



Build of Machine Learning Proxy Model for Prediction of Wax Deposition Rate in Two Phase Flow Water-Oil

Jalest Septiano, Amega Yasutra and Silvyia Dewi Rahmawati

Bandung Institute of Technology

Jl. Ganesha No. 10, Lebak Siliwangi, Bandung, Indonesia

Corresponding author: jalesttarigan@gmail.com

Manuscript received: January 20th, 2022; Revised: March 29th, 2022

Approved: April 29th, 2022; Available online: May 6th, 2022

ABSTRACT - Wax deposit is one of the major flow assurance experienced in the process of oil production and transportation from sub- surface to surface. Large amounts of data are required to perform modeling using existing thermodynamic models such as carbon number data from HGTC. In this paper, a machine learning algorithm using unified model approach from Huang (2008). Two types of input are implemented in order to simulate influence of feature selection used in training and testing machine learning which are input A consists of water volume fraction (fw), shear stress (τ_w), effective viscosity (μ_e), wax concentration gradient (dC/dT), and temperature gradient (dT/dR) and input B consists of water volume fraction (fw), shear stress (τ_w), effective viscosity (μ_e), wax concentration gradient (dC/dT), temperature gradient (dT/dR), shear stripping variable (SV) dan diffusion variable (DV). The random forest with $N_{tree} = 500$ known to be the best machine learning method compared to others. Based on accuracy parameter it achieves error parameter R-squared (R^2) for training, testing and total data for input A and B are 0.999, 0.992, 0.9975 and 0.999, 0.993, 0.9977, respectively.

Keywords: wax deposit, machine learning, wax deposition rate, two-phase water in oil flow.

© SCOG - 2022

How to cite this article:

Jalest Septiano, Amega Yasutra and Silvyia Dewi Rahmawati, 2021, Build of Machine Learning Proxy Model for Prediction of Wax Deposition Rate in Two Phase Flow Water-Oil, Scientific Contributions Oil and Gas, 45 (1) pp., 27-41.

INTRODUCTION

Wax deposit is one of the problems in the process of production and transportation of oil and gas. Wax molecules will precipitate from the oil when the oil temperature is below the wax appearance temperature (WAT) (Zheng, 2017; Obaseki & Elijah, 2020). The difference in temperature between the temperature in the bulk oil and around the inner wall causes a difference in the concentration of wax dissolved in the oil. Molecular diffusion can occur due to differences in concentration. This is the most reliable mechanism that causes wax to accumulate on the pipe walls (Burger, et al., 1981; Azevedo & Teixeira, 2003). The presence of wax deposits in the pipe walls and reservoirs can cause the pipe walls to become smaller, reduce oil production capacity,

increase handling costs, increase production downtime, reduce production efficiency and cause pipe damage when the pressure due to wax deposits is greater than the maximum yield strength of the pipe. (Kamari, et al., 2014; Chi, et al., 2017, Chu et al., 2017; Alnaimat & Ziaduddin, 2019; Sousa, et al., 2019; Hu, et al., 2019).

Massive research has been done in modeling the mechanism of deposition of wax on the pipe from thermodynamic and empirical model (Pedersen, et al., 1991; Singh, et al., 2000; Matzain, et al., 2000; Zhou, et al., 2015, Joshi, 2017; Obaseki & Elijah, 2020). There are 2 models that can be used in commercial software, namely the RRR model and the Matzain model. The RRR model is used in a multiphase model by considering 2 mechanisms, namely molecular

diffusion and shear dispersion. In the matzain model, 3 mechanisms are considered, namely molecular dispersion and shear dispersion as a mechanism for wax formation and shear stripping as a wax grinding mechanism (Giacchetta, et al., 2019). The two models above are 2 of 3 models available in commercial software. The use of the above model requires many and complex inputs such as oil and wax composition data, oil and wax physical properties data, and pipe specification data.

Huang (2008) developed a unified model using flow loop test data from 9 types of crude oil. The concept of effective wax precipitation (EWP) and effective deposition ratio (EDR) to calculate the wax deposition rate formed at a certain temperature. This model considers 4 parameters that affect wax deposition rate (W) on the pipe wall, namely shear stress (τ_w), viscosity (μ_e), wax concentration gradient (dC/dT) and temperature gradient near the pipe wall (dT/dR).

$$W = f \{ \tau_w, \mu_e, dC/dT, dT/dR \} \quad (1)$$

The wax deposition rate (W) here shows how dangerous the condition of wax deposits that form on the pipe wall is (Huang, 2008). Shear stress and viscosity are variables that show the shear stripping mechanism and the wax concentration gradient and the temperature gradient near the pipe wall indicate the molecular diffusion mechanism.

The development of computational technology increases the development of wax deposition rate prediction models using machine learning algorithms. Qiyu & Ma (2008) predict wax deposition rate using BP Neural network of 26 unified model data and it showed R-squared (R^2) of training and testing data set 0.97 and 0.95. Wei, et al. (2010) proposed a seven hidden neurons back-propagation neural network to estimate wax deposition rate for single phase flow. The estimated result indicated that the error is less than 2%. Kamari, et al. (2014) proposed LSSVM with coupled simulated annealing as optimization strategy to predict wax deposition for single phase flow in 10 different data. This method has error parameter R-squared (R^2) in training and testing data with a value of 0.999. The last is from Xie & Xing (2017) which created random basic function neural network to calculate wax deposition rate which acquire relative error of predicted data to experimental values is 1.5%. In this paper, two phase water in oil flow wax deposition will be modeled using a unified model approach (Huang, 2008).

By considering shear stress, effective oil viscosity, wax concentration gradient, near wall temperature gradient and water volume fraction as input and wax deposition rate as the target supervised machine learning model is created. Supervised machine learning is known to have the ability in self-learning, self-adaptability and successive non-linear fitting. Several supervised machine learning algorithms such as artificial neural network, support vector machine, and random forest will be used to determine the best model to predict wax deposition rate. The best machine learning algorithm will be used as the next wax deposition rate prediction model.

DATA AND METHODS

Selection of good input parameters is very necessary for the development of machine learning in making accurate predictions. The use of 75 data from Kamari, et al. (2014), Fan, et al (2015), and Xie & Xing (2017) is selected. On that data wax deposition rate based on laboratory experiment only available until $fw = 0.3$. In addition to variate the data, a wax deposition model based on the commercial software hypothetical case was created with the existing wax components. The carbon distribution components of the gas chromatography results based on the hypothetical case are shown in Figure 1 While the wax precipitation curve in Figure 2 wax precipitation curve can be obtained from the results of differential scanning calorimetry (DSC) or other heat testing experiment. From that figure wax content is about 10.5% and wax appearance temperature is 63.83°C. Multiflash results from the known carbon number distribution are then prepared as input in commercial software.

Before conducting the simulation, it is necessary to know the basic assumptions used in this study are wax only diffuses through the oil molecules, condition of water in oil flow on horizontal pipe, pipe roughness

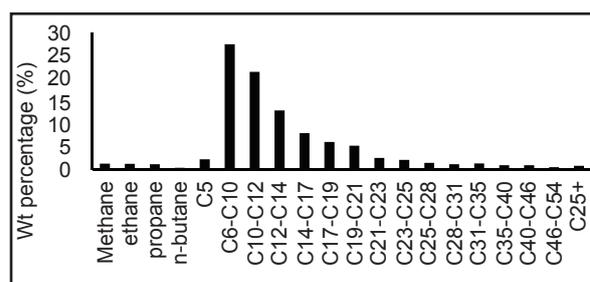


Figure 1
Carbon number distribution
of crude oil A (Multiflash, 2017).

is considered as smooth pipe, fw in deposit is the same as fw in liquid flow. PVT data and wax data is then used in the simulation. A horizontal pipe with a length of 8.2 km is made by discretization per 40 m. This follows a sample case from commercial software simulation. As the purpose of this paper is to build a wax deposition model for water inoil flow, the sensitivity to water cut is carried out from 0.1 to 0.9. Water inversion viscosity will be set at water volume fraction 0.9. In addition to the water cut, other parameters that were carried out by sensitivity analysis were liquid flowrate (Q), inlet temperature (T_{in}). Table 1 below shows the sensitivity of these variables.

The image below shows the simulation results of several scenarios used. Figure 3 shows the effect of fw on the same Q and T_{in} values. Figure 4 shows the wax deposition rate curve at different T_{in} but for the same Q and fw values. Figure 5 shows the wax deposition rate curve at different Q values but for the same T_{in} and fw. From all this sensitivity scenario fw and T_{in} has the bigger effect to shaped of wax deposition rate than T_{in}. Difference T_{in} only effect the front thickness of wax start to build up meanwhile fw and Q tend to reduce wax deposition rate when those two variable are increase.

After performing sensitivity with several cases selected variables such as water volume fraction (fw), shear stress at wall (τ_w), effective oil viscosity (μ_e), wax concentration gradient (dC/dT), temperature gradient near wall (dT/dR) and the wax deposition rate (W) which will be used as input for machine learning. The value of the water volume fraction and the effective viscosity of the oil can be directly obtained from the simulation results. Values for τ_w and dT/dR can be calculated based on the output data from simulation. τ_w is calculated using the Newtonian fluid flow equation assumption for the correlation of turbulent Darcy/Fanning fraction factor with Reynolds number (Wang & Huang, 2014).

$$\tau_w = 4.94 \times 10^{-3} RE^{0.75} \frac{8V}{ID} \mu_e \quad (2)$$

For laminar flow,

$$\tau_w = \frac{8V}{ID} \mu_e \quad (3)$$

Where τ_w is liquid shear stress at wall (Pa), RE is reynolds number, V is fluid velocity (m/s), μ_e is inside diameter of pipe (m), μ_e is effective viscosity of oil (cP). Meanwhile dT/dR can be calculated using the

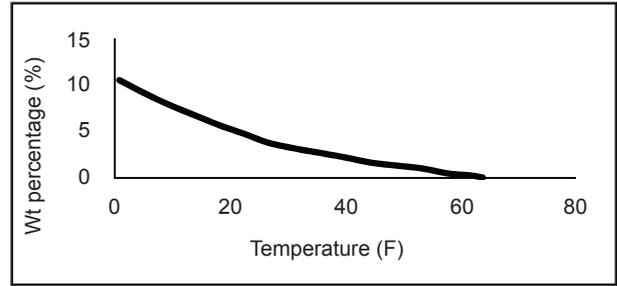


Figure 2

Wax precipitation curve of crude oil A (Multiflash, 2017).

Table 1
Simulation sensitivity analysis

No	fw (fraction)	Q (m ³ /d)	T _{in} (°C)
1	0-0.9	1891	60,70,80
2	0-0.9	2500	60,70,80
3	0-0.9	5000	60,70,80

equations given in the commercial software manual (Giacchetta, et al., 2019; Montero, 2020).

$$dT/dR = \left(\frac{T_b - T_{ws}}{k_{mix}} h h_{in} \right) / 1000 \quad (4)$$

Where dT/dR is radial temperature gradient (°C/mm), T_b is bulk temperature (°C), T_{ws} is wall section temperature (°C), h_{in} is heat capacity inside pipe (W/m²-C), k_{mix} is mix conductivity (W/m°C). Lastly, the dC/dT value can be calculated based on the equation of the line formed in Figure 2.

Based on the simulation results, there are a total of 13995 total input data which will be the final data in making machine learning. The distribution of data for each variable such as the minimum, maximum, standard deviation, and average values is shown in Table 2.

This input data will be divided using deterministic sampling into 80% as training data and 20% as testing data.

A. Feature Engineering Method

In producing robust machine learning, it is necessary to determine the input parameters that have the best relationship to the output or target being sought. The closer the value of correlation coefficient to 1 or -1, the better the relationship between the input and the output. Viscosity and shear stress known to have negative effect on build up of wax. Because of that, those two variables are divided into 1/τ_w and 1/μ_e. Meanwhile, the variables dC/dT and dT/dR are

Table 2
Input data distribution

	fw(-)	$\tau_w(\text{Pa})$	$\mu_e (\text{mPa}\cdot\text{s})$	$dC/dT (10^{-3}/^\circ\text{C})$	$dT/dR (^\circ\text{C}/\text{mm})$	W(g/m ² .h)
Amount	13995	13995	13995	13995	13995	13995
Max	0.9	87.9	780.5	5.073	29.54	18.7
Min	0	1	1.11	0.00007	0.011	0.023
SDV	0.29	16.01	182.84	0.65	4.1	2.16
Ave	0.44	13.65	84.26	0.41	3.31	2.60

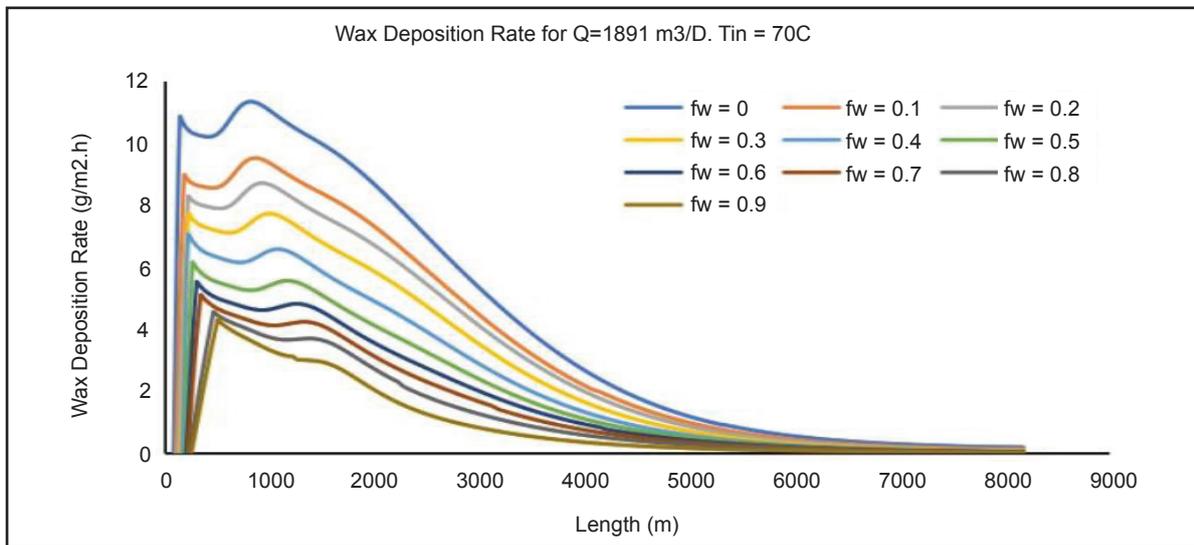


Figure 3
Wax deposition rate on different water fraction.

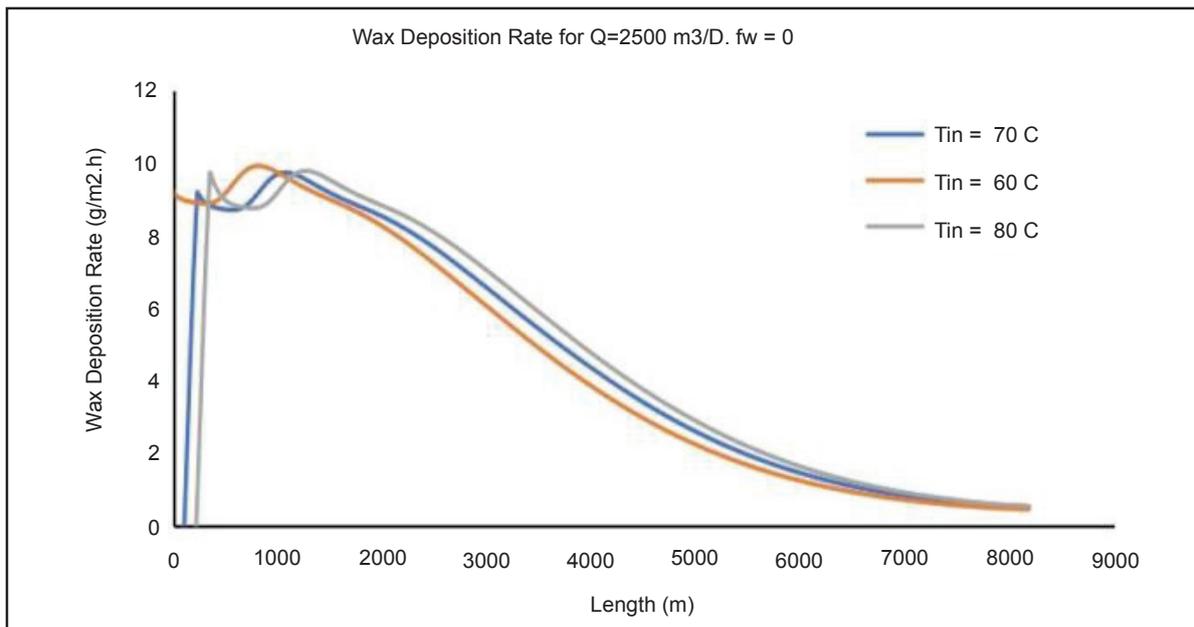


Figure 4
Wax deposition rate on different T_{in} .

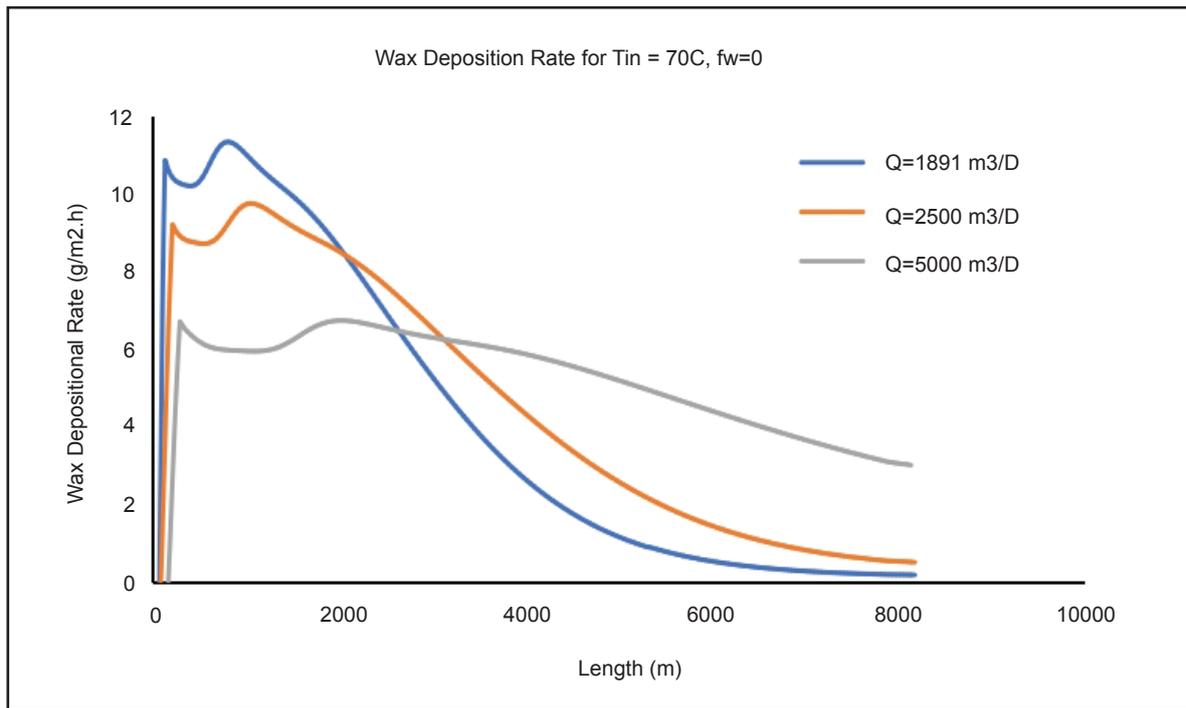


Figure 5
Wax deposition rate on different rate.

multiplied by the normal log so that they become $\ln(dC/dT)$ and $\ln(dT/dR)$. With this new data, it can be seen that there is an increase in the correlation value between the input and the target W as shown in Figure 6.

The relationship between $1/\mu_e$ and wax deposition rate (W) is up to 0.613. Meanwhile, $\ln(dC/dT)$ and $\ln(dT/dR)$ also showed high correlation to target which are 0.85 and 0.69, respectively. Nevertheless, this is not indicated by $1/\tau_w$ which has low correlation to target. The value of the relationship between $1/\tau_w$ and W is 0.16. Therefore, 2 additional features were introduced, namely shear stripping variable and diffusion variable. In the shear stripping variable, the new variables from the 2 basic variables are carried out mathematical operations as follows.

$$SV = \frac{1}{\tau_w \times \mu_e} \quad (5)$$

While the diffusion variable is formulated as follows.

$$DV = \ln(dC/dT \times dT/dR) \quad (6)$$

These two additional features are then included in the correlation coefficient analysis. It was found that these two variables had a fairly good correlation value to the target, namely the shear stripping variable and the diffusion variable 0.46 and 0.83, respectively.

The relationship between these two variables with other parameters can also be seen in Figure 7.

The results of these feature engineering strategy will be used in a supervised machine learning development strategy. Both will be compared in terms of the ability to predict train data and test data that have been generated previously. Table 3 shows 2 types of strategies used in making machine learning models based on the results of feature engineering that has been carried out.

B. Machine Learning Development

The data that will be used as machine learning input will be normalized into a simpler form that is -1 to 1. Min - max normalization of machine learning input data is done using the following equation.

$$\text{norm}_x = \frac{x_{\text{act}} - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} (\text{norm}_{\text{max}} - \text{norm}_{\text{min}}) + \text{norm}_{\text{min}} \quad (7)$$

Where X_{act} is variance value, X_{min} is minimum variance, X_{max} is maximum variance, norm_{max} is maximum normalization value, norm_{min} is minimum normalization value.

Making machine learning in predicting regression of existing data is done using 3 supervised machine learning algorithms, namely back propagation artificial neural network, random forest and support vector machine. Back Propagation Artificial Neural

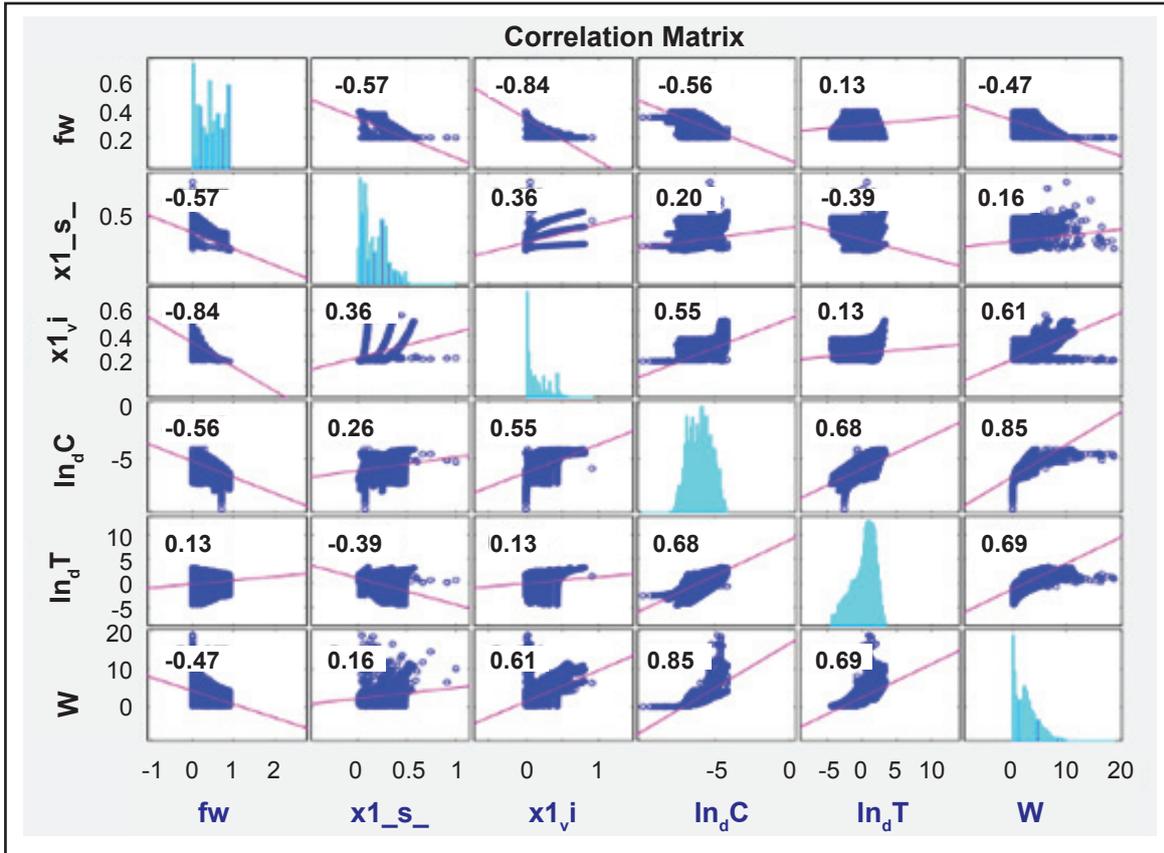


Figure 6
Correlation matrix after feature engineering A.

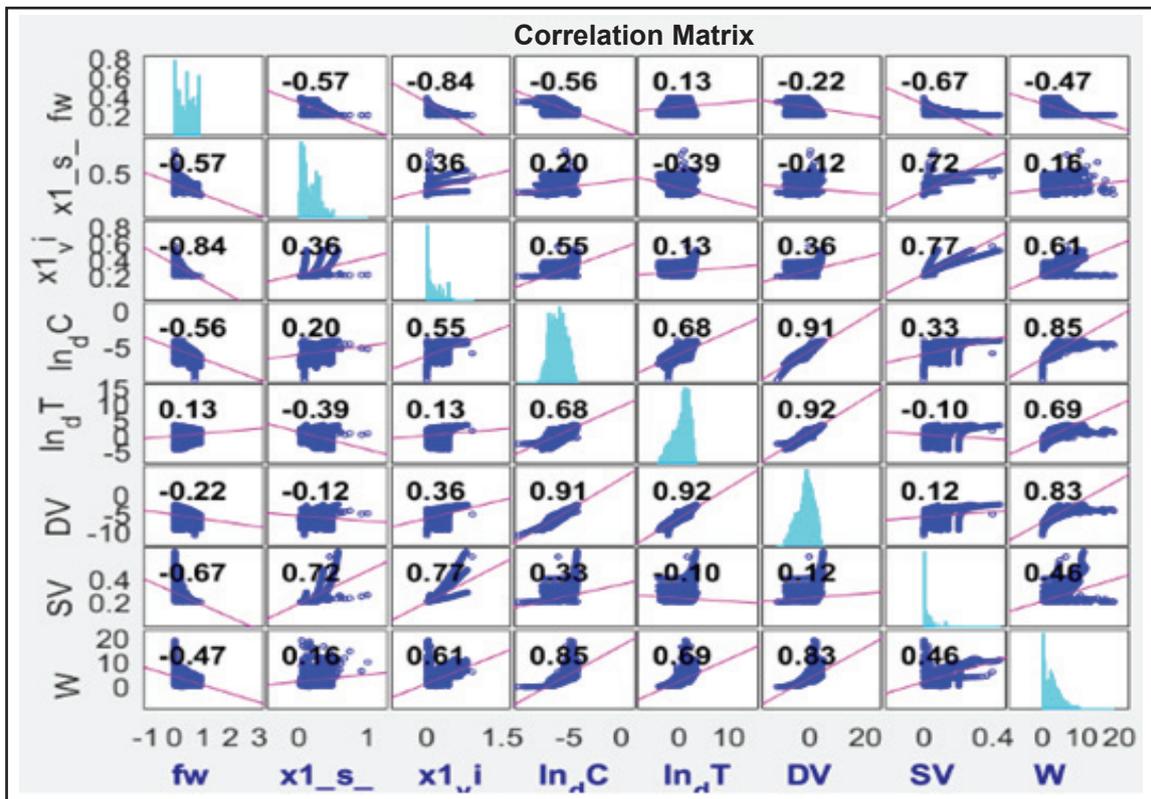


Figure 7
Correlation matrix after feature engineering B.

Network (BP-ANN) systems are made up of processing units called neurons. The BP-ANN training process is used to achieve a balance between memorization and generalization abilities. Thus, the results of the BP-ANN that have been created can be perfectly reused for other new data inputs whose quantities are still within the range of the previous training process (Agustin & Prahasto, 2012).

The first thing to do is to prepare a training sample consisting of inputs and ideal outputs. In this case the inputs are shown in Table 3. When the results of the calculation of the output match the ideal output, the training is stopped. This process can be repeated automatically with self- adaptation of the weight and activation value until the ideal output is obtained (Wei, et al., 2010).

Back propagation artificial neural network (BP-ANN) is used in this paper. In this algorithm, each unit that is in the input layer is associated with the hidden layer and continues to be interconnected with the output layer. This network can consist of 1 hidden layer or many (multilayer network). When this network is given input as training data, then the pattern goes to the hidden layer units to be forwarded to the units in the output layer. Then the output layer units will respond as the target/output of the neural network. When the output does not match the expected target, the output will be returned backward to the hidden layer and input layer and the output will be recalculated. This condition will continue to repeat until the error value between the target and the output is minimal.

Network created using Levenberg-marquardt training function and gradient descent adaptive learning. The use of Levenberg-Marquardt training function because it has a good fitting ability on the training set and also has the fastest training network (Huang, 2008). The amount of learning rate used is 0.01 with the activation function used, namely the tansig function for the input, hidden layer, and output layer. The number of hidden layers and hidden

neurons becomes the sensitivity parameter. Table 4 below shows the sensitivity of BP-ANN.

Random forest is a supervised machine learning development of an iterative decision tree so as to produce a better level of accuracy. Because random forest is an ensemble method from CART, random forest also has no assumptions or is good for use in nonparametric cases (Mulyahati, 2020). This method is used to build a decision tree consisting of root nodes, internal nodes, and leaf nodes by taking attributes and data randomly according to the applicable provisions (Sibirian & Mulyana, 2018). The root node variable which is also the separator variable is a determinant in making the splitting tree which can be done by using the MSE value, Gini ratio, and entropy/information ratio. After getting the separator attribute with the lowest MSE/gini ratio value, it will proceed to the next branch. This continues until the stopping criteria are met, namely the minimum leaf node is met (Mulyahati, 2020).

The final prediction results can be calculated by combining or aggregating the prediction results of each decision tree which can be shown in the following equation.

$$Y_i = \frac{1}{N_{tree}} \sum_{n=1}^{N_{tree}} Y_n \quad (8)$$

Where N_{tree} is number of trees. Random forest was created using Orange data mining software by varying the N_{tree} value. Based on Breiman (2001) the N_{tree} value of 50 has resulted in a satisfactory predictive value. Meanwhile, according to Sutton (2004) the value of $N_{tree} > 100$ will produce a low level of error. Therefore, the N_{tree} values of 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, are used to produce variations in accuracy.

The third machine learning is Support Vector Machine (SVM). Support vector machine basically uses the basic principle of linear classifier. By using kernel functions such as gaussian or polynomial, classification and prediction can be used in non-linear cases. In a space with high dimensions, a hyperplane

Table 3
Featured used on machine learning

Scenario		Input Machine Learning						
A	fw	$1/\tau w$	$1/\mu e$	$\ln(dC/dT)$	$\ln(dT/dR)$	W		
B	fw	$1/\tau w$	$1/\mu e$	$\ln(dC/dT)$	$\ln(dT/dR)$	DV	SV	W

Table 4
Sensitivity of BP-ANN

No	Hidden Layer	Hidden Neuron 1	Hidden Neuron 2
1	1	5	
2	1	7	
3	1	9	
4	1	10	
5	1	15	
6	1	20	
7	1	25	
8	1	30	
9	1	35	
10	1	40	
11	1	42	
12	1	44	
13	1	46	
14	1	48	
15	1	50	
16	1	52	
17	1	54	
18	1	56	
19	1	58	
20	1	60	
21	2	5	5
22	2	7	7
23	2	9	9
24	2	10	10
25	2	15	15
26	2	20	20
27	2	25	25
28	2	30	30
29	2	35	35
30	2	40	40
31	2	42	42
32	2	44	44
33	2	46	46
34	2	48	48
35	2	50	50
36	2	52	52
37	2	54	54
38	2	56	56
39	2	58	58
40	2	60	60

that can maximize the distance (margin) between data classes. Hyperplane in support vector machine can be defined as follows (Octaviani, et al., 2014):

$$f(x) = w^T x + b \quad (9)$$

Where w is the linear regression slope, x is the input vector of the model, b is the intercept linear regression and T is the transpose matrix (Kamari, et al., 2014).

Trial and error of cost and regression loss epsilon (σ) has been done to acquire best SVM prediction. Figure 8 shows the scheme used in making machine learning using Orange data mining.

These three machine learning models will be compared based on the mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and R-squared (R^2) error parameters using the following equation.

$$MAE = \frac{1}{n} \sum_{j=1}^n |W_{pred} - W_{act}| \quad (10)$$

$$MSE = \frac{1}{n} \sum_{j=1}^n (W_{pred} - W_{act})^2 \quad (11)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (W_{pred} - W_{act})^2} \quad (12)$$

$$R^2 = 1 - \frac{\sum_{j=1}^n (W_{(i)act} - W_{(i)pred})^2}{\sum_{j=1}^n (W_{(i)pred} - \text{average}W_{(i)pred})^2} \quad (13)$$

Where W_{pred} is prediction wax deposition rate (g/m^2h) and W_{act} is actual wax deposition rate (g/m^2h).

RESULTS AND DISCUSSION

In the experiment using BP-ANN using input scenario A, it was found that the BP-ANN sensitivity no. 38 at Table 4 where using 2 hidden layers with each hidden layer is 56 has the best error and accuracy.

The error and accuracy values for training, testing and total data using input scenario A are shown in Table 5. Using this model on input scenario B achieve error and accuracy better than input scenario A. The error and accuracy values for training, testing and total data are shown in Table 6. It was also found that the more hidden layers used the better adjustable parameters to define the nonlinear relationship between input parameters and output parameters. This has a positive impact on the network that is made, it will be better at storing memory about the predictions made. So that a more precise solution will be obtained when making predictions using other data (Wei, et al., 2010). Nonetheless, the training time will be longer when hidden layer increase. In this best parameter ANN, it takes 143.3 seconds of training time for input scenario A and 156.2 seconds for input scenario B. The addition of additional feature variables seems to only affect the training time not significantly.

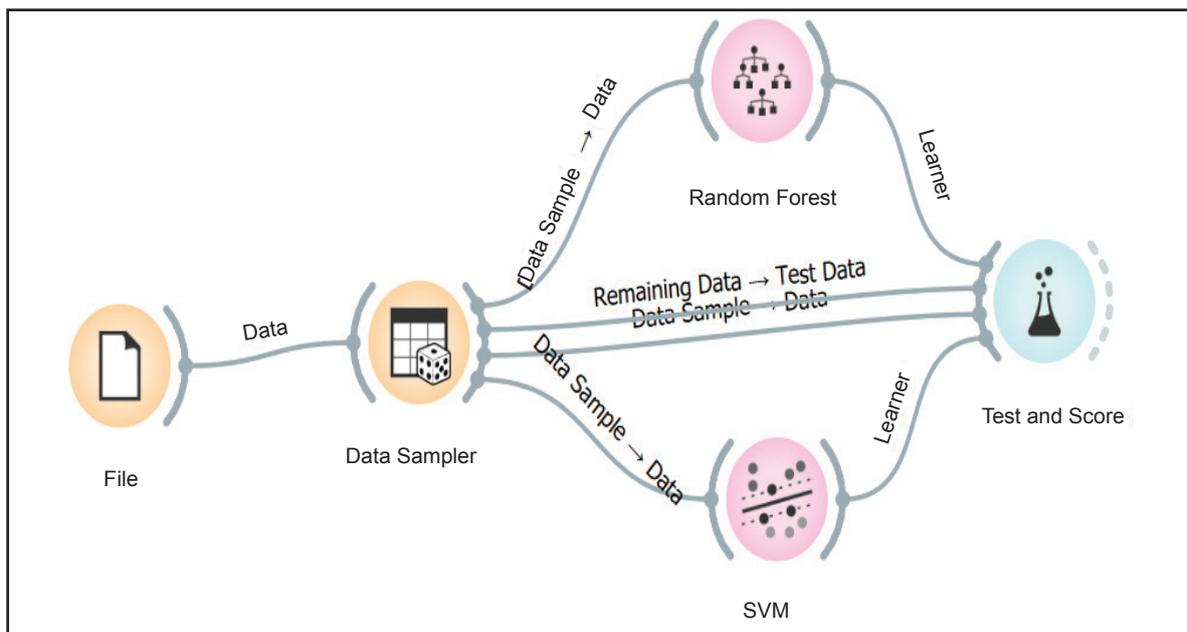


Figure 8
Model establishment of machine learning at Orange Data Mining.

Table 5
Input A scenario accuracy and error using best BP-ANN

	MSE	RMSE	MAE	R
Total	0.022	0.147	0.052	0.995
Training	0.0074	0.087	0.019	0.998
Testing	0.033	0.182	0.04	0.992

Table 6
Input B scenario accuracy and error using best BP-ANN

	MSE	RMSE	MAE	R
Total	0.014	0.117	0.042	0.997
Training	0.00684	0.083	0.018	0.998
Testing	0.031	0.175	0.04	0.992

Table 7
Input A scenario accuracy and error using best random forest

	MSE	RMSE	MAE	R
Total	0.012	0.109	0.022	0.9775
Training	0.006	0.081	0.017	0.999
Testing	0.033	0.182	0.04	0.992

The result of feature selection that has been done previously also shows that the presence of additional features as input for scenario B will result in a better level of predictability than input for scenario A. The increase in total data R-squared (R^2) from 0.995 to 0.997 indicates that the SV and DV parameters have a positive effect on machine learning's ability to recognize target.

Sensitivity of the N_{tree} value will be carried out when modelling using a random forest. The value of growth control on leaf nodes and individual trees is not limited. Based on the experimental results, it was found that $N_{tree} = 500$ produced the best error and accuracy values compared to the others with a training time of data is 21.06 seconds. The results of the accuracy and error of the random forest model $N_{tree} = 500$ can be seen in Table 7. The use of input B scenarios at the same $N_{tree} = 500$ also shows an increase in the accuracy and error obtained. It can be seen in Table 8.

This condition is similar to the experiment using the best parameter BP-ANN. Increasing the Ntree above 500 did not improve the accuracy and error parameter values obtained. The addition of N_{tree} will

only increase the training time so that it will take longer to perform calculations.

The last machine learning is Support Vector Machine (SVM). In making the best SVM model the cost value and regression loss epsilon (σ) values are varied. It was found that in the input scenario A, the best error and accuracy values were in the combination of cost and regression loss epsilon (σ) of 2.05 and 0.1, respectively. This trial and error and can be done by by change the value in red box on Figure 9. The error and accuracy values for training, testing and total can be seen in Table 9. Meanwhile in the input scenario B, the use of cost and regression loss epsilon (σ) of 2.05 and 0.1, respectively, produces better error and accuracy values than the input scenario A. This can be seen in Table 10.

When the cost value is fixed, an increase in epsilon regression loss will reduce prediction accuracy. Meanwhile, at a fixed epsilon regression loss value, an increase in cost value will increase the predictive ability of the data and then it will decrease that predictive ability. In this case the use of cost 2.05 is the optimum value.

Table 8
Input B scenario accuracy and error using best random forest

	MSE	RMSE	MAE	R
Total	0.011	0.104	0.022	0.9777
Training	0.006	0.083	0.017	0.999
Testing	0.029	0.173	0.04	0.993

Table 9
Input A scenario accuracy and error using best SVM

	MSE	RMSE	MAE	R
Total	0.25	0.5	0.369	0.948
Training	0.252	0.5	0.37	0.946
Testing	0.246	0.496	0.365	0.945

Table 10
Input B Scenario Accuracy and Error using Best S

	MSE	RMSE	MAE	R
Total	0.25	0.5	0.361	0.949
Training	0.253	0.5	0.363	0.946
Testing	0.239	0.489	0.355	0.946

Regardless type of machine learning used, it was found that the use of random forest produces the best error and accuracy values in the use of input scenario A and input scenario B. In addition, the use of input scenario B which defines additional parameters or features increases the predictive ability. As for each machine learning, slightly increase of accuracy observed from random forest which has total data R-squared (R^2) = 0.9975 to R-squared (R^2) = 0.9977. This value is quite good comparing to another machine learning model that previously created in another article to predict wax deposition rate using BP-ANN

or LSSVM which acquire R-squared at range 0.95 to 0.999 using test data. (Qiyu & Ma, 2008; Wei, et al., 2010; Kamari, et al., 2014).

Here figure 10 described actual vs predicted wax deposition rates for input scenario B using random

forest which is the best algorithm machine learning in this case.

To prove the ability of the random forest algorithm that has been made previously in making predictions, in this section predictions will be made in several

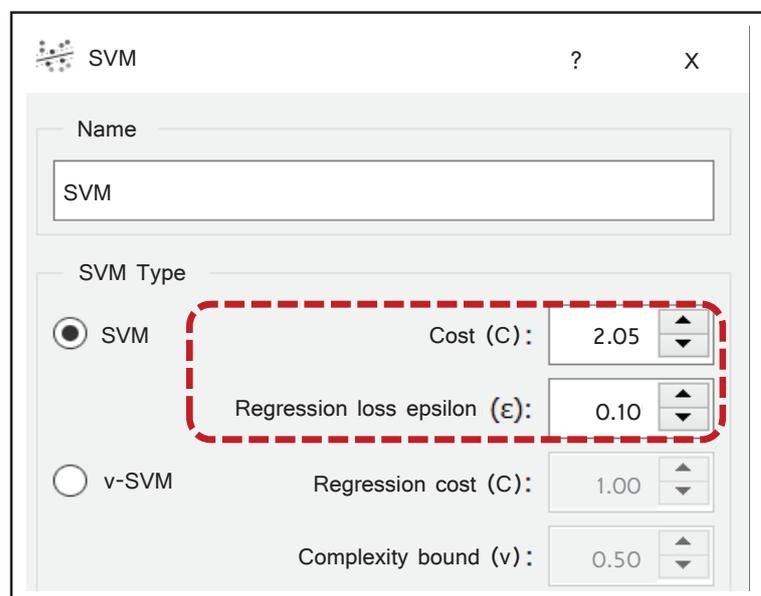


Figure 9
Sensitivity of cost and regression loss epsilon in SVM.

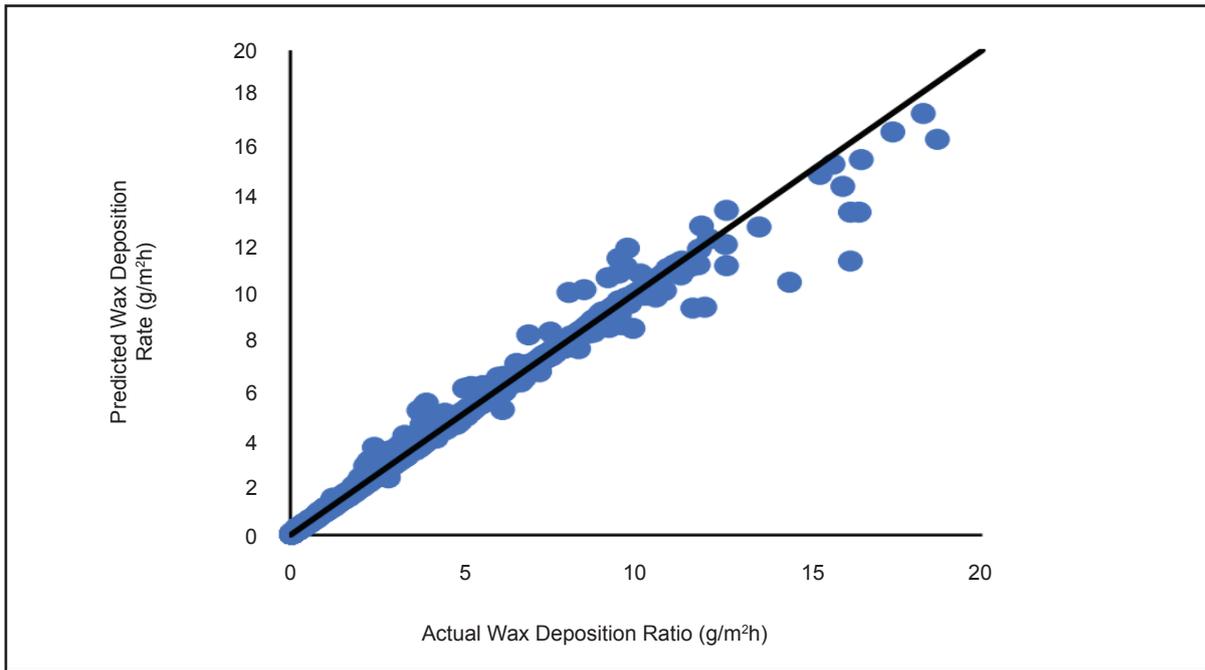


Figure 10
Input B scenario comparison between predicted and actual wax deposition rate for best random forest.

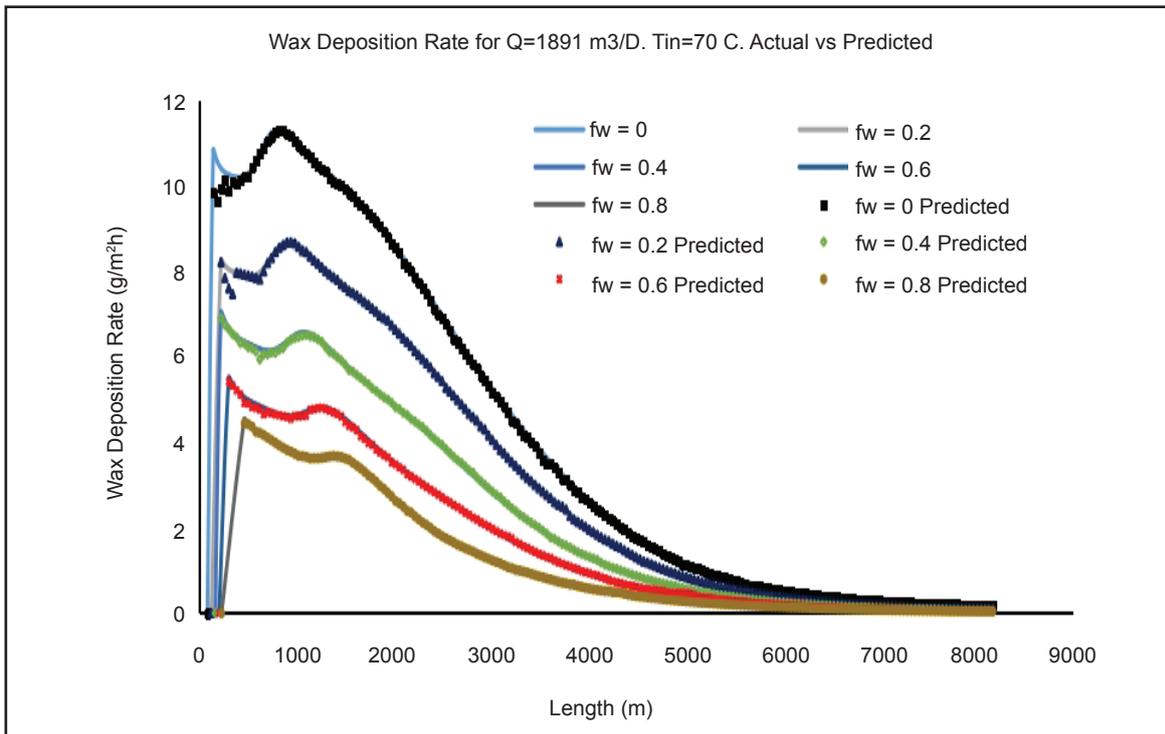


Figure 11
Comparisons of wax deposition ratio between dynamic multiphase flow and best machine learning (Case 1).

cases. The case used in this prediction is still the same as that described in the Data and Method section where wax content of waxy crude is 10.5%, wax appearing temperature is 63.83°C, and density of oil at standard condition is 800°C.

Figure 11 shows the results of predictions and calculations using dynamic multiphase flow at a flow rate of 1891 m³/D, an inlet temperature of 70 °C, and variations of water fraction at 0, 0.2, 0.4, 0.6, and 0.8 (case 1). The results of the predictions using

the random forest used show a similar pattern to the results of calculations using dynamic multiphase flow for each case. The mean absolute percentage error (MAPE) obtained also shows a value of 0.73% to 1.32%. Another case is shown in Figure 12 where the water fraction is 0 inlet temperature is 70°C and variations of Liquid rate (case 2). The good predictive ability of random forest can also be seen in this case. MAPE values ranging from 0.34 % - 1.1 %. This indicates that the predictive ability of the random

forest algorithm is quite good when compared to using a dynamic multiphase flow simulator.

As can be seen in Table 11. the value of MAPE not linear to the increasing value of water fraction. This condition can be mean that change of water fraction is not corellated linearly to feature obtained from calculation which used as an input of machine learning. Regardless of that, this condition still acceptable because error obtain in prediction is below 20%.

Table 11
MAPE of predicted wax deposition rate

Case	fw (fraction)	Q (m ³ /d)	Tin (°C)	MAPE (%)
1	0	1891	70	1.1
	0.2	1891	70	0.73
	0.4	1891	70	0.82
	0.6	1891	70	1.32
	0.8	1891	70	0.98
2	0	1891	70	1.1
	0	2500	70	0.43
	0	5000	70	0.34

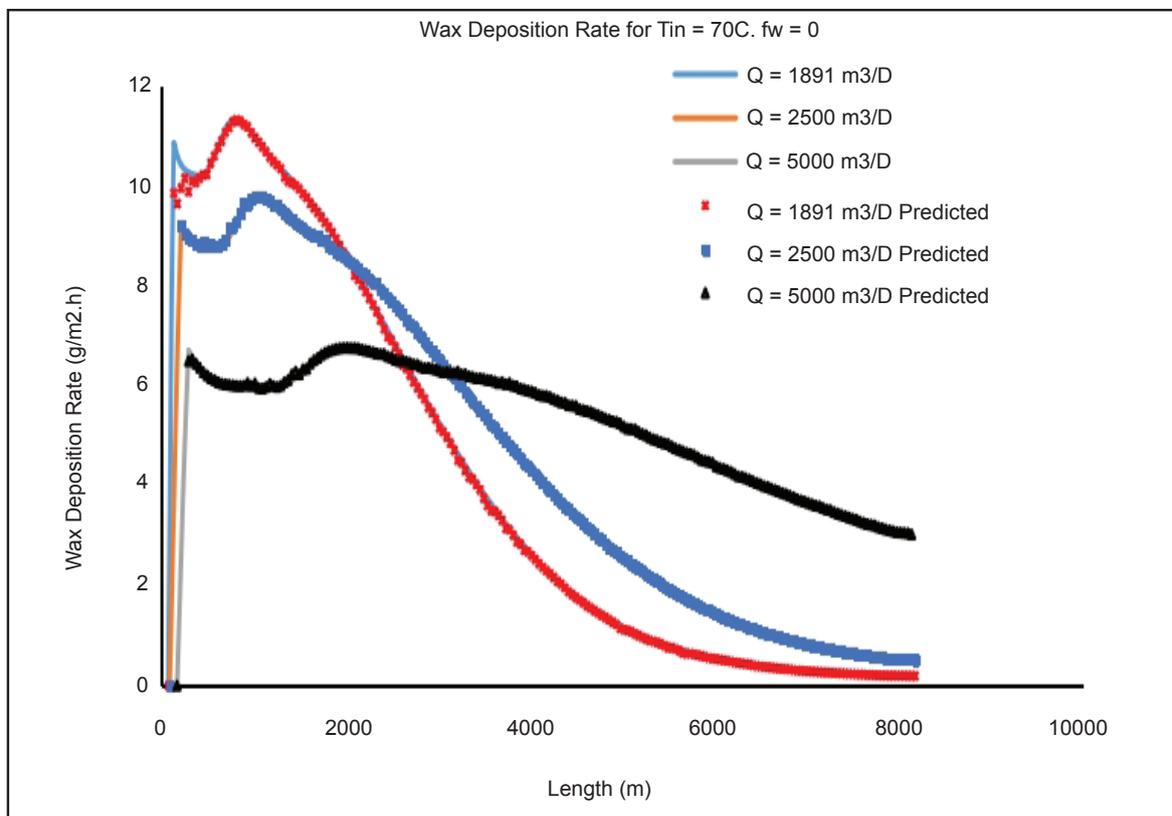


Figure 12
Comparisons of wax deposition ratio between dynamic multiphase flow and best machine learning (Case 2).

CONCLUSIONS

Machine learning has been made to predict wax deposition rate in this paper. Two types of scenario which are input A and input B have been made and compared each other. input A consists of water volume fraction (fw), shear stress (τ_w), effective viscosity (μ_e), wax concentration gradient (dC/dT), and temperature gradient (dT/dR) and input B consists of water volume fraction (fw), shear stress (τ_w), effective viscosity (μ_e), wax concentration gradient (dC/dT), temperature gradient (dT/dR), shear stripping variable (SV) dan diffusion variable (DV). Based on the data obtained from the results of simulations using commercial software and recent article, it is known that the use of random forests with $N_{tree} = 500$ produces the best predictive ability based on the error values obtained for both types of input scenarios. Using input scenario A It has R-squared (R^2) for training, testing and total data are 0.999, 0.992 and 0.9975, respectively. The strategy to include additional features is also known to improve predictability for best parameter random forests even on slightly values. Using the same parameter of best random forest for input B achieves R-squared (R^2) for training, testing and total data are 0.999, 0.993 and 0.9977, respectively. Beside that prediction using random forest while compare to the result from dynamic multiphase flow showed good prediction. It can be seen on the value of MAPE obtained is under 20%. Establishment of machine learning for next study can be made by varying input of simulation and apply another feature which has better correlation with target so it can have wider range of data that can be predict and also increase accuracy of predicted data.

REFERENCES

- Agustin, M. & Prahasto, T.**, 2012. Penggunaan Jaringan Syaraf Tiruan Backpropagation Untuk Seleksi Penerimaan Mahasiswa Baru Pada Jurusan Teknik Komputer Di Politeknik Negeri Sriwijaya. *Jurnal Sistem Informasi Bisnis*, 2(2), pp. 089-097.
- Azevedo, L. F. A. & Teixeira, A. M.**, 2003. A Critical Review of the Modeling of Wax Deposition. *Journal Petroleum Science & Technology*, 21(3), pp. 393-408.
- Burger, E. D., Perkins, T. K. & Striegler, J. H.**, 1981. Studies of Wax Deposition in the Trans Alaska Pipeline. *Journal Of Petroleum Technology*, 33(06), p. 1075–1086.
- Chi, Y., Daraboina, N. & Sarica, C.**, 2017. Effect of the Flow Field on the Wax Deposition and Performance of Wax Inhibitors: Cold Finger and Flow Loop Testing. *Energy Fuels*, 31(5), pp. 4915-4924.
- Chu, Z.-Q., Sasanipour, J., Saeedi, M., Baghban, A., & Mansoori, H.**, 2017. Modeling of Wax Deposition Produced in The Pipelines Using PSO-ANFIS Approach. *Journal Petroleum Science and Technology*, 35(20), pp. 1974-1981.
- Fan, K., Huang, Q., Li, S. & Yu, W.**, 2015. Wax Deposition Rate of Water-in-Crude Oil Emulsions Based on Laboratory Flow Loop Experiment. *Journal of Dispersion Science and Technology*, 38(1), pp. 1-53.
- Giacchetta, G., Marchetti, B., Leporini, M., Terenzi, A., Dall'Acqua, D., Capece, L., & Grifoni, R.C.**, 2019. Pipeline Wax Deposition Modeling: A Sensitivity Study on Two Commercial Software. *Petroleum*, 5(2), pp. 206-213.
- Huang, Q., Li, Y. & Zhang, J.**, 2008. Unified wax deposition model. Project: Physical properties of wax deposits on the walls of crude pipelines, Volume 29, pp. 459-462.
- Huang, Q. & Ma, J.**, 2008. Prediction of Wax Deposition of Crude Using Statistical and Neural Network Methods. Calgary, Alberta, Canada, International Pipeline Conference.
- Hu, Z., Meng, D., Liu, Y., Dai, Z., Jiang, N., & Zhuang, Z.**, 2019. Study of Wax Deposition Law by Cold Finger Device. *Journal Petroleum Science and Technology*, 37(15), pp. 1846-1853.
- Joshi, R. R.**, 2017. Modeling and Simulation of Wax Deposition in Crude Oil Pipeline. *International Journal of Engineering Research and Technology*, Volume 10, pp. 297-300.
- Kamari, A., Mohammadi, A. H., Bahadori, A. & Zendeheboudid, S.**, 2014. A Reliable Model for Estimating the Wax Deposition Rate During Crude Oil Production and Processing. *Journal Petroleum Science and Technology*, Volume 32, pp. 2837-2844.
- Matzain, A., Apte, M.S., Zhang, H.-Q., Volk, M., Redus, C.L., Brill, J.P., & Creek, J.L.**, 2000. Multiphase Flow Wax Deposition Modelling. Banff, Canada, ETCE, pp. 415-441.
- Montero, F.**, 2020. Wax Deposition Analysis for Oil and Gas Multiphase Flow in Pipelines. Copenhagen, Denmark: Aalborg University.
- Mulyahati, I. L.**, 2020. Implementasi Machine Learning Prediksi Harga Sewa Apartemen Menggunakan Algoritma Random Forest Melalui Framework Website Flask Python. Yogyakarta: Universitas Islam Indonesia.
- Obaseki, M. & Elijah, P. T.**, 2020. Dynamic Modelling and Prediction of Wax Deposition Thickness in Crude Oil Pipelines. *Journal of King Saud University - Engineering Sciences*, 33(6), pp. 437-445.

- Octaviani, P. A., Wilandari, Y. & Ispriyanti, D.**, 2014. Penerapan Metode Klasifikasi Support Vector Machine (SVM) Pada Data Akreditasi Sekolah Dasar (SD) Di Kabupaten Magelang. *Jurnal Gaussian*, 3(4), pp. 811-820.
- Pedersen, K. S., Skovborg, P. & Roenningsen, H. P.**, 1991. Wax precipitation from north sea crude oils-Thermodynamic modeling. *Energy Fuels*, 5(6), p. 924-932.
- Singh, P., Venkatesan, R., Fogler, H. S. & Nagarajan, N.**, 2000. Formation and aging of incipient thin film wax-oil gels. *AIChE Journal*, 46(5), pp. 1059-1074.
- Sousa, A. M., Matos, H. A. & Guerreiro, L.**, 2019. Wax Deposition Mechanisms and The Effect of Emulsions and Carbon Dioxide. *Journal Petroleum*, 6(3), pp. 216-226.
- Sutton, C. D.**, 2004. Classification and Regression Trees, Bagging, and Boosting. *Handbook of Statistics*, Volume 24, pp. 303-329.
- Wang, W. & Huang, Q.**, 2014. Prediction for Wax Deposition in Pipeline Validated by field Pigging. *Journal of the Energy Institute*, 87(3), pp. 196-207.
- Wei, Y., Wu, M. & Zhao, L.**, 2010. Wax Deposition Rate Model for Crude Oil Pipeline. Yantai, China, IEEE.
- Xie, Y. & Xing, Y.**, 2017. A Prediction Method for the Wax Deposition Rate Based on a Radial Basis Function Neural Network. *Petroleum*, 3(2), pp. 237-241.
- Zheng, S.**, 2017. Wax Deposition from Single-Phase Oil Flows and Water-Oil Two-Phase Flows in Oil Transportation Pipelines. Ann Arbor, Michigan: The University of Michigan.
- Zhou, Y., Gong, J. & Wang, P.**, 2015. Modeling of Wax Deposition for Water-In-Oil Dispersed Flow. *Asia-Pacific Journal Chemical Engineering*, 11(1), pp. 108-117.