The modeling of the Organic Molecules Rejection using the Bootstrap Aggregated Neural Networks for the evaluation of the Forward Osmosis Process performance

Fouad Kratbi¹*, Yamina Ammi¹, and Salah Hanini¹

¹Laboratory of Biomaterials and Transport Phenomena (LBMPT), University of Medea, Algeria

*Corresponding author: Email: kratbi.fouad@univ-medea.dz
Phone number: +213 06 71 18 49 86

GRAPHICAL ABSTRACT

Abstract

The forward osmosis process is currently more studied to be a replacement for another consuming-energy process, for this, many works show up the rejection of different molecules, energy consumption, and modeling of different objectives related to FO process. Our study consists to model the rejection of organic molecules (neutral and ionic) by FO process; however, this paper is the simultaneous applications of the single neural network based on quantitative-structure properties relationship (QSPR-SNN) and the bootstrap aggregated neural network (BANN) to predict the rejection of 53 OM. According to the results obtained, the coefficient correlation "$R$" is used to evaluate the performance of each model for the unseen data, the QSPR-BANN gives
$R$ value equal to 0.9909 higher than the value of the SNN which is 0.9401, the Root Mean Square Error of the QSPR-BANN is less than that of the QSPR-SNN with values equal to 0.5764% and 1.2826% respectively.

**Keywords:** Modeling; Organic molecules; Rejection; Bootstrap Aggregated Neural Networks; Forward Osmosis; Membranes.

1. Introduction

Membrane separation processes include a large number of techniques for performing separation in different phases (liquid, gas, and mixed-phase) under the action of different driving forces as the potential difference, the electrical, pressure, or chemical potential (Ammi *et al*., 2021; Cordier *et al*., 2020). Microfiltration (MF), Ultrafiltration (UF), Nanofiltration (NF), and Reverse Osmosis (RO) make the main separation membranes process using the gradient of pressure as the driving forces (Zhao *et al*., 2012; Lutchmiah *et al*., 2014). Recently, many works studied the Forward Osmosis (FO) process as the replacement of some techniques (RO and NF) to reduce the energy consumption used to generate the pressure difference (Ammi *et al*., 2021; Cath *et al*., 2006; Shaffer *et al*., 2014). The literature highlights several points, in the first place, it is essential to explain correctly the process of forward osmosis with the existence of many models of transfer namely; irreversible thermodynamics, solubilization, diffusion, and pores (Gur-Reznik *et al*., 2008; Lanteri *et al*., 2008). Indeed, forward osmosis presents several interactions between solutes, water, and membranes which complicate the explanation of these phenomena (Chung *et al*., 2012; Martinetti *et al*., 2009; McGinnis *et al*., 2007; Zhao *et al*., 2001; Bowen *et al*., 2000; Shetty *et al*., 2003. To solve these problems, new approaches have been proposed to describe nonlinear systems with artificial intelligence which is beneficial by minimization of the number of experiments, reducing financial expenses, saving time, and possibility of modeling systems without profound knowledge (Dornier *et al*., 1995). Artificial Neural Networks (ANN) (Single Neural Networks (SNN) and BANN) have been applied repeatedly to model nonlinear systems by studying the interactions between the molecules, water, and membranes (Ammi *et al*., 2021). However, for the FO process, there are few numbers of studies that use this type of modeling to study the FO system compared with other separation membrane processes such as nanofiltration and reverse osmosis.
The robustness and performance of neural networks have been made the subject of numerous types of research which have used the combination of several neural networks (Bootstrap Aggregated Neural Networks) to develop models in order to ensure the efficacy and reliability of neural networks generated by the limit of the training phase of the single neural network (Ammi et al., 2021; Sharma et al., 2009; Zhang et al., 2006; Mohammad et al., 2020).

Many works used the ANN (Single) to predict the rejection of OM by separation membranes, especially NF and RO membranes (Ammi et al., 2015; Yangali-Quintanilla et al., 2009), nevertheless, there are few studies that used the ANN to predict the rejection of OM by the FO membranes, (Pardeshi et al., 2016) used the ANN to determine the optimum conditions for the FO groundwater desalination, their work gave an honorable result and the ANN used can predict the optimum conditions for the FO system study. Jawad et al., 2020 have published a study about the modeling of FO process using an ANN to predict the permeate flux, they studied the effect of nine inputs (membrane (type, and orientation), feed solution (concentration, solution velocity, and solutions temperature) draw solution (concentration, molecule weight, and velocity) on the permeate flux with different parameters of the ANN used (number of neurons, number of the hidden layers), their results obtained were very satisfying and demonstrated its ability to predict the relationships between inputs and outputs in a way better than another simple learning machine such as Multiple Linear Regressions (MLR). Seong-Nam Nam et al., 2023 have modeled the permeate flux and the rejection of the Sulfamethoxazole by the FO membranes. In addition, Ibrar et al. 2023 have been used a learning machine to predict the permeate flux on the forward osmosis process. On the other hand, no study used the Bootstrap aggregated neural network to predict the FO process compared with the other separation membranes process such as nanofiltration and reverse osmosis processes which are developed with few studies like the studies of Ammi et al., 2018 ; 2021, where they used the QSAR-BANN to predict the membranes performance by treating the removal of pharmaceutical activate compounds, also Khaouane et al., 2017, have been studied the rejection of the organic compounds by nanofiltration and reverse osmosis using the BANN.

For the best of our knowledge, our work is the first one using the BANN to predict the rejection of OM by the FO membranes. In summary, many models will be generated
with to predict the rejection of OM by FO membranes, SNN will be developed firstly, with the study of the effect of inputs, training algorithms, activation functions, subdivision of the original database, this is the first section of this work, the second section contains the creation of new databases based on the original one with the resampling of its training set to which we will add the testing and validation sets. The QSPR-BANN (stacked of n networks) will be compared between them to get the best performance one, the third section consists to compare the individual neural networks (INNi) that constituted the best performance BANN model, and the last one is the comparison, in the first hand, between all the models developed in this work, the SNN, the best performance BANN, and the best performance INN, and in the second one, with the other works that used the BANN for the modeling of their works, especially which were about the membrane separation processes. At the end of this work, a conclusion summarizing the results of this work will be investigated.

2. Materials and Methods

2.1. Concepts of Artificial Neural Networks (ANN)

The neural network "NN" is a digital technique to bind inputs and outputs data of an external system (process) via a nonlinear regression model established for the best neural network architecture (Mohammad et al., 2020). This technique has the aptitude and viability to evaluate the relationships between the inputs and outputs similar to the biological neural networks (Ammi et al. 2021). Three stages named input, hidden, and output makes the simple architecture of the NN, the inputs signal received from external sources (bias, \( b \)) are multiplied by weights (\( W \)), and if the results of multiplying (\( y \)) beat the threshold, the signal will be released and sent to the output depending on the NN activation function. In this respect, three steps (training, testing, and validating) are applied to attain the chosen target through NN (Mohammad et al., 2020; García-Alba et al. 2019). Multi-layer perceptron is a feed-forward backpropagation neural network (FFNN) described by a particular configuration. These neurons are ordered in successive layers and the information runs in one direction, from the input to the output layer, and the neurons of the same layer are not interconnected (Ammi et al., 2021; Barello et al., 2014; Darwish et al. 2007; Madaeni et al. 2015; Si-Moussa et al. 2008; Maouz et al. 2019).

2.2. Bootstrap aggregated neural networks (BANN)
The SNN models are limited by their absence of generalization when applied to unseen data, a good performance has been given by the trained algorithm on the training data and poor performance with data that are not used in this training process. In recent years, many studies have developed new techniques to improve neural network generalization capability such as regularization (Bishop 1991), early stopping (Bishop 1995), Bayesian learning (Mackay, 1992), and combining multiple networks (Sridhar et al., 1996; Wolpert., 1992; Zhang et al., 1997) (Zhang et al., 2008). Among these techniques, the last one (combining multiple networks) is very promising to improve models’ prediction on unseen data, this technique consists to develop several neural network models with the aim to model the same relationship and combine them together to progress robustness and performance of the model. The new databases of the bootstrap aggregated neural networks model have been created by sampling the training databases using a function Matlab (bootstrap re-sampling) with replacement to forms 10, 15, 20, 25, and 30 (Ammi et al., 2020; Khouane et al., 2017; Zhang et al., 2008). Figure 1 shows a BANN, where numerous neural network models are assembled to model the relationship between inputs and outputs, and they are aggregated afterward. The individual networks are learned through using different training data and from different initial weights. The output of the BANN is a weighted combination of the individual neural outputs (Ammi et al., 2021).

**Figure 1.** Bootstrap aggregated neural network (Ammi et al., 2021).

### 2.3. Modeling procedure
This work contains four (04) sections to design and optimize the architecture of the NN as demonstrated in Figure 2. Firstly, the collection, pre-treatment, and analysis of the database. Secondly, the generation a (SNN) model using the original data base with different subdivisions, and different algorithms and functions. The original database is used to create new databases using the re-sampling of its training sets. The new training test assembles with testing and validation sets of the original data to show new data bases. Thirdly, many Individual Neural Networks (INN_i) have been created with the new databases to describe the BANN model. Finally, the comparison and the analysis between the SNN model and BANN model in the first place, and with other works in the second one.

**Figure 2.** Procedure of design and optimization of the SNN, INN_i, and BANN

2.3.1. Database; Collection, pre-treatment, and analysis

We have collected a database from the available literature (Alturki et al., 2013; Cui et al., 2016; Heo et al., 2020; Im et al., 2021; Jamil et al., 2015; Kim et al., 2017; Kong et al., 2014; Li et al., 2020; Lee et al., 2019; Linares et al., 2011; Madsen et al. 2015; Rostgar et al., 2020; Xie et al., 2014; 2013; Zhang et al., 2019), intending to group all the characteristics of the studied system, the size of this database was 193 points of 53 OM. The software "Get data Graph" was used to extract the values of the rejection from the diagram that presented the results. The selection of the input and output
variables was based on interactions between the organic molecule’s properties, membrane characteristics, and filtration operating conditions for the rejection of the OM by FO membranes. The inputs considered in this work are molecule descriptors (which can be access on the supplementary data, Table S1) (the effective diameter of an organic molecule in water \(d_c\), molecular length "Length", molecular width "Width", molecular depth "Depth", Dipole moment, The logarithm of the octanol-water partition coefficient "Log P", membrane characteristics (Surface membrane charge as" zeta potential", and the Hydrophobicity "as Contact angle"), and operating conditions (pH, Crossflow Velocity, and the water flux). The values of the effective diameter of an organic molecule in water \(d_c\), molecular width "Width", and molecular depth "Depth" are calculated with the equations (1), (2), and (3) respectively (Ammi et al., 2018; Dolar et al., 2012; Santos et al., 2006).

\[
d_c = 0.065 \times M_w^{0.0438}
\]

\[
Width = \frac{1}{2} \sqrt{S_{min}}
\]

\[
Depth = \frac{1}{2} \sqrt{S_{max}}
\]

Where:

\(M_w\) = Molecule Weight.
\(S_{min}\), \(S_{max}\) represent the minimum and maximum surfaces area.

The software Hypercham has been used to compute the Log p, Dipole moment, surfaces area and Chembio to calculate the molecular length.

The Statistical analysis is preliminary (standard deviations (STD), minimum, maximum, and mean) is shown in Table 1.

**Table 1.** Statistical analysis of the Inputs and Outputs for all databases.

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Max</th>
<th>STD</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>(d_c) (g/mol)</td>
<td>0.0914</td>
<td>1.1000</td>
<td>0.1291</td>
<td>0.7917</td>
</tr>
<tr>
<td>Log P</td>
<td>0.1612</td>
<td>6.0000</td>
<td>1.2105</td>
<td>2.4347</td>
</tr>
<tr>
<td>Dipole moment (Debye)</td>
<td>-2.2200</td>
<td>9.3000</td>
<td>2.9023</td>
<td>2.3499</td>
</tr>
<tr>
<td>Length (nm)</td>
<td>0.1468</td>
<td>0.2000</td>
<td>0.0119</td>
<td>0.1579</td>
</tr>
<tr>
<td>Width (nm)</td>
<td>0.5000</td>
<td>1.3000</td>
<td>0.1116</td>
<td>0.9969</td>
</tr>
<tr>
<td>Depth (nm)</td>
<td>0.3795</td>
<td>1.4000</td>
<td>0.1126</td>
<td>1.0509</td>
</tr>
</tbody>
</table>
For this database, we used a matrix correlation to edit the interactions between the variables (organic molecule properties, membrane characteristics, and operating conditions) and to reduce its size which can be accessed as Supplementary Data (Table S2).

2.3.2. Development of the QSPR-SNN model

From the original database, several networks have developed with different subdivisions of the database to find the most performance model of SNN, four (04) subdivisions were used; 80% for the training phase, 10% for the testing phase, and 10% for the validation phase for the first subdivision, the second one was 70% for training, 15% for each other, the third one divided the original data for 60% for the training phase, and 20% for each other, the last one was 90 % for the training phase , 5% for testing, and 5% for validation phase.

For each subdivision, a NN was developed, at first, the training algorithm used was BFGS (Broyden–Fletcher–Goldfarb–Shanno) (Ammi et al., 2021; Khaouane et al., 2017) quasi-Newton training algorithm and the activations functions in the hidden layer were variants, for this purpose, the Hyperbolic tangent (tanh), the Logarithmic sigmoid (logistic), the sin, and the exponential were used as activation functions for the hidden layer, the Pure linear (purelin) transfer function was fixed and used in the output layer. The number of neurons in the hidden layer was between 3 and 25 neurons for each neural network model.

Software STATISTICA was used for the QSPR-SNN and QSPR-INNi modeling of the rejection of the OM by the "FO" membranes.

2.3.3. Development of the QSPR-INNi and BANN (Stacked of N networks) models

The development of each QSPR-INNi was built with a new database, this database was obtained by resampled the training data of the subdivision that gives the most performance of the QSPR-SNN, using the bootstrap re-sampling with replacement to
forms 10, 15, 20, 25, and 30. For each one we added the testing and validation sets in aim to obtain a new database.

For each new database, a QSPR-INN$_i$ was modeled using the same steps that used for the NN model development (same training algorithm, same activation functions, and same various of the neurons number). For each model, several executions were established to extract the final best model, and each execution has been run with a high number of repetitions, sometimes more than 1000 repetitions by execution. The median of the outputs of the QSPR-INN$_i$ gives the stacked neural network QSPR-BANN, the output of the QSPR-BANN was obtained by the following equation (Ammi et al., 2021; 2018).

$$ y = \frac{\sum_{i=1}^{n} y_i}{n} $$

(4)

Where: $y$ represents the output BANN, $y_i$ is the outputs of the individual neural networks, and $n$ is the number of the INN. The software MATLAB is used for resampling and STATISTICA for the creation of the QSPR-INN$_i$.

3. Results and discussion

3.1. Creation of the QSPR-SNN Model

The original database was divided into three (03) subsets, the main part where the connection weights of the neurons are used to obtain the understanding of the neurons, this is the training data, the generalization ability is proved with the validation data, and the performance of the neural network is established by the testing data.

In order to estimate the influence of the database on the performance of the QSPR-SNN generated, two (02) parameters have been investigated, the correlation coefficient "R", and the Root Mean Square Error "RMSE", which is given by the next equation (Ammi et al., 2021; 2018).

$$ RMSE = \sqrt{\frac{\sum_{i=1}^{n} (Y_{i,exp} - Y_{i,cal})^2}{n}} $$

(5)

$Y_{i,exp}$, $Y_{i,cal}$ represent the experimental and the calculated values respectively; $n$ is the data number.
Table 2 shows the results of the effect of subdivision of the database on the rejection of the organic molecules by forward osmosis "FO" membrane, it is clear that the last subdivision with a percentage of 90% for the training phase, 05% for testing phase, and 05% for validation phase, gives higher correlation coefficient more than 0.9400 for the testing phase, and with less value of the Root Mean Square Error which equal to 1.2826 for the testing phase. The other subdivisions give acceptable values of the correlation coefficients and RMSE which mean the excellent choice of the inputs of our system studied. For the coming of this work, the subdivision that gives the best performance will be used for the modeling.

Table 2. Effect of the subdivision of database.

<table>
<thead>
<tr>
<th>Phases</th>
<th>Percentage (%)</th>
<th>Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>RMSE</td>
</tr>
<tr>
<td>Division 1</td>
<td>Processing phase</td>
<td>60%</td>
</tr>
<tr>
<td></td>
<td>Processing phase</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td>Processing phase</td>
<td>20%</td>
</tr>
<tr>
<td>Division 2</td>
<td>Processing phase</td>
<td>70%</td>
</tr>
<tr>
<td></td>
<td>Processing phase</td>
<td>15%</td>
</tr>
<tr>
<td></td>
<td>Processing phase</td>
<td>15%</td>
</tr>
<tr>
<td>Division 3</td>
<td>Processing phase</td>
<td>80%</td>
</tr>
<tr>
<td></td>
<td>Processing phase</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>Processing phase</td>
<td>10%</td>
</tr>
<tr>
<td>Division 4</td>
<td>Processing phase</td>
<td>90%</td>
</tr>
<tr>
<td></td>
<td>Processing phase</td>
<td>05%</td>
</tr>
<tr>
<td></td>
<td>Processing phase</td>
<td>05%</td>
</tr>
</tbody>
</table>

The QSPR-SNN model developed is characterized by BFGS quasi-Newton as the training algorithm, the exponential as the activation function in the hidden layer, the Identity (purelin) is the activation function in the output layer, the number of the neurons in the hidden layer is 25, and the number of neurons in the input layers and output layer is 11 and 1 respectively.

3.2. Creation of the QSPR-INNi, QSPR-BANN Models

3.2.1. Performance and comparison between QSPR-BANN
The QSPR-INN models were developed, as cited previously with the subdivision which has given the best performance of the QSPR-SNN, for this, several new databases have been created using the bootstrap resampling of the training data (90% of the original database), with replacement to form 10, 15, 20, 25, and 30, to which we have been added the testing and the validation data to obtain the new databases. The QSPR-BANN is given by the assembly of the QSPR-INN models created with the new databases. Table 3 shows the values of the coefficient correlation and the Root Mean Square Error of the testing phase and validation phases for each QSPR-BANN obtained.

For the testing phase (edit the performance of the QSPR-BANN), all the QSPR-BANN have excellent performance with values of "R" more than 0.9500, and small values of the RMSE, the QSPR-BANN (stacking 20 nets) is the model that gives the best performance with correlation coefficient very near to the ideal (R=1) and equal to 0.9909, the Root Mean Square Error equal to 0.5674%. For the validation phase, all the models created have been high values of the correlation coefficient, and small values of the RMSE which demonstrate the generalization ability for predicting the organic molecules rejection by the forward osmosis "FO" membranes of all the QSPR-BANN (stacking of n nets) generated.

Table 3. Values of "R" and "RMSE" of testing phase and validation phase for each QSPR-BANN

<table>
<thead>
<tr>
<th>QSPR-BANN</th>
<th>Errors</th>
<th>Testing Phase</th>
<th>Validation phase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R</td>
<td>RMSE</td>
</tr>
<tr>
<td>BANN (stacking 10 nets)</td>
<td>0.9618</td>
<td>0.9465</td>
<td>0.9741</td>
</tr>
<tr>
<td>BANN (stacking 15 nets)</td>
<td>0.9850</td>
<td>0.9727</td>
<td>0.9778</td>
</tr>
<tr>
<td>BANN (stacking 20 nets)</td>
<td>0.9909</td>
<td>0.5764</td>
<td>0.9646</td>
</tr>
<tr>
<td>BANN (stacking 25 nets)</td>
<td>0.9823</td>
<td>0.8201</td>
<td>0.9782</td>
</tr>
<tr>
<td>BANN (stacking 30 nets)</td>
<td>0.9839</td>
<td>0.7998</td>
<td>0.9774</td>
</tr>
</tbody>
</table>

In addition to the previous discussion, other values of errors are used to evaluate the performance of the QSPR-BANN (stacking 20 nets). For that, a comparison with the other QSPR-BANN developed using the mean absolute error (MAE), the root mean squared error (RMSE), the model predictive error (MPE), and the standard error of prediction (SEP) of the QSPR-BANN obtained on the testing phase, and validation phase. The performance of the model is based on the values of these errors, where for the values
less than 10% the performance is excellent, between 10% and 20% the performance is good, fair performance for values between 20% and 30%, and low performance if the values are more than 30% (Amiri et al., 2020; Despotovic et al., 2015; Fissa et al., 2019). These errors are obtained with the following equations (Ammi et al., 2021; 2020; Zhang et al., 2019).

\[
\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |(y_{i,\text{exp}} - y_{i,\text{cal}})| \quad (06)
\]

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (y_{i,\text{cal}} - y_{i,\text{exp}})^2}{N}} \quad (07)
\]

\[
\text{MPE(\%)} = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{y_{i,\text{exp}} - y_{i,\text{cal}}}{y_{i,\text{exp}}} \right| \quad (08)
\]

\[
\text{SEP(\%)} = \frac{\text{RMSE}}{Y_e} \times 100 \quad (09)
\]

\(Y_{i,\text{exp}}\), \(Y_{i,\text{cal}}\) represent the experimental and the calculated (predicted) values respectively, \(n\) is the number of data, and \(Y_e\) is the mean value of the experimental data.

As shown in Figure 3. For the testing phase, it is clear that the QSPR-BANN (stacking 20 nets) has a small value of error than the error QSPR-BANN with MAE equal to 0.5001%, RMSE equivalent to 0.5764%, MPE equal to 0.5179%, and SEP value is 0.5986%. These results give a higher performance of the QSPR-BANN (stacking 20 nets) generated.

![Figure 3. Parameters of performance of the QSPR-BANN for the testing phase.](image)

For the validation phase, as mentioned in Figure 4, the QSPR-BANN (stacking 20 nets) has a value of the errors near the smaller value (QSPR-BANN (stacking 15 nets)) and less than the other QSPR-BANN with values equal to 3.6997%, 4.8274%, 4.3967%, and...
5.2276% for the MAE, RMSE, MPE, and SEP respectively, which means that the generalization ability of the QSPR-BANN (stacking 20 nets) and (QSPR-BANN (stacking 15 nets)) is the best in comparison with other QSPR-BANN developed.

**Figure 4.** Parameters of performance of the QSPR-BANN for the validation phase.

### 3.2.2 Performance and comparison between QSPR-INN

The previous section 3.2.1 has demonstrated that the QSPR-BANN (stacking of 20 nets) gives the most performance model among the other QSPR-BANN.

The structures of the INN developed change from one model to another. It is clear that the log sigmoid present the activation function the most using as the transfer function with eight (08) individual neural networks, seven (07) individuals were used the tangent hyperbolic (tanh) as the activation function in the hidden layer, and the exponential function which gives the performance of the SNN model as cited previously has been used by five (05) individual neural networks, and while neural networks used the sin as the activation function in the hidden layer. It can be summarized that the two activation functions (tanh and logistic) are in the majority in comparison with the two other functions (Exponential and sin), these results are in accordance with the other works on the same targets.

The INN (5, 6, 8, 10, 14, 15, 17, and 19) have a number of the neurons in the hidden layer (determined by empirical rules) identical to or less than the number of the neurons in the input layer, the other individual neural networks have a number of neurons more than the number of the input (11) similar as the QSPR-SNN model. In
addition to the previous discussion, a rule of Dames 2005 gives a relationship between the number of the inputs and the number of neurons in the hidden layer, it can be equal to the number of inputs (11 neurons in our case); equal to 75% of the number of the inputs (8 neurons in our work); and equal to the square root of the input and output (11 neurons in our study), the INN with a number of neurons less than 11 are harmony with this rule. Also, to guarantee that the variables of the neural models do not surpass the size of the database we test the next rule cited by ((the inputs * the number of the hidden layer) + (the number of the hidden layer* the outputs)) ≤ the dimension of the database, for this, the twenty individuals neural network give respectively (192, 218, 156, 251, 36, 66, 192, 132, 165, 132, 253, 218, 240, 108, 36, 180, 36, 144, 84, 192) where the size of the database is more than 2123 (Fissa et al., 2019). Table S3 represents the values of the previous parameters of evaluation of the performance (MAE, RMSE, MPE, and SEP) in addition to the correlation coefficient of the all QSPR-INN; for the training phase, testing phase, validation phase, and total phase.

According to this Table, it is obvious that each INN has different values of previous parameters (R, MAE, RMSE, MPE, and SEP), which means the not dependable of the individual neural networks developed. Figure 5 shows the descending order of the coefficient correlation values, the INN_{14} has the highest value of "R" more than 0.9900 and equal to 0.9915, and five (05) INN give excellent performance with values of "R" more than 0.9600, eleven (11) INN have a good performance with values of R between 0.92 and 0.96, two (02) have acceptable performance with values of R from 0.91 to 0.92, and one (01) individual neural has poor performance with a value of "R" equal to 0.8642.
Figure 5. Coefficient correlation of twenty (20) INN; for the testing phase.

The precision of the INN can be extracted from the following figures, where, for the MAE as cited in Figure 6, it can be shown that only one individual neural network has a value of MAE more than 2%, eleven (11) between 2% and 1%, and the others have a good precision with values less than 1%. The best performance individual neural network INN$_{14}$ has the best precision with a value of MAE equal to 0.9749%.

![Figure 5](image)

Figure 6. MAE of twenty (20) INN; for the testing phase.

The ascending order present in Figure. 7 explains that the majority of the INN have good precision, except for two individuals which have values of the RMSE greater than 2%. The best performance individual INN$_{14}$ gives good precision with a value of RMSE equal to 1.1552%.

![Figure 6](image)

Figure 7. RMSE of twenty (20) INN; for the testing phase.
The values of MPE as mentioned in following Figure 8 explain that the minority of the individuals have higher values of MPE (more 1.5%) which demonstrates the good precision of the INN generated, the most performance INN\textsubscript{14} has been represented with a small value of MPE equal to 1.007%.

![Figure 8. MPE of twenty (20) INN\textsubscript{i} for the testing phase.](image)

The twenty (20) neurals developed have different values of SPE as mentioned in Figure 9, where, three (03) INN\textsubscript{i} have values more than 2%, eleven (11) between 1% and 2%, and six (06) INN\textsubscript{i} with SPE values less than 01%. The best performance individual neural network has a value of SPE equal to 1.1997%.

![Figure 9. SPE of twenty (20) INN\textsubscript{i} for the testing phase.](image)

3.3. Performance, comparison, and analysis of the NN models

3.3.1. Comparison between BANN, INN\textsubscript{i}, and SNN
According to the previous discussion, many NN models were developed (QSPR-SNN, QSPR-BANN (Stacking of 20 networks), and QSPR-INN) with the aim of modeling the rejection of OM by FO membranes.

The previous sections, 3.2.1 and 3.2.2 demonstrated that the QSPR-BANN (Stacking of 20 networks) and QSPR-INN are the most performance models between the QSPR-BANN (Stacking of n networks) and the QSPR-INN_i respectively. In this next section, a comparison between these two models (QSPR-BANN (Stacking of 20 networks) and QSPR-INN) and QSAR-SNN is carried out.

The scheme and the factors of the linear regression have been, directly, given by the Matlab function "Postreg". Figure 10 depicts a comparison between experimental and calculated rejections for each model with agreement vectors approaching the ideal [i.e. α= 1 (slope), β= 0 (intercept), R= 1 (correlation coefficient)] in the adjustment of the profiles of the neural networks, for SNN ( [α, β, R] = [0.6200, 34, 0.89173] for validation phase and [α, β, R] = [0.7700, 22.00, 0.94008] for test phase, for QSPR-BANN (Stacking of 20 networks) ([α, β, R] = [1.3, -31, 0.96462] for validation phase and [α, β, R] = [0.89, 10, 0.99095] for test phase), and for QSPR-INN ([α, β, R] = [1, -3.5, 0.91837] for validation phase and [α, β, R] = [1.2, -21, 0.99154] for test phase). The slope α is equal to 1 for the validation phase in QSPR-INN and it is close to 1 for the validation phase in the QSPR -SNN and QSPR-BANN (Stacking of 20 networks) models, it is very close to 1 for the testing phase in both neural networks models. The intercept β is very far from 0 for the validation phase and testing phase in the QSPR -SNN, QSPR-BANN (Stacking of 20 networks), and QSPR-INN models except that the value of the intercept β for the validation phase in the QSPR-INN which is close to 0.

Correlation coefficients have generally reflected the excellence of models where their values between (0.90 ≤ R ≤ 1.00), for these neural networks generated (QSAR-SNN, QSPR-BANN (Stacking of 20 networks), and QSAR-INN), the values of the correlation coefficients are very close to the ideal value (R=1) and lead us to show the good robustness of the established neural models and the possibility of predicting the different parameters that characterize the rejection of organic molecules during forward osmosis process.
Figure 10. Plot of linear regression: (a) testing phase for SNN, (b) testing phase for BANN (stacking of 20 networks), (c) testing phase for INN\textsubscript{14}, (d) validation phase for SNN, (e) validation phase for BANN (stacking of 20 networks), and (f) validation phase for INN\textsubscript{14}

Figure 11 shows the comparison between the three models, it can be seen that for all the errors used in the comparison, the BANN (stacking of 20 networks) is the model which has the best precision than the others (SNN, and INN\textsubscript{14}) for testing phase with values of MAE less than 1% for all models, RMSE equal to 0.5764% for BANN and more than 1% for SNN and INN\textsubscript{14}, the MPE has near values for SNN and INN\textsubscript{14}, the BANN (stacking of 20 networks) has fewer values than them with values equal to 0.5179%. The SEP as shown in the same Fig. 11 takes the same rule with values equal to 0.5986% for the BANN (stacking of 20 networks), between 1% and 1.2% for INN\textsubscript{14}, and more than 1.2% for SNN.
The comparison between the three (03) models for the validation phase has been demonstrated in the next Figure 12. It can be seen that for the validation phase, the errors have near values, and the SNN gives the best precision with less value than the other models (BANN (stacking of 20 networks), and INN14). The RMSE and SPE demonstrate that the INN14 has the best precision in comparison with the other models with values of more than 4% for all the models. MPE values are very close between them for the SNN and INN14, for the BANN (stacking of 20 networks) the MPE value is more than 4%.

From the previous discussion, this comparison indicates the robustness, reliability, and efficiency nature of neural network models (QSPR-BANN (Stacking of 20 networks), QSPR-SNN, and QSPR-INN); it has demonstrated the superiority of the QSPR-
BANN (Stacking of 20 networks) model. For the performance of the models, the QSPR-BANN (Stacking of 20 networks) model has the excellent precision and can predict the organic molecules rejection by the forward osmosis membranes, nevertheless, this model has less precision than the other models (QSPR-SNN and QSPR-INN) with the aim of the generalization ability of the NN models.

3.4. Comparison with other works

Many studies have been used the artificial neural network to develop the separation membranes process, for the best of our knowledge, our work is the first one consists to apply the bootstrap aggregated neural networks (BANN) to predict the rejection of the OM by the FO membranes, for this purpose, a comparison with other works applied these models (SNN and BANN) in other separation membranes process such as nanofiltration and reverse osmosis membrane is supported. Table 4 shows the aim of each study. This Table demonstrates that the Bootstrap aggregated neural network (BANN) gives results with higher performance and precision than the single neural networks in the first hand and proves the robustness and the accuracy of our model than the other works in the second hand.

**Table 4.** Overview of various works on models (BANN).

<table>
<thead>
<tr>
<th>Method</th>
<th>Membrane type</th>
<th>Database</th>
<th>Number of MO</th>
<th>R</th>
<th>RMSE (%)</th>
<th>MAE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Khaoune et al., 2017</td>
<td>BANN</td>
<td>NF/RO</td>
<td>436</td>
<td>0.8156</td>
<td>7.7058</td>
<td>4.7350</td>
</tr>
<tr>
<td>Ammi et al., 2018</td>
<td>BANN</td>
<td>NF/RO</td>
<td>278</td>
<td>0.9836</td>
<td>2.5882</td>
<td>0.9878</td>
</tr>
<tr>
<td>Ammi et al., 2021</td>
<td>BANN</td>
<td>NF/RO</td>
<td>599</td>
<td>0.9603</td>
<td>1.0105</td>
<td>0.7916</td>
</tr>
<tr>
<td>Our work 2023</td>
<td>BANN</td>
<td>FO</td>
<td>193</td>
<td>0.9909</td>
<td>0.5764</td>
<td>0.5001</td>
</tr>
</tbody>
</table>

4. Conclusion

This work explores the application of the BANN in forward osmosis membranes process separation to predict the organic molecules rejection. However, many models have been developed, single Neural Network has been created with eleventh input including the properties of organic molecules, membrane characteristics, and
operating conditions. The study of the effect of training algorithms, transfer functions, hidden neurons, and subdivisions of the database present the main of the first section.

The bootstrap aggregated neural networks were found by many resampling of the original database. Moreover, a comparison between the BANN (stacked of n networks) obtained describes that the BANN (stacked of 20 networks) is the most performance one than the others, the twenty (20) individual neural networks developed have different architectures which demonstrated the non-consistent of these models, among all of these models, the INN14 is the most performance one. The INN14, the BANN (stacked of 20 networks), and the SNN were compared between them using the Coefficient Correlation "R", the Root Mean Squared Error "RMSE" with (R= 0.9401, and RMSE= 1.2826%) for the SNN, INN14 gives R equal to 0.9915 and "RMSE" with 1.1526%, and the BANN (stacked of 20 networks) offers R =0.9904, and RMSE= 0.5764% for the testing phase. This work also shows the precision, the robustness, reliability, and efficiency nature of each model generated using the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE), the Model Predictive Error (MPE), and the Standard Error of Prediction (SEP) with the outstanding superiority of the QSRR-BANN (Stacking of 20 networks) model for the unseen data. This work demonstrates the excellence of our BANN model in comparison with other works that applied the Bootstrap aggregated neural networks method for the predicting of the organic molecules rejection by other separation membranes process such as nanofiltration and reverse osmosis. The bootstrap aggregated neural network gives honorable results for the prediction of the rejection of the organic molecules by the forward osmosis membranes in accordance with the other results of the same objective by other membranes (NF/RO) and with the advantage of the BANN compared to the Single Neural Networks for the unseen data.

Acknowledgments

The authors gratefully acknowledge the Ministry of Higher Education of Algeria (PRFU Projects N° A16N01UN260120220004) and the group of Laboratory of Biomaterials and Transport Phenomena in university of Medea.

Compliance with ethical standards

Conflict of interest Authors declares that there is no conflict of interest to declare.
REFERENCES


