© IWA Publishing 2020
1238

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.095

Application of machine learning algorithms in MBR simulation under big data platform

Weiwei Li[a,*], Chunqing Li[a] and Tao Wang[b]

[a] School of Computer Science and Technology, Tiangong University, Tianjin 300387, China

[b] School of Environmental and Chemical Engineering, Tiangong University, Tianjin 300387, China

*Corresponding author. E-mail: liweiwei0725@163.com

Abstract

Membrane bioreactors (MBRs) are a sewage treatment process that combines membrane separation with bioreactor technology. It has great advantages in sewage treatment. Membrane fouling hinders MBR process development, however. Studies have shown that the degree of membrane fouling can be judged using the membrane flux rate. In this study, principal component analysis was used to extract the main factors affecting membrane fouling, then the random forest algorithm on the Hadoop big data platform was used to establish an MBR membrane flux prediction model, which was tested. In order to verify the model's effectiveness, BP neural network and SVM support vector machine models were established using the same experimental data. The experimental results from the different models were compared, and the results showed that the random forest algorithm gave the best MBR membrane flux predictions.

Key words: big data platform, machine learning, membrane bioreactor, random forest algorithm, simulation

Highlights

- We used the principal component analysis method to obtain the main influence factors of MBR membrane flux.
- We used the random forest algorithm on the Hadoop big data platform to establish a simulation prediction model of MBR membrane flux, and realized the membrane flux prediction.
- Through comparison with other algorithms, our algorithm works better.

INTRODUCTION

Humans depend on water resources but, with the development of society, such resources have suffered serious pollution and water shortages have become a reality to be faced. Water pollution reduces product quality, hinders industrial development, affects the ecological environment, and harms human health. For a long time, traditional aerobic biological sewage treatment technology – for example, the activated sludge process – has played an important role in industrial and domestic sewage treatment. Due to the various disadvantages of traditional activated sludge, improved technologies have been developed, like membrane bioreactor technology (MBR), a high-efficiency process combining membrane separation and biological treatment technology (Chang *et al.* 2011; Li *et al.* 2019). MBR can separate mud and water efficiently through the membrane module, and the effluent can be used directly. It has the advantages of a small footprint, easy automation, and good effluent quality. With the development of wastewater treatment technology, the membrane bioreactor market has boomed (Braak *et al.* 2011; Wang *et al.* 2014; Meng *et al.* 2017).

Membrane contamination hinders MBR wastewater treatment development and is a direct cause of decreases in membrane flux, the magnitude of which is a measure of the degree of contamination

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.095

(Zhang *et al.* 2006; Alibardi *et al.* 2014). To help improve the MBR process, intelligent computer simulation models can be used to predict membrane flux. At present, MBR flux prediction is focused mostly on mathematical models, BP (back propagation) neural network models and other methods (Lee *et al.* 2002; Barello *et al.* 2014). Mathematical simulation helps to improve the MBR process. Kapumbe *et al.* (2019) used the ASM3 model to simulate MBR wastewater treatment (Kapumbe *et al.* 2019), for instance, while Khan *et al.* (2009) established a mathematical model to predict the degree of membrane pollution (Khan *et al.* 2009). These methods have been useful but have not achieved membrane flux prediction. Jun *et al.* (2019) established a regression model to predict flux through a ceramic membrane treating wastewater, based on Colin Maclaurin mathematical principles (Jun *et al.* 2019). Their model does not reflect the mechanism of membrane fouling well and the physical meaning of the parameters is not precise, because of which the model is generally poor and does not predict fouling accurately. Li *et al.* (2014) established a genetic algorithm optimized BP neural network to predict MBR membrane flux (Li *et al.* 2014). BP neural networks learn on the basis of gradient descent, which is prone to overfitting and requires a lot of test data to ensure prediction accuracy.

The random forest algorithm is a machine learning algorithm based on a classification tree. It has the advantages of fast classification speed and low parameter adjustment, and can effectively avoid overfitting while processing large, multi-dimensional data sets efficiently (Strobl *et al.* 2008; Lee *et al.* 2010; Ross & Allen 2014). The algorithm has achieved good results in the fields of medical treatment (Chen & Liu 2005; Lin *et al.* 2011), fire protection (Oliveira *et al.* 2012), fault diagnosis (Cerrada *et al.* 2016), agriculture (Naidoo *et al.* 2012; Wang *et al.* 2016) and other fields. Hadoop is an open source distributed computing platform developed by Apache. Its main core includes the distributed file system HDFS (Hadoop Distributed File System) and MapReduce computing framework (Ghazi & Gangodkar 2015). The random forest algorithm based on the Hadoop platform has been used very extensively in applied research. Wu *et al.* (2019) proposed an improved random forest algorithm combined with the MapReduce computing framework (Wu *et al.* 2019). Masarat *et al.* (2016) combined the random forest and Hadoop platforms to design an intrusion detection system (Masarat *et al.* 2016). Fan *et al.* (2018) used the pair to build a big data analysis platform for highway travel time prediction (Fan *et al.* 2018). To achieve better membrane flux prediction, the random forest algorithm is also used on the Hadoop platform. The main factors are used as the algorithm's input layer with the membrane flux as the output layer.

ANALYSIS OF FACTORS INFLUENCING MEMBRANE FOULING

An MBR is complicated. All parameters involved in sewage treatment can cause membrane fouling and, with changes in operating methods and conditions, the factors affecting the MBR change constantly. Microbial, inorganic and organic pollution cross and affect each other, between them constituting the main types of MBR membrane pollution. Because of this, membrane flux is affected by a combination of factors comprising mainly mixed liquor suspended solids (MLSS), temperature, operating pressure, total resistance, COD, etc (Kulesha *et al.* 2018). Due to the numerous operating conditions and many factors affecting membrane flux in the MBR process, principal component analysis (PCA) is used to simplify the MBR membrane flux prediction model and improve the effectiveness of membrane flux prediction.

PCA is a statistical analysis method based on the idea of dimensionality reduction to convert multiple indicators into a few comprehensive indicators. It can combine multiple relevant original variables linearly into unrelated principal component factor (Liu *et al.* 2016). It converges fast, does not require a basis function, and can solve irregular data distribution in a limited area. PCA is

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.095

used widely in communication technology, statistical analysis, and image processing. Its use in selecting membrane fouling factors includes several steps:

(1) Determine the main variables affecting MBR membrane flux and construct a matrix of influencing factors: MLSS, total resistance, operating pressure, COD, pH, temperature. Set to X.
(2) Normalize the matrix X by zero-mean normalization to obtain a new matrix A:

$$A_{ij} = (X_{ij} - \overline{X_j})/s_{ij},\ i = 1, 2, \ \ldots l.j = 1, 2, \ \ldots, n.$$

where i is the number of samples, j the number of sample components, and s_{ij} the standard deviation of the variable X_{j}.
(3) Solve matrix A to obtain the covariance matrix (S) of A, sort the characteristic roots, obtain the eigenvector matrix (U) and eigenvalue matrix (V) of S by solving.
(4) Decompose the matrix $A = TU^T$ and obtain the principal component matrix (T).
(5) Analyze the eigenvalue matrix V, determine the correlation between the indicators, and obtain three principal components.
(6) Determine the three principal component elements – MLSS, total resistance, and operating pressure – as the random forest algorithm input.

HADOOP AND RANDOM FOREST ALGORITHM

Hadoop platform

The core of Hadoop architecture is HDFS and MapReduce. HDFS is used to implement the underlying distributed file storage, and the MapReduce framework to implement distributed parallel computing. The HDFS cluster belongs to a master-slave structure (Master and Slave), and usually consists of one Namenode and multiple Datanodes. The Namenode is the master server, which coordinates the client's access to files and manages the file system. Datanodes are used for storage and data management data (Slave) (Li *et al.* 2016). Users can store files directly on HDFS, which are then divided into multiple small file blocks and stored on a set of data nodes. The Namenode coordinates the Datanodes' work, performing file system operations like file opening, closing, and renaming. Figure 1 is a Hadoop cluster deployment diagram.

MapReduce, a software framework in the Hadoop architecture, uses a parallel programming model. Software developers can write parallel distributed programs based on it and distribute them to a cluster containing thousands of machines for execution. At the same time, it can ensure that the processing of a large number of data sets is correct. The Jobtracker running independently on the management node and the TaskTracker running on multiple slave nodes jointly form the MapReduce framework. After receiving the tasks, the management node publishes them to slave nodes, and the master node coordinates their operation. Figure 2 is a MapReduce work flow chart.

Random forest algorithm

Random forest is a non-linear modeling tool based on statistical learning theory. It uses Bootstrap resampling to extract multiple samples from the original, model a decision tree for each sample, and then obtain the result by voting (Breiman 2001; Smith *et al.* 2010). Random forest consists of a set of decision tree classifiers, which can be expressed as:

$$h(X, \theta_k), k = 1, 2, \cdots, K \quad (1)$$

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.095

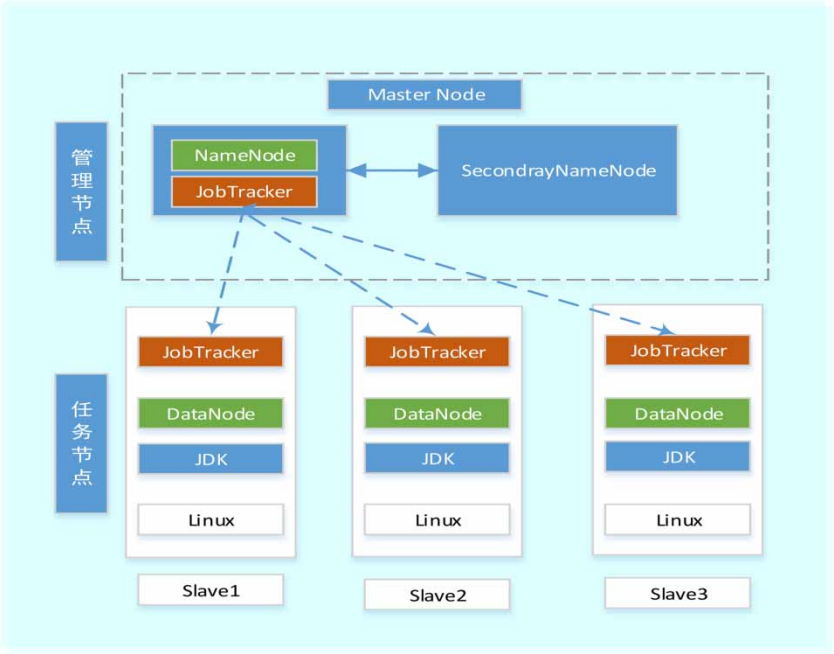

Figure 1 | Hadoop cluster deployment diagram.

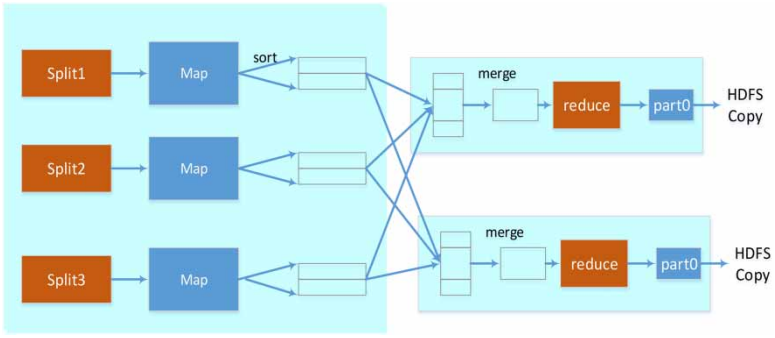

Figure 2 | MapReduce work flow.

where X is the input independent variable, K the number of decision trees, θ_k a random variable, and $h(X, \theta_k)$ a classification regression tree without minus branches constructed by the CART algorithm. CART is a very effective classification and regression method, which uses a bisection recursive segmentation technique to divide the current sample set into two sub-sample sets, so that all non-leaf nodes of the generated decision tree contain two branches. The classification formula of random forest is:

$$f(x_m) = majorityvote\{h_i(x_i)\}_{i=1 \cdots n\_tree} \tag{2}$$

where majorityvote is the number of votes, and n_tree the number of trees in the random forest.

The generalization error that defines the random forest algorithm is:

$$P_{X,Y}(av_k I(h(X, \theta_k) = Y) - \max_{j \neq Y} av_k I(h(X, \theta_k) = j) < 0) \tag{3}$$

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.095

As the number of decision trees in a random forest approaches infinity, the following convergence relationship holds:

$$P_{X,Y}(P_\theta(h(X,\theta)=Y)-\max_{j\neq Y}P_\theta(h(X,\theta)=j)<0) \tag{4}$$

By derivation and proof, the generalization error of the random forest can be obtained as:

$$PE^* \leq \frac{\overline{\rho}(1-s^2)}{s^2} \tag{5}$$

where s is the classification strength and $\overline{\rho}$ the correlation coefficient. Using Equation (5), it can be shown that the larger s and the smaller p, the more accurate the random forest model is.

Mahout

Mahout is an open source project under the Hadoop platform. It implements many classic machine learning algorithms based on the MapReduce mode, mainly including classification, clustering, and recommendation engine algorithms (Bagchi 2015). As the random forest algorithm can be run in parallel, construction of the decision tree can be assigned to different Map tasks, and the trained model is output to HDFS through the reduce process, thereby completing the random forest model's construction.

EXPERIMENTAL MODEL ESTABLISHMENT

Random forest model

Five computers were used to build a Hadoop cluster, one NameNode node, four DataNode nodes, and four Map and Reduce tasks set on each node. The size of the HDFS file block was set at 64M. Figure 3 shows the membrane flux prediction model.

The experimental data for MBR process operation came from a municipal sewage treatment plant. There were 90 data sets. MLSS, total membrane resistance, and operating pressure are used as the model's input, and its output is MBR membrane flux. Two very important random forest parameters are the number of variables (m_try) pre-selected by the tree nodes and the number of trees (n_tree)

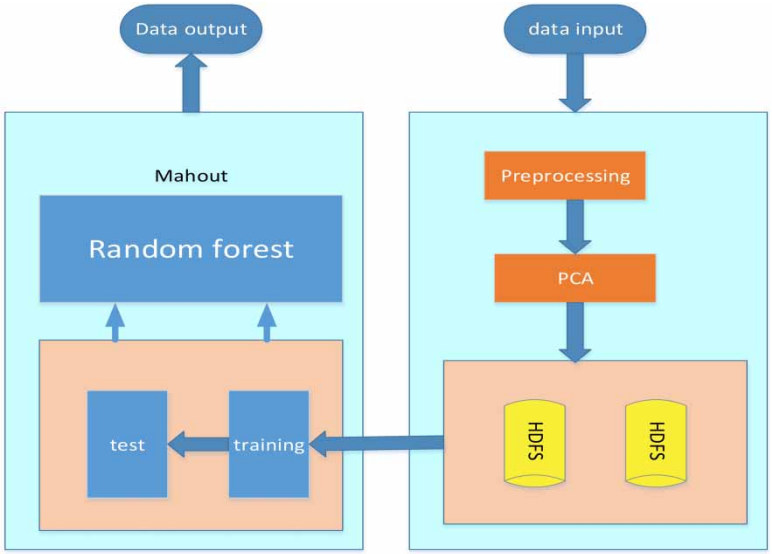

Figure 3 | Membrane flux prediction model.

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.095

in it. On the basis of experiments, values of 2 and 300 were selected for m_try and n_tree, respectively. 84 experimental data sets were generated randomly as training samples, and a nonlinear relationship was established between influencing factors and membrane flux by training. The algorithm steps were:

(1) Resampling using the Bootstrap method to generate the training set.
(2) For each of the M features in the training set, m were selected ($m < M$). For each tree node, the best of the m features for branch growth were selected based on the Gini coefficient.
(3) Steps (1) and (2) were repeated to generate a decision tree, using the training set, to form a random forest. The final result came from voting on each decision tree.

Random forest selects the training set from the original data using the bagging method (Strobl *et al.* 2007). If the original data set contains n records, the probability that a record is not selected is $(1–1/n)^n$; the records that are not selected are called Out of Bag (OOB). When n is very large, $(1–1/n)^n$ converges to 0.368; that is, the proportion of data that may not be selected is close to 36.8%, which guarantees sample diversity. The more decision trees there are in a random forest, the more accurate it will be. In other words, as the number of trees increases, the prediction model's error decreases gradually and then stabilizes. The model error analysis diagram is shown in Figure 4.

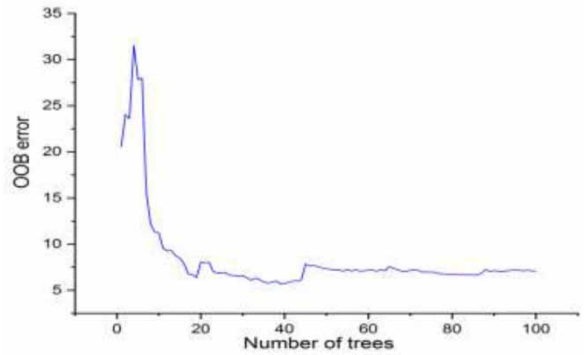

Figure 4 | Model error analysis diagram.

Once the random forest prediction model has been trained, six sets of test data are brought in to obtain the prediction, which is then compared with the experimental value. The experimental and predicted results are shown in Figure 5 and an analysis of the experimental results in Table 1. The average relative prediction error was 4.56%, a good result.

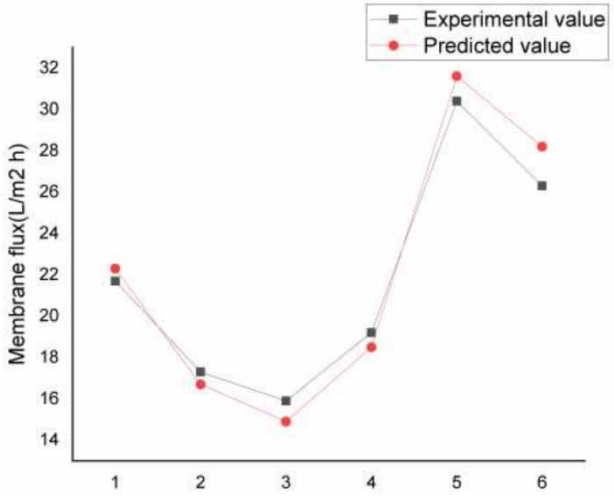

Figure 5 | Analysis of experimental results.

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.095

Table 1 | Analysis of experimental results

Experimental value (L/m^2 h)	21.7	17.3	15.9	19.2	30.4	26.3
Predicted value (L/m^2 h)	22.3	16.7	14.9	18.5	31.6	28.2
Relative error	0.0276	0.0347	0.0629	0.0365	0.0395	0.0722
Mean relative error	0.0456					

Comparative analysis of experimental results

In order to verify the random forest model's effectiveness, a BP neural network model and an SVM (support vector machine) model were established using the same samples. The multi-model comparison is shown in Figure 6.

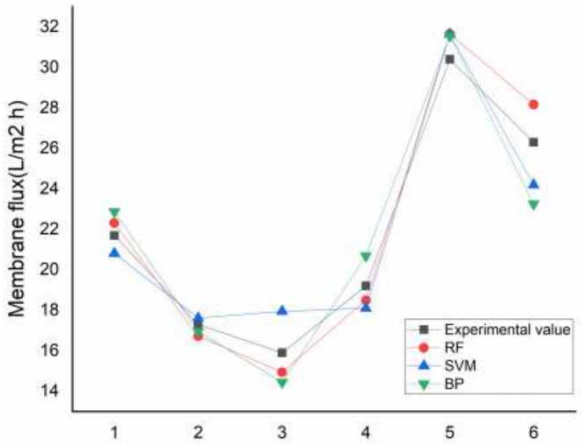

Figure 6 | Multi-model result comparison.

In order to evaluate and compare the models, their mean absolute error (MAE), root mean square error (RMSE), and determination coefficient (R^2) were calculated, using Equations (6)–(8):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)| \tag{6}$$

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{8}$$

where, y_i is the experimental value, $\hat{y}_i$ the predicted value, $\bar{y}_i$ the mean of the experimental values, and n the number of data samples. The better the model, the closer MAE and RMSE are to 0. The value range for R^2 is [0,1]. The numerator represents the sum of the squared differences between the experimental and predicted values, and the denominator the sum of the squared differences of the experimental value and the mean. MAE, RMSE and R^2 were calculated for each model, and the results are compared in Figure 7, with the specific comparison data in Table 2. The random forest

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.095

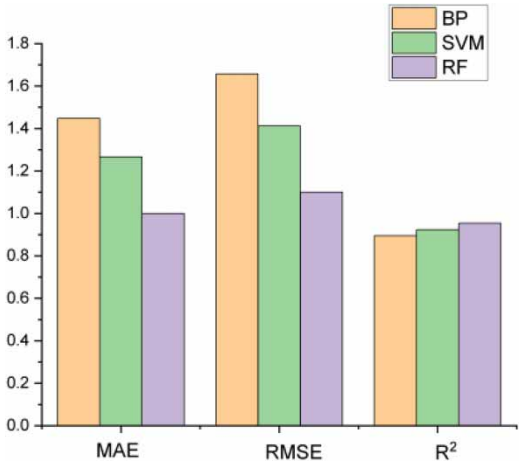

Figure 7 | Comparison of results from multiple models.

Table 2 | Comparison of multiple model results

Methods	MAE	RMSE	R2
BP neural network	1.4500	1.6568	0.8945
Support Vector Machine	1.2667	1.4119	0.9234
Random forest	1.0000	1.1000	0.9535

model has the lowest MAE and RMSE values, and its R^2 value is closest to 1, indicating that it predicts MBR membrane flux more accurately than the others.

CONCLUDING REMARKS

Membrane fouling has always hindered the widespread use of MBR. In this study, PCA was used to reduce the dimensions of the MBR membrane fouling factor set, and enable the selection of three indicators – MLSS, resistance, and pressure – as the main factors affecting MBR membrane flux. At present, the development of big data cloud computing is getting faster and faster. This study used the Hadoop platform and its two cores, HDFS and MapReduce, to build an actual distributed cluster environment. Due to shortcomings in MBR membrane flux prediction in common mathematical models and traditional simulation algorithms, prediction accuracy and prediction time were considered, and the random forest algorithm was applied to MBR membrane flux prediction on the big data platform. The MBR membrane flux prediction model based on random forest was used to predict MBR flux, and the predicted and experimental values compared. The random forest prediction model has higher prediction accuracy. In order to verify the model's validity, it was compared with BP neural network and SVM models, and was shown to have higher prediction accuracy than either. The random forest model is effective in MBR membrane pollution simulation prediction.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.095

REFERENCES

Alibardi, L., Cossu, R., Saleem, M. & Spagni, A. 2014 Development and permeability of a dynamic membrane for anaerobic wastewater treatment. *Bioresource Technology* **161**, 236–244.

Bagchi, S. 2015 Performance and quality assessment of similarity measures in collaborative filtering using Mahout. *Procedia Computer Science* **50**, 229–234.

Barello, M., Manca, D., Patel, R. & Mujtaba, I. M. 2014 Neural network based correlation for estimating water permeability constant in RO desalination process under fouling. *Desalination* **345**, 101–111.

Braak, E., Alliet, M., Schetrite, S. & Albasi, C. 2011 Aeration and hydrodynamics in submerged membrane bioreactors. *Journal of Membrane Science* **379**(1–2), 1–18.

Breiman, L. 2001 Random forests. *Machine Learning* **45**, 5–32.

Cerrada, M., Zurita, G., Cabrera, D., Sanchez, R. V., Artes, M. & Li, C. 2016 Fault diagnosis in spur gears based on genetic algorithm and random forest. *Mechanical Systems and Signal Processing* **70–71**, 87–103.

Chang, J. J., Liang, W., Xiao, E. R. & Wu, Z. B. 2011 Effect of intermittent aeration on the microbial community structure of activated sludge in a submerged membrane bioreactor. *Water and Environment Journal* **25**(2), 214–218.

Chen, X. W. & Liu, M. 2005 Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics* **21**(24), 4394–4400.

Fan, S. K., Su, C. J., Nien, H. T., Tsai, P. F. & Cheng, C. Y. 2018 Using machine learning and big data approaches to predict travel time based on historical and real-time data from Taiwan electronic toll collection. *Soft Computing* **22**(17), 5707–5718.

Ghazi, M. R. & Gangodkar, D. 2015 Hadoop, MapReduce and HDFS: a developers perspective. *Procedia Computer Science* **48**, 45–50.

Jun, X., Liying, Z., Hu, J., Shuili, Y., Binghan, X., Xuan, W. & Kunpeng, Z. 2019 Establishment and testing on membrane flux prediction regression model on treatment of polymer-flooding produced water with ceramic membrane. *Membrane Science and Technology* **39**(05), 112–118 + 135.

Kapumbe, D. J., Min, L., Zhang, X., Kisoholo, M. A. & Yongfeng, L. 2019 Modeling and simulation of membrane bioreactor model based on ASM3 for domestic wastewater treatment. *Applied Ecology and Environmental Research* **17**(5), 11395–11407.

Khan, S. J., Visvanathan, C. & Jegatheesan, V. 2009 Prediction of membrane fouling in MBR systems using empirically estimated specific cake resistance. *Bioresource Technology* **100**(23), 6133–6136.

Kulesha, O., Maletskyi, Z. & Ratnaweera, H. 2018 Multivariate chemometric analysis of membrane fouling patterns in biofilm ceramic membrane bioreactor. *Water* **10**(8), 982.

Lee, Y., Cho, J., Seo, Y., Lee, J. W. & Ahn, K. H. 2002 Modeling of submerged membrane bioreactor process for wastewater treatment. *Desalination* **146**(1–3), 451–457.

Lee, S. L. A., Kouzani, A. Z. & Hu, E. J. 2010 Random forest based lung nodule classification aided by clustering. *Computerized Medical Imaging and Graphics* **34**(7), 535–542.

Li, C., Yang, Z., Yan, H. & Wang, T. 2014 The application and research of the GA-BP neural network algorithm in the MBR membrane fouling. *Abstract and Applied Analysis* **2014**, 673156.

Li, Z., Yang, C., Liu, K., Hu, F. & Jin, B. 2016 Automatic scaling Hadoop in the cloud for efficient process of Big geospatial data. *ISPRS International Journal of Geo-Information* **5**(10), 173.

Li, C., Deng, W., Gao, C., Xiang, X., Feng, X., Batchelor, B. & Li, Y. 2019 Membrane distillation coupled with a novel two-stage pretreatment process for petrochemical wastewater treatment and reuse. *Separation and Purification Technology* **224**, 23–32.

Lin, W. Z., Fang, J. A., Xiao, X. & Chou, K. C. 2011 iDNA-Prot: Identification of DNA binding proteins using random forest with grey model. *Plos One* **6**(9), e24756.

Liu, G., Gao, X., You, D. & Zhang, N. 2016 Prediction of high power laser welding status based on PCA and SVM classification of multiple sensors. *Journal of Intelligent Manufacturing* **30**(2), 821–832.

Masarat, S., Sharifian, S. & Taheri, H. 2016 Modified parallel random forest for intrusion detection systems. *Journal of Supercomputing* **72**(6), 2235–2258.

Meng, F., Zhang, S., Oh, Y., Zhou, Z., Shin, H. S. & Chae, S. R. 2017 Fouling in membrane bioreactors: an updated review. *Water Research* **114**, 151–180.

Naidoo, L., Cho, M. A., Mathieu, R. & Asner, G. 2012 Classification of savanna tree species, in the greater Kruger national park region, by integrating hyperspectral and LiDAR data in a random forest data mining environment. *ISPRS Journal of Photogrammetry and Remote Sensing* **69**, 167–179.

Oliveira, S., Oehler, F., San-Miguel-Ayanz, J., Camia, A. & Pereira, J. M. C. 2012 Modeling spatial patterns of fire occurrence in Mediterranean Europe using multiple regression and random forest. *Forest Ecology and Management* **275**, 117–129.

Ross, J. C. & Allen, P. E. 2014 Random forest for improved analysis efficiency in passive acoustic monitoring. *Ecological Informatics* **21**, 34–39.

Smith, A., Sterba-Boatwright, B. & Mott, J. 2010 Novel application of a statistical technique, random forests, in a bacterial source tracking study. *Water Research* **44**(14), 4067–4076.

Strobl, C., Boulesteix, A. L., Zeileis, A. & Hothorn, T. 2007 Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* **8**, 21.

Water Practice & Technology Vol 15 No 4
doi: 10.2166/wpt.2020.095

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T. & Zeileis, A. 2008 Conditional variable importance for random forests. *BMC Bioinformatics* **9**, 307.

Wang, Z., Ma, J., Tang, C. Y., Kimura, K., Wang, Q. & Han, X. 2014 Membrane cleaning in membrane bioreactors: a review. *Journal of Membrane Science* **468**, 276–307.

Wang, L. A., Zhou, X., Zhu, X., Dong, Z. & Guo, W. 2016 Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *Crop Journal* **4**(3), 212–219.

Wu, Q., Wang, H., Yan, X. & Liu, X. 2019 MapReduce-based adaptive random forest algorithm for multi-label classification. *Neural Computing & Applications* **31**(12), 8239–8252.

Zhang, K., Choi, H., Dionysiou, D. D., Sorial, G. A. & Oerther, D. B. 2006 Identifying pioneer bacterial species responsible for biofouling membrane bioreactors. *Environmental Microbiology* **8**(3), 433–440.