



Comparative modelling of membrane filtration performance using response surface methodology, artificial neural network, and random forest regression: Implications for food and bioprocessing applications

Awang Bono¹, Zykamilia Kamin^{2,*}, Dona Stacy Petrus², Muhamad Afif Naqiudien Aladin² and Mohd Hardyianto Vai Bahrin³

¹ Faculty of Business Management and Information Technology, Universiti Muhammadiyah Malaysia, Padang Besar, 02100, Perlis, Malaysia

² Oil and Gas Engineering Programme, Faculty of Engineering, Universiti Malaysia Sabah, Jalan UMS, 88400, Kota Kinabalu, Sabah, Malaysia

³ Chemical Engineering Programme, Faculty of Engineering, Universiti Malaysia Sabah, Jalan UMS, 88400, Kota Kinabalu, Sabah, Malaysia

*Corresponding author: zykamilia@ums.edu.my

Received 31 January 2026

Revised 13 July 2026

Accepted 18 July 2026

Abstract

Membrane filtration plays a critical role in the food and bioprocessing industries. Despite its widespread application, accurately predicting membrane performance remains challenging due to the non-linear and inter-dependent influence of multiple operational parameters, including transmembrane pressure (TMP), crossflow velocity (CFV), temperature, pH, and solute concentration. This study evaluates the predictive performance of response surface methodology (RSM), artificial neural network (ANN), and random forest (RF) regression for permeate flux prediction. The models were trained and validated using 46 experimental data points compiled from a published study on water and wastewater filtration. Although derived from a different domain, the underlying flux dynamics and fouling mechanisms are conceptually analogous to those in food and bioprocessing systems, making this dataset a valid proxy for benchmarking model architecture and predictive capability. The RSM model identified TMP and CFV as the most statistically significant predictors and achieved an excellent goodness-of-fit ($R^2 = 0.9993$). The machine learning models also demonstrated strong predictive accuracy, with the ANN achieving an R^2 of 0.995 under cross-validation and the RF model reaching an R^2 of 0.933. The ANN effectively captured complex non-linear relationships in the data, whereas the RF model exhibited good generalization capability and interpretability. Collectively, these findings offer a conceptual framework for future membrane modelling applications in food and bioprocessing applications, with predictive accuracy expected to improve as larger, domain-specific datasets become available.

Keywords: Membrane filtration, response surface methodology, artificial neural network, random forest regression

1. Introduction

Membrane-based separation processes, including microfiltration, ultrafiltration, nanofiltration, and reverse osmosis [1,2], are widely employed in the food and bioprocessing industries for applications such as milk protein concentration, fermentation broth clarification, enzyme recovery, and fruit juice purification. Compared to conventional unit operations, membrane technology offers distinct advantages, including reduced sludge generation, higher-quality permeate, minimal chemical usage, and the potential for closed-loop water recycling. Combined with the compact footprint, relatively low capital investment, and operational simplicity, these attributes position membrane filtration as a cost-effective alternative or complement to traditional processing methods [3,4].

Despite these advantages, membrane processes are inherently limited by concentration polarization and fouling at the membrane surface, which cause a progressive decline in permeate flux. This performance degradation is highly sensitive to operational parameters, including transmembrane pressure (TMP), crossflow velocity (CFV), temperature, feed pH, and solute concentration. In food and bioprocessing applications, the presence of complex matrices, which is characterized by high viscosity, variable ionic strength, and emulsified fats or proteins, further exacerbates hydrodynamic interactions, concentration polarization, and fouling mechanisms. Consequently, accurately predicting permeate flux remains a complex, multivariate challenge [5,6].

Response surface methodology (RSM) has long been employed to analyse and optimize multivariate processes due to its string interpretability and its ability to model parameter interactions [7,8]. However, RSM often requires extensive experimental runs to adequately capture non-linear system behaviours, and it may struggle with highly complex datasets. Alternatively, machine learning (ML) algorithms, particularly artificial neural networks (ANNs) and random forest (RF), excel at modelling high-dimensional non-linear relationships without requiring pre-defined mechanistic assumptions, making them highly effective for permeate flux prediction and process optimization [9,10]. Integrating RSM with ML offers a synergistic approach, whereby RSM provides a structured experimental design and aids process understanding, while ML enhances predictive accuracy and model adaptability [11].

While numerous studies have applied either statistical or ML approaches to model membrane processes, few have systematically compared these methodologies under identical experimental conditions and validation protocols, particularly when working with limited datasets. The inherent non-linearity and strong variable interactions in membrane systems often compromise the predictive accuracy of conventional models when applied in isolation. To address this gap, a robust modelling framework is needed to rigorously evaluate both the predictive performance and the stability of statistical and data-driven approaches. Accordingly, this study presents a direct comparison of RSM, ANN, and RF models using a unified dataset and validation protocol. By using 46 experimental data points, an evaluation was conducted of each model's ability to predict permeate flux, quantify variable effects, and generalize across unseen data through cross-validation. Model performance was assessed using the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE).

The experimental dataset utilized in this work was compiled from a published study on water and wastewater membrane filtration. Although food and bioprocessing feed streams typically contain more complex constituents that increase viscosity and fouling intensity, the fundamental separation mechanisms, which are dictated by mass transport, concentration polarization, and fouling layer development, remain conceptually analogous across domains. Consequently, this dataset represents a reliable benchmark for evaluating model architecture and predictive capability. The findings presented here should be interpreted as a foundational modelling framework for membrane process prediction in food and bioprocessing applications, rather than a direct operational model for specific industrial systems.

2. Materials and methods

2.1 Dataset description

A dataset comprising 46 experimental observations was compiled from a published study on membrane filtration for water and wastewater treatment [12]. Although these data originate from non-food systems, the underlying transport phenomena and fouling mechanisms are conceptually analogous to those encountered in food and bioprocessing operations, such as whey protein concentration, fermentation broth clarification, and beverage purification. This study leverages these conceptual similarities to benchmark modelling methodologies, rather than the direct quantitative transferability of parameter effects across domains. Fundamental mechanisms, including concentration polarization, shear-dependent mass transport, and membrane surface fouling, are conserved across aqueous and complex food matrices, thereby supporting the use of this dataset to evaluate predictive model architecture and performance.

The dataset includes five independent operational variables, including temperature ($^{\circ}\text{C}$), transmembrane pressure (TMP, bar), crossflow velocity (CFV, m/s), pH, and salt concentration (g/L), with permeate flux (PFlux, L/m²/h) as the dependent response variable. These parameters critically influence fouling behaviour and hydrodynamic conditions in industrial bioprocessing. Specifically, temperature, TMP, pH, and CFV serve as adjustable parameters to mitigate the flux decline associated with concentration polarization, while pH and ionic strength modulate the membrane surface charge, electrostatic interactions, and protein solubility [13]. Such physicochemical effects are directly relevant to microbial fermentation, dairy processing, and enzymatic reaction systems.

Prior to the model development, data pre-processing was performed to ensure analytical rigour and minimize bias, including addressing potential outliers. All the input features for the ML model development were standardized using z-score normalization to account for disparate units and magnitudes. Given the limited sample size ($n = 46$), the ANN and RF models employed five-fold cross-validation to assess generalization performance,

reduce overfitting risk, and provide robust estimates of predictive accuracy [14]. This validation strategy ensures that model evaluation reflects performance on unseen data rather than memorization of the training set. Furthermore, to mitigate overfitting and assess generalization, the dataset was partitioned into 80% training and 20% testing sets.

2.2 Response surface methodology (RSM)

RSM analysis was conducted using Design-Expert 13 software (Stat-Ease Inc., USA). The experimental data were fitted to a second-order polynomial model based on a Box-Behnken design (BBD), as expressed in Eq. (1).

$$Y = \beta_0 + \sum \beta_i X_i + \sum \beta_{ii} X_i^2 + \sum_{i < j} \beta_{ij} X_i X_j + \varepsilon \quad (1)$$

where Y denotes the predicted permeate flux, X_i represents the coded independent variables, β_0 , β_i , β_{ii} , and β_{ij} correspond to the intercept, linear, quadratic, and two-factor interaction coefficients, respectively, and ε represents the residual error term. Analysis of variance (ANOVA) was performed to assess the statistical significance of the main effects and interactions at a 95% confidence level ($p < 0.05$). Model adequacy was evaluated using R^2 , adjusted R^2 , predicted R^2 , and lack-of-fit tests. The model assumptions were verified using diagnostic charts, which show the normal probability of residuals and projections against actual values [15].

It is important to clarify that the RSM model served solely as a comparative statistical framework. No synthetic data generated from the RSM polynomial equation were used to train the ML models; both ML models were developed exclusively using the original 46 experimental observations. This approach ensured a fair comparison of predictive methodologies without introducing potential bias from model-generated data.

2.3 Artificial neural network (ANN) model

The ANN model was implemented in Python 3.10 using the scikit-learn library [16]. Hyperparameter optimization was performed via GridSearchCV with five-fold cross-validation on the training set. The search space included hidden layer architectures [(25,), (50,), (100,), (25,50), (50,100), (25,100)], activation functions ['relu', 'tanh', 'logistic'], alpha values [0.0001, 0.001, 0.01, 0.1], learning rates [0.001, 0.01, 0.1], and optimizers ['adam', 'sgd']. Following optimization, the final architecture comprised an input layer with five neurons (corresponding to the operational variables), two hidden layers with 50 and 100 neurons, respectively, and a single output neuron for permeate flux prediction. The rectified linear unit (relu) activation function, Adam optimizer, alpha value of 0.001, and a learning rate of 0.001 were selected. While hyperparameter tuning favoured a relatively complex architecture (two hidden layers, with 50 to 100 neurons each), the limited dataset size might have increased the risk of overfitting. The architectures tested herein were intended to explore the relationship between network complexity and predictive performance; however, since shallower networks may offer better generalization for small experimental datasets, they warrant further investigation.

2.4 Random forest (RF) regression

An RF regression model was implemented to enhance interpretability and quantify variable importance, using the identical train/test split protocol as the ANN [17]. RF constructs an ensemble of decorrelated decision trees via bootstrap aggregation (bagging) and random feature subsampling; it then aggregates predictions to reduce variance and improve robustness. Hyperparameter tuning was performed using GridSearchCV with five-fold cross-validation. The search space included $n_estimators$ [100, 200], max_depth [5, 10, None], $min_samples_split$ [4, 6, 8], $min_samples_leaf$ [1, 2, 4], and $max_features$ ['sqrt', 0.8]. The optimized hyperparameters were: $n_estimators = 100$, $max_depth = None$, $max_features = 0.8$, $min_samples_split = 4$, and $min_samples_leaf = 1$.

2.5 Model evaluation metrics

Model performance was quantified using three standard regression metrics— R^2 , MAE, and RMSE—as expressed in Eqs. (2) – (4), respectively.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4)$$

where y_i , $\hat{y}_i$, and $\bar{y}_i$ denote the observed, predicted, and mean observed permeate flux values, respectively, and n is the number of samples. Higher R^2 values (approaching 1) indicate better explanatory power, whereas lower MAE and RMSE values reflect higher predictive accuracy. To assess the prediction uncertainty and model stability, cross-validation statistics were reported.

3. Results and discussion

3.1 RSM model fitting and statistical analysis

RSM was first applied to analyse the experimental dataset, establishing a statistically grounded baseline for comparison with ML approaches and identifying the most influential operational parameters. The fitted second-order polynomial model achieved an R^2 of 0.9993, an adjusted R^2 of 0.9987, and a predicted R^2 of 0.9979, indicating excellent goodness-of-fit and predictive capability within the experimental design space. The ANOVA identified TMP and CFV as the most statistically significant factors ($p < 0.001$), followed by temperature ($p < 0.05$). Enhanced mass transfer at higher shear rates reduces concentration polarization, thereby increasing permeate flux, which is consistent with prior studies on protein separation and fermentation broth clarification [18]. The strong interaction between TMP and CFV further underscores the importance of hydrodynamic control in mitigating fouling.

As indicated in Figures 1(a) and 1(b), the permeate flux showed an increasing trend with an increase in TMP; this continued until around 4 bar, when it started decreasing. The temperature and CFV variables positively impacted the flux by lowering the liquid viscosity and preventing fouling of the membrane; meanwhile, pH and salt concentration each had a major effect on osmotic pressure and electrostatic forces. The diagnostic plots provided in Figures 1(c) and 1(d) indicated that the model was adequate, since the residuals were normally distributed. In general, RSM is efficient at identifying key factors, although this efficiency declines as processes become more complex, highlighting the value of ML for modelling high-dimensional, non-linear systems.

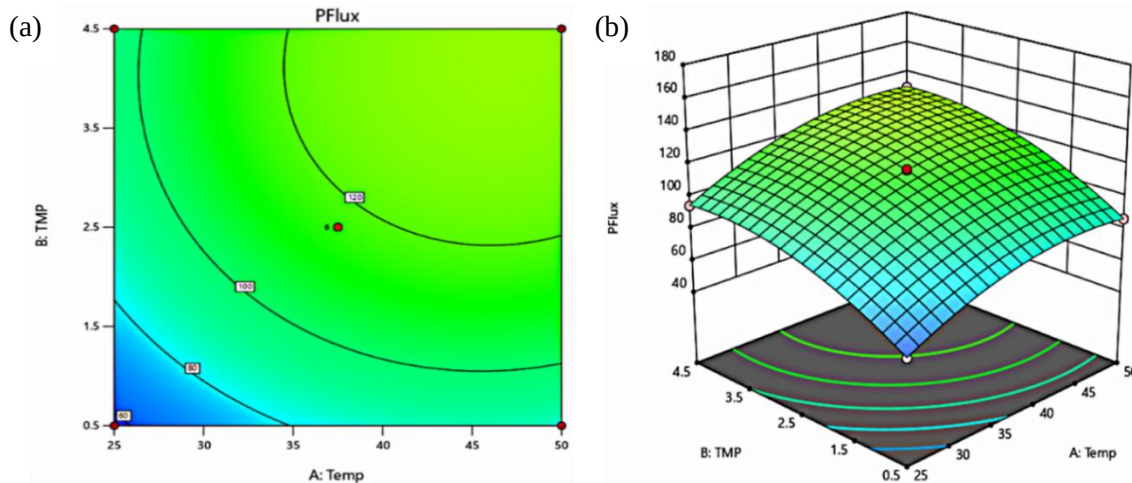


Figure 1 RSM results — (a) contour and (b) 3D surface plots showing the effects of TMP and CFV on permeate flux; (c) normal probability plot of residuals; (d) predicted versus actual permeate flux validating the model fit.

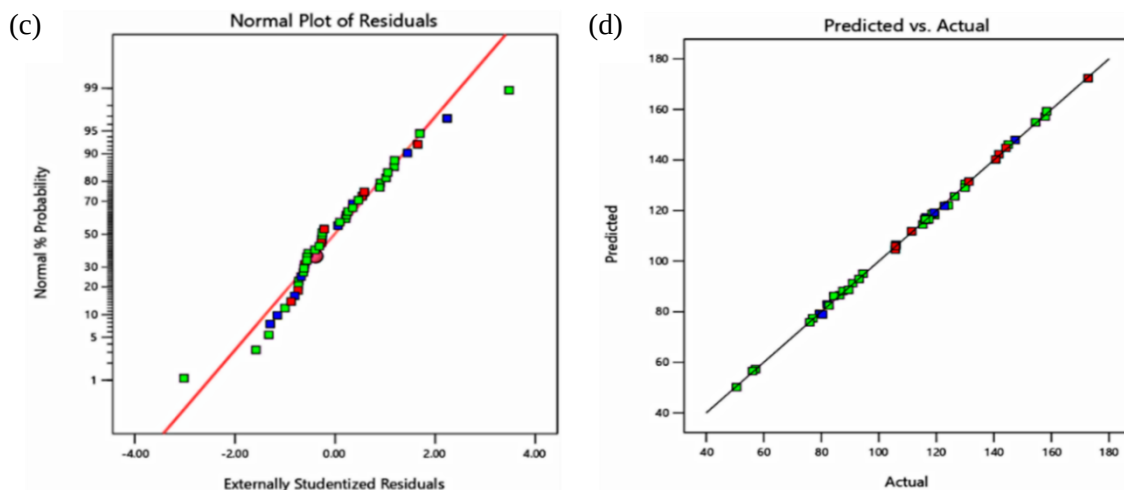


Figure 1 RSM results — (a) contour and (b) 3D surface plots showing the effects of TMP and CFV on permeate flux; (c) normal probability plot of residuals; (d) predicted versus actual permeate flux validating the model fit. (Conf.)

3.2 ANN model performance

The optimized ANN model achieved strong predictive performance on the test dataset, with an R^2 of 0.995, MAE of 0.733 L/m²/h, and RMSE of 2.02 L/m²/h. Five-fold cross-validation on the training subset yielded a mean R^2 of 0.995 ± 0.006 , indicating stable performance and low variance across data folds. The parity plot depicted in Figure 2(a) shows that the data points closely aligned with the diagonal line, confirming excellent agreement between the experimental and predicted flux values. The ANN model effectively captured non-linear phenomena, including the characteristic flux plateau at elevated TMP attributable to membrane compaction and fouling resistance. Meanwhile, Figure 2(b) presents a Pareto chart of the absolute prediction errors across individual samples. The errors were symmetrically distributed around zero, with most values falling within a narrow band, demonstrating consistent predictive accuracy across the full flux range. This robustness suggests the ANN architecture successfully generalized from the limited dataset without overfitting to specific regions of the design space.

When contrasting RSM and ANNs, a key distinction is that RSM derives predictions from an explicit polynomial equation fitted to the data, whereas ANNs learn complex input-output mappings directly from datasets through iterative optimization. This data-driven flexibility is particularly advantageous for food and bioprocessing applications, where complex fluid compositions and dynamic fouling mechanisms generate highly non-linear flux responses. A recognized limitation of ANN models, however, remains their limited intrinsic interpretability. Despite this, the ANN demonstrated exceptional predictive accuracy, positioning it as a promising tool for real-time process monitoring, adaptive control, and optimization of membrane-based food and bioprocessing systems.

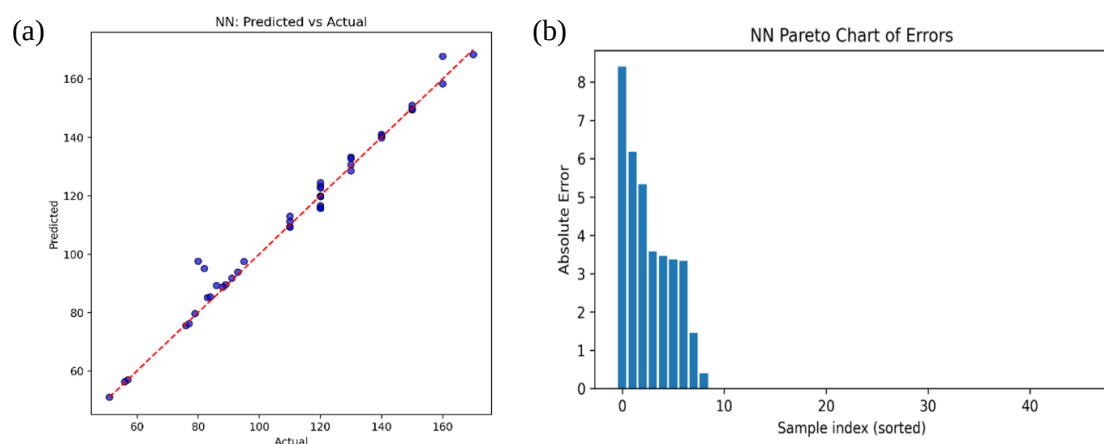


Figure 2 ANN model performance — (a) parity plot comparing predicted and experimental permeate-flux values, showing strong model calibration; (b) Pareto chart of errors plot illustrating prediction errors across the sample numbers and confirming consistent high performance.

3.3 RF regression results

The RF regression model achieved satisfactory predictive performance, with an R^2 of 0.933, MAE of 4.91 L/m²/h, and RMSE of 7.33 L/m²/h. The five-fold cross-validation yielded a mean R^2 of 0.933, confirming the model's stability across data folds. Although the optimized RF model showed good predictive performance, the selected hyperparameters may increase the model complexity when using a limited dataset. The hyperparameters were selected using five-fold cross-validation to reduce the overfitting risk during model optimization. However, shallower tree depths and stricter splitting conditions may provide better generalization performance for small datasets, and this should be explored in future studies.

The built-in feature importance of RF provided preliminary insights into variable significance. More advanced explainability techniques such as Shapley Additive exPlanations (SHAP) analysis could offer a better understanding of the magnitude and directional contribution that each process variable makes to permeate flux prediction. Therefore, SHAP-based interpretation should be implemented in future ML modelling studies to improve the model transparency and interpretability [19]. In this study, the RF feature importance was used as an initial and computationally simpler approach to evaluate the influence of each variable within the limited scope of the work [20]. The ranking for each factor is as follows, starting with TMP and ending with salt concentration: TMP $\approx$ 0.42, CFV $\approx$ 0.31, temperature $\approx$ 0.14, pH $\approx$ 0.08, and salt concentration $\approx$ 0.05. Both hydrodynamic parameters directly influenced the permeate flow and fouling behaviour during filtration.

Higher transmembrane pressure increased the driving force across the membrane, while higher crossflow velocity helped reduce foulant deposition and concentration polarization near the membrane surface. Due to this direct effect on flux behaviour, both variables contributed more strongly to the prediction model. In comparison, temperature and pH were of lower importance, perhaps because their investigated operating ranges were relatively small. Therefore, their effects on permeate flux changes were less obvious compared to those of pressure and flow conditions. This suggests that these outcomes are highly consistent with the findings from both the RSM and ANN models. Meanwhile, this ranking confirms that TMP and CFV are the most significant process variables in terms of influencing permeate flux, in parallel with the fundamental transport and fouling mechanisms in food and bioprocessing systems.

Figure 3a displays the parity plot comparing predicted and experimental flux values; it demonstrates a strong linear correlation and confirms accurate model calibration. Figure 3b presents the Pareto chart of errors plot, showing that prediction errors were skewed towards the left, with only two samples being significantly higher than the remaining samples. This indicates that only two samples could not be predicted accurately by the RF model, while the remaining prediction errors for the other samples are mostly below ten. Compared with the ANN model, RF displayed higher sensitivity to outliers and a slightly smaller generalization error, reflecting the robustness of its ensemble averaging approach [21].

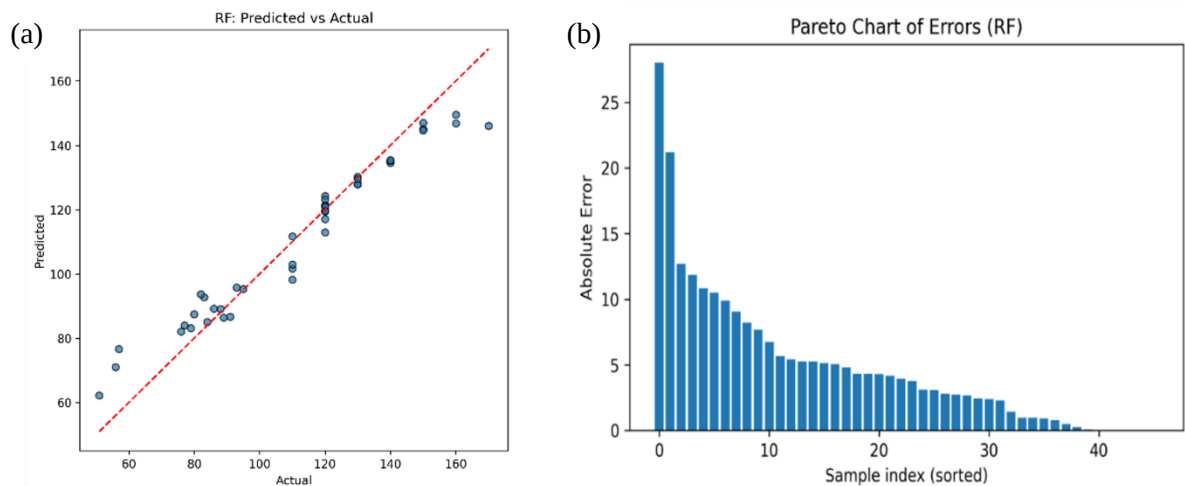


Figure 3 RF model performance — (a) parity plot of predicted versus observed flux, illustrating model accuracy; (b) Pareto chart of errors plot showing prediction errors.

3.4 Comparative analysis of RSM, ANN, and RF models

A summary of the model performance indicators is presented in Table 1, illustrating both the predictive accuracy and interpretability of the three approaches.

Table 1 Summary of model performance indicators.

Model	R^2	MAE (L/m ² /h)	RMSE (L/m ² /h)	Interpretability	Generalization
RSM	0.9993 (Curve fitting performance)	2.96	4.21	High	Moderate
ANN	0.995 (Cross-validated)	0.77	2.02	Low	High
RF	0.933 (Cross-validated)	4.91	7.33	Moderate	High

Due to the limited dataset size ($n = 46$), both polynomial regression (RSM) and machine learning models (ANN and RF) may show optimistic performance when evaluated on training data. Therefore, model comparison should be interpreted with caution, especially when dealing with an RSM model fitted using the full dataset without cross-validation. In this study, the RSM model is based on a deterministic second-order polynomial regression fitted using the full dataset, and the reported R^2 value (0.9993) represents the goodness-of-fit on the training data. The RSM model R^2 represents curve fitting performance and should not be directly compared with cross-validated metrics from the ANN and RF models. Additionally, the high R^2 performance of RSM might reflect the high-quality dataset collected by Salahi et al. [12]. In contrast, the ANN and RF models were evaluated using a cross-validation-based procedure to assess generalization performance.

Model performance was evaluated using R^2 , MAE, and RMSE, with the different models showing varying strengths depending on the metric considered. RSM has the highest R^2 value, but it simply cannot predict performance outside the experimental limit because it strongly relies on direct curve fitting from the polynomial model [22]. The second most accurate model is the ANN, which has performance metric of $R^2 = 0.995$ and the lowest error metrics (MAE = 0.77 L/m²/h; RMSE = 2.02 L/m²/h), confirming its ability to capture nonlinear and multivariate interactions effectively. The RF model also achieved satisfactory predictive accuracy, with a relatively high R^2 (0.933), as well as strong robustness and interpretability through feature-importance analysis, making it a versatile model option. Although random forest is usually deemed a strong method for small datasets, the RF model in this study gave a higher prediction error (higher MAE and RMSE) and a lower R^2 test compared to the ANN model. One possible reason is that the dataset contained only 46 samples, which may be insufficient for stable tree splitting in RF. The selected RF parameters may also have made the model too complex and increased overfitting, causing it to learn too well. Hence, a better approach may be to use simpler RF hyperparameters on a small dataset. On the other hand, ANN models are well-known for being better at learning nonlinear relationships between independent and dependent variables, regardless of the complexity of the datasets. Meanwhile, the small dataset size and random train-test split may also have affected the consistency of the model performance results. In practical terms, RSM excels in experimental design, factor screening, and statistical significance testing, while ANNs are ideal for predictive modelling and RF models are well-rounded, balancing interpretability and accuracy [23].

3.5 Implications for food and bioprocessing applications

While the dataset employed herein originates from water and wastewater filtration studies, this work is intended as a methodological framework applicable to food and bioprocessing systems. The core transport phenomena, such as pressure-driven flow, concentration polarization, and fouling due to solute accumulation, are conceptually conserved across aqueous and complex food matrices. Nevertheless, it is acknowledged that food and bioprocessing feeds introduce additional complexities such as organic fouling by proteins, lipids, and polysaccharides, and variable viscosity, which are not represented in the current dataset.

Therefore, the models in this study provide conceptual validation for the membrane processes of statistical and data-driven modelling strategies. This illustrates how such tools can be applied to domain-specific datasets. Based on the ranking of all five investigated factors, TMP and CFV were consistently identified as significant parameters; therefore, they are equally critical in protein ultrafiltration, microbial biomass recovery, and juice clarification, where controlling shear and minimizing pressure fouling and energy consumption are often necessary [24].

ANNs and RF can be used to make digital process control systems to ensure real-time permeate flux prediction and adaptive operation in food plants, especially in membrane separation. For instance, an ML-based digital twin could adjust the TMP or CFV values on the spot during whey concentration or enzyme purification to maintain optimal flux while preventing irreversible fouling [25]. However, this study presents a preliminary machine learning-based modelling framework using an ANN and RF for permeate flux prediction. While the results show good predictive capability within the studied dataset, the current models are trained on simplified water/wastewater filtration data, and they do not capture the full physicochemical complexity of food and bioprocessing systems. As such, the proposed approach should be interpreted as a proof-of-concept rather than a fully validated digital twin or real-time control system for industrial food processing applications. Moreover, food and bioprocessing systems typically produce many datasets because they can be influenced by many factors, making them suitable for ML model training. As more experimental data become available globally, the predictive

performance of these models can be further improved through transfer learning or hybrid RSM–ML modelling, especially for ANNs, which rely heavily on learning complex patterns from large datasets [26,27].

4. Conclusion

This study compared RSM, ANN, and RF regression models for predicting permeate flux in membrane filtration systems. Utilizing 46 experimental data points from water and wastewater filtration literature, three distinct modelling frameworks were developed and evaluated for predictive accuracy and interpretability. The results revealed that RSM provides a robust statistical baseline for identifying the most influential operational parameters. The ANN model excelled in capturing complex, non-linear relationships among multiple variables, achieving the highest predictive accuracy. Meanwhile, the RF model offered a favourable balance of predictive performance and interpretability. Collectively, this comparative analysis confirms that integrating statistical and ML approaches yields comprehensive insights for data-driven membrane process modelling. The proposed framework serves as a foundational step towards developing advanced predictive tools for food and bioprocessing membrane systems, pending further validation with domain-specific, larger-scale datasets.

Future research should prioritize the expansion of high-quality, domain-specific datasets, as the efficacy of these data-driven models is inherently tied to data volume and diversity. Additionally, subsequent studies should incorporate additional physicochemical variables under industrial bioprocessing conditions, such as feed viscosity, total solids, protein and lipid content, and specific membrane material properties. Finally, integrating dynamic time-series data, fouling indices, and membrane resistance metrics would facilitate the development of advanced digital twins for membrane processes. In the context of Industry 4.0, such intelligent data-driven modelling frameworks are essential for advancing energy-efficient, sustainable, and highly optimized food and bioprocessing operations.

5. Acknowledgements

We would like to express our sincere gratitude to Universiti Malaysia Sabah for giving Muhamad Afif Naqiudien Aladin the opportunity to obtain his research fellowship with financial support. This provision greatly motivates him to continue his research and overcome any unexpected challenges that occur throughout the journey.

6. Author contributions

Muhamad Afif Naqiudien Aladin: Validation, Visualization, Writing – review & editing; Zykamilia Kamin: Project administration, Supervision; Dona Stacy Petrus: Resources, Software; Awang Bono: Supervision, Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft; Mohd Hardyianto Vai Bahrin: Resources, Software, Writing – review & editing.

7. Conflicts of interest

The authors declare no competing interest.

8. References

- [1] Baker RW. Membrane technology and applications. John Wiley & Sons; 2023.
- [2] Judd S. The MBR book: principles and applications of membrane bioreactors for water and wastewater treatment. Elsevier; 2010.
- [3] Shahabadi SMS, Reyhani A. Optimization of operating conditions in ultrafiltration process for produced water treatment via the full factorial design methodology. *Sep Purif Technol.* 2014;132:50-61.
- [4] Fakhru'l-Razi A, Pendashteh A, Abdullah LC, Biak DRA, Madaeni SS, Abidin ZZ. Review of technologies for oil and gas produced water treatment. *J Hazard Mater.* 2009;170:530-551.
- [5] Bowen WR, Welfoot JS. Modelling the performance of membrane nanofiltration—critical assessment and model development. *Chem Eng Sci.* 2002;57:1121-1137.
- [6] Field RW, Wu D, Howell JA, Gupta BB. Critical flux concept for microfiltration fouling. *J Memb Sci.* 1995;100:259-272.
- [7] Montgomery DC. Design and analysis of experiments. John Wiley & Sons; 2017.
- [8] Myers RH, Montgomery DC, Anderson-Cook CM. Response surface methodology: process and product optimization using designed experiments. John Wiley & Sons; 2016.
- [9] Haykin S. Neural networks and learning machines, 3/E. Pearson Education India; 2009.

- [10] Maier HR, Jain A, Dandy GC, Sudheer KP. Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environ Model Softw.* 2010;25:891-909.
- [11] Molnar C. Interpretable machine learning. Lulu. com; 2020.
- [12] Salahi A, Noshadi I, Badrmezhad R, Kanjilal B, Mohammadi T. Nano-porous membrane process for oily wastewater treatment: Optimization using response surface methodology. *J Environ Chem Eng.* 2013;1:218-225.
- [13] Kovacs DJ, Li Z, Baetz BW, Hong Y, Donnaz S, Zhao X, et al. Membrane fouling prediction and uncertainty analysis using machine learning: A wastewater treatment plant case study. *J Memb Sci.* 2022;660:120817.
- [14] Garakani SS, Chew JW. Development of physics-informed machine-learning models to enhance understanding and prediction of membrane fouling. *J Memb Sci.* 2025;124133.
- [15] Jang H, Lee CS, Kim JH, Kim J. Optimization of photocatalytic ceramic membrane filtration by response surface methodology: Effects of hydrodynamic conditions on organic fouling and removal efficiency. *Chemosphere* 2024;356:141885.
- [16] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res.* 2011;12:2825-2830.
- [17] Breiman L. Random forests. *Mach Learn.* 2001;45:5–32.
- [18] Owusu Sekyere A, Baidoo MF, Ohemeng-Boahen G, Essandoh HMK. Design optimization of ceramic membrane filters based on a response surface method. *Compos Adv Mater.* 2024;33:1-19.
- [19] Lu Q, Zhang H, Fan R, Wan Y, Luo J. Machine learning-based Bayesian optimization facilitates ultrafiltration process design for efficient protein purification. *Sep Purif Technol.* 2025;363:132122.
- [20] Hengl T, Nussbaum M, Wright MN, Heuvelink GBM, Gräler B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ.* 2018;6:e5518.
- [21] Wang L, Li Z, Fan J, Lu G, Liu D, Han Z. Prediction of membrane purification by membrane fouling based on mathematic and machine learning models combined with image processing technology. *J Environ Chem Eng.* 2023;11:111154.
- [22] Eteba A, Bassyouni M, Elzahar MH, El-shekhby MZ. Development and optimization of biodegradable chitosan-based nanofiltration membranes for effective textile wastewater treatment. *Sep Purif Technol.* 2025;367:132852.
- [23] Deng B, Deng Y, Liu M, Chen Y, Wu Q, Guo H. Integrated models for prediction and global factors sensitivity analysis of ultrafiltration (UF) membrane fouling: statistics and machine learning approach. *Sep Purif Technol.* 2023;313:123326.
- [24] Al-Amoudi AS. Factors affecting natural organic matter (NOM) and scaling fouling in NF membranes: A review. *Desalination.* 2010;259:1-10.
- [25] Alzahrani S, Mohammad AW. Challenges and trends in membrane technology implementation for produced water treatment: A review. *J Water Proc. engineering.* 2014;4:107-133.
- [26] Lim SJ, Kim YM, Park H, Ki S, Jeong K, Seo J, et al. Enhancing accuracy of membrane fouling prediction using hybrid machine learning models. *Desalination Water Treat.* 2019;146:22-28.
- [27] Afif M, Aladin N, Kamin Z, Chel-Ken C, Hardyianto Vai Bahrin M, Bono A, et al. Applications of Machine Learning in Modelling and Optimization of Breakthrough Curve Analysis: A Focus on Artificial Neural Network and their Comparison. *J Adv Res Comput Appl.* 2026;42:69-76.