The efficiency of soft sensors modelling in advanced control systems in oil refinery through the application of hybrid intelligent data mining techniques

Ali Al-Jlibawi¹, Mohammed Lutfi Bin Othman¹, Muayed S. Al-Huseiny², Ishak Bin Aris¹ and Samsul Bahari¹

¹Center of Advanced Power and Energy Research (CAPER), Dept. of Electrical and Electronic Engineering, Faculty of Engineering, Universiti Putra Malaysia, Selangor, Malaysia
²Faculty of Engineering, University of Wasit, Kut, Iraq

e-mail: alihhumod@gmail.com

Abstract. Quality variables cannot be automatically measured to all nor be measured at a high cost, infrequently, nor with high delays, such as laboratory analysis and online analyser. Therefore, data-driven soft sensors are inferential models which use online available sensors, such as temperature, pressure, and flow rate among others, to predict the quality variables. Soft sensors which are built using historical data of the processes are normally developed from the supervisory control and data acquisition (SCADA) systems connected with PLC or DCS (distribution control systems) as the daily reports on the oil refinery processes. These systems are then obtained from laboratory observation/measurements. Notably, the main issue in the development of the soft sensor is the treatment of missing data, outlier detection, selection of input variables, model training, validation, and soft sensor maintenance to adopt the heavy-duty oil refineries to improve the products of the crude oil and increase yield. In this article, the improvement in the virtual sensor based on hybrid soft computing methods (FLS and NN), which are combined into ANFIS, will be employed to construct the soft sensor model. Moreover, RST will be used to reduce the fuzzy rules and discretisation method to optimise and deal with the large continuous data. It was found from the implementation of rough set theory and discretisation methods that these two methods solved the complexity and nonlinearity of the soft sensor model. This model was employed for the refining process measurements data of the oil refinery from two different crude oil sources, in which the database of the measurements and processes was combined to improve the quality of data and discover the knowledge stored in the data pattern. It was indicated from this study result that the ANFIS model is able to manage the complex data to predict two important parameters of light naphtha (API and RVP) compared to the simple regression model. Additionally, controlling and monitoring the process are crucial actions performed to achieve the 4th industrial revolution and IoT. This study has contributed to the assistance in breaking the barriers of privacy between oil industries and the applicability of soft sensors modelling in the changes of data sources to achieve remarkable data analysis. The analyses result of RVP show the efficiency of ANFIS compare with linear regression regarding the generalization and overfitting.
1. Introduction

Although industrial processing plants are normally highly instrumented with various sensors, a low number of reliable and accurate sensors are present, which perform precise online measurement of the quality variables, particularly those associated with the composition. Sensors mainly aim to distribute data for process monitoring and control [1]. This research article primarily aims to improve data quality and reduce the redundant information in manufacturing data. Furthermore, data mining techniques are used in knowledge discovery, with each technique possessing both advantages and disadvantages. Meanwhile, soft computing is used to perform data mining projects. The development of intelligent sensors is according to the functional association of the product quality variables with other process variables, which could be evaluated online [2].

The vast development of modern process control methods has an impact on process monitoring and control in all aspects. Furthermore, the constant focus on new challenges could be seen from the development of applications methods, soft sensors modelling, intelligent sensors, modern actuator, new regulatory methods, and control theory and aspects. A Supervisory Control and Data Acquisition (SCADA) and Programmable Logic Control (PLC) systems are constantly applied to manage small industries, such as irrigation systems, electric power stations, water treatment stations, and oil and gas refineries, which are mostly dependent on Distribution Control Systems (DCS) [3]. Moreover, petroleum refinery process control is performed using control technologies, such as conventional control and modern control technologies. To be specific, the conventional method involves the use of microcontrollers for limited control parameters and small-scale control applications. Meanwhile, modern technologies, including DCS, PLC and SCADA are implemented in a large geographical area across the process control with multiple process parameters.

Crude oil distillation tower refers to the process element, which implements the separation of petroleum cuts, making it the center of any crude refinery. The petroleum cuts are later processed in other operation units so that the light and heavy naphtha, gas oil, gasoline, and other commercial products are refined and blended. The economic objective of the distillation tower operation is according to the acquirement of the highest number of product with the optimum quality set in the specifications [4]. This research paper will focus on the process performed by the industries for petroleum production and the development of modelling for soft sensors. Accordingly, the construction of virtual sensors, especially in an oil refinery, will be discussed in detail due to the higher effectiveness of this industry on the local revenue in Malaysia refineries and other countries refineries.

Adjustment should be made on the combination of the overall product and process development chains through the communication between the design, manufacture, marketing, and management in chemical and petroleum industries in terms of simulation and modelling [5]. Essentially, the soft sensor delivers additional online data for the process control. [6] And the soft sensor-based control system could calculate the control action using the adaptive model adjustment [7]. The performance of soft sensors is influenced by the quality of the information used to acquire the knowledge in the identification procedure. This information is saved in the historical plant databases, which are essential for other motives besides modelling.

Due to the common usage of computer and networking technologies, including the emergence of new data acquisition systems in manufacturing organizations, large electronic databases are developed, which store the product, manufacturing process, or equipment related information. The analysis could be performed on this information to determine the potential pattern in the parameters, which influenced the manufacturing procedures or product quality [8]. Furthermore, monitoring and evaluation need to be made on the polymerization, fermentation, and refineries in terms of the processes and the extreme and risky environment. In this case, processing variables create the essential value of the production line procedures. Notably, there are various phases of operations, with many variables present to contribute to an understanding of the association of the parameter variables of the system with the processes.
The soft sensor is used as a solution to various issues, including the measurement of system backup, real-time estimation for plant control, and the methods of sensor validation and fault diagnosis. It is noteworthy that acquiring historical data using industrial procedures to develop the model is a challenging task. The profit of soft sensor used in the industrial field is based on its characteristics in the closed-loop inferential control schemes. Moreover, the main soft sensors could be described as the backup for device measurement, real-time prediction of control and monitoring, and the reduction of the measurement of hardware specifications, what-if analysis, including the fault identification in sensor validation [9].

Several works of literature have investigated the soft sensors in oil refineries and the cooperation of data between the manufacturers. The challenge of the modelling of soft sensors to adopt the changing data is due to the various feedstocks, especially those in the oil refineries sectors. This is followed by the changing process conditions and work environment. Furthermore, the complexity, nonlinearity, and the data quality of industrial processes. Accordingly, the creation of a robust and efficient soft sensor requires data sharing between the industries, overcoming the privacy barriers, and cooperation between the manufacturers. However, the soft sensing in crude oil distillation rowsers with diverse feedstocks is still a challenging issue due to the association of the efficient measurement of the process variables with the challenging measurement. These variables could be distinguished from one another based on the type of crude oil processed, which state is influenced by the suppliers. Additionally, different hydrocarbon content might be present in the crude from the same supplier, and the operation of numerous refineries takes place with various sources of crude oil and a wide range of blending ratios.

This research aims to optimize the performance of soft sensors for the hybrid neuro-fuzzy model (soft computing methodology) according to the rough set theory and various discretization methods in the event of a high number of continuous data in oil refinery control system. It also aims to predict and identify the quality of cuts (fractions) of the light naphtha and adopt the changed in the data process. Another objective of this research is to enhance the quality of data by combining the decision tables of soft computing model from different sources and the database of the case studies. To prove the validity of the soft sensor model, an improvement would be made by comparing this model with another regression model.

The scope of this study is the investigation on the refined crude oil units in the petroleum refinery, which convert the crude oil into lighter hydrocarbon products, such as light diesel fuel, gasoline, and heavy naphtha. These units are crucial in the overall economic performance of the refinery, and they are particularly focused on the control of the API (American Petroleum Institution) which is the unit to measure the density of petroleum and their products, while RVP (Reid Vapor Pressure) refers to the criteria of measuring the quality control of the light naphtha. Furthermore, the soft sensors become a trend for many industrial field and large plants due to their potential applicability despite several issues including difficulty in measuring the process parameters, delay in measurement, severe working environment for the evaluation of device survival, and difficulty in managing the requirements.

This research article is organized into four sections. To be specific, the first section of this article reviews the research on the soft sensor modelling in oil refinery control systems and the manufacturing application. The second section presents the development of the soft sensors, which implemented the neuro-fuzzy soft computing methodology to predict the quality of crude oil in distillation towers and improved the quality of the soft sensors data based on the discretization methods and rough set theory. Following that is the third section, which illustrates the result of the suggested model for the data process of the refined unit in the oil refinery. Lastly, the fourth section performs a review on the goal of the industry 4.0 and its relation with soft sensor, including the validity of the combination of the same models for the soft sensors in different refineries to improve the intelligent data mining techniques worldwide.
2. Development of soft sensor

Sensors are the visions of the manufacturers. The term Soft sensor was derived from the incorporation of the word “software” as the models were normally computer programmers, while the term “sensors” emerged due to the same information distributed by the models as their hardware counterparts. Other common terms for predictive sensors in the process industry are observed-based sensors, on-line analyzer, virtual, and inferential sensors. However, traditional sensors would be faced with difficulties in seeing various targets. Many sensors are often needed by the complex operations to perform their tasks.

Soft sensor modelling techniques according to the empirical models consisted of Kalman filter, artificial neural networks (ANN), fuzzy logic, multivariate statistical, regression-based model, and hybrid methods. The accuracy and robustness of soft sensing model are strongly influenced by the availability of training data. Moreover, it was previously stated that it is crucial for soft sensors to not only exhibit precise measurement, the measurement should also be significant to the new plant data [10]. Soft sensors were created by Rogina et al. (2011) for continuous good measurement of light naphtha as the crude distillation unit (CDU) product. It was found that the nonlinear soft sensors models using MLP and RBF NN displayed an acceptable range of deviations, while it was implied by the statistical indicators that these models could be incorporated in property monitoring and process product quality [11]. In this study, a significant amount of information was evaluated and retained in the process industry through the development of the related predictive model.

Small data sets were used to compare among the soft sensors design methods for industrial plants. This method was based on the combination of neural models, which were trained on various training data sets. Noise injection and bootstrap resampling were applied to obtain the data sets. Then, a comparison was made between the methods used in this study and those used in an industrial case study regarding the backup soft sensors design for a thermal cracking unit in the refinery in Sicily, Italy. The analysis was also conducted in a real case study on the noise injection bootstrap and stacking to enhance the generalization abilities of a neural network [12].

The mathematical models of processes, which were created based on the experimental data through the system identification procedures, could significantly reduce the necessity of measuring the devices and develop secure control policies. Among the mathematical models created as the representations of the computing devices for crucial cases in industries were soft, inferential, or virtual sensors. Furthermore, a neural network-based soft sensor was created so that the effluent concentration in active sludge process could be estimated online in terms of the primary difficulty in measurement variables. These variables included total nitrogen content, chemical oxygen demand, and total suspended solids, which were initiated by the secondary efficiency in online measurement variables, such as the concentrations of oxygen and nitrogen compounds in biological tanks, alkalinity, and input flow rate among others.

In the selection of the optimal net input vector for the soft sensor, an algorithm based on the principal component analysis (PCA) was implemented, and an adequate number of samples of the secondary variables was used. Meanwhile, the soft sensors were based on the use of the PCA with an artificial neural network [13]. The hybrid soft computing model was proposed in numerous case studies, and the Neurofuzzy system based on genetic algorithm and rough set theory was used to estimate the freezing point in FCCU (Fluid Catalytic Cracking Unit). This was followed by the implementation of the soft sensor, which was based on a Neurofuzzy system, to integrate the qualitative reasoning of fuzzy logic with the quantitative numeric processing of ANNs. While RST was used to produce the reductive rule set of the Neurofuzzy system, while GA was used to obtain the optimal discretization value [14].

2.1. Method of the study

Soft sensors refer to the mathematical models which enable the inference of the relevant variables based on the influence of the set of inferential variables on the relevant variables. The processes of designing and applying the soft sensors started with the selection of historical data from the outlier
detection and data filtering, plant database, model structure and regressor selection, and model estimation and validation. Before the model performed its functions, the size, quality, and nature of data, importance of the task, current computational method, and the actions to be performed on the data were the elements to be aware of.

The methodology of this research started with the definition of the manufacturing issue in a precise statement, namely the processes of CDU and the classification of the variable input and output. This was followed by the definition of the data model and requirement, where the data from all repositories, namely refinery (1) and refinery (2) were extracted either from DCS or SCADA. Then, preparation of data was performed, in which the data might be relational or be located in a flat, files, data warehouse, and calculated on the site or purchased from another party. Upon the evaluation of the data quality, soft sensor model was selected. This model referred to the hybrid soft sensor model fuzzy logic system, which was combined with the neural network based on the rough set theory and discretization methods. Following that, the results of this research were elaborated and new information (prediction) were identified. The results and the new knowledge were then applied in the manufacturing field. Meanwhile, the types of data collected from different sources were the process variables data which are secondary variables and expert’s knowledge. Notably, the data set should be split into two parts, namely the training or generalization data and the validation or testing data. With the classic split, the dataset was divided into either four or five datasets based on the data size.

2.1.1. Case of the study. This research investigated on oil refineries, especially crude distillation unit. The most essential element in the distillation tower is the measurement of the quality of crude oil products. The suggested model in this study used a hybrid soft computing fuzzy logic system combined with a neural network based on rough set theory and discretisation methods. Following that, the fuzzy inferential systems (FIS) were developed by reducing the rules based on the rough set theory, which was used to reduce the rules. Although this study aims to identify useful knowledge inside the pattern for each data, the data stores important information for the suggested model. Notably, the advantage of fuzzy logic lies in the efficient incorporation of the prerequisite knowledge regarding a system into a direct fuzzy rule base. Expert’s knowledge would play an important role in this method.

2.1.2. Control systems in the oil refinery. Most of the control systems of an oil refinery are distribution control system, which store several recorded data and process it for the control and monitoring of the process in the petroleum product. The measurement facilitated the values of the process variables (PV’s) to DCS, which transferred the information to the graphical user interface (GUI). The operator of the technology is capable of obtaining data regarding the system through GUI and managing the process by changing the set points (SP’s). Furthermore, an advanced model-based process control (process computer) computer figures the operation set points (OP’s) to DCS.

Although the information retained in DCS is capable of distributing information for the product and process design, including control and monitoring, there is limited access to the information in the process control computer. Defective data could be produced when they are archived retrospectively due to failure in measurement and inconsistent storage. The DCS and transmitters used in ALDORA oil refinery originated from the Japan Yokogawa Company, which performs critical operation infrastructure and improves business performance. Meanwhile, PID represents the standard control loop for the essential parameter of the crude distillation unit (flow, level, temperature, pressure) of the research targets (heavy and light naphtha). The majority of the control functions is Cascade slave-controller PID regulator, Cascade Master- controller PID regulator, Standard PID regulation, and Alarm, while the remote setpoint is the output from PID controller.

2.1.3. The platform of the simulation and calculation of the model. MATLAB software was the platform used to implement the simulation of soft sensors modelling for FLS, ANN, and discretisation method due to its potential in analysing and implementing the theories previously mentioned in this
study. In the case of RST, RStudio software was used as the programming platform for RST, which included all the RST packages and calculation tools. In this case, the script was linked to MATLAB to achieve one block of soft sensor modelling.

2.1.4. Process variables of oil refinery. The main variables of the distillation unit process included column pressure, column feed temperature, column reflux and pumparound, column stripping steam, and column product withdrawal. In this research, the processes of the fraction naphtha began with the crude oil refining processes and ended by storing the light naphtha before proceeding to the next process. Upon the discussion with engineers in the refining department of ALDORA oil refinery, the engineers highlighted that the crude oil originated from various resources, with each resource contributing to diverse CDU procedures impacts on their parameters (temperature, pressure, and flow rate). Furthermore, two critical parameters were tested for the crude oil (API and RVP), followed by three important parameter effects on the cuts and quality of products of each cut in terms of temperature, pressure, and flow rate.

Notably, heavy naphtha and light naphtha were important products with various applications, namely CDU, which consisted of crude distillation column at atmospheric pressure, stripping unit, preheating section which used top and bottom pumparound, and overhead condensing system splitter and stabiliser units. The two key factors in the distillation process, namely temperature and pressure, must be correct as they function together. Meanwhile, to achieve an efficient distillation process, the temperature was the process variable to be prioritised. This was followed by a pressure parameter of light naphtha, API, RVP, flash point, I.B.P, E.P, and colour. The tower pressure was fixed and the products tested daily at 2 a.m., 8 a.m., 2 p.m., 8 p.m., and the collection of laboratory testing report was performed on the 1st of September for 60 days. Subsequently, the DCS daily report was simultaneously provided when the sample for the light naphtha was obtained for analysis and pre-processing.

2.1.5. Data collection. The collection of data regarding the system from the control system and operating system in oil refinery was a critical stage in the research, and the data were stored in DCS. These processing data were collected from the daily report for the actions of the unit. As mentioned previously, the CDU consisted of four main sections, namely atmospheric tower CO1, stripper CO2, stabiliser CO3, splitter CO4, and other complementary equipment, including heater, cooler (air and water), and condenser. This equipment played its roles in the operation processes and variables. Furthermore, the data were divided into two groups. To be specific, the first group consisted of the data collected from the refining processes, the process variables (independent variables), or directly from the DCS of the control rooms, which stored all the daily variables for all the processes. Meanwhile, the second part of data consisted of the dependent variables collected in every six hours from the daily reports on the quality control department of the refinery. The completed data arranged in an Excel datasheet were then converted into a CSV file with two groups consisting of variables of the processes, namely pressure and temperature. Pressure and temperature were the most effective variables on the target data sought by this research, while the output variables are the API and RVP of the light naphtha. The next important step of this research after the arrangement of the collected data was the filtering of the raw data, which was conducted to identify the malfunctioned data or the outlier using the pre-processing and filtering algorithm. As a result, the total data collected amounted to two hundred and six samples from the light naphtha testing reports. The sample of system information is presented in the Table 1 below:
Table 1. system information (Decision table) of the collected data for oil refinery processes

| T0331 | T032 | T034A | T034B | T0337 | T0339 | T0335 | T0338A | T0338B | T0329 | T0330 | RVP | IBP |
|-------|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-----|-----|
| 60.00 | 70.00 | 150.00 | 150.00 | 175.00 | 135.00 | 6.60  | 7.20  | 136.00 | 165.00 | 100.00 | 150.00 | 150.00 | 1.10 | 1.00 | 38 | 0.74 |
| 59.90 | 76.61 | 126.41 | 126.15 | 150.58 | 113.00 | 6.60  | 7.20  | 98.00  | 129.38 | 91.34  | 113.12 | 112.99 | 1.20 | 0.99 | 37 | 0.85 |
| 60.00 | 76.62 | 130.88 | 127.00 | 140.12 | 117.08 | 6.24  | 7.82  | 100.22 | 130.81 | 100.43 | 140.64 | 135.24 | 1.10 | 1.10 | 38 | 0.84 |
| 61.00 | 76.00 | 130.00 | 129.79 | 140.50 | 117.05 | 6.40  | 7.79  | 100.12 | 130.58 | 100.31 | 140.96 | 135.12 | 1.10 | 1.10 | 35 | 0.74 |
| 59.80 | 80.00 | 130.68 | 129.96 | 140.96 | 117.53 | 6.10  | 7.64  | 100.30 | 130.18 | 100.16 | 140.24 | 135.61 | 1.11 | 1.12 | 33 | 0.70 |
| 60.00 | 78.00 | 130.28 | 129.66 | 140.34 | 117.78 | 6.13  | 7.38  | 100.32 | 130.24 | 100.18 | 140.68 | 135.45 | 1.12 | 1.12 | 38 | 0.76 |

3. Prediction model
The quality of light naphtha was authenticated by RVP and API laboratory, and the measured RVP and API values should not exceed the limitation set during the design the CDU. The constant observation was conducted on the variables, which impacted the light naphtha RVP and API. According to the knowledge of the expert, specific variables were selected as the input variables, which would function as the estimators of light naphtha RVP and API reported in the laboratory tests by using the analysis on the standard (ASTM, IP). The aforementioned variables are as showing in details in table 2

Table 2. showing the details of the process variables for column CO3 and CO4

| Symbols | Definition in the refining processing |
|---------|--------------------------------------|
| T031    | Temperature of the top of the stabiliser column. |
| T032    | Temperature of the middle of the stabiliser column. |
| T034A   | Temperature of the bottom of the stabiliser column. |
| T034B   | Temperature of the bottom of the stabiliser column. |
| T037    | Temperature of stabiliser naphtha. |
| T033    | Temperature of the feeding naphtha. |
| P028    | Pressure of the head of the stabiliser column. |
| P027    | Pressure of the bottom of the stabiliser column. |
| T036    | Temperature of stabilised naphtha from E12 to the splitter. |
| T039    | Temperature of heavy naphtha from the bottom of the splitter column. |
| T035    | Temperature of the head of the splitter column. |
| T038A   | Temperature of the bottom of the splitter column. |
| T038B   | Temperature of the bottom of the splitter column. |
| P029    | Pressure of the bottom of the splitter column. |
| P030    | Pressure of the head of the splitter column. |

An increase was observed in the RVP light naphtha, while a decrease in the I.B.P was found due to the reduction in the temperatures levels of the top and bottom of Column CO3, including the reduction in the temperature levels of the top and bottom of Column CO4. However, an increase in E.B.P occurred when the temperature levels of the top and bottom for Columns CO4 increased.
3.1. Statistical data processing

Various software, which had been used in this research (Microsoft office Excel, MATLAB, and RStudio), was utilised in the statistical analysis, leading to similar results. Data collection was conducted at the beginning of September 2018 to the end of October 2018 from various resources of crude oil; Kirkuk, Medium Basra, Basra. The collection of laboratory data was performed from the refining department four times daily. Subsequently, the data gathered from the distributed control system were matched in terms of time delays and sampling. Data filtration and pre-processing would be conducted after obtaining the descriptive statistics mean value, median, minimal and maximal value, lower and upper quartile, and variance and standard deviation were pre-processed.

3.1.1. Data filtering and pre-processing. This sub-section presents the low quality of data led to unsatisfied data mining results, and the quality of estimation and prediction, which is according to the quality data. The objective of this procedure is to improve the effectiveness of the data processing by the actual model through data modification. In the case of the data created in the process industry, several pre-processing stages were involved, namely missing data, outlier detection and replacement, choice of relevant variables, management of drifting data, and identification of delays between the particular variables. Data pre-processing is normally performed in an iterative way, while missing value treatment and standardisation are only conducted once. Moreover, data pre-processing involved other processes, namely data integration, cleaning, transformation, reduction, and discretisation. Following this process, several samples were reduced to 200 for the collected data based on the data quantity. Meanwhile, outlier removal methods and feature selection were applied on repeat until the model data were ready to be used for model training and assessment of the actual model.

3.1.2. Data and variable reduction. Rough set theory is among the most crucial theories which reduce the set of rules in the fuzzy systems. In this section, improvement and reduction of rules were performed using RStudio software, which is the most powerful software to implement RST. This was followed by a more accurate examination of the approximation. This phase started with the definition of a data set, which was also known as an information system. The representation of the data set acquired and produced in this study was significantly smaller in volume despite the similar or almost similar analytical results. Data reduction methodology consisted of data cube integration, data compression, the decrease in dimensionality, numerosity reduction, concept hierarchy generation, and discretisation. Furthermore, the temperature and pressure variables in CO3 and CO4 were represented by the condition attributes, while API and RVP were represented by the decision attributes. Moreover, the globally semi-optimal cuts were implemented using the maximum discernibility heuristics. The Quick-Reduce algorithm was used to reduce the number of genes from expression data, and the rough set-based Quick-Reduce algorithm was an efficient algorithm to find the reduce and select the heuristic attributes.

It is illustrated in the RST analysis that the conditional attributes were reduced to 137 rules, with the final selection attributes consisting of TI031, PI027, TI036, TI035, and RVP decision attributes. The reduce rules implemented by RST were applied by RStudio software and incorporated the Quick-Supper Reduct algorithm.

3.2. Adaptive Neural Fuzzy Inferential System (ANFIS)

Adaptive Neural Fuzzy Inferential System (ANFIS) is a data learning method which incorporates fuzzy logic for the transformation of the provided input into the targeted output using highly interconnected information connection and Neural Network processing elements. The weights of these two components were measured to form the numerical inputs into an output. Meanwhile, provided that RST was used in the previous section to reduce the number of Fuzzy rules, this section displays the attempt of this study to use the Fuzzy logic with neural network in a Hybrid model known as ANFIS. The advantage of this model could be seen from its capability to simulate the reasoning ability of Fuzzy rules, which would interpret the expert’s knowledge. This was followed by the adaptation of the
knowledge in rules and the computational and learning abilities of Neural Network. Moreover, the input data of this model would be portioned out into training and checking sets. As a result, the training set constituted 70 % of the pre-processed data, while the test set constituted 30 % of the data. In some cases, random portioning was used based on the needs of the data needing in each level of the process.

3.3. Input selection
For the selection of the set of the input which posted the strongest impact on API and RVP, a thorough search was conducted within the available input. As a result, an ANFIS model was developed for each combination, which was trained for one epoch before its performance was recorded. Meanwhile, a heuristic approach to input selection is known as the sequential forward search, where each input was sequentially chosen to achieve an optimised total squared error.

3.4. The discretisation of new inputs
The discretisation is a data pre-processing task, which transforms continuous variables into the discrete variables. It is also a process of quantising continuous attribute to apply several data mining algorithms. In this section, new inputs were obtained from the attribute defined in input selections. The extraction of the chosen input attribute from the original training was performed, and the datasets were checked. With the use of MATLAB, code as a present to discretise the continuous variables of the inputs. The freezing for the discretisation step was performed in the suggested model, in which this step would be active in a significant amount of data.

3.5. Training of the ANFIS model
Following the identification and fixing of the input, the training of the ANFIS model with 40 epochs became the main focus of this section. An initial FIS from the training data was generated, which was then fine-tuned by ANFIS to develop the final model. Gaussian function was the membership function to be used for FIS, while the Sugeno-type fuzzy inference system was tuned using the training data. In this case, the error was returned through ANFIS in terms of data training and checking in the list of the output parameters. As shown in Figure 1 below, the plot of the error contributes to valuable information regarding the training process:

![Figure 1. Training (green) and checking (red) of error curve.](image)

The error curves for 40 epochs of ANFIS training are illustrated in the figure above. Specifically, the green curve represented the training errors, while the red curves referred to the checking errors. The minimal checking error could be seen at epoch 20, which was represented by a circle. It could be seen that the checking error curve increased after 10 epochs and dampened, implying data good fitting
through further training and good generalisation. Notably, the ANFIS information in this model included the 34 nodes, 32 linear parameters, 18 nonlinear parameters, 50 overall parameters, 97 training data pairs, 42 checking data pairs, and eight fuzzy rules. The minimal training of root mean square error (RMSE) amounted to 0.07.

3.6. **ANFIS model vs linear regression**

By checking the performance of the ANFIS model with a linear regression model was examined as a result, a comparison could be made between ANFIS prediction and a linear regression model through the contrast between the respective RMSE values and data checking. It could be observed that the value of RMSE against the data checking for ANFIS amounted to 0.168, while the linear regression amounted to 0.036 and training RMSE equal to 0.045, indicating that ANFIS model had a good performance compared to the linear regression model regarding the fitting of the data.

3.7. **Analysis of the ANFIS model**

It was indicated by the ANFIS model performance that the Reduct of the fuzzy rules was enhanced with prediction time. Positive results were achieved from this model compared to the regression when the checking and training of data were performed. The comparison among various modelling approaches. The modelling of linear regression spends the least amount of to reach the underfitting with limited number of samples or size of data and the ANFIS model-based RST and quantile discretization method takes the most amount of time to reach the best precision. In the words, if fast and easy design with low number of data the goal, then linear regression is the right choice. But if precious with is the utmost concern, then we should go with ANFIS, which is designed for big data modelling and higher precision. Figure 2 show the scatterplot of the actual versus prediction of RVP and the checking and training of data, including the ANFIS prediction of RVP.

![Figure 2. Comparison of predicted and actual for ANFIS modelling in training set A and checking set B.](image-url)
The figure 2 for both set A and Set B demonstrates that the model performance is accurate in the training phase where all most of the predicted data fall on the actual data. Meanwhile, a satisfactory model performance for testing phase.

4. Soft sensor and Industry 4.0
Following the emergence of Industry 4.0 and big data movement which increased the industrial motion, the manufacturing has now achieved unique opportunities in terms of the key enablers to enhance its performance to a new level. Notably, the model suggested in this study was able to manage a large amount of data. Meanwhile, as the embedded theory of the ANFIS model, RST could control the fuzzy rules set to transform millions of data, which stored a large amount of information, into a limited amount of data with rich knowledge. Provided that the neural network was able to improve the learning and computational ability of the suggested model in this study, it was an ideal choice for the calculation of the number of fuzzy rules. These fuzzy rules were converted from an expert’s knowledge and prediction of the targeted output.

5. Summary
Soft sensors are the trends in the various industrial field and large plants due to their potential applications despite several issues, including difficulty in measuring process parameter, delay in measurement, severe working environment for the evaluation of device survival, and difficulty in managing the requirements. Furthermore, machine learning techniques and theories reduced the difficulty in soft sensors easier, making them suitable in a severe environment. However, oil refinery, which represented the complex manufacturing processes, consisted of many complex parameters in terms of measurement. Moreover, API and RVP were two different parameters, which indicated the quality of light naphtha from the refining stage, and these parameters were estimated using statistical and machine learning techniques. The two methods displayed positive results in terms of their predictions of the two parameters, in which each parameter displayed different performance of API and RVP when the linear regression managed a low amount of data. It was also found through ANFIS that the soft sensor model based on the rough set theory and quantile discretisation method displayed high performance to function with a large amount of data. Finally, the contribution of this study could be seen from how it filled the gap between the industries by sharing data processes with an aim to progress towards Industry 4.0 in oil refinery industries.

References
[1] P. Kadlec, B. Gabrys, and S. Strandt, “Data-driven Soft Sensors in the process industry,” Comput. Chem. Eng., vol. 33, no. 4, pp. 795–814, 2009.
[2] D. Devogelaere, M. Rijckaert, O. G. Leon, and G. C. Lemus, “Application of feedforward neural networks for soft sensors in the sugar industry,” Proc. - Brazilian Symp. Neural Networks, SBRN, vol. 2002-Janua, pp. 2–6, 2002.
[3] I. Morsi and L. M. El-Din, “SCADA system for oil refinery control,” Meas. J. Int. Meas. Confed., vol. 47, no. 1, pp. 5–13, 2014.
[4] J. Macias-Hernandez, P. Angelov, and Z. X., “Soft Sensor for Predicting Crude Oil Distillation Side Streams using Takagi Sugeno Evolving Fuzzy Models.,” Int. Conf. Syst. Man, Cybern., vol. 44, no. 1524, pp. 3305–3310, 2007.
[5] B. Balasko, S. Nemeth, G. Nagy, and J. Abonyi, “Integrated process and control system model for product quality control-application,” Chem. Prod. Process Model., vol. 3, no. 1, p. Article ID: 50, 2008.
[6] P. Kadlec and B. Gabrys, Soft sensors: Where are we and what are the current and future challenges?, vol. 2, no. PART 1. IFAC, 2009.
[7] S. Sensors, “( Efficient of soft sensors modelling in advanced control systems for improving the performance of processes and sensors in Advance manufacturing systems by applying hybrid Intelligent data mining technique ),” pp. 0–54.

[8] H. Sadoyan, A. Zakarian, and P. Mohanty, “Data mining algorithm for manufacturing process control,” Int. J. Adv. Manuf. Technol., vol. 28, no. 3–4, pp. 342–350, 2006.

[9] L. Fortuna, S. Graziani, A. Rizzo, and M. G. Xibilia, Soft Sensors for Monitoring and Control of Industrial Processes. 2007.

[10] Y. Chang, Z. Lv, F. Wang, Z. Mao, and X. Wang, “Soft sensing modeling based on stacked least square-support vector machine and its application,” Proc. World Congr. Intell. Control Autom., vol. 1, pp. 4846–4850, 2006.

[11] A. Rogina, I. Si, I. Mohler, and N. Bolf, “Soft sensor for continuous product quality estimation (in crude distillation unit),” Chem. Eng. Res. Des., vol. 9, no. June 2010, pp. 2070–2077, 2011.

[12] L. Fortuna, S. Graziani, and M. G. Xibilia, “Comparison of soft-sensor design methods for industrial plants using small data sets,” IEEE Trans. Instrum. Meas., vol. 58, no. 8, pp. 2444–2451, 2009.

[13] J. Fernandez de Canete, P. Del Saz-Orozco, R. Baratti, M. Mulas, A. Ruano, and A. Garcia-Cerezo, “Soft-sensing estimation of plant effluent concentrations in a biological wastewater treatment plant using an optimal neural network,” Expert Syst. Appl., vol. 63, pp. 8–19, 2016.

[14] J. X. Luo and H. H. Shao, “Developing soft sensors using hybrid soft computing methodology: A neurofuzzy system based on rough set theory and genetic algorithms,” Soft Comput., vol. 10, no. 1, pp. 54–60, 2006.

[15] N. S. Ali, K. Mohd-Yusof, M. F. Othman, R. A. Latip, and M. S. N. Ismail, “Adaptive Neuro Fuzzy Inference System (ANFIS) modelling for quality estimation in palm oil refining process,” J. Mech. Eng., vol. 8, no. Specialissue1, pp. 36–47, 2019.

[16] M. R. Saybani and T. Y. Wah, “Data mining and data gathering in a refinery,” Int. Conf. Appl. Comput. Sci. - Proc., no. October 2010, pp. 62–66, 2010.

[17] M. S. Reis and G. Gins, “Industrial process monitoring in the big data/industry 4.0 era: From detection, to diagnosis, to prognosis,” Processes, vol. 5, no. 3, 2017.