Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2023.29328

286 (2023) 64–72
February

Towards machine learning in water treatment: a diagnostic tool for assessing water quality

Jaydev Zaveri[a], Shankar Raman Dhanushkodi[a,*], Lalit Bansal[b]

[a]*Department of Chemical Engineering, Vellore Institute of Technology, Katpadi, India, Tel.: +91 9626845903; email: shankarraman.d@vit.ac.in*
[b]*School of Mechanical Engineering, Vellore Institute of Technology, Katpadi, India, Tel.: +91 9945167343; email: lalit.bansal@vit.ac.in*

Received 13 September 2022; Accepted 24 January 2023

ABSTRACT

Saltwater from the ocean constitutes 96.5% of the total water available on the Earth. The remaining 3.5% is freshwater, which is a crucial resource for living organisms. Increasing population and activities related to climate change have led researchers to develop new methods to maximize freshwater resources. Solar desalination is an environment-benign method that can fulfil the requirement of freshwater. However, the efficiency of desalination cells is limited by the fouling phenomenon. The efficiency of the desalting process decreases because the pores of the membrane are clogged by fouling. Therefore, methods for detecting and diagnosing the fouling phenomenon by using mathematical models are required. We propose a machine learning modelling framework comprising of K-nearest neighbor, random forest, artificial neural network, and support vector machine algorithms to monitor the onset of fouling in desalination cells individually. Permeate datapoints from the filtration process were collected using a lab on a chip device. The datapoints were used to validate all four models. Furthermore, model responses for permeate data points were used as an indicator or soft sensor to grade the fouling level and potability of the treated water. The modelling framework can be used to detect the onset of fouling and erosion in desalination cells with high precision.

Keywords: Predictive maintenance; Machine learning model; Fouling; Soft sensor

1. Introduction

The majority of the Earth's 11.5 million m^3 of freshwater is in the form of glaciers. Water reservoirs, such as ponds, lakes, and underground aquifers, contain less than 5 million cubic miles of freshwater. Approximately 1.74 million species in both terrestrial and marine environments require freshwater to survive. The irregular supply of drinking water is becoming a major problem. Major water sources, such as oceans and seas, cannot be used to fulfil the demand for drinking water [1]. Moreover, disruptive weather patterns caused by climate change have resulted in extreme conditions that are exacerbating water scarcity and polluting water sources. Ingestion of sea water can be harmful for the millions of species, including humans. Although the kidneys can remove salt from water, the processing of salted water is substantially limited by cellular tissues and kidney functioning. Therefore, producing fresh water from seawater is essential for the humankind. Most water treatment plants use state-of-the-art membrane filtration system that captures undesirable particles that build up with salt components.

Desalination by reverse osmosis (RO) has become ubiquitous for producing freshwater from salt water. The operating cost of RO is lower than that of alternative methods, such as multistage flash. Pre-treated brackish water is pumped at high pressure through a thin microporous membrane composed of polyamide rolls. Water molecules flow through the membrane and are purified. Potable water is collected on the permeate side, whereas salt is retained at the feed side

* Corresponding author.

of the membrane. Many plants have been designed to provide million liters of freshwater based on the daily demand. The capital and maintenance cost of plants is high due to energy lost through evaporation. Adopting energy reuse strategies and installing state-of-the-art membranes that can withstand high pressure can reduce cost. Because the membrane is the integral component of the desalination process, fouling or erosion of the membrane reduces process efficiency. Microbial and bacterial adhesion, solute adhesion, and gel-layer formation are some of the predominant examples of fouling [2]. Bacterial adhesion is the severe form of membrane fouling because it requires chemicals to clean biofilms. Operating cells with fouling cause a decline in pressure across the membrane and thus reduces the quality of the permeate flux and increases power consumption.

Apart from fouling, process efficiency is affected by the degradation of membranes caused by the addition of disinfectants such as chlorine. Disinfectants are mixed with feed water to remove water hardness. Chlorine in the disinfectant reacts with the polyamide membrane and erodes it. To differentiate fouling from erosion, membrane autopsy is required. In this process, a small component of the membrane is removed from the membrane module and analysed using various techniques, such as nuclear magnetic resonance, scanning electron microscopy, and confocal laser scanning microscopy [3]. These techniques are invasive and require sophisticated equipment and trained personnel. The total cost incurred to produce desalinated water from saltwater has considerably increased in the last decade because of the high energy consumption and capital cost of the membrane. Increasing membrane efficiency and optimizing desalination cell components can reduce OEM spending in the plant. Thus, the mathematical modelling of the desalination process can help in understanding and predicting the fouling process.

Robust mathematical models are required to examine the characteristics of various processes. Although two-dimensional mechanistic models provide details regarding the interaction between the membrane and water interfaces [4–6], models that can account for fouling mechanisms in industry-level plants [7,8] are yet to be developed. Factors such as changes in the feed composition, fouling mechanism, and diurnal variations should be included in deterministic or machine learning models. Thus, nonlinear equations are required to model desalination processes. These models should account for the decline in pressure and membrane permeability data so that it can auto adapt to changes in conditions [9]. Developing a data-driven model or its algorithms can help to determine the corroboration behaviour of the plant. However, this approach needs information on plant history with process parameters [10]. Table 1 summarises data-driven models used to predict the performance and fouling of the plant. Several studies have used both steady and unsteady state models to predict or auto adjust plant parameters on the basis of extensively collected operational data. However, developing diagnosis systems to control the plant requires a comparison of various ML algorithms that can accurately assess or interfere with plant functioning. Such approach can be useful for developing model-based fault management techniques. This study (a) compared among four ML-based data-driven models to diagnose fouling based on the properties of permeates in a water desalination

Table 1
Literature findings

Models and method	Findings and research need	References
Lab-on-a-chip	Reports data related to chemical and biological contaminants and water quality parameters using a lab on a chip.	[11]
Lab-on-a-chip	Reviews various lab on chip systems in this research, which are cost-effective, free of fouling, and clogging problems.	[12]
Deterministic models of RO plant behaviour	- Provides fundamental understanding on how to run desalination plants; - Reports the working pattern of sensors and actuators must be captured; - Uses predictive models to explain membrane fouling and scaling.	[13]
Mechanistic models	- Reports flow behavior across membrane channel geometries using governing equations; - Correlate the fouling with flow dynamics; - Accounts plant hydraulics to assess fouling.	[14,16]
Data driven models	- Describe nonlinearity in the desalination system; - Effective to describe the plant behavior; - Obtains process parameters from the deterministic model; - Models need to be trained and auto-adjusted to operating conditions; - Must be integrated with a control system.	[7,15–17]
Supervised machine learning	- Develops protocols to assess water quality; - Demonstrates how to estimate the water quality index.	[18]
Artificial neural networks	Predicts water quality.	[19]
SVM, artificial neural networks	Compares the accuracy of different models for water quality.	[6,20]
Membrane fouling	Discusses factors affecting fouling and their diagnostic and mitigation techniques.	[21]

plant under different operating conditions; (b) used the fault detection approach to identify factors causing system failure induced by fouling; (c) determine how different models react to fouling when the properties of permeate water change in the desalination cell.

2. Materials and methods

2.1. Data collection

Data for training, testing, and validating the ML model were collected from the open-source database [22]. Python v.2.10.8 software was used to develop all the machine learning models developed in this study. The dataset includes pH, hardness, total dissolved solids, sulphates, conductivity, trihalomethanes, and turbidity as key features. The label for our model was the measure of potability in terms of binary 1 and 0, where 1 indicated potable and 0 indicated non-potable. The ratio of the training to testing dataset was 80:20.

2.1.1. Data pre-processing and scaling

The dataset collected had missing data points. These points were filled using the mean of that column instead of removing those rows. Removing those rows would have reduced the size of our dataset, thus affecting our results. The data were scaled using standard-scaling function, which was used to find the mean of features and scales to unit variance.

2.2. Machine learning model

Modelling scheme and methodology adopted is given in Fig. 1. Permeate water properties were used to formulate the model. The rationale for the formulation is given:

- Properties of feed water are not consistent. Therefore, we have not considered the feed water characteristics in the model formulation.
- We have used the potability of the water as a soft sensor or indicator to assess the fouling which eliminates the need for input data related to fouling.

2.2.1. Support vector machine

The classifier was generated using the Lagrangian Eq. (1).

$$L = \frac{1}{2}|w|^2 - \sum_i \alpha_i \left(Y_{\text{SVi}} \left(\vec{w} \cdot \overrightarrow{X_{\text{SVi}}} + b \right) - 1 \right) \tag{1}$$

where w vector is the weight, b vector is the bias, X_{SVi} are the features, α_i is the Lagrange multiplier, and Y_{SVi} are the labels. The aforementioned equation was deducted to formulate the best-fit hyperplane. LSVM has an inbuilt kernel function that can transform the vector space to feature space.

$$K(X_1, X_2) = \exp\left(-\frac{\|X_1 - X_2\|^2}{2\sigma^2} \right) \tag{2}$$

where σ is the variance and the hyperparameter $||X_1 - X_2||$ is the Euclidean distance between the data points X_1 and X_2. The hyperplane is created in support vector machine on the basis of the Lagrangian equation [23]. The aim is to maximize the distance between the support vector and hyperplane such that the maximum separation of features is possible. Support vectors are imaginary (dotted/dashed) lines and pass through the boundary. In our binary classification model, the two classes are mainly binary 0 (non-potable) and binary 1 (potable); for simplicity, we can consider them as positive and negative sides, respectively. The positive side can be used to identify the most favourable features associated with the potability measurement. When a dataset is treated in batches, we can treat data points accumulated below the hyperplane or on the negative side of the plane as non-potable water. Once this hyperplane is created, test points are categorized on the basis of the location of these points in respect to the hyperplane. This is known as the predicted value of our model (the output provided by our model for the particular set of input features from the test set). Fig. 2 presents the working of the SVM model. The predicted value is compared with the true value (actual value), which we already had determined for the given test set, and its accuracy was calculated [18].

2.2.2. K-nearest neighbor

K-nearest neighbor (K-NN) is a supervised learning algorithm where learning is based on how similar the data are.

$$d = \sqrt{\left(\left(x_i - x_j \right)^2 + \left(y_i - y_j \right)^2 \right)} \tag{3}$$

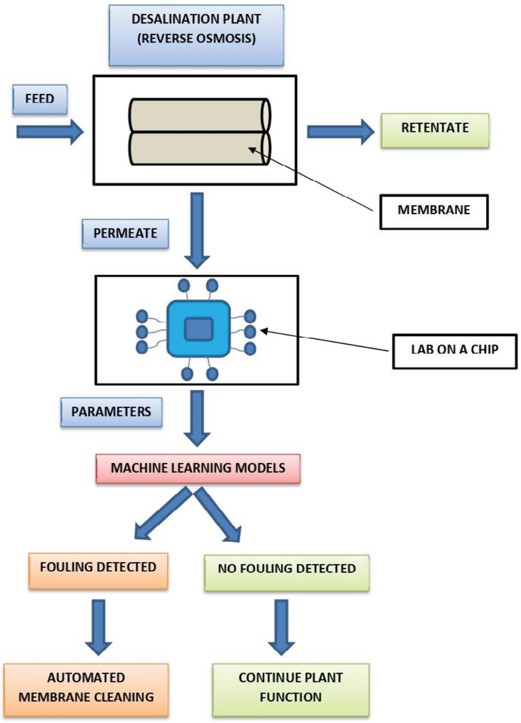
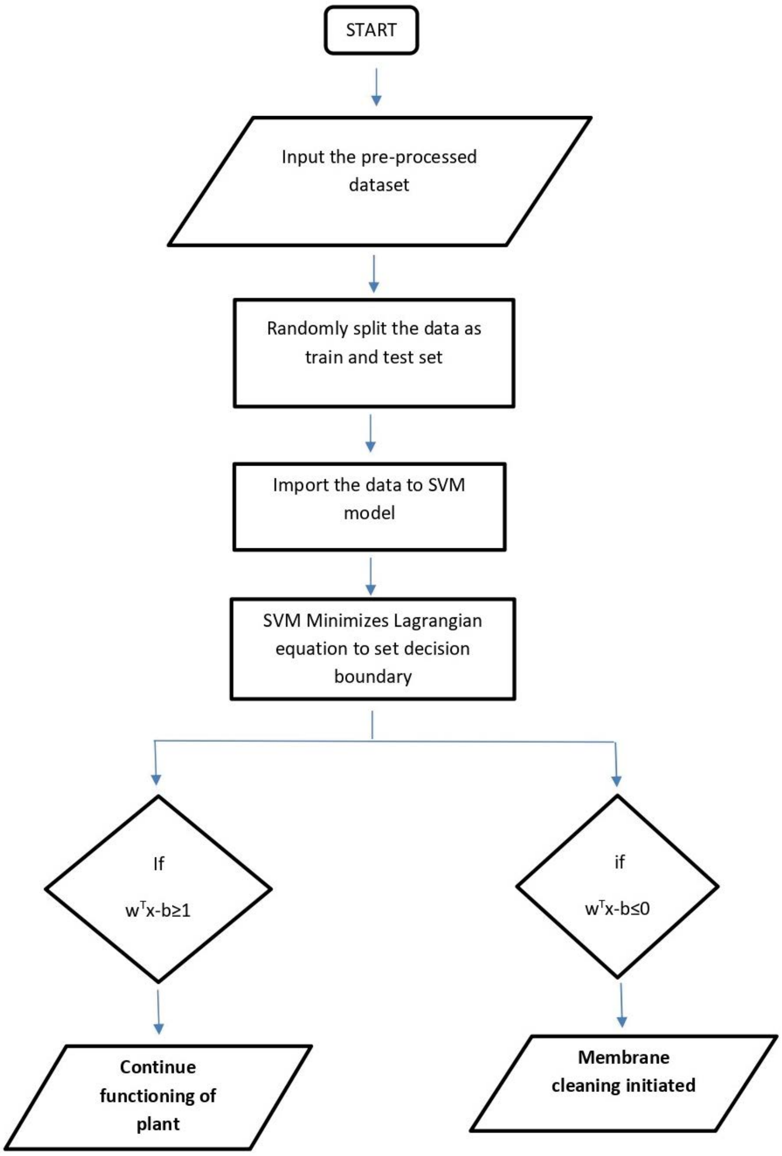


Fig. 1. Modelling scheme and methodology adopted.


START
Input the pre-processed dataset
Randomly split the data as train and test set
Import the data to SVM model
SVM Minimizes Lagrangian equation to set decision boundary
If $w^Tx-b\geq1$
if $w^Tx-b\leq0$
Continue functioning of plant
Membrane cleaning initiated

Where w^T is the weight and b is the bias and X are the datapoints (features).

Fig. 2. SVM algorithm.

where x_i and y_i are the feature and label, respectively, of unclassified points and x_j and y_j are the feature and label, respectively, of classified points. The Euclidean distance is the most common method used to determine the point on the basis of the minimum distance between the unclassified and classified points. The unclassified point belongs to the class that appears the most (defined by parameter *K*) [24].

2.2.3. Random forest

A random forest comprises many individual decision trees that work collectively as an ensemble [25].

The technique of combining several models is known as an ensemble. Various models are used to produce forecasts instead of only one model. Each tree in the random forest produces a class prediction. The class with the most votes is the model's prediction. Bagging and boosting are the two basic techniques for an ensemble; these techniques are used to transform poor learners into strong learners by developing continuous models with the best accuracy as the final model. In our model, we employed bagging. The branching of nodes in the decision tree was determined using entropy whose equation is provided as follows:

$$\sum_{i=1}^{C} -p_i \times \log_2\left(p_i\right) \qquad (4)$$

where p_i is the frequentist probability of an element/class *i* in data. In a random forest, *n* random records are selected from the dataset of k records. Then, for each sample, a distinct decision tree is created, generating a unique output. The result is based on the majority voting or averaging, respectively. The higher the number of trees (in the forest) is, the higher is the accuracy; thus, the problem of overfitting can be avoided.

2.2.4. *Artificial neural network*

An artificial neural network is composed of numerous nodes that are identical to actual neurons in the brain [26]. The neurons are linked and interact with each other. Nodes can receive input data and perform simple operations. These activities produce a result, which is then transferred on to other neurons. Each node's output is indicated as its node or activation value. An ANN consists of units and connections between units. The output of unit i is the input to unit j. Unit i is considered the predecessor of j. Each connection is assigned a weight w_{ij}. Each node has an activation threshold. The negation of the threshold is termed bias (b for node $j = b_j$). The weighted inputs are summed together with the bias.

$$\sum w_{ij} \times a_i + b_j > 0 \tag{5}$$

The output of unit j is calculated as:

$$a_j = f\left(w_{ij} \times a_i + b_j\right) \tag{6}$$

where f is the Relu activation function.

The predicted results are compared with actual data, and the error is measured. This error is backpropagated, and weights are adjusted depending on how much they contribute to inaccuracy. We update weights on the basis of the learning rate [19].

2.3. *ML procedures*

The dataset was pre-processed, which involved filling missing datapoints and scaling the dataset. After pre-processing, a heatmap was plotted to determine the correlation. Feature selection was performed, which involved removing columns. Then, the dataset was split in an 80:20 ratio, where 80% of randomly selected data were used as the training set for the models and the remaining 20% were used as the testing set. Hyperparameters and kernel function of the models were finetuned by using the GridSearchCV function, and the best fit values were inputted in the model. These models were trained using the training set. Finally, the accuracy of the models was calculated using the test set. The accuracy of all the four models (SVM, KNN, random forest, and ANN) was calculated using various matrices, namely confusion matrix, precision recall, and receiver operating characteristics. The results of these models are presented in Table 2, and the accuracy was compared among the four models.

3. Results and discussion

3.1. *Comparison of four ML-based data-driven models*

The values of activation functions and hyperparameters were fed, and from those, the best value was determined and used for building the model. The radial base function (RBF) was used as the kernel for the SVM model. For the KNN model, 35 nearest neighbors were selected, and the Euclidean distance metrics were used. A total of 500 trees were used for the random forest method. The sequential modular approach was employed for the ANN model. The number of parameters for input, hidden 1, and hidden 2 layers was 52,130. Relu was selected as the activation function for input and hidden layers and SoftMax for the output layer. Adam's optimizer was used, and binary cross-entropy loss was determined. To evaluate the accuracy of ML models, various matrices (measures) were used, namely the F1-score, accuracy, recall, precision, and receiver operating characteristic (ROC) curve [27]. The model's accuracy can be calculated as the number of predictions made by the model over observed values.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{7}$$

Precision can be calculated as the proportion of the accurately classified instances of a positive class out of the total classified instances of that class.

$$\text{Precison} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

The results of precision and recall ensure that the fractions of actual positives are classified accurately. The proportion of the instances of a specific positive class that was correctly identified is known as recall or sensitivity.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

True positive (TP), false positive (FP), false negative (FN), and true negative (TN) values were calculated using the test dataset. Specificity refers to the number of instances of a specific negative class that were accurately identified. The F1 score is the harmonic mean of precision and recall because precision and recall do not cover all the aspects of accuracy. Table 2 presents the comparison of all the four models.

The random forest model outperformed among all the models because instead of searching for the most crucial dataset feature, it searches for the best feature in a given random subset of features. Additional randomness is thus added to the model, making it more robust. The accuracy of the random forest model was the highest, followed by

Table 2
Comparison of all four models

Model	Accuracy	Precision	Recall	F1 score
SVM	0.70	0.68	0.63	0.64
KNN	0.66	0.66	0.56	0.54
Random forest	0.88	0.89	0.84	0.86
ANN	0.70	0.68	00.66	0.67

that of the SVM, KNN, and ANN models. The ROC curve can be used to measure the separability of a model and to determine how accurately a model can quantify between the water being potable and non-potable in our case. The trade-off between the TP and FP rates as the criterion for positivity is changed. The concave nature of the curve (Fig. 4) can be due to the monotonically increasing likelihood ratio (distribution of the separator variable in potable and non-potable water) [28]. The area under the curve is the combined measure of specificity and sensitivity that

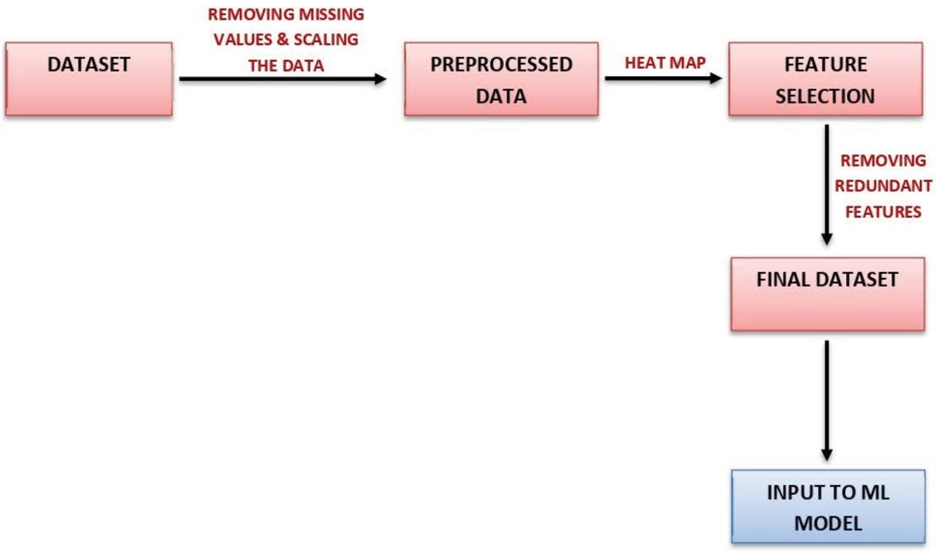


Fig. 3. Schematic of modelling and data analysis approaches employed to investigate the use of ML methods.

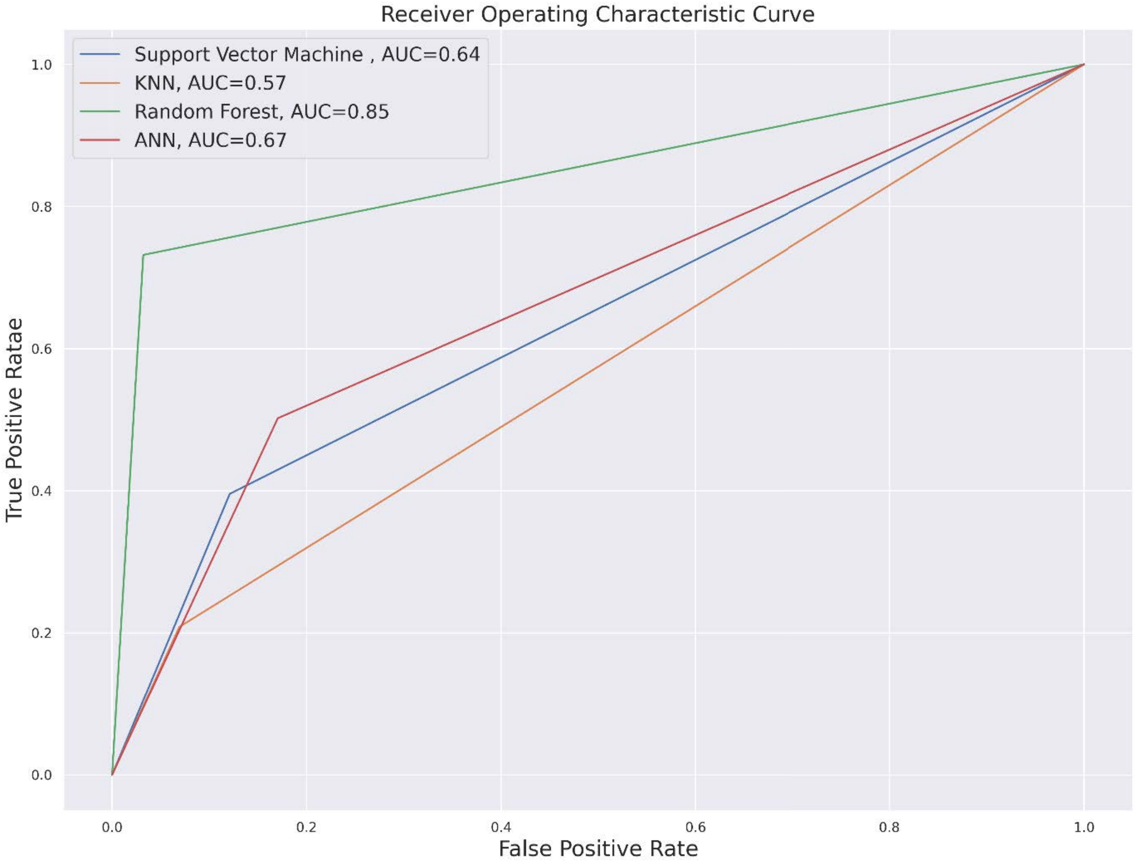


Fig. 4. Receiver operating characteristic curves for all the models.

indicates whether water is potable. The greater the area under the ROC curve is, the better is the model. The area under the curve for the random forest model was the highest among all the models. A higher value of the area under the curve represents higher separability, indicating the suitability of model classification.

As presented in Fig. 4, the random forest model had a greater discrimination capacity than did the other models. The random forest was the optimal model because it has the highest accuracy and required less computational time.

3.2. Use of the fault detection approach to determine factors causing system failure induced by fouling

To examine the fault, one of the four models can be employed to obtain the property matrix. We adopted the support vector classification (SVC) method. The property matrix obtained from the dataset using the SVC model is presented in Fig. 5. The precision recall value (Table 2) and ROC-AUC are plotted in Fig. 4 to indicate their correlation with the property matrix. Because the SVC model is based on structural risk minimization, all data points converge near the local minima and prevent overfitting (Fig. 5). The main aim of the SVC model is to determine the deviation in the function (caused by fouling) with respect to changes in permeate conductivity, such as TDS and other parameters. Thereby this model can be used as the precursor to minimize or eliminate the use of the anti-scalant chemicals when the membrane shows no sign of fouling. This will reduce the OEM cost and increase the durability of the membrane. Potability can be effectively gauged using the predictive data science model using only the parameters that are readily available, that is, already measured without the need of any additional tests.

A kernel function (rbf in our case) was used to map nonlinear to linear trends for converting the vector space to feature space that can be separated using a hyperplane. Nonlinear data points related to properties can be obtained by mapping process variables. Fouling could be identified by changes in the hardness of water and the number of solids present in the feed and permeate water. Fig. 5 predicts the fouling with an acceptable error. A nonlinear optimization method was used to maximize the accuracy to 70%. Slack variables were added within the SVC formula to minimize the error.

3.3. Determine how different models react to fouling when the properties of permeate water change in the desalination cell

Our data-driven models detected fouling when changes in modelling parameters, namely pH, conductivity, hardness, solid, surface, and carbon content, were not in accordance with water quality standards (IS 10500:2012). To determine the potability of water, we use colour coding in

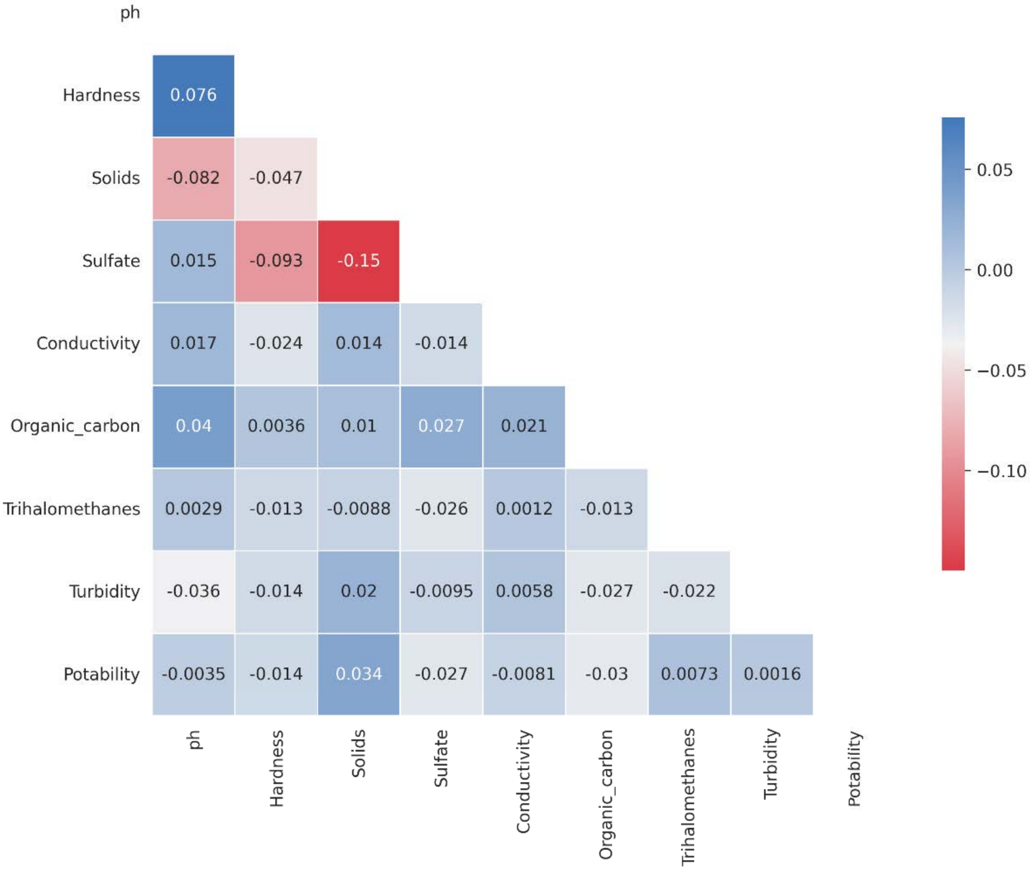


Fig. 5. Property matrix.

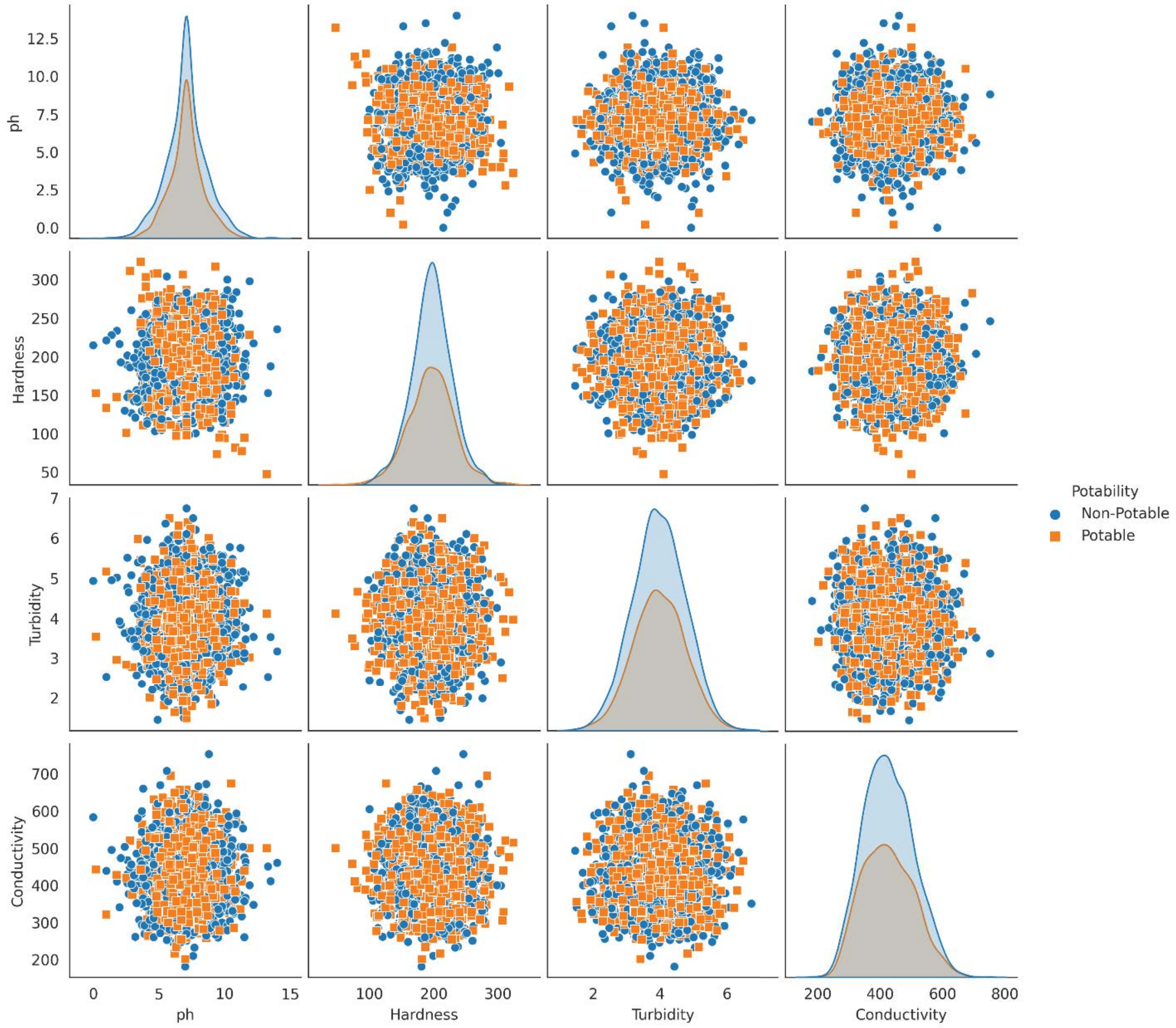


Fig. 6. Identifying the portability of water when fouling occurred.

Fig. 6. Although all the four modelling algorithms were applied to evaluate the deductive nature of individual models with respect to permeate parameters, the random forest model exhibited less average absolute relative errors (Table 2). This model provided a long-term response to notify potability but did not reveal any details related to a short or sudden surge in the feed flow rate. We will focus on this in our future study and detect faults in operating plants in real time.

4. Conclusion

This study demonstrates the development of supervised learning data-driven models by using the SVM, KNN, random forest, and ANN modelling approaches. Models were developed on the basis of permeate property data obtained from the literature. The random forest model developed for fouling by using pH, solid, carbon, and conductivity of permeate streams as features exhibited a high accuracy (88%). The random forest model could predict the potability of water with a precision and recall score of 0.89 and 0.84, respectively. The F1 score of the random forest model suggests that property-based performance forecasting models can be used as a diagnosis tool for desalination plants. Furthermore, the model can help to purge faulty data points. The data-driven models used in the study are directly applicable to distributed control systems in the plant, and strategies employed for model training can yield higher performance when new permeate property data are applied. However, the models cannot predict the retentate properties. The model uses classifiers to assess the potability using the permeate water characteristics. This model can further be improved if fouling layer formation over time is included as a parameter. Moreover, a time series model can more accurately predict membrane fouling and erosion.

Acknowledgements

The authors would like to acknowledge the contribution provided by VIT undergraduate interns: G Pareek, M Singh, S Naik, D Kabra, K Kumari and Dr. Asha Srinivasan, University of British Columbia, Vancouver, Canada. Lalit Bansal would like to acknowledge funding from DST SERB via file number SRG/2020/002159.

Abbreviation

ANN	—	Artificial neural network
AUC	—	Area under curve
FN	—	False negative
FP	—	False positive
KNN	—	K-nearest neighbor
LSVM	—	Lagrangian support vector machine
ML	—	Machine learning
OEM	—	Original equipment manufacturer
RBF	—	Radial basis function
RO	—	Reverse osmosis
ROC	—	Receiver operating characteristic
SVM	—	Support vector machine
TDS	—	Total dissolved solids
TN	—	True negative
TP	—	True positive

References

[1] W.A. Jury, H.J. Vaux Jr., The emerging global water crisis: managing scarcity and conflict between water users, Adv. Agron., 95 (2007) 1–76.

[2] N. AlSawaftah, W. Abuwatfa, N. Darwish, G. Husseini, A comprehensive review on membrane fouling: mathematical modelling, prediction, diagnosis, and mitigation, Water, 13 (2021) 1327, doi: 10.3390/w13091327.

[3] H.F. Ridgway, A. Kelly, C. Justice, B.H. Olson, Microbial fouling of reverse-osmosis membranes used in advanced wastewater treatment technology: chemical, bacteriological, and ultrastructural analyses, Appl. Environ. Microbiol., 45 (1983) 1066–1084.

[4] G. Belfort, R.H. Davis, A.L. Zydney, The behavior of suspensions and macromolecular solutions in crossflow microfiltration, J. Membr. Sci., 96 (1994) 1–58.

[5] M.F.A. Goosen, S.S. Sablani, D. Jackson, Fouling of reverse osmosis and ultrafiltration membranes: a critical review, Sep. Sci. Technol., 39 (2005) 2261–2297.

[6] S. Shirazi, C.-J. Lin, D. Chen, Inorganic fouling of pressure-driven membrane processes — a critical review, Desalination, 250 (2010) 236–248.

[7] Q.-F. Liu, S.-H. Kim, Evaluation of membrane fouling models based on bench-scale experiments: a comparison between constant flowrate blocking laws and artificial neural network (ANNs) model, J. Membr. Sci., 310 (2008) 393–401.

[8] S. Gray, R. Semiat, M.C. Duke, A. Rahardianto, Y. Cohen, Seawater Use and Desalination Technology, In: Treatise on Water Science, Elsevier, 2011, pp. 73–109.

[9] J.-L. Dirion, M. Cabassud, M.V. Le Lann, G. Casamatta, Development of adaptive neural networks for flexible control of batch processes, Chem. Eng. J. Biochem. Eng. J., 63 (1996) 65–77.

[10] W. Richard Bowen, M.G. Jones, J.S. Welfoot, H.N.S. Yousef, Predicting salt rejections at nanofiltration membranes using artificial neural networks, Desalination, 129 (2000) 147–162.

[11] A. Kapoor, S. Balasubramanian, P. Muthamilselvi, V. Vaishampayan, S. Prabhakar, Lab-on-a-Chip Devices for Water Quality Monitoring, Inamuddin, A. Asiri, Eds., Nanosensor Technologies for Environmental Monitoring. Nanotechnology in the Life Sciences, Springer, Cham, 2020.

[12] A. Jang, Z. Zou, K.K. Lee, C.H. Ahn, P.L. Bishop, State-of-the-art lab chip sensors for environmental water monitoring, Meas. Sci. Technol., 22 (2011) 032001, doi: 10.1088/0957-0233/22/3/032001.

[13] X. Pascual, H. Gu, A.R. Bartman, A. Zhu, A. Rahardianto, J. Giralt, R. Rallo, P.D. Christofides, Y. Cohen, Data-driven models of steady state and transient operations of spiral-wound RO plant, Desalination, 316 (2013) 154–161.

[14] A. Abdelrasoul, H. Doan, A. Lohi, A mechanistic model for ultrafiltration membrane fouling by latex, J. Membr. Sci., 433 (2013) 88–99.

[15] N. Peña, S. Gallego, F. del Vigo, S.P. Chesters, Evaluating impact of fouling on reverse osmosis membranes performance, Desal. Water Treat., 51 (2012) 958–968.

[16] B. Gu, X.Y. Xu, C.S. Adjiman, A predictive model for spiral wound reverse osmosis membrane modules: the effect of winding geometry and accurate geometric details, Comput. Chem. Eng., 96 (2017) 248–265.

[17] R. Rivas-Perez, J. Sotomayor-Moriano, G. Pérez-Zuñiga, M.E. Soto-Angles, Real-time implementation of an expert model predictive controller in a pilot-scale reverse osmosis plant for brackish and seawater desalination, Appl. Sci., 9 (2019) 2932, doi: 10.3390/app9142932.

[18] U. Ahmed, R. Mumtaz, H. Anwar, A.A. Shah, R. Irfan, J. García-Nieto, Efficient water quality prediction using supervised machine learning, Water, 11 (2019) 2210, doi: 10.3390/w11112210.

[19] G. Hayder, I. Kurniawan, H.M. Mustafa, Implementation of machine learning methods for monitoring and predicting water quality parameters, Biointerface Res. Appl. Chem., 11 (2021) 9285–9295.

[20] A.H. Haghiabi, A.H. Nasrolahi, A. Parsaie, Water quality prediction using machine learning methods, Water Qual. Res. J., 53 (2018) 3–13.

[21] N. AlSawaftah, W. Abuwatfa, N. Darwish, G. Husseini, A comprehensive review on membrane fouling: mathematical modelling, prediction, diagnosis, and mitigation, Water, 13 (2021) 1327, doi: 10.3390/w13091327.

[22] A. Kadiwal, Water Quality, Kaggle, 25 Apr. 2021, Available at: https://www.kaggle.com/datasets/adityakadiwal/water-potability

[23] Y. Zhang, Support Vector Machine Classification Algorithm and Its Application, C. Liu, L. Wang, A. Yang, Eds., Information Computing and Applications, ICICA 2012, Communications in Computer and Information Science, Vol. 308, Springer, Berlin, Heidelberg, 2012, pp. 179–186. Available at: https://doi.org/10.1007/978-3-642-34041-3_27

[24] K. Taunk, S. De, S. Verma, A. Swetapadma, A Brief Review of Nearest Neighbor Algorithm for Learning and Classification, 2019 International Conference on Intelligent Computing and Control Systems (ICCS), IEEE, Madurai, India, 2019, pp. 1255–1260. Available at: https://doi.org/10.1109/ICCS45141.2019.9065747

[25] P.O. Gislason, J.A. Benediktsson, J.R. Sveinsson, Random Forest Classification of Multisource Remote Sensing and Geographic Data, IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium, IEEE, Anchorage, AK, USA, 2004, pp. 1049–1052. Available at: https://doi.org/10.1109/IGARSS.2004.1368591

[26] K.P. Singh, A. Basant, A. Malik, G. Jain, Artificial neural network modeling of the river water quality—a case study, Ecol. Modell., 220 (2009) 888–895.

[27] J. Davis, M. Goadrich, The Relationship Between Precision-Recall and ROC Curves, Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006, pp. 233–240. Available at: https://doi.org/10.1145/1143844.1143874

[28] J.N. Mandrekar, Receiver operating characteristic curve in diagnostic test assessment, J. Thoracic Oncol., 5 (2010) 1315–1316.