# Evaluation of Machine Learning Algorithms to Predict Internal Concentration Polarization in Forward Osmosis

Ibrar Ibrar[1], Sudesh Yadav[1], Ali Braytee[2], Ali Altaee[1,*], Ahmad HosseinZadeh[1], Akshaya K. Samal[3], John L Zhou[1], Jamshed Ali Khan[1], Pietro Bartocci[4,5], Francesco Fantozzi[4]

1: School of Civil and Environmental Engineering, University of Technology Sydney, Ultimo, NSW 2007, Australia

2: School of Computer Science, University of Technology Sydney, Ultimo, NSW 2007, Australia

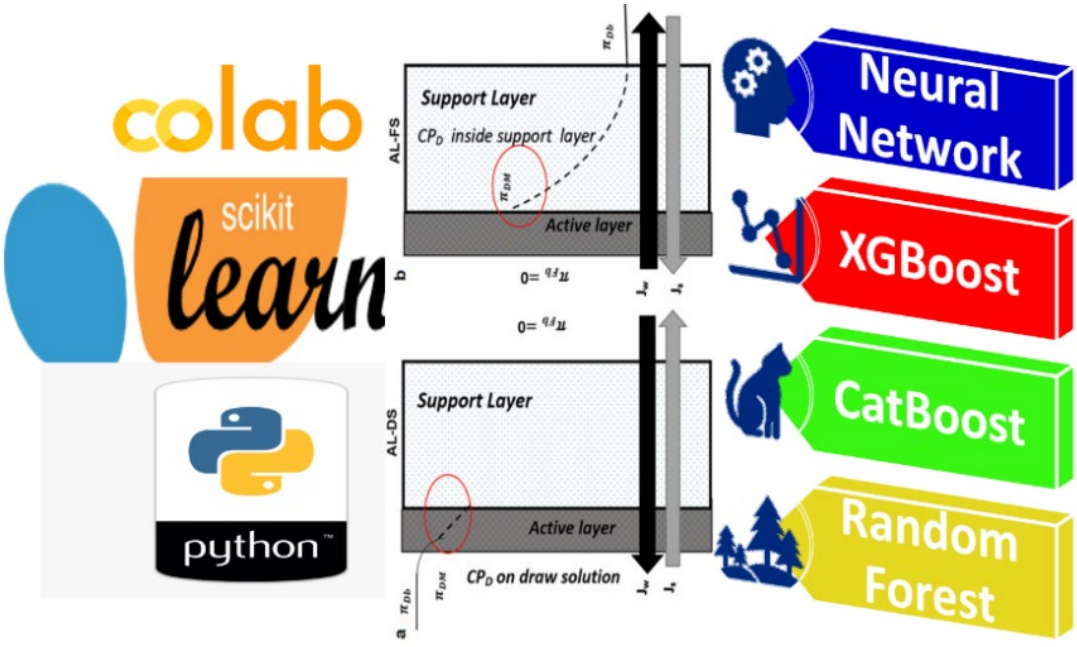
3: Centre for Nano and Material Sciences, Jain University, Bangalore - 562112, India

*Corresponding author email: *Ali.altaee@uts.edu.au*

4: Department of Engineering, University of Perugia, Via G. Durante 67, Perugia 06125, Italy

5: Instituto de Carboquímica (C.S.I.C.), Miguel Luesma Castán 4, 50018 Zaragoza, Spain

## Graphical Abstract



**Abstract**

Internal concentration polarization (ICP) is currently a major bottleneck in the forward osmosis process. Proper modelling of the internal concentration polarisation is therefore vital for improving the process performance and efficiency. This study assessed the feasibility of several machine learning methods for internal concentration polarisation prediction, including artificial neural networks, extreme gradient boosting (XGBoost), Categorical boosting (CatBoost), Random forest, and linear regression. Among the many algorithms evaluated, the CatBoost regression outperformed other methods in terms of coefficient of determination ($R^2$) and the mean square error. The CatBoost algorithm's prediction power was then evaluated using non-training (user-provided) data and compared to solution diffusion models. The results indicated that the machine learning algorithms could predict ICP in the process with high accuracy for the provided dataset and excellent generalizability for future testing data. Furthermore, machine learning algorithms may offer insights into the input features that majorly affect ICP modelling in the forward osmosis process.

**Keywords**: Forward osmosis (FO), internal concentration polarization (ICP), machine learning modelling, artificial neural network, and wastewater treatment

## 1. Introduction

Forward osmosis (FO) membrane-based desalination has attracted tremendous attention due to its numerous advantages over pressure-driven membrane processes, particularly for treating complex wastewaters [1, 2]. The FO process advantages include low membrane fouling and a high-water recovery rate [3, 4]. However, the process being driven by osmosis or concentration differences between the feed and the draw solutions, is hindered by the problem of concentration polarization (CP) [5-7]. In the FO process, two types of CP co-occur,

viz. external concentration polarization (ECP) and internal concentration polarization (ICP) [5]. The ECP has been reported in all membrane processes, whereas the ICP is unique to the forward osmosis process. Both ICP and ECP lead to a decrease in the concentration at the boundary layer near the membrane surface compared to the actual concentration in the bulk solutions, leading to a decrease in the net osmotic driving force in the FO process [5]. However, ECP is not a severe issue compared to ICP in the FO process since ECP can be mitigated by varying hydrodynamic conditions.

Earlier studies on the FO process revealed that ICP accounts for about 80% of flux reduction in the FO process [7-9]. Since CP is such a significant issue in the forward osmosis process, modelling water flux behaviour in the presence of the CP is of utmost importance to understand the process performance and possible optimisation. Several models for CP measurements and predictions have been reported in the literature [10-15]. The first FO water flux model incorporating concentration polarization was proposed by Lee et al. [16]. The model has been revised and witnessed a lot of improvements by several researchers over time. This model was later updated by McCutcheon and Elimelech [10] by considering external and internal CPs on both sides of the FO membrane. This model was widely used in the FO literature; however, the impact of reverse salt flux ( RSF) from the draw solution to the feed solution, osmotic swelling of the FO membrane, as well as the mass transfer resistance provided by the support layer was neglected. Yip et al. [13] presented an updated model for FO transport in the presence of external and internal CPs along with the impact of the RSF. Details equation of these models is presented in the methodology section. Several other models are also presented in the literature, such as Bui et al. [15] model, resistance in the series model by Nagy [14], empirical model by [17] and several new models based on CFD

(computational fluid dynamics)[18], and two dimensional-FEM (finite element analysis) models [19-21].

Each model given in the literature has limitations and was developed based on certain assumptions or ideal circumstances. The measurement, prediction and estimation of concentration polarization, water flux and RSF are based on iterative procedures in these models. Furthermore, these models are based on exacting methodologies, are mathematically sophisticated and are time-consuming. Despite these drawbacks, these models are still widely employed in FO studies; however, there is a lack of universality since each model is limited to a certain range of parameters. For instance, if the draw solution is changed from NaCl (sodium chloride) to a mixture (e.g. NaCl + $MgSO_4$ (magnesium sulphate) + $MgCl_2$ (magnesium chloride)), most FO flux models would be inaccurate because the diffusion coefficient for the mixture would have to be determined separately through a lengthy series of exhaustive experiments. Recently, there has been a surge in ML modelling due to their high precision and accuracy [22-24], and ML algorithms have recently been used for FO's performance in several studies [23, 25]. ML is the study and interpretation of patterns and structures in data to automate learning, reasoning, and prediction [23]. Further, ML enables users to feed massive amounts of data to a computer algorithm, then analyses and makes data-driven recommendations and predictions based entirely on the input data. Moreover, unlike solution diffusion models, ML models can model complex and non-linear systems with high accuracy [26, 27].

Very few studies have investigated ML techniques to model forward osmosis performance. For instance, Jawad et al. [23] employed artificial neural networks to predict forward osmosis permeate flux. K et al. [28] employed artificial neural networks and ANFIS (Adaptive-Neuro

Fuzzy Inference System) to model forward osmosis treatment of textile wastewater. Hosseinzadeh et al. [29] employed ANN and the ANFIS to model water flux in an osmotic membrane bioreactor. The prediction of all of these studies is limited to water flux prediction, which otherwise can also be predicted accurately using solution diffusion or empirical models. Fundamentally, the problem is with the prediction of internal concentration polarisation, where solution diffusion models or empirical models have limitations. As a result, the use of ML techniques can be a viable tool to predict the ICP in the forward osmosis process.

This research aims to assess and compare various ML models for ICP prediction and compare the prediction to solution diffusion (SD) models. ML predictions can be superior to SD models since ML algorithms can learn from the input data, find hidden patterns and predict the output with high precision and accuracy. On the other hand, using solution diffusion-based models, for example, if (NH4)HCO3 (ammonium bicarbonate) or Ca(NO3)2 (calcium nitrate) is used as a draw solution, the diffusion coefficient of these compounds is not accessible in the literature, and the solution diffusion model would fail to estimate the severity of the ICP. Furthermore, when draw solutions mixture is employed, calculating the diffusion coefficient becomes difficult. As a result, we propose and analyse several ML models for ICP prediction based on various inputs and FO process parameters for the first time. By applying different ML algorithms to the dataset, two critical questions will be addressed; which ML algorithm is the most effective for the prediction of ICP modulus, and what is the potential of ML models, in general, to give insights into the most important features that impact ICP on the FO process. For assessing the modulus of ICP, the findings are compared to the SD model.

## 2. Theory and methods

## 2.1. Study design

Initial data were collected from published papers that reported values for K (solute resistance to diffusion), "S," predicted water flux using SD models, and experimental water flux. The rest of the data for input parameters were gathered from the experimental design and results of each study after the published studies were chosen (**Appendix Figure A.3 and A.4, heatmaps**) [5, 8, 11-15, 17, 30-33]. ML modelling was done in Python using Google Collaboratory or Google Collab on the Google research platform, a free Jupyter notebook in the Google cloud. Modelling was done following each ML algorithm-specific methodology without major feature engineering.

## 2.2. Data collection and pre-processing

The data collected for the ML models included a total of 438 instances and 19 features, including output features for CICP (concentrative internal concentration polarization) and DICP (dilutive internal concentration polarization) [10, 17, 34-36]. The total number of data points for DICP were 351 instances and 19 features (Heatmap is shown in Figure A.3, Appendix A. On the other hand, the total number of data points of CICP was 84 instances and 19 features (Heatmap shown in Figure A.4, Appendix A). To the best of our knowledge, no study recommends a minimum quantity of data for the ANN and other ML algorithms to perform well; however, some studies have reported good performance even with datasets ranging from 20-80 data points [37, 38]. The data from images were extracted using an online software known as “WebPlotDigitizer 4.5”. The data included all the input parameters that impact internal concentration polarization in the FO process, except reverse salt flux (RSF) since RSF data was scarce in the published studies. The forward osmosis system used throughout all the studies was operated in the FO mode, also termed the AL-FS mode, and the PRO mode, also known as the AL-DS mode (**Table 1**). The crossflow velocities ranged from

129 0.0192 m.sec$^{-1}$ to a maximum of 1.22 m.sec$^{-1}$. Recent studies have revealed that flow
130 arrangement impacts the ICP [15]; thus, it was also considered an input parameter. Most of
131 the data were collected using commercial CTA, TFC and $NaNoH_2O$ membranes provided by
132 different suppliers of FO membranes. All of the studies' pure water permeability A value, salt
133 rejection R-value, and salt permeability B value were collected. Later, rejection rate data were
134 not considered in the input since most studies reported a similar rejection for the FO
135 membrane. The osmotic pressure data calculated by OLI software was taken directly from the
136 studies or calculated by Van't Hoff equation when it was not directly available. The data was
137 normalized in Python by utilising the data's minimum and maximum values.

138 $$X_{sc} = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

139 **Table 1:** Input and output parameters for the collected data

| Input/output Parameters | Mean | Maximum value | Minimum value | Standard deviation |
|---|---|---|---|---|
| Flow arrangement (Cocurrent or counter current) | N/A | N/A | N/A | N/A |
| Membrane orientation AL-FS | N/A | N/A | N/A | N/A |
| Membrane orientation AL-DS | N/A | N/A | N/A | N/A |
| FS cross-flow velocity (CFV) | 0.21 | 122.22 | 0.56 | 26.64 |
| DS CFV | 0.21 | 122.22 | 1.92 | 26.26 |
| FS molarity | 0.13 | | | 0.21 |
| FS Osmotic Pressure | 6.59 | 52.02 | 0 | 10.76 |
| DS molarity | 1.36 M | 9 | 0.04 | 1.03 |

| | | | | |
|---|---|---|---|---|
| DS osmotic pressure | 71.98 | 1.96 | 483.14 | 54.71 |
| FS type | N/A | N/A | N/A | N/A |
| DS type | N/A | N/A | N/A | N/A |
| FS temperature | 300.98 K | 323 | 273 | 6.61 |
| DS temperature | 300.98 K | 323 | 273 | 6.23 |
| DS Molecular weight | 67.47 g | 119.02 | 47.60 | 12.02 |
| Osmotic pressure difference | 71.98 atm | 0 | 472 | 52.54 |
| Membrane pure water permeability (A) value | 3.01E-07 m/sec.atm | 7.12E-07 | 1.1E-07 | 1.24E-07 |
| Membrane salt permeability (B) value | 6.24E-07 m/s | 9.8E-08 | 5E-08 | 4.18E-07 |
| Water flux | 3.92E-06 m/s | 2.1E-05 | 2.7E-10 | 3.41E-06 |
| DICP (output) | 0.38 | 1 | 0 | 0.33 |
| CICP (output) | 5.72 | 1 | 147.10 | 19.26 |
| Mass transfer coefficient (*k)* | 2.81E-05 | 6.3E-05 | 1.7E-05 | 1.93E-05 |
| Solute resistance to diffusion (***K)*** | 2.6E05 | 3.5E05 | 2.5E05 | 3.1E05 |

*RSF data was not available for most studies in the literature and CP models, therefore was excluded.

**Data for solution diffusion and empirical model**

The ICP modulus calculation requires the value solute resistance to diffusion. The ***k*** value (mass transfer coefficient of the FO channel) and ***K*** (solute resistance to diffusion) must be calculated for the SD model. This was either taken directly from the studies where it was available or calculated accordingly to the two models of [10] and [13]. According to [10], water flux in the FO process operating in the PRO mode (active membrane layer faces the draw solution) is given by Equation (2). For the FO membrane operating in the FO mode (active membrane layer faces the feed solution), water flux is presented by Equation (3).

149 $$J_w^{PRO} = A\left[\pi_{Db}\, exp\left(\frac{-J_w}{k}\right) - \pi_{Fb} exp(J_w K)\right] \quad (2)$$

150 $$J_w^{FO} = A\left[\pi_{Db}\, exp(J_w K) - \pi_{Fb}\, exp\left(-\frac{J_w}{k}\right)\right] \quad (3)$$

151 Where $J_w$ is the experimental flux, $\pi_{Db}$and $\pi_{Fb}$ are the bulk feed concentration of draw and
152 feed solution, respectively, ***k*** is the convective mass transfer coefficient, and ***K*** is the salt
153 resistivity. Unfortunately, the proposed model neglects the effects of salt transport and the
154 external mass transfer resistance on the support layer [14]. Yip et al. [13] presented a
155 modified mathematical model in Equations (4) and (5) for the FO and the PRO orientation,
156 respectively.

157 $$J_w^{FO} = A\left[\frac{\pi_{Db}\, exp(-J_w K) - \pi_{Fb} exp\left(\frac{J_w}{k}\right)}{1+\frac{B}{J_w}\left\{exp\left(\frac{J_w}{k}\right) - exp\,-(J_w K)\right\}}\right] \quad (4)$$

158 $$J_w^{PRO} = A\left[\frac{\pi_{D,b}\, exp\left(-\frac{J_w}{k}\right) - \pi_{F,b}\, exp(J_w K)}{1+\frac{B}{J_w}\left\{exp(J_w K) - exp\left(-\frac{J_w}{k}\right)\right\}}\right] \quad (5)$$

159 The mass transfer resistance at the porous support layer, on the other hand, was not
160 considered in equations (4) and (5), but it's by far the most widely used model for the
161 calculation of ***K*** value. The experimental flux was taken as reported directly from the studies.
162 The modelling studies considered the initial water flux (30 minutes), and constant draw
163 solution concentration is maintained, or the dilution of draw solution is ignored and not
164 considered.

165 **2.3. Design settings and performance evaluation metrics**

166 Two different models were designed and evaluated for DICP and CICP, respectively (**Table 2**).
167 The former CP occurs in the AL-FS mode and the latter CP in the AL-DS mode, and both do not
168 occur simultaneously in the FO process, so one was made redundant when modelling for the
169 other parameter. The efficacy of the constructed ML models can be determined using the
170 correlation coefficient ($R^2$), which is defined as:

$$R^2 = \frac{\sum_{i=1}^{n}(P_i - \bar{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y_i})^2} \quad (6)$$

Where $P_i$ denotes the predicted output of the ML model, $\bar{y_i}$ denotes the mean of the sample data, and $y_i$ represent the actual output. In general, R-square values range from 0 to 1, with a higher R-squared value indicating a better model. The MSE (mean square error) for each model was calculated using equation (7).

$$MSE = \frac{1}{n}\sum(y_i - P_i)^2 \quad (7)$$

The MSE indicates how close the regression model line (or fit) is to the data. The MSE of the training data in supervised ML shows how good is the model in detecting the anomaly in the dataset. However, it is the MSE of the test data, which indicates the performance of the ML model. In this study, the MSE for the test data was considered the performance indicator of the predictive power of the ML models. In general, the higher the $R^2$ value and lower the MSE, the better the model's predictive power. The different models were built and run on the Google Colaboratory platform using Python (Python 3.9) using the built-in libraries. Python offers numerous advantages over MATLAB for ML as Python is free, open-source and has various inbuilt libraries for ML such as NumPy, Matplotlib, Keras and Pandas. Further, the back-end libraries in Python, such as TensorFlow, chose the best and efficient way to define and train an ML model with a few lines of code. The model's input data comprised the FO's independent variables, and the output layer consisted of either CICP or DICP based on the membrane orientation.

**Table 2:** Different ML approaches used in this study

| Model Name | Output | Validation method | Python Library |
|---|---|---|---|
| ANN | CICP/DICP | R square/MSE | TensorFlow |

| | | | |
|---|---|---|---|
| Extreme Gradient boosting (XGBoost) | CICP/DICP | R square/MSE | XGBoost |
| CatBoostRegressor | CICP/DICP | R square/MSE | CatBoost |
| RandomForestRegressor | CICP/DICP | R square/MSE | Sklearn.ensemble |
| Linear regression | CICP/DICP | R square/MSE | Scikit-Learn |

## 2.4. Artificial neural network

This study designed a neural network in Python, using the Keras (A Python library for neural networks). After loading the dataset, the model was one-hot encoded to create dummies from all the categorical variables such as flow arrangement (countercurrent or co-current flow), membrane orientation (AL-FS or AL-DS), types of FS and DS. This method of encoding was chosen as there was no direct relationship between the categorical input variables. Customized models were trained for both datasets of DICP and CICP. In the datasets, various categories of input variables were processed into dense layers after the sequential model was implemented.

## 2.5. Gradient Boosting Models Theory

Three gradient tree boosting models were employed in this study, known as extreme gradient boosting or XGBoost, Categorical boosting or CatBoost and RandomForest. The most common functions (predictive learners) used in the Gradient Boosting framework are Decision Trees [39]. Let's say we have a set of input variables $x = \{x_1, x_{2,}\, x_3 \dots \dots \dots \dots x_p\}$. The Decision Trees are built in a greedy manner, with the best split points chosen based on purity scores such as Gini or to minimise the loss. In this model, we have all the parameters which have an impact on ICP as input variables. Our output variable $y$ is either CICP or DICP. In Gradient Boosting

models, a single terminal tree is used to create an initial basis function $F_0\,(x)$ given by equation (9) [40, 41].

$$F_0\,(x) = argmin_\beta \sum_{i=1}^{n} L(y_i, \beta) \qquad (9)$$

In equation (6) $L(y_i, \beta)$ represents the loss function, $\beta$ represents the set of split points for the tree's internal nodes, and n represents the number of input and output variables. Using some samples from a training dataset, the goal is to find the values of $\beta$ that minimize the loss function to the minimum. Detailed methodology and equation derivation can be found in [40].

## 3. Results and Discussions

In the FO mode or the AL-FS orientation, the DS is diluted inside the support layer leading to DICP output for the ML model. On the other hand, when the membrane orientation is the AL-DS mode, the model's output is the CICP. The modulus of ICP is always less than 1, whereas the modulus of CICP is always greater than 1. The "K" value reported for SD models is usually higher in the AL-DS mode than the AL-FS mode.

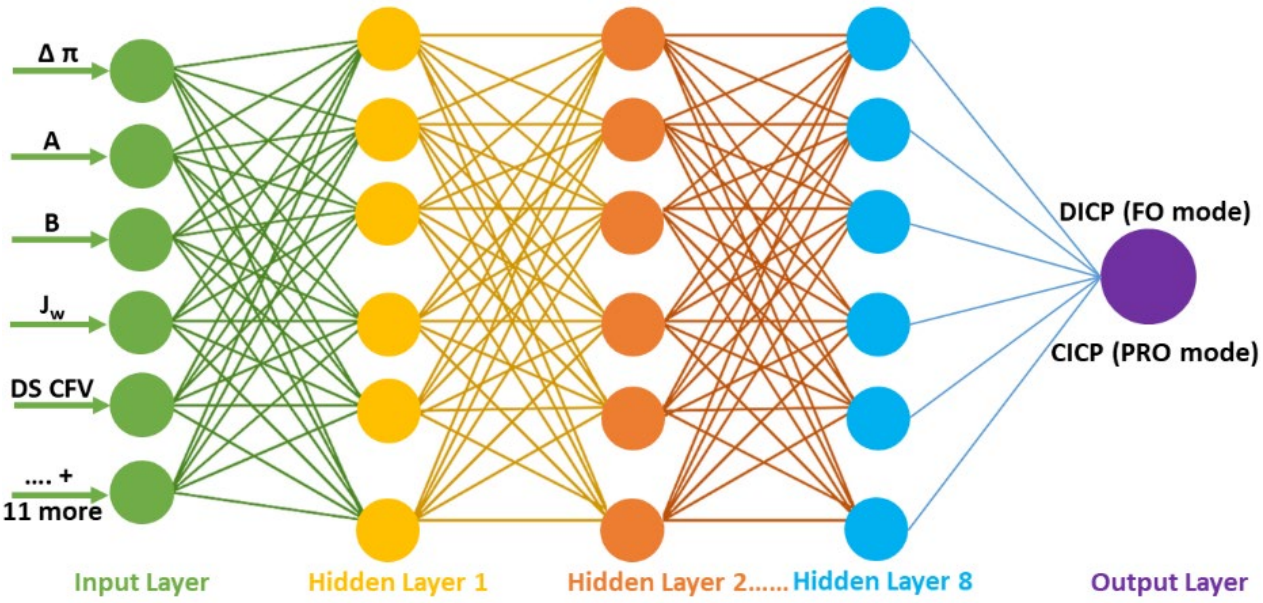
### 3.1. Prediction of DICP and CICP using ANN

A neural network using a deep learning approach was designed in Python to predict DICP and CICP, respectively. Since there is no linear relationship between the input and output parameters, to create an efficient neural network, we must carefully choose the appropriate or the optimum number of layers, the learning rate of the algorithm, and the number of training epochs. For instance, if the hidden layer contains insufficient neurons, the ANN will not reflect nonlinearity in the training data and lead to inaccurate predictions. Initially, setting the hidden layers to one, the optimal number of hidden layers was determined through trial

and error. An initial neural network was designed by considering all the input parameters listed in Table 1 without hyperparameter optimisation. The designed neural network had an input layer, three hidden layers, and an output layer (for CICP or DICP, depending on the membrane orientation). The neural network exhibited good predictive power for the CICP; however, it was found that this neural network was prone to overfitting and didn't exhibit good predictive power for the DICP dataset (Fig.A.1 for DICP and Fig.A.2 for CICP (Appendix A.1)).

Further optimisation was performed by selecting appropriate inputs to the ANN model based on the correlation matrix indicated in Fig. A.3 and Fig.A.4 for the DICP (blue) and CICP (pink), respectively (Appendix A.1). It was found from the heatmap that several parameters have correlated to each other with a correlation factor of 1, such as the correlation of FS CFV with DS CFV and FS molarity with FS osmotic pressure. It meant one of those should not be considered as input and should be removed from the input dataset, which may increase the ANN model performance. As a result, FS CFV and FS molarity were removed from the main dataset. In general, DS CFV has more impact on the ICP than FS CFV, and FS osmotic pressure is generally considered in all ICP equations rather than molarity. The hyperparameters tuning of the ANN model was further performed through Grid search (GridSearchCV in Python), leading to selecting the best parameters for the model performance. Additionally, batch normalisation was added after the activation functions for the normalisation of the output of each layer. The resulting optimised neural network had an input layer, eight hidden layers, and an output layer (for CICP or DICP, depending on the membrane orientation). It should be noted that the optimal number of hidden layers was determined through trial and error. Using Python Scikit learn, 80 % of the input data was used as the training set for the neural network, 10% was used to validate the data, and the rest of the 10 % was utilized for testing

255 the ANN for the prediction of DICP and CICP. The ANN was implemented using the TensorFlow
256 Sequential API, and the Relu (rectified linear activation function) and LeakyRelu activated
257 function was chosen for the DICP and the CICP models. The categorical data include
258 membrane orientation (AL-FS or AL-DS), flow arrangement (countercurrent or cocurrent), FS
259 types such as NaCl, DI water, KCl, DS types and membranes in the dataset was transformed
260 into numerical values using Python "LabelEncoder". To reduce the learning rate and avoid
261 overfitting of the model, "early stopping" was implemented. After some trial and error with
262 the number of epochs based on the validation R square and mean square error, the neural
263 network with the best possible results was chosen and presented here in this study (**Fig.1**).




264

265 **Figure 1:** ANN network for prediction of DICP and CICP

266 For each phase of the simulation, the $R^2$ value was calculated for the training, validation,
267 testing, and all data. For the DICP model, The $R^2$ values for training and test data are 0.90 and
268 0.88, respectively (**Fig 2a and 2c**). The testing data $R^2$ (0.82) shows a good performance of the
269 optimised ANN compared to the initial one (reported in Appendix A.1).

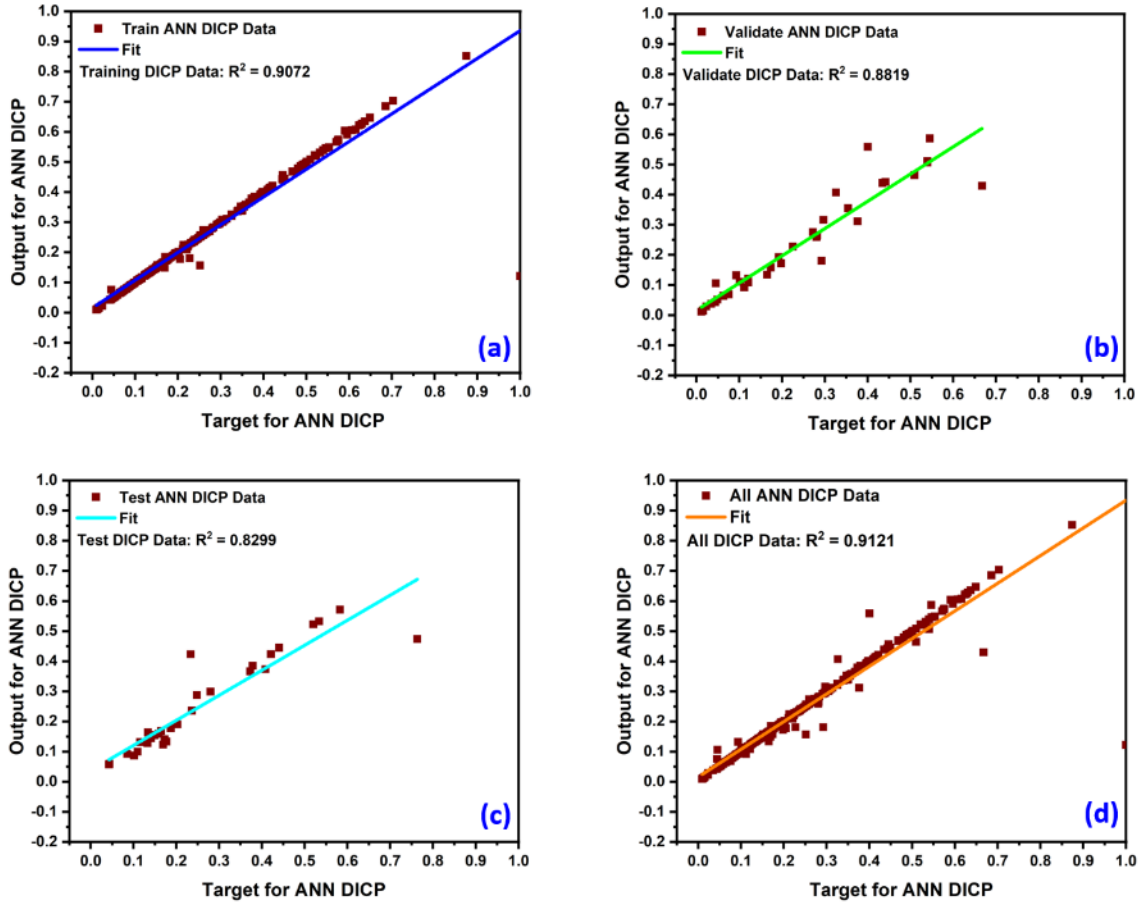



**Figure 2**: Predictive power of the ANN using $R^2$ square metric for the prediction of DICP, **a)** For training data**, b)** For validating the data, **c)** for testing the data, d) for all data.

Since the $R^2$ value for the testing set is rational, the neural network shows good generalization overall for the prediction of DICP using the given dataset (**Fig 2c**). The MSE for the testing data was calculated to be 0.075, which shows the model good performance for the DICP prediction for the testing data. The MSE of the training data for the DICP model was 0.0029. Thus, the test MSE reveals that the model has good generalisation ability and can predict output variables of the test data with good accuracy.

Compared to the DICP, the neural network for CICP gave a high $R^2$ value for the training, validation and testing data **(Fig. 3a, 3b, 3c**). Overall, the $R^2$ value for all the data was quite high (0.99), as presented in **Fig. 3d**. The $R^2$ value for the training data was almost closer to 1;

on the other hand, the $R^2$ value of the testing data (0.9730) was higher than that of the DICP prediction for the neural network. The MSE for the CICP test data was evaluated to be 5.07.

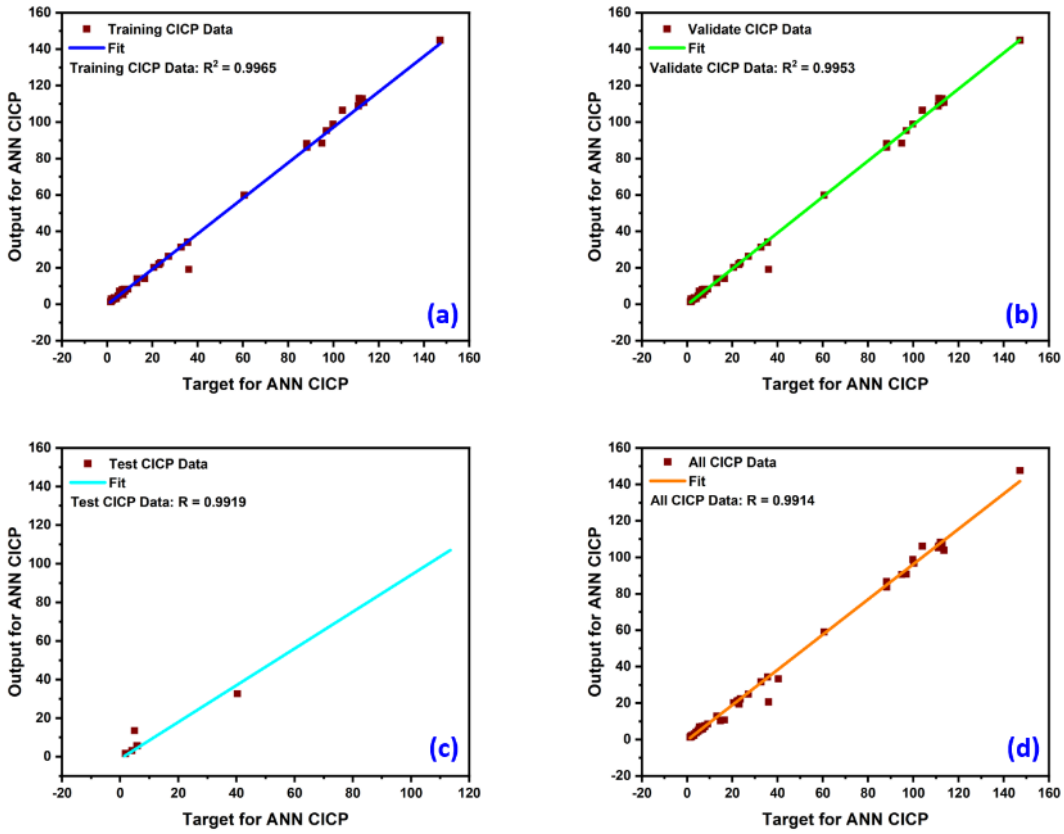



**Figure 3**: Predictive power of the ANN using $R^2$ square metric for the prediction of CICP, **a)** For training data, **b)** For validating the data, **c)** for testing the data, **d)** for all data.

On the other hand, the MSE of the training data for the CICP model was 2.74. The higher MSE for the CICP may be due to the smaller data points for the CICP dataset than the DICP. Since CICP is more prevalent in the PRO mode [42], ANN can be a viable tool to predict the CICP in the FO process with great accuracy In the PRO mode. Further optimisation can improve the accuracy of the ANN by minimising the number of inputs further to the neural network. However, further optimisation was not performed in this study. Although ANN can be a powerful tool and can predict CICP with high accuracy, it should be noted that the ANN

network consists of ten thousand neurons with hidden relationships between the input and output functions. With such modelling, users can get no insights into which feature of the data has a high or low impact on shaping the neural network's output. Further, a major issue in the ANN is the overfitting of the model [43]. Overfitting occurs when the model's training data shows good performance but has poor predictive power for the test data, especially for larger datasets. It should also be emphasised that the modulus of the ICP in the FO process does not give much information alone unless we know what input parameters to the FO process affect its severity. Therefore, it is important to evaluate other advanced models to better predict the ICP in the FO mode and the CICP in the PRO mode.

### 3.2. Prediction of DICP and CICP using Gradient tree Boosting models

Different tree-based models were also evaluated in this study for the prediction of DICP and CICP data. In tree-based machine learning models, the residuals are utilised to correct the previous prediction at each iteration of gradient boosting, allowing the stated loss function to be optimised [44]. Three different tree-based models, XGBoost, CatBoost and Random Forest, were evaluated to predict DICP and CICP modulus, respectively. The initial model was implemented using Extreme Gradient boosting, also known as XGBoost, to predict DICP and CICP and compared to the optimised ANN. The same dataset was used for the tree models like the one used for optimised ANN (removing FS CFV and FS molarity). XGBoost is an ensemble learning algorithm that uses boosted trees, and by weighting, the learning rate aids in shrinking the boosting process, making fitting more conservative [45]. The main advantage of using this algorithm is that XGBoost can provide insights into the importance of various inputs that impact the ICP in the FO process. The XGBoost algorithm also offers several other advantages in ML, such as cross-validation, high flexibility and speed of operation compared to the neural network approach. For XGBoost, 80 % of the data was used for training, and the

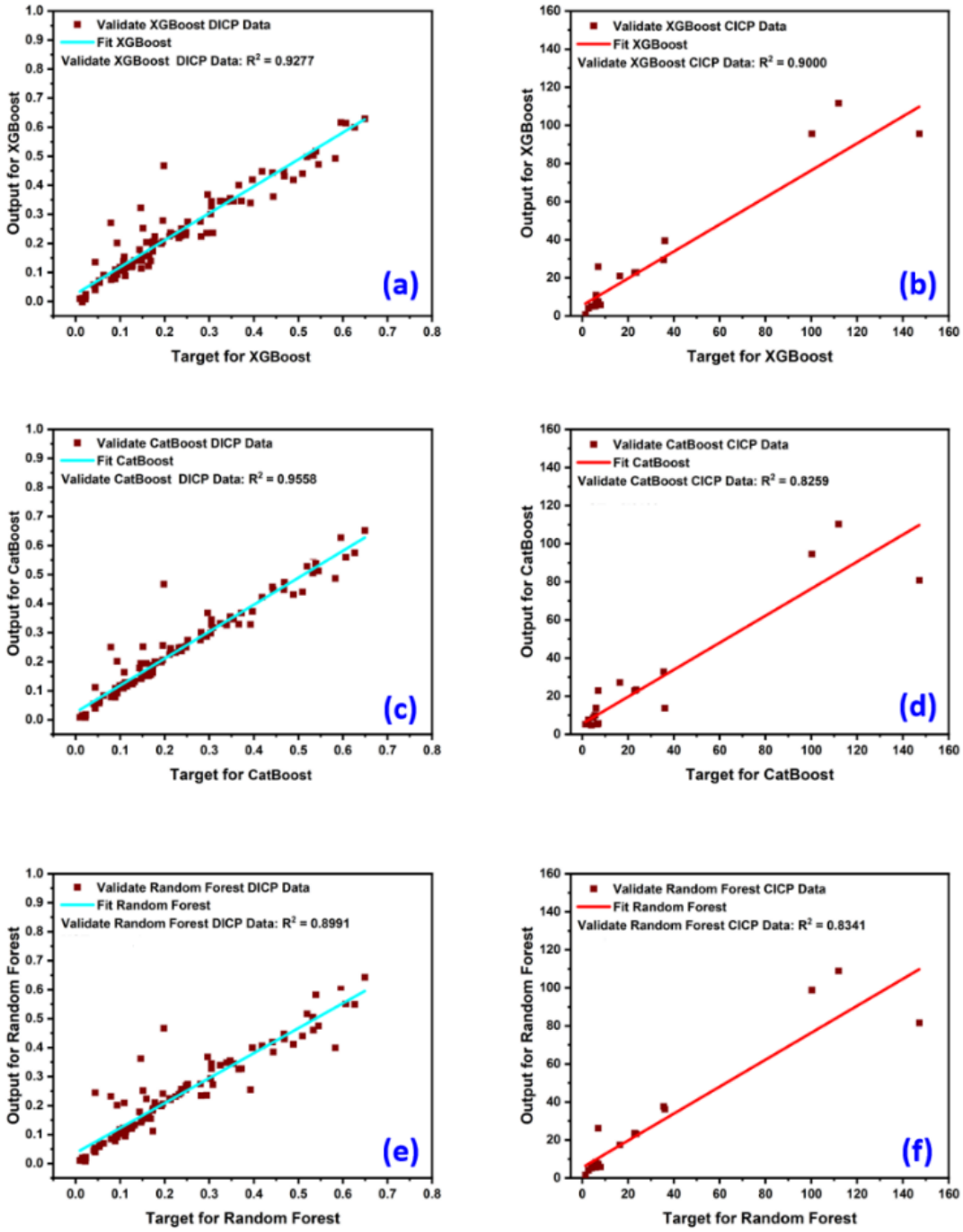
318 remaining 10 % was used for validation and testing (10% for validation, 10 % for testing). The
319 models were trained on the training set and the parameters were adjusted based on their
320 performance on the validation set. The predictive power was evaluated based on the metrics
321 of validation and test data $R^2$ and MSE. The results from the XGBoost simulation
322 demonstrated that the model using this dataset is rational for predicting DICP with an $R^2$ value
323 of 0.94 for the training data, 0.92 for the validation data and 0.81 for the test data (**Fig.4a**).
324 The MSE for DICP prediction on the test data was 0.076, confirming the model's efficient
325 predictive power for the DICP data. On the other hand, the $R^2$ value for the prediction of CICP
326 was 0.92 for the validation data and 0.88 for the test data **(Fig.4b)**. For the prediction of the
327 DICP data, the XGBoost shows almost similar predictive power as the ANN in terms of $R^2$ and
328 MSE; however, the test data $R^2$ for this model for the CICP model was lower than the neural
329 network modelling for the CICP, and hence it was not optimised any further. It should be
330 noted that XGBoost performs better than ANN generally in terms of computational time.
331 Furthermore, XGBoost requires less coding and hyperparameter tuning compared to the ANN
332 program implemented in this study.

333 A slightly different approach to XGBoost is the categorical boosting, also known as the
334 CatBoost algorithm, which was evaluated to predict DICP and CICP in this study. The main
335 advantage of Categorical boosting or CatBoost over other models is that it can learn from its
336 mistakes during the learning phase of the simulation. Furthermore, it is easy to implement,
337 has good stability, has less computational requirements and workload [46], and can
338 sometimes outperform XGBoost in terms of accuracy. Furthermore, we do not have to pre-
339 process any categorical variables such as membrane orientation, types of draw and feed
340 solutions, flow arrangement and membrane types. The basic idea behind boosting is to
341 combine numerous weak models sequentially and create a strong competitive predictive

342 model via greedy search. Unlike other models, this ML method does not require large data to
343 learn from and perform efficiently, even with a smaller dataset. Similar to the XGBoost
344 simulation, the data was split into 80 % training data, and the remaining 10 % was used to
345 validate and test the model on a dataset not seen by the model. It should be noted that the
346 model simulated here was done using minimal feature engineering. **Fig. 4c** and **Fig.4d** show
347 the model performance for DICP and CICP, respectively. For both the CICP and DICP, the
348 learning rates of the models were set to default. For CICP, the learning rate obtained from the
349 program was 0.17065, whereas for the DICP model, the learning rate was 0.2142. The best
350 performance was obtained for DICP predictions with an $R^2$ value of 0.99 for the training data,
351 0.95 for validation data, and 0.92 for the test data. The MSE for the DICP model was 0.049,
352 which shows the model superiority over the Xgboost algorithm and the ANN. The $R^2$ value of
353 CICP was also superior to both ANN and XGBoost ($R^2$=0.99 for training data and 0.99 for the
354 test data). It should be noted that the $R^2$ can be boosted further by doing some feature
355 engineering, later discussed in the next section. For both the DICP prediction and CICP, the
356 CatBoost provides the highest $R^2$ and lower MSE compared to the ANN and Xgboost.

357 To overcome the limitations of boosting ML algorithms for the prediction of ICP, the study
358 checked the validity of the Random Forest regressor, which uses multiple decision trees (that
359 is why called forest) and can perform both regression and classification tasks using a
360 technique known as bagging. Instead of relying on individual decision trees, the basic idea of
361 Random Forest ML is to combine multiple decision trees to determine the final output, in this
362 case, CICP or DICP. Random Forests, in general, produce better results, perform well on large
363 amounts of data, and work with missing data by generating estimates for it. Further, it avoids
364 the overfitting of the model since overfitting is one of the most serious problems in machine
365 learning. Still, it is rarely a problem with the random forest method. If the forest contains a

sufficient number of trees, the classifier will avoid overfitting the model. The RandomForest regressor was used in Python for fitting the Random Forest regression to the forward osmosis data. The ". fit ()" function fits the training input and output data to the RandomForest regressor. The hyperparameter of the model (n_estimators) was set to 100. This is simply the number of trees the algorithm constructs prior to performing maximum voting or calculating prediction averages. In general, increasing the number of trees improves performance and increases the stability of the predictions. **Fig. 4e** and **4f** show the $R^2$ value for the DICP and CICP for the Random Forest Regressor with 0.81 $R^2$ for the DICP test data and 0.93 for the CICP test data. The MSE for the DICP and CICP calculated on the test data for this model was 0.076 and 12.16, respectively. Overall, the Random Forest regression provides an acceptable $R^2$ and MSE score for CICP and DICP; however, the DICP $R^2$ score obtained is less than the CatBoost model. On the other hand, the MSE for the DICP of this model is also slightly higher than the CatBoost. While the model simulation performance was almost comparable with the Xgboost and ANN, it should be noted that if there are many trees in a random forest model, the model's predictive power can become very slow and inefficient, hindering the model application in the real world. For this reason, the model was not further optimised. Table 3 lists the tree-based algorithm training data $R^2$, testing data $R^2$ and MSE of the testing data.




385

386 **Figure 4:** Predictive power of the different gradient tree models using $R^2$ square metric for
387 validation data **a)** XGBoost for validating the data of DICP, **b)** XGBoost for validating the data
388 of CICP, **c)** CatBoost for validating the data of DICP, **d)** CatBoost for validating the data of CICP,
389 **e)** Random Forest for validating the data of DICP, **f)** Random Forest for validating the data of
390 CICP.

391 Table 3: Tree-based algorithms $R^2$ for training, validation and testing data and MSE for testing
392 data

| | Training Data $R^2$ | Testing Data $R^2$ | MSE (Test Data) |
|---|---|---|---|
| XGBoost DICP | 0.94 | 0.81 | 0.0771 |
| XGBoost CICP | 0.99 | 0.88 | 13.32 |
| CatBoost DICP | 0.99 | 0.92 | 0.049 |
| CatBoost CICP | 0.99 | 0.99 | 3.61 |
| Random Forest DICP | 0.96 | 0.81 | 0.076 |
| Random Forest CICP | 0.98 | 0.90 | 12.16 |

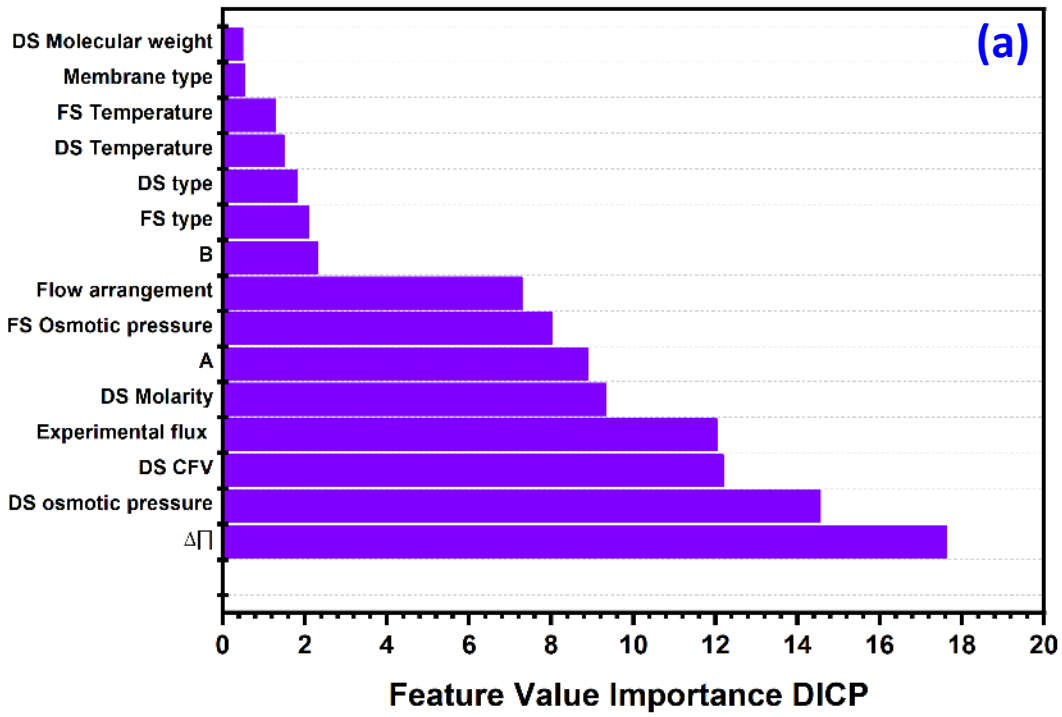
393

394 **3.3. Feature importance Scores for the Catboost model**

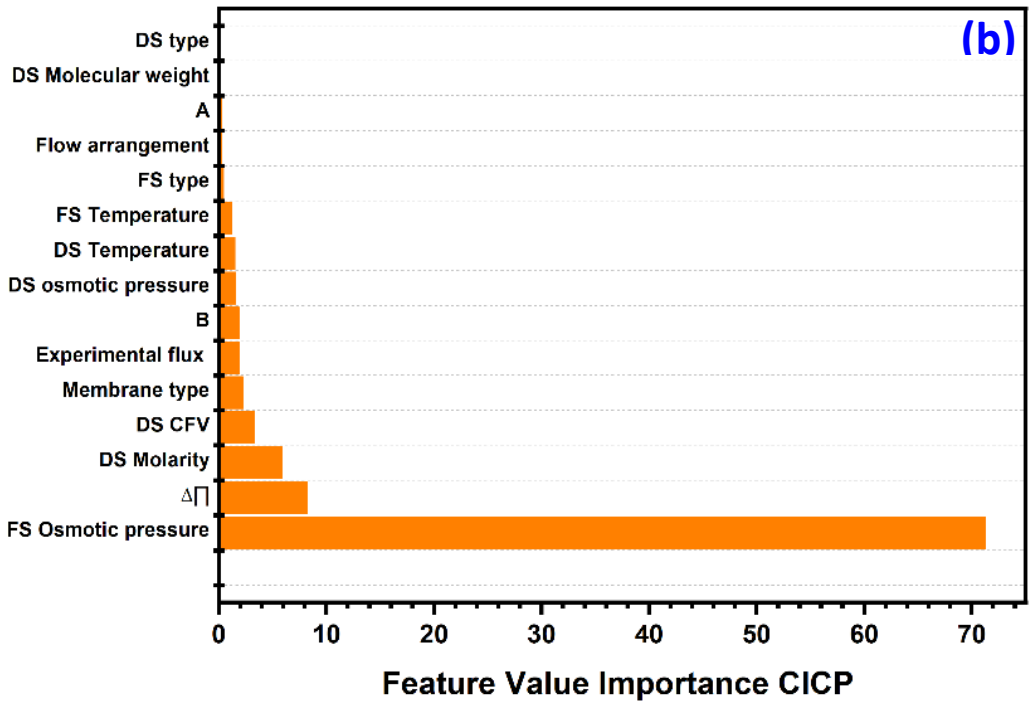
395 Tree-based algorithms such as Xgboost, CatBoost and Random Forest can provide further
396 insights about the importance of the input parameters or the input features using a metric
397 called "Feature Importance" or FI score. Since there are many complex relationships between
398 the input parameters to the FO process and the ICP, the importance of various input
399 parameters should be considered when building a model for ICP. This will allow us to improve
400 the model further for accurate predictions and only consider important input parameters.
401 Feature importance was evaluated for the best performing model, CatBoost. **Fig. 5a** and **Fig.**
402 **5b** present the FI for the training data of CatBoost for DICP and CICP, respectively. For DICP,
403 it can be observed that the highest score is for the osmotic pressure difference (FI score of
404 17.64) followed by the DS osmotic pressure ( FI score of 14.55). The result implies that the
405 osmotic pressure difference and DS osmotic pressure values have a significant impact on the
406 model, whereas the features with the lowest FI score values have a small impact on the
407 modelling performance, such as DS molecular weight and membrane type. Generally, the
408 highest the osmotic pressure difference between the FS and the DS, the higher the dilution

409 and thus the higher the DICP value. On the other hand, the higher the DS osmotic pressure,
410 the higher the permeation drag of the DS enabling more water transport across the FO
411 membrane, and therefore, more dilution of the DS, leading to higher values of the DICP. DS
412 CFV and experimental flux have also a significant impact on the DICP. For instance, higher DS
413 CFV will result in increased flux in the forward osmosis process, and thus reducing the impacts
414 of DICP. Recent studies have also confirmed that increasing CFV will result in minimizing the
415 impacts of DICP as well as CICP [15]. For CICP (**Fig.5b**), the highest score is for the osmotic
416 pressure of the FS and is significantly higher than all the other features. Generally, when the
417 FS is DI water (FS osmotic pressure=0), a higher water flux is obtained in the PRO mode, as
418 there is no CICP on the FS of the membrane. On the other hand, when the FS in the PRO mode
419 is FS with high osmotic pressure, such as wastewater, the flux in the PRO mode is less than
420 the FO mode due to the severity of the CICP. The lowest FI score for the CICP CatBoost was
421 obtained for the DS type and the DS molecular weight. The lowest FI score for the CICP
422 CatBoost model were obtained for the DS type and the DS molecular weight. As CICP occurs
423 on the FS of the FO membrane, the DS type and the DS molecular weight may have a negligible
424 effect on the CICP. DS weight and the type of DS shows to have no significant impact on the
425 modelling of the CICP process, and hence, these can be ignored when modelling for the CICP
426 process in the PRO mode. This will allow for more accurate predictions of CICP in the FO
427 process in the PRO mode.

428




429




430

431 **Figure 5: a**) Feature importance score of the CatBoost model for DICP, b**)** Feature importance

432 score of the CatBoost model for CICP.

These can be made redundant when modelling the effects of CICP on the FO process. As CICP occurs on the FS of the FO membrane, the DS type and the DS molecular weight may have a negligible effect on the CICP.

The CatBoost FI score optimises the model by including the highest scores and neglecting those with the lowest scores. This enables the ML algorithm to predict the output more accurately. Features that are unimportant or relevant to the mode introduce noise in the modelling process, decreasing accuracy [47]. Since ML-based predictive models do not perform well when exposed to noisy data, particularly in the training phase. Furthermore, the model built on noisy data tends not to generalise well to data outside the training set, possibly leading to over-fitting the model. It is interesting to see that the feature importance of the CatBoost model for the test data is similar to the one obtained for the training data. **Fig. A.4** and **Fig.A.5 (Appendix A.1)** present the feature importance for DICP and CICP, respectively, for the CatBoost model, respectively. The feature importance is an important metric in the CatBoost model, and it tells us how much the prediction in CICP or DICP would change if any of the feature's change. Even though we use all of the features of the given dataset based on self-collected user data, not all of them contribute equally to the prediction of the CICP and the DICP. Modelling the most important features of the predictive model can lead us to a high $R^2$ value and lower MSE for the CatBoost algorithm. The model for the DICP or CICP can be further optimised by cleaning the data and removing the features with a very low score, such as type of membrane and DS molecular weight.

### 3.4. Prediction of additional test samples for CatBoost

**Fig.6a** and **Fig.6b** present additional test samples tested on the CatBoost model for DICP and optimised CatBoost model for CICP. The model shows very good agreement with the actual

values. It should be noted here that the test data here was outside of the training dataset (not used for training data or unseen data), so overall; the model exhibits very good generalisation.

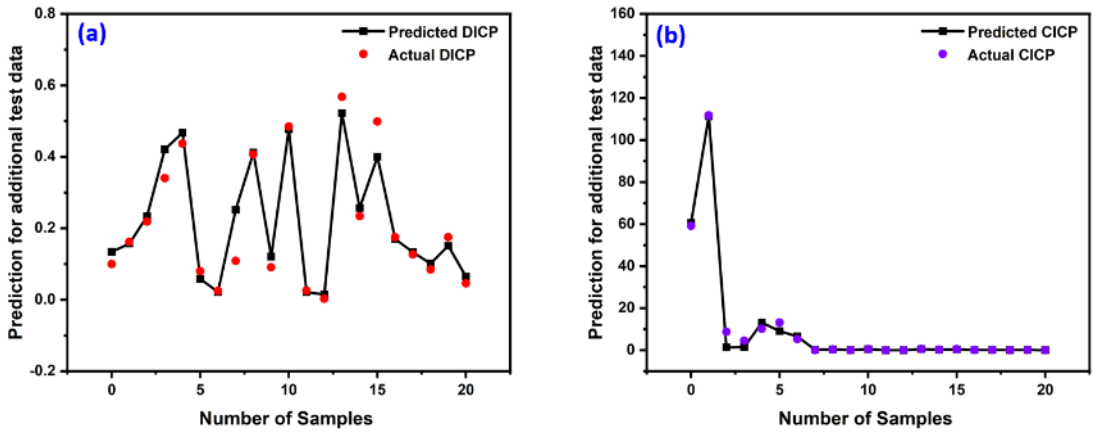



**Figure 6**: The predictive power of the CatBoost model using $R^2$ square metric **a)** For validating the data of DICP, **b)** For validating the data of CICP.

### 3.5. Linear Regression

Linear regression was also evaluated in Python to test the prediction of the ICP. The $R^2$ value for the DICP model for linear regression was 0.48, and the lowest $R^2$ value amongst all the models of 0.27 was for the CICP prediction (**Fig.7a** and **Fig.7b)**. This implies that the predictive power of linear regression is very bad for the CICP prediction and is lower than CatBoost for ICP predictions. This could be due to the linear regression model's inability to predict the ICP with different inputs with no linear relationships. As a result, the $R^2$ is low, resulting in inaccurate results.

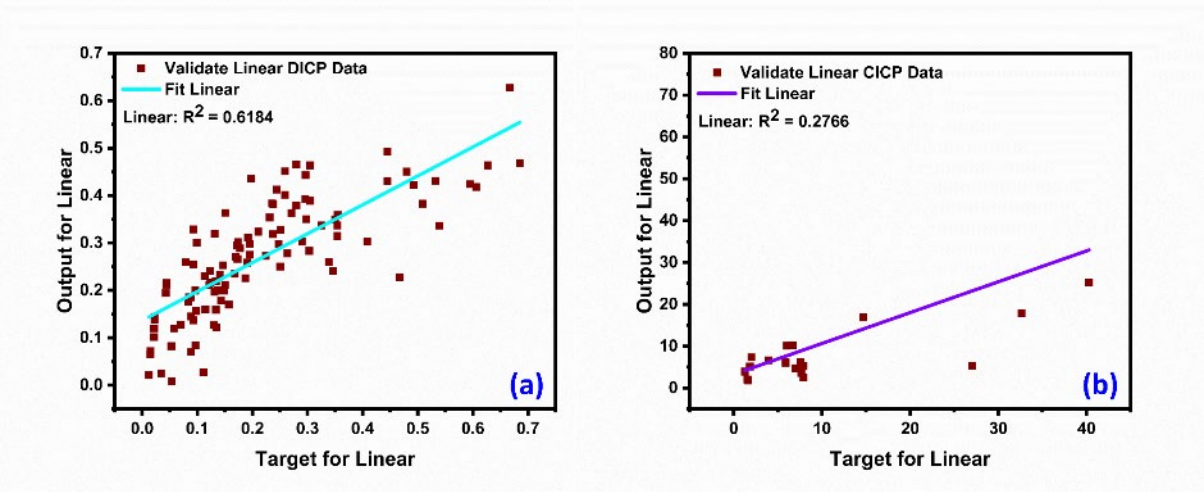



**Figure 7:** Predictive power of the linear regression model using $R^2$ square metric, **a)** For validating the data of DICP, **b)** For validating the data of CICP

### 3.6. Summary of Comparison for ML models

Amongst all the models, the best performance was obtained by CatBoost, followed by ANN. **Fig.8a** and **8b** show the comparison of the MSE and $R^2$ value for each model except linear regression, which was excluded since the performance was very poor. It should be emphasised that ANN requires data normalisation compared to CatBoost, which can handle categorical values such as membrane orientation, DS type, FS type, flow direction and type of membrane used in the FO study. CatBoost also has the additional ability to handle missing values or information in the data by using estimates. Finally, in terms of execution and design, CatBoost is much efficient and requires less computational and processing time.

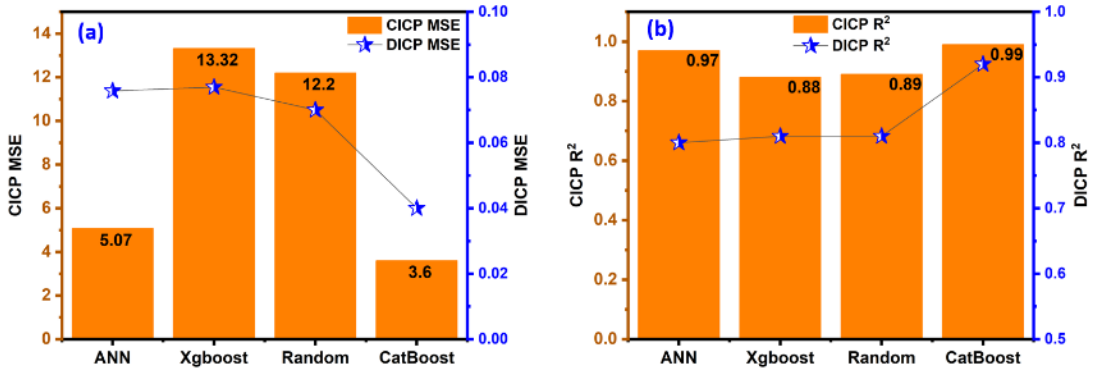



**Figure 8:** Comparison of mean square error (MSE) and $R^2$ value for different models for CICP and DICP, **a)** MSE for CICP and DICP, **b)** $R^2$ value comparison for different models for CICP and DICP.

### 3.7. Comparison of CatBoost vs prediction by literature models

The prediction of the CatBoost was also compared with the solution diffusion model to show the applicability of the ML models for CP prediction. As discussed before, the CatBoost model for CICP was optimised further to improve the accuracy. The initial CatBoost model had a good score for the prediction of DICP and the CICP in terms of $R^2$ and MSE. Although the evaluation metrics can be further optimised with hyperparameter tuning and data cleaning, the initial CatBoost model was not optimised further as the performance was efficient compared to other models. Furthermore, major feature engineering is avoided here so that environmental engineers do not have to apply complex codes, which otherwise would require advanced programming skills in Python. The results are presented in **Table 4** with the absolute error between the values obtained for the solution diffusion model and the CatBoost algorithm.

**Table 4:** Additional test data for comparison against SD models.

| **AL-DS Mode, Temperature of FS and DS =25, CFV of FS and DS = 0.36 Concentrative ICP Modulus** | | | | | | |
|---|---|---|---|---|---|---|
| DS Osmotic Pressure atm | FS Osmotic pressure atm | Water flux m/s | K (experimental) s/m | K (average) s/m | CICP SD | CICP CatBoost |
| 70.04 | 2.33 | 8.82E-06 | 2.02E+05 | 2.2 E05 | 5.93 | 7.2 |
| 70.04 | 4.67 | 7.35E-06 | 2.12 E+05 | 2.2 E05 | 4.75 | 5.18 |
| 70.04 | 23.35 | 3.11E-06 | 2.36 E+05 | 2.2 E05 | 2.08 | 2.00 |
| 70.04 | 46.70 | 1.14E-06 | 2.47 E+05 | 2.2 E05 | 1.32 | 1.29 |
| **AL-FS Mode, Temperature of FS and DS = 25, CFV of FS and DS = 45.3 cm/s** | | | | | **DICP SD** | **DICP CatBoost** |
| 70.04 | 0.00 | 5.05E-06 | 2.83E05 | 2.7E05 | 0.25 | 0.25 |
| 70.04 | 3.06 | 4.70E-06 | 2.80E05 | 2.7E05 | 0.28 | 0.28 |
| 70.04 | 6.00 | 4.36E-06 | 2.81E05 | 2.7E05 | 0.30 | 0.30 |
| 70.04 | 27.11 | 2.59E-06 | 2.56E05 | 2.7E05 | 0.49 | 0.47 |
| 70.04 | 49.63 | 1.06E-06 | 2.52E05 | 2.7E05 | 0.75 | 0.71 |

It is evident that the ML method used here can predict the internal concentration polarization based on the process input parameters. A user can simulate and predict the ICP in the FO process, as long as the membrane's operating parameters and the A and B are known. Changes in the operating parameters would also reflect the corresponding changes in ICP and the process performance. This will allow for improvement in membrane design without long

experimentation. On the other hand, with the solution diffusion model, modelling and prediction are done using a constant value of solute resistance to diffusion "***K***".

## 4. Conclusion

Accurate measurement and mitigation of internal concentration polarization are vital for the FO process to compete against the existing state of the art RO and other technologies. ML models can accurately predict ICP in the FO process based on the input variables or process conditions. The following are the main findings from this study.

a) ML models can predict both dilutive and concentrative ICP in the FO process without the need for structural or solute resistance to diffusion inputs.
b) The models can also be applicable where the diffusivity of the draw solution is unknown.
c) Data captured from ML models can be used to optimize the FO process at the pilot scale.
d) The CatBoost algorithm best predicts dilutive and concentrative internal polarization in the FO process among the different ML models used in this study.
e) The feature importance score in the ML models can help select the best input variables for modelling and redesign the model based on the score.
f) The CatBoost model can predict ICP with an $R^2$ value of 0.92 and has more predictive power than the ANN.

The model designed here is solely reliant on the input or trained data collected from the literature. The accuracy of the data is directly proportional to the accuracy of the model. Any discrepancies in the methodologies of the user data from the literature studies might lead to inaccurate predictions in the DICP or CICP occurring in the FO process. ML models also perform well when there is a large amount of data available.

**Acknowledgements**

Thanks to the Australian government for providing a research training scholarship to Ibrar.

**References**

[1] S. Yadav, H. Saleem, I. Ibrar, O. Naji, A.A. Hawari, A.A. Alanezi, S.J. Zaidi, A. Altaee, J. Zhou, Recent developments in forward osmosis membranes using carbon-based nanomaterials, Desalination, 482 (2020) 114375.
[2] S. Yadav, I. Ibrar, S. Bakly, D. Khanafer, A. Altaee, V. Padmanaban, A.K. Samal, A.H. Hawari, Organic Fouling in Forward Osmosis: A Comprehensive Review, Water, 12 (2020) 1505.
[3] B. Corzo, T. de la Torre, C. Sans, R. Escorihuela, S. Navea, J.J. Malfeito, Long-term evaluation of a forward osmosis-nanofiltration demonstration plant for wastewater reuse in agriculture, Chemical Engineering Journal, 338 (2018) 383-391.
[4] S.K. Singh, C. Sharma, A. Maiti, A comprehensive review of standalone and hybrid forward osmosis for water treatment: Membranes and recovery strategies of draw solutions, Journal of Environmental Chemical Engineering, 9 (2021) 105473.
[5] I. Ibrar, S. Yadav, A. Altaee, A.K. Samal, J.L. Zhou, T.V. Nguyen, N. Ganbat, Treatment of biologically treated landfill leachate with forward osmosis: Investigating membrane performance and cleaning protocols, Science of The Total Environment, 744 (2020) 140901.
[6] I. Ibrar, O. Naji, A. Sharif, A. Malekizadeh, A. Alhawari, A.A. Alanezi, A. Altaee, A Review of Fouling Mechanisms, Control Strategies and Real-Time Fouling Monitoring Techniques in Forward Osmosis, Water, 11 (2019) 695.
[7] S. Yadav, I. Ibrar, A. Altaee, S. Déon, J. Zhou, Preparation of novel high permeability and antifouling polysulfone-vanillin membrane, Desalination, 496 (2020) 114759.
[8] G.T. Gray, J.R. McCutcheon, M. Elimelech, Internal concentration polarization in forward osmosis: role of membrane orientation, Desalination, 197 (2006) 1-8.
[9] D. Khanafer, I. Ibrahim, S. Yadav, A. Altaee, A. Hawari, J. Zhou, Brine reject dilution with treated wastewater for indirect desalination, Journal of Cleaner Production, (2021) 129129.
[10] J.R. McCutcheon, M. Elimelech, Influence of concentrative and dilutive internal concentration polarization on flux behavior in forward osmosis, Journal of Membrane Science, 284 (2006) 237-247.
[11] C.H. Tan, H.Y. Ng, Modified models to predict flux behavior in forward osmosis in consideration of external and internal concentration polarizations, Journal of Membrane Science, 324 (2008) 209-219.
[12] C.H. Tan, H.Y. Ng, Revised external and internal concentration polarization models to improve flux prediction in forward osmosis process, Desalination, 309 (2013) 125-140.
[13] N.Y. Yip, A. Tiraferri, W.A. Phillip, J.D. Schiffman, L.A. Hoover, Y.C. Kim, M. Elimelech, Thin-Film Composite Pressure Retarded Osmosis Membranes for Sustainable Power Generation from Salinity Gradients, Environmental Science & Technology, 45 (2011) 4360-4369.
[14] E. Nagy, A general, resistance-in-series, salt- and water flux models for forward osmosis and pressure-retarded osmosis for energy generation, Journal of Membrane Science, 460 (2014) 71-81.

570 [15] N.-N. Bui, J.T. Arena, J.R. McCutcheon, Proper accounting of mass transfer resistances in
571 forward osmosis: Improving the accuracy of model predictions of structural parameter,
572 Journal of Membrane Science, 492 (2015) 289-302.
573 [16] K.L. Lee, R.W. Baker, H.K. Lonsdale, Membranes for power generation by pressure-
574 retarded osmosis, Journal of Membrane Science, 8 (1981) 141-171.
575 [17] I. Ibrar, S. Yadav, A. Altaee, A. Hawari, V. Nguyen, J. Zhou, A novel empirical method for
576 predicting concentration polarization in forward osmosis for single and multicomponent
577 draw solutions, Desalination, 494 (2020) 114668.
578 [18] M. Kahrizi, N. Kasiri, T. Mohammadi, S. Zhao, Introducing sorption coefficient through
579 extended UNIQAC and Flory-Huggins models for improved flux prediction in forward
580 osmosis, Chemical Engineering Science, 198 (2019) 33-42.
581 [19] A. Sagiv, R. Semiat, Finite element analysis of forward osmosis process using NaCl
582 solutions, Journal of Membrane Science, 379 (2011) 86-96.
583 [20] M.F. Gruber, C.J. Johnson, C.Y. Tang, M.H. Jensen, L. Yde, C. Hélix-Nielsen,
584 Computational fluid dynamics simulations of flow and concentration polarization in forward
585 osmosis membrane systems, Journal of Membrane Science, 379 (2011) 488-495.
586 [21] M. Kahrizi, J. Lin, G. Ji, L. Kong, C. Song, L.F. Dumée, S. Sahebi, S. Zhao, Relating forward
587 water and reverse salt fluxes to membrane porosity and tortuosity in forward osmosis: CFD
588 modelling, Separation and Purification Technology, 241 (2020) 116727.
589 [22] A. Salgado-Reyna, E. Soto-Regalado, R. Gómez-González, F. Cerino-Córdova, R. García-
590 Reyes, M. Garza-González, M. Alcalá-Rodríguez, Artificial neural networks for modeling the
591 reverse osmosis unit in a wastewater pilot treatment plant, Desalination and Water
592 Treatment, 53 (2015) 1177-1187.
593 [23] J. Jawad, A.H. Hawari, S. Zaidi, Modeling of forward osmosis process using artificial
594 neural networks (ANN) to predict the permeate flux, Desalination, 484 (2020) 114427.
595 [24] A.T. Mohammad, M.A. Al-Obaidi, E.M. Hameed, B.N. Basheer, I.M. Mujtaba, Modelling
596 the chlorophenol removal from wastewater via reverse osmosis process using a multilayer
597 artificial neural network with genetic algorithm, Journal of Water Process Engineering, 33
598 (2020) 100993.
599 [25] A. Hosseinzadeh, J.L. Zhou, A. Altaee, M. Baziar, D. Li, Effective modelling of hydrogen
600 and energy recovery in microbial electrolysis cell by artificial neural network and adaptive
601 network-based fuzzy inference system, Bioresource Technology, 316 (2020) 123967-123967.
602 [26] Q.-F. Liu, S.-H. Kim, Evaluation of membrane fouling models based on bench-scale
603 experiments: A comparison between constant flowrate blocking laws and artificial neural
604 network (ANNs) model, Journal of Membrane Science, 310 (2008) 393-401.
605 [27] I. Ibrar, S. Yadav, N. Ganbat, A.K. Samal, A. Altaee, J.L. Zhou, T.V. Nguyen, Feasibility of
606 H2O2 cleaning for forward osmosis membrane treating landfill leachate, Journal of
607 Environmental Management, 294 (2021) 113024.
608 [28] A. K, A. Mungray, S. Agarwal, J. Ali, M. Chandra Garg, Performance optimisation of
609 forward-osmosis membrane system using machine learning for the treatment of textile
610 industry wastewater, Journal of Cleaner Production, 289 (2021) 125690.
611 [29] A. Hosseinzadeh, J.L. Zhou, A. Altaee, M. Baziar, X. Li, Modeling water flux in osmotic
612 membrane bioreactor by adaptive network-based fuzzy inference system and artificial
613 neural network, Bioresource Technology, 310 (2020) 123391.
614 [30] E. Nagy, I. Hegedüs, E.W. Tow, J.H. Lienhard V, Effect of fouling on performance of
615 pressure retarded osmosis (PRO) and forward osmosis (FO), Journal of Membrane Science,
616 565 (2018) 450-462.

[31] J.-G. Gai, X.-L. Gong, Zero internal concentration polarization FO membrane: functionalized graphene, Journal of Materials Chemistry A, 2 (2014) 425-429.
[32] Q. She, R. Wang, A.G. Fane, C.Y. Tang, Membrane fouling in osmotically driven membrane processes: A review, Journal of Membrane Science, 499 (2016) 201-233.
[33] C.-Y. Wu, H. Mouri, S.-S. Chen, D.-Z. Zhang, M. Koga, J. Kobayashi, Removal of trace-amount mercury from wastewater by forward osmosis, Journal of Water Process Engineering, 14 (2016) 108-116.
[34] G.Q. Chen, S.L. Gras, S.E. Kentish, The application of forward osmosis to dairy processing, Separation and Purification Technology, 246 (2020) 116900.
[35] Y. Chun, S.-J. Kim, G.J. Millar, D. Mulcahy, I.S. Kim, L. Zou, Forward osmosis as a pre-treatment for treating coal seam gas associated water: Flux and fouling behaviour, Desalination, 403 (2017) 144-152.
[36] K. Lutchmiah, A.R.D. Verliefde, K. Roest, L.C. Rietveld, E.R. Cornelissen, Forward osmosis for application in wastewater treatment: A review, Water Research, 58 (2014) 179-197.
[37] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, N. Khovanova, Machine Learning for Predictive Modelling based on Small Data in Biomedical Engineering, IFAC-PapersOnLine, 48 (2015) 469-474.
[38] D.C. Elton, Z. Boukouvalas, M.S. Butrico, M.D. Fuge, P.W. Chung, Applying machine learning techniques to predict the properties of energetic materials, Scientific Reports, 8 (2018) 9059.
[39] S. Badirli, X. Liu, Z. Xing, A. Bhowmik, K. Doan, S.S. Keerthi, Gradient boosting neural networks: Grownet, arXiv preprint arXiv:2002.07971, (2020).
[40] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics, (2001) 1189-1232.
[41] C.S.H. Yeo, Q. Xie, X. Wang, S. Zhang, Understanding and optimization of thin film nanocomposite membranes for reverse osmosis with machine learning, Journal of Membrane Science, 606 (2020) 118135.
[42] A. Altaee, A. Sharif, Pressure retarded osmosis: advancement in the process applications for power generation and desalination, Desalination, 356 (2015) 31-46.
[43] F. Ghasemi, A. Mehridehnavi, A. Pérez-Garrido, H. Pérez-Sánchez, Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks, Drug Discovery Today, 23 (2018) 1784-1790.
[44] D. Zhang, L. Qian, B. Mao, C. Huang, B. Huang, Y. Si, A Data-Driven Design for Fault Detection of Wind Turbines Using Random Forests and XGboost, IEEE Access, 6 (2018) 21020-21031.
[45] X. Shi, Y.D. Wong, M.Z.-F. Li, C. Palanisamy, C. Chai, A feature learning approach based on XGBoost for driving assessment and risk prediction, Accident Analysis & Prevention, 129 (2019) 170-179.
[46] G. Huang, L. Wu, X. Ma, W. Zhang, J. Fan, X. Yu, W. Zeng, H. Zhou, Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions, Journal of Hydrology, 574 (2019) 1029-1041.
[47] M. Rzychoń, A. Żogała, L. Róg, Experimental study and extreme gradient boosting (XGBoost) based prediction of caking ability of coal blends, Journal of Analytical and Applied Pyrolysis, 156 (2021) 105020.

662

663

664 **Appendix A**

665 **Evaluation of Machine Learning Algorithms to Predict Internal Concentration Polarization**

666 **(ICP) in Forward Osmosis (FO)**

667 Ibrar Ibrar[1], Sudesh Yadav[1], Ali Braytee[2], Ali Altaee[1,] *, Ahmad HosseinZadeh[1], Akshaya K.

668 Samal[3], John L Zhou[1], Jamshed Ali Khan[1], Pietro Bartocci[4,5], Francesco Fantozzi[4]

669 1: School of Civil and Environmental Engineering, University of Technology Sydney, Ultimo,

670 NSW 2007, Australia

671 2: School of Computer Science, University of Technology Sydney, Ultimo, NSW 2007, Australia

672 3: Centre for Nano and Material Sciences, Jain University, Bangalore - 562112, India

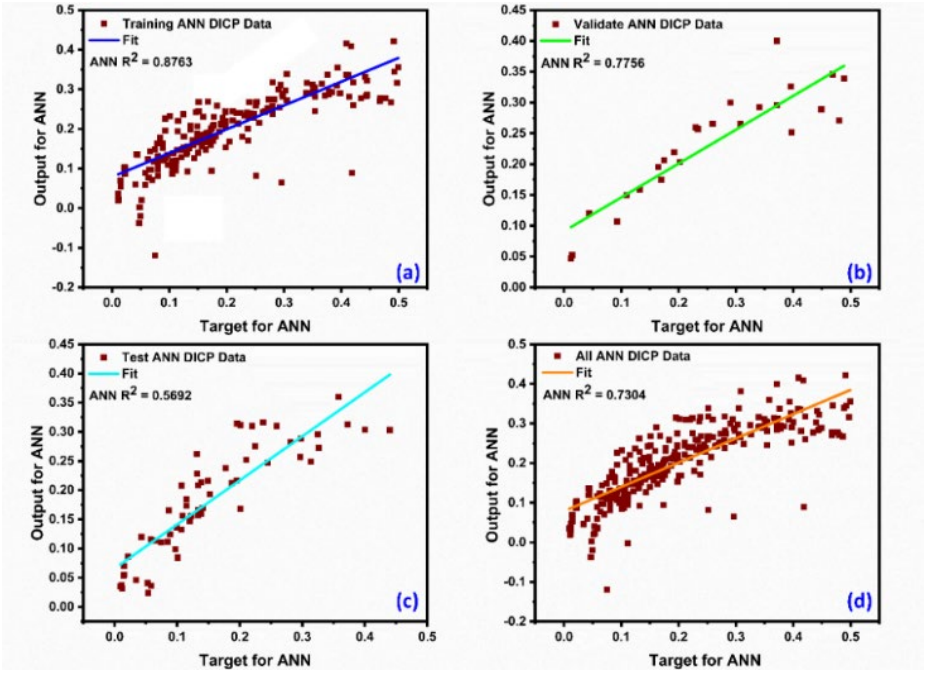
673 *Corresponding author email: *Ali.altaee@uts.edu.au*

674 4: Department of Engineering, University of Perugia, Via G. Durant 67, Perugia 06125, Italy

675 5: Instituto de Carboquímica (C.S.I.C.), Miguel Luesma Castán 4, 50018 Zaragoza, Spain

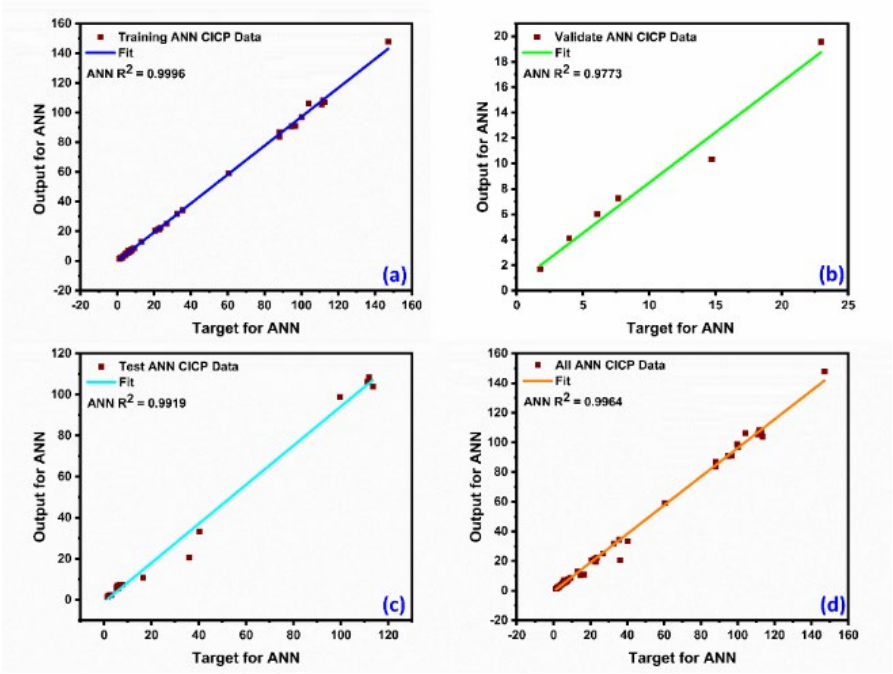
676 **A.1. Prediction of DICP and CICP with ANN without optimisation**

677 Initially, all inputs were considered without any consideration, and the ANN performance for
678 the training, validation data, testing data and all the data is presented here in Fig.A.1 and
679 Fig.A.2. For the DICP data, the neural network was prone to overfitting. However, the
680 performance was acceptable for the CICP data.




681

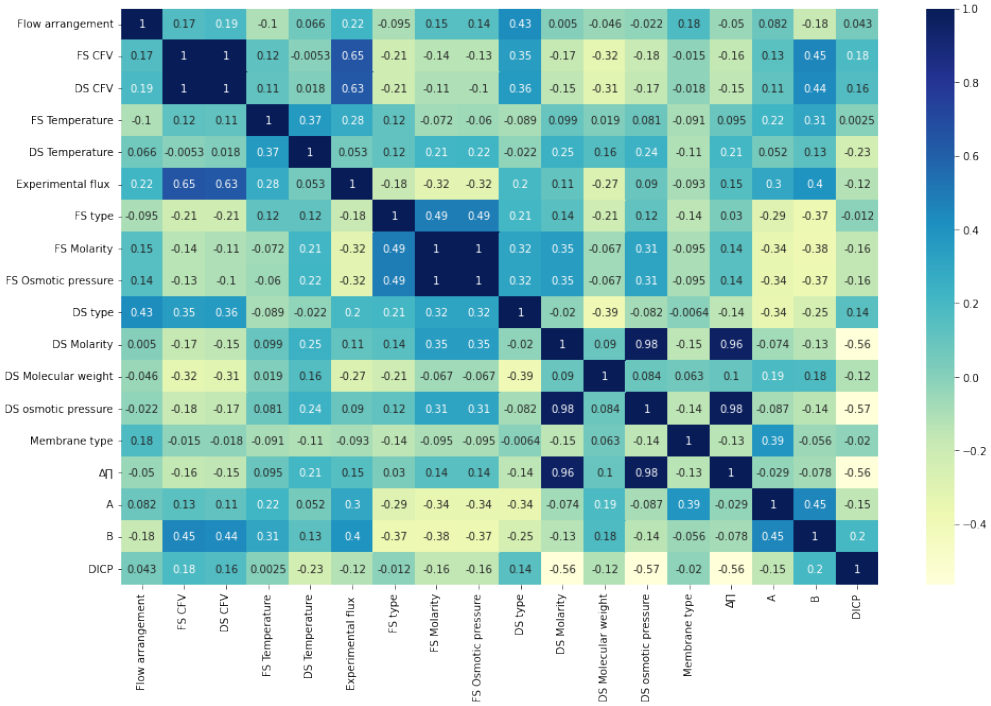
682 **Figure A.1.**: Predictive power of the ANN (no optimisation) using $R^2$ square metric for the
683 prediction of DICP, **a)** For training data**, b)** For validating the data, **c)** for testing the data, d)
684 for all data.




685

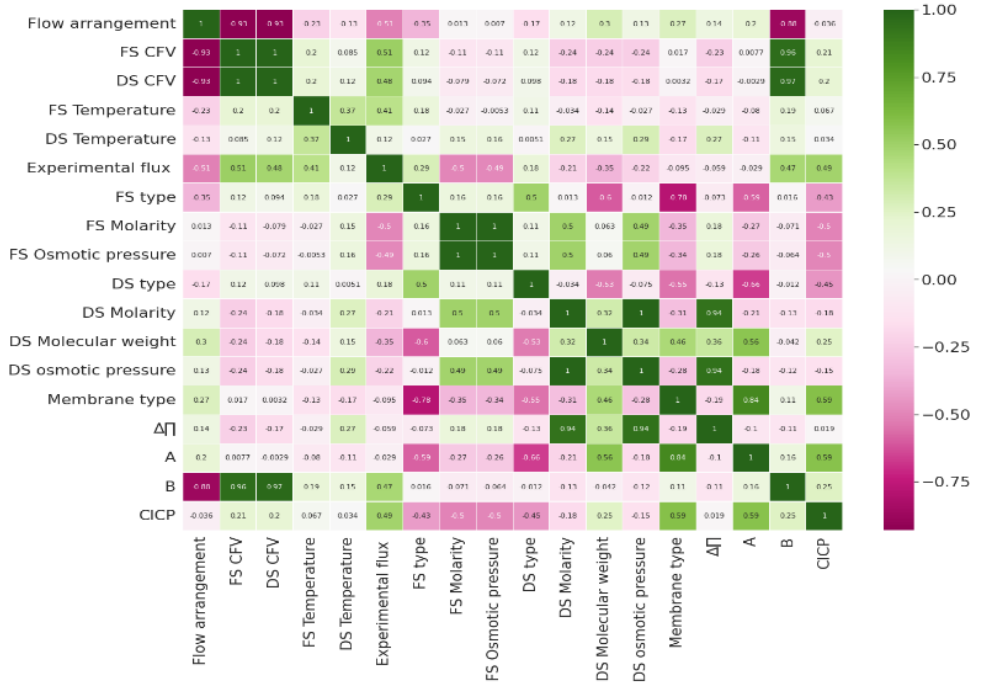
686 **Figure A.2.**: Predictive power of the ANN (no optimisation) using $R^2$ square metric for the
687 prediction of CICP, **a)** For training data**, b)** For validating the data, **c)** for testing the data, d)
688 for all data.

689




690

691 **Figure A.3:** Heat map of the input parameters to the FO process for DICP model.




692

693 **Figure A.4:** Heat map of the input parameters to the FO process for CICP model.

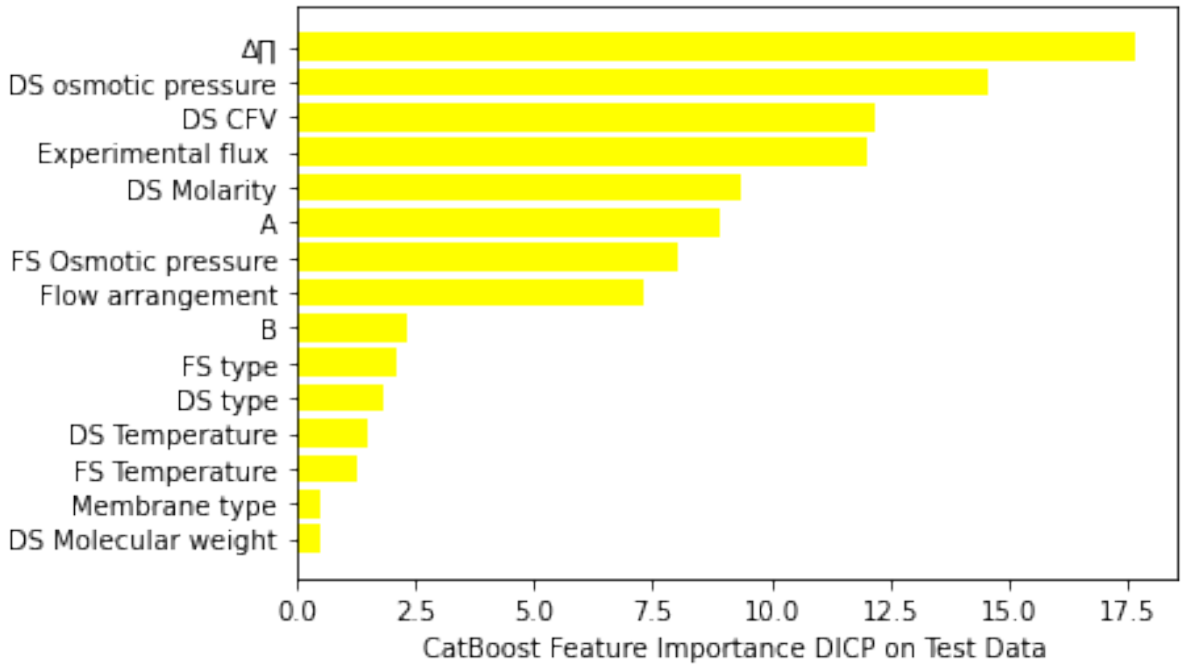



**Figure. A.5. Catboost Feature Importance based on the test dataset for the DICP predictions, (Figure directly exported from the program)**

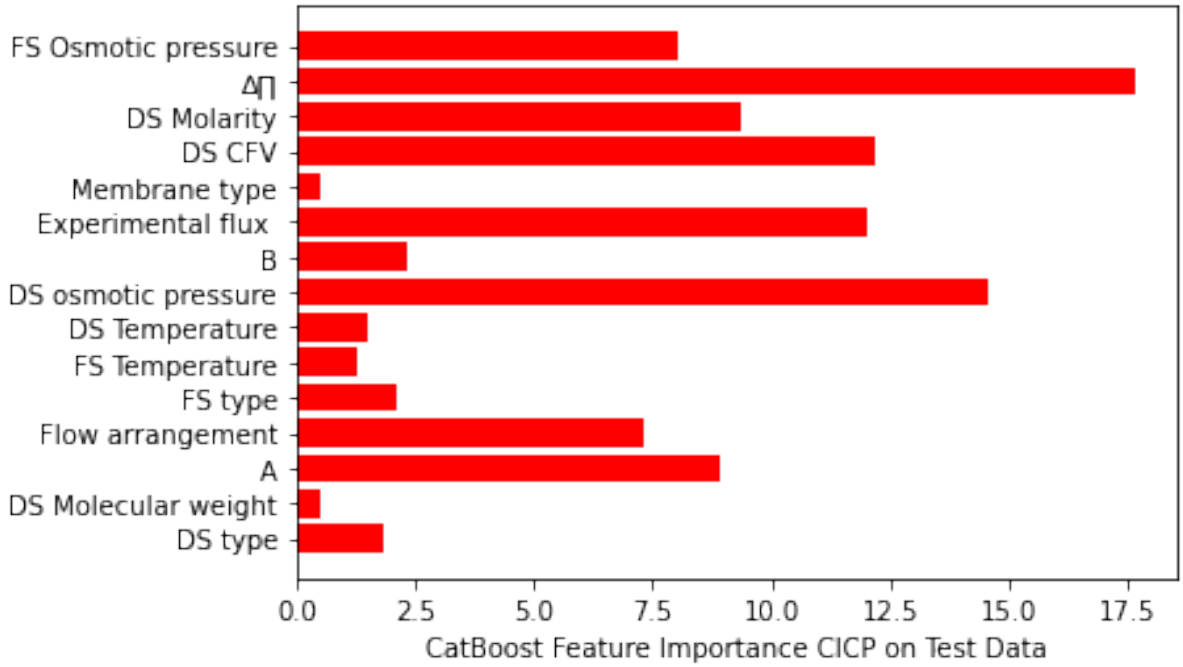



**Figure. A.6. Catboost Feature importance score based on the test data for the CICP predictions (Figure directly exported from the program).**