processes

MDPI

Article

PSO-Optimized Data-Driven and Mechanism Hybrid Model to Enhance Prediction of Industrial Hydrocracking Product Yields Under Data Constraints

Zhenming Li, Kang Qin *, Yang Zhang, Peng Yang, Yue Lou and Mingfeng Li *

Research Institute of Petroleum Processing Co., Ltd., Sinopec, Beijing 100083, China; lizhenming.ripp@sinopec.com (Z.L.); zhangyang.ripp@sinopec.com (Y.Z.); yangpeng.ripp@sinopec.com (P.Y.); louyue.ripp@sinopec.com (Y.L.)
* Correspondence: qinkang.ripp@sinopec.com (K.Q.); limf.ripp@sinopec.com (M.L.)

Abstract: The accurate prediction of hydrocracking product yields is crucial for optimizing resource allocation and improving production efficiency. However, the prediction of product flowrates in hydrocracking units often faces challenges due to insufficient data and weak correlations between input and output variables. This study proposes a hybrid framework combining a Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM) model, mechanism modeling, and Particle Swarm Optimization (PSO) to address these issues. The CNN-LSTM captures spatiotemporal dependencies in operational data, while the mechanism model incorporates domain-specific physical constraints. The hybrid model is structured in both series and parallel configurations, with PSO optimizing key hyperparameters to enhance its predictive performance. The results demonstrate significant improvements in prediction accuracy, with determination coefficients (R^2s) reaching 0.896 (kerosene), 0.879 (residue), 0.899 (heavy naphtha), and 0.78 (light naphtha). Shapley Additive Explanations (SHAP) and Mutual Information Coefficient (MIC) analyses highlight the mechanism model's role in improving feature interpretability. This study underscores the efficacy of integrating kinetics modeling, deep learning, and metaheuristic optimization for complex industrial processes under data constraints, offering a robust approach to enhance hydrocracking yield prediction.

Keywords: industrial hydrocracking; enhancing predictions; hybrid model; PSO-optimized

check for updates

Academic Editor: Iqbal M. Mujtaba

Received: 23 February 2025
Revised: 25 March 2025
Accepted: 28 March 2025
Published: 8 April 2025

Citation: Li, Z.; Qin, K.; Zhang, Y.; Yang, P.; Lou, Y.; Li, M. PSO-Optimized Data-Driven and Mechanism Hybrid Model to Enhance Prediction of Industrial Hydrocracking Product Yields Under Data Constraints. *Processes* **2025**, *13*, 1118. https://doi.org/10.3390/pr13041118

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Hydrocracking is capable of converting heavy oil into lighter products with higher values, such as naphtha, diesel, and kerosene [1–3]. Industrial hydrocracking processes are becoming more complex due to the continuous evolution of market requirements and consist of multi-stage tandem operations. Consequently, based on the intellectualization trends, modeling this complex reaction system has garnered widespread interest [4–8]. In industrial hydrocracking units, daily monitoring data only can be obtained; these data consist of macro reaction conditions which are insufficient for accurately predicting the product yield. Inaccurate production forecasts of industrial hydrocracking units can bring about many negative effects, including a loss of economic benefits, waste of resources, and disturbance of production plans. The kinetics model is constrained by insufficient data, relying on distillation range and density as foundational parameters. The monitored raw material properties cannot be fully incorporated into the kinetics model, making it challenging to achieve a precise prediction accuracy. Thus, it is necessary to develop a

new strategy to capture key parameter information efficiently to enhance the prediction precision of industrial hydrocracking products.

Data-driven models can reveal the hidden rules in data and effectively develop the model when the relationships between input and output are difficult to explain directly [7,9–12]. In the field of hydrogenation reaction, commonly used data-driven modeling methods consist of artificial neural networks, partial least squares, support vector machine, and ensemble learning [13–16]. Regarding the highly complex relationships between reactants and products, deep neural networks have been applied in the hydrocracking reaction system [17]. Therefore, the data-driven model is suitable for high-dimensional complex reaction systems, but the data-driven model also has the problem of poor interpretability. The mechanism model is based on the physical method and has the characteristics of strong interpretability and high stability. At the same time, some researchers have integrated data-driven models with mechanism models to enhance the accuracy and extrapolation of data-driven models [18–20]. The core principle of the integration model is comprehensive advantages by combining the scientific interpretability of the mechanism model and the computational efficiency of the data-driven model. The target of the data-driven model's calculation is taken from the error between the mechanism model's predictions and actual values. In the field of hydrocracking, an adaptive weighted hybrid model has been proposed to obtain a better extrapolation capability compared to other combination models [21]. The key to establishing a hybrid model is the ability to combine the advantages of the two model types [22].

The hyperparameters of data-driven models are numerous and usually difficult to adjust manually. Using certain optimization algorithms to optimize hyperparameters can ensure that the model has an optimal prediction effect [23,24]. The optimization algorithms can be divided into local optimization algorithms and global optimization algorithms. Global optimization algorithms are more conducive to obtaining the optimizing parameters in a data-driven model's strong adaptability and high robustness conditions. Global optimization algorithms typically adjust model key hyperparameters such as weights, learning rates, and dropout rates, which can enhance the model's predictive performance [25,26].

However, these studies mentioned above used sufficient input variable information to build the models. This specific solution strategy is lacking in terms of insufficient input variables and weak input–output connections in industrial hydrocracking units. In this paper, an industrial hydrocracker was modeled, with the constraint that it can only provide daily monitoring data. The main calculation steps are shown in Figure 1.

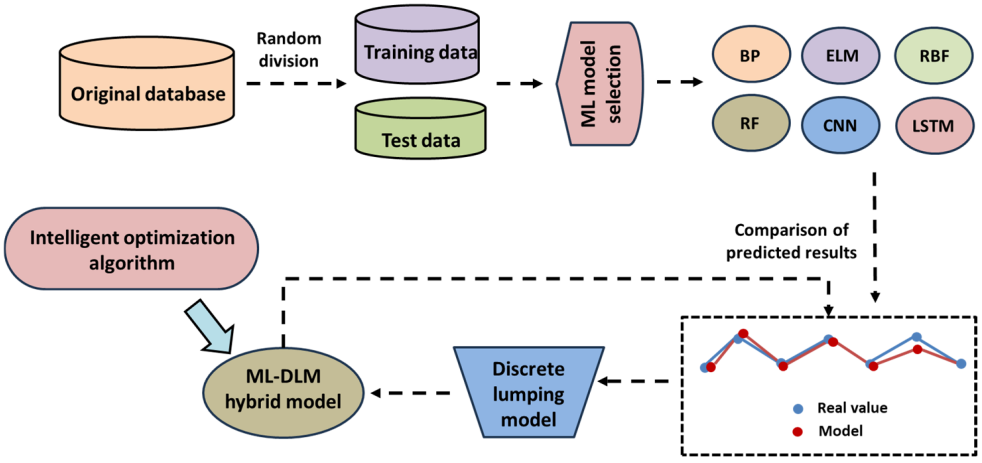


Figure 1. The design strategy of the computing framework.

- First, the collected data are divided into training and testing sets according to a certain ratio and then fed into various types of data-driven models, including Backpropagation Neural Networks (BPs), Extreme Learning Machine (ELM), Random Forest (RF), Radial Basis Function Neural Networks (RBFs), Convolutional Neural Network (CNNs), the Long Short-Term Memory (LSTM) model, and CNN-LSTM model. The optimal data-driven model structure was selected from multiple data-driven models;
- A mechanism model is established for industrial hydrocracking units based on the discrete lumped method to reinforce the learning pattern's ability from the input data. Subsequently, the mechanism model is combined with the optimal data-driven model to establish a hybrid model structure. To fully clarify the contribution of the mechanism model in the hybrid model, the results are interpreted through Shapley Additive Explanations (SHAP) analysis and Mutual Information Coefficient (MIC) analysis [27–29];
- Due to difficulty in obtaining the optimal hyperparameter values by manual adjustment, a PSO algorithm is introduced to optimize further important hyperparameters affecting the data-driven model prediction effect [30–32].

After the above steps, the issues of insufficient data availability and the low correlation between input and output variables in industrial hydrocracking units can be effectively addressed. This is achieved by selecting the optimal data-driven model, combining it with the mechanism model to form a hybrid model, and then further optimizing the hybrid model.

2. Materials and Methods

2.1. Data Preprocessing

The hydrocracking unit can only provide some daily monitoring data relevant to modeling. Some available input data were collected, including feedstock properties, operating variables, and catalyst information. This study focuses on an industrial hydrocracking unit with a capacity of two million tons per year. To ensure the generalization capability of the model, the data were collected from the unit over approximately one year. Reaction conditions (such as temperature, pressure, flow, etc.) were collected from the process control system (DCS), the properties of the feedstock (such as density, sulfur content, distillation range, etc.) were collected from the laboratory information management system, and the catalyst use time was obtained from the catalyst management system. These data were mined and invalid entries removed. Finally, a total of 13 relevant input variables were identified as shown in Table 1.

Table 1. Collected input variables from the hydrocracking unit.

Variable Categories	Number of Variables	Variables Name
Raw oil	4	Vacuum gas oil (VGO) ratio, feed flowrate, fresh hydrogen rate, density
Operating variable	8	Hydrogen partial pressure, inlet pressure, inlet temperature, initial distillation points of light naphtha, light and heavy naphtha cutting point, heavy naphtha and kerosene cutting point, kerosene and residue cutting point, final distillation point of residue
Catalyst	1	Catalyst use time

For model construction, factors are explored concerning outcomes influenced by the input variables [33]. The MIC method is used to assess the relationship between input variables and product yields. MIC is a data mining technique, which can enable the calculation of correlations between different features and analyze their contribution [34]. In the ranges of MIC value from 0 to 1, a higher value indicates a strong correlation between

the two variables, while a lower value suggests a weak correlation or even irrelevance. The MIC values between the product yield and the input variables is shown in Figure 2.

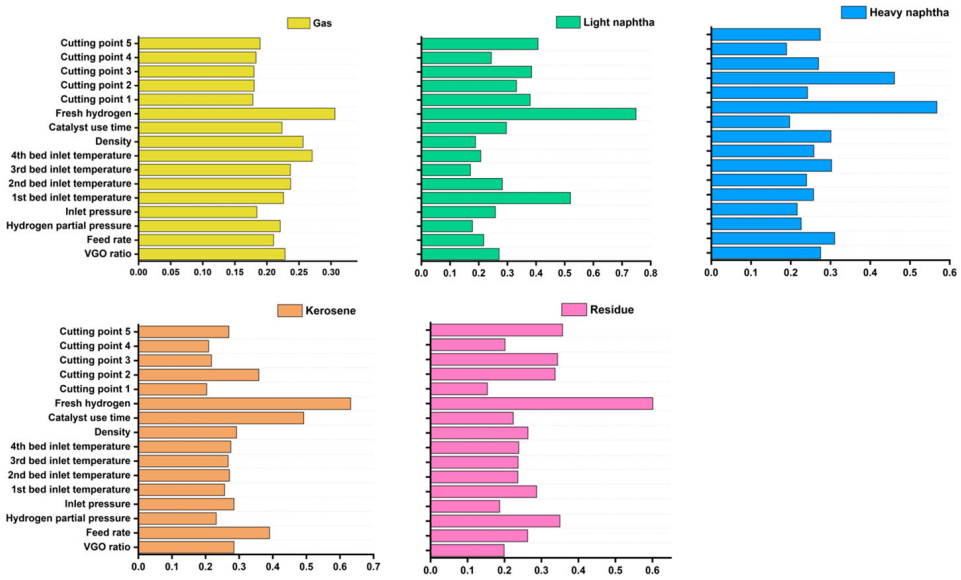


Figure 2. MIC values between the product yield and the input variables.

Based on the MIC analysis, the weak correlation between the model's input and output variables suggests that conventional modeling methods may not yield a highly accurate model with the given data. Consequently, it becomes imperative to adopt suitable measures to establish a robust and reliable model.

2.2. Model Construction and Evaluation

2.2.1. Data-Driven Model Introduction

The CNN is a feedforward neural network that includes convolutional operations and a deep structure. Convolutional Neural Networks adopt a layer-by-layer pre-training system, which can effectively extract spatial features from data through components such as the convolutional layer, pooling layer, and fully connected layer [17,35,36]. For the CNN, the height h_o and width w_o of the output can be denoted as

$$h_0 = (h_i - h_k + 2p)/s + 1 \quad (1)$$

$$w_0 = (w_i - w_k + 2p)/s + 1 \quad (2)$$

Among them, h_i and w_i are the height and width of the input data; h_k and w_k are the height and width of the convolutional kernel; p is the padding size; and s is the stride size.

LSTM networks are an essential and widely used algorithm for time series analysis. They belong to a special type of Recurrent Neural Network (RNN). Through the introduction of gating units, the LSTM effectively addresses the issues of long-term dependence and gradient disappearance that are commonly encountered in traditional RNN models [37–39].

Whereas an RNN has only one transfer state, the LSTM has two transfer states. $f(t)$, $i(t)$, and $o(t)$ are the input, forget, and output gates, respectively; $c(t)$ is the memory cell; $h(t)$ is hidden states; σg is a sigmoid function; and σc is a hyperbolic tangent function. The flow chart of LSTM is shown in Figure 3.

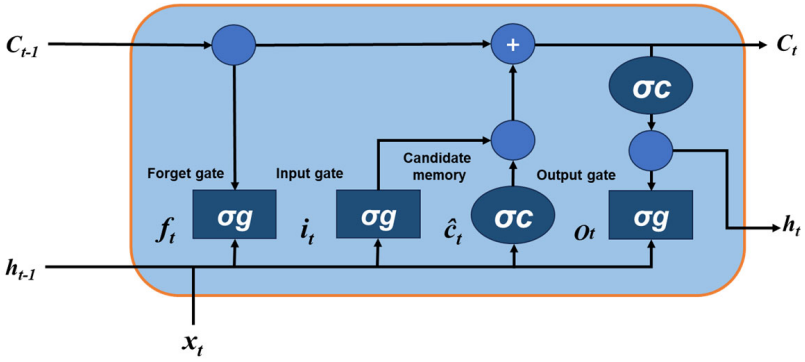


Figure 3. Typical architecture of LSTM networks.

The CNN-LSTM model is a powerful deep learning architecture that integrates the strengths of CNN and LSTM networks. This model seamlessly integrates the spatial feature extraction capabilities of CNNs and the temporal information capabilities of LSTMs. The CNN component is responsible for effectively processing the spatial features present in the input data, while the LSTM component focuses on capturing and learning the long-term dependencies within the time series data [37,40,41]. The basic structure of the CNN-LSTM is shown in Figure 4. Compared to other deep learning frameworks, Transformer performs well in time series modeling, but its high computational complexity and the large amount of data required for training may not be suitable for industrial data. A multi-layer perceptron (MLP) is suitable for processing static data, but it cannot effectively model spatiotemporal dependencies.

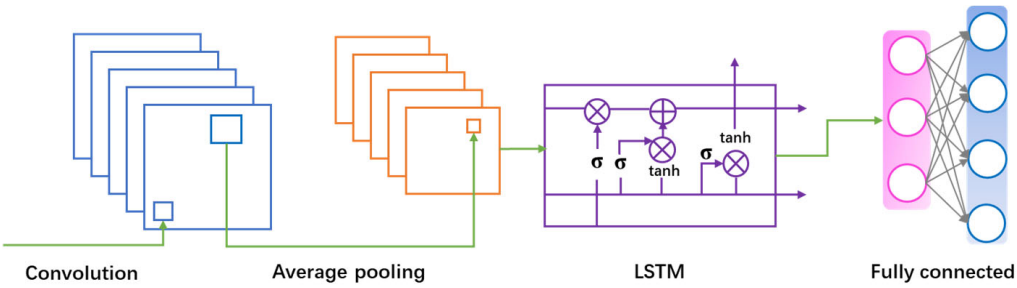


Figure 4. The basic structure of the CNN-LSTM.

2.2.2. Mechanism Model Introduction

Compared with the data-driven model, the advantage of the mechanism model is that it can quickly adapt to the refinery data with different cutting schemes, without the need for great model adjustments [42]. It is a challenge to analyze the complex feedstock composition in industrial units; employing a discrete lumped model is an effective approach to modeling this. The hydrocracking reaction system is divided into N components at intervals of 10 K from high to low true boiling point. Based on the relationship between actual reactions, we assume that components larger than 400 K (corresponding to carbon atom number 9) can be cracked into light components but cannot be cracked into sub-light components. Based on the above assumptions, the reaction network obtained is shown in Figure 5.

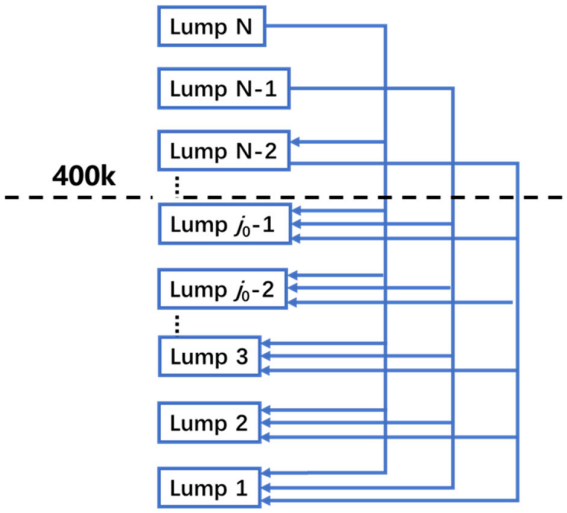


Figure 5. Reaction networks for mechanism model.

The discrete lumped model is based on reaction kinetics and reactor principles, which can provide stable and reliable modeling results. Material balance is essential for calculating hydrocracking product yields. The corresponding model calculation formulas are shown in Table 2. Utilizing the established reaction network, the reaction kinetics equations are formulated as Equations (3) and (4). Building on operational data from existing hydrocracking units, kinetic equations are developed based on the Stangeland model. The distribution coefficients $P_{i,j}$ critical modeling parameters directly influence the final product distribution. The remaining distribution coefficients for oil-related lumps are determined using Equations (9)–(11); B_1 and B_2 are oil-related distribution regulation parameters.

Table 2. Discrete lumping model equations.

Section	Equations	
Mass balance equations	$G\frac{dC_i}{dW} = r_i = -k_iC_i + \sum_{j=i+2}^{N} k_jP_{i,j}C_j, i = j_0, j_{0+1}, N$	(3)
	$G\frac{dC_i}{dW} = r_i = \sum_{j=j_0}^{N} k_jP_{i,j}C_j, i = 1, 2, j_0$	(4)
Kinetic equations	$k = A_he^{-E_h/RT}$	(5)
	$K_i = T_i + DT_i^3 - DT_i$	(6)
	$T_i = \frac{1.8Tb_i+32}{1000}$	(7)
Distribution coefficient equations	$P_{ij} = \frac{C\times e^{[-\omega(1.8tb_j-229.5)]}}{q}, i = 1, q$	(8)
	$P_{ij} = P'_{ij} - P'_{i-1,j}, i = q+1, N$	(9)
	$P'_{ij} = [y_{ij}^2 + (B_1y_{ij}^3 - B_2y_{ij}^2)] \times [1 - \sum_{i=1}^{q} P_{ij}], i = q+1, j-2; j = p+1, N$	(10)
	$y_{ij} = \frac{T_{bi}-2.5}{(T_{bi}-20)-2.5}, i = q+1, j-2; j = p+1, N$	(11)

Among them, B is the related product distribution parameter; C is the related gas distribution parameter; C_i is the fraction of the ith lump; D is the related boiling point parameter; G is inlet oil mass flow, kg/h; K_i is the rate coefficient related to the boiling point; k is the reference rate constant; k_i is the reaction rate constant of the ith lump, (kg/kg catalyst); $P_{i,j}$ is the partition coefficient; $P_{i,j}{}'$ is the sum of the q + 1th lump to the ith lump generated from the j lump; r_i is the reaction rate, (kg/(kg catalyst·h)); T_{bi} is the average boiling point, °C; W is the catalyst loading, ton; and y_{ij} is the normalized boiling point.

The single data-driven model can provide rapid results, but it may experience a significant increase in prediction errors when faced with data not present in the training set. To address this limitation, combining data-driven models with mechanism models generates many benefits. The combination allows for the inclusion of additional input variable information, thereby enhancing the predictive width of the model. Additionally, it can achieve more accurate prediction results, while mitigating the impact of varying response conditions between the training and testing sets.

2.3. Model Measure Metric

In the error evaluation of data-driven models, commonly used assessment metrics include Root Mean Square Error (RMSE), coefficient of determination (R^2), and Mean Absolute Error (MAE) [43–45]. These metrics are loss functions that use d to measure the difference between the predicted values and the actual target values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{12}$$

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \overline{y})^2} \tag{13}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{14}$$

Among them, y_i is the real value of hydrocracking production; $\hat{y}_i$ is the predicted value of hydrocracking production; and n is the total amount of hydrocracking production data in the test dataset. For each product, the deviations between the calculated values of the model and the actual values were calculated.

3. Results and Discussions

3.1. Prediction Effects of Data-Driven Models

The process and results of the parameter optimization of the data-driven model are shown in Tables A1 and A2. The collected hydrocracking product data were brought into the seven data-driven models to analyze and compare various assessment methods. The calculation process of the data-driven model is shown in Figure 6.

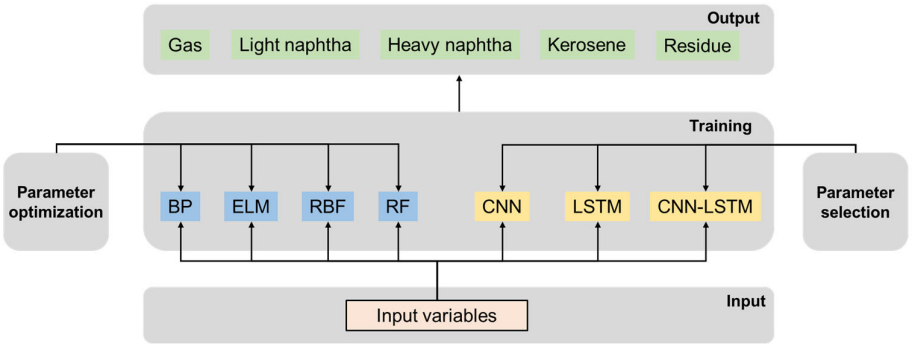


Figure 6. Data-driven model calculation process.

A total of 104 sets of operational data collected from industrial hydrocracking units were fed into the model and divided into 80 sets for training and 24 sets for testing. The input data, after being normalized and with the introduction of optimized and selected

parameters, were used for model-training iterations. The calculation results of various models on the test set are shown in Table 3.

Table 3. Comparison of prediction effect of data-driven models.

Project	Model	RMSE	R^2	MAE
Gas flowrate (t/h)	BP	1.04	−0.44	0.78
	ELM	0.82	0.097	0.68
	RBF	0.60	0.52	0.49
	RF	0.76	0.23	0.58
	CNN	0.80	0.15	0.44
	LSTM	0.67	0.41	0.51
	CNN-LSTM	0.74	0.27	0.47
Light naphtha flowrate (t/h)	BP	1.36	−0.27	0.97
	ELM	1.14	0.40	0.87
	RBF	0.89	0.45	0.69
	RF	0.95	0.38	0.73
	CNN	0.69	0.67	0.58
	LSTM	1.28	−0.13	0.92
	CNN-LSTM	0.69	0.68	0.49
Heavy naphtha flowrate (t/h)	BP	2.52	0.61	2.09
	ELM	3.31	0.086	2.46
	RBF	2.12	0.72	1.47
	RF	1.69	0.69	1.62
	CNN	2.39	0.64	1.35
	LSTM	1.94	0.76	1.49
	CNN-LSTM	1.91	0.77	1.04
Kerosene flowrate (t/h)	BP	3.54	0.20	2.74
	ELM	5.86	0.46	3.90
	RBF	1.83	0.79	1.43
	RF	2.06	0.73	1.48
	CNN	2.55	0.69	1.54
	LSTM	2.89	0.47	2.01
	CNN-LSTM	1.56	0.85	1.14
Residue flowrate (t/h)	BP	3.58	0.39	2.97
	ELM	5.24	0.14	3.69
	RBF	2.51	0.69	1.75
	RF	2.77	0.63	2.00
	CNN	2.55	0.69	1.55
	LSTM	3.55	0.39	2.56
	CNN-LSTM	2.04	0.80	1.44

Based on the analysis of RMSE, R^2, and MAE values, it can be concluded that the optimal data-driven model for gas prediction is RBF. For all other hydrocracking products, CNN-LSTM emerges as the optimal choice. Although the input data do not fully meet the prediction requirements of the model, relatively favorable results are achieved for heavy naphtha, kerosene, and residue. These results can be attributed to the learning of spatiotemporal information enabled by the CNN-LSTM model. The CNN component extracts spatial features, while the LSTM component processes time series data, thus improving the accuracy of the model.

According to the calculation results in Table 3 and Figure 7 and considering that the gas product exhibits a lower flowrate and is not the primary product of hydrocracking, the CNN-LSTM model was selected as the optimal data-driven model type in this paper. Subsequent optimization and adjustments will be built upon the CNN-LSTM model.

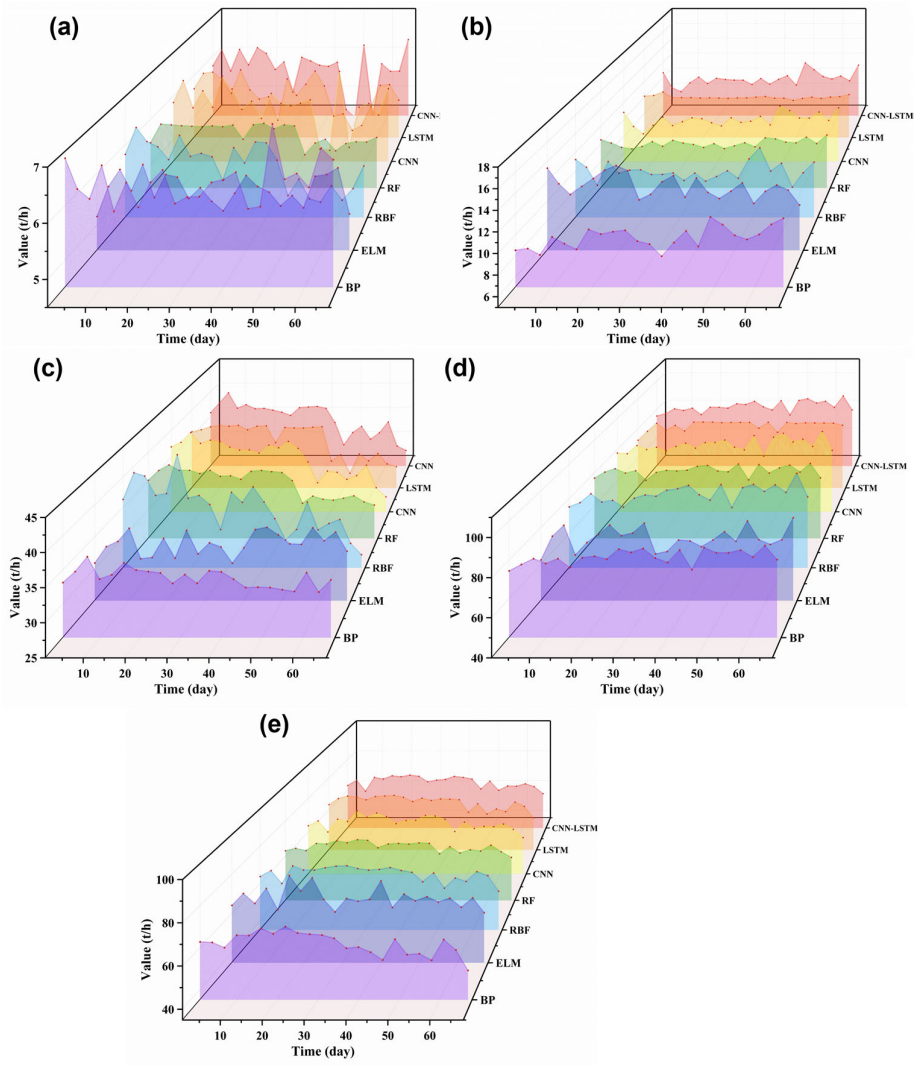


Figure 7. Comparison of the calculated value and the true value of the data-driven model test set: (**a**) Gas; (**b**) light naphtha; (**c**) heavy naphtha; (**d**) kerosene; (**e**) residue.

3.2. Combined Model of CNN-LSTM and Discrete Lumping

The mechanism model was solved by the Runge–Kutta method. And the quasi-Newton method and a Genetic Algorithm were used to find the accurate optimal parameter estimation. The objective function for optimization was the sum of the squares of the absolute errors between the calculated and actual values. The kinetic parameters obtained by optimizing the solution are shown in Table 4.

Table 4. The estimated kinetic parameters in mechanism model.

Kinetic Parameters		Value
Pre-exponential factor (vol feed/vol cat h)	A_1	7.27×10^7
	A_2	3.32×10^7
	A_3	1.26×10^7
	A_4	3.64×10^7
Activation energy (kJ/mol)	E_1	84.34
	E_2	84.88
	E_3	84.99
	E_4	84.01
$\Delta H_{H,con}$ (J/mol)		1.03×10^5

Multiple sets of data that differ from the parameter estimation were used to verify the reliability of the discrete lumped model. The verification results of the calculated and actual values are shown in Figure 8. The relative errors calculated after statistics are shown in Table 5. The calculation errors were basically less than 10% for products prediction, demonstrating the relatively high accuracy of the discrete lumped model. Compared with CNN-LSTM, the error of the mechanism model was still relatively high, especially for the prediction of heavy naphtha flowrate and residue flowrate, mainly due to the limited input information learned by the single mechanism model. A mechanical data-driven combined model can be established to combine the advantages of the two models.

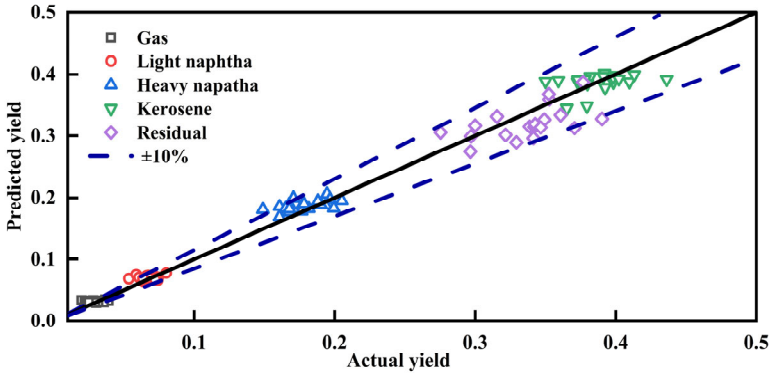


Figure 8. Discrete lumping model verification result.

Table 5. The relative errors of products flowrate predicted by the mechanism model.

Program (t/h)	Average Error (%)
Gas flowrate	10.57
Light naphtha flowrate	5.64
Heavy naphtha flowrate	7.16
Kerosene flowrate	2.70
Residue flowrate	5.49

By coupling the data-driven model and mechanism model, a new hybrid model network comprising 14 layers is proposed, including double convolutional layers, double pooling layers, LSTM layers, fully connected layers, and a data processing section. The hybrid model network is shown in Figure 9.

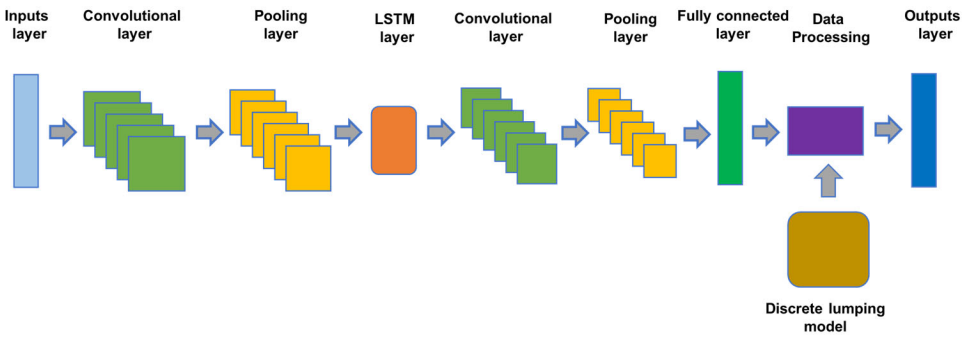


Figure 9. Hybrid model network coupling CNN-LSTM and discrete lumping model.

The main method to establish the mechanism and data-driven hybrid model was to further strengthen the transmission of input variable information. The above work brought the input variables into the CNN-LSTM model to obtain the calculated values. The calculated value of the mechanism model was obtained by entering the input variables

into the discrete lumped model, which contained the additional input variable information. The information processing of the input variables in the mechanism model differs from that in the data-driven model. The Convolutional Neural Network–Long Short-Term Memory Network–Discrete Lumping Model (CNN-LSTM-DLM) model was constructed by incorporating the mechanism model calculated results into the CNN-LSTM model. There are two main ways to combine the mechanism model and the data-driven model, which is shown in Figure 10. One method is that the calculation result of the mechanism model is taken as an input variable into the data-driven model to establish a series structure mode. The other method is that the residual between the calculation result of the mechanism model and the actual value is taken as the output variable of the data-driven model to establish a parallel structure model [20]. The initial values of the hyperparameters used by the CNN-LSTM-DLM are consistent with the CNN-LSTM model.

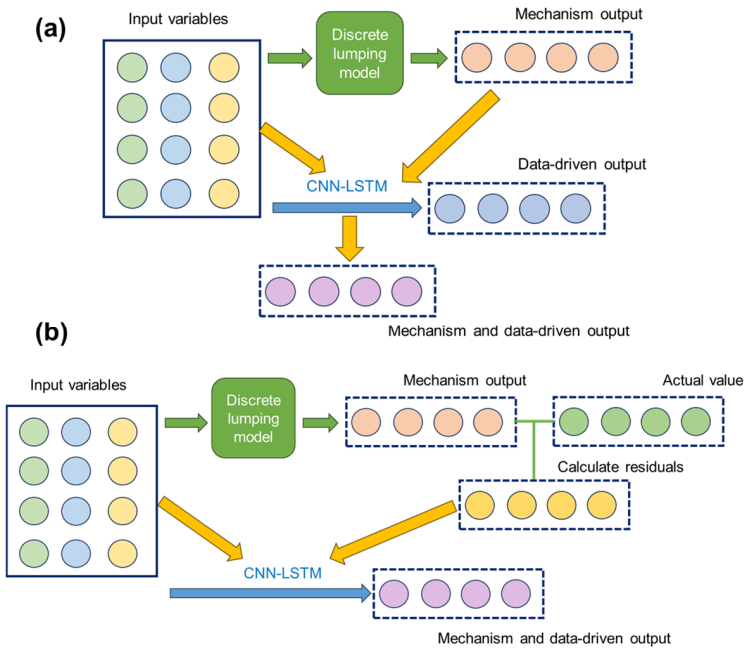


Figure 10. CNN-LSTM-DLM model construction process: (**a**) Series structure model; (**b**) parallel structure model.

The industrial hydrocracking data were brought into the established CNN-LSTM-DLM model, and its performance was compared with the single data-driven model to verify the effectiveness of the hybrid model. The correlation coefficient R^2 of the two types of models was compared and the results are shown in Figure 11.

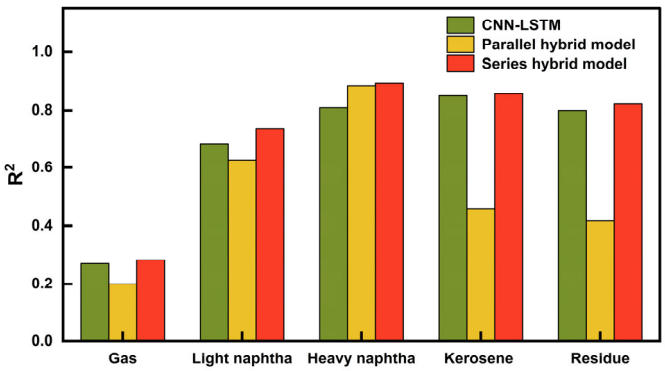


Figure 11. The correlation coefficient of the two types of hybrid models.

According to the preliminary calculation results, the prediction effect of the mechanism data-driven combined model in series form was significantly better than that of the single data-driven model and the parallel model. The poor prediction effect of the parallel model may be due to the large calculation bias of some mechanism model results, which leads to a deviation in the final fitting direction. Therefore, the series structure model is the mechanism data-driven hybrid model for subsequent research. The comparison of prediction effects is shown in Figure 12.

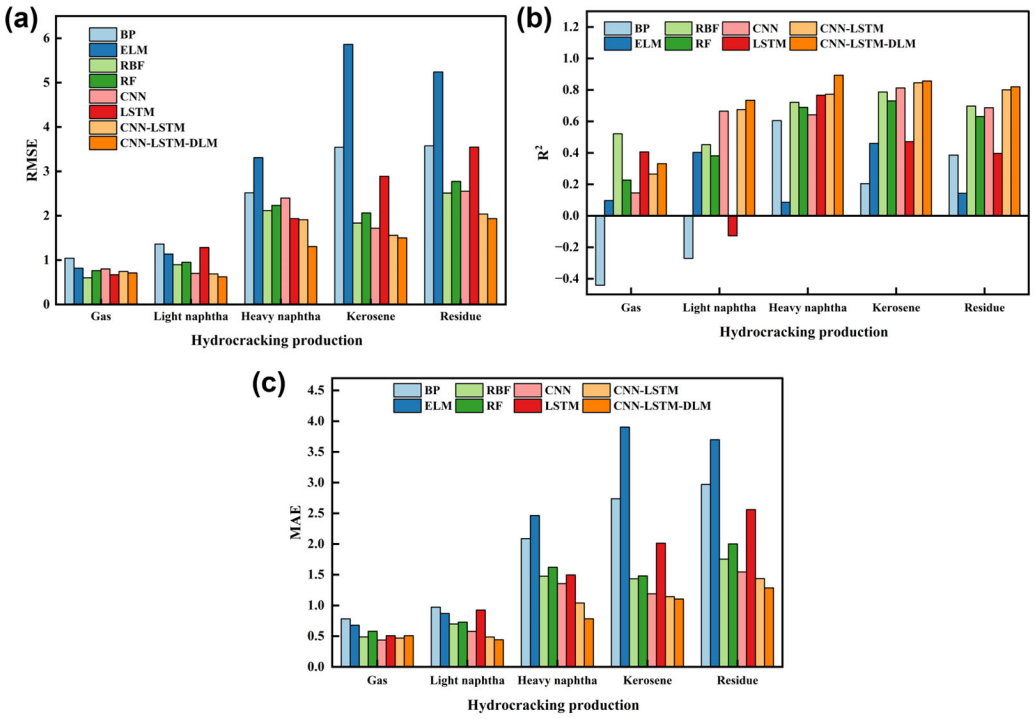


Figure 12. Comparison of prediction effect between CNN-LSTM-DLM model and data-driven model: (**a**) RMSE values; (**b**) R^2 values; (**c**) MAE values.

According to the RMSE values, R^2 values, and MAE values of the key indicators used in assessing the model performance, the accuracy of the established CNN-LSTM-DLM model has been significantly improved compared with the single data-driven model, mainly because the existing data information is extracted more effectively after the discrete lumped model is added. The improved accuracy of the prediction can help operators to more accurately adjust reaction conditions (such as temperature, pressure, amount of catalyst, etc.) to optimize the process. Meanwhile, the required training calculation time of various models is shown in Figure 13.

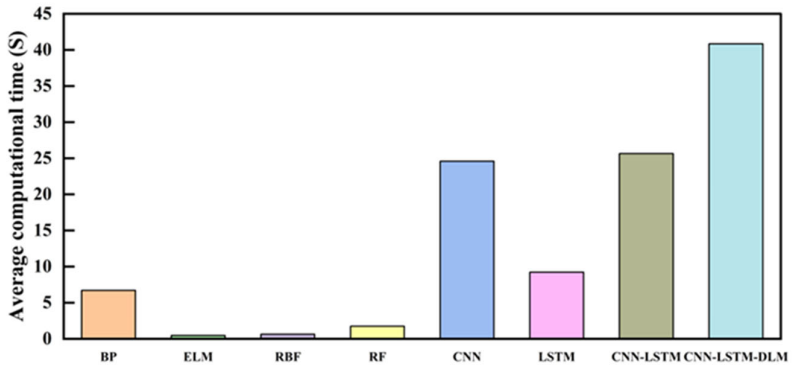


Figure 13. Comparison of training time for different models.

In the calculation cost part, the calculation time of the CNN-LSTM-DLM model is longer than the data-driven model. The computer configuration used is 2.30 GHz, and the calculation time of the training model is about 40 s. Although the computational time has increased, it still meets the efficiency requirements of real-time calculation and does not increase the calculation cost too much.

To clarify the impact of the mechanism model calculated values on the hybrid model results, a SHAP analysis and MIC analysis are conducted on the input variables of the CNN-LSTM-DLM model. SHAP is an interpretability method based on cooperative game theory, used to quantify the contribution of each feature to the model's predictions. It is grounded in Shapley values, which fairly allocate the impact of predictions by calculating the marginal contribution of features across all possible feature combinations. MIC is a statistical method for measuring nonlinear relationships between two variables. It is based on mutual information, which quantifies the reduction in uncertainty of one variable given knowledge of another. These analyses deepen the understanding of the relationship between the mechanism model calculated values and the hybrid model results, shedding light on the significance and role of these calculated values within the CNN-LSTM-DLM model.

Positive values indicate promotion (yellow points), and negative values indicate inhibition (blue points) in Figure 14. The newly introduced calculated value features from the mechanism model exhibit a highly significant influence on the final calculation results. The input features of the mechanism model calculation values were consistently among the top two most important features across all the variables. This result can be attributed to the fact that the mechanism model calculations effectively extract relevant information, causing a strong correlation with the product yield. The SHAP analysis provides evidence of the mechanism model's crucial role in the CNN-LSTM-DLM model calculations. This finding emphasizes the importance of the mechanism model and highlights that its integration contributes to the enhanced prediction accuracy achieved by the CNN-LSTM-DLM model. The MIC values predicted by the mechanism model calculation results were all greater than 0.2 (dashed line in Figure 14f) in the hybrid model, especially for light naphtha, kerosene, and residue, which were all above 0.6. It showed that the addition of the mechanism model improves the correlation of input and output variables and increases the feature information required for prediction. This was conducive to revealing the correlation of important data features and models and improving the prediction accuracy of the model. By applying a SHAP and MIC analysis to the CNN-LSTM-DLM model, insights are gained into how the calculated values from the mechanism model influence the overall predictions.

The prediction of reactor bed temperature is of great significance for improving the operation efficiency of the reactor, ensuring production safety, improving product quality, realizing energy savings, and reducing consumption. The prediction results and evaluation effects are shown in Figure 15. These established models were applied to the bed temperature prediction of the hydrocracking reactor. Compared with the prediction effect of hydrocracking product yields, various data-driven models and the CNN-LSTM-DLM model had better effects on the bed outlet temperature prediction. The main reason is that the outlet temperature is highly correlated with the inlet temperature, thus it is relatively easy to extract data features.

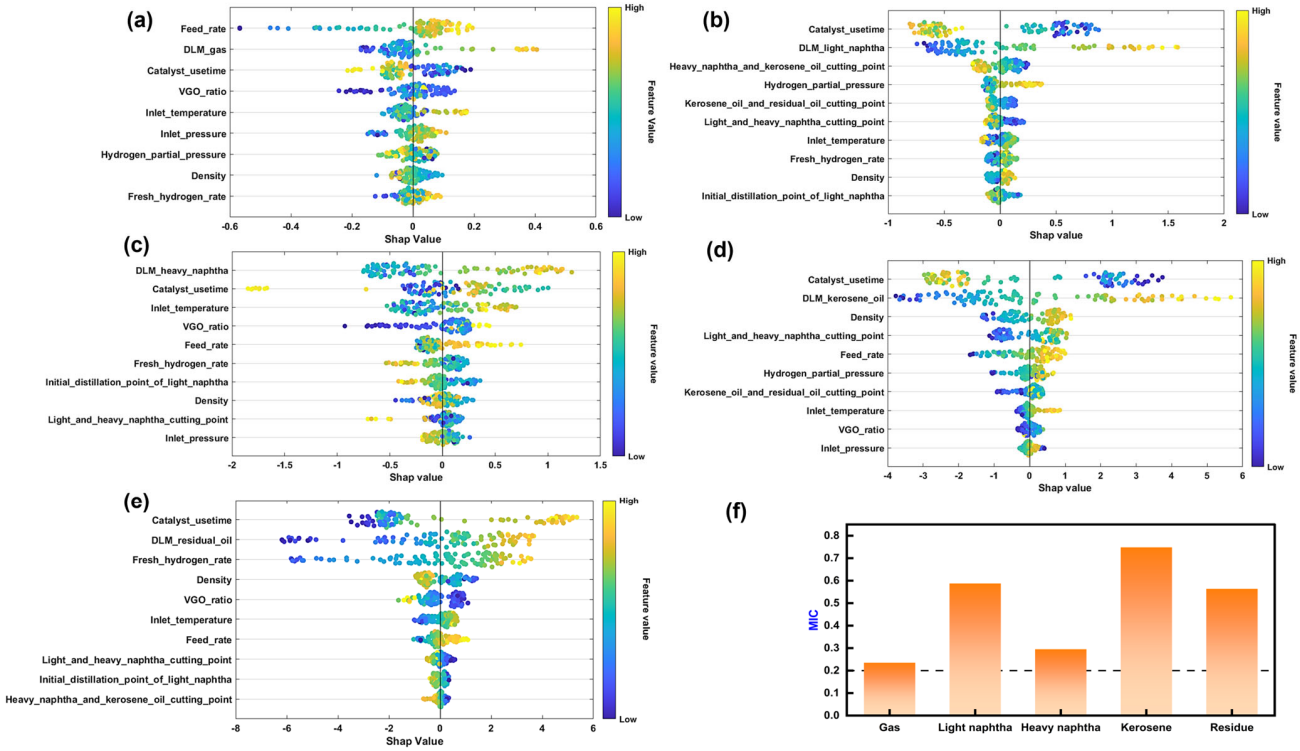


Figure 14. SHAP value of each input feature in the database: (**a**) Gas; (**b**) light naphtha; (**c**) heavy naphtha; (**d**) kerosene; (**e**) residue. (**f**) The influence of adding a discrete lumped model to predict target MIC values.

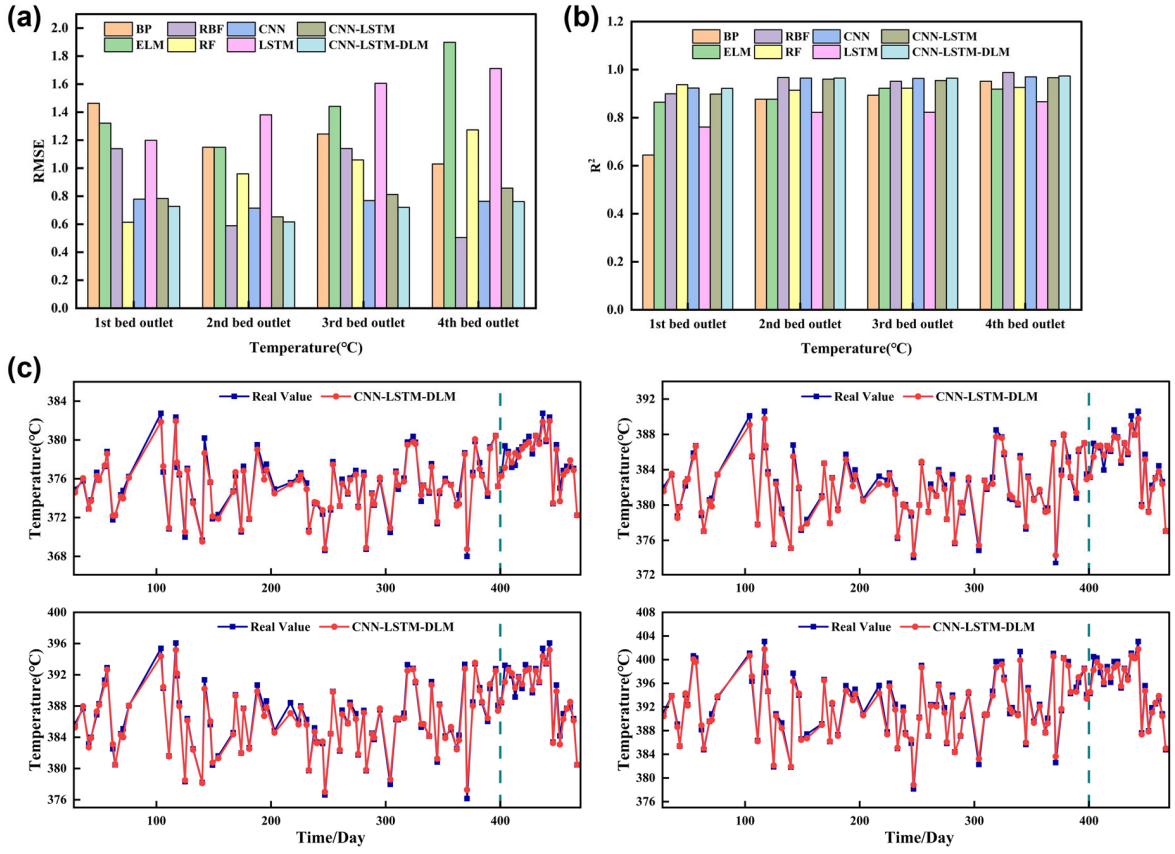


Figure 15. Comparison of prediction effect of different models on hydrocracking reactor bed outlet temperature: (**a**) RMSE; (**b**) R^2; (**c**) Compare the calculated values with the actual values.

By comparing the prediction effects of the established models, it can be found that the comprehensive effect of the established CNN-LSTM-DLM has the best performance. The CNN-LSTM-DLM model showed the minimum average RMSE value and the maximum average R^2 value, indicating that the hybrid model can maximize the extraction of information. The calculated results of the CNN-LSTM-DLM model have a good consistency with the actual values of the hydrocracking reactor bed temperature in both the test set and training set.

3.3. PSO Optimization Combination Model

The combination of the CNN-LSTM model and mechanism model greatly improves the prediction accuracy of the model, but the prediction effect is still not very good at some marginal points, especially for gas products prediction. One of the reasons is that the same learning rate, number of hidden neurons, regularization coefficient, and dropout rate was set for each calculation. These parameters have a significant impact on the fitting effect of the simulation, but they are difficult to adjust manually.

The Particle Swarm Optimization (PSO) algorithm is a heuristic optimization technique. The PSO algorithm is known for its simplicity in implementation and high computational efficiency, making it a popular choice in various fields. Over recent years, it has gained widespread adoption and has proved to be a valuable tool in solving complex optimization problems [46]. The specific process of PSO optimization and the hybrid model is shown in Figure 16.

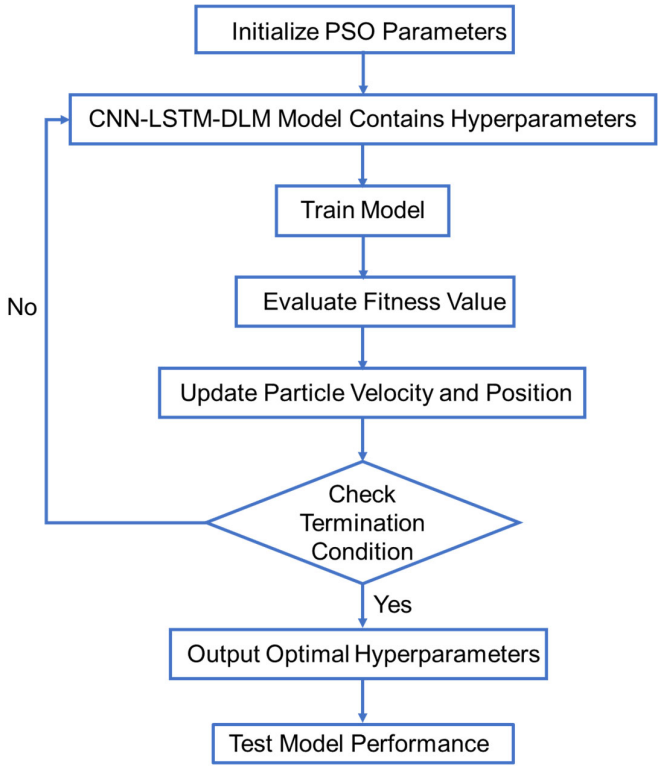


Figure 16. Flowchart for PSO optimization combination model.

For the gas yield prediction model, Table 6 shows the prediction range and the corresponding optimized values of its hyperparameters. This optimization process ensures that the CNN-LSTM-DLM model was fine-tuned to achieve the best possible performance in predicting gas product yield.

Table 6. PSO optimal parameters of the gas flowrate prediction model.

Parameter	Range	Optimal Value
Number of hidden neurons	[10, 30]	12
Initial learning rate	$[1 \times 10^{-4}, 3 \times 10^{-3}]$	0.00154
Regularization coefficient	$[1 \times 10^{-10}, 1 \times 10^{-2}]$	2.62×10^{-10}
Drop rate	[0.1, 0.7]	0.31208

Increasing the number of hidden neurons may improve the ability of the model to capture the complex relationship of the data, but too many hidden neurons can cause the overfitting of the model and reduce the generalization ability of the model [47]. The appropriate initial learning rate can help the model converge to the optimal solution faster [39]. If the initial learning rate and the parameter update are too large, the model has difficulty converging during training. Too small a learning rate results in slow model convergence and a longer training time. The regularization coefficient penalizes excessive weight values by introducing the L2 norm of weights into the loss function, which can make the model tend to learn smaller weight values and avoid overfitting [48]. A dropout rate was applied in the fully connected layer following the convolutional layer and the pooling layer of the CNN module, and this can reduce overfitting and improve the model's stability. Therefore, a suitable choice of key hyperparameter scopes can enhance the flexibility and adaptability of the model, enabling it to effectively handle diverse product prediction scenarios. The results of PSO optimized gas flowrate prediction are shown in Figure 17.

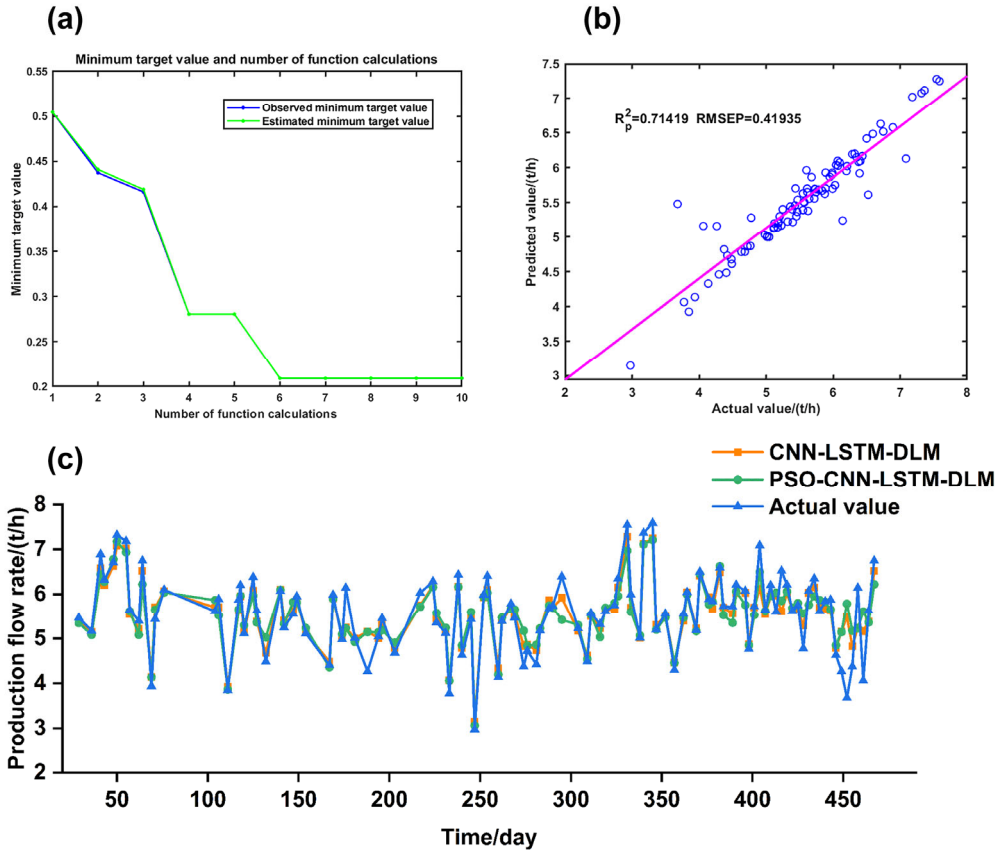


Figure 17. PSO optimization gas flowrate prediction model: (**a**) Optimization function decline curve. (**b**) Comparing the calculated value with the actual value. (**c**) The calculated value of the PSO-optimized CNN-LSTM-DLM model and the calculated value of CNN-LSTM-DLM model were compared with the actual value.

After optimization by the PSO algorithm, the R^2 value of the gas product test set increased from 0.332 to 0.492, effectively improving the prediction accuracy. According to the analysis of the optimization parameters, it is obvious to focus on overfitting for the data with a small overall amount, so the accuracy of part of the training set needs to be sacrificed. Considering the limitations, such as insufficient input variables and a weak correlation between many inputs and product yields, the computational results of the new strategy were satisfactory.

The CNN-LSTM-DLM models of kerosene, residue, heavy naphtha, and light naphtha were also optimized using the PSO algorithm and compared with the test set data. The predicted results are shown in Figure 18 and Table 7.

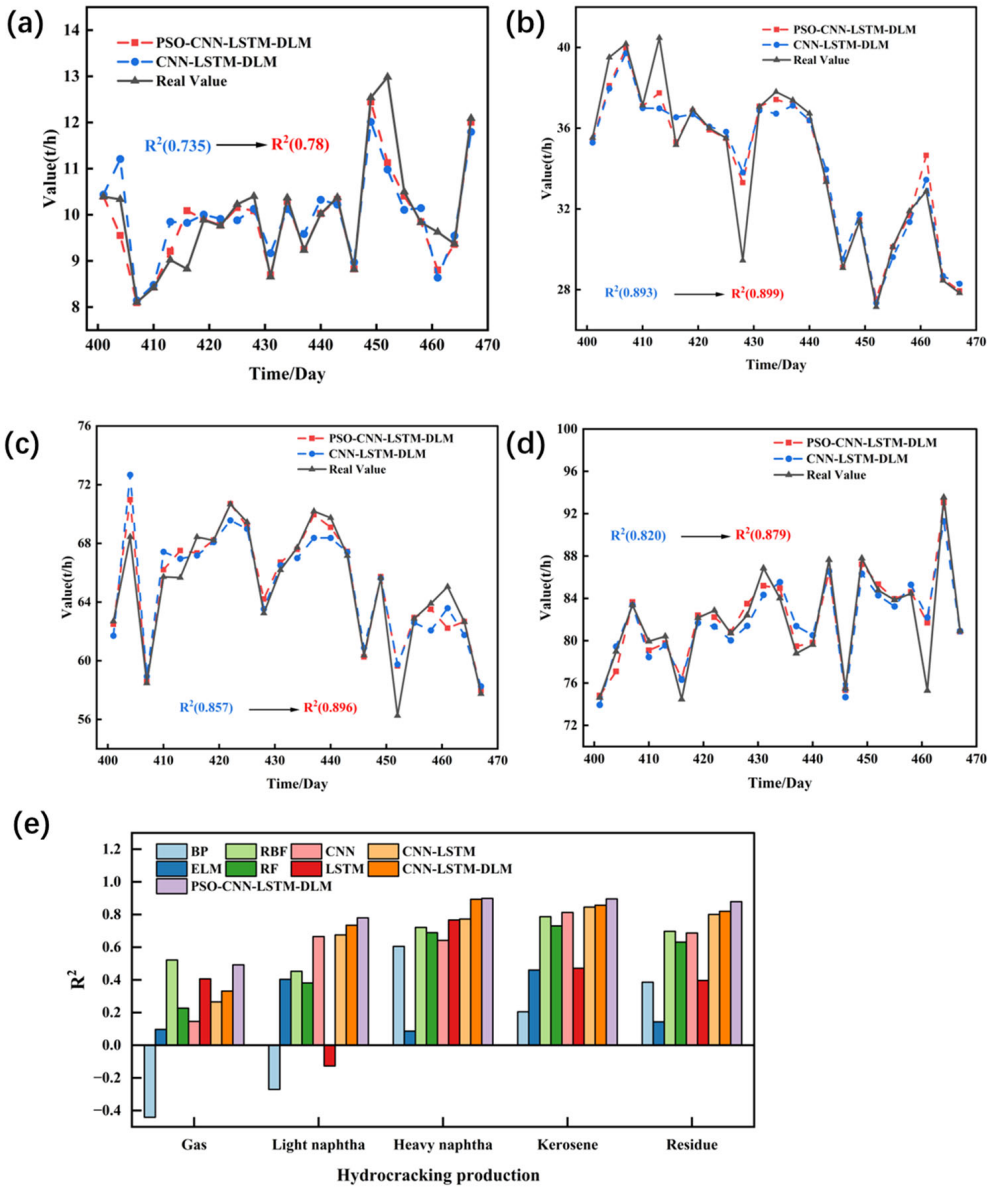


Figure 18. The calculated value of the PSO-optimized CNN-LSTM-DLM oil prediction model and the calculated value of the CNN-LSTM-DLM oil prediction model were compared with the actual value: (**a**) Light naphtha; (**b**) heavy naphtha; (**c**) kerosene; (**d**) residue. (**e**) Comparison of R2 values before and after PSO optimization.

Table 7. PSO-optimized CNN-LSTM-DLM evaluation results.

Objective	RMSE		MAE	
	Before	After	Before	After
Light naphtha	0.6226	0.5667	0.4412	0.2776
Heavy naphtha	1.3050	1.2741	0.7824	0.5796
Kerosene	1.5002	1.2814	1.1044	0.7319
Residue	1.9355	1.5833	1.2847	0.9094

The prediction accuracy improved to a certain extent, and the R^2 value was basically above 0.8. Given the low correlation of input and output variables in the characteristic analysis, these prediction results were satisfactory for industrial hydrocracking prediction. The PSO can effectively avoid the overfitting or insufficient training problem when manually adjusting parameters, thereby improving the prediction accuracy and the generalization ability of the model.

4. Conclusions

This paper proposed a new strategy to enhance the prediction accuracy for a weak correlation between input and output variables and insufficient data in industrial hydrocracking units. The CNN-LSTM can extract and retain features from the space and time domains, and it has much better performance than other data-driven models. The mechanism model was established based on the discrete lumped method and combined with the CNN-LSTM model to effectively increase the information required for predicting hydrocracking product yields. A SHAP analysis and MIC analysis were used to reasonably explain the effect of the mechanism model in the hybrid model. A PSO algorithm was used to adjust the hidden neuron number, initial learning rate, regularization coefficient, and dropout rate, which further improved the prediction effect. Through the proposed new strategy, the coefficient of determination of the predicted hydrocracking gas yield increased to 0.492, and the coefficients of determination of the predicted hydrocracking kerosene, residue, heavy naphtha, and light naphtha yields increased to 0.896, 0.879, 0.899, and 0.78 separately. Based on insufficient data conditions, the prediction effect of the model fully meets the demand for industrial hydrocracking units. The modeling strategy is also applicable to other refining processes, and the model can be applied in industrial environments by collecting relevant industrial data. Hence, this paper presents an efficacious approach to modeling complex industrial units, specifically addressing the challenge of modeling with insufficient data. The data from industrial hydrocracking units used in the current study are limited, especially high-quality, high-frequency real-time data, which restricts the training and validation of data-driven models. This can lead to inadequate model generalization capabilities. In the future, more sensors and monitoring devices should be installed to collect higher-frequency operation data in real time and capture more subtle process changes, which will further improve the quality of the training data to enhance the extensibility of the model.

Author Contributions: Conceptualization, Z.L. and K.Q.; methodology, Z.L.; software, Z.L.; validation, Z.L. and K.Q.; formal analysis, K.Q.; investigation, M.L.; resources, K.Q.; data curation, Z.L.; writing—original draft preparation, Z.L.; writing—review and editing, Y.Z., P.Y., and Y.L.; visualization, K.Q.; supervision, M.L.; project administration, K.Q.; funding acquisition, M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Development projects of SINOPEC, China, 323063.

Data Availability Statement: The datasets presented in this article are not readily available because of technical limitations. Requests to access the datasets should be directed to contact authors.

Conflicts of Interest: Authors Zhenming Li, Kang Qin, Yang Zhang, Peng Yang, Yue Lou and Mingfeng Li were employed by the company Research Institute of Petroleum Processing Co., Ltd., Sinopec. The company had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

The optimal parameter selection for the four traditional data-driven models, BP, ELM, RBF, and RF, was achieved through the least squares method. Each model type has specific parameters that play a crucial role in determining its performance. For the BP neural networks, the primary parameter to be optimized was the number of hidden neurons. The ELM requires the optimization of the number of hidden neurons and the selection of activation functions. The key optimization parameter of RBF networks is the expansion speed of the radial basis function. For the RF model, the main optimization parameters include the number of decision trees and the minimum number of leaves. These parameters significantly affect the prediction effects of these data-driven models, with their optimized values presented in Table A1.

The optimized objective functions were all the minimum sum of squares of the errors between the calculated yields and the actual yields in the training dataset, and the optimized calculated objective function is shown in Equation (A1).

$$Obj = \min \sqrt{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \quad \text{(A1)}$$

where y_i is the actual yields of hydrocracking production; $\hat{y}_i$ is the predicted yield of hydrocracking production; and n is the total amount of hydrocracking production data in the training dataset.

Table A1. BP, ELM, RBF, and RF model optimal parameters.

Projects	Hidden Neurons	Activation Function	Spread	Number of Decision Trees	Minimum Leaf Number
BP	12	/	/	/	/
ELM	11	sig	/	/	/
RBF	/	/	150	/	/
RF	/	/	/	124	1

The three types of deep structure networks, CNN, LSTM, and CNN-LSTM, encompass a substantial number of parameters, posing challenges for direct optimization with traditional algorithms. For the CNN and LSTM, the appropriate empirical parameters are chosen for calculation according to the literature [49]. The parameters of the CNN-LSTM align with those of the CNN and LSTM to facilitate effective comparisons.

Table A2. CNN, LSTM, and CNN-LSTM setting parameters.

Projects	Convolution Kernel Size	Initial Learning Rate	Learning Rate Decline Factor	Drop Rate	Regularization Coefficient	Hidden Neurons
CNN	3×1	0.001	0.1	0.35	1×10^{-10}	/
LSTM	/	0.001	0.1	0.35	/	64

References

1. Pham, D.V.; Nguyen, N.T.; Kang, K.H.; Seo, P.W.; Yun, D.; Phan, P.D.; Park, Y.-K.; Park, S. Comparative study of single- and two-stage slurry-phase catalytic hydrocracking of vacuum residue for selective conversion of heavy oil. *Catal. Today* **2024**, *426*, 114391.
2. Prajapati, R.; Kohli, K.; Maity, S.K. Slurry phase hydrocracking of heavy oil and residue to produce lighter fuels: An experimental review. *Fuel* **2021**, *288*, 119686.
3. Žula, M.; Grilc, M.; Likozar, B. Hydrocracking, hydrogenation and hydro-deoxygenation of fatty acids, esters and glycerides: Mechanisms, kinetics and transport phenomena. *Chem. Eng. J.* **2022**, *444*, 136564.
4. Pang, Z.; Huang, P.; Lian, C.; Peng, C.; Fang, X.; Liu, H. Data-driven prediction of product yields and control framework of hydrocracking unit. *Chem. Eng. Sci.* **2024**, *283*, 119386.
5. Li, Z.; Wang, X.; Du, W.; Yang, M.; Li, Z.; Liao, P. Data-driven adaptive predictive control of hydrocracking process using a covariance matrix adaption evolution strategy. *Control Eng. Pract.* **2022**, *125*, 105222.
6. Oh, D.-H.; Adams, D.; Vo, N.D.; Gbadago, D.Q.; Lee, C.-H.; Oh, M. Actor-critic reinforcement learning to estimate the optimal operating conditions of the hydrocracking process. *Comput. Chem. Eng.* **2021**, *149*, 107280.
7. Liu, C.; Wang, K.; Ye, L.; Wang, Y.; Yuan, X. Deep learning with neighborhood preserving embedding regularization and its application for soft sensor in an industrial hydrocracking process. *Inform. Sci.* **2021**, *567*, 42–57.
8. Zhong, W.; Qiao, C.; Peng, X.; Li, Z.; Fan, C.; Qian, F. Operation optimization of hydrocracking process based on Kriging surrogate model. *Control Eng. Pract.* **2019**, *85*, 34–40.
9. Alhajree, I.; Zahedi, G.; Manan, Z.A.; Zadeh, S.M. Modeling and optimization of an industrial hydrocracker plant. *J. Petrol. Sci. Eng.* **2011**, *78*, 627–636.
10. Elkamel, A.; Al-Ajmi, A.; Fahim, M. Modeling the hydrocracking process using artificial neural networks. *Petrol. Sci. Technol.* **1999**, *17*, 931–954.
11. Sepehr, S.; Gholam, R. Comparison of Kinetic-based and Artificial Neural Network Modeling Methods for a Pilot Scale Vacuum Gas Oil Hydrocracking Reactor. *Bull. Chem. React. Eng.* **2013**, *8*, 125–136.
12. Sharifi, K.; Safiri, A.; Asl, M.H.; Adib, H.; Nonahal, B. Development of a SVM model for Prediction of Hydrocracking Product Yields. *Petrol. Chem.* **2019**, *59*, 233–238.
13. Rahimpour, M.R.; Shayanmehr, M.; Nazari, M. Modeling and Simulation of an Industrial Ethylene Oxide (EO) Reactor Using Artificial Neural Networks (ANN). *Ind. Eng. Chem. Res.* **2011**, *50*, 6044–6052.
14. Ghosh, D.; Moreira, J.; Mhaskar, P. Application of data-driven modeling approaches to industrial hydroprocessing units. *Chem. Eng. Res. Des.* **2022**, *177*, 123–135.
15. Shokri, S.; Marvast, M.A.; Sadeghi, M.T.; Narasimhan, S. Combination of data rectification techniques and soft sensor model for robust prediction of sulfur content in HDS process. *J. Taiwan Inst. Chem. E* **2016**, *58*, 117–126.
16. Li, X.; Chan, C.W.; Nguyen, H.H. Application of the Neural Decision Tree approach for prediction of petroleum production. *J. Pet. Sci. Eng.* **2013**, *104*, 11–16.
17. Song, W.; Mahalec, V.; Long, J.; Yang, M.; Qian, F. Modeling the Hydrocracking Process with Deep Neural Networks. *Ind. Eng. Chem. Res.* **2020**, *59*, 3077–3090.
18. Gaddem, M.R.; Kim, J.; Matsunami, K.; Hayashi, Y.; Badr, S.; Sugiyama, H. Roles of mechanistic, data-driven, and hybrid modeling approaches for pharmaceutical process design and operation. *Curr. Opin. Chem. Eng.* **2024**, *44*, 101019.
19. Ye, Q.; Duan, P.; Kuang, S.; Ji, L.; Zou, R.; Yu, A. Multi-objective optimization of hydrocyclone by combining mechanistic and data-driven models. *Powder Technol.* **2022**, *407*, 117674.
20. Zhuang, Y.; Liu, Y.; Ahmed, A.; Zhong, Z.; del Rio Chanona, E.A.; Hale, C.P.; Mercangöz, M. A hybrid data-driven and mechanistic model soft sensor for estimating CO_2 concentrations for a carbon capture pilot plant. *Comput. Ind.* **2022**, *143*, 103747.
21. Song, W.; Du, W.; Fan, C.; Yang, M.; Qian, F. Adaptive Weighted Hybrid Modeling of Hydrocracking Process and Its Operational Optimization. *Ind. Eng. Chem. Res.* **2021**, *60*, 3617–3632.
22. Zhao, C.; Medeiros, T.X.; Sové, R.J.; Annex, B.H.; Popel, A.S. A data-driven computational model enables integrative and mechanistic characterization of dynamic macrophage polarization. *iScience* **2021**, *24*, 102112. [PubMed]
23. Huang, M.; Peng, C.; Du, Z.; Liu, Y. A power regulation strategy for heat pipe cooled reactors based on deep learning and hybrid data-driven optimization algorithm. *Energy* **2024**, *289*, 130050.
24. Tian, Y.; He, Y.-L.; Zhu, Q.-X. Soft Sensor Development Using Improved Whale Optimization and Regularization-Based Functional Link Neural Network. *Ind. Eng. Chem. Res.* **2020**, *59*, 19361–19369.
25. Yeung, C.; Pham, B.; Tsai, R.; Fountaine, K.T.; Raman, A.P. DeepAdjoint: An All-in-One Photonic Inverse Design Framework Integrating Data-Driven Machine Learning with Optimization Algorithms. *ACS Photonics* **2023**, *10*, 884–891.
26. Demir, S.; Sahin, E.K. Predicting occurrence of liquefaction-induced lateral spreading using gradient boosting algorithms integrated with particle swarm optimization: PSO-XGBoost, PSO-LightGBM, and PSO-CatBoost. *Acta Geotech.* **2023**, *18*, 3403–3419.

27. Yang, C.; Guan, X.; Xu, Q.; Xing, W.; Chen, X.; Chen, J.; Jia, P. How can SHAP (SHapley Additive exPlanations) interpretations improve deep learning based urban cellular automata model? *Comput. Environ. Urban Syst.* **2024**, *111*, 102133.
28. Wang, J.; Xu, P.; Ji, X.; Li, M.; Lu, W. MIC-SHAP: An ensemble feature selection method for materials machine learning. *Mater. Today. Commun.* **2023**, *37*, 106910.
29. Chen, Y.; Liu, C.; Guo, G.; Zhao, Y.; Qian, C.; Jiang, H.; Shen, B.; Wu, D.; Cao, F.; Sun, H. Machine-learning-guided reaction kinetics prediction towards solvent identification for chemical absorption of carbonyl sulfide. *Chem. Eng. J.* **2022**, *444*, 136662.
30. M'dioud, M.; Bannari, R.; Elkafazi, I. A novel PSO algorithm for DG insertion problem. *Energy Syst.* **2024**, *15*, 325–351.
31. Alavi Nezhad Khalil Abad, S.V.; Hazbeh, O.; Rajabi, M.; Tabasi, S.; Lajmorak, S.; Ghorbani, H.; Radwan, A.E.; Mudabbir, M. Determination of the Rate of Penetration by Robust Machine Learning Algorithms Based on Drilling Parameters. *ACS Omega* **2023**, *8*, 46390–46398. [PubMed]
32. Bharti, A.; Prerna Banerjee, T. Applicability of Cuckoo Search Algorithm for the Prediction of Multicomponent Liquid–Liquid Equilibria for Imidazolium and Phosphonium Based Ionic Liquids. *Ind. Eng. Chem. Res.* **2015**, *54*, 12393–12407.
33. Lin, T.-H.; Chiu, S.-H.; Tsai, K.-C. Supervised Feature Ranking Using a Genetic Algorithm Optimized Artificial Neural Network. *J. Chem. Inf. Model.* **2006**, *46*, 1604–1614.
34. Kim, S.-H.; Park, S.Y.; Seo, H.; Woo, J. Feature selection integrating Shapley values and mutual information in reinforcement learning: An application in the prediction of post-operative outcomes in patients with end-stage renal disease. *Comput. Meth. Prog. Bio.* **2024**, *257*, 108416.
35. Park, H.S.; Oh, B.K. CNN-based model updating for structures by direct use of dynamic structural response measurements. *Eng. Struct.* **2024**, *307*, 117880.
36. Gao, S.; Zhang, Q.; Tian, R.; Ma, Z.; Dang, X. Horizontal Data Augmentation Strategy for Industrial Quality Prediction. *ACS Omega* **2022**, *7*, 30782–30793. [PubMed]
37. Xie, Z.; Gu, X.; Shen, Y. A Machine Learning Study of Predicting Mixing and Segregation Behaviors in a Bidisperse Solid–Liquid Fluidized Bed. *Ind. Eng. Chem. Res.* **2022**, *61*, 8551–8565.
38. Kazi, M.-K.; Eljack, F. Practicality of Green H2 Economy for Industry and Maritime Sector Decarbonization through Multiobjective Optimization and RNN-LSTM Model Analysis. *Ind. Eng. Chem. Res.* **2022**, *61*, 6173–6189.
39. Zhong, H.; Yu, X.; Wei, Z.; Zhang, J.; Ding, L.; Niu, B.; Tang, R.; Xiong, Q.; Zhang, Y.; Kong, X. Development and Evaluation of Deep Learning Models for Predicting Instantaneous Mass Flow Rates of Biomass Fast Pyrolysis in Bubbling Fluidized Beds. *Ind. Eng. Chem. Res.* **2023**, *62*, 17158–17167.
40. Pan, S.; Yang, B.; Wang, S.; Guo, Z.; Wang, L.; Liu, J.; Wu, S. Oil well production prediction based on CNN-LSTM model with self-attention mechanism. *Energy* **2023**, *284*, 128701.
41. Shah, J.; Vaidya, D.; Shah, M. A comprehensive review on multiple hybrid deep learning approaches for stock prediction. *Intell. Syst. Appl.* **2022**, *16*, 200111. [CrossRef]
42. Li, G.; Xia, Y.; Zeng, W. Kinetic mechanism research of an industrial hydrocracker based on strict calculation of stoichiometric coefficients. *Fuel* **2013**, *103*, 285–291. [CrossRef]
43. Çay, Y.; Çiçek, A.; Kara, F.; Sağiroğlu, S. Prediction of engine performance for an alternative fuel using artificial neural network. *Appl. Therm. Eng.* **2012**, *37*, 217–225. [CrossRef]
44. Wang, X.; Bai, Y.; Liu, X. Prediction of railroad track geometry change using a hybrid CNN-LSTM spatial-temporal model. *Adv. Eng. Inform.* **2023**, *58*, 102235.
45. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90.
46. Zhang, C.; Wang, S.; Yu, C.; Xie, Y.; Fernandez, C. Novel Improved Particle Swarm Optimization-Extreme Learning Machine Algorithm for State of Charge Estimation of Lithium-Ion Batteries. *Ind. Eng. Chem. Res.* **2022**, *61*, 17209–17217.
47. Fujita, O. Statistical estimation of the number of hidden units for feedforward neural networks. *Neural Netw.* **1998**, *11*, 851–859. [CrossRef]
48. Pang, M.; Sun, H. Distributed regression learning with coefficient regularization. *J. Math. Anal. Appl.* **2018**, *466*, 676–689.
49. Shaohu, L.; Yuandeng, W.; Rui, H. Prediction of drilling plug operation parameters based on incremental learning and CNN-LSTM. *Geoenergy Sci. Eng.* **2024**, *234*, 212631.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.