

# A Hybrid Intelligent Method for Predicting Gasoline Octane Number and Optimising Operation Parameters

Jian Du<sup>b</sup>, Jianqin Zheng<sup>b</sup>, Yongtu Liang<sup>b</sup>, Qi Liao<sup>b</sup>, Bohong Wang<sup>a,\*</sup>

<sup>a</sup> National-Local Joint Engineering Laboratory of Harbor Oil & Gas Storage and Transportation Technology/Zhejiang

Provincial Key Laboratory of Petrochemical Pollution Control, Zhejiang Ocean University, Zhoushan, 316022, PR China

<sup>b</sup> National Engineering Laboratory for Pipeline Safety/ MOE Key Laboratory of Petroleum Engineering/ Beijing Key

Laboratory of Urban Oil and Gas Distribution Technology, China University of Petroleum-Beijing, Fuxue Road No. 18,

Changping District, Beijing 102249, PR China

wangbh@zjou.edu.cn

The Research Octane Number (RON) is a significant indicator for reflecting the combustion performance of gasoline, and the products of gasoline combustion provide a major influence on the atmospheric environment. It is crucial to measure RON during the process of gasoline refining. However, current methods generally use the instrument in the laboratory to measure RON, which is time-consuming and expensive. This paper proposes a hybrid intelligent method for predicting RON and optimising operating parameters to reduce RON loss, aiming to improve combustion performance and reduce pollutant emissions. Considering that the feature engineering and model optimisation are time-consuming, the RON prediction model is generated and optimised automatically via automatic machine learning technique. To reduce RON loss, the particle swarm optimisation algorithm is applied to optimise the operation parameters based on predicted RON. Subsequently, a real-world industrial dataset is taken as an example to test the accuracy and effectiveness of the proposed model. The results suggest that the proposed model is more accurate than other advanced models, such as Support Vector Machine (SVM) and Decision Tree (DT), with mean absolute error and root mean squared error being 0.162 and 0.225, 46 % and 44 % reduction of the SVM and DT. The experiment also shows that the RON loss of data samples is reduced by more than 40 % after optimisation.

## 1. Introduction

### 1.1 Background

As one of the most widely utilised petroleum products, gasoline provides the fuel supply for small vehicles (Dabbagh et al., 2013). With the rapid development of society and economic, the gasoline demand increases continuously in recent years (Demirbas, 2007). The increase in gasoline consumption has also brought serious environmental problems, caused by the exhaust emissions from gasoline combustion (Cui et al., 2021). Stringent gasoline quality standards have been increasingly formulated by many countries (Pan et al., 2018). Many countries supply domestic crude oil dependent on imports, particularly China. In China, there are about 40 %-60 % of imported crude oil is heavy oil, which contains sulphur and olefins and is hard to employ directly. For this reason, China has rapidly developed lightweight technology to convert heavy oil into gasoline, diesel, and light olefins.

The research octane number (RON) is the most significant indicator which reflects the combustion performance of gasoline. Each unit reduction of RON results in an economic decrease of 22 \$/t. In China, the catalytic cracking has produced more than 70 % of gasoline and brings more than 95 % of sulphur and olefins in refined gasoline. As a consequence, it is necessary to refine the catalytic cracking gasoline to reduce the content of sulphur and olefins in gasoline. However, the RON is decreased while reducing the sulphur and olefins in the process of catalytic cracking.

To improve the quality of refined gasoline, catalytic gasoline adsorption desulphurisation technology (S-Zorb) is obtained. Nevertheless, the RON of refined gasoline is difficult to require due to the complexity of operation conditions and measurement equipment. Given this circumstance, it is worthwhile to develop an accurate model

for the gasoline refining process to predict the RON of refined products. The refining process is complex and measurement is diversified, which leads to the high nonlinearity and strong coupling of involved control variables. The historical operating data of the industrial refining process is under-used.

In summary, the difficulty to model the refining process of gasoline can be concluded as follows:

- (1) The refining equipment contains many important control variables, the influence of control variables on the RON of refined products is hard to quantify.
- (2) It is difficult to establish an accurate model to illustrate the refining process due to the high nonlinearity and strong coupling relationship of control variables.

## 1.2 Literature review

In the past decades, a lot of research has been conducted to model the industrial process of gasoline refining. Among these, the infrared spectroscopy (IR) technique, which has become the most common technique, is combined with chemometric algorithms for remote monitoring and acquiring high information content of spectra. Kelly et al. (1989) applied multiple linear regression (MLR) and partial least squares (PLS) to demonstrate the possibility of predicting RON from the IR data on the 660-1,215 nm spectral range. Based on selected wavenumbers from the mid-IR spectrum, Rashid et al. (1989) proposed a predicting model to determine clear RON, the heat of formation, gravity, and mole fraction. Parisi et al. (1990) employed Near-IR in combination with multivariate calibration to measure online gasoline characteristics, such as RON, cetane number, paraffin, and olefin. Though the traditional model has obtained certain results, the prediction results show less satisfactory due to the high nonlinearity and strong coupling relationship of control variables.

In the context of industrial 4.0 and Big Data, many researchers show increasing interest in the huge amount of industrial process data (Dias et al., 2020). The excellent nonlinear approximate function of machine learning can be used to model complex and nonlinear chemical processes. Doicin and Onutu (2014) proposed a prediction method using a neural network to predict the octane number of petroleum mixtures. However, the influencing factors were not included broadly in input properties. Based on the chemical features acquired from IR, Ibrahim and Farooq (2019) utilised an artificial neural network (ANN) and Principal Component Analysis (PCA) to predict octane number. Kubic et al. (2017) developed an ANN-based model on special chemical composition to predict the octane number of hydrocarbons and oxygenated. Though the machine learning models can better demonstrate the nonlinear relationship between control variables and RON, feature engineering is time-consuming and few researchers had provided the optimisation plan to increase the RON of refined products.

Despite the great progress made in the above-mentioned research, there are still some limitations summarised as follows:

- (1) The feature engineering is time-consuming due to lots of operation parameters in the refined process, and the process of optimising control variables further increases the complexity and difficulty of modelling.
- (2) Few researchers focused on process optimisation to increase the RON of refined products based on prediction models and intelligent algorithms.

To overcome the drawback of conventional chemical modelling of gasoline refining, this paper proposes a novel hybrid intelligent framework for RON predicting and process parameters optimisation. At first, the Pearson correlation coefficient (PCC) is calculated to select the input variables. Subsequently, based on the automatic machine learning (AML) technique, the RON prediction model is established. To increase the RON of refined products, the prediction model is combined with the PSO algorithm to optimise control variables.

## 2. Methodology

In this section, the hybrid framework is developed to predict RON and optimise operating parameters of the refined process. The fundamental theories of AML and PSO are briefly introduced initially. Then a hybrid framework in combination with the RON prediction model and parameters optimisation method is proposed.

### 2.1 Basic principles

#### 2.1.1 Automatic machine learning

Despite the advancement obtained in the field of machine learning, the model construction process is time-consuming and the model performance is related to the goodness of parameters optimisation (Zhao et al., 2021). Given this circumstance, the AML technique is proposed to perform repetitive works, such as model selection, model combination, and model optimisation, aiming to improving modelling efficiency (Balaji and Allen, 2018). In this paper, AML technique is applied for automatic model construction. To find the best pipeline for the research dataset, genetic programming is used to automate the most tedious part of machine learning.

### 2.1.2 PSO algorithm

PSO algorithm is first proposed by Eberhart and Kennedy in 1995 (Poil et al., 2007) and has received increasing awareness and application in recent years. In this work, PSO is applied to optimise the operation parameters of the refined process. At first, the size of the population is determined to form the first generation. The dimensional of each individual is determined based on the input features of the RON predicted model. Then, the RON loss (Eq (1)), which is the difference between the RON of material gasoline and the RON of product gasoline, is calculated to conduct a fitness evaluation. The optimal individual and the optimal population value are acquired to update the particles relying on the less RON loss. To this end, the best individual of the group will be found, as the best value will converge within the maximum iteration. The operation parameters are optimised by it.

$$RON_{loss} = RON_{material} - RON_{product} \quad (1)$$

where  $RON_{material}$  represents the RON of refining material, which is the catalytic cracking gasoline.  $RON_{product}$  is the RON of gasoline products.

### 2.2 A hybrid framework for RON prediction and parameters optimisation

The main process of the proposed hybrid framework is depicted in Figure 1. The historical operating data of S-Zorb equipment is collected to form a database. The operating data is pre-processed to select input control variables, the output variable is product gasoline's RON. The process of data pre-processing mainly contains data dimensionality reduction, data normalisation, and data division. Then the prediction model is generated by automating model selection, model combination, model optimisation, and the historical operating data is used to train the model. After the training process is finished, the RON predicted model is constructed. Then, the real-time operating parameters are obtained to select input control variables for the predicted model, and the RON of refined products is estimated through it.

Based on the prediction model and intelligent algorithm, the operating parameters are optimised by maximising RON loss reduction. To increase the RON of refined products to improve economic benefits, the initial values of control variables are utilised to initialise the particle swarm. After several iterations, the best value of operating parameters and RON loss reduction tend to converge. Eventually, the operating parameters of the refining process are acquired to decrease RON loss.

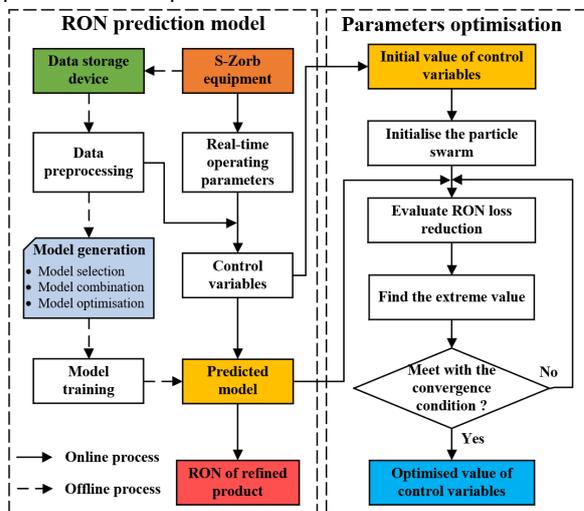


Figure 1: The hybrid intelligent framework of this work

### 2.3 Evaluation metrics

To evaluate the performance of prediction models, several evaluation metrics are applied, such as mean absolute percentage error (MAPE), root mean square error (RMSE), mean absolute error (MAE), and the R-squared ( $R^2$ ). The formulations of these metrics are as follows:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100, \% \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (\bar{y}_i - y_i)^2} \quad -\infty \leq R^2 \leq 1 \quad (5)$$

Where  $y_i$ ,  $\hat{y}_i$  and  $\bar{y}_i$  are observed values, predicted values, and average values of RON product.

### 3. Case study

In this section, a real-world industrial dataset is established by collecting operating data of S-Zorb refining equipment. Then, the performance of the proposed hybrid intelligent framework is verified by testing the accuracy of the RON prediction model and optimising the operating parameters to increase product RON.

#### 3.1 Data description

The historical operating data accumulated over 4 y from the S-Zorb refining equipment of a petrochemical corporation in China is collected for experiments. The dataset contains the following information: 7 material properties, 2 raw adsorbent properties, 2 regenerated adsorbent properties, 2 product properties, and 354 operation parameters.

Before constructing the proposed framework, it is necessary to carry out data pre-processing, which includes data normalisation and data splitting. In this paper, the Min-Max normalisation method (as shown in Eq (6)) is employed to normalised the input data to values between 0 and 1.

$$\tilde{x}_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (6)$$

where  $x_i$  is the input original value,  $\tilde{x}_i$  is the normalised value.  $x_{\min}$  and  $x_{\max}$  denote the minimum and maximum value of the input original value. After normalising the input data, the data splitting is conducted to ensure the robustness of the proposed framework. In general, the input data will be divided into two data sets, namely, 80 % for training the prediction model (260 data samples), and 20 % for verifying the prediction performance after the training process is finished (65 data samples).

To predict the product RON, the operation variables ought to be analysed to determine the main control variables. The PCC between operation variables and product RON is calculated for dimension reduction. A total of 16 variables with high correlation ( $PCC \geq 0.4$ ) are selected as the input variables. To this end, the input data is prepared for model construction.

#### 3.2 RON prediction via the hybrid framework

After acquiring the input data for RON prediction, the experiments are conducted ten times to obtain the average predicted errors of different models. The programming environment of the prediction models is Python. The model construction of ANN, DTR, and SVR are optimised by the means of trial and errors. For ANN, the number of hidden layers is set as 2, the number of nodes is set as 15, 20, learning rate is 0.001, and Adam optimiser is used. The optimal model constructed by AML technique is an ensemble of extreme gradient boost, stochastic gradient descent, and k-neighbour. The predicted errors of different models are shown in Table 1. By comparing the results predicted by SVR and DTR, it can be determined that the DTR obtains the worst prediction performance and the highest predicted errors. It is suggested that the DTR is not suitable for RON prediction. From the RON curves shown in Figure 2, it can be seen that the RON predicted by ANN are closer to the observed RON, and the predicted errors of ANN are lower than that of SVR. The proposed RON prediction method constructs the model automatically, and the model parameters are optimised by using genetic programming. This not only reduces the time cost in feature engineering and parameters optimisation, also provides better performance to the proposed model, which acquires the lowest predicted errors compared to other machine learning models and previous work. The RON results predicted by the proposed model are much closer than that of other comparative models.

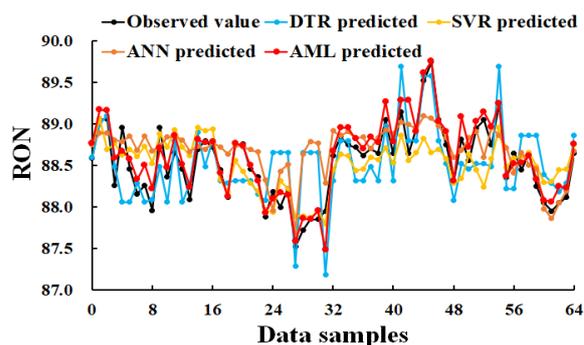


Figure 2: The prediction results of different models

Table 1: The predicted errors of different models

Metrics	AML	ANN	SVR	DTR	Previous work (Li and Qin, 2021)
MAPE / %	0.207	0.310	0.328	0.347	0.305
RMSE	0.225	0.344	0.390	0.374	0.335
MAE	0.175	0.274	0.290	0.307	-
$R^2$	0.934	0.785	0.445	0.258	0.860

### 3.3 Parameters optimisation for increasing product RON

The decrease in product RON brings economic loss and reduces the combustion performance. To improve the product RON, the control variables are adjusted to increase product RON. At first, the original values of control variables are used to initialise the particle swarm, and the percentage of RON loss reduction (as shown in Eq (7)) is set to optimise the control variables. The experimental results of the three cases are depicted in Figure 3 and Table 2. For case A, the original value of product RON is 88.65. To increase the product RON, the reduction percentage is set as 40 % to obtain the expected optimised RON, as shown in Eq (8). With the optimisation of operation variables, the product RON is increasing gradually. When the objective function tends to converge, the optimised product RON is obtained.

Similar to case A, the experiments of cases B and C set the product RON reduction as 50 % and 60 %. The comparisons of original values and optimised values are shown in Table 2. Each operation variable is adjusted within its corresponding value range. Through the parameter optimisation process, the operation variables are optimised to increase the product RON.

$$\text{reduction} = \frac{RON_{\text{optimised}} - RON_{\text{product}}}{RON_{\text{material}} - RON_{\text{product}}} \quad (7)$$

$$\text{Objective function} = \min(|RON_{\text{optimised}} - RON_{\text{expected}}|) \quad (8)$$

where  $RON_{\text{optimised}}$  is the optimised product RON.  $RON_{\text{expected}}$  is the expected optimised product RON.

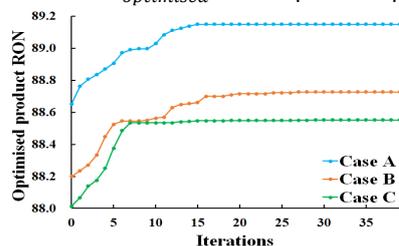


Figure 3: The iteration curves of product RON for these three cases

## 4. Conclusions

The accurate and efficient intelligent model for petrochemical refining process analysing is of great significance for improving the combustion performance and reducing pollutant emissions. This paper proposes a novel intelligent framework for RON prediction of refining products and increasing product RON by optimising operation parameters. The proposed framework overcomes the shortcoming of conventional prediction methods which are time-consuming and have poor accuracy in manual optimisation and fills the gap in operation

parameters optimisation. Compared to other advanced models, such as ANN, SVR, and DTR, the proposed framework achieves better prediction results, with MAPE being 0.207 %. Through the iterative optimisation of intelligent algorithms, the RON loss reductions of three cases are improved by at least 40 % higher. Applied in the industrial process of gasoline refining, the proposed method can estimate the product RON in real-time accurately. To further improve the economic efficiency and combustion performance, the operators can employ the proposed framework to increase product RON.

Table 2: The optimisation results of model variables

Model variables		Case A		Case B		Case C	
		Original values	Optimal values	Original values	Optimal values	Original values	Optimal values
Operation variables	S-ZORB.FC_1203.PV	9.86 (t/h)	11.22 (t/h)	10.41 (t/h)	13.02 (t/h)	10.00 (t/h)	12.11 (t/h)
	S-ZORB.FT_1004.PV	72.24 (t/h)	77.03 (t/h)	75.09 (t/h)	58.63 (t/h)	76.54 (t/h)	79.81 (t/h)
...	...	...	...	...	...	...	...
Product property	Product RON	88.65	<b>89.15</b>	88.12	<b>88.67</b>	87.95	<b>88.54</b>

### Nomenclature

RON<sub>material</sub> – RON of refining material

RON<sub>product</sub> – RON of refining products

RON<sub>optimised</sub> – The optimised product RON

y<sub>i</sub> – Observed RON

x<sub>i</sub> – The input original values

x<sub>max</sub> – The maximum of the input value

x<sub>min</sub> – The minimum of the input value

### Acknowledgments

This work was funded by the Science Foundation of Zhejiang Ocean University (11025092122).

### References

- Ibrahim A.E., Farooq A., 2019. Octane prediction from infrared spectroscopic data. *Energy & Fuels*, 34, 817-826.
- Balaji A., Allen A., 2018. Benchmarking automatic machine learning frameworks. *arXiv preprint arXiv:1808.06492*.
- Cui S., Qiu H., Wang S., Wang Y., 2021. Two-stage stacking heterogeneous ensemble learning method for gasoline octane number loss prediction. *Applied Soft Computing*, 113, 107989.
- Dabbagh H.A., Ghobadi F., Ehsani M.R., Moradmand M., 2013. The influence of ester additives on the properties of gasoline. *Fuel*, 104, 216-223.
- Demirbas A., 2007. Importance of biodiesel as transportation fuel. *Energy Policy*, 35, 4661-4670.
- Dias T., Oliveira R., Saraiva P., Reis M.S., 2020. Predictive analytics in the petrochemical industry: Research Octane Number (RON) forecasting and analysis in an industrial catalytic reforming unit. *Computers & Chemical Engineering*, 139, 106912.
- Doicin B., Onuțu I., 2014. Octane number estimation using neural networks. *Revista de Chimie*, 65, 599-602.
- Kelly J.J., Barlow C.H., Jinguji T.M., Callis J.B., 1989. Prediction of gasoline octane numbers from near-infrared spectral features in the range 660-1215 nm. *Analytical Chemistry*, 61, 313-320.
- Kubic Jr W.L., Jenkins R.W., Moore C.M., Semelsberger T.A., Sutton A.D., 2017. Artificial neural network based group contribution method for estimating cetane and octane numbers of hydrocarbons and oxygenated organic compounds. *Industrial & Engineering Chemistry Research*, 56, 12236-12245.
- Li B., Qin C., 2021. Predictive Analytics for Octane Number: A Novel Hybrid Approach of KPCA and GS-PSO-SVR Model. *IEEE Access*, 9, 66531-66541.
- Pan L., Liu P., Li Z., 2018. A discussion on China's vehicle fuel policy: Based on the development route optimization of refining industry. *Energy Policy*, 114, 403-412.
- Parisi A.F., Nogueiras L., Prieto H., 1990. On-line determination of fuel quality parameters using near-infrared spectrometry with fibre optics and multivariate calibration. *Analytica Chimica Acta*, 238, 95-100.
- Poli R., Kennedy J., Blackwell T., 2007. Particle swarm optimization. *Swarm Intelligence*, 1, 33-57.
- Rashid H.A., Dekran S.B., Fakhri N.A., Aziz H.J., Hamoudi N.A., 1989. Determination of several physical properties of light petroleum products using IR. *Fuel science & Technology International*, 7, 237-250.
- Zhao W., Zhang H., Zheng J., Dai Y., Huang L., Shang W., Liang Y., 2021. A point prediction method based automatic machine learning for day-ahead power output of multi-region photovoltaic plants. *Energy*, 223, 120026.