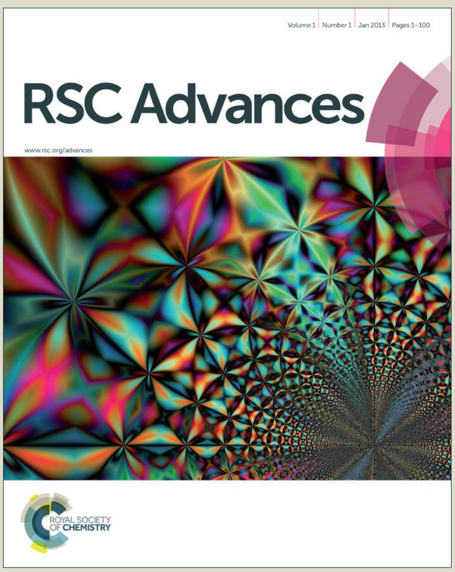
RSC Advances



Modeling and Optimizing Performance of PVC/PVB Ultrafiltration Membranes Using Supervised Learning Approaches

Lina Chi [a,b,c*], Jie Wang [b], Tianshu Chu [b], Yingjia Qian [a], Zhenjiang Yu [a], Deyi Wu [a], Zhenjia Zhang [a], Zheng Jiang [c], James O. Leckie [b]

[a] *School of Environmental Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200240, PRC*
[b] *The Center for Sustainable Development and Global Competitiveness (CSDGC), Stanford University, Stanford, CA 94305, USA*
[c] *Faculty of Engineering and the Environment, University of Southampton, Southampton, SO17 1BJ, UK*

* Correspondent author. Tel: +86 13816632156; E-mail address: lnchi@sjtu.edu.cn

Abstract

Mathematical models plays an important role in performance prediction and optimization of ultrafiltration (UF) membranes fabricated via dry/wet phase inversion in an efficient and economical manner. In this study, a systematic approach, namely, a supervised, learning-based experimental data analytics framework, is developed to model and optimize the flux and rejection rate of Poly (vinyl chloride) (PVC) and Polyvinyl butyral (PVB) blend UF membranes. Four supervised learning (SL) approaches, namely, the multiple additive regression tree (MART), the neural network (NN), the linear regression (LR), and the support vector machine (SVM), are employed in a rigorous fashion. The dependent variables representing membrane performance response with regard to independent variables representing fabrication conditions are systematically analyzed. By comparing the predicting indicators of the four SL methods, the NN model is found to be superior to the other SL models with training and testing R-squared values as high as 0.8897 and 0.6344, respectively, for the rejection rate, and 0.9175 and 0.8093, respectively, for the flux. The optimal combination of processing parameters and the most favorable flux and rejection rate for PVC/PVB ultrafiltration membranes are further predicted by NN model and verified by experiments. We hope the approach is able to shed light on how to systematically analyzing multi-objective optimization issues for fabrication conditions

to obtain the desired ultrafiltration membrane performance based on complex experiment data characteristics.

Key words: Poly (vinyl chloride) (PVC); Polyvinyl butyral (PVB); Supervised Learning （SL）; Neural network (NN); modeling; Membrane fabrication optimization

1. Introduction

Poly (vinyl chloride), or PVC, is commonly used to produce relatively inexpensive ultrafiltration (UF) membranes due to its relative low cost, robust mechanical strength, and other favorable physical and chemical properties, such as abrasive resistance, acid and alkali resistance, microbial corrosion resistance, and chemical performance stability[1]. Moreover, PVC membranes can usually maintain a longer membrane life and remain intact after repeated cleaning with a wide variety of chemical agents. However, the hydrophobic nature of PVC always leads to severe fouling, thereby impeding its applications[1,2]. Thus, a critical challenge is to improve the hydrophilicity of PVC membranes without interfering with their positive characteristics so that PVC-based membranes can comply with industry requirements for a wider range of applications.

In recent years, considerable research has been conducted in order to overcome this problem. Among all available methods, polymer blends often exhibit superior properties when compared with a standalone, individual component polymer; in addition, the polymer blend method also has the advantages of a simple procedure for preparation and easy control of physical properties for various compositional changes. There are several polymers that have been studied as functional polymer pairs of PVC, such as PMMA[1], PU[3], EVA[4], PEO[5], and PVB[6] among others. In most previous studies[7,8], PVB is found to be one of the ideal polymers to blend with PVC due to its well-predicted miscible properties, chemical similarity, and less unfavorable heat while mixing. In addition, owing to the –OH bond, the PVC/PVB blend demonstrates more hydrophilicity than the original PVC membrane [6,9] .

The selection of membrane material is essential for developing high-performance membranes. However, due to the complexities of the fabrication process, even more critical—especially when the membranes are made via a complex dry/wet phase inversion—is a consistent and robust data analysis procedure for effectively analyzing these membranes for better performance. Pure water flux (PWF) and rejection rate of Bull Serum Albumin (BSA) are the most important performances for UF membranes[10,11], depending not only upon the composition of the casting solution but also upon the technical conditions used in the fabrication process. Typical variables of importance for membrane development include the types and amounts of polymer, additive, and the pore-forming agents used in the casting solution, the kind and concentration of gelation medium, the evaporation time and temperature of the spread-casting solution, the length of gelation period, and the temperature of gelation bath[12] etc.. Some of the above mentioned variables have to be classified as categorical variables, such as the type of the polymer, the pore-forming reagent, or the gelation medium used, since they cannot be quantified. Remaining variables are quantitative ones, including the temperature of evaporation or gelation, the amount of pore-forming reagent added, and the duration of evaporation or gelation. Generally, these complex influential factors in the membrane fabrication process would greatly delay the development cycle and increase research and development (R&D) costs. Therefore, it is worthwhile to investigate efficient statistical and computational methods to optimize experiment design and to minimize the number of experiments.

Traditionally, statistically-based design of experiments (DOE) has been widely used as a proper approach to optimize membrane parameters in membrane fabrication processing[13-15]. However, DOE is based on the assumption that interactions between factors are not likely to be significant[16,17], which is usually not the case in the real world. When reducing the number of runs, a fractional factorial DOE becomes insufficient to evaluate the impact of some of the factors independently[16]. Moreover, it is also beyond the ability of DOE in dealing with categorical factors in experiments. As a result, DOE has limitations in modeling a membrane fabrication process and in optimizing the filtration performance of the membrane.

Recently, the supervised learning (SL) approach—a powerful method in analyzing complex, but data-rich problems—has found strong application in diverse engineering fields such as control, robotics, pattern recognition, forecasting, power systems, manufacturing, optimization, and signal processing, etc. [18-20]. Although the idea of solving engineering problems using SL has been around for decades, it has been introduced only recently into the field of material studies[21]. There are several publications discussing the application of SL to the modeling and optimization of membrane fabrication. S. S. Madaeni modeled and optimized PES- and PS-membrane fabrication using artificial neural networks[22], while Xi and Wang [23]reported that the Support Vector Machine (SVM) model could be an efficient approach for optimizing fabrication conditions of homemade VC-co-VAc-OH microfiltration membranes. Yet, there are still a couple of key issues that need to be investigated. A systematic framework for using SL approaches is required to discover the relationships between membrane performance and complicated fabrication conditions.

The purpose of this research is to develop such a framework. More specifically, we need first to evaluate experimental data quality, which is important in making valid assumptions and selecting proper models for analyzing complex data. Secondly, we need to develop an approach for efficiently employing reliable analysis models, including the decision tree approach, neural network method, linear regression, and support vector machine, for thoroughly analyzing all features and all responses of the membranes, as opposed to current approaches that analyze only a single response with regard to either one feature or all of the features. Finally, we need to select the most suitable SL approach to predict the optimal combination of features for membrane fabrication.

2. Experimental

2.1. Chemicals and materials

Unless otherwise specified, all reagents and chemicals used were of analytical grade. More specifically, PVC resin (Mw = 1.265×10^5 g/mol, and [η] = 240 mPa·s) was supplied by Shanghai Chlor-Alkali Chemical Co., Ltd. Mw = 1.265×10^5 g/mol, and [η] = 240 mPa·s. PVB (Mw = 3.026×10^4 g/mol, and [η] = 40 mPa·s) was purchased from Tianjin Bingfeng Organic Chemical Co., Ltd.. N,N-dimethylacetamide (DMAc) was purchased from Shanghai Lingfeng Chemical Reagent Co., Ltd. PEG 600, PVP K90, and Ca(NO3)2 were purchased from Aladdin Industrial Inc. BSA(Mw=67,000 g/mol) was supplied by Shanghai Huamei Biological Engineering Company.

2.2 Membrane fabrication

PVC/PVB composite membranes were prepared by the non-solvent induced phase inversion. The casting solutions, containing PVC, PVB, DMAc, and additives, were prepared in a 250 mL conical flask and heated to approximately 30-80 °C in a water bath while being stirred at 600 rpm using a digital stirring machine (Fluko, GE).

After the polymers had been dissolved completely and stirred for at least 24 h, the resulting solution was degassed for at least 30 min until no gas bubbles were visible. The solution was cast on a glass plate using an 8-inch wide doctor blade with a gap of 200 μm between the glass plate and blade. The temperature of the blade and the glass plate was controlled between 30-80 °C. After a predetermined evaporation period, ranging from 5 to 120 seconds, the film was immersed in a pure water or DMAc (with volume concentration ranging from 10–80%) gelation bath maintained at 20°C. The film was then removed from the glass plate and leached overnight in water in order to completely remove any traces of solvent. Table S1 listed the various combination of composition of casting solutions and corresponding processing parameters.

2.3 Membrane characterization

The pure water flux of the PVC/PVB blend ultrafiltration membranes was measured at a temperature of 25 °C and under an operating pressure of 0.1 MPa after pre-operating for 30 min. The flux of permeate was calculated according to Eq.(1):

$$J_w = V/(A \cdot t) \tag{1}$$

where J_w (L/(m^2 • hr)) is the pure water flux, V (L) is the volume of the collected permeate, and A (m^2)is the area of the membrane. In our study, the effective membrane area is 0.0342 m^2 and t (hr) is the separation time.

Membrane retention ability was tested using 100 mg/L BSA at a temperature of 20 °C and under an operating pressure of 0.1 MPa. The concentrations of both the feed water and the permeation water were determined using an ultraviolet spectrophotometer (TU-1810, Beijing Purkinje Genera, China) at a wavelength of 280 nm. The percentage of the observed rejection solutes BSA phosphate buffer for each permeate collected was calculated as the following Eq.(2):

$$R = (1 - C_p/C_f) \times 100\% \tag{2}$$

where C_p is the permeate concentration and C_f is the feed concentration.

3. Analyzing membrane performance by SL approaches

In both this section and in Section 4, we describe a systematic framework for modeling and optimizing performance of PVC/PVB ultrafiltration membranes using supervised learning approaches, consisting of the following: (1) methods for analyzing raw datasets and their dependencies, (2) a general procedure and algorithms of SL-based data processing, (3) detailed results analysis and comparisons among all

SL approaches, and (4) selection of the best learning approach for optimally predicting experimental performance for analyzing membrane performance.

3.1 Data structures and characteristics

To better understand the potential inherent structures among independent and dependent variables, in this section, we first describe data structures and characteristics of experimental data sets. As listed in Table S1, there are a total of 68 valid experimental measurements. For each measurement, we have initially identified and employed 9 processing parameters that are regarded as independent variables and 2 performance indicators that are regarded as dependent variables. Specifically, the processing parameters are PVC Wt%, DMAc Wt%, Additive Wt%, Additive type (PEG600, PVPk90, $Ca(NO_3)_2$), Casting solution temperature (°C), Evaporation time (sec), Blade temperature (°C), Gelation bath type (Water, DMAc), and Bath concentration (solute concentration in gelation bath) (mg/L). Note that the types of additives and the gelation bath are categorical variables. The performance indicators, including the rejection rate of BSA (%) and the flux ($L/(m^2 \cdot h)$), are numerical variables. Through our preconditioning analysis, we find that the Wt% of polymers and the Wt% of PVB have to be removed from the processing parameters because they are dependent on, and correlated with, the change of PVC Wt%, DMAc Wt%, and Additive Wt%. We introduce *k* as the ratio of PVC Wt%/ Polymer Wt%, giving us $0 < k < 1$. There exist following relationships:

$$\text{PVC Wt\%}/k=\text{Polymer Wt\%} \tag{3}$$

$$\text{DMAc wt\%}+\text{Polymer Wt\%}+\text{Additive Wt\%}=100\% \tag{4}$$

Before the data analysis process, we briefly verify the characteristics of the data by scattering the measurement points under different parameter-indicator pairs in Fig. 1. If the processing parameters are categorical, box-plots are used instead of scatter plots. Obviously, the rejection rate and the flux are negatively correlated. For numerical parameters, PVC Wt% and DMAc Wt% have the strongest correlations with flux and rejection rate, respectively, while evaporation time and blade temperature have cross-like scatterings, thus indicating very weak correlations. Both categorical parameters can provide considerable information for performance prediction. This is especially true for the additive type, where the significant differences of indicators are shown between different groups of additives. In general, useful information can be found in the data for performance prediction, but there are not enough measurements to estimate how the indicators are distributed with regard to processing parameters. In other words, our predicted indicators using SL tools will have a low bias but high variance, and we need to carefully balance the accuracy and stability of modeling.

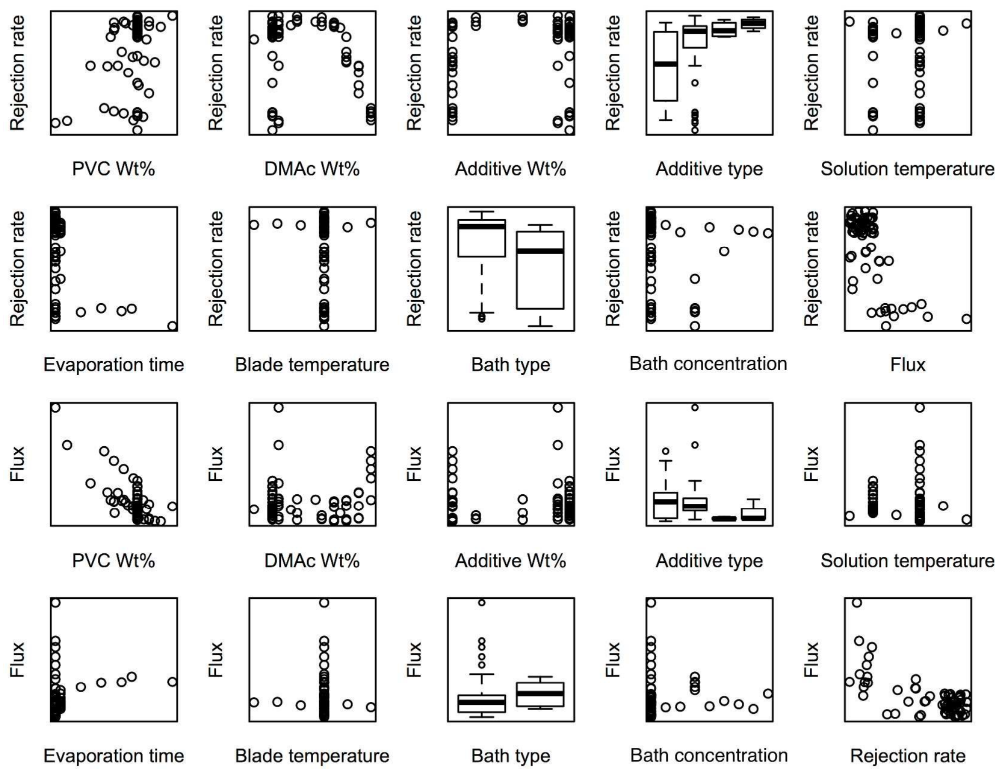


Fig. 1 Scatter plots over measurements.

3.2 Supervised learning and data analysis procedures

3.2.1 General description and criteria

Different SL algorithms, including linear regression (LR), a multiple additive regression tree (MART), a neural network (NN), and a support vector machine (SVM), were introduced and implemented to find the potential influence of processing parameters (predictors) on performance indicators (responses). The advantages, limitations and assumptions when utilizing each SL algorithm were described in Supporting information. To analyze the results, we train each SL algorithm over the whole data.

Furthermore, to estimate the accuracy of each SL algorithm, we apply the Monte Carlo method by repeating the learning processes 50 times on our measurement data. During each learning process, we first randomly split the data into a training set and a testing set, with the ratio 50/18. Next, we train each SL model based on the predictors of the training set with cross-validation and make predictions of responses over the training and testing sets using the trained learning model. Finally, we estimate the accuracy of each model by R-squared over the training and testing sets, computed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right)^2}{\sum_{i=1}^{m}\left(\bar{y} - y^{(i)}\right)^2} \qquad (5)$$

where m denotes the size of the data over which we perform predictions, $\hat{y}$ denotes the prediction of each response for each array of predictors, and $\bar{y}$ denotes the mean of true responses in the data. Usually, higher training and testing R-squared values imply lower bias and variance in the predictions, respectively. Fig. 2 shows the whole SL process. Once we select the best SL model with the highest prediction accuracy, we can train it again with all 68 data points for the further analysis.

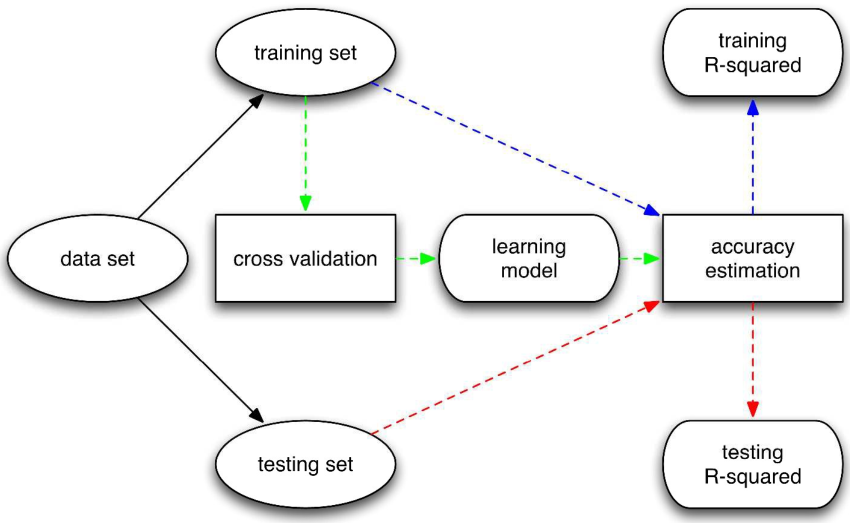


Fig. 2 Data analysis procedure for each SL model, where ovals and rounded rectangles denote the input and estimated variables, respectively

3.2.2 Implementation of supervised learning

Since our data size is small compared to the number of predictors, to avoid over-fitting of NN and SVM, only statistically significant predictors are used for training. Here we apply LR and MART (which are robust to irrelevant predictors) to analyze and extract significant predictors. Also, cross validation is implemented to determine appropriate controlling parameters of NN and SVM, for optimizing the learning performance.

3.2.2.1 Analysis of predictors' significance

According to LR analysis, the coefficients of PVC Wt% and evaporation time, those of DMAc Wt% and Additive Wt%, and those of additive and bath types are statistically significant at level 0, level 0.01, and level 0.05 for the rejection rate. However, there are only two statistically significant coefficients: one of PVC Wt% at level 0 and another of DMAc Wt% at level 0.1 for the flux. In other words, only a few processing parameters can provide significant information on the predictions; especially for the flux, PVC Wt% and DMAc Wt% are the two that carry the most amount of information. The low statistical significances are partially due to the small number of measurements. The linearity assumption on the relationship can be tested with R-squared values, which we will discuss later. In addition, the identical and independent distribution assumption on the noise can be tested by residual versus predicted response plots, which are shown in Fig. 3. Although the mean of residuals is indeed zero, the variance does not follow the null plot; this may be because our data is collected via a controlled parameter method.

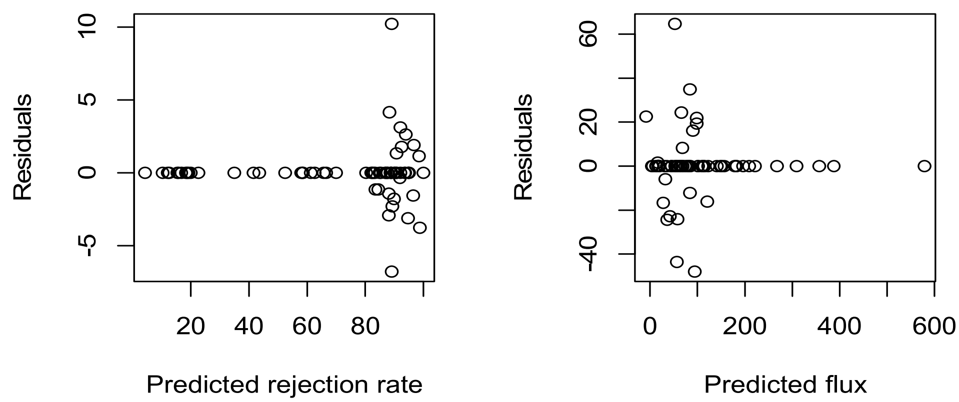


Fig. 3 Residuals versus Predicted values plots for rejection rate and flux

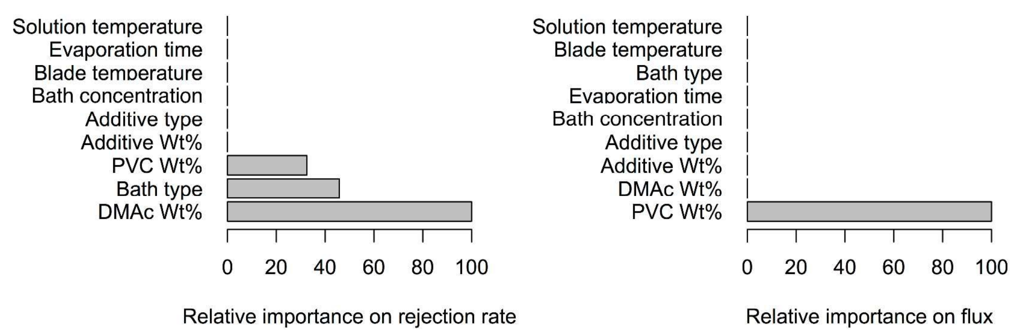


Fig. 4 Importance plots of predictors on each indicator

In case of MART analysis, the resulting importance rankings of each predictor for predictions are shown in Fig. 4. We can see that the number of significant predictors is even fewer than that in LR for each indicator. The importance order is

DMAc Wt% > Bath type > PVC Wt% for rejection rate, and only PVC Wt% determines the regression tree for flux.

In sumary, LR suggests that PVC Wt% and DMAc Wt% are the two most significant predictors. MART claims that the importance order of predictors is DMAc Wt% > Bath type > PVC Wt% for rejection rate, while only PVC Wt% determines the regression tree for flux. Based on the results of LR and MART, we remove the insignificant predictors (solution temperature) and then train NN and SVM with the appropriate controlling parameters determined by cross validation.

3.2.2.2 Selection of appropriate controlling parameters for NN and SVM

As shown in Fig.1, the responses in our data are correlated, so NN is more appropriate than any other SL model, which can only predict the rejection rate and the flux separately. To apply NN, we should first assume that the categorical predictors (additive type and bath type) are numerical. In addition, we remove the unimportant predicator (solution temperature) and normalize all input predictors to zero-mean and one-standard-deviation.

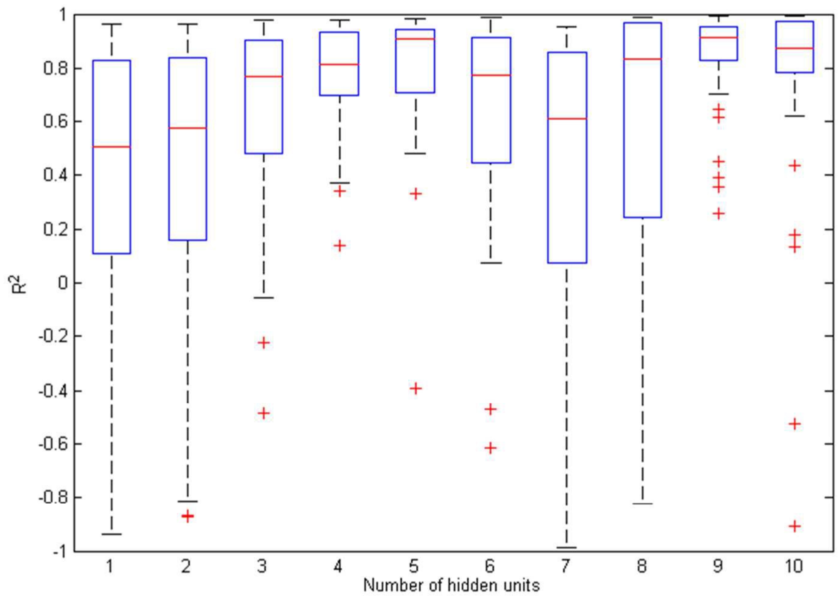


Fig. 5 Box-plots of testing R-squared values over 50 training processes with different hidden layer sizes

Furthermore, we select appropriate controlling parameters. Usually, one hidden layer is sufficient for a small training set. To select the optimal number of hidden units, we repeat the learning processes 50 times for each, and then select the one with a high mean and a low variance of testing R-squared values. During each process, we randomly split the data into a training set, a validation set, and a testing set, with the ratio 51/10/7, and then select the best number of epochs through cross-validation. The resulting box-plots are shown in Fig. 5. We can see the optimal number of hidden units is 9, with both the highest mean (0.8218) and the lowest variance of testing R-squared values.

As regard to SVM, since our data size is small, we select only the statistically significant 6 predictors in LR and MART to avoid overfitting. Furthermore, we choose the appropriate controlling parameters with five-fold cross-validation. The resulting support vectors are from all measurements except the 43rd or 18th measurements for the rejection rate or the flux, implying the risk of over-fitting.

4. Results and discussion

4.1 Performance of SL models and selections

The training and testing R-squared values of all SL models introduced above are listed in Table 1, where Rm and Rn denote the training and testing R-squared values, respectively, and y1 and y2 denote the rejection rate and the flux. We can see NN is the best SL model, with the highest Rm and Rn for both y1 and y2. The second best SL model is SVM, which performs considerably worse for y2 and Rn.

Table 1 Summary of performance of different SL models

	MART	**NN**	**LR**	**SVM**
Rm(y1)	0.2122	0.8897	0.6577	0.8065
Rm(y2)	0.0725	0.9175	0.6887	0.6583
Rn(y1)	0.0784	0.6344	0.3104	0.4344
Rn(y2)	-0.0329	0.8093	0.1800	0.6583

By combining the performance results in Table 1 and the properties of each SL model, we can reveal some interesting underlying characteristics of the data. We begin with the worst SL model, MART, which has very low R-squared values for all conditions. In other words, the piecewise constant approximation does not work on this data, partially due to the small number of controlled measurements. However, we find that both the bias and variance are lower for the rejection rate. Thus, compared to the flux, the rejection rate has relatively high order interactions with processing parameters. This argument can be verified with the performance of LR. Both training R-squared values are relatively high. Especially for the flux, this value is even higher than that of SVM. Furthermore, SVM has much higher training R-squared of the rejection rate, and testing R-squared of both rejection and flux than those of LR. Therefore, the relationship between the flux and the processing parameters is approximately linear, but the rejection rate may have more complex and higher order interactions between the processing parameters. In addition, the noise of the measurement data is relatively high. Finally, although the testing R-squared values of SVM are much higher than LR due to the noise reduction in the higher dimensional feature space, they are still much lower than those of NN. This verifies the overfitting of SVM on small data, even when the regularization cost is set as high as 2^5.

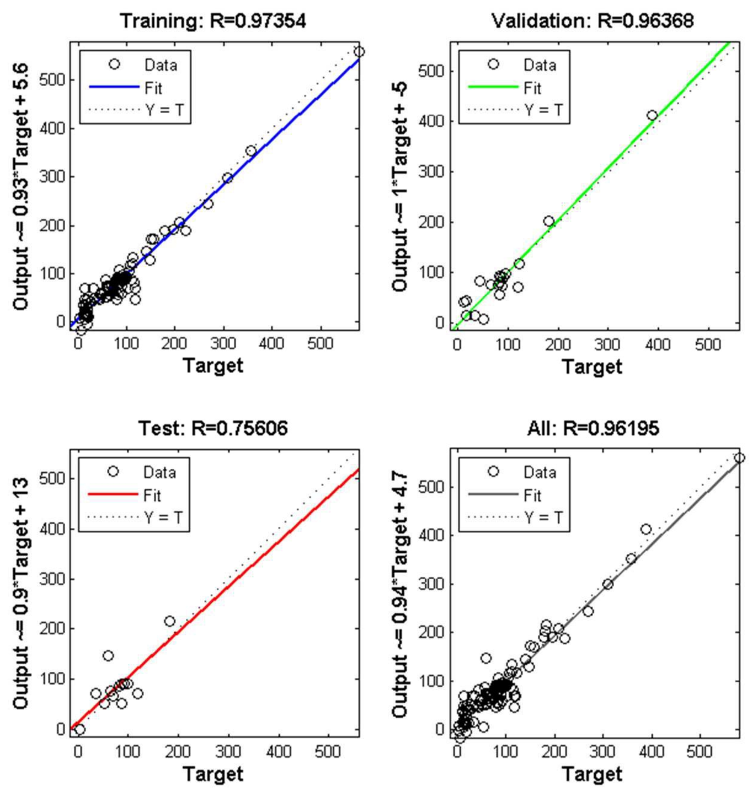


Fig. 6 Prediction versus response plots for training, validation, testing, and the whole data set; target and output denote the true response and the predicted response by NN, respectively

NN beats all other SL models in all aspects, and if the whole data is used for training, it has training R-squared values as high as 0.8992 and 0.9559 for the rejection rate and the flux. Thus, compared to the numerical approximation on categorical predictors, the correlation between the rejection rate and the flux is much more important in our predictions. To visualize the performance of NN, we plot the prediction versus the true response in Fig. 6. The performance is considered perfect if the point lies on the line with intersection 0 and slope 1. Furthermore, we plot the training data points and fitting curves of SVM and NN inside the predictor subspace of PVC Wt% and DMAc Wt% in Fig. 7 and Fig. 8 by fixing all other predictors as Additive Wt% = 0%, Additive type = None, Evaporation time = 5 sec, Blade temperature = 60 °C, Bath type = Water, and volume concentration of solute in gelation bath= 0 mg/L. We can see that the fitting curves of NN are smoother and fit the training data better. In summary, because our data set is very small and noisy, the complex relationship between the rejection rate and the processing parameters is hard to fit with a good trade-off between bias and variance. Fortunately, we have the helpful information that tells us that it is correlated with the flux, which has a much simpler linear relationship, so we can apply NN to fit these two indicators.

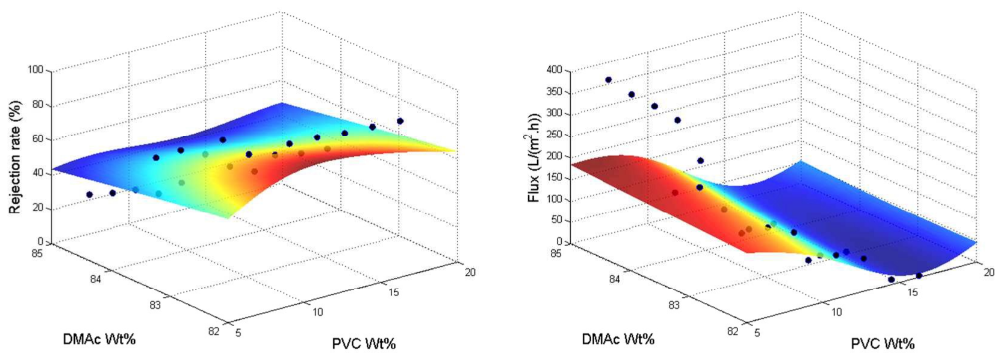


Fig. 7 Training data and fitting curves of rejection rate and flux in the subspace of PVC Wt% and DMAc Wt% using SVM

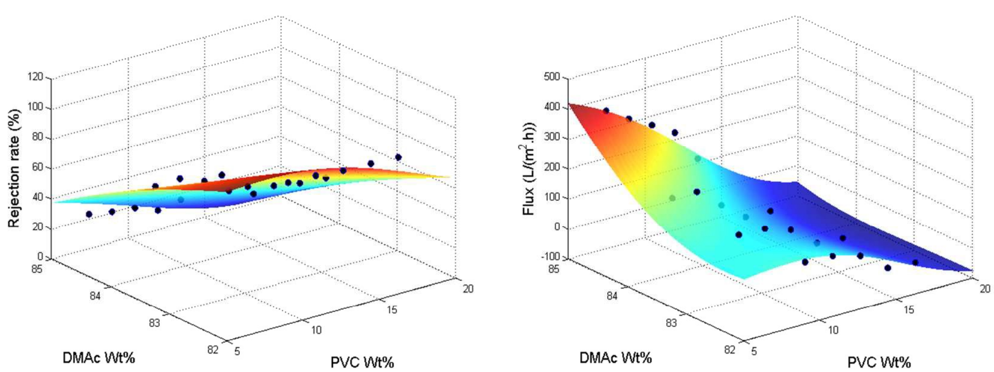


Fig. 8 Training data and fitting curves of rejection rate and flux in the subspace of PVC Wt% and DMAc Wt% using NN

4.2 Optimization with NN

In this section, we use NN model to find the optimal combinations of processing parameters to maximize the flux under the constraint that the rejection rate of BSA should be no less than 80%. The idea is very simple: we search over the predictor space to find certain combinations that achieve the maximum predicted flux under the constraint regarding the predicted rejection rate by NN. For example, when we fix Additive Wt% = 1%, Additive type = PEG600, Evaporation time = 35 sec, Blade temperature = 70 °C, Bath type = Water, and Bath concentration= 0 mg/L, the possible combinations of PVC Wt% and DMAc Wt% satisfying rejection rate >= 80%, flux >= 200 $L/(m^2 \cdot h)$ are scattered in Fig. 9. It is noticed that the combinations are almost impossible in reality in the case of DMAc Wt%<40% or DMAc Wt%>85%. Therefore, a question is raised here on how to perform an efficient and reliable search. As a matter of fact, regarding to the problem, there exist two main difficulties: (1) when searching over a high-dimensional predictor space, the computation cost is very high; and (2) the predictions have high variance since the size of the training data is small. To overcome these difficulties, we first narrow down the search space by utilizing additional knowledge about the experiments and constraints on predictors. There are several obvious constraints, such as if Additive type = None, then Additive Wt% = 0%; if Bath type = water, then Bath concentration= 0 mg/L. In addition, our focus is on estimating how the addition of PVB into PVC improves the performance of membranes, so we introduce k as the ratio of PVC Wt%/ Polymer Wt%, giving us $0 < k < 1$. Furthermore, we should keep the Polymer Wt% at no greater than 21%. Note that DMAc Wt% can be easily calculated using Eq.3 and Eq.4.

So we can use k instead of DMAc Wt%. On the other hand, although the prediction accuracy is not guaranteed over the whole predictor space, both training and testing R-squared are very high within the data set. This means that if the search points are not too far away from the measurement points, the corresponding predictions are reliable. In particular, we have the search space PVC Wt% = 7.5:0.5:18 (%), $k = (\lceil \mathrm{PVC\ Wt\%}/21 \rceil)$, 0.05:0.9 , and Additive Wt% = 1:1:5 (%) if Additive type is not None, Evaporation time = 5:15:110 (sec), Blade temperature = 30:10:80 (°C), and Bath concentration = 10:10:80 (mg/L).

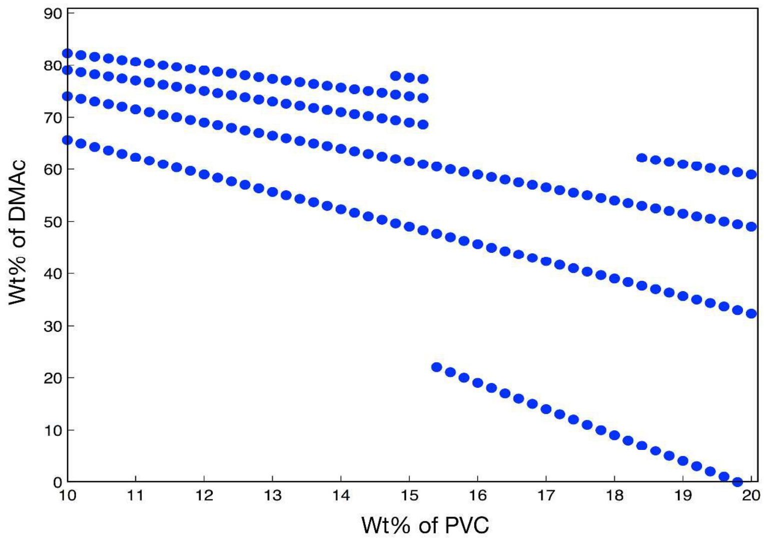


Fig. 9 Possible combinations of PVC Wt% and DMAc Wt% for specific constraints on indicators fixing all other processing parameters

Finally, we select the combination of processing parameters that have the maximum flux under the constraint 80% ⩽ rejection rate ⩽100%. We find with the water bath that the optimal combination of processing parameters is PVC Wt% = 7.5%, DMAc Wt% = 84%, Additive Wt% = 1%, k = 0.5 (PVB Wt%=7.5%), Additive type = PEG600, Evaporation time = 5 (sec), and Blade temperature = 30 (°C), leading to the rejection rate = 80.03% and the flux = 329.88 (L/(m^2·h)). Similarly, in the DMAc bath, we find that when PVC Wt% = 16%, DMAc Wt% = 78%, Additive Wt% = 2%, k = 0.8 (PVB Wt%=4%), Additive type = PVP k90, Evaporation time = 5 (sec), Blade temperature = 30 (°C), and Bath concentration= 80 (mg/L), we have the rejection rate = 81.39% and the maximum flux = 271.61 L/(m^2·h). Although our results are not guaranteed to be globally optimal, they are much robust than the best measurement, which has the rejection rate = 82.07% and the flux = 122.70 L/(m^2·h) (with the processing parameters PVC Wt% = 12.6%, DMAc Wt% = 77%, Additive Wt% = 5%, k = 0.7 (PVB Wt%=5.4%), Additive type = PEG600, Evaporation time = 10 sec, Blade temperature = 60 °C, Bath type = DMAc, and Bath concentration= 80 mg/L). To check the accuracy of the models used to optimize membrane performance, we fabricated PVC/PVB flat sheet membranes strictly under the above optimized parameters. Fig.10 shows the surface and cross-section morphology and the contact angle of the as-prepared membranes. In the case of pure water gelation bath, the rejection rate of the as-prepared membrane was 80.2% and the flux was 318.27

L/(m^2·h), while in the case of DMAc as the solute of gelation bath, the as-prepared membrane has the rejection rate of 86.2% and the flux of 298.5 L/(m^2·h). The results showed that there was a very good agreement between the model predictions and experimental data.

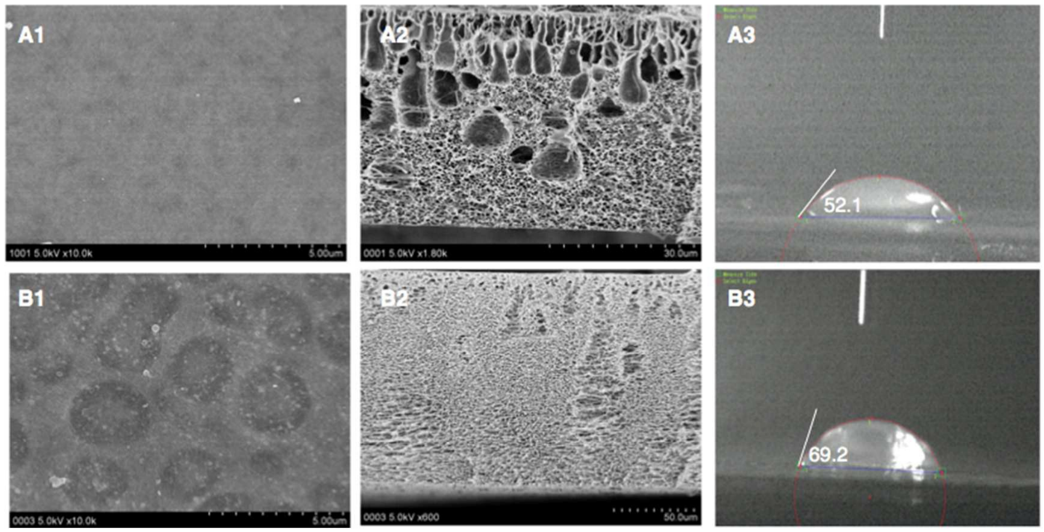


Fig.10 Morphology and the contact angle of PVC/PVB composite membranes (A: the membrane prepared under optimized parameters in the case of using pure water as gelation bath, B: the membrane prepared under optimized parameters in the case of using DMAc as the solute of gelation bath. 1. Suface structure 2. Cross-section structure 3. Contact angle)

5. Conclusions

In this paper, we provide a systematical approach, namely, an SL-based framework for experimental data analytics, for modeling and optimizing membrane responses for complex combinations of membrane features during fabrication. This approach consists of the following procedures. First, control experiments are established to get various membranes with differing performances by combining various fabrication conditions. Second, the characteristics of the feature variables are analyzed in order to ascertain the quality of the data, as well as the data dependencies among the variables. Third, four SL approaches (MART, NN, LR, SVM) are employed to systematically analyze membrane performance and fabrication conditions in a rigorous fashion. Finally, the most reliable and trustful SL model is selected to optimize the fabrication conditions and predict the most favorable performance of PVC/PVB ultrafiltration membranes. During this last step, we analyze multiple responses simultaneously with multiple input feature variables. In this way, we eliminate most unnecessary assumptions that are traditionally proposed by other methods. In addition, this approach simplifies the analysis process by using a unified SL framework that has been thoroughly investigated by machine learning communities[24]. This advantage surpasses previously reported DOE approaches in that these standard SL approaches provide smaller biases and variances for data analysis. Thus, the SL approaches offer us a more standard method not only in procedure but also with more rigorous results.

Additionally, we glean several interesting findings from this research. One is how to find the optimal mixture of feature compounds for the fabrication processes more effectively and efficiently. Another is that among the tested SL approaches, the NN method provides the most reliable and trusted results. In the future, we will investigate how to develop a recursive and automated data-driven experimental analytics approach to design performance-specific membranes more effectively and efficiently.

Acknowledgments

We acknowledge financial support from the Division of International Cooperation & Exchange of Shanghai Jiao Tong University, China and the Newton Research Collaboration Award from Royal Academy of Engineering, UK (Reference: NRCP/1415/261). In addition, we are grateful to sponsors of the Center for Sustainable Development & Global Competitiveness (CSDGC) at Stanford University for additional financial support. Special thanks are given to the researchers at CSDGC for their technical support. Additionally, we thank Weimin Wu and Ting Wang for their valuable discussion.

References:

1 S. Ramesh, A. H. Yahaya and A. K. Arof, *Solid State Ionics*, 2002, **148**, 483–486.
2 Z. Yu, X. Liu, F. Zhao, X. Liang and Y. Tian, *J. Appl. Polym. Sci.*, 2015, **132**, 41267.
3 N. Wang, A. Raza, Y. Si, J. Yu, G. Sun and B. Ding, *Journal of Colloid and Interface Science*, 2013, **398**, 240–246.
4 S. Chuayjuljit, R. Thongraar and O. Saravari, *Journal of Reinforced Plastics and Composites*, 2008, **27**, 431–442.
5 M. Jakic, N. S. Vrandecic and I. Klaric, *Polymer Degradation and Stability*, 2013, **98**, 1738–1743.
6 X. Zhao and N. Zhang, *Journal of Tianjing University of Science and Technology*, 2007, **22**, 36–39.

7 J. zhu, L. Chi, Y. Zhang, A. Saddat and Z. Zhang, *Water Purification Technology*, 2012, **31**, 46–54.
8 Y. Peng and Y. Sui, *Desalination*, 2006, **196**, 13–21.
9 Y. Sui, *Beijing University of Technology Thesis for Master Degree*, Beijing, China, 2004, 24–26.
10 X. Zhao and K. Xu, *Plastics Sci.& Technology*, 2010, 1–6.
11 E. Corradini, A. F. Rubira and E. C. Muniz, *European polymer journal*, 1997, **33**, 1651–1658.
12 S.Y. L. Leung, W.H. Chan and C.H. Luk, *Chemometrics and Intelligent Laboratory Systems*, 2000, **53**, 21–35.
13 S.Y. Lam Leung, W.H. Chan, C.H. Leung and C.H. Luk, *Chemometrics and Intelligent Laboratory Systems*, 1998, **40**, 203–213.
14 W.H. Chan and S.C. Tsao, *Chemometrics and Intelligent Laboratory Systems*, 2003, **65**, 241–256.
15 M. Khayet, C. Cojocaru, M. Essalhi, M. C. García-Payo, P. Arribas and L. García-Fernández, *DES*, 2012, **287**, 146–158.
16 L. Wenjau and O. Soonchuan, *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference*, 2010, **2**, 50–54.
17 P. W. Araujo and R. G. Brereton, *Trends in Analytical Chemistry*, 1996, **15**, 63–70.
18 K. I. Wong, P. K. Wong, C. S. Cheung and C. M. Vong, *Energy*, 2013, **55**, 519–528.
19 Y. Reich and S. V. Barai, *Artificial Intelligence in Engineering*, 1999, **13**, 257–272.
20 B. L. Whitehall, S.-Y. Lu and R. E. Stepp, *Artificial Intelligence in Engineering*, 1990, **5**, 189–198.
21 F. J. Alexander and T. Lookman, *Novel Approaches to StatisticalLearning in Materials Science*, Informatics for Materials Science and Engineering, 2013.
22 S. S. Madaeni, N. T. Hasankiadeh, A. R. Kurdian and A. Rahimpour, *Separation and Purification Technology*, 2010, **76**, 33–43.
23 X. Xi, Z. Wang, J. Zhang, Y. Zhou, N. Chen, L. Shi, D. Wenyue, L. Cheng and W. Yang, *DESALINATION AND WATER TREATMENT*, 2013, **51**, 3970–3978.
24 C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Verlag, 2006.

Figures

Fig. 1 Scatter plots over measurements

Fig. 2 Data mining procedure for each SL model, where ovals and rounded rectangles denote the input and estimated variables, respectively

Fig. 3 Residuals versus Predicted values plots for rejection rate and flux

Fig. 4 Importance plots of predictors on each indicator

Fig. 5 Box-plots of testing R-squared values over 50 training processes with different hidden layer sizes

Fig. 6 Prediction versus response plots for training, validation, testing, and the whole data set; target and output denote the true response and the predicted response by NN, respectively

Fig. 7 Training data and fitting curves of rejection rate and flux in the subspace of PVC Wt% and DMAc Wt% using SVM

Fig. 8 Training data and fitting curves of rejection rate and flux in the subspace of PVC Wt% and DMAc Wt% using NN

Fig. 9 Feasible combinations of PVC Wt% and DMAc Wt% for specific constraints on indicators fixing all other processing parameters

Fig. 10 Morphology and the contact angle of the as-prepared optimized membranes (A: the membrane prepared under optimized parameters in the case of using pure water as gelation bath, B: the membrane prepared under optimized parameters in the case of using DMAc as the solute of gelation bath. 1. Suface structure, 2. Cross-section structure, 3. Contact angle.)

Figure 1

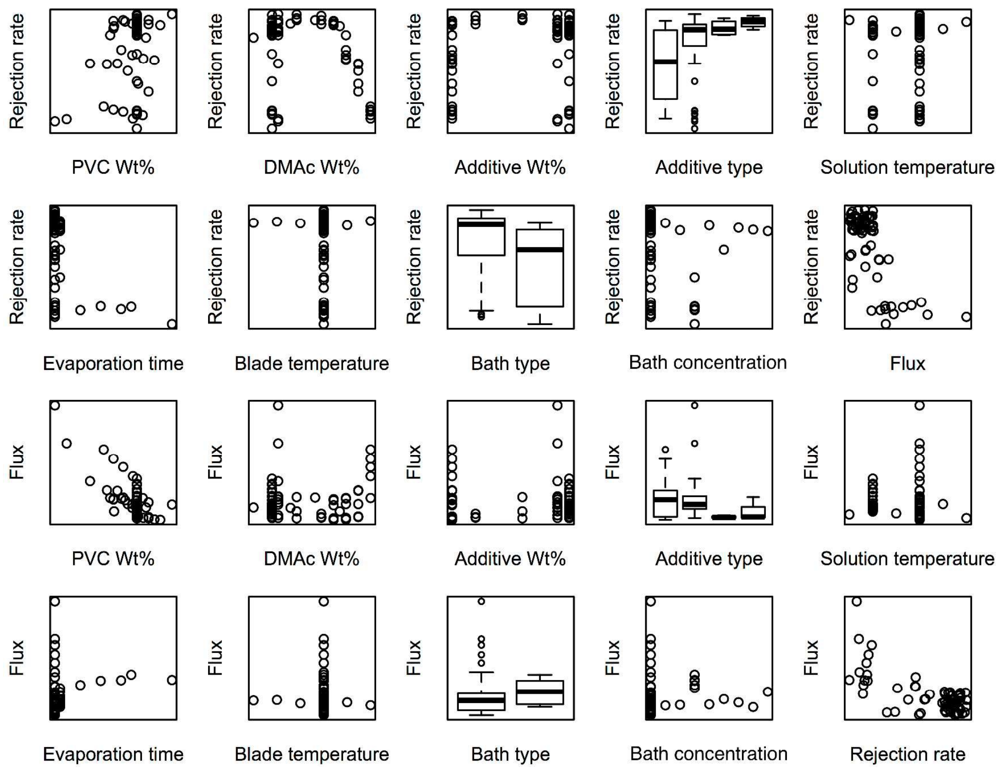

Rejection rate
PVC Wt%
DMAc Wt%
Additive Wt%
Additive type
Solution temperature
Evaporation time
Blade temperature
Bath type
Bath concentration
Flux
Rejection rate

Figure 2

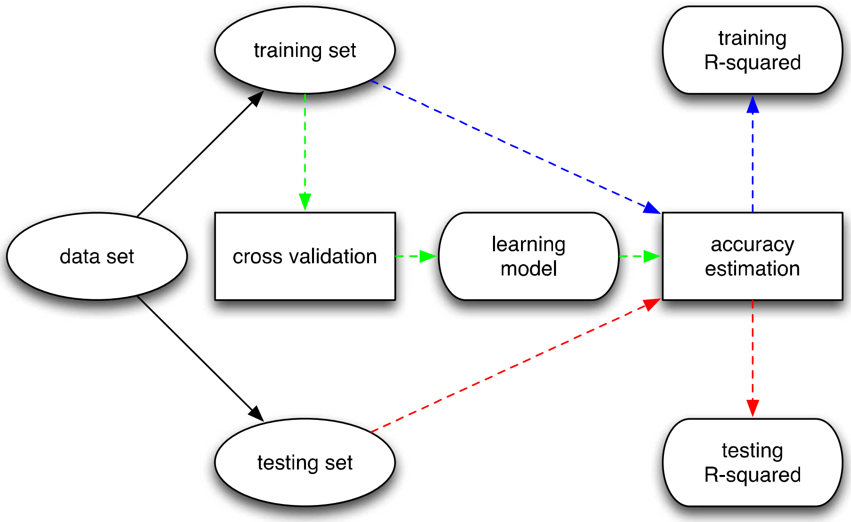

training set
training
R-squared
data set
cross validation
learning
model
accuracy
estimation
testing set
testing
R-squared

Figure 3

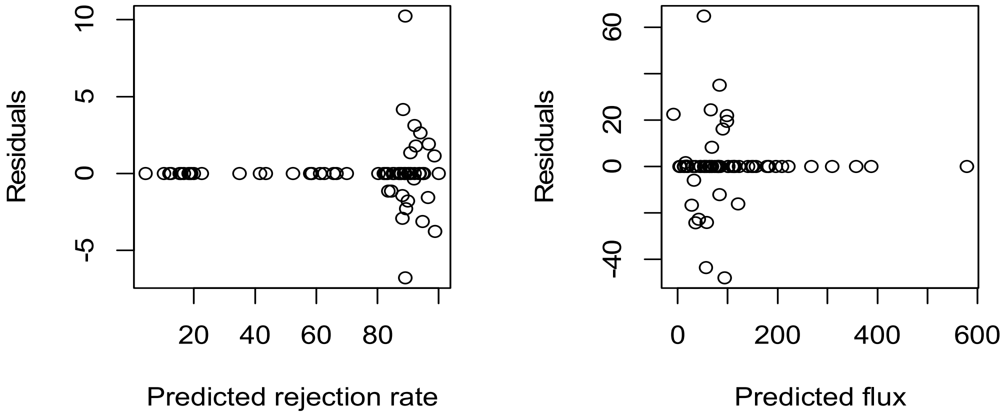
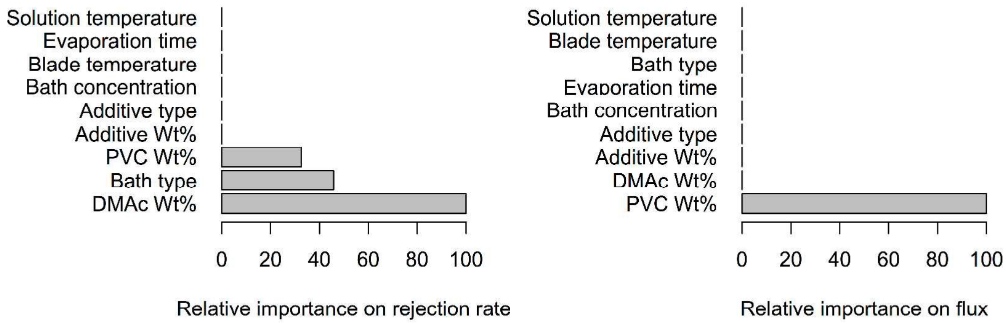

Residuals
10
5
0
-5
20
40
60
80
Predicted rejection rate
Residuals
60
20
0
-40
0
200
400
600
Predicted flux

Figure 4



Solution temperature
Evaporation time
Blade temperature
Bath concentration
Additive type
Additive Wt%
PVC Wt%
Bath type
DMAc Wt%
0
20
40
60
80
100
Relative importance on rejection rate
Solution temperature
Blade temperature
Bath type
Evaporation time
Bath concentration
Additive type
Additive Wt%
DMAc Wt%
PVC Wt%
0
20
40
60
80
100
Relative importance on flux

Figure 5

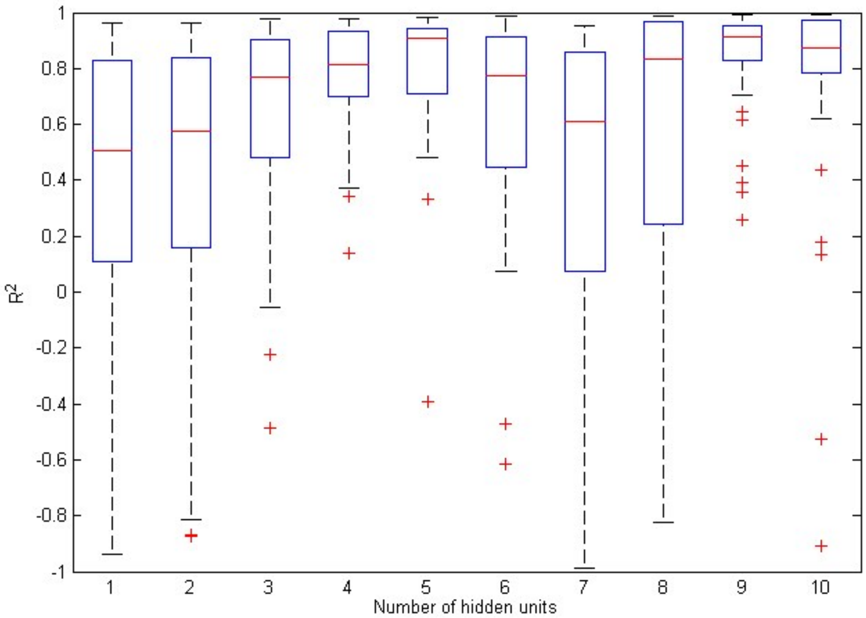

1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
R2
1 2 3 4 5 6 7 8 9 10
Number of hidden units

Figure 6

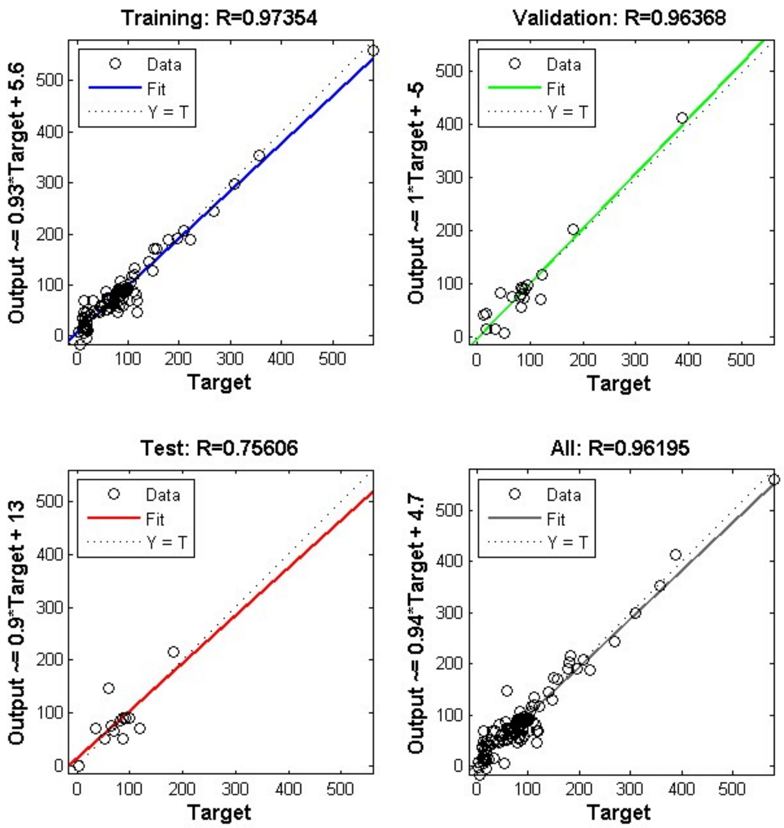
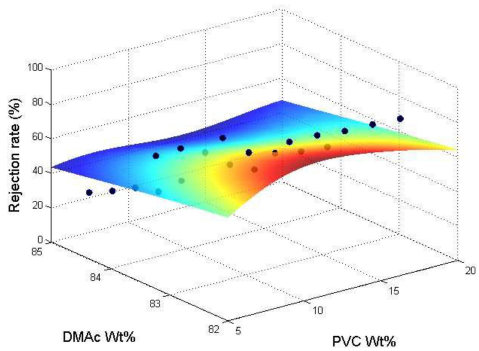

Training: R=0.97354
Output ~= 0.93*Target + 5.6
Data
Fit
Y = T
Target
Validation: R=0.96368
Output ~= 1*Target + -5
Test: R=0.75606
Output ~= 0.9*Target + 13
All: R=0.96195
Output ~= 0.94*Target + 4.7

Figure 7


Rejection rate (%)
100
80
60
40
20
0
85
84
83
82
DMAc Wt%
5
10
15
20
PVC Wt%

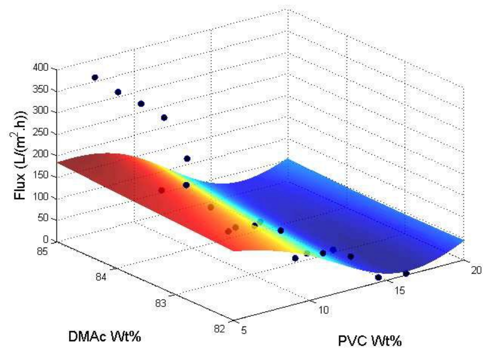

Flux (L/(m^2.h))
400
350
300
250
200
150
100
50
0
85
84
83
82
DMAc Wt%
5
10
15
20
PVC Wt%

Figure 8

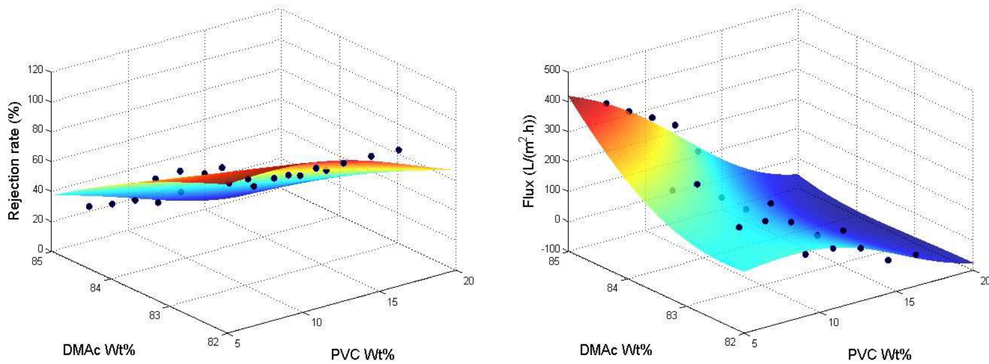

Rejection rate (%)
120
100
80
60
40
20
0
85
84
83
82
DMAc Wt%
5
10
15
20
PVC Wt%
Flux (L/(m^2.h))
500
400
300
200
100
0
-100
85
84
83
82
DMAc Wt%
5
10
15
20
PVC Wt%

Figure 9

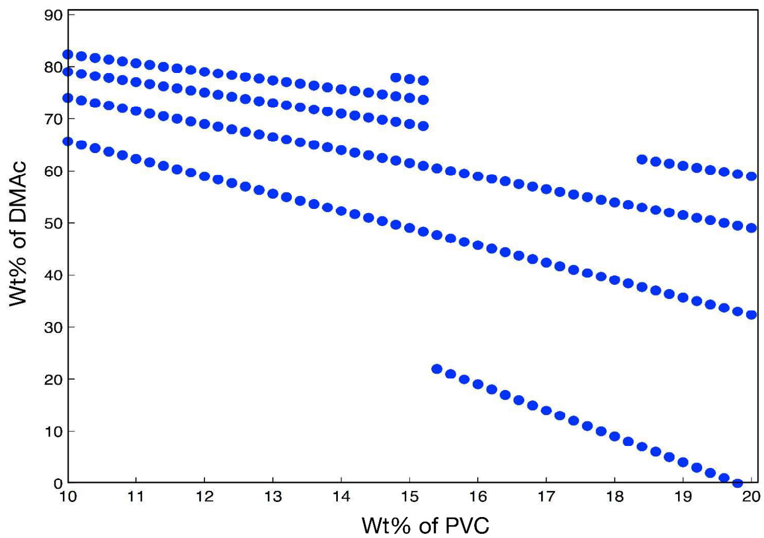
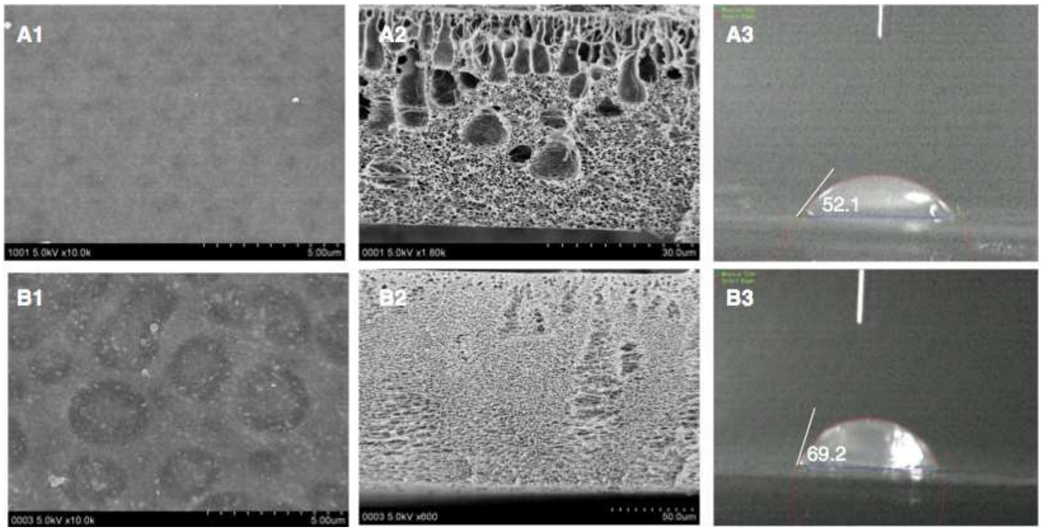

90
80
70
60
50
40
30
20
10
0
10
11
12
13
14
15
16
17
18
19
20
Wt% of DMAc
Wt% of PVC

Fig. 10



Tables

Table.1 Summary of performance of different SL models

Table 1

	MART	**NN**	**LR**	**SVM**
Rm(y1)	0.2122	0.8897	0.6577	0.8065
Rm(y2)	0.0725	0.9175	0.6887	0.6583
Rn(y1)	0.0784	0.6344	0.3104	0.4344
Rn(y2)	-0.0329	0.8093	0.1800	0.6583

Graphical abstract