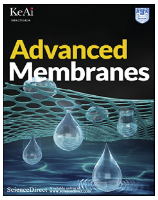
Advanced Membranes 3 (2023) 100072

Contents lists available at ScienceDirect

Advanced Membranes

journal homepage: www.keaipublishing.com/en/journals/advanced-membranes



Machine learning toward improving the performance of membrane-based wastewater treatment: A review

Panchan Dansawad [a,b], Yanxiang Li [a,b], Yize Li [c], Jingjie Zhang [d], Siming You [c,**], Wangliang Li [a,b,*], Shouliang Yi [e,***]

[a] CAS Key Laboratory of Green Process and Engineering, Institute of Process Engineering, Chinese Academy of Sciences, Beijing, 100190, China
[b] University of Chinese Academy of Sciences, Beijing, 100049, China
[c] James Watt School of Engineering, University of Glasgow, Glasgow, G12 8QQ, UK
[d] NUS-Environmental Research Institute, National University of Singapore, 5A Engineering Drive 1, 117411, Singapore
[e] U. S. Department of Energy National Energy Technology Laboratory, Pittsburgh, PA, 15236, USA

ARTICLE INFO

Keywords:
Machine learning
Algorithm
Artificial neural network
Wastewater treatment
Membrane separation

ABSTRACT

Machine learning (ML) is a data-driven approach that can be applied to design, analyze, predict, and optimize a process based on existing data. Recently, ML has found its application in improving membrane separation performance for wastewater treatment. Models have been developed to predict the performance of membranes to separate contaminants from wastewater, design optimum conditions for membrane fabrication for greater membrane separation performance and predict backwashing membranes and membrane fouling. This review summarizes the progress of ML-based membrane separation modeling and explores the direction of the future development of ML in membrane separation-based wastewater treatment. The strengths and drawbacks of the ML algorithms extensively used in membrane separation-based wastewater treatment are summarized. Artificial neural network (ANN) was the most used algorithm for modeling membrane separation-based wastewater treatment. Future research is recommended to focus on the development of integrated ML algorithms and on combining ML algorithms with other modeling approaches (*e.g.*, process-based models and statistical models). This will serve to achieve higher accuracy and better performance of the ML application.

1. Introduction

Big data and data analytics are receiving increasing attention due to the potential advantages of improved data management, analysis, and creation of prediction models from large volumes of data for wide applications [1,2]. Recently, relevant approaches have been applied to model and optimize wastewater treatment processes, with machine learning (ML)-based modeling being one of the most popular choices. ML is one typical type of data-driven modeling approach, and it starts with training an algorithm with a dataset to explain the phenomena of the data [3]. ML has been applied to predict relevant phenomena or processes for wastewater treatment modeling, featured by high efficiency and reasonable accuracy [4,5].

Conventional approaches for wastewater treatment modeling often rely on process-based models that involve combinations of process-based mathematical equations [6]. Compared with ML, the process-based model's disadvantages include relatively low efficiency, low accuracy, and time-consuming [7–9]. Process-based models assume that predictors or input features are known, and models are parametric. In comparison, the ML model is usually based on a non-parametric model in which the structure is not specified. ML modeling does not need assumptions about distributions or linearity [10,11]. ML aims to achieve prediction using learning algorithms to find patterns within a given dataset without relying on previous understanding of underlying structures [12,13].

Wastewater can contain various contaminants, such as pharmaceutical compounds, heavy metals, oil-water emulsions, microorganisms, disinfection byproducts, and pesticides [14]. Some contaminants are poorly degraded and can remain dissolved in water for long periods,

* Corresponding author. CAS Key Laboratory of Green Process and Engineering, Institute of Process Engineering, Chinese Academy of Sciences, Beijing, 100190, China.
** Corresponding author.
*** Corresponding author.
E-mail addresses: Siming.You@glasgow.ac.uk (S. You), wlli@ipe.ac.cn (W. Li), shouliang.yi@hotmail.com (S. Yi).

https://doi.org/10.1016/j.advmem.2023.100072
Received 1 September 2023; Received in revised form 3 October 2023; Accepted 7 October 2023
Available online xxxx

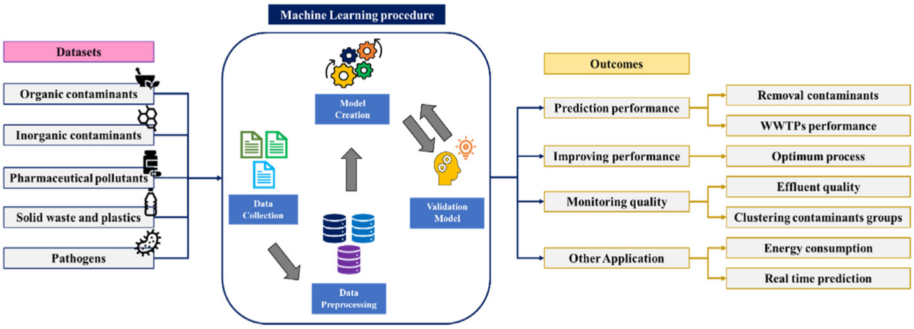


Fig. 1. Schematic diagram of the procedure of ML for wastewater treatment.

posing long-term environmental and health risks [15]. Membrane technologies are one kind of wastewater treatment technology that has developed significantly over the past two decades. Due to their advantages in wastewater treatment, such as small equipment size, decreased energy demand, reduced cost of operation, and higher efficiency, membrane technologies have been applied in various ways, such as filtration, desalination, coagulation-flocculation, maturation lagoons, and membrane biological reactors [16,17]. However, applying membrane technologies for wastewater treatment still faces challenges, such as easy fouling, low permeability rate, low flux rate, and limited membrane fabrication efficiency [18,19].

ML can be applied for designing, analyzing, and predicting the performance of membrane technologies, including determining the optimum operation conditions for maximum membrane performance, alleviating the need for costly, time-consuming experimental research and pilot studies. Specifically, it has been applied to predict the performance of membranes for separating contaminants from wastewater [20, 21], designing optimum conditions for membrane fabrication and applications to achieve optimum separation performance [22,23], and predicting membrane fouling [24,25].

In recent years, many reviews have been published regarding the application of ML for membrane separation-based wastewater treatment, such as ANN-based modeling for wastewater treatment [26,27] and general application of ML for wastewater treatment [28,29]. However, there is lacking systematic summary of the trend of ML development for membrane separation-based wastewater treatment. This review focused on and summarized the advantages and disadvantages of the most used ML algorithms for membrane-based wastewater treatment. The application of ML and the trend of the most widely used ML algorithms for membrane separation-based wastewater treatment in the last five years (2018-present) were discussed. Finally, a detailed conclusion and our perspectives on the current constraints and viable remedies for future research and advancements were provided.

2. Machine learning for wastewater treatment

As a subset of artificial intelligence (AI), ML is an interdisciplinary field that integrates computer science, mathematics, and statistics to perform intelligent analysis with increasingly accurate and efficient models, algorithms, and more data processing through computational learning theory and pattern recognition [30,31]. ML is developed by imitating the human ability to modify model parameters automatically, continually learning and upgrading the model over time, solving problems, and finding relationships between input and output features [32, 33].

A typical ML procedure consists of the following steps. The first step is data collection - a dataset of input and output features is obtained from publications, open-source databases, and laboratory databases. Unwanted data (*e.g.*, missing values, duplicate values) in the dataset will be preprocessed, typically involving a normalization process and/or data cleaning. After preprocessing, the data structure in the relationship between data features will be visualized, and the dataset will be separated into training and testing sets. Second, ML algorithms are selected and trained for autonomous learning using the training dataset to find

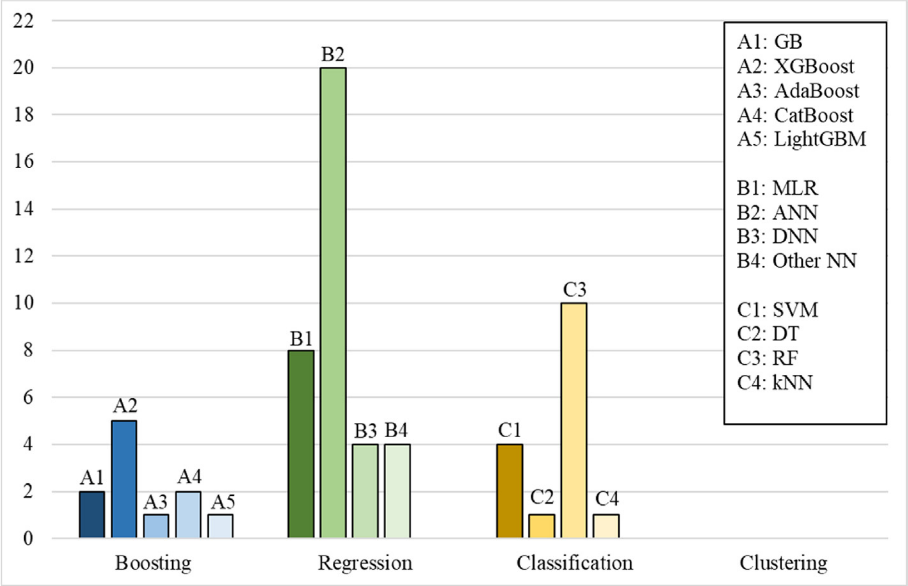


Fig. 2. Machine learning algorithms used for membrane separation-based wastewater treatment in recent five years (2018- present).

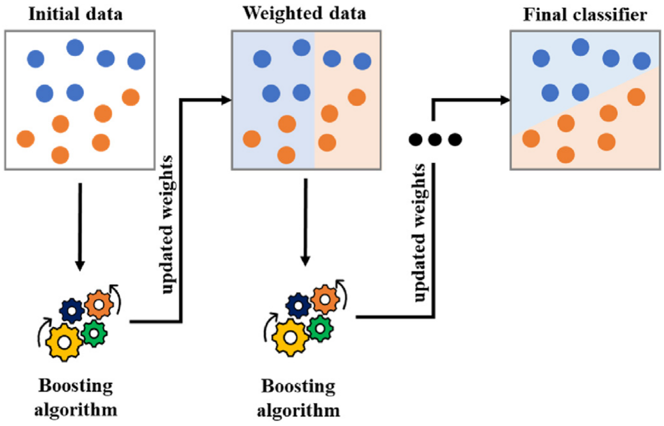


Fig. 3. Schematic diagram of boosting algorithm.

patterns and make predictions. During autonomous learning, ML algorithms will optimize and adjust model parameters. Third, the trained model will be validated by a comparison based on the testing dataset. Accuracy, precision, and sensitivity are the three main metrics used to evaluate the performance of developed ML model and associated metrics include mean square error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) [34,35]. The model will then be adjusted by tuning the parameters until achieving good performance if overfitting or underfitting. Finally, the developed optimum model is applied for new scenario prediction or identifying patterns of datasets [36,37].

Wastewater treatment is a complicated, nonlinear system involved with varying flow rates, pollutant loads, chemical environment, and hydraulic conditions, making the management of wastewater treatment environments challenging [38,39]. ML has been applied in various aspects of wastewater treatment (Fig. 1), such as predicting effluent quality [40,41], estimating pathogens in the wastewater [42,43], predicting contaminant removal [44,45], predicting wastewater treatment plants (WWTPs) operation [46,47], and predicting energy cost and consumption of WWTPs [48]. The existing applications of ML algorithms for membrane separation-based wastewater treatment vary depending on the data types and structures, learning techniques, and expected output. The details will be discussed in the next section.

3. Machine learning for membrane separation-based wastewater treatment

ML has been applied in various wastewater treatment processes, especially for membrane separation-based processes, due to its advantages and the capacity to improve membrane separation performance in wastewater treatment areas. ML is increasingly applied for membrane separation-based wastewater treatment in recent years. For example, it has been applied for permeate flux prediction [49,50], increasing in predicted transmembrane pressure (TMP) prediction [51], appropriate selection of membranes for treating wastewater [52], membrane fouling solution prediction [53], and prediction of WWTPs performance and pollutant removal [54,55]. 36 publications that are on the application of ML for membrane separation-based wastewater treatment in the past five years were gathered from Web of Science and Google Scholar and shown in Fig. 2. Regression algorithms, especially ANN are the most used ML algorithm, followed by boosting algorithms. In contrast, clustering algorithms are the lowest-studied ML algorithms in membrane separation-based wastewater treatment.

3.1. Boosting algorithms

Boosting algorithms are built for adjusting model parameters to obtain high accuracy and computation speed, particularly for large and complex datasets. Gradient Boosting (GB) is an approach that builds models sequentially, with each model attempting to decrease the error of the previous model. It updates weights in the model to decrease the error and builds a new model based on the errors or residuals from the previous model, which can be applied for regression problems, as shown in Fig. 3 [56]. Adaptive Boosting (AdaBoost) operates through an iterative process wherein weak models are successively trained based on the training data. In each iteration, the weights of the samples are adjusted to prioritize those misclassified in the preceding iteration. The final classifier is determined by conducting a weighted majority vote among all the trees within the ensemble of decision tree [57]. At the same time, eXtreme Gradient Boosting (XGBoost) is the improved version of GB for tree-based learning algorithms by increasing functions of algorithm, such as multiple learning, regularized learning, and automatic sparse awareness of missing values, which is widely used for regression and classification problems [58]. Light Gradient Boosting Machine (LightGBM) is developed based on decision tree algorithms. It is suitable for numerous datasets with fast training speed and higher performance in missing values awareness as compared to XGBoost [59]. Categorical Boosting (CatBoost) has strong compatibility with categorical data, which is effective in managing categorical information through the utilization of a systematic encoding technique for the categorical attributes, wherein the order information is included in the learning procedure. The utilization of this methodology yields models of higher precision and mitigates overfitting to a greater extent as compared to conventional encoding techniques [60].

Boosting algorithms are used to achieve highly effective, accurate, and flexible modelling and prediction in the wastewater treatment field [61,62]. AdaBoost, GB, and random forest (RF) models have been used to predict the removal performance of perfluorooctane sulfonate (PFOS) from contaminated water using nanofiltration (NF) membranes with different environmental and operating parameters. It was found that AdaBoost, GBM, and RF performed high performance in predicting PFOS removal with MSE of 2.8794, 2.450, and 4.726, respectively, and R^2 of 0.968, 0.975, and 0.930, respectively [63]. Gao et al. (2022) used XGBoost and CatBoost to develop ML by treating missing values of the membrane datasets of permeability and salt rejection. The efficacy of these models in membrane design was demonstrated by identifying the most favorable combinations of membrane materials and production parameters [64]. Additionally, in real applications, GB is optimal for

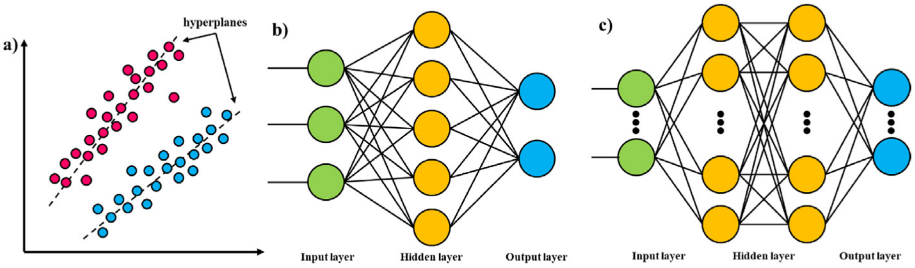


Fig. 4. Schematic diagram of regression algorithms: a) multiple linear regression, b) artificial neural network, c) deep learning neural network.

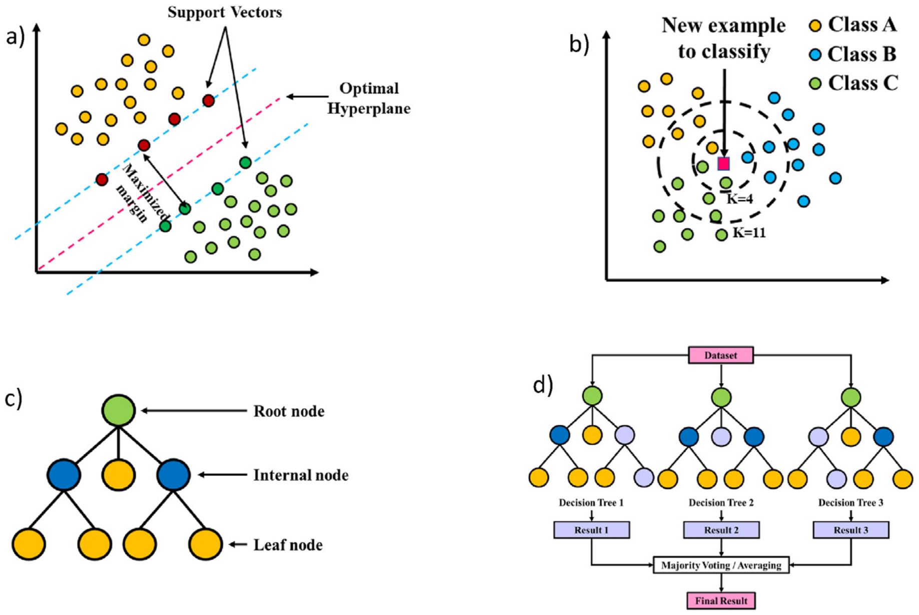


Fig. 5. Schematic diagram of classification algorithms: a) support vector machine, b) K-nearest neighbors, c) decision tree, and d) random forest.

neural networks (NNs) that create the gradient of the loss function to modify the neuronal weight [65]. GB may never converge if it is excessively slow because it is difficult to identify a precise local minimum [59].

3.2. Regression algorithms

Regression algorithms are supervised learning algorithms that find and relate predictive relationships between input and output features [66,67]. The most used regression algorithms for wastewater treatment include multiple linear regression (MLR) and NNs.

3.2.1. Multiple linear regression

MLR connects various independent and dependent variables, and the input features are potentially related to both output and input features, known as multicollinearity (Fig. 4a). In Fig. 4a, there are two straight hyperplanes, known as predictive relationships between input and output features as linearity, representing the relationship between multiple independent and one dependent variable. MLR has been applied to develop and predict the heavy metal permeate flux in a complexation-microfiltration, which was compared with ANN and polynomial neural network (PNN) models. The result showed that the MLR model, ANN and PNN performed well in flux prediction with R^2 of 0.9648 [68].

3.2.2. NN

NN is a mathematical model that imitates biological neural in terms of structure and function. NN is suitable for complicated nonlinear modeling and prediction using example-based learning. NN also has multiple sub-categories, such as ANN, PNN, recurrent neural network (RNN), and deep learning neural network [69,70]. ANN is one of the most widely used algorithms because of its capability of learning and summarizing automatically, reliability, parallel processing capability, and robust nonlinear fitting capability with feed-forward network from one layer to another without revisiting a node [71]. ANN generally consists of three layers: input, hidden, and output layer, represented by green, yellow, and blue points in Fig. 4b. The input layer is composed of artificial input neurons, which are configured by initial data or input dataset added. The hidden layer is between the input and output layers and utilizes a set of weighted inputs to create output by an activation function. In comparison, RNN possesses a distinctive characteristic whereby they can effectively handle sequential information by incorporating both historical data and current input data. This attribute enables RNNs to retain and recall information, thus addressing the limitations of the feed-forward network [72]. PNN is a self-organizing network with a flexible neural network that is developed through the learning process; it does not fix the number of layers but rather adapts neural network dynamically throughout the training process [73]. Deep learning neural network has similar procedures to ANN but comprises more hidden layers than ANN (*i.e.*, multiple hidden layers) (Fig. 4c). Deep learning neural networks and ANN can achieve similar accuracy in modeling complex nonlinear systems [70,74,75]. ANN has been applied for predicting reverse osmosis membrane performance using feed characteristics and operational parameters in comparison with XGBoost, random forest, and MLR. It was found that the pressure difference during reverse osmosis (RO) operation, the salt passage of RO membrane, and the permeate flow rate were well predicted by ANN, RF, and MLR, respectively [76]. Yaqub et al. (2020) used ANN to predict the removal efficiency of Hg from simulated wastewater by polyacrylonitrile membrane in a micellar enhanced ultrafiltration (MEUF) process. It was found that ANN was reliable in optimizing the MEUF process with MSE values for training, validation, and testing as 0.00083, 0.00096, and 0.0025, respectively [77].

3.3. Classification algorithms

Classification algorithms use input datasets to find a predictive relationship of the pattern of the output dataset which is classified into predetermined categories from the same pattern of the input dataset [78,79]. The widely used classification algorithms applied in the field of wastewater treatment include logistic regression, Naïve Bayes, Bayesian network, support vector machine, decision tree, RF, and K-nearest neighbors.

3.3.1. Support vector machine (SVM)

SVM can accurately manage classification and regression problems and categorize unlabeled datasets into two groups, using the pattern of the input dataset to predict the output dataset. The SVM process generally consists of two steps: first, the SVM model is constructed by training a labeled dataset, and the hyperplane and decision boundary of the SVM model is determined, for which the decision boundary is established as a hyperplane to separate different classes of unlabeled datasets with kernel functions. Kernel functions serve to increase the accuracy and efficiency in determining the hyperplane and decision boundary. If the input dataset represents a nonlinear problem, kernel functions transform the data to obtain linear classifiers for determining the hyperplane and decision boundary. Fig. 5a is an illustration of SVM, the green and yellow points are data points from the dataset, and the blue line is the decision boundary determined based on the nearest data points of each group of the dataset (represented by the red and dark green points, and thus called support vectors). SVM has been utilized to forecast the grafting of maleic anhydride and hyperbranched polyethylene glycol (PEG) onto the surface of polyethersulfone (PES) membranes. This prediction was based on experimental datasets encompassing oil-water separation and permeation flux, and antifouling properties. The study revealed that the SMV model demonstrated accurate prediction ability of benzo-phenone formation on the PES membrane [80].

3.3.2. *Decision tree (DT)*

DT can be used to manage classification and regression problems by dividing data based on a classification tree. DT has three types of nodes in structure: the root, internal, and leaf [81,82]. Fig. 5c shows a diagram of DT within a hierarchical structure that is composed of root nodes, internal nodes, and leaf nodes, represented by the green, blue, and yellow points, respectively. The start point of DT is the root nodes, each internal node is a test on a feature, and the leaf nodes represent the test feature's outcomes. DT has many advantages, such as simple algorithm structure, great interpretability, straightforward implementation, and ease of handling categorical and quantitative values, including the ability to fill missing attribute values with the most probable value. Nevertheless, the structure of DT can be unstable or complex, with difficulty in controlling the tree size. A single DT model can be susceptible to noisy data and overfitting. DT has been applied to predict bilgewater emulsion stability based on the image processing of separation experiments of 360 emulsion samples. The result showed that the predictive performance of DT was superior for emulsion stability, as evidenced by the average MAE value of 0.1611 [83].

3.3.3. *Random forest (RF)*

RF is a method of ensemble learning for classification and regression problems. RF is a solution to the problem of overfitting that can be encountered using DT. Fig. 5d illustrates the RF structure: the green points are root nodes, the dark blue points are internal nodes, and the yellow points are leaf nodes of each DT; the light purple points are the direction decision of each DT. The result of RF will come from the majority voting or averaging of the result from all DTs. The advantages of RF include reduced overfitting and flexibility than DT. Typical disadvantages include being time-consuming, requiring a large dataset, and being more complex than DT [84,85]. RF has been utilized to forecast the performance of the backwashing technique of ultrafiltration membranes by considering environmental and operational variables, including temperature, hydraulic pressure, and water turbidity. It was shown that RF outperformed both linear regression and ANN in terms of prediction accuracy as assessed by MSE [86]. Moreover, Henry et al. employed the RF algorithm to identify the primary factor that governs the critical flux of oil-in-water emulsions in crossflow microfiltration. The analysis based on 223 datasets revealed that the crossflow velocity emerged as the most influential variable for the critical flux of oil-in-water emulsions [87].

3.3.4. *K-nearest neighbors (kNN)*

kNN is a non-parametric and supervised learning classifier using the distance between each data point. Labeled data are organized into multiple groups or classes, and unlabeled data are categorized according to the hypothesis that similar points can be near one another. Fig. 5b illustrates kNN: the green, blue, and yellow points are different classes of the dataset, and the red square is a classified dataset of new example points, which are obtained based on kNN hypothesis. For example, if k value (the number of data points around new example point) is 4 (the small circle which measures the circle radius from new example point), kNN will assign the data points in this circle to be class c (*i.e.* two green, one blue, and one yellow data points). If k value is 11 (the bigger circle), kNN will assign the data point in this circle to be class c (*i.e.* five green, three blue, and three yellow data points). The model creation of kNN is inexpensive, and it includes a technique of categorization that is adaptable and well-suited for multimodal classes. However, the maximum error rate of kNN is twice the Naïve Bayesian (NB), or when the size of the training dataset increased. The kNN needs to calculate the distance between each data point again, leading to decreasing of noisy or irrelevant parameters [88,89]. Eight ML models, *i.e.* kNN, MLR, SVM, ANN, RF, gradient-boosted decision tree (GBDT), XGBoost, and LightGBM were applied for prediction of organic contaminant removal from contaminated effluents by NF and RO membranes. It was found that RF, GBDT, XGBoost, and LightGBM performed better than the other four models in accuracy and model robustness with high R^2 of 92.4 %, 95 %, 99.5 %, and 92.9 %, respectively [90].

3.4. *Clustering algorithms*

Clustering algorithms are unsupervised learning for grouping related data without regard to the specific outcome. It is usually used to identify interesting trends or patterns in data. The algorithms will organize datasets into the same category, known as a cluster, which is more similar than other categories [91,92]. Clustering algorithms aim to divide the data into separate clusters, where the observations within each cluster exhibit similarity, while observations in other clusters demonstrate dissimilarity. However, most datasets on membrane separation-based wastewater treatment are quantitative data with known data types or data clusters, which leads to clustering of the datasets is irrelevant. Moreover, some datasets in this field are known as time-series or long-term operations, so clustering is unsuitable for data analysis. For these reasons, there are few studies using clustering algorithms for modelling the performance of membrane separation in wastewater treatment.

3.5. *Applications of ML algorithms for membrane separation-based wastewater treatment*

The applications of ML algorithms for membrane separation-based wastewater treatment were focused on the prediction of membrane performance in wastewater treatment and membrane designing for improving membrane performance. ML algorithms can be applied to predict optimum conditions for reducing the cost of operation, decreasing energy demand, and increasing the efficiency of wastewater treatments. Numerous studies have used ML algorithms for predicting membrane-based separation for wastewater treatment, which can classified into three sub-topics: contaminant removal, such as organic, heavy metal, and salt contamination in wastewater; impacts of operating parameters such as feed rate, temperature, and pressure of inflow of wastewater; and other processes, such as the performance of emulsion removal, flux, and backwashing performance [63,68,76]. Zahmatkesh et al. (2022) determined optimal conditions for reducing biological oxygen demand (BOD) and chemical oxygen demand (COD) with polymeric membranes, using ANN, achieved highly appropriate for predicting the removal of BOD and COD with R^2 and RMSE as 0.99 and 0.05 %, and 0.99 and 0.99 %, respectively [93]. Odabaşi et al. (2021) predicted RO membrane performance from feed characteristics of municipal wastewater, using RF, XGBoost, ANN, and MLR, achieved MLR more effective than other methods [94].

ML algorithms can be applied to facilitate the design and fabrication of membranes to improve performance by identifying materials for membrane fabrication, fabrication parameters, and modification membrane methods [64,80]. which are the most important ways to gain the best results of wastewater treatment with using membrane separation. Gao et al. (2023) designed a high-performance ultrafiltration membrane for wastewater treatment and resource recovery using XGBoost and

Table 1
Advantages and disadvantages of the most used machine learning algorithms for membrane separation-based wastewater treatment.

Algorithms	Typical used	Advantages	Disadvantages	Ref.
1. Boosting Algorithms				
Gradient Boosting	Regression and classification	Simple implementation, easy to understand, high accuracy, fast computation for larger datasets, achieving low error with small datasets.	Easy to be overfitting models and sensitive to outliers.	[56,97]
2. Regression Algorithms				
Multiple Linear Regression	Regression	Suitable with a linear relationship between one independent and more than one dependent variable and the ability to identify outliers.	Not good to explain of nonlinear relationship between independent and dependent variables, easily resulting in prediction errors.	[98,99]
Neural Network	Regression and classification	High efficiency, good with nonlinear data, continuous and long learning, multitasking and multiple results simultaneously, and flexible and wide applications.	Complex and difficult to explain, invisible, requires lots of data, needs great attention in data preparation, and adjusting optimization models can be challenging.	[100,101]
3. Classification Algorithms				
Support Vector Machine	Regression and classification	Simple training, good with high dimensional data, capable of handling both continuous and categorical data, and high prediction accuracy.	Not suitable for large datasets, performs poorly when the dataset has more noise and overlapping classes, and cannot provide probability estimates.	[102–104]
Decision Tree	Regression and classification	Simple implementation, highly interpretable, simple algorithm structure, requires less data preparation, no need for data normalization and scaling, and missing values do not affect the models.	Requires large memory during computation, small data changes affect algorithm structure, unstable algorithm structure, easy to be the overfitting models.	[105,106]
Random Forest	Regression and classification	Good with high dimensional data, capable of handling both continuous and categorical data, high prediction accuracy, no need for data normalization and scaling, and missing values do not affect the models.	Complex algorithm, high computational cost, and requires much time for computation.	[107–109]

CatBoost, while Fetanat et al. (2021) designed an ultrafiltration polymeric membrane using ANN, which guided the design of separation membranes suited to their intended purpose [95,96].

Each ML method has its own strengths and drawbacks (Table 1), leading to selecting suitable algorithms as one of the most significant aspects of ML research. Membrane separation-based wastewater treatment mainly involves quantitative and non-stationary data. Using ML for membrane separation-based wastewater treatment can face some challenges, such as data characteristics, discontinuous of datasets, and ML algorithms selection and suitability. Integration of ML with other modeling approaches is a potential way to overcome these challenges. The details will be discussed in the next section.

4. Integration of ML with other modeling approaches

When the datasets are continuously increased, also known as time-series or long-term operation data, non-stationary data distributions of experiments will lead to interruption of model computation and a decrease of model accuracy [110,111]. Consequently, researchers have undertaken studies aimed at enhancing and advancing the efficacy of ML techniques in the context of long-term wastewater treatment operations. For instance, an existing study has paid attention to predicting the changing trends of chemical oxygen demand (COD) in the outflow of wastewater treatment, considering various temperature and water inflow data over 20 months in real-time operations with support vector regression (SVR), long short-term memory neural networks, and RNN [97]. Shi et al. (2021) predicted the performance of municipal wastewater treatment by two anaerobic membrane bioreactors (1 year operation) using the approach convolutional neural network [49]. Nevertheless, the application of ML approach for predicting long-term operation of membrane separation-based wastewater treatment has not been reported yet.

The ML approach's expansion in the wastewater treatment domain to encompass other applications can be achieved by the coordination or integration of ML with alternative modeling approaches. For instance, the combination of ML with the response surface method (RSM) can be employed to identify the optimal operating parameters for wastewater treatment. Aghilesh et al. (2021) used RSM, ANN, and adaptive neuro-fuzzy inference systems (ANFIS) to model and optimize a forward osmosis (FO) process for textile industry wastewater treatment [112]. Bhatti et al. (2011) used RSM and ANN to optimize the copper removal efficiency and minimize the energy consumption for a copper wastewater treatment process [113]. Li et al. (2022) combined RSM and ANN to optimize wastewater treatment membrane fabrication [114]. It is expected that choosing the appropriate algorithms and operation features is key to achieve high efficiency and accuracy in modelling and designing membrane separation-based wastewater treatment.

5. Conclusions and future perspective

ML has been studied and applied to predict and improve the separation performance and efficiency of membranes, reduce energy consumption and cost of fabrication and operation, and design and find appropriate for membrane fabrication and operation. This work reviewed commonly used ML algorithms for membrane separation-based wastewater treatment and their advantages and disadvantages. The commonly used ML algorithms for the membrane separation of wastewater can be categorized into three groups: boosting, regression, and classification. Artificial neural networks (ANN) have been widely applied to predict the performance of membranes includingseparation efficiency, and for membrane designing and fabrication.

Due to the quantitative and non-stationary data characteristics of membrane separation-based wastewater treatment, ML algorithms were commonly used to predict and classify datasets to enhance the performance of membranes, as opposed to clustering datasets. There are still challenges on the use of ML for modelling membrane separation-based wastewater treatment, in particular, when it is used to deal with time-series or long-term operation data. Integrating ML with other modeling approaches for long-term operation is a potential direction ahead.

There are several important challenges to the use of ML for modelling membrane separation-based wastewater treatment. First, selecting a suitable ML algorithm for the intended application can be tricky because of the wide range of algorithms available. With the same input of the dataset, the prediction outcomes may change depending on the different algorithms used. For example, DT is a simple algorithm, and missing values do not affect the models, but the models are easily overfitting, and the algorithm structure will affect when there are small data changes. On the other hand, RF is a high prediction accuracy algorithm capable of handling continuous and categorical data, and missing values do not affect the models. Therefore, after data preprocessing, choosing appropriate algorithms is key and it is necessary to create more knowledge about the suitability of different models for different applications. Second, combining ML algorithms with more than one algorithm or with other models (e.g., physics-informed models) can potentially help to improve efficiency and accuracy of models and outcomes. Nevertheless, the utilization of integrated process-based models and combining long-term operation and intensive datasets are limited by the lack of appropriate data and relevant collection and management strategies.

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.

Acknowledgments

This work is financially supported by National Energy R&D Center of Petroleum Refining Technology (RIPP, SINOPEC), National Natural Science Foundation of China (22078347), Key Research and Development Program of Hebei Province, China (21373303D), Science and Technology Support Project of Guizhou Province ([2019]2839), Distinguished Professor of Henan Province (220508001). Siming You would like to thank the financial support from the Royal Society International Exchange Scheme (EC\NSFC\211175).

References

[1] B.T. Hazen, C.A. Boone, J.D. Ezell, L.A. Jones-Farmer, Data quality for data science, predictive analytics, and big data in supply chain management: an introduction to the problem and suggestions for research and applications, Int. J. Prod. Econ. 154 (2014) 72–80, https://doi.org/10.1016/j.ijpe.2014.04.018.

[2] K. Vassakis, E. Petrakis, I. Kopanakis, Big Data Analytics: Applications, Prospects and Challenges, 2018, pp. 3–20, https://doi.org/10.1007/978-3-319-67925-9_1.

[3] B. Knüsel, C. Baumberger, Understanding climate phenomena with data-driven models, Stud. Hist. Philos. Sci. 84 (2020) 46–56, https://doi.org/10.1016/j.shpsa.2020.08.003.

[4] D. Wang, S. Thunéll, U. Lindberg, L. Jiang, J. Trygg, M. Tysklind, N. Souihi, A machine learning framework to improve effluent quality control in wastewater treatment plants, Sci. Total Environ. 784 (2021) 147138, https://doi.org/10.1016/j.scitotenv.2021.147138.

[5] O. Icke, D.M. van Es, M.F. de Koning, J.J.G. Wuister, J. Ng, K.M. Phua, Y.K.K. Koh, W.J. Chan, G. Tao, Performance improvement of wastewater treatment processes by application of machine learning, Water Sci. Technol. 82 (2020) 2671–2680, https://doi.org/10.2166/wst.2020.382.

[6] P. Jiang, X. Liu, J. Zhang, X. Yuan, A framework based on hidden Markov model with adaptive weighting for microcystin forecasting and early-warning, Decis. Support Syst. 84 (2016) 89–103, https://doi.org/10.1016/j.dss.2016.02.003.

[7] A. Singh, M. Imtiyaz, Hydrological Modelling Using Process Based and Data Driven Models, Scholars' Press, 2013.

[8] A. Bhusal, U. Parajuli, S. Regmi, A. Kalra, Application of machine learning and process-based models for rainfall-runoff simulation in DuPage river basin, Illinois, Hydrology 9 (2022) 117, https://doi.org/10.3390/hydrology9070117.

[9] S. Petruseva, V. Zileska-Pancovska, D. Car-Pušić, Implementation of process-based and data-driven models for early prediction of construction time, Adv. Civ. Eng. 2019 (2019) 1–12, https://doi.org/10.1155/2019/7405863.

[10] R.J. Desai, S.V. Wang, M. Vaduganathan, T. Evers, S. Schneeweiss, Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes, JAMA Netw. Open 3 (2020) e1918962, https://doi.org/10.1001/jamanetworkopen.2019.18962.

[11] L.N. Grendas, L. Chiapella, D.E. Rodante, F.M. Daray, Comparison of traditional model-based statistical methods with machine learning for the prediction of suicide behaviour, J. Psychiatr. Res. 145 (2022) 85–91, https://doi.org/10.1016/j.jpsychires.2021.11.029.

[12] J. Wu, X.Y. Chen, H. Zhang, L.D. Xiong, H. Lei, S.H. Deng, Hyperparameter optimization for machine learning models based on Bayesian optimization, J. Electron. Sci. Technol. 17 (2019) 26–40, https://doi.org/10.11989/JEST.1674-862X.80904120.

[13] L. Yang, A. Shami, On hyperparameter optimization of machine learning algorithms: theory and practice, Neurocomputing 415 (2020) 295–316, https://doi.org/10.1016/j.neucom.2020.07.061.

[14] J.P. Vareda, A.J.M. Valente, L. Durães, Assessment of heavy metal pollution from anthropogenic activities and remediation strategies: a review, J. Environ. Manag. 246 (2019) 101–118, https://doi.org/10.1016/j.jenvman.2019.05.126.

[15] S. Deng, X. Yan, Q. Zhu, C. Liao, The utilization of reclaimed water: possible risks arising from waterborne contaminants, Environ. Pollut. 254 (2019) 113020, https://doi.org/10.1016/j.envpol.2019.113020.

[16] E. Obotey Ezugbe, S. Rathilal, Membrane technologies in wastewater treatment: a review, Membranes 10 (2020) 89, https://doi.org/10.3390/membranes10050089.

[17] D. Yokoyama, S. Suzuki, T. Asakura, J. Kikuchi, Chemometric analysis of NMR spectra and machine learning to investigate membrane fouling, ACS Omega 7 (2022) 12654–12660, https://doi.org/10.1021/acsomega.1c06891.

[18] C. Algieri, V. Pugliese, G. Coppola, S. Curcio, V. Calabro, S. Chakraborty, Arsenic removal from groundwater by membrane technology: advantages, disadvantages, and effect on human health, Groundw. Sustain. Dev. 19 (2022) 100815, https://doi.org/10.1016/j.gsd.2022.100815.

[19] Y. Liu, H. Liu, Z. Shen, Nanocellulose based filtration membrane in industrial waste water treatment: a review, Materials 14 (2021) 5398, https://doi.org/10.3390/ma14185398.

[20] T. Bonny, M. Kashkash, F. Ahmed, An efficient deep reinforcement machine learning-based control reverse osmosis system for water desalination, Desalination 522 (2022) 115443, https://doi.org/10.1016/j.desal.2021.115443.

[21] P. Priya, T.C. Nguyen, A. Saxena, N.R. Aluru, Machine learning assisted screening of two-dimensional materials for water desalination, ACS Nano 16 (2022) 1929–1939, https://doi.org/10.1021/acsnano.1c05345.

[22] P. Behnam, A. Shafieian, M. Zargar, M. Khiadani, Development of machine learning and stepwise mechanistic models for performance prediction of direct contact membrane distillation module- A comparative study, Chem. Eng. Process. - Process Intensif. 173 (2022) 108857, https://doi.org/10.1016/j.cep.2022.108857.

[23] T. Liu, L. Liu, F. Cui, F. Ding, Q. Zhang, Y. Li, Predicting the performance of polyvinylidene fluoride, polyethersulfone and polysulfone filtration membranes using machine learning, J. Mater. Chem. A 8 (2020) 21862–21871, https://doi.org/10.1039/D0TA07607D.

[24] D.J. Kovacs, Z. Li, B.W. Baetz, Y. Hong, S. Donnaz, X. Zhao, P. Zhou, H. Ding, Q. Dong, Membrane fouling prediction and uncertainty analysis using machine learning: a wastewater treatment plant case study, J. Membr. Sci. 660 (2022) 120817, https://doi.org/10.1016/j.memsci.2022.120817.

[25] M. Bagheri, A. Akbari, S.A. Mirbagheri, Advanced control of membrane fouling in filtration systems using artificial intelligence and machine learning techniques: a critical review, Process Saf. Environ. Protect. 123 (2019) 229–252, https://doi.org/10.1016/j.psep.2019.01.013.

[26] J. Jawad, A.H. Hawari, S. Javaid Zaidi, Artificial neural network modeling of wastewater treatment and desalination using membrane processes: a review, Chem. Eng. J. 419 (2021) 129540, https://doi.org/10.1016/j.cej.2021.129540.

[27] N.D. Viet, D. Jang, Y. Yoon, A. Jang, Enhancement of membrane system performance using artificial intelligence technologies for sustainable water and wastewater treatment: a critical review, Crit. Rev. Environ. Sci. Technol. 52 (2022) 3689–3719, https://doi.org/10.1080/10643389.2021.1940031.

[28] M. Lowe, R. Qin, X. Mao, A review on machine learning, artificial intelligence, and smart technology in water treatment and monitoring, Water 14 (2022) 1384, https://doi.org/10.3390/w14091384.

[29] S. Safeer, R.P. Pandey, B. Rehman, T. Safdar, I. Ahmad, S.W. Hasan, A. Ullah, A review of artificial intelligence in water purification and wastewater treatment: recent advancements, J. Water Process Eng. 49 (2022) 102974, https://doi.org/10.1016/j.jwpe.2022.102974.

[30] H. Jiang, Machine Learning Fundamentals : A Concise Introduction, 2021.

[31] I.H. Sarker, M.H. Furhad, R. Nowrozy, AI-driven cybersecurity: an overview, security intelligence modeling and research directions, SN Comput. Sci. 2 (2021), https://doi.org/10.1007/s42979-021-00557-0.

[32] Z. Shi, W. Yang, X. Deng, C. Cai, Y. Yan, H. Liang, Z. Liu, Z. Qiao, Machine-learning-assisted high-throughput computational screening of high performance metal-organic frameworks, Mol. Syst. Des. Eng. 5 (2020) 725–742, https://doi.org/10.1039/d0me00005a.

[33] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine learning in materials informatics: recent applications and prospects, npj Comput. Mater. 3 (2017) 54, https://doi.org/10.1038/s41524-017-0056-5.

[34] A. Kassambara, Machine Learning Essentials: Practical Guide in R, CreateSpace Independent Publishing Platform, 2018.

[35] Z.H. Zhou, Machine Learning, Springer Singapore, Singapore, 2021, https://doi.org/10.1007/978-981-15-1967-3.

[36] S.J. Qin, L.H. Chiang, Advances and opportunities in machine learning for process data analytics, Comput. Chem. Eng. 126 (2019) 465–473, https://doi.org/10.1016/j.compchemeng.2019.04.003.

[37] C. Shang, F. You, Data analytics and machine learning for smart process manufacturing: recent advances and perspectives in the big data era, Engineering 5 (2019) 1010–1016, https://doi.org/10.1016/j.eng.2019.01.019.

[38] V. Vučić, C. Süring, H. Harms, S. Müller, S. Günther, A framework for P-cycle assessment in wastewater treatment plants, Sci. Total Environ. 760 (2021) 143392, https://doi.org/10.1016/j.scitotenv.2020.143392.

[39] S. Borzooei, G. Campo, A. Cerutti, L. Meucci, D. Panepinto, M. Ravina, V. Riggio, B. Ruffino, G. Scibilia, M. Zanetti, Optimization of the wastewater treatment plant: from energy saving to environmental impact mitigation, Sci. Total Environ. 691 (2019) 1182–1189, https://doi.org/10.1016/j.scitotenv.2019.07.241.

[40] H. Guo, K. Jeong, J. Lim, J. Jo, Y.M. Kim, J. Park, J.H. Kim, K.H. Cho, Prediction of effluent concentration in a wastewater treatment plant using machine learning models, J. Environ. Sci. 32 (2015) 90–101, https://doi.org/10.1016/j.jes.2015.01.007.

[41] F. Granata, S. Papirio, G. Esposito, R. Gargano, G. De Marinis, Machine learning algorithms for the forecasting of wastewater quality indicators, Water 9 (2017) 105, https://doi.org/10.3390/w9020105.

[42] O.M. Abdeldayem, A.M. Dabbish, M.M. Habashy, M.K. Mostafa, M. Elhefnawy, L. Amin, E.G. Al-Sakkari, A. Ragab, E.R. Rene, Viral outbreaks detection and surveillance using wastewater-based epidemiology, viral air sampling, and

machine learning techniques: a comprehensive review and outlook, Sci. Total Environ. 803 (2022) 149834, https://doi.org/10.1016/j.scitotenv.2021.149834.

[43] W. Cai, F. Long, Y. Wang, H. Liu, K. Guo, Enhancement of microbiome management by machine learning for biological wastewater treatment, Microb. Biotechnol. 14 (2021) 59–62, https://doi.org/10.1111/1751-7915.13707.

[44] B. Caglar Gencosman, G. Eker Sanli, Prediction of polycyclic aromatic hydrocarbons (PAHs) removal from wastewater treatment sludge using machine learning methods, Water, Air, Soil Pollut. 232 (2021) 87, https://doi.org/10.1007/s11270-021-05049-8.

[45] N. Taoufik, W. Boumya, M. Achak, H. Chennouk, R. Dewil, N. Barka, The state of art on the prediction of efficiency and modeling of the processes of pollutants removal based on machine learning, Sci. Total Environ. 807 (2022) 150554, https://doi.org/10.1016/j.scitotenv.2021.150554.

[46] P.M.L. Ching, R.H.Y. So, T. Morck, Advances in soft sensors for wastewater treatment plants: a systematic review, J. Water Process Eng. 44 (2021) 102367, https://doi.org/10.1016/j.jwpe.2021.102367.

[47] D. Wang, S. Thunéll, U. Lindberg, L. Jiang, J. Trygg, M. Tysklind, Towards better process management in wastewater treatment plants: process analytics based on SHAP values for tree-based machine learning methods, J. Environ. Manag. 301 (2022) 113941, https://doi.org/10.1016/j.jenvman.2021.113941.

[48] D. Torregrossa, U. Leopold, F. Hernández-Sancho, J. Hansen, Machine learning for energy cost modelling in wastewater treatment plants, J. Environ. Manag. 223 (2018) 1061–1067, https://doi.org/10.1016/j.jenvman.2018.06.092.

[49] Y. Shi, Z. Wang, X. Du, B. Gong, V. Jegatheesan, I.U. Haq, Recent advances in the prediction of fouling in membrane bioreactors, Membranes 11 (2021) 381, https://doi.org/10.3390/membranes11060381.

[50] C. Quezada, H. Estay, A. Cassano, E. Troncoso, R. Ruby-Figueroa, Prediction of permeate flux in ultrafiltration processes: a review of modeling approaches, Membranes 11 (2021) 368, https://doi.org/10.3390/membranes11050368.

[51] M. Kamali, L. Appels, X. Yu, T.M. Aminabhavi, R. Dewil, Artificial intelligence as a sustainable tool in wastewater treatment using membrane bioreactors, Chem. Eng. J. 417 (2021) 128070, https://doi.org/10.1016/j.cej.2020.128070.

[52] R. Maleki, S.M. Shams, Y.M. Chellehbari, S. Rezvantalab, A.M. Jahromi, M. Asadnia, R. Abbassi, T. Aminabhavi, A. Razmjou, Materials discovery of ion-selective membranes using artificial intelligence, Commun. Chem. 5 (2022) 132, https://doi.org/10.1038/s42004-022-00744-x.

[53] M.T. Gaudio, G. Coppola, L. Zangari, S. Curcio, S. Greco, S. Chakraborty, Artificial intelligence-based optimization of industrial membrane processes, Earth Syst. Environ. 5 (2021) 385–398, https://doi.org/10.1007/s41748-021-00220-x.

[54] G. Wang, Q.S. Jia, M. Zhou, J. Bi, J. Qiao, A. Abusorrah, Artificial neural networks for water quality soft-sensing in wastewater treatment: a review, Artif. Intell. Rev. 55 (2022) 565–587, https://doi.org/10.1007/s10462-021-10038-8.

[55] Z.M. Yaseen, An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: review, challenges and solutions, Chemosphere 277 (2021) 130126, https://doi.org/10.1016/j.chemosphere.2021.130126.

[56] A. Manoharan, K.M. Begam, V.R. Aparow, D. Sooriamoorthy, Artificial neural networks, gradient boosting and support vector machines for electric vehicle battery state estimation: a review, J. Energy Storage 55 (2022) 105384, https://doi.org/10.1016/j.est.2022.105384.

[57] Y. Cao, Q.G. Miao, J.C. Liu, L. Gao, Advance and prospects of AdaBoost algorithm, Acta Autom. Sin. 39 (2013) 745–758, https://doi.org/10.1016/S1874-1029(13)60052-X.

[58] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: KDD '16 Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min, 2016, https://doi.org/10.1145/2939672.2939785. New York, NY.

[59] J. Korstanje, Gradient boosting with XGBoost and LightGBM, in: Adv. Forecast. With Python, Apress, Berkeley, CA, 2021, pp. 193–205, https://doi.org/10.1007/978-1-4842-7150-6_15.

[60] A.V. Dorogush, V. Ershov, A. Gulin, CatBoost: Gradient Boosting with Categorical Features Support, 2018. http://arxiv.org/abs/1810.11363.

[61] W. Zhang, R. Zhang, C. Wu, A.T.C. Goh, S. Lacasse, Z. Liu, H. Liu, State-of-the-art review of soft computing applications in underground excavations, Geosci. Front. 11 (2020) 1095–1106, https://doi.org/10.1016/j.gsf.2019.12.003.

[62] C. Bentéjac, A. Csörgő, G. Martínez-Muñoz, A comparative analysis of gradient boosting algorithms, Artif. Intell. Rev. 54 (2021) 1937–1967, https://doi.org/10.1007/s10462-020-09896-5.

[63] A. Hosseinzadeh, J.L. Zhou, J. Zyaie, N. AlZainati, I. Ibrar, A. Altaee, Machine learning-based modeling and analysis of PFOS removal from contaminated water by nanofiltration process, Sep. Purif. Technol. 289 (2022) 120775, https://doi.org/10.1016/j.seppur.2022.120775.

[64] 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, Environ. Sci. Technol. 56 (2022) 2572–2581, https://doi.org/10.1021/acs.est.1c04373.

[65] X. Zhou, Understanding the convolutional neural networks with gradient descent and backpropagation, J. Phys. Conf. Ser. 1004 (2018), https://doi.org/10.1088/1742-6596/1004/1/012028.

[66] D. Maulud, A.M. Abdulazeez, A review on linear regression comprehensive in machine learning, J. Appl. Sci. Technol. Trends. 1 (2020) 140–147, https://doi.org/10.38094/jastt1457.

[67] Y. Liu, O.C. Esan, Z. Pan, L. An, Machine learning for advanced energy materials, Energy AI 3 (2021) 100049, https://doi.org/10.1016/j.egyai.2021.100049.

[68] Z. Sekulić, D. Antanasijević, S. Stevanović, K. Trivunac, The prediction of heavy metal permeate flux in complexation-microfiltration process: polynomial neural network approach, Water, Air, Soil Pollut. 230 (2019) 23, https://doi.org/10.1007/s11270-018-4072-y.

[69] O. Oyebode, D. Stretch, Neural network modeling of hydrological systems: a review of implementation techniques, Nat. Resour. Model. 32 (2019) e12189, https://doi.org/10.1111/nrm.12189.

[70] O.I. Abiodun, A. Jantan, A.E. Omolara, K.V. Dada, N.A.E. Mohamed, H. Arshad, State-of-the-art in artificial neural network applications: a survey, Heliyon 4 (2018) e00938, https://doi.org/10.1016/j.heliyon.2018.e00938.

[71] Y. Wu, J. Feng, Development and application of artificial neural network, Wirel. Pers. Commun. 102 (2018) 1645–1656, https://doi.org/10.1007/s11277-017-5224-x.

[72] P. Rodriguez, J. Wiles, J.L. Elman, A recurrent neural network that learns to count, Connect. Sci. 11 (1999) 5–40, https://doi.org/10.1080/095400999116340.

[73] S. Das, The polynomial neural network, Inf. Sci. 87 (1995) 231–246, https://doi.org/10.1016/0020-0255(95)00133-6.

[74] S.F. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A.R. Varkonyi-Koczy, U. Reuter, T. Rabczuk, P.M. Atkinson, COVID-19 outbreak prediction with machine learning, Algorithms 13 (2020), https://doi.org/10.3390/a13100249.

[75] R.Y. Choi, A.S. Coyner, J. Kalpathy-Cramer, M.F. Chiang, J. Peter Campbell, Introduction to machine learning, neural networks, and deep learning, Transl. Vis. Sci. Technol. 9 (2020) 1–12, https://doi.org/10.1167/tvst.9.2.14.

[76] Ç. Odabaşı, P. Dologlu, F. Gülmez, G. Kuşoğlu, Ö. Çağlar, Investigation of the factors affecting reverse osmosis membrane performance using machine-learning techniques, Comput. Chem. Eng. 159 (2022) 107669, https://doi.org/10.1016/j.compchemeng.2022.107669.

[77] M. Yaqub, S.H. Lee, Micellar enhanced ultrafiltration (MEUF) of mercury-contaminated wastewater: experimental and artificial neural network modeling, J. Water Process Eng. 33 (2020) 101046, https://doi.org/10.1016/j.jwpe.2019.101046.

[78] R.C. Chen, C. Dewi, S.W. Huang, R.E. Caraka, Selecting critical features for data classification based on machine learning methods, J. Big Data 7 (2020) 52, https://doi.org/10.1186/s40537-020-00327-4.

[79] D. Sisodia, D.S. Sisodia, Prediction of diabetes using classification algorithms, Procedia Comput. Sci. 132 (2018) 1578–1585, https://doi.org/10.1016/j.procs.2018.05.122.

[80] H. Adib, A. Raisi, B. Salari, Support vector machine-based modeling of grafting hyperbranched polyethylene glycol on polyethersulfone ultrafiltration membrane for separation of oil–water emulsion, Res. Chem. Intermed. 45 (2019) 5725–5743, https://doi.org/10.1007/s11164-019-03931-z.

[81] S. Lamichhane, L. Kumar, B. Wilson, Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: a review, Geoderma 352 (2019) 395–413, https://doi.org/10.1016/j.geoderma.2019.05.031.

[82] S. Rizvi, B. Rienties, S.A. Khoja, The role of demographics in online learning; A decision tree based approach, Comput. Educ. 137 (2019) 32–47, https://doi.org/10.1016/j.compedu.2019.04.001.

[83] W.H. Lee, C.Y. Park, D. Diaz, K.L. Rodriguez, J. Chung, J. Church, M.R. Willner, J.G. Lundin, D.M. Paynter, Predicting bilgewater emulsion stability by oil separation using image processing and machine learning, Water Res. 223 (2022) 118977, https://doi.org/10.1016/j.watres.2022.118977.

[84] A. Sekulić, M. Kilibarda, G.B.M. Heuvelink, M. Nikolić, B. Bajat, Random forest spatial interpolation, Rem. Sens. 12 (2020) 1687, https://doi.org/10.3390/rs12101687.

[85] P. Probst, M.N. Wright, A. Boulesteix, Hyperparameters and tuning strategies for random forest, WIREs Data Min. Knowl. Discov. 9 (2019), https://doi.org/10.1002/widm.1301.

[86] B. Zhang, G. Kotsalis, J. Khan, Z. Xiong, T. Igou, G. Lan, Y. Chen, Backwash sequence optimization of a pilot-scale ultrafiltration membrane system using data-driven modeling for parameter forecasting, J. Membr. Sci. 612 (2020) 118464, https://doi.org/10.1016/j.memsci.2020.118464.

[87] H.J. Tanudjaja, J.W. Chew, Application of machine learning-based models to understand and predict critical flux of oil-in-water emulsion in crossflow microfiltration, Ind. Eng. Chem. Res. (2021), https://doi.org/10.1021/acs.iecr.1c04662.

[88] Z. Pan, Y. Wang, Y. Pan, A new locally adaptive k-nearest neighbor algotithm based on discrimination class, Knowl. Base Syst. 204 (2020) 106185, https://doi.org/10.1016/j.knosys.2020.106185.

[89] I. Triguero, D. García-Gil, J. Maillo, J. Luengo, S. García, F. Herrera, Transforming big data into smart data: an insight on the use of the k-nearest neighbors algorithm to obtain quality data, WIREs Data Min. Knowl. Discov. 9 (2019), https://doi.org/10.1002/widm.1289.

[90] 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) 159348, https://doi.org/10.1016/j.scitotenv.2022.159348.

[91] C. Yuan, H. Yang, Research on K-value selection method of K-means clustering algorithm 2 (2019) 226–235, https://doi.org/10.3390/j2020016. J.

[92] M.Z. Rodriguez, C.H. Comin, D. Casanova, O.M. Bruno, D.R. Amancio, L. da F. Costa, F.A. Rodrigues, Clustering algorithms: a comparative approach, PLoS One 14 (2019) e0210236, https://doi.org/10.1371/journal.pone.0210236.

[93] S. Zahmatkesh, Y. Rezakhani, A. Arabi, M. Hasan, Z. Ahmad, C. Wang, M. Sillanpää, M. Al-Bahrani, I. Ghodrati, An approach to removing COD and BOD based on polycarbonate mixed matrix membranes that contain hydrous manganese oxide and silver nanoparticles: a novel application of artificial neural network based simulation in MATLAB, Chemosphere 308 (2022) 136304, https://doi.org/10.1016/j.chemosphere.2022.136304.

[94] Ç. Odabaşi, P. Dölöğlu, F. Gülmez, G. Kuşoğlu, Ö. Çağlar, Machine Learning Analysis of the Feed Water Parameters Affecting Reverse Osmosis Membrane Operation, 2021, pp. 235–240, https://doi.org/10.1016/B978-0-323-88506-5.50038-3.

[95] H. Gao, S. Zhong, R. Dangayach, Y. Chen, Understanding and designing a high-performance ultrafiltration membrane using machine learning, Environ. Sci. Technol. (2023), https://doi.org/10.1021/acs.est.2c05404.

[96] M. Fetanat, M. Keshtiara, Z.-X. Low, R. Keyikoglu, A. Khataee, Y. Orooji, V. Chen, G. Leslie, A. Razmjou, Machine learning for advanced design of nanocomposite ultrafiltration membranes, Ind. Eng. Chem. Res. 60 (2021) 5236–5250, https://doi.org/10.1021/acs.iecr.0c05446.

[97] D. Paul, A.K. Goswami, R.L. Chetri, R. Roy, P. Sen, Bayesian optimization-based gradient boosting method of fault detection in oil-immersed transformer and reactors, IEEE Trans. Ind. Appl. 58 (2022) 1910–1919, https://doi.org/10.1109/TIA.2021.3134140.

[98] A. Hosseinzadeh, M. Baziar, H. Alidadi, J.L. Zhou, A. Altaee, A.A. Najafpoor, S. Jafarpour, Application of artificial neural network and multiple linear regression in modeling nutrient recovery in vermicompost under different conditions, Bioresour. Technol. 303 (2020) 122926, https://doi.org/10.1016/j.biortech.2020.122926.

[99] F.H.M. Salleh, S. Zainudin, S.M. Arif, Multiple linear regression for reconstruction of gene regulatory networks in solving cascade error problems, Adv. Bioinformatics. 2017 (2017) 1–14, https://doi.org/10.1155/2017/4827171.

[100] W. Cao, X. Wang, Z. Ming, J. Gao, A review on neural networks with random weights, Neurocomputing 275 (2018) 278–287, https://doi.org/10.1016/j.neucom.2017.08.040.

[101] S. Manzhos, T. Carrington, Neural network potential energy surfaces for small molecules and reactions, Chem. Rev. 121 (2021) 10187–10217, https://doi.org/10.1021/acs.chemrev.0c00665.

[102] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, A. Lopez, A comprehensive survey on support vector machine classification: applications, challenges and trends, Neurocomputing 408 (2020) 189–215, https://doi.org/10.1016/j.neucom.2019.10.118.

[103] V.G. Maltarollo, T. Kronenberger, G.Z. Espinoza, P.R. Oliveira, K.M. Honorio, Advances with support vector machines for novel drug discovery, Expet Opin. Drug Discov. 14 (2019) 23–33, https://doi.org/10.1080/17460441.2019.1549033.

[104] W.L. Chu, C.J. Lin, K.N. Chang, Detection and classification of advanced persistent threats and attacks using the support vector machine, Appl. Sci. 9 (2019) 4579, https://doi.org/10.3390/app9214579.

[105] H. Zhou, J. Zhang, Y. Zhou, X. Guo, Y. Ma, A feature selection algorithm of decision tree based on feature weight, Expert Syst. Appl. 164 (2021) 113842, https://doi.org/10.1016/j.eswa.2020.113842.

[106] S.J. Park, C.W. Lee, S. Lee, M.J. Lee, Landslide susceptibility mapping and comparison using decision tree models: a case study of Jumunjin area, Korea, Rem. Sens. 10 (2018) 1545, https://doi.org/10.3390/rs10101545.

[107] J.J. Klemeš, Y. Van Fan, P. Jiang, Plastics: friends or foes? The circularity and plastic waste footprint, Energy Sources, Part A Recover, Util. Environ. Eff. 43 (2021) 1549–1565, https://doi.org/10.1080/15567036.2020.1801906.

[108] Y. Ao, L. Zhu, S. Guo, Z. Yang, Probabilistic logging lithology characterization with random forest probability estimation, Comput. Geosci. 144 (2020) 104556, https://doi.org/10.1016/j.cageo.2020.104556.

[109] J. Magidi, L. Nhamo, S. Mpandeli, T. Mabhaudhi, Application of the random forest classifier to map irrigated areas using google earth engine, Rem. Sens. 13 (2021) 876, https://doi.org/10.3390/rs13050876.

[110] G.I. Parisi, R. Kemker, J.L. Part, C. Kanan, S. Wermter, Continual lifelong learning with neural networks: a review, Neural Network. 113 (2019) 54–71, https://doi.org/10.1016/j.neunet.2019.01.012.

[111] X. Wang, J. Zhang, V. Babovic, K.Y.H. Gin, A comprehensive integrated catchment-scale monitoring and modelling approach for facilitating management of water quality, Environ. Model. Software 120 (2019) 104489, https://doi.org/10.1016/j.envsoft.2019.07.014.

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

[113] M.S. Bhatti, D. Kapoor, R.K. Kalia, A.S. Reddy, A.K. Thukral, RSM and ANN modeling for electrocoagulation of copper from simulated wastewater: multi objective optimization using genetic algorithm approach, Desalination 274 (2011) 74–80, https://doi.org/10.1016/j.desal.2011.01.083.

[114] B. Li, R. Yue, L. Shen, C. Chen, R. Li, Y. Xu, M. Zhang, H. Hong, H. Lin, A novel method integrating response surface method with artificial neural network to optimize membrane fabrication for wastewater treatment, J. Clean. Prod. 376 (2022) 134236, https://doi.org/10.1016/j.jclepro.2022.134236.

Panchan Dansawad is a Ph.D. candidate at CAS Key Laboratory of Green Process and Engineering, Institute of Process Engineering, Chinese Academy of Sciences, Beijing, China, and University of Chinese Academy of Sciences, Beijing, China.

Yanxiang Li is a Professor at CAS Key Laboratory of Green Process and Engineering, Institute of Process Engineering, Chinese Academy of Sciences, Beijing, China, and University of Chinese Academy of Sciences, Beijing, China. Contact: yxli@ipe.ac.cn

Yize Li is a PhD candidate at James Watt School of Engineering, University of Glasgow, Glasgow, UK.

Jingjie Zhang is a Professor at NUS-Environmental Research Institute, National University of Singapore, Singapore.

Siming You is a Professor at James Watt School of Engineering, University of Glasgow, Glasgow, UK. Contact: Siming.You@glasgow.ac.uk

Wangliang Li is a Professor at CAS Key Laboratory of Green Process and Engineering, Institute of Process Engineering, Chinese Academy of Sciences, Beijing, China, and University of Chinese Academy of Sciences, Beijing, China. Contact: +86-10-82544976, wlli@ipe.ac.cn

Shouliang Yi is a Professor at U. S. Department of Energy National Energy Technology Laboratory, Pittsburgh, USA. Contact: shouliang.yi@hotmail.com