Arabian Journal of Chemistry (2022) **15**, 104306

King Saud University

Arabian Journal of Chemistry

www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Hybrid mechanistic approach in the estimation of flow properties in cylindrical membrane modules

Fang Peng [a,*], **Wei Yang** [a], **Yi Liu** [b]

[a] *Department of Mechanical and Electrical Engineering, Inner Mongolia Vocational College of Chemical Engineering, Hohhot, Inner Mongolia 010000, China*
[b] *Hangzhou Xinya Low Temperature Technology Co., Ltd, Hangzhou, Zhejiang 310000, China*

Received 26 February 2022; accepted 24 September 2022
Available online 29 September 2022

KEYWORDS

CFD;
Computational simulation;
Fluid flow;
Separation;
Machine learning;
Environmental

Abstract A hybrid computational approach was employed for simulation of molecular separation using polymeric membranes. The considered system is a cylindrical membrane module in which the mass transfer equations were solved numerically using CFD (Computational Fluid Dynamics) to obtain the concentration of the species, and then the simulation results were used in machine learning models. Indeed, the CFD simulation results were used as the inputs for several machine learning models to obtain the hybrid model. We have a dataset with more than 2000 data points and two input features (r and z). Also, the only output is C which is the concentration of the species in the feed channel of membrane module. KNN (K nearest neighbor), PLSR (Partial Least Square Regression), and SGD (Stochastic Gradient Descent) are the models employed in this research to analyze the mentioned data set. Models were optimized with their hyper-parameters and finally evaluated with different statistical metrics. MAE error metric is 3.4, 5.1, and 5.5 for KNN, SGD, and PLSR. Also, they have 0.998, 0.997, 0.896 coefficient of determination (R^2) respectively. Finally, based on the overall results, KNN with K = 8 is selected as the best model in this study for simulation of the membrane system. The final maximum error is also 1.35E+02.

1. Introduction

Separation and removal of water pollutants is a great subject of interest which can be done using molecular separation methods via different separation techniques. From the environmental perspective, pollutant molecules must be removed from effluents prior to discharge to the environment, as they might cause severe health problems (Fan et al., 2022; Ge et al., 2019; Guan et al., 2021; Liu et al., 2020, 2022; Bai et al., 2021). The process of molecular separation has also wide applications in different fields such as nanotechnology-based areas, biochemical, pharmaceuticals, food, etc. (Wang et al., 2022a, 2022b; Xu et al., 2022; Yang et al., 2021; Yu et al., 2022a, 2022b). The

* Corresponding author.
E-mail address: pfyouxiang82@126.com (F. Peng).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.arabjc.2022.104306
1878-5352

selection and design of appropriate method is of fundamental importance to achieve the best and optimum conditions for the separation yield (Wang et al., 2022c, 2021; Zhang et al., 2022).

The molecular separation based on membrane technology offers superior advantages in process engineering field which makes this novel process efficient for separation and removal of target molecules from a mixture (Han et al., 2022; Chen and Yang, 2022; Wang et al., 2022d). The main advantage of membrane processes in molecular separation is that these processes can be regarded as green technology for separation, as no organic solvent is required in most of the membrane processes. In some cases, organic solvents are used for the separation such as those in membrane contactors for CO_2 capture and liquid extraction (Sabzekar et al., 2022; Petukhov et al., 2022; Liu et al., 2022).

To design and optimize the membrane separation process for a given target molecule, computational simulation can be employed in this regard to improve the separation process (Mou et al., 2022; Liu et al., 2022). The membrane-based computational simulation can be performed mainly using either computational fluid dynamics (CFD) or machine learning models. CFD methods are considered to be mechanistic models which provide process understanding, while the machine learning (ML) models are black-box models and use dataset for training algorithms. Integration of both models would be a novel idea in molecular separation which is the main focus of the current research.

A growing number of scientific disciplines are transitioning to machine learning (ML) methodologies that are progressively taking on the roles of classical computing methods. These techniques can be applied to solve problems in a number of ways. There is some correlation between inputs and outputs derived from these models, which produce some predictions (Alpaydin, 2020; Bishop, 2006; Carbonell et al., 1983).

KNN (K nearest neighbor) is one of the models that have been selected for this study to model the molecular separation case study. As well as being used for classification problems, it can also be used for regression problems. The application of this method is, however, more widely used in the field of classification problems. As the name implies, The K nearest neighbor algorithm is a simple algorithm that stores all of the available cases. New cases are estimated by the K closest neighbors based on their majority votes or average. In a distance function, the case being assigned to a class will be the one that has the highest frequency among its K nearest neighbors (Song et al., 2017; Kohli et al., 2021).

In this study, partial least square regression is also used as one of the models. In many problems, the partial least square regression (PLSR) model is a method that combines spectral measurements and chemical composition or physical properties of solids and charred materials to produce a rapid multivariate analysis with several variables (Xie and Chen, 2022; Boulesteix and Strimmer, 2007).

It is also relevant to note that Stochastic Gradient Descent (SGD) is a straightforward and effective technique for classifying and regressing linear data with convex loss functions, such as Support Vector Machines (linear) and Logistic Regression (linear). In the field of machine learning, the idea of SGD has been around for a long time; however, it is only relatively recently that it has received a significant amount of attention in the context of large-scale learning (Ighalo et al., 2020; Bottou, 1998).

In the current research, we have employed for the first time, the three mentioned machine learning methods for integration with CFD simulation to build hybrid modeling for simulation of a molecular separation case study via polymeric membrane technique.

2. Methodology

2.1. CFD method of modeling

As stated before, we used CFD simulation results in modeling of membrane separation process using machine learning models. In the first step of this work, CFD was performed to simulate removal of a species using polymeric cylindrical shape membrane. A simple geometry was designed and the mass transfer equations including convection and diffusion was solved for the feed, membrane, and permeate channels of the membrane module (Shirazian et al., 2012). Cylindrical coordinate was used for deriving and solving the governing equations. The main mass transfer equation can be expressed as (Rezakazemi et al., 2019):

$$D_{AB}\left[\frac{\partial^2 C}{\partial r^2}+\frac{1}{r}\frac{\partial C}{\partial r}+\frac{\partial^2 C}{\partial z^2}\right]=V\frac{\partial C}{\partial z} \tag{1}$$

where D_{AB} refers to the diffusion coefficient and C refers to the species concentration. Also, r and z refer to the radial and axial coordinate, while V is the velocity vector. Finite element method was used to solve Eq. (1) numerically (Rezakazemi et al., 2019).

2.2. Partial least squares regression (PLSR)

We have used a number of machine learning models to integrate to the CFD results for simulation of the membrane process in which PLSR is the first model which is utilized in this research. Using the covariance matrix, partial least squares regression is an efficient, quick and optimal regression technique. Partial Least Squares Regression, or PLS Regression, is the most common Partial Least Squares model. All other models in the family of PLS models are based on partial least squares regression. Regression models are applicable when you have numerical dependent variables (Bjørn-Helge et al., 2019). In the field of engineering and science, PLSR acts as a bilinear factor model which is extensively used in a variety of engineering and science fields. As a result, it performs well on matrices X that contain many noisy and collinear factors. According to the following Equations, the PLSR should have an optimal factor number (Abdi, 2010; Guo et al., 2021):

$$S_{SS,h}=\sum_{i=1}^{n}\left(y_i-\hat{y}_{hi}\right)^2 \tag{2}$$

$$S_{PRESS,h}=\sum_{i=1}^{n}\left(y_i-\hat{y}_{h(-i)}\right)^2 \tag{3}$$

$$Q_h^{\,2}=1-\frac{S_{PRESS,h}}{S_{SS,h}} \tag{4}$$

where y_i is the original data and $\hat{y}_{hi}$ is the fitted value of the i-th data point calculated from all data points and the h component of the regression model. By removing I as a data point from the modeling process and extracting the h component from the resulting regression model, we get the $\hat{y}_{h(-i)}$ notation, which indicates that y_i was determined using the latter. In the presence of $Q_h^{\,2}\geq 0.0975$, the addition of a new component greatly enhances the predictive power of the constructed model; in the other cases, the added new component has no meaningful significance (Guo et al., 2021; Elden, 2004).

2.3. KNN method

This model is a straightforward classification or regression model that uses the k-Nearest Neighbor estimation technique

as an instance-based machine learning model. A second reason why this is an inefficient technique is that it does not generalize from the subset of data used for training. This means that when testing and for unknown data, the full training subset is retained in addition to the testing subset. In KNN regression, the test examples are identified by comparing them to the training examples, in order to learn the regression model (Cover, 1968).

Suppose $T = [(x_1, y_1), \cdots, (x_N, y_N)]$ indicate distance parameterized training data and d, $x_i = (x_{i1}, \cdots, x_{im})$ is a representation of the i-th instance based on m input variables together with corresponding target output y_i. The number N represents the quantity of examples that have been provided. The d_i must be calculated between a test data point (x) and all other points $x_i \in T$. Then sort the d_i distances for a test data point x. If d_i is ranked as the i^{th} position, the d_i matching sample is referred to as the i^{th} nearest neighbor $NNi(x)$, and its target is denoted by y_i (x). Lastly, $\widehat{y}$ denotes the average of the regression outputs of KNN to x, i.e., $\widehat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i(x)$. Following are the steps involved in the KNN regression algorithm (Song, 2017):

- Inputs: training samples $\{x_i, y_i\}, x_i$: input features, y_i: output value, a single test data point x
- Algorithm:
  - o Compare each x_i with the query data point(x) and calculate the distance between them.
  - o Search for k closet data points $x_{i1} \cdots x_{ik}$ and their target outputs $y_{i1} \cdots y_{ik}$
  - o output:

$$\widehat{y} = \frac{1}{K}\sum_{j=1}^{k} y_{i_j}$$

2.4. Stochastic Gradient Descent (SGD)

SGD estimates the objective function gradient using a subset of training samples and dynamically adjusts the parameters. Batch size refers to the number of training samples employed. In the case of larger batches, the SGD simply changes into a gradient descent when the batch size increases. Although this is a very slow solution to convergence, it can be sped up by changing parameters more often with a small batch size (Haider et al., 2021).

The stochastic gradient descent (SGD) method is an algorithmic simplification of GD that uses only a sample from the dataset to estimate the gradient on each update iteration:

$$w_{t+1} = w_t - \gamma_t \nabla_w Q(z_i, w_t) \tag{5}$$

Under sufficient regularity conditions, it usually achieves a fast convergence when the learning rate is $\gamma\ t^{-1}$.

Each time the SGD iterates, it picks random samples from a finite training set. Most training sets have a bigger number of iterations than the number of training iterations. In this way, most of the data points in the training subset can be traversed by the average traversal of the data. Additionally, some of them will be run multiple times so that the expected risk can be directly optimized (Johnson and Zhang, 2013).

3. Results and discussions

3.1. CFD results

As explained earlier, we have performed CFD simulations for removal of species using a cylindrical membrane module. At the beginning of this work, the concentration distribution of the species was calculated using the CFD which uses finite element method. The results of the concentration (C) distribution are illustrated in Fig. 1 as 2D plot in cylindrical coordinate. The feed enters the membrane module from the bottom where the concentration is the highest (~1000) and leaves the module from the top side where the concentration decreases due to the mass transfer of species towards the polymeric membrane surface. It should be noted that we presented only the feed channel of the membrane module, and the membrane itself, and the permeate side is not represented in Fig. 1. Also, we used the feed channel data for training and validation of the machine learning models.

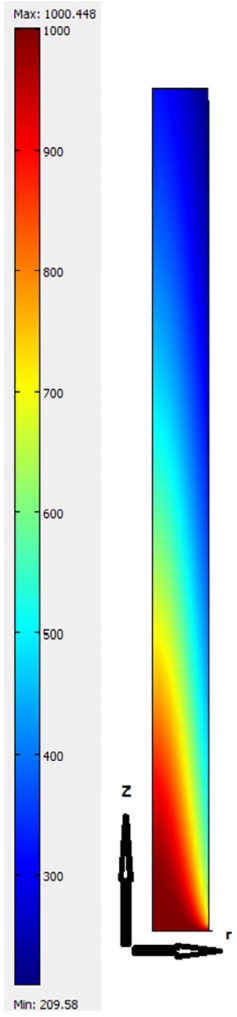

Fig. 1 Concentration (C) distribution of species in the feed channel of membrane calculated using CFD method.

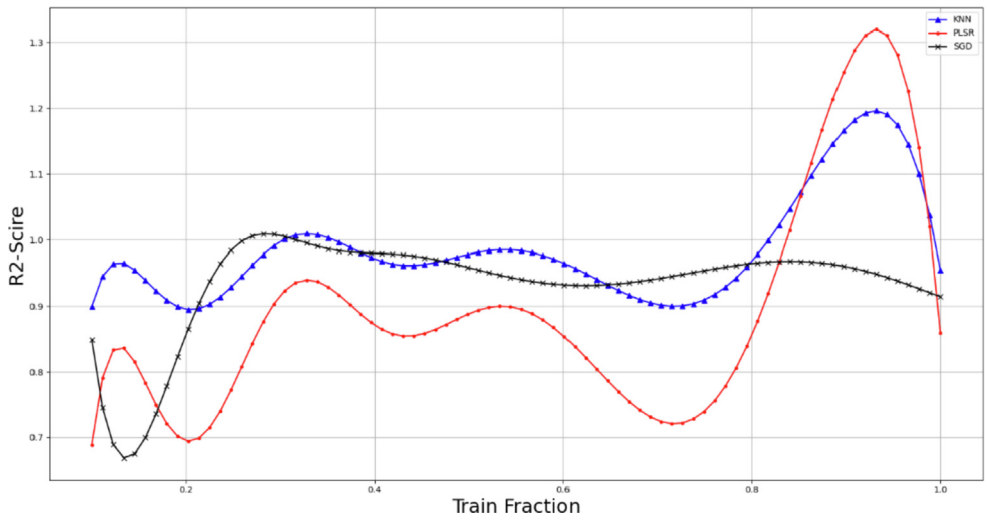

Fig. 2 Impact of different train fractions on model accuracy.

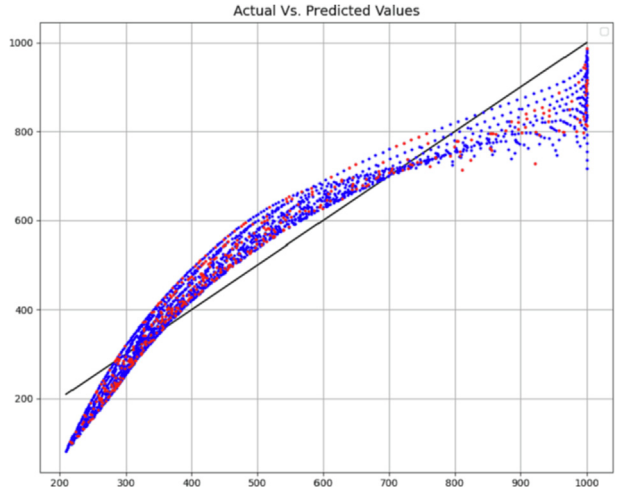

Fig. 3 0.5 training set – Comparison of predicted and actual C values.

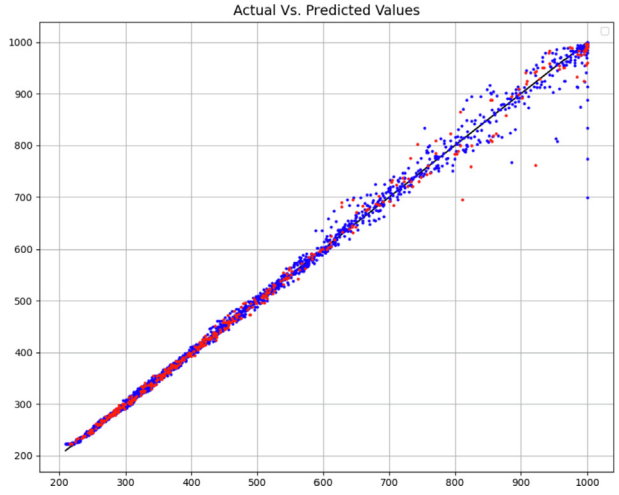

Fig. 4 0.9 training set – Comparison of predicted and actual C values.

3.2. Machine learning results

In order to obtain a suitable evaluation, the selection of the ratio of training and testing in the data is of particular importance. For this purpose, we tested this work with different values and evaluated the accuracy of the models (on average) in order to choose this configuration accurately. In Fig. 2, the effect of the ratio of the data of the train and the test on the accuracy of the model is examined, and according to this figure, 90% of the data is considered as the train and 10% as the test. Figs. 3 and 4 also show the comparison of the most 50% and 90% data in terms of the proximity of the expected and predicted data, which confirms the above fact.

For the purpose of comparing and analyzing final results, three models' parameters are optimized, and final models are implemented based on these configurations. For example, in Fig. 5, as an example of the hyper-parameter optimization result, the value of K (number of neighbors) in the KNN is shown for different values, for which the value of 8 was finally selected.

In regression modeling, a model error represents a deviation of the dataset from a reference set. A regression line is defined as the line that is closest to the observable data points. If there are multiple data points, the model's error is estimated using the below parameters (Hu et al., 2022):

- MAE: A mean is found by averaging the absolute values of the errors (Willmott and Matsuura, 2005; Botchkarev, 2018):

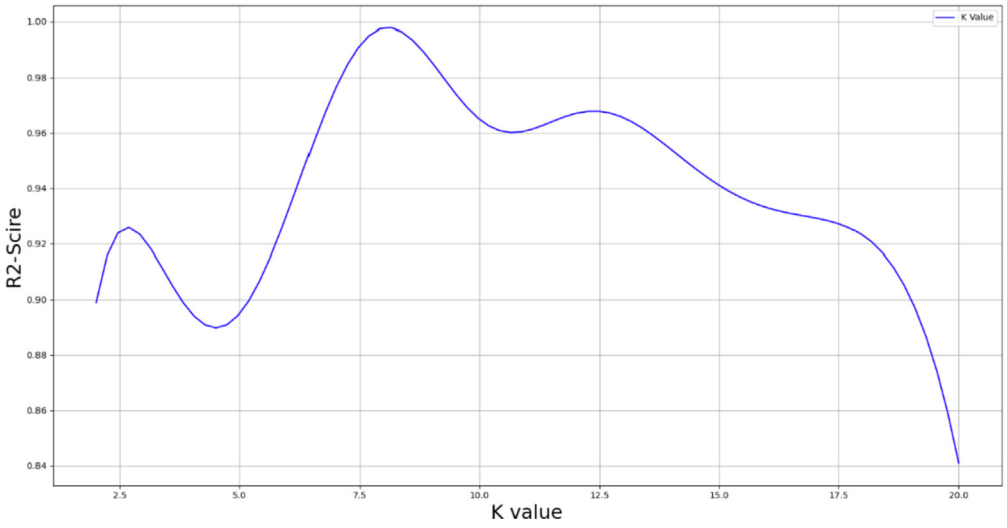

Fig. 5 Impact of value of K in KNN model on performance of model.

Table 1 The model results and their comparisons.

Models	MAE	RMSE	Max Error	R^2
KNN	3.43514E + 00	8.7266E + 00	1.34717E + 02	0.9983
SGD	5.13564E + 00	1.1170E + 01	1.64051E + 02	0.9973
PLSR	5.51127E + 01	6.5923E + 01	1.97509E + 02	0.8969

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{6}$$

- RMSE: The root mean square error is one of the most commonly used model performance evaluation statistics which is expressed as (Willmott and Matsuura, 2005):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{n}} \tag{7}$$

- R^2-Score

As can be seen from Table 1, the numerical results of the introduced models can be compared with each other. We can clearly see from the table that the KNN model is the most accurate model. This is because it has the most accurate results in terms of all the metrics, thus making it the most general model. There is also a confirmation of this issue in the validation Figs. 6–8.

The plot of prediction surface in the final model is shown in Fig. 9 for the best model which is KNN. Also, this diagram is similarly shown for two other models that are not the best case in Figs. 10 and 11. It is clearly seen that the developed machine learning models are capable of capturing the variations of the species concentration in the membrane feed channel with great accuracy.

In order to have a better view of the result of the final model, two heat map diagrams are shown in color in Figs. 12 and 13, in which the inputs are scaled in the first and not in the second illustration. Having looked at the results, it is concluded that the machine learning models can predict the CFD simulation results with acceptable accuracy and thereby can minimize the computational expenses for CFD simulation of large-scale systems which require high computational expenses.

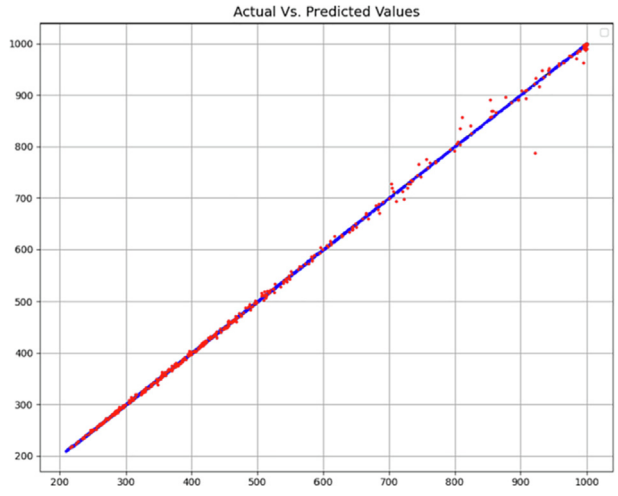

Fig. 6 KNN model – Comparison of predicted and actual C values.

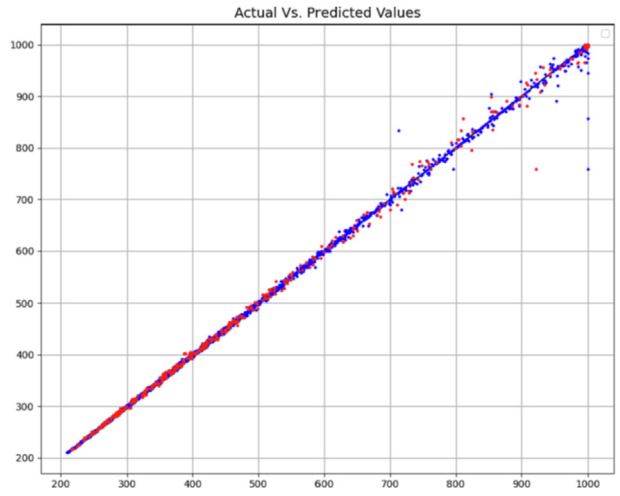

Fig. 7 SGD model – Comparison of predicted and actual C values.

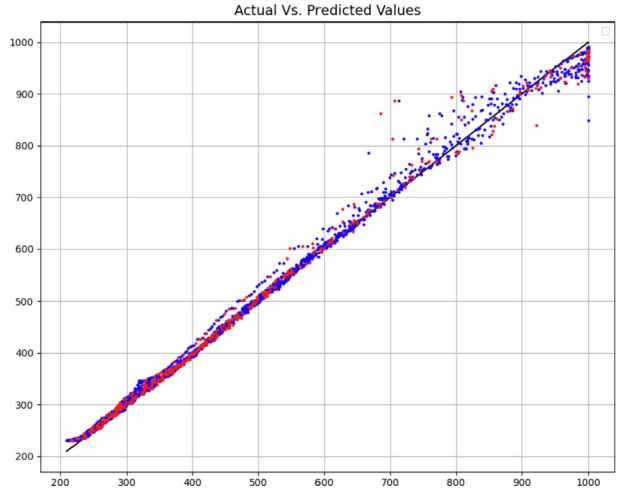

Fig. 8 PLSR model – Comparison of predicted and actual C values.

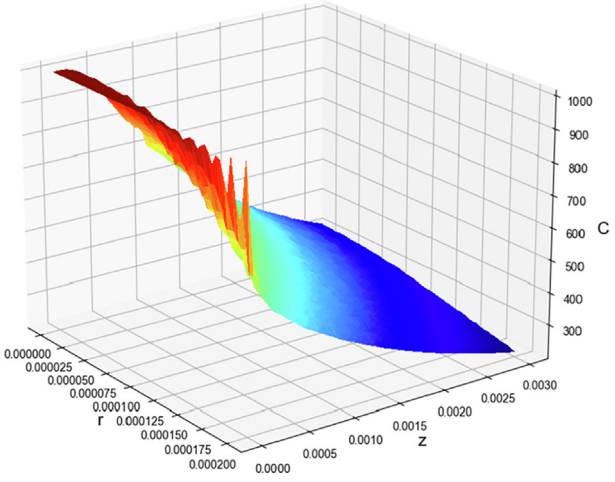

Fig. 9 Final KNN model. 3D projection of output with model.

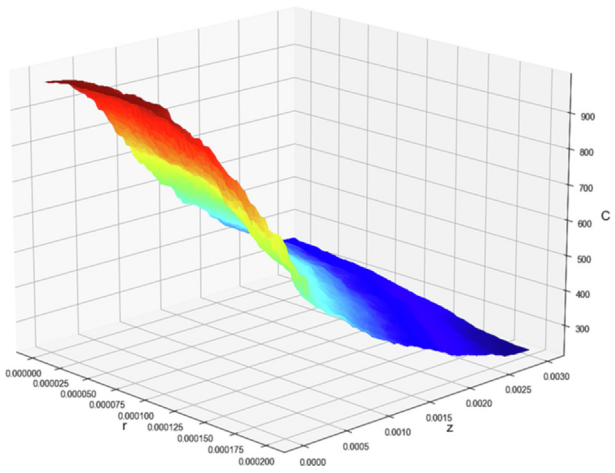

Fig. 10 Final SGD model. 3D projection of output with model.

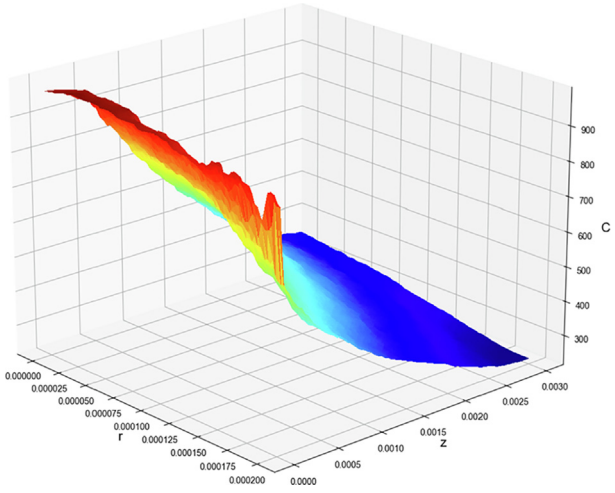

Fig. 11 Final PLSR model. 3D projection of output with model.

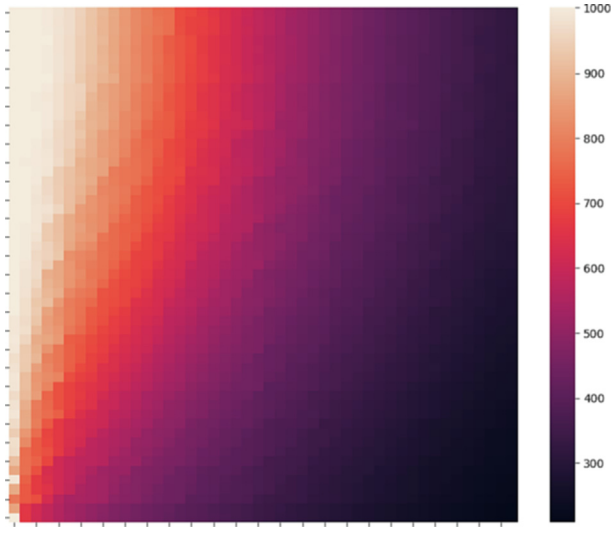

Fig. 12 The heat map plot of final model. X-ax is r and Y-ax is z.

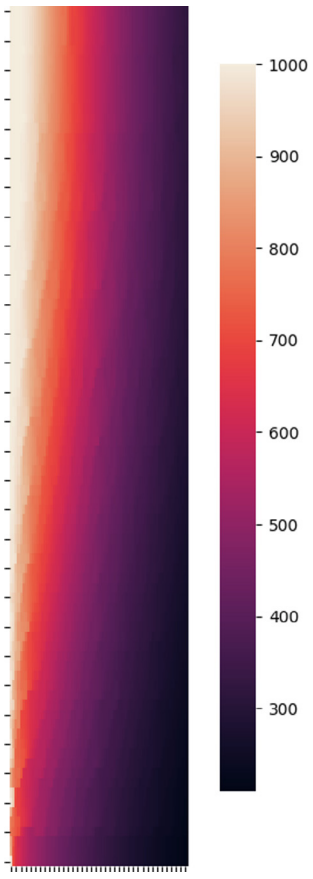

Fig. 13 The unscaled heat map plot of final model. X-ax is r and Y-ax is z.

4. Conclusion

In this investigation, we have a dataset that contains over 2,000 individual data points and two distinct input features (r and z) which have been obtained from a CFD simulation of membrane system. Additionally, the only output is the concentration distribution of the species in the feed channel, i.e., C. KNN, which stands for "K nearest neighbor," PLSR, which stands for "Partial Least Square Regression," and SGD, which stands for "Stochastic Gradient Descent," are the models that were used in this research to analyze the data set that was extracted from the CFD simulation of membrane. Following the optimization of the models with their hyper-parameters, the models were finally evaluated using a variety of statistical metrics. The MAE error metric for KNN, SGD, and PLSR are respectively 3.4, 5.1, and 5.5. In addition to this, the coefficient of determination (R^2) for each of them is respectively 0.998, 0.997, and 0.896. In conclusion, the KNN model with a K value of 8 was chosen for process analysis, because it provided the best results across the board in this investigation. The maximum error possible in the final analysis is also 1.35E + 02.

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.

Acknowledgements

Study on the pulsation characteristics of non-stable pressure of centrifugal pumps (No. NJZY17450) is acknowledged.

References

Abdi, H., 2010. Partial least squares regression and projection on latent structure regression (PLS Regression). Wiley Interdiscip. Rev.: Comput. Stat. 2 (1), 97–106.
Alpaydin, E., 2020. Introduction to Machine Learning. MIT Press.
Bai, B. et al, 2021. Cotransport of heavy metals and SiO2 particles at different temperatures by seepage. J. Hydrol. 597, 125771.
Bishop, C.M., 2006. Pattern Recognition. Mach. Learn. 128, 9.
Bjørn-Helge, M., Wehrens, R., Hovde Liland, K., 2019. pls: Partial least squares and principal component regression. R package version, p. 2.7-2.
Botchkarev, A., 2018. Evaluating performance of regression machine learning models using multiple error metrics in azure machine learning studio. Available at SSRN 3177507.
Bottou, L., 1998. Online learning and stochastic approximations. Online Learn. Neural Netw. 17 (9), 142.
Boulesteix, A.-L., Strimmer, K., 2007. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings Bioinf. 8 (1), 32–44.
Carbonell, J.G., Michalski, R.S., Mitchell, T.M., 1983. An overview of machine learning. Mach. Learn., 3–23
Chen, Y., Yang, X., 2022. Molecular simulation of layered GO membranes with amorphous structure for heavy metal ions separation. J. Membr. Sci. 660, 120863.
Cover, T., 1968. Estimation by the nearest neighbor rule. IEEE Trans. Inf. Theory 14 (1), 50–55.
Elden, L., 2004. Partial least-squares vs. Lanczos bidiagonalization—I: analysis of a projection method for multiple regression. Comput. Stat. Data Anal. 46 (1), 11–31.
Fan, S. et al, 2022. A novel water-free cleaning robot for dust removal from distributed photovoltaic (PV) in water-scarce areas. Sol. Energy 241, 553–563.
Ge, D. et al, 2019. Insight into the enhanced sludge dewaterability by tannic acid conditioning and pH regulation. Sci. Total Environ. 679, 298–306.
Guan, Q. et al, 2021. Ultrasonic power combined with seed materials for recovery of phosphorus from swine wastewater via struvite crystallization process. J. Environ. Manage. 293, 112961.
Guo, Y. et al, 2021. Prediction of rice yield in East China based on climate and agronomic traits data using artificial neural networks and partial least squares regression. Agronomy 11 (2), 282.
Haider, S.A., Sajid, M., Iqbal, S., 2021. Forecasting hydrogen production potential in islamabad from solar energy using water electrolysis. Int. J. Hydrogen Energy 46 (2), 1671–1681.
Han, M.-C. et al, 2022. Single-Side Superhydrophobicity in Si3N4-Doped and SiO2-Treated Polypropylene Nonwoven Webs with Antibacterial Activity. Polymers 14 (14), 2952.
Hu, X. et al, 2022. Predictive modeling and computational machine learning simulation of adsorption separation using advanced nanocomposite materials. Arabian J. Chem. 15, (9) 104062.
Ighalo, J.O., Adeniyi, A.G., Marques, G., 2020. Application of linear regression algorithm and stochastic gradient descent in a machine-learning environment for predicting biomass higher heating value. Biofuels Bioprod. Biorefin. 14 (6), 1286–1295.
Johnson, R., Zhang, T., 2013. Accelerating stochastic gradient descent using predictive variance reduction. Adv. Neural Inf. Process. Syst. 26.
Kohli, S., Godwin, G.T., Urolagin, S., 2021. Sales prediction using linear and KNN regression. In: Advances in Machine Learning and Computational Intelligence. Springer, pp. 321–329.
Liu, W. et al, 2020. Different Pathways for Cr(III) Oxidation: Implications for Cr(VI) Reoccurrence in Reduced Chromite Ore Processing Residue. Environ. Sci. Technol. 54 (19), 11971–11979.
Liu, Y. et al, 2022. Modeling insights into the role of support layer in the enhanced separation performance and stability of nanofiltration membrane. J. Membr. Sci. 658, 120681.

Liu, Y. et al, 2022. Novel method for high-performance simultaneous removal of NOx and SO2 by coupling yellow phosphorus emulsion with red mud. Chem. Eng. J. 428, 131991.

Liu, M. et al, 2022. Continuous separation and recovery of high viscosity oil from oil-in-water emulsion through nondispersive solvent extraction using hydrophobic nanofibrous poly(vinylidene fluoride) membrane. J. Membr. Sci. 660, 120876.

Mou, J. et al, 2022. An effective hybrid collaborative algorithm for energy-efficient distributed permutation flow-shop inverse scheduling. Future Generation Computer Systems 128, 521–537.

Petukhov, D.I. et al, 2022. Nanoporous polypropylene membrane contactors for CO2 and H2S capture using alkali absorbents. Chem. Eng. Res. Des. 177, 448–460.

Rezakazemi, M., Mosavi, A., Shirazian, S., 2019. ANFIS pattern for molecular membranes separation optimization. J. Mol. Liq. 274, 470–476.

Sabzekar, M. et al, 2022. Cyclic olefin polymer membrane as an emerging material for CO2 capture in gas-liquid membrane contactor. J. Environ. Chem. Eng. 10, (3) 107669.

Shirazian, S. et al, 2012. Hydrodynamics and mass transfer simulation of wastewater treatment in membrane reactors. Desalination 286, 290–295.

Song, Y. et al, 2017. An efficient instance selection algorithm for k nearest neighbor regression. Neurocomputing 251, 26–34.

Wang, X. et al, 2021. Radium and nitrogen isotopes tracing fluxes and sources of submarine groundwater discharge driven nitrate in an urbanized coastal area. Sci. Total Environ. 763, 144616.

Wang, J. et al, 2022a. Performance synergism of pervious pavement on stormwater management and urban heat island mitigation: A review of its benefits, key parameters, and co-benefits approach. Water Res. 221, 118755.

Wang, Y. et al, 2022b. Mo-modified band structure and enhanced photocatalytic properties of tin oxide quantum dots for visible-light driven degradation of antibiotic contaminants. J. Environ. Chem. Eng. 10, (1) 107091.

Wang, X. et al, 2022c. Highly selective membrane for efficient separation of environmental determinands: Enhanced molecular imprinting in polydopamine-embedded porous sleeve. Chem. Eng. J. 449, 137825.

Wang, Z. et al, 2022d. Experimental Investigation on Fracture Properties of HTPB Propellant with Circumferentially Notched Cylinder Sample. Propellants Explos. Pyrotech. 47 (7), e202200046.

Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30 (1), 79–82.

Xie, Z., Chen, X., 2022. Subsampling for partial least-squares regression via an influence function. Knowl.-Based Syst. 245, 108661.

Xu, D. et al, 2022. Coupling of sponge fillers and two-zone clarifiers for granular sludge in an integrated oxidation ditch. Environ. Technol. Innovation 26, 102264.

Yang, Y. et al, 2021. Construction of a novel lanthanum carbonate-grafted ZSM-5 zeolite for effective highly selective phosphate removal from wastewater. Microporous Mesoporous Mater. 324, 111289.

Yu, C. et al, 2022a. Ag3PO4-based photocatalysts and their application in organic-polluted wastewater treatment. Environ. Sci. Pollut. Res. 29 (13), 18423–18439.

Yu, F. et al, 2022b. A redox-active perylene-anthraquinone donor-acceptor conjugated microporous polymer with an unusual electron delocalization channel for photocatalytic reduction of uranium (VI) in strongly acidic solution. Appl. Catal. B 314, 121467.

Zhang, T. et al, 2022. Improving the humification and phosphorus flow during swine manure composting: A trial for enhancing the beneficial applications of hazardous biowastes. J. Hazard. Mater. 425, 127906.