COMPARATIVE ANALYSIS ON DIFFERENT MODELLING TECHNIQUES OF C5 TOP COMPOSITION FOR NAPHTHA STABILIZER COLUMN

Lee Mei Yan[1], Ramasamy Marappa Gounder[2]
Chemical Engineering Department, Universiti Teknologi PETRONAS,
32610 Seri Iskandar, Perak, Malaysia

Abstract:* Product quality monitoring is an important element in industrial process control. In reality, most product qualities are difficult to be measured online due to technical or economic restrictions. Empirical inferential model is an effective solution to provide real-time analysis on product quality. Thus, this project analyses different modelling techniques for C5 top composition of naphtha stabilizer column using historical plant data. Seven statistical and machine learning techniques are considered, which include multiple linear regression (MLR), stepwise linear regression (SLR), principle component regression (PCR), partial least squares regression (PLSR), regression tree (RT), gradient boosting regression tree (GBRT) and artificial neural networks (ANN). The aims of this project are to develop an inferential model of C5 top composition using the proposed modelling techniques, to compare the model prediction accuracy and to evaluate the limitations of each modelling technique for industrial application. The developed models are then assessed based on prediction accuracy and model development ease. The order of model performance obtained from this study is ANN (tangent sigmoid) > SLR > MLR > ANN (linear) > GBRT > PLSR > RT > PCR. Meanwhile, the assessment based on model prediction accuracy and development ease reveals that SLR is the most recommended modelling technique for industrial application due to its reliable predictive performance and simplicity in developing a model.

***Keywords:* Empirical inferential model; Soft sensor; Statistical modelling; Machine learning; Distillation;**

I. INTRODUCTION

In chemical process industries, the top and bottom products of a distillation column usually have a very tight specification as stated in Sales and Purchase Agreement (SPA). Violation of SPA leads to monetary losses. Though off-specification products can be blended with over-purified products to meet the requirement, production of over-purified products is a waste of energy. An ideal distillation process is to maintain product composition at the required specification while minimizing energy consumption. Therefore, continuous quality monitoring is important. However, measuring product quality online is difficult and requires high capital cost. A solution to the problem is by implementing empirical inferential model.

An empirical inferential model, also defined as a data-driven soft sensor provides real-time estimation of product quality using other online available measurements. Several attractive properties of empirical inferential models over conventional instrumentations are as follows [1]:

i. Provide process insight through capturing the information concealed in data.
ii. Allow real-time estimation of product quality.
iii. Improve productivity and business profitability by reducing production cost due to off-spec products.
iv. Easy implementation on existing hardware.
v. Require little or no capital costs for installation, management of the required infrastructure and commissioning

Numerous modelling techniques are used to construct inferential model in process control applications. Generally, there are three types of inferential model which are first principle model, empirical model and semi-empirical model. The development of first principle model requires knowledge on chemistry and physics of the process. This type of model may not be practical for complex processes because the model needs a vast number of equations, process variables and unknown parameters, for example, the chemical and physical properties of a mixture [2]. Pearson [3] analysed that first principle models perform excellently in approximation accuracy and physical interpretation compared to empirical and semi-empirical models. Nevertheless, first principle models are bad in terms of controllability and ease of development.

Empirical models are categorized into two, which are linear and nonlinear models. Nonlinear inferential models are suitable for inferring quality of chemical processes that exhibit nonlinear behaviour. Artificial neural networks (ANN) are very popular for complex processes with large input and output data. For distillation process, an inferential model is developed to estimate the top or bottom composition, or both. The product composition is being inferred from other online measured process variables such as tray temperatures, column pressure, feed flow, reflux flow and etc. During model development, information carried in the datasets is important because the goodness of data dictates the predictive performance of an inferential model. In reality, data collected from industrial processes is associated with problems such as sampling time, missing data, outliers, operating conditions and accuracy [4]. Hence, it is crucial to pre-process raw data in advance. The general procedures of building an empirical inferential model are first collecting and pre-processing data, then, selecting significant input variables followed by training a model and validate it subsequently, lastly, maintaining the model at good performance.

The implementation of inferential models for monitoring product quality has gained momentum in process industries due to its superiority of online estimation of product quality with little requirement of a priori knowledge on the process. A broad variety of statistical modelling and machine learning techniques has been employed. The widespread modelling techniques are principal component regression (PCR), partial least squares regression (PLSR), support vector machine (SVM) and artificial neural networks (ANN).

One of the challenging issues for effective inferential models is correlation between process variables. Latent variable regressions (LVR) like PCR and PLSR are designed to deal with collinearity through projection of principal components or latent variables orthogonally while describing maximum variance of original process data. In the past decades, several research papers on empirical inferential model using PCR and PLSR have been reported. Mejdell and Skogestad [5] studied the prediction performance of steady-state PCR and PLS model for multicomponent distillation column using multiple temperature measurements. Both PCR and PLS models estimated well despite multicomponent mixtures, nonlinearity and pressure variations. Kano, et al. [6] employed dynamic PLS regression to investigate the inferential control system of distillation compositions. The dynamic PLS model performed excellently in cascade control system of product composition and the performance was better than usual tray temperature control. Zhang [7] reported that the inferential feedback control performance of a distillation composition was enhanced by using PCR and PLS models. A modified PLS model that integrated a bias update scheme and advanced cross-validation method was proposed by Kim, et al. [8] for inferential quality control. The proposed PLS model was proven to be more robust without updating the model parameters frequently. Lin, et al. [9] proposed a systematic approach that detected outliers by a univariate approach, followed by PCA for developing robust PCR and dynamic PLS inferential models. The prediction accuracy was reasonably accurate for both PCR and dynamic PLS models via the proposed method. A comparative analysis of PCR, PLSR, regularized canonical correlation analysis (RCCA), ridge regression (RR) and ordinary least squares (OLS) regression was carried out by Madakyaru, et al. [10] using synthetic data and simulated distillation column data. Latent variable regression techniques, which were PCR, PLSR and RCCA performed better than OLS and RR techniques because LVR techniques reduced the noise effect on model prediction by discarding latent variables or principal components with small eigenvalues. Nevertheless, PCR and PLSR are linear modelling methods which cannot function excellently for nonlinear processes. Thus, Jin, et al. [11] designed an adaptive inferential model based on just-in-time (JIT) learning and kernel PLSR (KPLSR) for nonlinear multiphase batch processes and compared with single-model, multi-model, Bayesian model averaging based multi-model and JIT learning based soft sensors. The paper revealed that MJIT-KPLS provided an outstanding performance whilst conventional single model PLS model performed poorly. This signified that nonlinear approaches such as ANN, SVM, KPLS and neuro-fuzzy system are preferable for nonlinear processes.

With the advent of digital computers, proliferation of research on empirical inferential models using machine learning techniques has been observed in recent years. Typical machine learning techniques used for model development are SVM and ANN. ANN consists of a huge class of model structures. Multilayer perceptron (MLP) and radial basis function networks (RBFN) are the common types. Singh, et al. [12] designed an artificial neural networks-based estimator for distillate composition. It was reported that ANN model predicted distillate composition excellently with simulation result. A comparison between backpropagation neural network (BPNN), generalized regression neural network (GRNN) and RBFN was analysed by Pani, et al. [13] for cement clinker quality estimation. It was found that RBFN had the best estimation capabilities. Desai, et al. [14] compared the predictive performance of support vector regression (SVR), MLP and RBF neural networks models. The results indicated that SVR was an attractive alternative to ANN based inferential models. Jain, et al. [15] developed an inferential model for batch distillation column using SVR. The SVR-based model described the process accurately for varying reflux ratio. Xuefeng [16] proposed a hybrid artificial neural network (HANN) integrating back propagation algorithm (BP) with PLSR model to predict the product concentration in p-xylene oxidation reaction. This HANN model was more robust and overfitted less compared to typical artificial neural network models. Several linear and nonlinear inferential models using multiple linear regression (MLR) and ANN were developed by Rogina, et al. [17] for light naphtha RVP estimation of a crude distillation unit. Linear models that developed by MLR and ANN predicted poorly whilst the prediction accuracy of nonlinear models using MLP and RBF neural networks were within acceptable range. Rani, et al. [18] presented the application of Levenberg-Marquardt (LM) and adaptive linear network (ADALINE) based inferential models of a multicomponent distillation process. The model prediction accuracy, training time and memory space of ADALINE were better than LM model, and dynamic inferential control scheme was more robust and efficient than static inferential control scheme. An MLP model that incorporated least absolute shrinkage and selection operator (LASSO) was proposed by Sun, et al. [19]. The designed model had better predictive performance than usual MLP and other neural network-based models with input variable selection algorithms. Pani, et al. [20] employed MLR, PCR and back propagation neural networks (BPNN) to develop inferential models of a debutaniser column. It was found that BPNN with LM algorithm performed better than MLR and PCR models. Urhan, et al. [21] analysed 16 statistical learning techniques for inferential models of a crude distillation unit. PLS modelling was found to have a relatively low prediction efficiency.

Machine learning algorithms like classification and regression tree (CART) and boosted regression tree (BRT) are gaining popularity in prediction fields such as traffic prediction, ecology, soil science and pharmaceutical industries due to its ability to predict nonlinear and complex data. Elith, et al. [22] demonstrated the use of BRT in an ecology study and discussed the advantages. It was discussed that BRT can estimate the relative importance of predictor variables, which made this method differs from other "black box" machine learning algorithm. Zhang and Haghani [23] employed gradient boosted regression tree (GBRT) method to predict free-way travel time. They concluded that gradient boosted tree had superior prediction performance and it can handle sharp discontinuities. Ding, et al. [24] investigated the influences of built environment on driving distance using gradient boosting decision tree. Keskin, et al. [25] assessed the predictive performance of CART, bagged regression tree, BRT, random forest, SVM, PLSR, regression kriging and ordinary kriging on soil carbon fractions. The results revealed that random forest performed the best, followed by SVM and BRT while PLSR performed better than CART. Deconinck, et al. [26] employed stepwise MLR, PLSR, BRT and multivariate adaptive regression splines (MARS) to estimate blood-brain barrier passage. The results showed that PLS-MARS yielded the lowest prediction error while PLS performed slightly better than BRT and stepwise MLR.

In light of all the literature review, machine learning techniques like regression tree and gradient boosting regression tree are yet to be explored in the development of empirical inferential model for chemical processing industry. Therefore, this paper aims to discover the application of regression tree and gradient boosting regression tree in a refinery plant using historical plant data. Besides that, model prediction accuracy and limitations of different statistical and machine learning techniques are analysed and compared. Seven modelling techniques, that are multiple linear regression (MLR), stepwise linear regression (SLR), principle component regression (PCR), partial least squares regression (PLSR), regression tree (RT), gradient boosting regression tree (GBRT) and artificial neural networks (ANN) are investigated in this paper. Section II explains briefly on the process description and data of naphtha stabilizer column. Section III and IV describe the procedures of data pre-processing and model development, respectively. Section V presents and discusses the performance of each modelling technique. Lastly, section VI summarizes the findings of this study.

II. PROCESS DESCRIPTION AND DATA

A. *Process Description*

Crude oil refinery is a process plant where crude oil is refined into useful petroleum products, for instance, liquid petroleum gas (LPG), gasoline, diesel fuel, kerosene, asphalt base and heating oil. Crude distillation unit (CDU) is one of the main process units in a refinery plant. The products of CDU consist of LPG, naphtha, kerosene, diesel, atmospheric gas oil and waxy residue. Naphtha stabilizer column falls under CDU which serves as a purpose of removing C4 and lighter hydrocarbons from naphtha. Naphtha stabilizer is a two-cut distillation column with LPG as top product and unstabilized naphtha as bottom product. C5 and heavier hydrocarbons are impurities at top. The unstablized naphtha is then routed to a naphtha splitter to separate light and heavy naphtha. A simplified process flow diagram of naphtha stabilizer is shown in Fig. 1. Naphtha feed is preheated by the bottom stream of naphtha stabilizer while the heat supplied to reboiler of the column is by hot diesel pump-around from upstream crude tower.

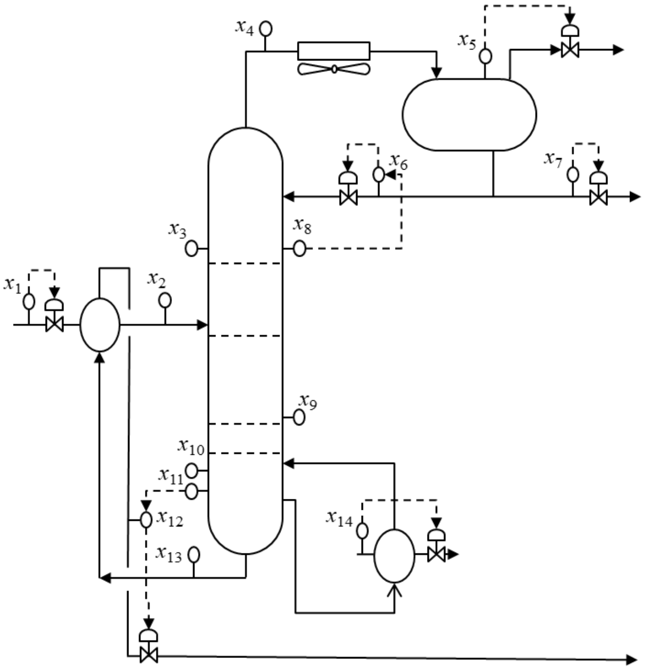

Fig. 1. Naphtha stabilizer column configuration.

TABLE I. DESCRIPTION OF PROCESS VARIABLES.

Tag	Description	Unit
x_1	Naphtha feed flow	kg/h
x_2	Naphtha feed temperature	°C
x_3	Tray 7 temperature	°C
x_4	Overhead temperature	°C
x_5	Condenser pressure	bar
x_6	Reflux flow	kg/h
x_7	Net overhead liquid flow	kg/h
x_8	Tray 7 temperature	°C
x_9	Tray 33 temperature	°C
x_{10}	Bottom pressure	bar
x_{11}	Column level	%
x_{12}	Bottom flow	kg/h
x_{13}	Bottom temperature	°C

Tag	Description	Unit
x_{14}	Diesel pump-around	kg/h
x_{15}	CDU load	kbpd

B. *Process Data*

Steady-state historical data of a naphtha stabilizer column from a refinery plant was collected. Process variables that labelled in Fig. 1 were considered and the description of each process variable was tabulated in Table I. An additional process variable x_{15} which represented the CDU load was taken into account as well. The average value of x_3 and x_8 was considered as both transmitters are measuring same tray temperature. All process variables were averaged value of one hour before and one hour after sampling time to consider factors of process lag time and actual site sampling time. The response variable (y) of this process is C5 top composition in volume percentage. The measured C5 top composition was attained from laboratory analyser. The collected raw data contained 1056 observations with 14 process variables and one response variable.

III. DATA PRE-PROCESSING

Observations that contained missing values were first removed. Then, observations that contained zero values were assessed by comparing with neighbouring observations. Observations with unreasonable zero values were removed. Thereafter, the first-cleaned data was shuffled randomly and segregated into training data and validation data. 60% of the data was labelled as training data while the remaining 40% was labelled as validation data. Outliers in training data were detected using 3σ-rule. Process variables that were outside the bounds of $\mu \pm 3\sigma$ were considered as outliers, where μ is the mean and σ is the standard deviation. Observations with outliers were removed. After pre-processing, training data contained 536 observations while validation data contained 381 observations.

IV. MODEL DEVELOPMENT

Seven modelling techniques, including multiple linear regression (MLR), stepwise linear regression (SLR), principle component regression (PCR), partial least squares regression (PLSR), regression tree (RT), gradient boosting regression tree (GBRT) and artificial neural networks (ANN) were employed for model development. MATLAB2018b software was implemented to train model with MLR, SLR, RT and ANN techniques whilst Python 3.7 was utilized for GBRT technique. Meanwhile, SIMCA-P software was used to train both PCR and PLSR models.

The prediction of trained model was tested using validation data and the prediction accuracy was evaluated by computing root mean squared error (RMSE) as shown in (1).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(y_m(i)-y_p(i)\right)^2}{N}} \qquad (1)$$

where y_m is measured y value, y_p is predicted y value, i is index number of observation and N is total number of observations.

A plot of actual against the predicted values from each modelling technique was constructed and the goodness of fit, R-squared value was calculated to assess the correlation. Besides that, line graphs of actual and predicted values were plotted simultaneously to examine the variation. The trained model was accepted when the model prediction accuracy was satisfactory, else, the modelling parameter(s) was (were) adjusted and model was retrained.

A. *Multiple Linear Regression (MLR)*

All input variables were selected for model training during the first trial. Process variables that were insignificant to the model based on F-test evaluation (p-value > 0.05) were removed individually. Subsequently, the prediction accuracy of trained model was tested by using validation data. Variable removal and validation step were repeated until a satisfactory model prediction was attained.

B. *Stepwise Linear Regression (SLR)*

Both constant and linear starting model types were examined and three model specifications describing the largest set of terms in fit, i.e. linear, interactions (bivariate term) and quadratic (squared term) were investigated via MATLAB2018b. The criterion for adding and trimming a model was p-value for an F-test of the change in sum of squared error. The p-value requirement for adding and removing a term into or from the model were 0.05 and 0.1, respectively. The procedures of SLR were summarized as below:

1. Constant or linear initial model was fitted.
2. Term that not in the model and had the smallest p-value (p-value < 0.05) was added to the model and this step was repeated. Otherwise, step 3 was carried out.
3. Term that in the model and had the largest p-value (p-value > 0.1) was removed and step 2 was repeated. Otherwise, end.

C. *Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR)*

SIMCA-P software was applied to train both PCR and PLSR models. The clean data was imported, and all input variables were selected as X variables while C5 composition was set as Y variable. The optimum number of principal components for PCR and latent variables for PLSR were determined through 7-fold cross validation method by default.

D. *Regression Tree (RT)*

Standard CART algorithm was applied to grow a regression tree in MATLAB2018b with the following steps:

1. All input data and possible binary splits on every predictor were examined.
2. A split that produced the least mean squared error (MSE), subjected to the minimum number of leaf observations constraint was selected.

3. The split was imposed, and two child nodes were formed.
4. Steps 1 to 3 were repeated for the two child nodes.
5. Steps 1 to 4 were stopped when one of the stopping criteria was fulfilled. The criteria were:
   - The impurity of a node is less than the threshold 10^{-6}, meaning the MSE in a node was lower than the MSE of the entire data computed multiplied by 10^{-6}.
   - The child nodes contained observations fewer than the specified minimum leaf size.
   - Maximal number of decision splits (number of observations – 1) was met

The minimum leaf size was determined by computing the regression error of a tree through 10-fold cross validation. The leaf size that yielded the least error was selected for regression tree model training.

TABLE II. BOOSTING PARAMETERS FOR GBRT MODEL TRAINING.

Parameter	Symbol/Abbreviation	Initial value
Number of iterations	M	100
Minimum leaf observations to split	min_split	2
Minimum observations in a terminal node	min_node	1
Maximum depth of a tree	max_depth	3
Learning rate	λ	0.1

E. *Gradient Boosting Regression Tree (GBRT)*

GBRT model was trained and validated by using Python 3.7 software. The loss function for GBRT model training was least square regression. The boosting parameters for model training were tuned using 10-fold cross validation method. The parameters considered were tabulated in Table II together with respective initial values. Parameter M was first tuned, followed by max_depth, min_split and min_node. Lastly, parameter λ was tuned in proportional to parameter M. The node splitting criteria were according to the specified min_split, min_node and max_depth for each iteration. Also, the node impurity threshold was set at 10^{-7} by default. The general procedures of GBRT were as follows:

1. An optimal constant model was initialized.
2. Negative gradient of loss function was calculated for each observation.
3. A regression tree was fitted to the negative gradients.
4. A gradient descent step size was calculated to prevent over-stepping and missing the local minima.
5. The regression tree was multiplied with the calculated step size and a learning rate parameter, λ.
6. The regression tree was then added to the previous models.
7. Steps 2 to 6 were repeated until the number of iterations, M was met.

F. *Artificial Neural Network (ANN)*

Feedforward ANN model was trained by applying Levenberg-Marquardt (LM) backpropagation algorithm in MATLAB2018b. A three-layer-feedforward network with single input layer, single hidden layer and single linear output layer was developed. The pre-processed training data was distributed into 70% training, 25% validation and 5% testing for ANN model training. Validation set was used to stop training when any one of the following conditions was achieved:

- The maximum number of epochs (repetitions), 1000 was reached.
- Performance gradient was less than minimum performance gradient, 1 x 10^{-7}.
- The maximum number of validation checks (number of successive iterations that the validation performance fails to decrease), 6 was reached.
- Performance was minimized to a specified goal, 0.

Both tangent sigmoid and linear activation functions on hidden layer were considered and compared. The number of neurons in hidden layer was determined by evaluating the RMSE of the trained network using the pre-processed training data.

V. RESULTS AND DISCUSSION

A. *MLR*

The selection of process variables for MLR model are based on F-test evaluation. Table III tabulates the prediction accuracy of MLR models with different process variables being removed. It is seen from that MLR model 6 that excludes x_9 (Tray 33 temperature) and x_{15} (CDU load) has the highest adjusted R^2 (0.681) with lowest RMSE (0.1879) and highest R^2 (0.60979) on validation data. Fig.2 shows the C5 composition line plot of actual and predicted values by MLR model 6. It is observed that the prediction line does not coincide with the actual line plot especially on the extreme values of actual composition. Fig. 3 illustrates the predicted values of MLR model 6 against the actual C5 composition. It is apparent that the plots are scattered and the best fit line deviates from 45 degree. This indicates that the MLR model prediction accuracy is unsatisfactory.

TABLE III. MLR MODEL PREDICTION ACCURACY WITH DIFFERENT NUMBER OF VARIABLES.

No.	Removed variables	Training		Validation	
		R^2	*Adjusted-R^2*	***RMSE***	R^2
1	None (benchmark)	0.678	0.669	0.1885	0.60843
2	x_{15}	0.689	0.681	0.1887	0.60664
3	x_9	0.681	0.673	0.1883	0.60843
4	x_{11}	0.682	0.674	0.1893	0.604
5	x_{10}	0.681	0.673	0.1887	0.60666
6	$x_9 + x_{15}$	0.688	0.681	0.1879	0.60979
7	$x_{11} + x_{15}$	0.688	0.681	0.1889	0.60557
8	$x_{10} + x_{15}$	0.687	0.68	0.1883	0.60816
9	$x_9 + x_{10}$	0.681	0.673	0.1884	0.60771

No.	Removed variables	Training		Validation	
		R^2	Adjusted-R^2	*RMSE*	R^2
10	$x_{11} + x_{15}$	0.681	0.674	0.1885	0.60752
11	$x_9 + x_{10} + x_{15}$	0.687	0.681	0.1881	0.60906
12	$x_9 + x_{10} + x_{11} + x_{15}$	0.687	0.681	0.1883	0.60833

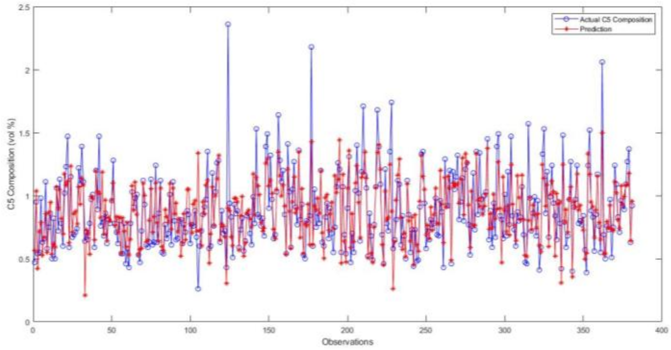

Fig. 2. Line plot of actual and predicted C5 composition by MLR model 6.

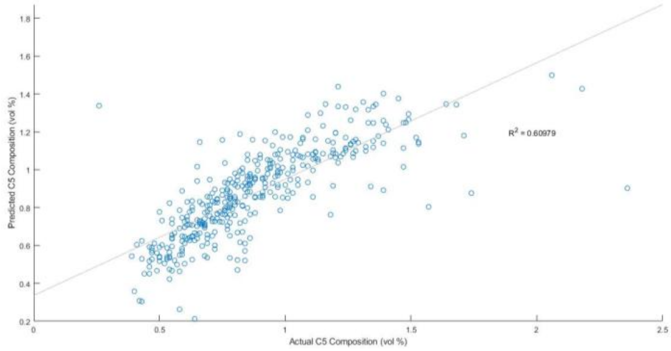

Fig. 3. Scatter plot of actual and predicted C5 composition by MLR model 6.

B. SLR

Stepwise linear regression in MATLAB 2018b is employed to investigate model prediction accuracy in consideration of linear, bivariate and squared terms. Table IV summarizes the SLR model prediction accuracy with different number and type of terms present in the model. It is observed that SLR model 5 has the highest adjusted R^2 (0.763) with lowest RMSE (0.1876) and highest R^2 (0.61509) on validation data. In fact, SLR model 5 is the most complicated model that contains 26 combination of linear, bivariate and squared terms. Yet, overfitting issue has not been aroused. This implies that the complexity of the model fairly describes the nonlinear behaviour of distillation. Fig. 4 portrays the line plot of actual C5 composition and predicted values from SLR model 5. It is seen that some of the predictions do not fit with the actual values. Fig. 5 shows the predicted values against actual C5 compositions by SLR model 5. The scattered plots and a non-diagonal best fit line are indications of bad prediction. Thus, this signifies that SLR model prediction accuracy is poor.

TABLE IV. SLR MODEL PREDICTION ACCURACY.

No.	Number of terms	Training		Validation	
		R^2	Adjusted-R^2	*RMSE*	R^2
1	20 linear and bivariate terms	0.759	0.75	0.1887	0.61047
2	11 linear terms	0.679	0.673	0.1882	0.60852
3	12 linear terms	0.681	0.674	0.1882	0.60869
4	26 linear and bivariate terms	0.77	0.759	0.1912	0.6001
5	26 linear, bivariate and squared terms	0.774	0.763	0.1876	0.61509

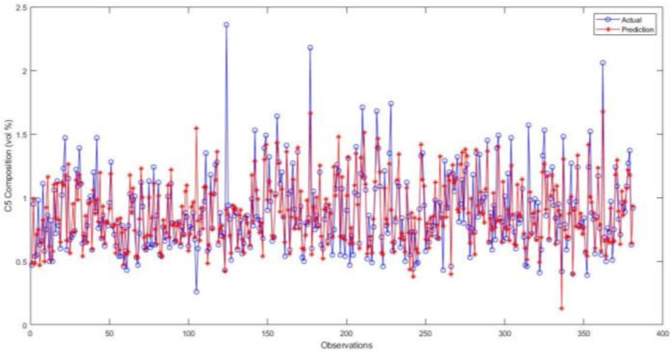

Fig. 4. Line plot of actual and predicted C5 composition by SLR model 5.

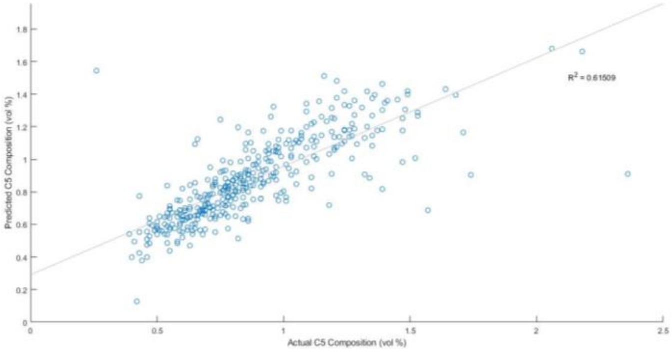

Fig. 5. Actual plot of actual and predicted C5 composition by SLR model 5.

C. PCR

The efficacy of PCR on model prediction accuracy is analysed. Three principal components are selected with 68.7% of variance in input variables being explained. Fig. 6 shows the variations of actual and predicted C5 composition by PCR model. It is observed that variation of predicted C5 composition is minimal compared to the variation of actual composition. This implies that predictions based on this model is not reliable. A predicted versus actual C5 composition scatter graph is plotted in Fig.7. It reveals that that plots are scattered randomly, and the best fit line with a R2 value of 0.30372 deviates a lot from the 45-degree line. The calculated RMSE is 0.2511. It is deduced that PCR model has low descriptive and predictive power.

D. PLSR

Model prediction accuracy of PLSR modelling technique is examined. The optimum number of latent variables obtained is four. These four latent variables explain 64.1% variance of response variable. Fig. 8 illustrates the deviation of actual and predicted C5 composition by PLSR model. It is observed that the predictions do not fit with the actual values perfectly. The calculated RMSE is 0.19504. Fig. 9 depicts the actual C5 composition versus predicted values by PLSR model. It is learnt that the predictivity of PLSR model is poor because the plots are scattered, and the best fit line with a R2 value of 0.57956 deviates from 45-degree line.

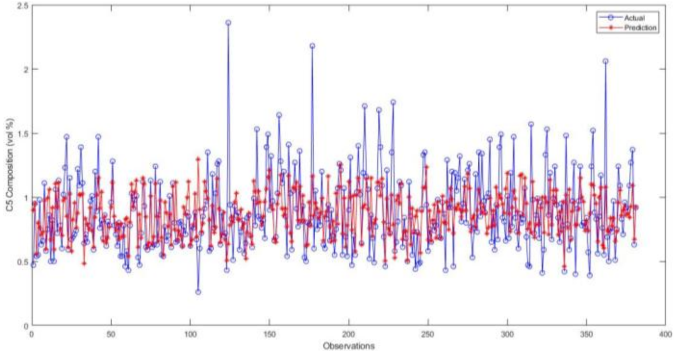

Fig. 6. Line plot of actual and predicted C5 composition by PCR model.

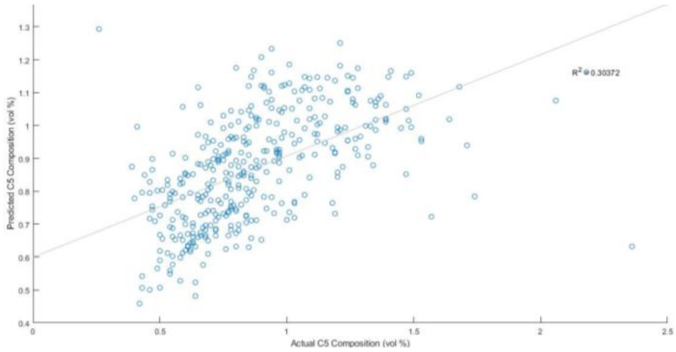

Fig. 7. Scatter plot of actual and predicted C5 composition by PCR model.

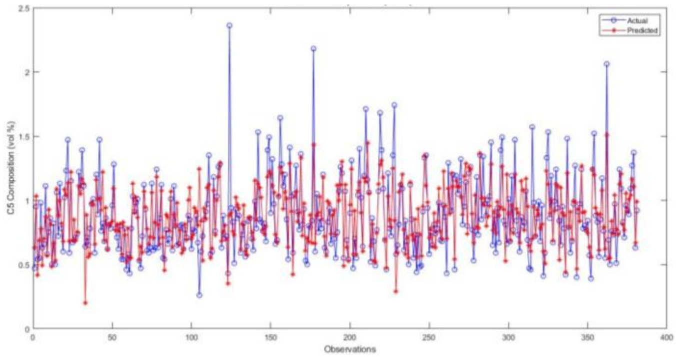

Fig. 8. Line plot of actual and predicted C5 composition by PLSR model.

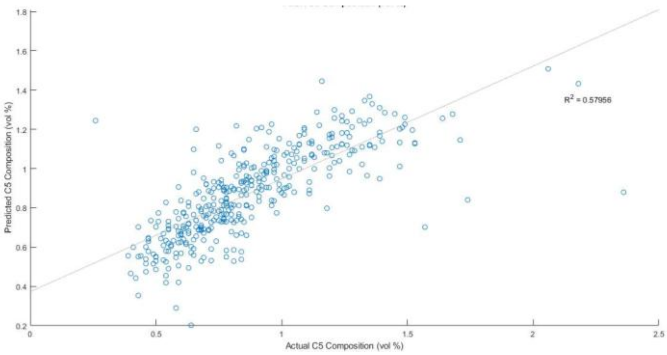

Fig. 9. Scatter plot of actual and predicted C5 composition by PLSR model.

E. RT

In order to validate the attractive properties of RT as mentioned by Elith, et al. [22], particularly, insensitivity towards outliers and ability to deal with missing data using surrogates, model prediction accuracy of RT using both raw and clean data are analysed. Firstly, the sensitivity of RT towards outliers is tested. Through cross validation approach, the minimum leaf size chose for RT model training is seven as shown in Fig. 10. The RMSE and R2 value of RT model with and without outliers are tabulated in Table V. Table V reveals that the predictive performance of RT changes insignificantly with 3.09 % change of RMSE and 0.12% change of R2. This concludes that RT is insensitive to outliers.

The ability of dealing with missing data through surrogate splitting is investigated. Table VI shows the RMSE and R2 value of RT model with and without missing data. It is seen that RT model with missing data and surrogate splitting can predict response variable fairly similar to RT model without missing data. In fact, the predictivity of RT model with missing data and surrogate splitting is better. The percent differences of RMSE and R2 value are merely 1.67% and 1.62%, respectively. Therefore, this result reveals that RT has an advantage over other typical statistical modelling techniques that it can predict the response variable with missing values on unseen data.

For comparative analysis in the latter section (section H), RT model is trained and validated using clean data. Fig. 11 illustrates the variations of both actual and predicted C5 compositions by RT model. It is observed that the predictions fail to comply with the actual C5 compositions impeccably. Fig. 12 demonstrates the actual versus predicted C5 composition by RT model. It is apparent that the prediction accuracy is poor as the plots are scattered and the R2 value is just 0.43893 with a calculated RMSE of 0.2297. In short, the predictivity of this RT model is inadequate for estimating product quality in industry.

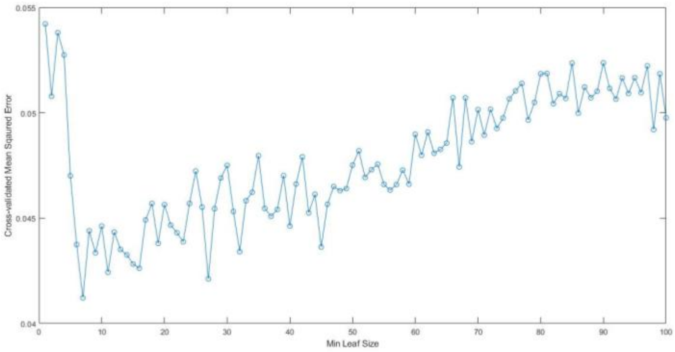

Fig. 10. Cross validated mean squared error against minimum leaf size.

TABLE V. PREDICTIVE PERFORMANCE OF RT WITH AND WITHOUT OUTLIERS.

Assessment	RT Model		Percent difference (%)
	With outliers	*Without outliers*	
RMSE	0.2368	0.2297	3.09
R^2	0.43840	0.43893	0.12

TABLE VI. PREDICTIVE PERFORMANCE OF RT WITH AND WITHOUT MISSING DATA.

Assessment	RT Model		Percent difference (%)
	With missing data	*Without missing data*	
RMSE	0.2242	0.2280	1.67
R^2	0.58972	0.58033	1.62

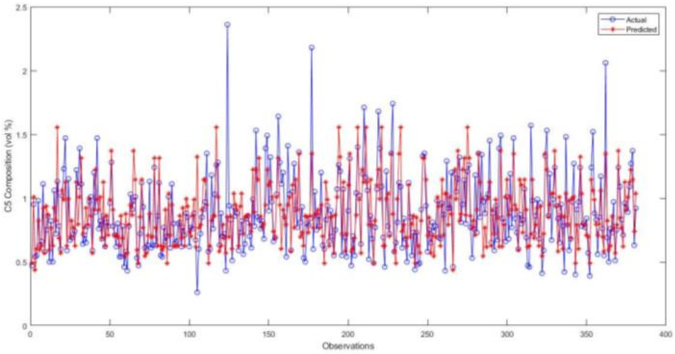

Fig. 11. Line plot of actual and predicted C5 composition by RT model.

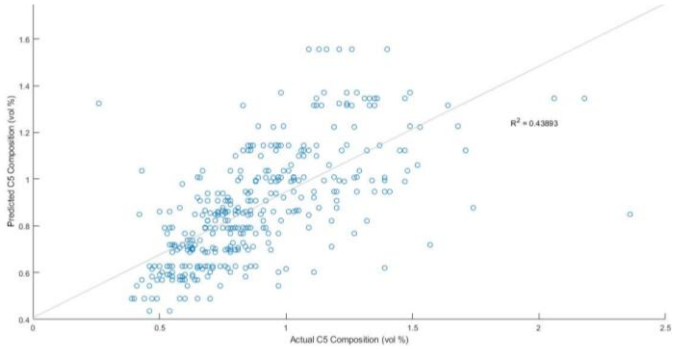

Fig. 12. Scatter plot of actual and predicted C5 composition by RT model.

F. GBRT

Tuning of boosting parameters, i.e. number of iterations (M), minimum leaf observations to split (min_split), minimum observations in a terminal node (min_node), maximum depth of a tree (max_depth) and learning rate (λ) is carried out through several iterations. As a result, the optimum boosting parameters are summarized in Table VII. The line plot of actual and predicted C5 composition by GBRT model are portrayed in Fig. 13. It is seen that predictions do not fit with the actual compositions. Fig. 14 displays the actual against predictions by GBRT model. The scattered plots and non-diagonal best fit line with R2 value of 0.58091 are signs of unsatisfactory model prediction. The calculated RMSE for this GBRT model is 0.1947.

TABLE VII. OPTIMUM BOOSTING PARAMETERS FOR GBRT MODEL TRAINING.

Parameter	Notation	Initial value
Number of iterations	M	4830
Minimum leaf observations to split	min_split	16
Minimum observations in a terminal node	min_node	1
Maximum depth of a tree	max_depth	3
Learning rate	λ	0.01

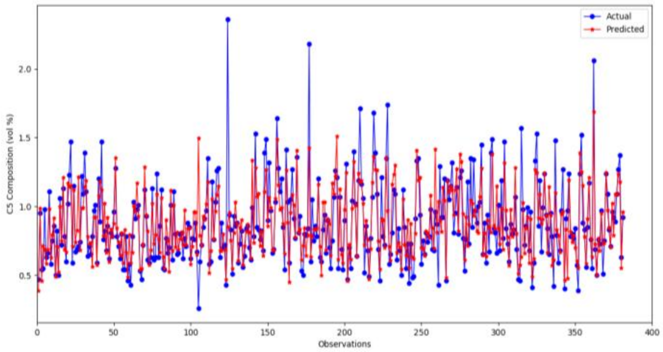

Fig. 13. Line plot of actual and predicted C5 composition by GBRT model.

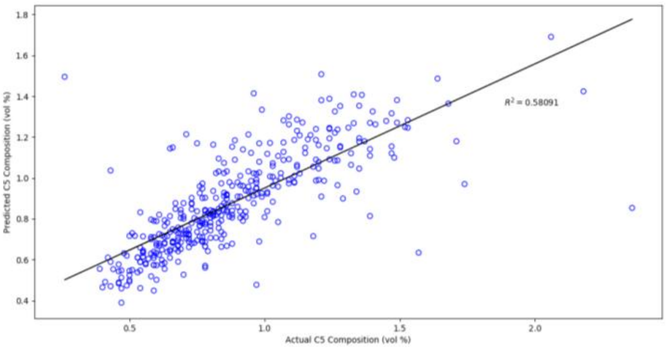

Fig. 14. Scatter plot of actual and predicted C5 composition by GBRT model.

G. ANN

A feed forward network is developed with 14 nodes in input layer, one hidden layer and a single node at linear output layer. Two types of activation function for hidden layer, including tangent sigmoid and linear activation functions are considered. The number of hidden nodes is determined through computing the RMSE of the trained network on training data. Fig. 15 shows the RMSE versus number of nodes in hidden layer using tangent sigmoid activation function. Eight hidden nodes are selected as they yield the least RMSE. Therefore, the topology of this network is 14-8-1. Fig. 16 demonstrates the deviation between actual and predicted C5 compositions by ANN model with tangent sigmoid activation function. It is observed that the predictions follow the trend of actual compositions except for the spike values. Fig. 17 illustrates the scatter plot of actual and predicted C5 composition by ANN model with tangent sigmoid activation function. It is shown that majority of the plots falls around the best fit line with a R^2 value of 0.65401. The calculated RMSE is 0.1795. This ANN model prediction accuracy is considerably good.

Fig. 18 shows the RMSE of ANN model using linear activation function with different number of hidden nodes. Five hidden nodes are preferred for training the network because they exhibit the least RMSE. Hence, the architecture of this network is 14-5-1. The variations of actual and predicted C5 composition by ANN model with linear activation function are depicted in Fig. 19. It is seen that several predictions are notably under-estimated. Fig. 20 portrays the correlation between actual and predicted C5 compositions by ANN model with linear activation function. The calculated RMSE is 0.19 and the best fit line with R^2 value of 0.60114 is seemed to be deviated from the 45-degree line. This indicates that the model prediction is unsatisfactory. Fig. 21 compares the RMSE and R^2 of ANN models with tangent sigmoid and linear activation functions of hidden nodes. It reveals that nonlinear mapping between process inputs and output using tangent sigmoid activation function outperforms linear mapping by 5.85% of RMSE and 8.08% of R^2. This signifies that nonlinear mapping manages to describe the nonlinear behaviour of naphtha stabilizing process and thus, nonlinear model is preferable.

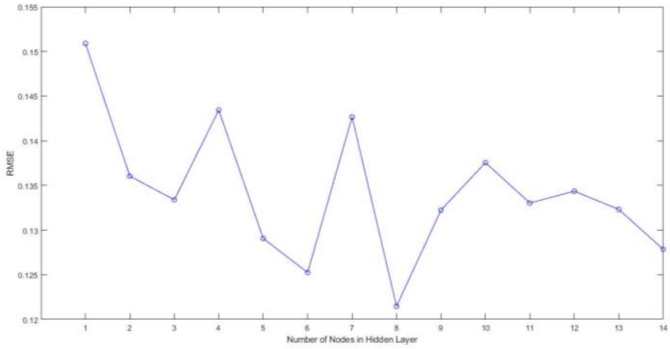

Fig. 15. RMSE against number of nodes in hidden layer using tangent sigmoid activation function.

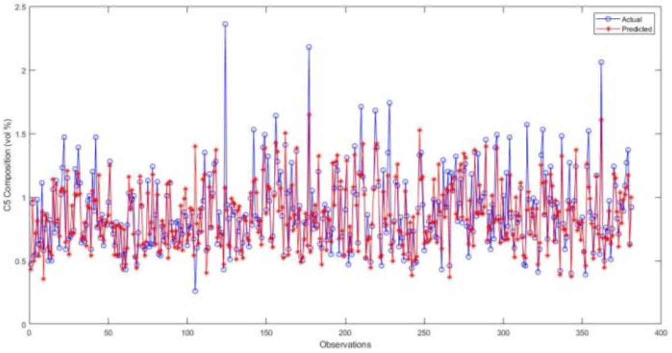

Fig. 16. Line plot of actual and predicted C5 composition by ANN model with tangent sigmoid activation function.

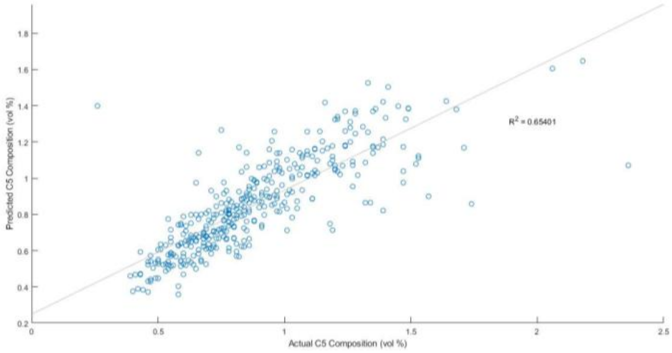

Fig. 17. Scatter plot of actual and predicted C5 composition by ANN model with tangent sigmoid activation function.

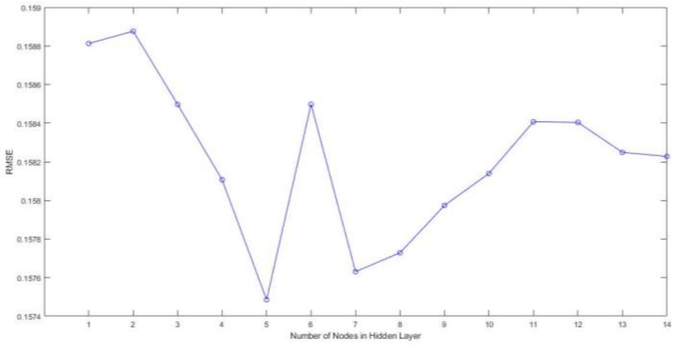

Fig. 18. RMSE against number of nodes in hidden layer using linear activation function.

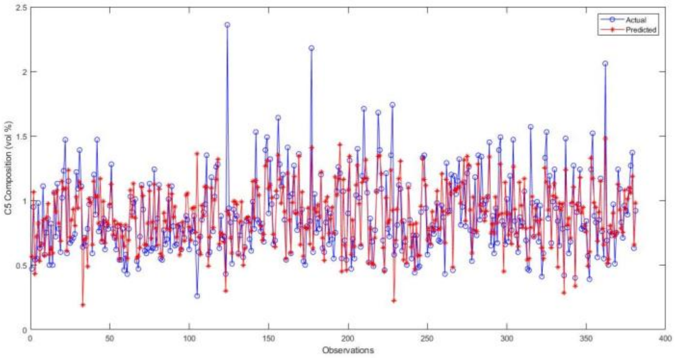

Fig. 19. Line plot of actual and predicted C5 composition by ANN model with linear activation function.

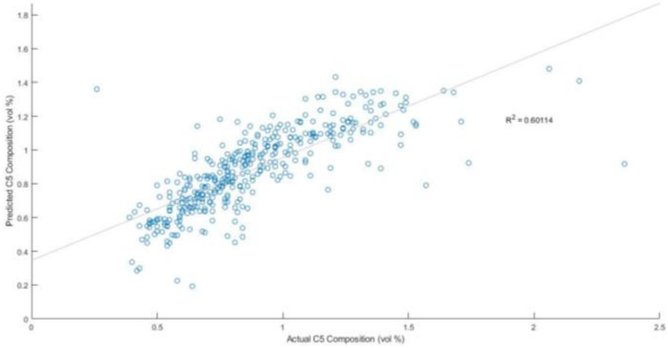

Fig. 20. Scatter plot of actual and predicted C5 composition by ANN model with linear activation function.

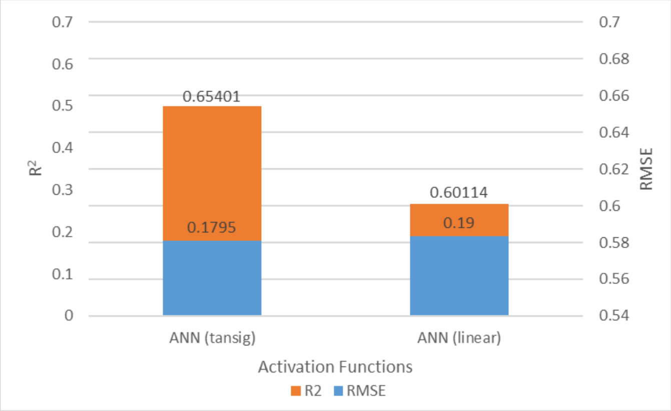

Fig. 21. RMSE of ANN model with different activation functions.

TABLE VIII. OVERALL ASSESSMENT OF DIFFERENT MODELLING TECHNIQUES.

Criterion	MLR	SLR	PCR	PLSR	RT	GBRT	ANN
Prediction Accuracy	2.25	2.5	1	1.5	1.25	2	3
Model development ease	2.5	2.5	2.5	3	2	1	1
Total	**4.75**	**5**	**3.5**	**4.5**	**3.25**	**3**	**4**

H. Comparative Analysis

Model prediction accuracy of all seven modelling techniques (MLR, SLR, PCR, PLSR, RT, GBRT and ANN) are compared based on RMSE and R^2. Fig. 22 depicts the performance of all modelling techniques. The results show that ANN model with tangent sigmoid activation function performs the best whereas PCR model performs the worst. The predictivity of SLR model is the second best, closely followed by MLR model and ANN model with linear activation function. The order of model performance with respect to RMSE and R^2 is as follows: ANN (tansig) > SLR > MLR > ANN (linear) > GBRT > PLSR > RT > PCR. The performance ranking of GBRT, PLSR and RT conforms with the results discovered by Keskin, et al. [25]. Also, the superior performance of nonlinear ANN model with Levenberg-Marquardt algorithm over MLR, PCR and PLSR models are in accordance with the results obtained by Pani, et al. [20]. However, they reported that PCR and PLSR models outperform MLR, which contradict with the results obtained in this study. Likewise, Madakyaru, et al. [10] presented that latent variable regression such as PCR and PLSR had better predictivity properties than MLR. Theoretically, PCR and PLSR should give better prediction than MLR model due to collinearity among the input variables. The possible reasons that cause such contradiction could be [27]:

- Low signal-to-noise ratio of data, where signal is the correlation of variances between independent and dependent variables whereas noise is the variance in independent variables that does not correlate with dependent variable.
- The noise in data is less impactful to MLR because variables that are not highly related to the response variable have no influence on the goodness of the resulting model.
- Latent variable regression is prone to Type I errors, which means overlooking structural factors that are related to the response.
- Collinearity of the input variables are not significant.

Similarly, the performance of SLR, PLSR and GBRT models are not in line with the findings published by Deconinck, et al. [26]. They reported that PLSR model outperformed BRT and SLR model. Nevertheless, the reported SLR model contained linear terms only while the SLR model developed in this study contains a combination of linear, bivariate and squared terms. Therefore, SLR model in this study has relatively outstanding performance because of the complex terms that successfully capture the nonlinear behaviour of the process. In short, nonlinear model such as ANN with tangent sigmoid activation function and complex SLR model are recommended for predicting C5 top composition of a naphtha stabilizer column from the perspective of prediction accuracy.

Table VIII presents the scores assigned to each modelling technique based on two criteria which are prediction accuracy and model development ease. Each criterion is rated using a scale ranges from 1 to 3 whereby the highest score represents the best while the lowest score represents the worst. The score ranking of the seven modelling techniques is as follows: SLR > MLR > PLSR > ANN > PCR > RT > GBRT. Overall, SLR is the most recommended modelling technique for industrial application due to its prominent prediction accuracy and simplicity in developing an empirical model.

VI. CONCLUSION

This paper presents a comparative analysis of seven different modelling techniques for estimating C5 top composition of a naphtha stabilizer column using historical plant data. The seven modelling techniques are multiple linear regression (MLR), stepwise linear regression (SLR), principle component regression (PCR), partial least squares regression (PLSR), regression tree (RT), gradient boosting regression tree (GBRT) and artificial neural networks (ANN). Several important findings are summarized as below:

- RT and GBRT have advantages over the other statistical modelling techniques. They are insensitive to outliers and able to make prediction with missing value present in unseen data.

- The order of model predictive performance is ANN (tansig) > SLR > MLR > ANN (linear) > GBRT > PLSR > RT > PCR.
- Feedforward ANN with tangent sigmoid activation function has the highest prediction accuracy and descriptive power because of its nonlinear mapping between input and output data that fits with the nonlinear nature of distillation process.
- SLR outperforms MLR, PCR and PLSR due to the complex terms in the model that fairly describe the nonlinear behaviour of the process.
- MLR performs better than PCR and PLSR could be due to the low signal-to-noise ratio of the input data, in which MLR modelling is less sensitive towards noise.
- The overall score ranking based on the criteria of model prediction accuracy and development ease is as follows: SLR > MLR > PLSR > ANN > PCR > RT > GBRT.
- SLR is the most preferable modelling technique for industrial application.

It is important to note that this paper focuses merely on steady state data. Therefore, future work can emphasise on dynamic response of the process by considering the time lag of each process variable. Hybrid of two different modelling techniques is another option that future work can opt for.

ACKNOWLEDGMENT

The authors would like to acknowledge PETRONAS for providing the required data and information for the research. Also, the authors would like to thank Universiti Teknologi PETRONAS for giving supports.

REFERENCES

[1] S. Khatibisepehr, B. Huang, and S. Khare, "Design of inferential sensors in the process industry: A review of Bayesian methods," *Journal of Process Control,* vol. 23, no. 10, pp. 1575-1596, 2013.

[2] D. E. Seborg, D. A. Mellichamp, T. F. Edgar, and F. J. Doyle III, *Process dynamics and control*. John Wiley & Sons, 2010.

[3] R. K. Pearson, "Selecting nonlinear model structures for computer control," *Journal of Process Control,* vol. 13, no. 1, pp. 1-26, 2003/02/01/ 2003.

[4] F. A. A. Souza, R. Araújo, and J. Mendes, "Review of soft sensor methods for regression applications," *Chemometrics and Intelligent Laboratory Systems,* vol. 152, pp. 69-79, 2016.

[5] T. Mejdell and S. Skogestad, "Estimation of distillation compositions from multiple temperature measurements using partial-least-squares regression," *Industrial & Engineering Chemistry Research,* vol. 30, no. 12, pp. 2543-2555, 1991/12/01 1991.

[6] M. Kano, K. Miyazaki, S. Hasebe, and I. Hashimoto, "Inferential control system of distillation compositions using dynamic partial least squares regression," *Journal of Process Control,* vol. 10, no. 2, pp. 157-166, 2000/04/01/ 2000.

[7] J. Zhang, "Inferential feedback control of distillation composition based on PCR and PLS models," in *Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148),* 2001, vol. 2, pp. 1196-1201: IEEE.

[8] M. Kim, I.-S. Han, and C. Han, "Modified PLS method for inferential quality control," in *Computer Aided Chemical Engineering,* vol. 15, B. Chen and A. W. Westerberg, Eds.: Elsevier, 2003, pp. 876-881.

[9] B. Lin, B. Recke, J. K. H. Knudsen, and S. B. Jørgensen, "A systematic approach for soft sensor development," *Computers & Chemical Engineering,* vol. 31, no. 5-6, pp. 419-425, 2007.

[10] M. Madakyaru, M. N. Nounou, and H. N. Nounou, "Linear Inferential Modeling: Theoretical Perspectives, Extensions, and Comparative Analysis," *Intelligent Control and Automation,* vol. 03, no. 04, pp. 376-389, 2012.

[11] H. Jin, X. Chen, J. Yang, and L. Wu, "Adaptive soft sensor modeling framework based on just-in-time learning and kernel partial least squares regression for nonlinear multiphase batch processes," *Computers & Chemical Engineering,* vol. 71, pp. 77-93, 2014/12/04/ 2014.

[12] V. Singh, I. Gupta, and H. O. Gupta, "ANN based estimator for distillation—inferential control," *Chemical Engineering and Processing: Process Intensification,* vol. 44, no. 7, pp. 785-795, 2005/07/01/ 2005.

[13] A. K. Pani, V. K. Vadlamudi, and H. K. Mohanta, "Development and comparison of neural network based soft sensors for online estimation of cement clinker quality," *ISA Transactions,* vol. 52, no. 1, pp. 19-29, 2013/01/01/ 2013.

[14] K. Desai, Y. Badhe, S. S. Tambe, and B. D. Kulkarni, "Soft-sensor development for fed-batch bioreactors using support vector regression," *Biochemical Engineering Journal,* vol. 27, no. 3, pp. 225-239, 2006/01/01/ 2006.

[15] P. Jain, I. Rahman, and B. D. Kulkarni, "Development of a Soft Sensor for a Batch Distillation Column Using Support Vector Regression Techniques," *Chemical Engineering Research and Design,* vol. 85, no. 2, pp. 283-287, 2007.

[16] Y. Xuefeng, "Hybrid artificial neural network based on BP-PLSR and its application in development of soft sensors," *Chemometrics and Intelligent Laboratory Systems,* vol. 103, no. 2, pp. 152-159, 2010/10/15/ 2010.

[17] A. Rogina, I. Šiško, I. Mohler, Ž. Ujević, and N. Bolf, "Soft sensor for continuous product quality estimation (in crude distillation unit)," *Chemical*

Engineering Research and Design, vol. 89, no. 10, pp. 2070-2077, 2011/10/01/ 2011.

[18] A. Rani, V. Singh, and J. R. P. Gupta, "Development of soft sensor for neural network based control of distillation column," *ISA Transactions,* vol. 52, no. 3, pp. 438-449, 2013/05/01/ 2013.

[19] K. Sun, S.-h. Huang, S.-S. Jang, and D. S.-H. Wong, "Development of soft sensor with neural network and nonlinear variable selection for crude distillation unit process," in *Computer Aided Chemical Engineering*, vol. 38, Z. Kravanja and M. Bogataj, Eds.: Elsevier, 2016, pp. 337-342.

[20] A. K. Pani, K. G. Amin, and H. K. Mohanta, "Soft sensing of product quality in the debutanizer column with principal component analysis and feed-forward artificial neural network," *Alexandria Engineering Journal,* vol. 55, no. 2, pp. 1667-1674, 2016/06/01/ 2016.

[21] A. Urhan, N. G. Ince, R. Bondy, and B. Alakent, "Soft-Sensor Design for a Crude Distillation Unit Using Statistical Learning Methods," in *Computer Aided Chemical Engineering*, vol. 44, M. R. Eden, M. G. Ierapetritou, and G. P. Towler, Eds.: Elsevier, 2018, pp. 2269-2274.

[22] J. Elith, J. R Leathwick, and T. Hastie, *A Working Guide to Boosted Regression Trees*. 2008, pp. 802-13.

[23] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 308-324, 2015/09/01/ 2015.

[24] C. Ding, X. Cao, and P. Næss, "Applying gradient boosting decision trees to examine non-linear effects of the built environment on driving distance in Oslo," *Transportation Research Part A: Policy and Practice*, vol. 110, pp. 107-117, 2018/04/01/ 2018.

[25] H. Keskin, S. Grunwald, and W. G. Harris, "Digital mapping of soil carbon fractions with machine learning," *Geoderma,* vol. 339, pp. 40-58, 2019/04/01/ 2019.

[26] E. Deconinck *et al.*, "Boosted regression trees, multivariate adaptive regression splines and their two-step combinations with multiple linear regression or partial least squares to predict blood-brain barrier passage: a case study," *Anal Chim Acta*, vol. 609, no. 1, pp. 13-23, Feb 18 2008.

[27] R. Cramer, *Partial Least Squares (PLS): Its Strengths and Limitations*. 1993, pp. 269-278.