Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2022.29110

277 (2022) 30–39
November

Prediction of permeate water flux in forward osmosis desalination system using tree-based ensemble machine learning models

Yinseo Song[a], Jeongwoo Moon[b], Joon Ha Kim[b], Kiho Park[a,*]

[a]*School of Chemical Engineering, Chonnam National University, 77 Yongbong-ro, Buk-gu, Gwangju 61186, Republic of Korea, Tel. +82 (0)62 530 1878; email: kiho138@jnu.ac.kr (K. Park)*
[b]*School of Earth Sciences and Environmental Engineering, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, Republic of Korea*

Received 15 August 2022; Accepted 14 October 2022

ABSTRACT

This study developed tree-based ensemble machine learning models to predict forward osmosis (FO) permeate flux using extreme gradient boosting (XGBoost) and light-gradient boosting machine (LGBM) methods. The models were trained by approximately 700 data points from the FO experimental data. The results showed that LGBM and XGBoost could predict the FO permeate flux with very high accuracy (>0.95 of R^2) in the test set. Feature analysis using Shapley additive explanations values was performed to identify the influences of input variables on the model output and the correlation between the input variables. The results revealed that water permeability and pressure difference have the most significant variables on the FO permeate flux. The correlation between the operating conditions and water permeability cannot be neglected. In this study, we clarified the applicability of ensemble machine learning models for FO systems and suggested directions for future data collection.

Keywords: Forward osmosis; Machine learning models

1. Introduction

Water scarcity is one of the most significant obstacles to maintain a sustainable human society [1]. While the water demand has increased due to climate change caused by global warming, rapid urbanization, and industrialization, the amount of freshwater on Earth cannot easily meet this demand [2,3]. Because 97% of water resources on Earth are seawater, seawater desalination technology is vital for resolving the water shortage problem and securing a sustainable freshwater supply system. Although reverse osmosis (RO) is regarded as an efficient water treatment technology, its high fouling/scaling propensity and high energy consumption due to high-pressure operation still have severe weaknesses [4]. Forward osmosis (FO) has been proposed as an alternative for resolving RO technology's shortcomings. The main characteristic of FO is spontaneous water permeation by a concentration-driven system, not by a pressure-driven system [5]. The osmotic pressure difference between the low-concentration feed solution (FS) and the highly concentrated draw solution (DS) is the main driving force of FO [6]. Therefore, these concentrations should be considered when investigating FO systems.

Water flux and reverse salt flux (RSF) are typically utilized to evaluate the performance of FO systems [7]. High water flux and low RSF are desirable in FO systems. However, the water flux and RSF influence the effective osmotic pressure difference because these fluxes change the concentration near the FO membrane compared with the bulk concentration. This phenomenon, called concentration polarization (CP), should be identified because the main driving force of the FO system might be heavily

[*] Corresponding author.

influenced by the CP [8]. Standard FO membranes have an asymmetric structure that includes a thin active layer and a porous support layer. Thus, there are two kinds of CPs: internal concentration polarization (ICP), which appears in the porous support layer, and external concentration polarization (ECP), which can be found at the surface of the active layer and the support layer [9]. These CPs should be precisely calculated to estimate the water and reverse salt fluxes in the FO system. However, CPs and their corresponding FO performance are affected by various variables, such as the DS and FS concentration, temperature, speed, and flow direction [7]. Because these variables are implicitly correlated, complex relationships between them should be revealed to accurately evaluate the performance of the FO system [1]. Therefore, developing an accurate model and elucidating the correlation between variables is necessary to estimate FO performance [6] accurately.

Recently, predictive models using machine learning and deep learning have focused on various fields, such as medicine [10], gene manipulation [11], chemical reactions [12], and desalination [13]. An artificial neural network (ANN) was used to optimize the FO process based on the water flux and RSF [14]. A simulation approach using an ANN was conducted to analyze the cause of membrane contamination in the RO process and predict membrane fouling [15]. In addition, a recent study developed an algorithm to predict permeate flux using a multi-layered neural network that improved the ANN model [16]. Although many studies have predicted the FO permeate flux using machine learning models such as ANN, the relationship between the membrane parameters and operation variables have not been clearly revealed. Further, state-of-the-art machine learning methods have not been implemented in the flux estimation of FO systems. Therefore, it is necessary to implement state-of-the-art machine learning algorithms to improve the prediction performance of FO fluxes and understand the implicit correlations and sensitivities of the variables based on the prediction results.

In this study, we developed FO water flux prediction models using tree-based machine learning algorithms that typically have a higher regression performance than an ANN model. This study used two different algorithms: extreme gradient boosting (XGBoost) and light-gradient boosting machine (LGBM). A multiple linear regression (MLR) model was also implemented as a base case study to analyze the nonlinear behavior between the variables. Training and testing were conducted using approximately 700 lab-scale FO process data points obtained in previous studies. The optimal hyperparameters in the tree-based machine learning models were determined using a grid-search. A quantitative evaluation was performed to analyze the prediction performance of the models. Using the Shapley additive explanation (SHAP) method, we analyzed the effect of each operation variable on the permeate flux.

2. Methodology

2.1. FO process data collection

Experimental FO process data for training the machine learning models were obtained from a previous study [7]. Usually, the design and existence of a spacer affect FO performance [7]. Thus, FO experimental data were collected only in the case of a flat-sheet membrane module without spacers [17].

The data characteristics were identified and categorized to utilize FO data in the training of machine learning models. The total collected data was 692 with 18 different types of variables (11 numeric, 1 date, and 6 categorical). The data included membrane characteristics, such as the type of active layer, manufacturer, membrane orientation, reported year (partially correlated with the manufactured year), water permeability, and flow direction. The conditions of the solutions were included in the data set, such as the type of solute, concentration, temperature, osmotic pressure, and cross-flow velocity of the FS and DS. The water flux and hydraulic pressure differences across the FO membrane were also included. The water permeate flux was designed as a target variable for model prediction. Further, the reverse salt flux and salt permeability were excluded from the model development because of a lack of sufficient data. A simple imputer replaced partially missing data (FS and DS velocities) with an average value.

For model training, categorical data were converted using the one-hot encoding (OHE). The OHE was expressed as number (digit) 1. The OHE can cause a problem of curse of dimensionality if the categorical data contains many different types. In this study, however, only membrane manufacture has six different types of data, and the other data consist of just two types. Therefore, the problem of OHE can be minimized. Numeric data were preprocessed using logarithmic transformation (np.log1p) in the Python NumPy library to reduce deviations and increase regularity. The preprocessed data were split at an 8:2 ratio to use the training/test data.

2.2. Model

Among the tree-based ensemble machine learning models in Python, the XGBoost and LGBM models were used to predict the permeate flux according to the input variables. MLR was also used to predict FO permeate flux.

2.2.1. Multiple linear regression

Regression analysis is a statistical method used to identify the relationships between variables. The MLR model is an extension of simple regression analysis and a machine learning technique that expresses the linear relationship between two or more independent variables [18]. It has been widely used to predict various system variables, such as the thermal efficiency [19] and compressive strength of recycled aggregate concrete [20].

$$Y = B_0 + B_1X_1 + B_2X_2 + ... + B_nX_n + \varepsilon \quad (1)$$

The dependent variable Y was predicted based on several independent variables X. Parameter B_i denotes a regression coefficient (or weighting factor) weighted on the X value. ε is a residual part that cannot be explained by the MLR, such as noise [21]. Through MLR, the weighting factors and noise were quantified, and the prediction model's performance was evaluated.

2.2.2. XGBoost

XGBoost has been widely used in tree-based ensemble models using gradient boosting, as it has a very high accuracy in predicting system variables [22]. In addition, the XGBoost model applies to small-size training data sets [23,24]. This advantage is helpful because obtaining extensive and sufficient data is difficult. Because approximately 700 data points were used in this study, the XGBoost model is appropriate for use in FO permeate flux prediction.

XGBoost can handle classification and regression tree (CART) problems based on a decision-tree algorithm. This model creates two branches (sets) according to the condition of one variable, up to the maximum depth of the specified tree, as shown in Fig. 1. This approach is expressed in Eq. (2) as follows [22].

$$R_1\left(j,s\right)=\left\{x\middle|x^j\le s\right\}\text{and}R_2\left(j,s\right)=\left\{x\middle|x^j\ge s\right\} \tag{2}$$

where x^j is the observation of the jth feature component corresponding to the training dataset. A decision tree split (j, s) is represented by a splitting feature component j and split point s, where the two leaves are divided.

XGBoost learns by creating a new tree to minimize the residual error, as shown in Fig. 1. After repeating the tree algorithm, the predicted output value is obtained using Eq. (3) [25]:

$$\hat{y}_i=\sum_{k=1}^{k} f_k\left(x_i\right) \tag{3}$$

where $\hat{y}_i$ is the variable used to predict $f_k(x_i)$ K is the number of trees. $f_k(x_i)$ is the result of the kth tree according to x_i. In this study, y_i is the permeate flux.

2.2.3. Light-gradient boosting machine

The LGBM is a gradient-boosting decision tree algorithm developed by Microsoft to obtain a faster learning speed than the other models [26]. XGBoost is a level-wise method that uniformly creates trees on both sides. LGBM creates an asymmetric tree using a leaf-wise method (Fig. 2) [24,27]. In LGBM, a gradient-based one-side sampling (GOSS) method creates a leaf-wise structure [28]. In addition, exclusive feature bundling (EFB), which minimizes the sparsity of the training dataset by bundling exclusive features, was utilized in this study. The LGBM has the advantages of fast learning and fewer memory requirements during learning. Thus, the LGBM is recommended for developing a machine learning model when the dataset size is large. However, it is more prone to overfitting problems than other tree-based ensemble algorithms when the data size is small.

In this study, the permeate flux was predicted by increasing the available number of datasets using k-fold cross-validation to avoid overfitting. In addition, hyperparameters such as max_depth, min_sample_split, and lambda were adjusted using grid-search. Finally, the n_estimator is tuned

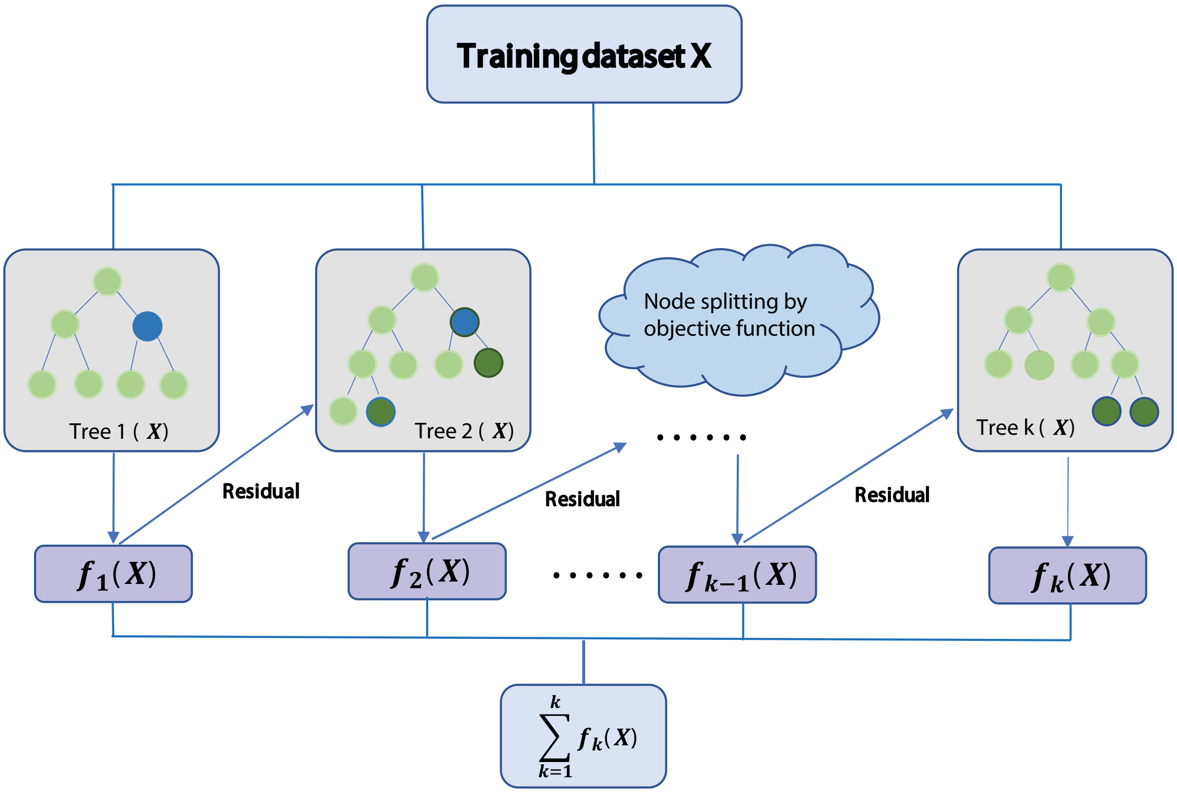


Fig. 1. Schematic diagram of the XGBoost model.

to a significant value with a small learning rate to find the optimal points of the hyperparameters.

2.2.4. Model validation

We used the coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE) to verify whether the machine learning algorithms used in this study (MLR, XGBosst, and LGBM) were well-trained. The test data, divided from the original data before training, were used to compare the predicted and actual values. Eqs. (4)–(7) are used for each verification method.

$$R^2 = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)^2 - \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)^2}} \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{Y}_i - Y_i\right| \tag{5}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2} \tag{6}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{7}$$

R^2 is an indicator of how well the statistical model explains the data. The closer the number is to 1, the better the model performance [29]. The other indicators (MAE, RMSE, and MSE) were correlated with the magnitude of the errors between the data and model predictions. Thus, if these indicators are close to zero, the regression model exhibits high prediction performance. MAE considers the absolute value of the error between the predicted and actual values. Although relatively straightforward, an outlier in the prediction model cannot be easily found. Unlike MAE, MSE and RMSE utilize squared error values instead of absolute values. Thus, outliers can be easily found because the errors between the outliers and actual values are amplified. As the squared treatment of the errors amplifies the overall values of the summation of errors, RMSE converts the size of the error value to be similar to the actual value by rooting the MSE for intuitive analysis [26].

2.2.5. Cross-validation and finding optimal hyperparameters

The k-fold cross-validation method was used to supplement the insufficient dataset, improve accuracy, and prevent overfitting. As shown in Fig. 3, the entire dataset was divided into k groups of equal sizes. One of the divided groups was used as a validation set, and the remaining groups (the size is (k–1)/k to the total dataset) were used as a training set. This procedure is repeated k times. In other words, the MSE values of k different models were averaged to obtain CV(k), the final test error [30]. In this study, k is set to 5.

$$\text{CV}\left(k\right) = \frac{1}{k}\sum_{i=1}^{k}\text{MSE}_i \tag{8}$$

The tree-based ensemble machine learning models used a grid-search method to determine the best hyperparameters. The grid-search method was implemented in the Python scikit-learn library. This function was designed to find the best hyperparameter set from combinations of the potential candidates for the hyperparameters [31].

This study used a built-in function, 'GridSearchCV', which combines grid-search and k-fold cross-validation. The 'GridSearchCV' provides the best score which represents how much the model fits well to the training data set. The range of the best score is [0,1], and the model prediction becomes perfect if the best score approaches 1. The algorithm provided optimal hyperparameters such as the learning rate, number of trees, sampling, tree depth, and regularization coefficients (L1, L2), essential parameters for the XGBoost and LGBM models.

3. Results and discussion

3.1. Prediction results using MLR model

Although the MLR model provides good performance for prediction, any complexity and nonlinearity in the data set worsen the model prediction performance because the MLR model is inherently a linear regression [19,21]. The MLR regression model was used to compare with the tree-based ensemble models and identify the relationship between the input variables and target variable (permeate flux).

The MLR prediction performance was evaluated using R^2, MAE, MSE, and RMSE. Comparing the predicted results using the test data with the actual values, MAE was 3.5784, MSE was 32.9855, and RMSE was 5.7433. As shown in Fig. 4, the coefficient of determination (R^2) is 0.8252 for the training dataset and 0.7624 for the test data-set. The FO process dataset included many nonlinear correlations between variables. Thus, the prediction performance for estimating permeate flux was poor because of the inherent characteristics of the linear regression model, as discussed in the previous paragraph. In addition, the highly scattered

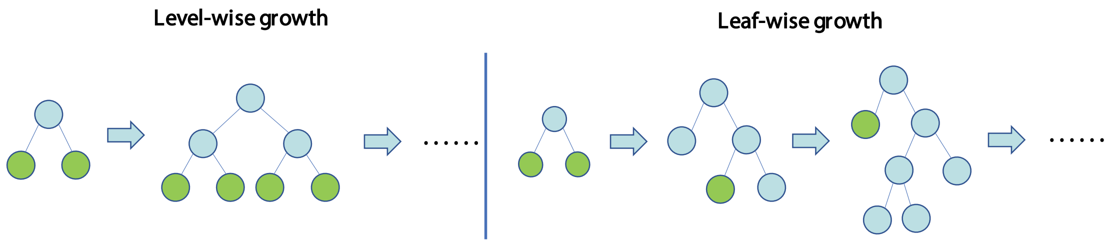


Fig. 2. Schematic diagram of level-wise growth (XGBoost) and leaf-wise growth (LGBM).

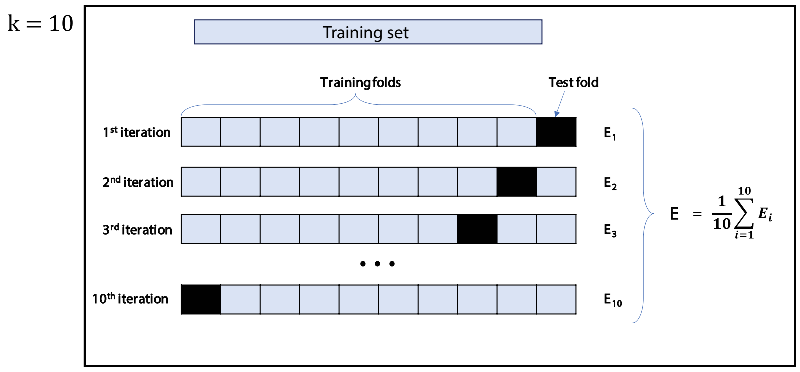


Fig. 3. Mechanism of *k*-fold cross-validation.

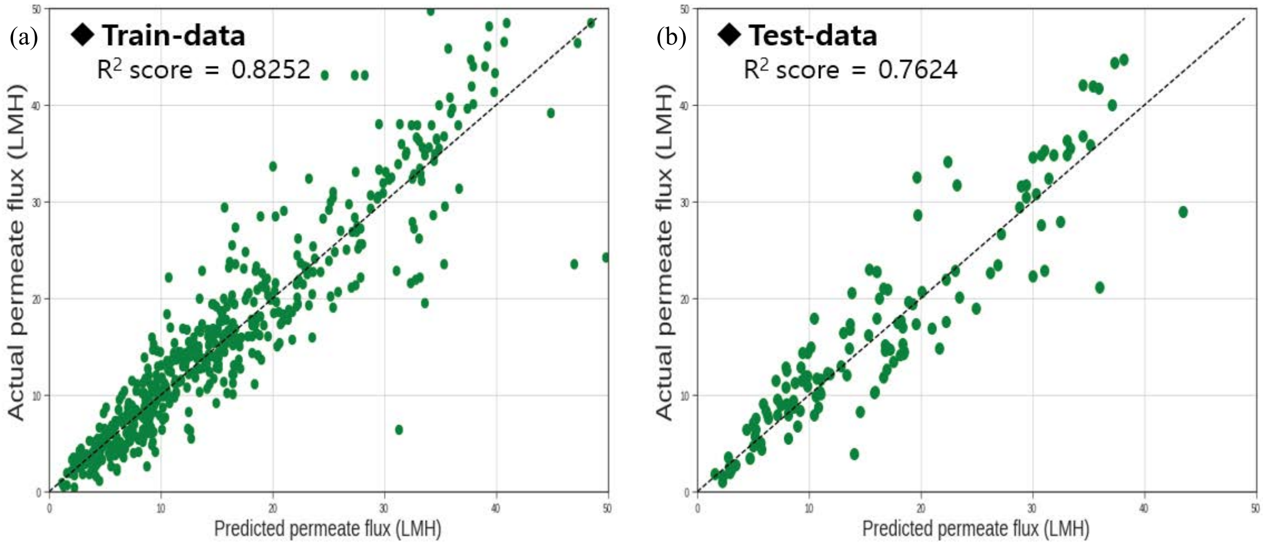


Fig. 4. Predicted permeate flux vs. actual permeate flux with R^2 by MLR model (a) train data and (b) test data.

data points in Fig. 4 indicate that MLR cannot be successfully utilized as a prediction model for the FO system.

3.2. Prediction results using the XGBoost model

The XGBoost model provides a good prediction performance by reducing the residuals of the predicted and actual values and can also quantify the importance of relationships between the input variable and the output variable through feature importance analysis. To obtain the highest performance of the XGBoost model, hyperparameter optimization should be conducted. The overfitting problem can be controlled by specifying regularization coefficients such as λ (L2 regularization) and α (L1 regularization). The grid-search algorithm obtained hyperparameters for the FO permeate flux prediction model, as listed in Table 1.

The model with the optimal parameters obtained by grid-search was evaluated using *k*-fold cross-validation. The GridSearchCV result was 0.9669, indicating that the model was well-trained without overfitting.

The performances of the XGBoost model for FO permeate flux prediction were 1.2768, 5.618, and 2.3702 for MAE, MSE, and RMSE, respectively. The R^2 was 0.9873 for the training dataset and 0.9544 for the test data-set, as shown in Fig. 5. The test data showed a higher R^2 value than the MLR model, and the scattering of the data points was narrower. The XGBoost model has been widely used because of its high prediction performance and because the structure of the model can express nonlinear characteristics. Therefore,

Table 1
Hyperparameters of XGBoost model optimized using grid-search

Hyperparameters	Values	Ranges
n_estimators	1,000	[100, 1000]
max_depth	3	[1, 5]
Learning rate	0.01	[0.01, 0.001]
min_sample_split	1	[1, 5]
γ	0	[0, 1]
α	0.6	[0, 2]
β	0.8	[0, 1]

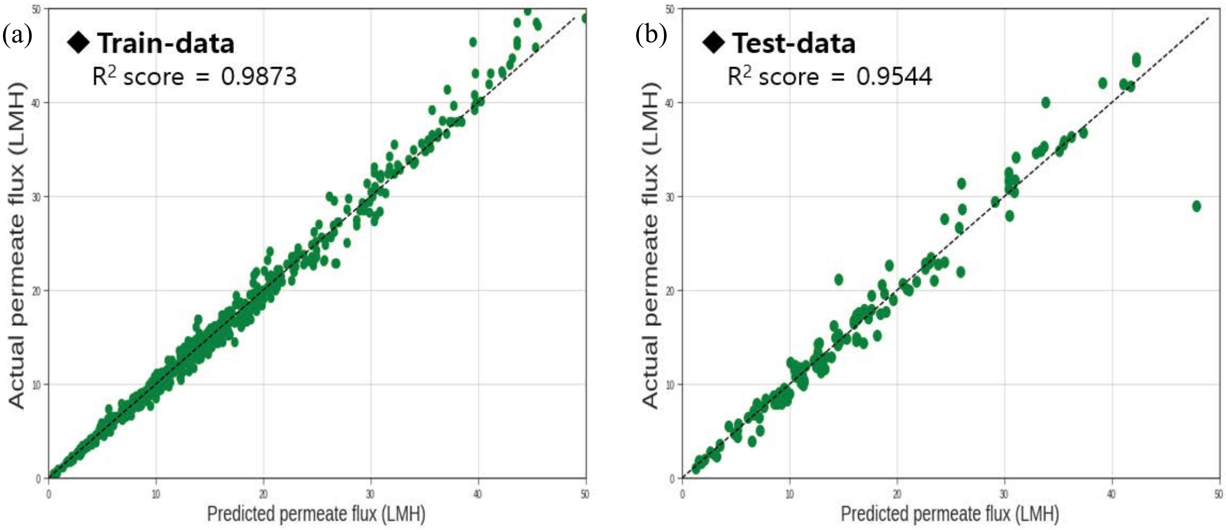


Fig. 5. Predicted permeate flux vs. actual permeate flux with R^2 by XGBoost model (a) train data and (b) test data.

the FO permeate flux prediction performance can be achieved using the XGBoost model, despite the dispersed FO experimental data obtained from various studies.

3.3. Prediction results using the LGBM model

The LGBM model was built based on many CARTs that won the Kaggle data analysis competition. The main advantage of the LGBM is that it requires a relatively shorter training time. In this study, the training time of the LGBM for FO permeate flux prediction was five times faster than that of the XGBoost model.

Because the LGBM uses the GOSS algorithm to construct a leaf-wise structure for fast and high-precision results, overfitting should be carefully checked if the dataset size is insufficient. The hyperparameter setting is significant in the LGBM model and the XGBoost model. Table 2 shows the optimal hyperparameter set used in the LGBM model in this study. The LGBM model with optimal hyperparameters was verified using the GridSearchCV function. The highest accuracy point of LGBM was 0.9682, which was slightly higher than that of XGBoost (0.9669). Hence, it was confirmed that the insufficient dataset was supplemented by *k*-fold cross-validation.

The performance of the LGBM for FO permeate flux prediction was 1.1184 for MAE, 4.8540 for MSE, and 2.2032 for RMSE. The R^2 values for the training and test data-sets were 0.9884 and 0.9606, respectively (Fig. 6). The prediction performance of the LGBM model was higher than that of MLR and XGBoost. Although the prediction performances of LGBM and XGBoost are comparable, the fast training speed of the LGBM is a significant advantage of the machine learning model. Therefore, the LGBM model is the most appropriate method for predicting the FO permeate flux.

3.4. Model comparisons and feature importance

To display the comparison results, the FO prediction performance of each model with the test data-set is summarized in Table 3. The LGBM showed the highest prediction performance, and MLR showed the lowest prediction performance among all indicators. R^2 > 0.95, using the model with good prediction performance is desirable. In addition, the learning times of the XGBoost model and the LGBM model are 0.251s and 0.034s, respectively. The LGBM model is 7 times faster than the XGBoost. It can be concluded that tree-based ensemble models using gradient boosting are suitable for learning FO process data and exhibit good performance.

Table 2
Hyperparameters of LGBM models optimized using grid-search

Hyperparameters	Values	Ranges
n_estimators	1,000	[100, 1000]
max_depth	3	[1, 5]
Learning rate	0.01	[0.01, 0.001]
min_sample_split	1	[1, 5]
γ	0	[0, 1]
α	1.0	[0, 2]
β	0.4	[0, 1]

In this study, we used SHAP values to describe the feature importance of the input variables to the FO permeate flux without the problems of inconsistency and high cardinality. In addition, a feature importance analysis was performed to identify the extent of its influence on the target variable. The feature importance of each input variable was obtained by calculating the average impurity reduction in the developed tree for each characteristic.

3.4.1. SHAP values using the LGBM model

Fig. 7a and b show the feature importance and summary plot of the SHAP values of the input variables using the LGBM model. As shown in Fig. 7a, water permeability

Table 3
Comparison of the FO permeate flux prediction performance by MLR, XGBoost, and LGBM models using test data

Models	MAE	MSE	RMSE	R^2 score
MLR	3.4833	29.3667	5.4191	0.8107
XGBoost	1.2768	5.6181	2.3702	0.9544
LGBM	1.1184	4.8540	2.2032	0.9606

has the highest SHAP value (+8.9), followed by the pressure difference (SHAP value = +6.12). Water permeability is directly correlated with the FO permeate flux, as expressed in a flux equation [32–34]. Thus, the highest SHAP value for water permeability is reasonable. Because FO is a concentration-driven water permeation system, DS osmotic pressure's relatively high SHAP value is acceptable. Although the high SHAP value of the pressure difference appears unusual, the pressure difference also affects the permeate flux directly, even in the FO system. Pressure-assisted FO is a relevant example of the significant influence of the pressure difference on the permeate flux [35]. The pressure difference affects the permeate flux more directly than the osmotic pressure difference because the CPs reduce the effective osmotic pressure difference. This might be the main reason for the higher SHAP value of the pressure difference than that of the DS osmotic pressure. The other input variables had similar impacts on the FO permeate flux. Although the effects of the other input variables were relatively minor, these factors should be carefully considered to improve the FO permeate flux prediction of the LGBM model.

Fig. 7b shows the SHAP summary plot of the impact of the features on the model output. Red points show high feature values, and blue points represent low values. Similar conclusions were drawn from the SHAP summary chart. The higher the values of the water permeability and pressure difference applied in the FO system, the larger the water permeate flux. Through Fig. 7b, the feature impact can be quantified comprehensively.

In addition, the correlation between the input variables should also be checked. Fig. 8 shows the heatmap results for analyzing the correlation between the input variables [36]. Water permeability has the highest positive correlation (0.602) with the target variable (FO permeate flux), followed by the pressure difference (0.189). Although most of the input variables have a low correlation, some of the inputs are significantly correlated owing to their inherent characteristics. For example, osmotic pressure and molarity are directly correlated, and the correlation values are higher than 0.9. The correlation between the input variables and water permeability was quite large. The correlation values of the DS osmotic pressure and pressure difference with water permeability are higher than 0.46. As suggested in a previous study, water permeability can vary depending on operating conditions [7]. Therefore, the possibility of variable water permeability under different operating conditions was investigated in this study.

However, some of the input variables are unnecessarily correlated owing to the conditions of the experimental data. The correlation between FS and DS temperatures was very high because most of the experimental data were obtained at the same temperature. Because the effects of different temperatures on FS and DS solutions have been reported, data treatment should consider the effects of these limitations [37]. In addition, other variables that might influence FO permeate flux should be included as much as possible during model development. Further considerations will improve the performance of machine-learning models in FO systems and expand their applicability to various membrane separation systems.

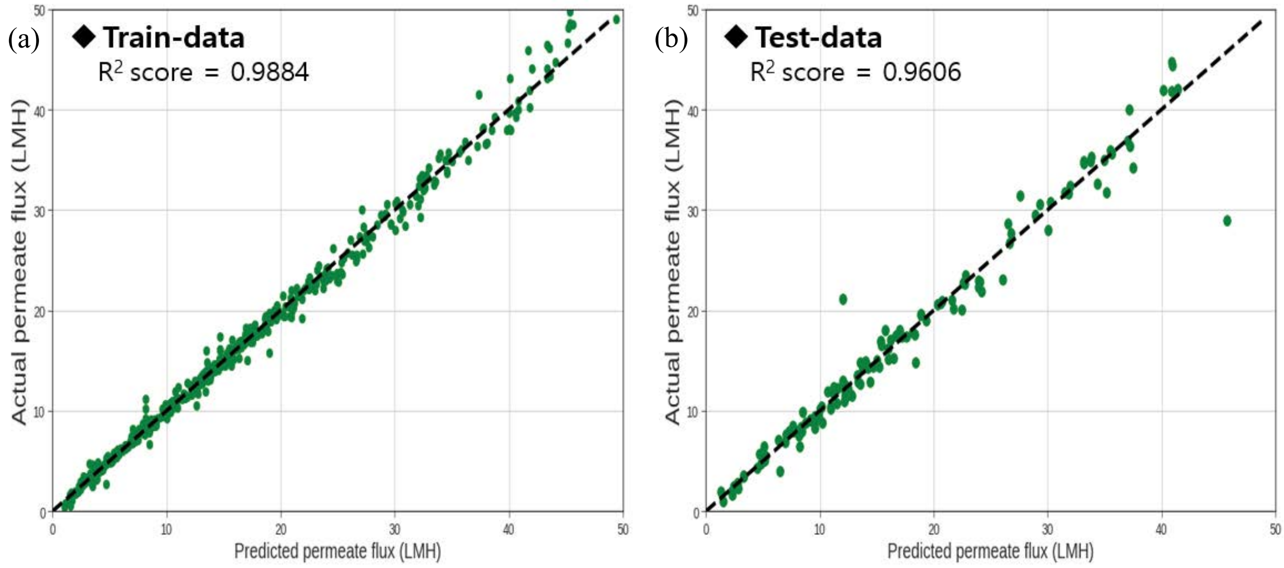


Fig. 6. Predicted permeate flux vs. actual permeate flux with R^2 by LGBM model (a) train data and (b) test data.

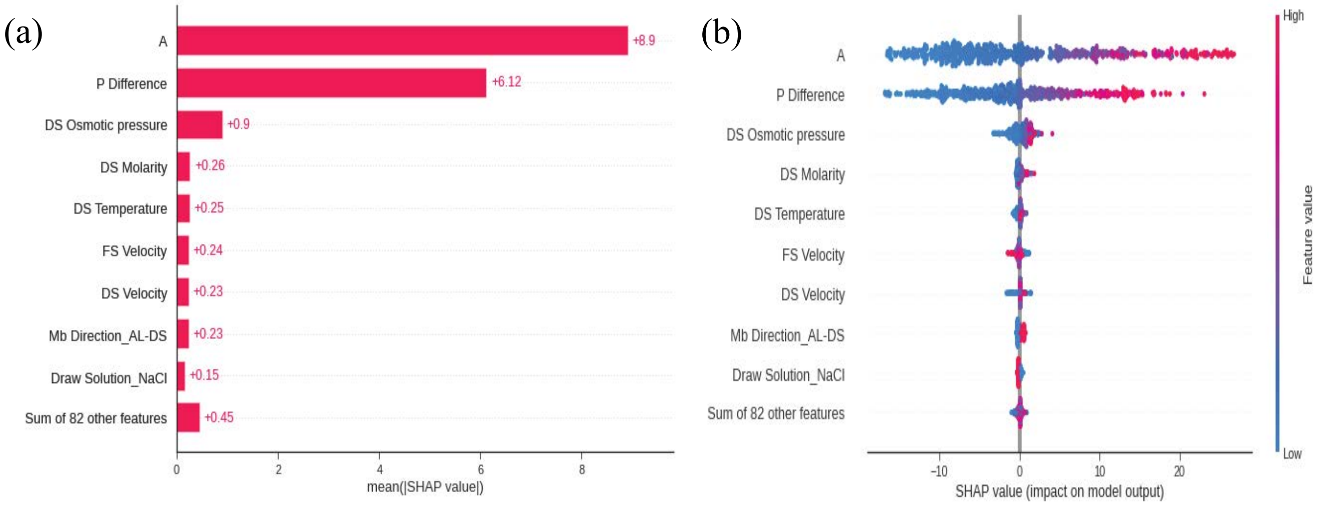
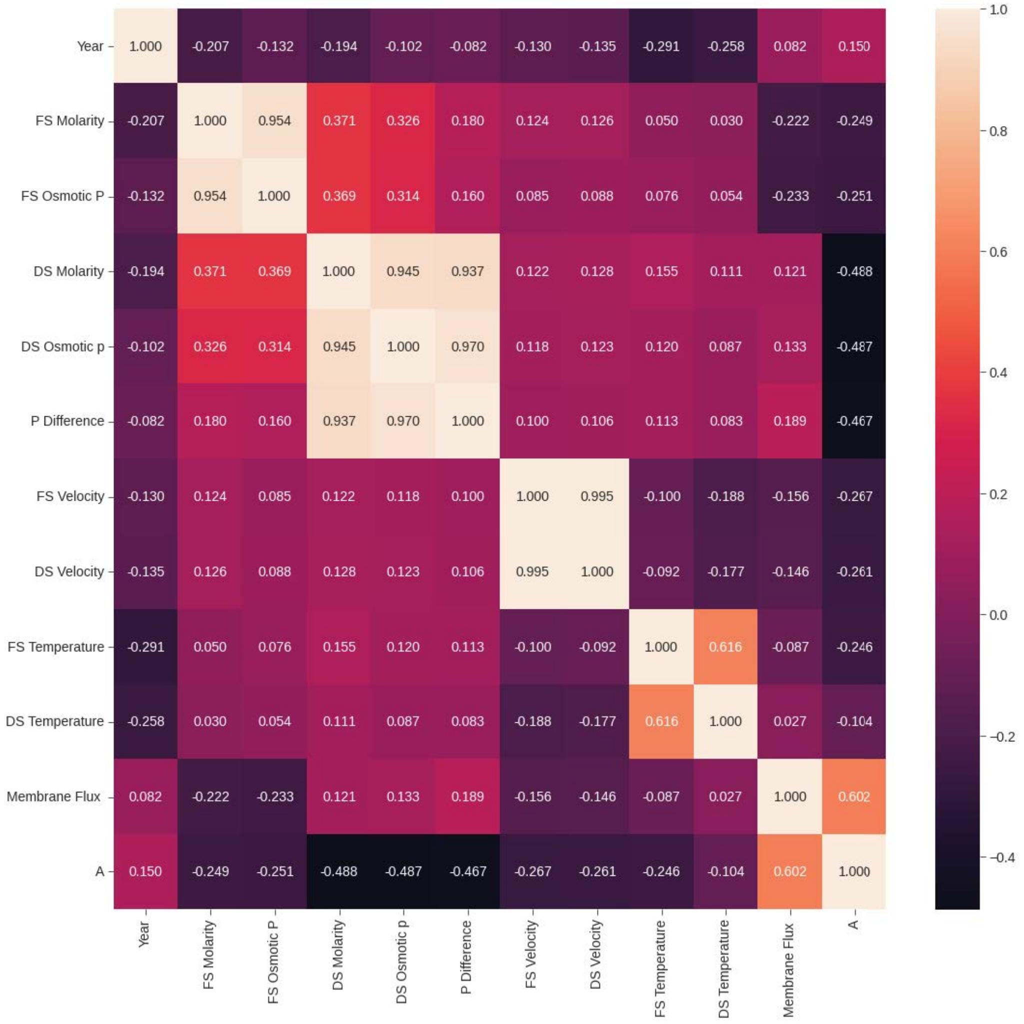


Fig. 7. SHAP values of the input variables using the LGBM model. (a) SHAP feature importance and (b) SHAP summary plot.



	Year	FS Molarity	FS Osmotic P	DS Molarity	DS Osmotic p	P Difference	FS Velocity	DS Velocity	FS Temperature	DS Temperature	Membrane Flux	A
Year	1.000	-0.207	-0.132	-0.194	-0.102	-0.082	-0.130	-0.135	-0.291	-0.258	0.082	0.150
FS Molarity	-0.207	1.000	0.954	0.371	0.326	0.180	0.124	0.126	0.050	0.030	-0.222	-0.249
FS Osmotic P	-0.132	0.954	1.000	0.369	0.314	0.160	0.085	0.088	0.076	0.054	-0.233	-0.251
DS Molarity	-0.194	0.371	0.369	1.000	0.945	0.937	0.122	0.128	0.155	0.111	0.121	-0.488
DS Osmotic p	-0.102	0.326	0.314	0.945	1.000	0.970	0.118	0.123	0.120	0.087	0.133	-0.487
P Difference	-0.082	0.180	0.160	0.937	0.970	1.000	0.100	0.106	0.113	0.083	0.189	-0.467
FS Velocity	-0.130	0.124	0.085	0.122	0.118	0.100	1.000	0.995	-0.100	-0.188	-0.156	-0.267
DS Velocity	-0.135	0.126	0.088	0.128	0.123	0.106	0.995	1.000	-0.092	-0.177	-0.146	-0.261
FS Temperature	-0.291	0.050	0.076	0.155	0.120	0.113	-0.100	-0.092	1.000	0.616	-0.087	-0.246
DS Temperature	-0.258	0.030	0.054	0.111	0.087	0.083	-0.188	-0.177	0.616	1.000	0.027	-0.104
Membrane Flux	0.082	-0.222	-0.233	0.121	0.133	0.189	-0.156	-0.146	-0.087	0.027	1.000	0.602
A	0.150	-0.249	-0.251	-0.488	-0.487	-0.467	-0.267	-0.261	-0.246	-0.104	0.602	1.000

Fig. 8. Heatmap results of the input variables using the LGBM model. The unit of input variables are Year (year), FS molarity (M), FS osmotic pressure (atm), DS molarity (M), DS osmotic pressure (atm), P difference (atm), FS velocity (cm/s), DS velocity (cm/s), FS temperature (°C), DS temperature (°C), and A (LMH/atm).

4. Conclusions

This study predicted the FO permeate flux using tree-based ensemble machine learning models. Two tree-based ensemble models (XGBoost and LGBM) and MLR models were used for FO permeate flux prediction, and a comparative analysis of flux prediction performance was conducted. The experimental data (approximately 700 points) were obtained from the literature on FO systems using a flat-sheet module. The machine learning models were trained using the dataset, and overfitting problems were minimized using *k*-fold cross-validation.

The results showed that the LGBM model had a higher performance of FO permeate flux prediction ($R^2 = 0.9606$ by test set). XGBoost was comparable to LGBM ($R^2 = 0.9544$ by test set). However, the longer training time makes the LGBM model more useful. Both tree-based ensemble models have better performance than MLR ($R^2 = 0.8107$), revealing that the nonlinear characteristics of the FO permeate flux cannot be neglected. In addition, the feature importance of the input variables was investigated, and the correlation between the input variables was quantified. The results showed that the water permeability and pressure difference significantly influenced the FO permeate flux. Most input variables were not highly correlated. However, the correlations in some input variables were unreasonable due to the skewed conditions of the experimental data. We revealed the current potential of tree-based ensemble machine learning models for FO permeate flux prediction and clarified the limitations of the dataset. Based on the conclusions of this study, we suggest future directions for FO data collection.

Acknowledgments

This study was financially supported by Korea Environment Industry & Technology Institute (KEITI) through Development of UltraPure Water Production Process Technology for Localization Project, funded by Korea Ministry of Environment (MOE) (Grant number: 2021003210006), and also funded by Chonnam National University (Grant number: 2022–2651).

References

[1] R. Colciaghi, R. Simonetti, L. Molinaroli, M. Binotti, G. Manzolini, Potentialities of thermal responsive polymer in forward osmosis (FO) process for water desalination, Desalination, 519 (2021) 115311, doi: 10.1016/j.desal.2021.115311.

[2] D.J. Johnson, W.A. Suwaileh, A.W. Mohammed, N. Hilal, Osmotic's potential: an overview of draw solutes for forward osmosis, Desalination, 434 (2018) 100–120.

[3] N. Ghaffour, T.M. Missimer, G.L. Amy, Technical review and evaluation of the economics of water desalination: current and future challenges for better water supply sustainability, Desalination, 309 (2013) 197–207.

[4] K. Park, J. Kim, D.R. Yang, S. Hong, Towards a low-energy seawater reverse osmosis desalination plant: a review and theoretical analysis for future directions, J. Membr. Sci., 595 (2020) 117607, doi: 10.1016/j.memsci.2019.117607.

[5] K. Park, Y.H. Jang, J.W. Chang, D.R. Yang, Membrane transport behavior characterization method with constant water flux in pressure-assisted forward osmosis, Desalination, 498 (2021) 114738, doi: 10.1016/j.desal.2020.114738.

[6] T.-S. Chung, S. Zhang, K.Y. Wang, J. Su, M.M. Ling, Forward osmosis processes: yesterday, today and tomorrow, Desalination, 287 (2012) 78–81.

[7] M.-k. Kim, J.W. Chang, K. Park, D.R. Yang, Comprehensive assessment of the effects of operating conditions on membrane intrinsic parameters of forward osmosis (FO) based on principal component analysis (PCA), J. Membr. Sci., 641 (2022) 119909, doi: 10.1016/j.memsci.2021.119909.

[8] G.T. Gray, J.R. McCutcheon, M. Elimelech, Internal concentration polarization in forward osmosis: role of membrane orientation, Desalination, 197 (2006) 1–8.

[9] B. Kim, G. Gwak, S. Hong, Analysis of enhancing water flux and reducing reverse solute flux in pressure assisted forward osmosis process, Desalination, 421 (2017) 61–71.

[10] A. Rajkomar, J. Dean, I. Kohane, Machine learning in medicine, New. Eng. J. Med., 380 (2019) 1347–1358.

[11] M.J. Volk, I. Lourentzou, S. Mishra, L.T. Vo, C. Zhai, H. Zhao, Biosystems design by machine learning, ACS Synth. Biol., 9 (2020) 1514–1533.

[12] S. Stocker, G. Csányi, K. Reuter, J.T. Margraf, Machine learning in chemical reaction space, Nat. Commun., 11 (2020) 1–11.

[13] G.N. Marichal Plasencia, J. Camacho-Espino, D. Ávila Prats, B. Peñate Suárez, Machine learning models applied to manage the operation of a simple SWRO desalination plant and its application in marine vessels, Water, 13 (2021) 2547, doi: 10.3390/w13182547.

[14] K. Aghilesh, A. Mungray, S. Agarwal, J. Ali, M.C. Garg, Performance optimisation of forward-osmosis membrane system using machine learning for the treatment of textile industry wastewater, J. Cleaner Prod., 289 (2021) 125690, doi: 10.1016/j.jclepro.2020.125690.

[15] E.A. Roehl Jr., D.A. Ladner, R.C. Daamen, J.B. Cook, J. Safarik, D.W. Phipps Jr., P. Xie, Modeling fouling in a large RO system with artificial neural networks, J. Membr. Sci., 552 (2018) 95–106.

[16] J. Jawad, A.H. Hawari, S. Zaidi, Modeling of forward osmosis process using artificial neural networks (ANN) to predict the permeate flux, Desalination, 484 (2020) 114427, doi: 10.1016/j.desal.2020.114427.

[17] Y. Xu, X. Peng, C.Y. Tang, Q.S. Fu, S. Nie, Effect of draw solution concentration and operating conditions on forward osmosis and pressure retarded osmosis performance in a spiral wound module, J. Membr. Sci., 348 (2010) 298–309.

[18] S. Sousa, F.G. Martins, M.C. Alvim-Ferraz, M.C. Pereira, Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations, Environ. Modell. Software, 22 (2007) 97–103.

[19] A.F. Mashaly, A. Alazba, MLP and MLR models for instantaneous thermal efficiency prediction of solar still under hyper-arid environment, Comput. Electron. Agric., 122 (2016) 146–155.

[20] F. Khademi, S.M. Jamal, N. Deshpande, S. Londhe, Predicting strength of recycled aggregate concrete using artificial neural network, adaptive neuro-fuzzy inference system and multiple linear regression, Int. J. Sustainable Built Environ., 5 (2016) 355–369.

[21] G.K. Uyanık, N. Güler, A study on multiple linear regression analysis, Procedia Soc. Behav. Sci., 106 (2013) 234–240.

[22] W. Dong, Y. Huang, B. Lehane, G. Ma, XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring, Autom. Constr., 114 (2020) 103155, doi: 10.1016/j.autcon.2020.103155.

[23] Y. Liang, J. Wu, W. Wang, Y. Cao, B. Zhong, Z. Chen, Z. Li, Product Marketing Prediction Based on XGBoost and LightGBM Algorithm, Proceedings of the 2nd International Conference on Artificial Intelligence and Pattern Recognition, Association for Computing Machinery, New York, NY, United States, 2019, pp. 150–153, doi: 10.1145/3357254.3357290.

[24] X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, X. Niu, Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGBoost algorithms according to different high dimensional data cleaning, Electron. Commer. Res. Appl., 31 (2018) 24–39.

[25] J. Brownlee, XGBoost With Python: Gradient boosted Trees with XGBoost and Scikit-Learn, Machine Learning Mastery, 2016.

[26] A. Shehadeh, O. Alshboul, R.E. Al Mamlook, O. Hamedat, Machine learning models for predicting the residual value of heavy construction equipment: an evaluation of modified decision tree, LightGBM, and XGBoost regression, Autom. Constr., 129 (2021) 103827, doi: 10.1016/j.autcon.2021.103827.

[27] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017.

[28] M.R. Machado, S. Karray, I.T. de Sousa, LightGBM: An Effective Decision Tree Gradient Boosting Method to Predict Customer Loyalty in the Finance Industry, 2019 14th International Conference on Computer Science & Education (ICCSE), IEEE, Toronto, ON, Canada, 2019, pp. 1111–1116.

[29] H. Gholami, A. Mohamadifar, A. Sorooshian, J.D. Jansen, Machine-learning algorithms for predicting land susceptibility to dust emissions: the case of the Jazmurian Basin, Iran, Atmos. Pollut. Res., 11 (2020) 1303–1315.

[30] F. Arcadu, F. Benmansour, A. Maunz, J. Willis, Z. Haskova, M. Prunotto, Deep learning algorithm predicts diabetic retinopathy progression in individual patients, npj Digit. Med., 2 (2019) 92, doi: 10.1038/s41746-019-0172-3.

[31] H.A. Fayed, A.F. Atiya, Speed up grid-search for parameter selection of support vector machines, Appl. Soft Comput., 80 (2019) 202–210.

[32] K. Park, H. Heo, D.Y. Kim, D.R. Yang, Feasibility study of a forward osmosis/crystallization/reverse osmosis hybrid process with high-temperature operation: modeling, experiments, and energy consumption, J. Membr. Sci., 555 (2018) 206–219.

[33] K. Park, Y.H. Jang, M.-g. Kim, D.R. Yang, S. Hong, Comprehensive analysis of a hybrid FO/crystallization/RO process for improving its economic feasibility to seawater desalination, Water Res., 171 (2020) 115426, doi: 10.1016/j.watres.2019.115426.

[34] W. Suwaileh, N. Pathak, H. Shon, N. Hilal, Forward osmosis membranes and processes: a comprehensive review of research trends and future outlook, Desalination, 485 (2020) 114455, doi: 10.1016/j.desal.2020.114455.

[35] T. Yun, Y.-J. Kim, S. Lee, S. Hong, G.I. Kim, Flux behavior and membrane fouling in pressure-assisted forward osmosis, Desal. Water Treat., 52 (2014) 564–569.

[36] M. Tang, Q. Zhao, S.X. Ding, H. Wu, L. Li, W. Long, B. Huang, An improved LightGBM algorithm for online fault detection of wind turbine gearboxes, Energies, 13 (2020) 807, doi: 10.3390/en13040807.

[37] M. Xie, W.E. Price, LD. Nghiem, M. Elimelech, Effects of feed and draw solution temperature and transmembrane temperature difference on the rejection of trace organic contaminants by forward osmosis, J. Membr. Sci., 438 (2013) 57–64.