiltration (NF) is becoming increasingly crucial in separating water and salts, requiring specialized membranes that meet particular demands in various application scenarios. The lack of clarity in the separation mechanisms of NF, such as membrane construction characteristics and functioning circumstances, obstructs the systematic development of adaptable NF Membranes (NFM). In this research, Machine Learning (ML) was employed to investigate the relationship between membrane structural variables and operating circumstances concerning water/salt selectivity. The analysis highlighted the significance of various features based on existing literature information. Two critical characteristics of polyamide NFM and two specific operational factors were identified and linked to the membrane's ability to separate water from salts selectively. The study utilized Random Forest (RF) and XGBoost (XGB) models to analyze critical variables and determine their significance. The membrane structure variables emphasized the selectivity of water and salts compared to the impact on operational conditions. This influence varied depending on the kind of salts, with symmetrical salts primarily affected by size sieving. The extensive insights obtained can be utilized to address the optimization of effectiveness and designing multi-scenario NFM.

Keywords: Machine Learning, Nanofiltration, Water/Salt Selectivity, Polyamide.

1. Introduction

Nanofiltration (NF) is a cost-effective and energy-efficient method of separating substances using a membrane [1]. Regarding its capabilities, it is positioned between Reverse Osmosis (RO) and Ultrafiltration (UF). NF is used in many applications such as desalting, ion

segregation, recovery of resources, and water softener [2]. The membrane efficiency criteria vary in various applications. In desalination, there is a need for higher rejection of monovalent salts. In ion separation, there is a need for a higher rejection of divalent salts while simultaneously having a low rejection of monovalent sodium salts [7]. In dye desalting, the focus is on minimal rejection of divalent salts.

Customizing NF membranes with varying structural characteristics is imperative to provide adjustable membrane capabilities in various application circumstances. To effectively produce NF Membranes (NFM) for multiple applications, it is essential to have a thorough understanding of the behavior of ionic transportation across the membrane. The polyamide NFM, created using Interfacial Polymerization (IP), has a high level of development [3]. These NFMs primarily rely on their sub-nanoscale pore dimensions and electrokinetic characteristics to regulate the movement of ions.

Machine Learning (ML) is a data-driven method that allows for using algorithms in activities such as categorization, regression forecasting, and grouping [12]. It has become an essential tool for sophisticated substance identification and efficiency prediction [8]. ML facilitates the NF procedure because of its exceptional capacity to address complex challenges [9]. ML approaches applied to NFM have made significant progress in various areas [4]. These include identifying the microstructure of membrane surfaces, predicting membrane movement and rejection efficiency, understanding the key factors that control ion selection, predicting the effectiveness of removing micro-pollutants, screening sophisticated membrane substances, and constructing model systems [6].

The research created a database using 100 literary works to apply ML techniques to separate the relationship between multiple factors and the selectivity of water and salts [11]. The four characteristics of the polyamide NFM, which include two structural variables and two operational variables, were retrieved and linked to the ability of the membrane to reject four different types of salts. The Random Forest (RF) [5] and XGBoost (XGB) [10] algorithms were utilized to train and assess relevant variables. 80% of the information in the database was allocated for model training, while the remaining 20% was used for forecasting performance and verification. Both models strongly aligned with the experimental data set, as indicated by the training performance.

2. Methods and Materials

2.1 Sample data collection

The empirical data on polyamide NFM were obtained from 100 scholarly articles, yielding 360 data points. The selection of polyamide NFM materials for inclusion in this study was based on the results of the Web of Science repository up until June 2023. Any empty or anomalous data points were eliminated during the information processing phase. Four characteristics, precisely two structural characteristics of pore diameter and zeta possibility, as well as two operational settings of pressure and feed attention, were carefully identified from the gathered information. These characteristics were then utilized as the target variable for machine learning. The pore diameter was determined using the measuring technique, whereas the zeta possibility was determined utilizing the streamed zeta possibility technique.

Four metrics to assess membrane efficiency are the rejection rates of symmetric salts such as NaCl and MgSO4 and asymmetrical salts like Na204 and MgCl. The polyamide NFM obtained was pure and manufactured using the interstitial polymerization process.

2.2 Boosting Tree Ensemble (BTE) model

This study used two designs, RF and XGBoost, to implement machine learning. RF is a supervised ensemble ML method that relies on several decision trees. It utilizes bagging and randomized sub-space strategies. A stochastic sampling of the training database produces these trees. Every node of one tree is divided using a user-defined number of randomly picked characteristics linked with a hyperrectangular cell. In regression problems, the final estimation is obtained by calculating the average efficiency of each tree. XGB, an ensemble ML method, is comparable to RF because it utilizes a boosting technique to combine several regression trees. The distinction lies in the fact that every regression tree generates a continuous rating, and the ultimate projection is the summation of the scores estimated by each tree.

This study enhanced the model's efficiency by splitting the training dataset into a validating set for fine-tuning the hyperparameters. The evaluating set was then used to evaluate the forecasting accuracy of the improved model. The dataset was divided in a proportion of 65:15:20 for the learning, validating, and assessing sets, respectively. The grid search technique was employed to do hyperparameter optimization. This involved methodically iterating over multiple values of the learning rates, maximum depth, and n estimations to choose the optimal variable. The R2, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) variables were assessed to determine the optimum variable.

The hyperparameters of the RF approach, particularly maximum depth and count of predictions, were rigorously improved. The learning rate was adjusted from 0.04 to 0.09, with an increment size of 0.01. The maximum depth was selected from the range of 5 to 7, with a step size of 1. The number of estimations was determined to be 100, 120, or 200. The ideal variables were selected by iterative calculation and fitting assessment.

Once the optimum variables were determined, the research employed a training set-testing set splitting ratio of 80:20 to train the algorithm and forecast its performance (Figs. 1). The ML libraries Scikit-learn and XGB were utilized for this goal.

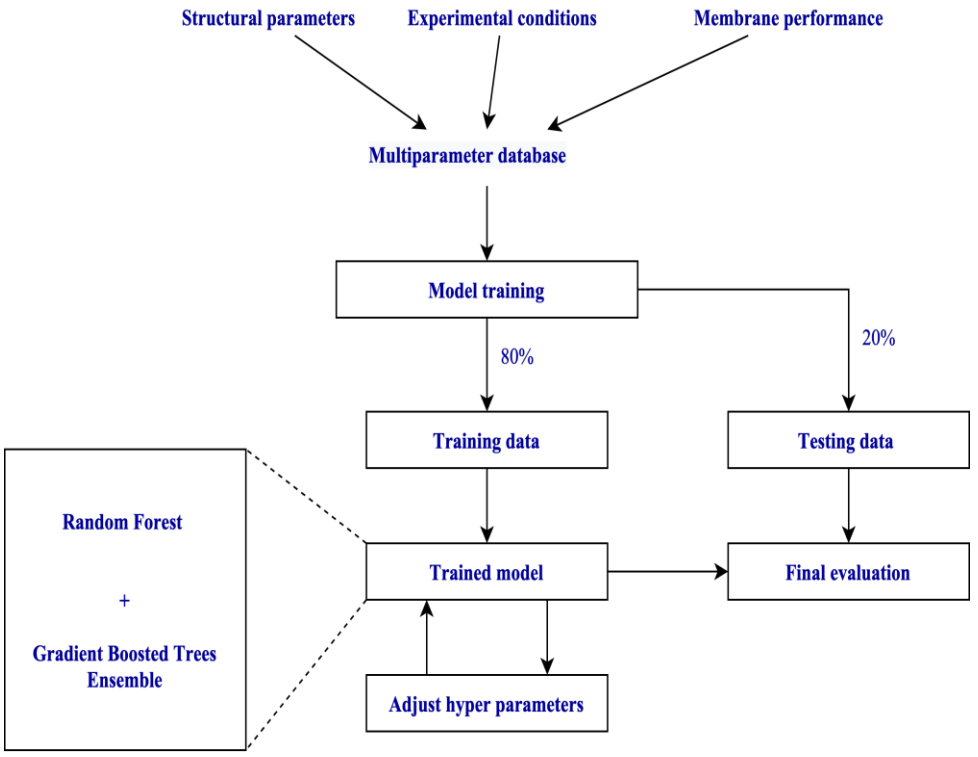

Fig. 1. Workflow of the analysis process

2.3 Partial Dependence Plots (PDP) Analysis

Feature importance assessment quantifies the significance of features in determining the outcomes, but it does not provide information on the specific influence of each feature on the findings. A PDP plot analysis illustrates the marginal impact of features on the ML projected outputs. This study can unveil the interconnection between characteristic parameters and outputs. Two variants of PDP were created - a single parameter PDP and a double factor PDP - to analyze and comprehend the framework's outcomes. The univariate primary influence of pore diameter and zeta potential on various salt rejections was visualized using a single parameter PDP. A contour plot displays the two related parameters' PDP values.

3. Results and Discussions

Following the data collection and statistical analysis, the process split the database into two subsets: training and evaluation sets. The learning set comprised 80% of the information and was employed for training the algorithm, while the remaining 20% was utilized for model verification and evaluation. The research used two standard methods for training and analyzing the databases to exploit the strong abilities of supervised tree ensemble techniques in data regression and essential assessment. These models are the RF approach, which uses bagging, and the XGB approach, which uses gradient boosting. Two methods were

developed using four input factors: pore diameter, zeta capacity, pressure, and concentrations. The output parameters were the rejection efficiency of NaCl, Na2S04, MgClz, and MgSO4. The hyperparameters were continuously adjusted throughout the algorithm's training procedure to improve the approach via iterative calculation and assessment of the fitting level.

Table 1. Salt selectivity analysis

Salt type	Random forest			XGBoost		
	R2	MAE (%)	RMSE (%)	R2	MAE (%)	RMSE (%)
NaCl	0.72	7.32	10.6	0.74	6.12	9.54
Na2SO4	0.84	3.42	6.32	0.85	4.02	6.32
MgCl2	0.79	7.85	12.7	0.85	7.95	13.65
MgSO4	0.71	6.21	8.42	0.94	4.52	5.42

The prediction findings for the testing set indicated that both approaches demonstrated a solid ability to anticipate the four salts' rejection efficiency accurately. R2 was explicitly selected for the study to assess the correlation between tested and forecast values. The findings obtained from the RF and XGB models showed a highly positive correlation between the testing and predicting values. Table 1 demonstrates that the R values for both models consistently exceeded 0.81, indicating a solid fit of linear regression for the assessment set. The proportions of forecasting errors were carefully constrained within a narrow range, further confirming the accuracy of the predictions. When comparing the two approaches, the XGB approach demonstrated improved predictive precision with a more excellent R-squared value for the forecasting outcomes for all four salts. The two tree-ensemble systems demonstrated exceptional predictive capability as a solid basis for the subsequent study of feature decoupling and significance.

Table 2. Salt analysis with different pore radius

Salt type	Pore radius	Random forest			XGBoost		
		R2	MAE (%)	RMSE (%)	R2	MAE (%)	RMSE (%)
NaCl	> 0.5 nm	0.67	5.63	9.32	0.76	3.64	5.32
	<= 0.5 nm	0.84	6.52	9.01	0.93	4.32	6.52
Na2SO4	> 0.5 nm	0.89	5.13	8.53	0.82	7.42	13.21
	<= 0.5 nm	0.73	2.34	2.05	0.78	4.93	3.64
MgCl2	> 0.5 nm	0.89	6.32	9.53	0.85	5.21	7.32
	<= 0.5 nm	0.82	7.83	13.21	0.93	4.96	9.42
MgSO4	> 0.5 nm	0.89	7.42	9.42	0.91	6.32	4.86
	<= 0.5 nm	0.84	2.31	3.06	0.96	1.54	2.31

Upon analyzing the diameters of commonly hydrated ions, the research discovered that the predominant hydrated diameter is focused within the range of 0.5 nanometers. The research partitioned the pore radius into two categories: tiny pores with a radius less than 0.5 nm and

large pores with a radius greater than 0.5 nm. This division allowed to investigate the significance of various characteristics within each range. The analysis focused on correlations within the boundaries of membrane pore diameters less than 0.5 nm. The findings of the model training demonstrated that both algorithms performed exceptionally well in predicting outcomes. The XGB approach outperformed the RF approach in accurately predicting salt rejections, as indicated by its better R2 values (Table 2).

4. Conclusion and Findings

To summarize, the research utilized ML techniques using a multi-parameter database to identify the primary factors determining the selective properties of polyamide NFM towards different salts. Two designs, RF and XGB, were utilized to achieve success in ML. The forecasting accuracy of the two systems was thoroughly compared. The correlation between structural variables and operating circumstances on the salt selectivity of NF membranes was separated, and the significance of distinct features in various pore dimension ranges was thoroughly assessed. The relationship between membrane pore diameter and zeta capability and their impact on water/salt selectivity was emphasized. This study established a correlation between operating conditions and shows membrane structures. It shows the potential of ML in investigating the intricate separation processes of NFM. It provides a technical foundation for the targeted design of NFM for various scenarios.

References

1. Shi, G. M., Feng, Y., Li, B., Tham, H. M., Lai, J. Y., & Chung, T. S. (2021). Recent progress of organic solvent nanofiltration membranes. Progress in Polymer Science, 123, 101470.
2. Jayapriya, R. (2021). Scientometrics Analysis on Water Treatment During 2011 to 2020. Indian Journal of Information Sources and Services, 11(2), 58–63.
3. Feng, X., Peng, D., Zhu, J., Wang, Y., & Zhang, Y. (2022). Recent advances of loose nanofiltration membranes for dye/salt separation. Separation and Purification Technology, 285, 120228.
4. Liloja and Ranjana, P. (2023). An Intrusion Detection System Using a Machine Learning Approach in IOT-based Smart Cities. Journal of Internet Services and Information Security, 13(1), 11-21.
5. Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. Nature reviews Molecular cell biology, 23(1), 40-55.
6. Robles, T., Alcarria, R., De Andrés, D.M., De la Cruz, M.N., Calero, R., Iglesias, S., & Lopez, M. (2015). An IoT based reference architecture for smart water management processes. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 6(1), 4-23.
7. He, S., Li, B., Peng, H., Xin, J., & Zhang, E. (2021). An effective cost-sensitive XGBoost method for malicious URLs detection in imbalanced dataset. IEEE Access, 9, 93089-93096.
8. Premakumari, R. N., et al. "Modeling the dynamics of a marine system using the fractional order approach to assess its susceptibility to global warming." Results in Nonlinear Analysis 7.1 (2024): 89-109.
9. Kamp, J., Emonds, S., Seidenfaden, M., Papenheim, P., Kryschewski, M., Rubner, J., & Wessling, M. (2021). Tuning the excess charge and inverting the salt rejection hierarchy of polyelectrolyte multilayer membranes. Journal of Membrane Science, 639, 119636.

10. Camgözlü, Y., & Kutlu, Y. (2023). Leaf Image Classification Based on Pre-trained Convolutional Neural Network Models. Natural and Engineering Sciences, 8(3), 214-232.
11. Gholizadeh, M., Jamei, M., Ahmadianfar, I., & Pourrajab, R. (2020). Prediction of nanofluids viscosity using random forest (RF) approach. Chemometrics and Intelligent Laboratory Systems, 201, 104010.
12. Sredić, S., Knežević, N., & Milunović, I. (2024). Effects of Landfill Leaches on Ground and Surface Waters: A Case Study of A Wild Landfill in Eastern Bosnia and Herzegovina. Archives for Technical Sciences, 1(30), 97-106.
13. Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. Nature reviews Molecular cell biology, 23(1), 40-55.