Mathematical Biosciences and Engineering

MBE, 20 (7): 12433–12453.
DOI: 10.3934/mbe.2023553
Received: 04 April 2023
Revised: 04 May 2023
Accepted: 10 May 2023
Published: 24 May 2023

http://www.aimspress.com/journal/MBE

Research article

C4 olefin production conditions optimizing based on a hybrid model

Yancong Zhou[1,*,#], Chenheng Xu[2,#], Yongqiang Chen[1] and Shanshan Li[3]

[1] School of Information Engineering, Tianjin University of Commerce, Tianjin 300134, China
[2] School of Economics, Tianjin University of Commerce, Tianjin 300134, China
[3] School of Science, Tianjin University of Commerce, Tianjin 300134, China

*** Correspondence:** Email: zycong78@126.com; Tel: +86-156-9222-1287; Fax: +86-022-2668-6251.

[#] The authors contributed equally to this work.

Abstract: The yield of C4 olefin is often low due to the complexity of the associated products. Finding the optimal ethanol reaction conditions requires repeated manual experiments, which results in a large consumption of resources. Therefore, it is challenging to design ethanol reaction conditions to make the highest possible yield of C4 olefin. This paper introduces artificial intelligence technology to the optimization problem of C4 olefin production conditions. A sample incremental eXtreme Gradient Boosting tree based on Gaussian noise (GXGB) is proposed to establish the objective function of the variables to be optimized. The Sparrow Search Algorithm (SSA), which has an improved advantage in the optimization efficiency, is used to combine with GXGB. Therefore, a kind of hybrid model GXGB-SSA that can solve the optimization of complex problems is proposed. The purpose of this model is to find the combination of ethanol reaction conditions that makes the maximum yield of C4 olefin. In addition, due to the insufficient interpretation ability of GXGB on the data, the SHAP (SHapley Additive exPlanations) value method is creatively introduced to investigate the effect of each ethanol reaction condition on the yield of C4 olefin. The constraints of each decision variable for optimization are adjusted according to the analysis results. The experimental results have showed that the proposed GXGB-SSA model obtained the combination of ethanol reaction conditions that maximized the yield of C4 olefin. (i.e., when the Co loading is 1.1248 wt%, the Co/SiO_2 and HAP mass ratio is 1.8402, the ethanol concentration is 0.8992 ml/min, the total catalyst mass is 400 mg, and the reaction temperature is 420.37 ℃, the highest C4 olefin yield is obtained as 5611.46%). It is nearly 25.46 % higher compared to the current highest yield of 4472.81 % obtained from manual experiments.

Keywords: C4 olefin production; Complex problem optimization; Gaussian noise; XGBOOST; SSA; SHAP

1. Introduction

C4 olefin is an important industrial feedstock and is widely used in the production of chemical products and pharmaceuticals. At present, there are many methods to produce C4 olefin [1–3]. For example, the preparation of C4 olefins using carbon-oxygen and hydrogen is a well-known method [4,5]. Among which, ethanol is more important as a base material for C4 olefin production because of its wide source and low pollution [6,7]. In the production of C4 olefin from ethanol, different catalyst combinations and reaction temperatures play an important role. Although important results have been achieved in the production of C4 olefin from ethanol [8,9], based on the chain growth reaction mechanism, it is known that ketones and aldehydes are inevitably generated during the reaction, resulting in a low selectivity and poor economy of target products.

Lu [10] studied the C4 olefin production conditions by using the controlled variable method and concluded that the maximum yield of C4 olefin was obtained when the Co loading was 1 wt%, the Co/SiO2 and HAP mass ratio were 1, and the reaction temperature was 400 ℃ by conducting several manual experiments. However, conducting repeated manual experiments can cause more human, material, and financial resources consumption, while applying artificial intelligence technology to find the optimal reaction conditions for ethanol can effectively reduce resource consumption. At present, more experts and scholars have applied swarm intelligence algorithms to optimize various chemical problems [11–13]. The swarm intelligence algorithm has the advantages of simplicity, parallelism and applicability compared with traditional optimization algorithms [14,15]. It is particularly suitable for solving optimization problems of various complex systems [16,17]. Wang and Zhang [18] used a response surface method and a quadratic regression model to simulate the process of synthesizing methyl chloroacetate in the next-door tower of reaction distillation and obtained the operating parameters that make the synthesis product of the highest purity. Gao et al [19] used BP neural networks and genetic algorithms to optimize the S-Zorb device, which effectively reduced the octane loss of gasoline in the cracking process. Zeng [20] et al. used genetic and sequential algorithms to optimize the biogas decarbonization process and obtained the CO_2 values that make the LNG yield reach the standard. However, the problem of optimization of C4 olefin production conditions has not been investigated.

Therefore, a hybrid model combining the advantages of GXGB and SSA (i.e., GXGB-SSA) is proposed for the first time in this paper. This model was established to obtain the highest possible ethanol reaction conditions for C4 olefin production and reduce the resource consumption of repeated manual experiments by mining the experimental data of C4 olefin production.

The first task to achieve optimization of C4 olefin production conditions is to establish the objective function of the variable to be optimized for C4 olefin yield. However, the chemical reaction mechanism for the production of C4 olefin is nonlinear and complex, leading to the difficulty of establishing this objective function using conventional fitting models. In addition, the experimental data of C4 olefin production for the study conducted in this paper was obtained from the China 2021 National Student Mathematical Modeling Competition; it has a small sample size of only 109 groups, so the overfitting problem will occur by using an artificial neural network model to establish the

objective function of the variables to be optimized. As an integrated learning algorithm based on the idea of Boosting, the extreme gradient boosting tree (XGB) [21,22], can effectively reduce the bias of the model. Additionally, its introduction of the regularization term effectively reduces the probability of overfitting the model [23,24]. Therefore, it is very suitable for solving the establishment problem of this objective function [25]. In order to further improve the fitting effect of XGB, this paper proposes a sample increment type limit gradient boosting tree based on Gaussian noise, namely GXGB. The experimental results show that it can obtain a better fitting effect compared with XGB without the improvement.

In the process of optimizing the C4 olefin production conditions, the traditional grid search algorithm will undoubtedly consume a lot of time due to the large range of values of each decision variable, while the swarm intelligence algorithm, as an emerging evolutionary computing technology, has a greater advantage in operational efficiency; therefore, it is a good choice to apply the swarm intelligence algorithm to solve the optimization problem of C4 olefin production conditions. The sparrow search algorithm [26–28] is a new swarm intelligence optimization algorithm proposed based on the feeding behavior of sparrows. The main idea of the sparrow search algorithm is to perform local and global search by imitating the foraging and anti-feeding behavior of sparrows, and the sparrow foraging process is the algorithm seeking process [29,30]. And it has a greater advantage in the efficiency of finding the best [31,32]. Therefore, in this paper, we combine it with GXGB to propose a hybrid model, namely GXGB-SSA, which can solve the complex problem of finding the optimal conditions, apply it to the problem of finding the optimal conditions for C4 olefin production, and obtain the combination of ethanol reaction conditions that makes the highest yield of C4 olefin.

In addition, since GXGB is a typical black-box model, it has a weaker ability to interpret the data when compared to the conventional fitting model [33–35]. Although the degree of influence of each reaction condition of ethanol on the yield of C4 olefin can be obtained, the positive and negative effects of the influence and the dynamic process of the influence are not available. Therefore, in this paper, the SHAP value [36–38] is creatively combined with it to investigate the effect of each reaction condition of ethanol on C4 olefin yield. And the constraints of each decision variable involved in the optimization search are adjusted according to the analysis results [39,40]. Therefore, in this paper, the SHAP value is creatively combined with it to investigate the effect of ethanol reaction conditions on the yield of C4 olefin, and the constraints of each decision variable of optimization are adjusted according to the analysis results.

The innovations of this paper are as follows: (1) combines Gaussian noise with XGB for the first time, and proposes a GXGB model that can solve the modeling problem with small samples and complex mechanisms; (2) proposes a GXGB-SSA hybrid model for the first time, and applies it to the optimization problem of C4 olefin production conditions, and the yield of C4 olefin is improved by nearly 25.46% compared with the manual experimental data; (3) creatively combines the SHAP value with GXGB to solve the problem of its insufficient ability to explain data as a black-box model, effectively analyzes the effect of each reaction condition of ethanol on the yield of C4 olefin, and limits the search range of the optimal solution.

The structure of this paper is as follows: (1) the second part introduces the chemical reaction principles of C4 olefin production and quantifies the ethanol reaction conditions to obtain 109 sets of experimental data for modeling; (2) the third part of describes the process of establishing the proposed hybrid model GXGB-SSA; (3) the fourth part describes the process of using GXGB-SSA to find the optimal C4 olefin reaction conditions and the combination of ethanol reaction conditions that

maximizes the yield of C4 olefin; and (4) the fifth part concludes the paper and provides an outlook for future work.

2. Materials and methods

2.1. Preliminaries

The chemical principle applied in the C4 olefin production experiment studied in this paper is the ethanol dehydration reaction, the reaction process of which is divided into three main stages as follows.

$$CH_3CH_2OH \rightarrow CH_2{=}CH_2{+}H_2O \quad (a)$$

$$2CH_3CH_2OH \rightarrow CH_3CH_2OCH_2CH_3{+}H_2O \quad (b)$$

$$2CH_2 = CH_2 \rightarrow CH_2 = CHCH_2CH_3 \quad (c)$$

The dehydration reaction of ethanol is capable of producing C4 olefin, ethylene, and ether. The main reaction (a) is a strong heat-absorbing reaction; reaction (b) is a weakly exothermic reaction; and reaction (c) is a strong exothermic reaction, and its change with temperature is not obvious. Therefore, a proper increase in temperature is beneficial for the reaction to produce C4 olefin.

In addition, during the dehydration reaction of ethanol, decreasing the ethanol concentration actually decreases the partial pressure of the ethanol involved in the reaction, which facilitates the increase of the molar coefficient. Therefore, appropriately reducing the ethanol concentration also facilitates the reaction to form C4 olefin.

The catalysts utilized for the ethanol dehydration reaction studied in this paper are Co and SiO_2-HAP, where Co is a metal with dehydrogenation activity to catalyze the dehydration reaction of ethanol and SiO_2-HAP is a catalyst with both acid and base active sites.

In this paper, the ethanol reaction conditions were quantified by splitting them into Co loading (independent variable X_1), the Co/SiO2 and HAP loading ratio (independent variable X_2), ethanol concentration (independent variable X_3), total catalyst mass (independent variable X_4), and reaction temperature (independent variable X_5), where Co loading is the ratio of Co to SiO_2 by weight and the Co/SiO_2 and HAP mass ratio is the mass ratio of Co/SiO_2 and HAP. By adjusting the Co loading and SiO_2/HAP loading ratio, the acidity and alkalinity of the catalyst surface can be adjusted.

Table 1. Quantitative results of some experimental data.

	Y	X_1	X_2	X_3	X_4	X_5
1	70.39	1	1	1.68	400	250
2	315.99	0.5	1	1.68	400	300
3	343.43	1	2.03	1.68	100	350
…	…	…	…	…	…	…
107	1161.88	1	1	2.10	100	400
108	2906.24	2	1	0.30	400	400
109	1195.62	1	0.49	1.68	100	400

Furthermore, in this paper, the efficiency of the ethanol dehydration reaction is expressed in terms of the C4 olefin yield (dependent variable, Y), whose value is equal to the ethanol conversion rate

multiplied by the C4 olefin selectivity, where, ethanol conversion is the one-way conversion of ethanol per unit time and C4 olefin selectivity is the percentage of C4 olefin in all products.

The quantitative results of some experimental data for the ethanol dehydration reaction are shown in Table 1.

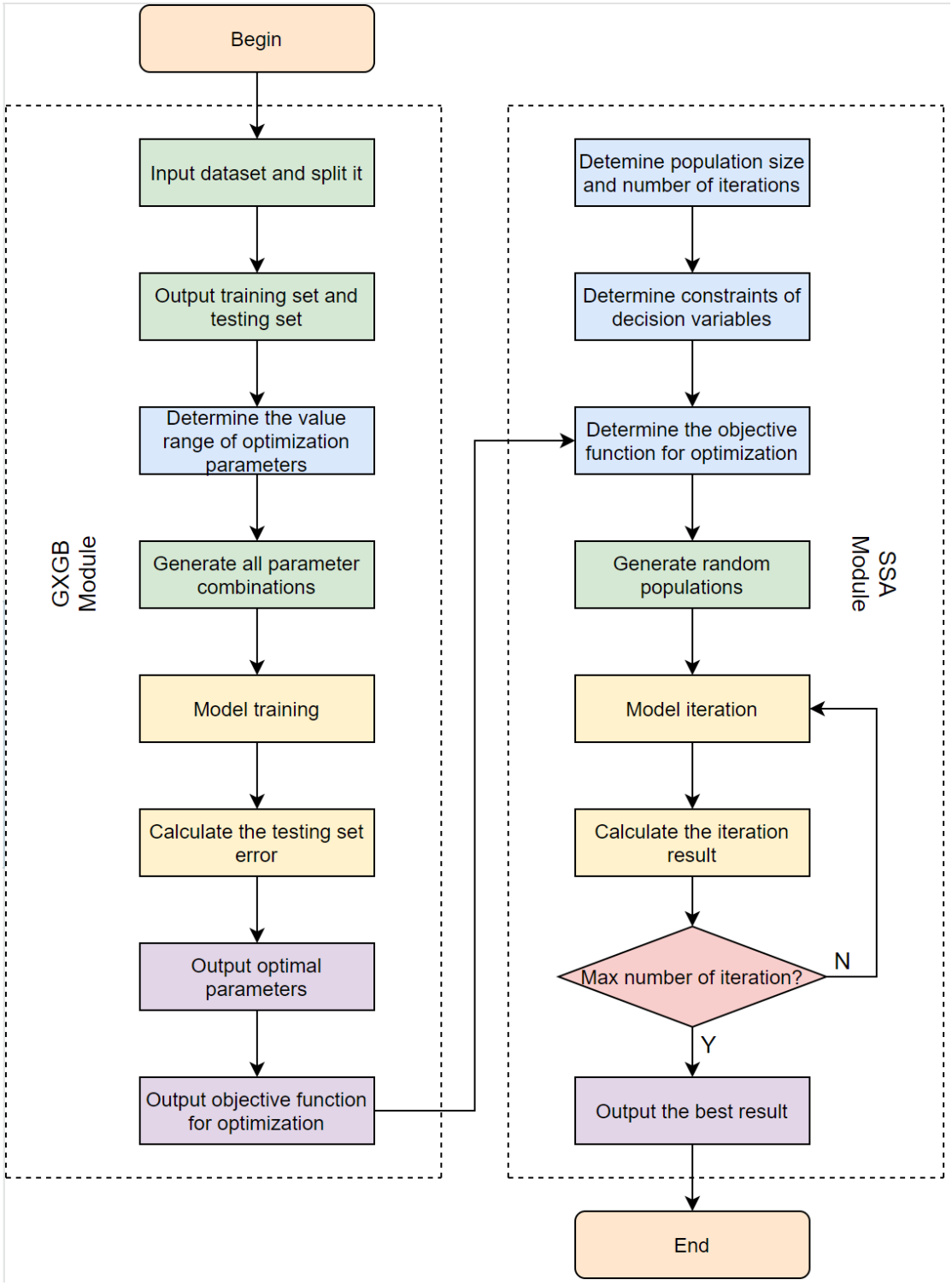

Figure 1. Flowchart of GXGB-SSA model.

Therefore, the tasks of this paper are described in two parts: (1) to establish the objective function *f(X)* of the variables to be optimized (i.e., the objective function with C4 olefin yield as the dependent variable and Co loading, the Co/SiO_2 and HAP mass ratio, ethanol concentration, total catalyst mass, and reaction temperature as the independent variables); and (2) to find the values of the independent variables that make the objective function *f(X)* the largest (i.e., to obtain the combination of ethanol reaction conditions with the highest C4 olefin yield).

2.2. Algorithms

2.2.1. Modeling ideas of GXGB-SSA

In this paper, we fully absorb the advantages of GXGB and SSA, and creatively fuse them to propose a model that can efficiently solve the optimization of complex problems (i.e., GXGB-SSA) and apply it to the optimization problem of reaction conditions for C4 olefin production. First, the basic idea of the model is to establish the objective function *f(X)* of the variables to be optimized for C4 olefin yield using GXGB, and then apply SSA to find the optimal value of the objective function to obtain the values of the decision variables that make the highest C4 olefin yield. The flow chart of the model is shown in Figure 1.

As can be seen in Figure 1, the establishment process of GXGB-SSA is divided into two main modules, namely, the GXGB module (the objective function establishment part) and the SSA module (the decision variable seeking part).

For the GXGB module, its main purpose is to establish the objective function of the variables to be optimized for C4 olefin yield, which is established as follows:

(1) Import a dataset and slice this dataset into a training set for model training and a test set for model testing, where the input variables to this dataset are Co loading, the Co/SiO2 and HAP mass ratio, ethanol concentration, total catalyst mass, reaction temperature, and the output variables are C4 olefin yields.

(2) The grid search method finds the optimal hyperparameter (i.e., the hyperparameter that minimizes the error in the test set) and the grid search range for each hyperparameter is shown in Table 2.

Table 2. Grid search range for hyperparameters of GXGB.

Hyperparameter	Grid search range
noise_level	[0.001, 0.005, 0.01]
increment_size	[0, 5, 10, 15, 20]
n_estimators	[200, 500, 700, 1000]
eta	[0.01, 0.05, 0.1, 0.2]
min_child_weight	[1, 2, 3, 4, 5]
max_depth	[5, 6, 7, 8, 9, 10]
gamma	[0, 0.1, 0.2, 0.3, 0.4, 0.5]
subsample	[0.5, 0.6, 0.7, 0.8, 0.9, 1]
colsample_bytree	[0.5, 0.6, 0.7, 0.8, 0.9, 1]

where noise level is the size of the noise level and increment size is a multiple of the sample increment.

(3) Make a model using the optimal model hyperparameters obtained in step (2) and output the objective function of the variable to be optimized for the C4 olefin yield.

In addition, GXGB is a typical black-box model because of its weaker ability to interpret data compared to the traditional fitting model. Therefore, in this paper, the SHAP value is combined with GXGB to investigate the effect of each reaction condition of ethanol on the yield of C4 olefin, and the constraints of each decision variable involved in the optimization search are adjusted according to the analysis results.

For the SSA module, the main objective is to find the values of each decision variable that makes the highest yield of C4 olefin, and the module is created as follows:

(1) Set the values of the optimization parameters (i.e., the number of populations (pop) and the maximum number of iterations (maxiter));

(2) Determine the constraints for each decision variable (i.e., the range of values available for Co loading, the Co/SiO2 and HAP mass ratio, ethanol concentration, total catalyst mass, and reaction temperature);

(3) Determine the objective function of the optimization (i.e., the objective function of the variable to be optimized for the C4 olefin yield obtained by the GXGB module);

(4) Iterate the model until the maximum number of iterations is reached and output the highest value of the C4 olefin yield and the value of each decision variable that makes the C4 olefin yield reach the maximum.

2.2.2. The establishment process of GXGB-SSA

For the GXGB module, the basic idea is to obtain a strong learner by continuously integrating weak learners, the learning result is the weighted mean value of each weak learner, and its flow chart is shown in Figure 2.

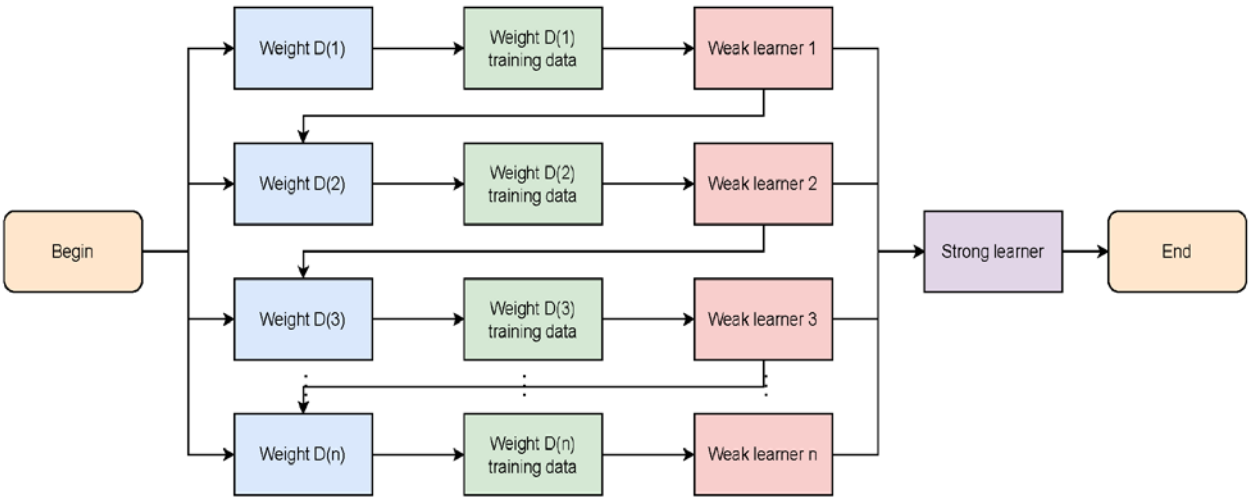

Figure 2. Flowchart of the GXGB module.

In the GXGB module, the added Gaussian noise samples obey the underlying normal distribution, whose expression is shown in Equation (1):

$$p(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(z-\bar{z})^2/2\sigma^2} \tag{1}$$

where z is a Gaussian random variable, $\bar{z}$ is the mean of z, and σ is the standard deviation of z. Gaussian noise samples can be added to the data samples by setting the Gaussian noise size and the sample increment multiplier.

The objective function established using the GXGB module consists of two main components, namely, the loss function and the regularization term. The introduction of the loss function can reduce the bias of the model, the introduction of the regularization term can reduce the probability of overfitting the model, and the expression of this objective function is shown in Equation (2):

$$Obj^{(t)} = \sum_{j=1}^{n} loss\left(y_j, \hat{y}_j^{(t-1)} + f_t(x_j)\right) + \Omega(f_t) + c \tag{2}$$

where $Obj^{(t)}$ is the objective function when integrating the t th weak learner; $\hat{y}_j^{(t-1)}$ is the objective value calculated by integrating the previous *t-1* weak learners; *loss()* is the loss function; c is the constant term; and $\Omega(f_t)$ is the regularization term, whose expression is shown in Equation (3):

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{o=1}^{T} w_o^2 \tag{3}$$

where, $\Omega(f_t)$ is the regularization term of the tth weak learner; γ and λ are the regularization coefficients; T is the number of all nodes of a certain weak learner; and w_o is the weight of the oth node of a certain weak learner.

A Taylor expansion of the objective function results in Equation (4) and (5) as follows:

$$\text{Obj}^{(t)} \simeq \sum_{j=1}^{n} \left[\text{loss}\left(y_j, \hat{y}_j^{(t-1)}\right) + g_j f_t(x_j) + \frac{1}{2} h_j f_t^2(x_j)\right] + \Omega(f_t) + c \tag{4}$$

$$g_j = \partial_{\hat{y}_j^{(t-1)}} \text{loss}\left(y_j, \hat{y}_j^{(t-1)}\right), h_j = \partial^2_{\hat{y}_j^{(t-1)}} \text{loss}\left(y_j, \hat{y}_j^{(t-1)}\right) . \tag{5}$$

Since the constant term does not affect the model solution, the constant term c and the fixed value $loss\left(y_j, \hat{y}_j^{(t-1)}\right)$ in the above equation are removed, and the result is shown in Equation (6):

$$Obj^{(t)} = \sum_{j=1}^{n} \left[g_j f_t(x_j) + \frac{1}{2} h_j f_t^2(x_j)\right] + \Omega(f_t) . \tag{6}$$

The objective function is deformed, and the result is shown in Equation (7):

$$Obj^{(t)} = \sum_{o=1}^{T} \left[\left(\sum_{j \in J_o} g_j\right) w_o + \frac{1}{2}\left(\sum_{j \in J_o} h_j + \lambda\right) w_o^2\right] + \gamma T . \tag{7}$$

Let $G_o = \sum_{j \in J_o} g_j, H_o = \sum_{j \in J_o} h_j$, then the objective function is as shown in (8):

$$Obj^{(t)} = \sum_{o=1}^{T} \left[G_o w_o + \frac{1}{2}(H_o + \lambda) w_o^2\right] + \gamma T . \tag{8}$$

The final objective function is obtained by taking partial derivatives of w_o, and its expression is shown in Equation (9):

$$Obj^{(t)} = -\frac{1}{2}\sum_{o=1}^{T}\frac{G_o^2}{H_o+\lambda} + \gamma T \quad . \tag{9}$$

In the GXGB module, assuming that the ith sample is xi, the jth feature of the ith sample is xij, the marginal contribution of the feature is $mcij$, and the weight is wi, then the ith expression for the SHAP value of the jth feature of the sample is shown in Equation (10).

$$f(x_{ij}) = mc_{ij}w_1 + \cdots \ldots + mc_{ij}w_n \tag{10}$$

Assuming that the predicted value of GXGB for this sample is yi and the baseline of the entire model (i.e., the mean of all sample target variables) is $ybase$, the expression for SHAP value is shown in Equation (11):

$$y_i = y_{\text{base}} + f(x_{i1}) + f(x_{i2}) + \cdots \ldots + f(x_{is}) \ . \tag{11}$$

$f(x_{ij})$ is the value of the contribution of the jth feature in the ith sample to the final prediction yi, and the SHAP value of each feature indicates the change in the model prediction when conditioned on that feature. When $f(x_{ij}) > 0$, it means that the feature boosts the prediction value, and vice versa, it means that the feature reduces the prediction value.

For the SSA module, the proposed basic idea is mainly inspired by the foraging behavior of sparrows. In the process of foraging, as explorers, sparrows provide the search direction and area for the population; as followers, sparrows search through the explorers' guidance; and as vigilantes, sparrows rely on anti-predation strategies to avoid the population from falling into local optimum.

During the iterative search, the expression for the explorer position update is shown in Equation (12):

$$x_{i,j}^{t+1} = \begin{cases} x_{i,j}^{t} \cdot exp\left(\frac{-i}{\alpha \cdot iter_{max}}\right) & if\ r_2 < ST \\ x_{i,j}^{t} + Q \cdot L & if\ r_2 \geq ST \end{cases} \tag{12}$$

where $x_{i,j}^{t}$ is the position of the ith sparrow in the jth dimension in the tth iteration, and $x_{i,j}^{t+1}$ is the position of the ith sparrow in the jth dimension in the $t+1$th iteration; r_2 is the warning value, whose value lies between $(0,1]$; ST is the safety value, whose value lies between $[0.5,1]$; α is a random number between $(0,1]$, whose value obeys uniform distribution ; $iter_{max}$ is the maximum number of iterations; Q is a random number of $(0,1]$ obeying normal distribution; and L is a matrix with element 1.

When the warning value r_2 is less than the safety value ST, the searcher performs a wide range of jump search; when the warning value r_2 is greater than the safety value ST, the searcher moves to other locations for search. Then, the expression of follower position update is shown in Equation (13):

$$x_{i,j}^{t+1} = \begin{cases} Q \cdot exp\left(\frac{x_{worst}^{t} - x_{i,j}^{t}}{i^2}\right) & if\ i > n/2 \\ x_P^{t+1} + \left|x_{i,j}^{t} - x_P^{t+1}\right| \cdot A^{+} \cdot L & otherwise \end{cases} \tag{13}$$

where $x_{i,j}^{t}$ is the position of the ith sparrow in the jth dimension in the tth iteration, and $x_{i,j}^{t+1}$ is the position of the ith sparrow in the jth dimension in the $t+1$th iteration; x_{worst} is the global worst position found by the discoverer's search; and $\mathbf{A}$ is a $1 \times d$ matrix where each element is randomly assigned to 1 or -1, and $A^{+} = A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}$.

If a sparrow with $i > n/2$ has a low fitness value and does not obtain food, then it is necessary to jump and move the search in the direction of the minimum value, and the other follower sparrows move towards the optimal position found by the explorer. Vigilantes are some sparrows randomly

selected from the sparrow population explorers and followers to avoid getting into local optimum by anti-predation strategy.

3. Results

3.1. Optimization of the objective function based on GXGB-SSA for C4 olefin variables

In this paper, with C4 olefin yield (Y) as the output variable and each reaction condition of ethanol as the input variable (i.e., Co loading (X_1), the Co/SiO_2 and HAP mass ratio (X_2), ethanol concentration (X_3), total catalyst mass (X_4), and reaction temperature (X_5)), the GXGB module was used to establish the objective function of the variables to be optimized for the C4 olefin yield.

The C4 olefin data set was sliced in a ratio of 4:1 to obtain a training set for model training and a test set for model testing, and a grid search method was applied to find the optimal hyperparameters of GXGB, and the combination of hyperparameters that minimized the fitting error of the test set was obtained, and the results are shown in Table 3.

Table 3. Grid search results for hyperparameters of GXGB.

Hyperparameters	Grid search results
noise_level	0.001
increment_size	5
n_estimators	500
eta	0.1
min_child_weight	5
max_depth	10
gamma	0
subsample	1
colsample_bytree	1

As seen from Table 3, the magnitude of the optimal Gaussian noise error level and the multiplicity of sample increments obtained by the grid search method are 0.001 and 5, respectively. Since the sample size of the experimental data for the production of C4 olefin studied in this paper is 109, the total sample size after the operation of sample increment is 545. Under the rule of dividing the training set and the test set with a ratio of 4:1, the final sample size involved in training is 436 groups, while the sample size involved in testing is 109 groups, where the GXGB is fitted on the test set as shown in Figure 3.

Figure 3 shows the fitting effect of GXGB on the testing set. Among them, the red circular marker represents the target value and the ×-shaped marker of the blue line represents the predicted value, which basically overlap. This shows that GXGB fits very well on the testing set.

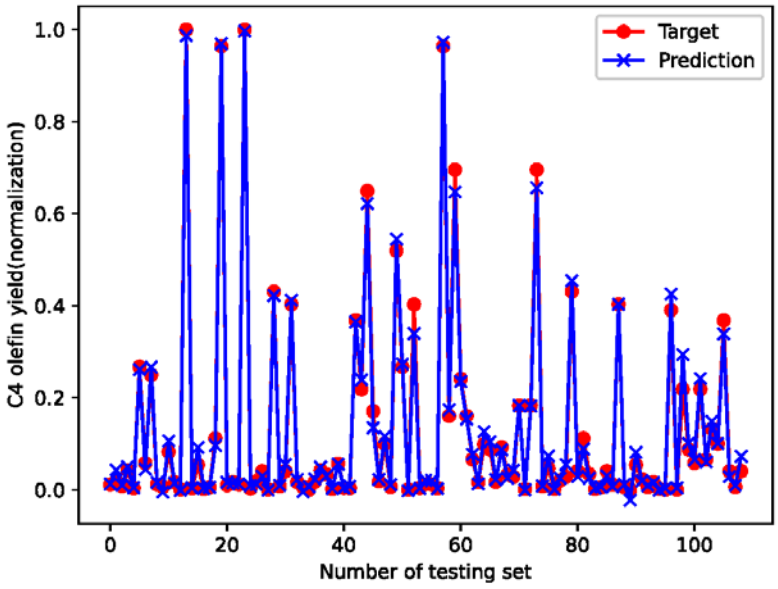

Figure 3. The fitting effect of GXGB on the testing set.

To further validate the performance of GXGB, a 10-fold cross-validation was conducted in this paper (i.e., the data samples from 10 different training and testing sets were trained 10 times). Additionally, the mean of the three metrics of goodness of fit (R^2), mean square error (MSE), and absolute error (MAE) obtained from the 10-fold cross-validation were used to evaluate the fitting accuracy of GXGB, and the variance (S^2) of the mean square error (MSE) and absolute error (MAE) obtained from the 10-fold cross-validation were used to evaluate the stability of GXGB. The stability of the GXGB was assessed by the variance (S^2) of the mean square error (MSE) obtained from the 10-fold cross-validation, where the expressions for the three indicators of goodness of fit (R^2), mean square error (MSE), and absolute error (MAE) were calculated as shown in Equations (14), (15), and (16):

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{m} (y_i - \bar{y})^2} \tag{14}$$

$$MSE = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \tag{15}$$

$$MAE = \frac{1}{m} \sum_{i=1}^{m} |(y_i - \hat{y}_i)| \tag{16}$$

where y_i is the target value of C4 olefin yield, $\hat{y}_1$ is the fitted value of the model, and m is the sample size.

For the goodness of fit (R^2), a larger value indicates a better model fit, and a smaller value indicates a worse model fit. For mean square error (MSE) and absolute error (MAE), the smaller the value, the better the model fit, and the larger the value, the worse the model fit. For the variance (S2), the smaller the value, the better the stability of the model, and the larger the value, the worse the stability of the model.

In this paper, GXGB is compared with XGB, random forest (RF) and support vector machines (SVM) without the introduction of Gaussian noise for improvement, and the results are shown in Tables 4 and 5.

Table 4. Comparison of GXGB with other algorithms on the training set.

	GXGB	XGB	RF	SVM
R^2	**0.9999**	0.9873	0.9642	0.9538
MSE	**3.5851 e-6**	4.6270 e-4	1.4086 e-3	6.3691 e-3
MAE	**8.2142 e-4**	1.2249 e-2	1.8544 e-2	5.9181 e-2
S^2	**4.2495 e-10**	2.9064 e-4	1.4463 e-4	1.1258 e-3

Table 5. Comparison of GXGB with other algorithms on the test set.

	GXGB	XGB	RF	SVM
R^2	**0.9953**	0.9664	0.9325	0.9274
MSE	**5.9207 e-4**	2.6855 e-3	6.2104 e-3	1.5241 e-2
MAE	**7.4447 e-3**	3.5297 e-2	5.4352 e-2	7.2451 e-2
S^2	**2.0628 e-5**	8.1356 e-4	6.9822 e-4	1.9799 e-3

As shown in Tables 4 and 5, GXGB outperforms XGB without the introduction of Gaussian noise for improvement in both the model fitting effect and stability. Among them, the goodness of fit (R^2) of GXGB is very high (very close to the ideal state 1) and its mean square error (MSE) and absolute mean error (MAE) are very small, indicating that the fitting error for this C4 olefin yield dataset is very small, and for assessing the variance (S^2) of model stability, GXGB also has a greater advantage.

In addition, this paper also compares GXGB with RF SVM, both of which have been used to adjust the hyperparameters using the grid search method. From Tables 4 and 5, it can be seen that GXGB outperforms RF and SVM in terms of model fitting effect and stability, which again verifies the good performance of GXGB.

3.2. Optimization of C4 olefin production conditions based on GXGB-SSA

In this paper, the objective function of the variables to be optimized for the C4 olefin yield established by the above GXGB module is used as the fitness function of the SSA module to find the optimal value, and the fitness value that makes the highest C4 olefin yield (i.e., the optimal value of each decision variable) is obtained. The constraints of each decision variable are shown in Table 6.

Table 6. Constraints on decision variables.

Decision variables	Constraints
Co loading(X_1)	$0.5 \leq X_1 \leq 5$
Co/SiO_2 and HAP mass ratio (X_2)	$0.5 \leq X_2 \leq 2$
Ethanol concentration (X_3)	$0.3 \leq X_4 \leq 2.1$
Total mass of catalyst (X_4)	$20 \leq X_3 \leq 400$
Reaction temperature (X_5)	$250 \leq X_5 \leq 450$

In this paper, the number of populations (pop) and the maximum number of iterations (maxiter) in the optimization parameters are uniformly set to 20 and 200, respectively, and the optimization search results of GXGB-SSA are compared with those of GXGB-GWO and GXGB-PSO using the

Gray Wolf Optimization algorithm. The iterative process of the three algorithms is shown in Figure 4.

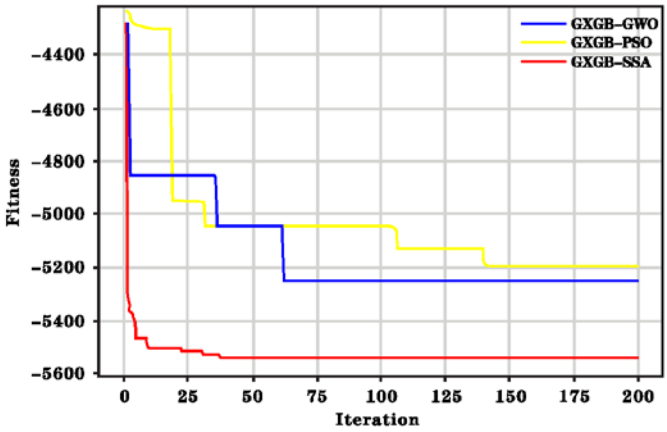

Figure 4. Iterative process diagram of the three algorithms.

As can be seen in Figure 4, all three algorithms converge after 200 iterations, and GXGB-SSA has a better convergence speed and optimization results than GXGB-GWO and GXGB-PSO for the optimization of C4 olefin production conditions. The optimization results of the three algorithms and their comparison with the optimal results obtained from manual experiments (ME) are shown in Table 7.

As can be seen from Table 7, the values of each decision variable that makes the highest C4 olefin yield are obtained for GXGB-SSA compared to GXGB-GWO and GXGB-PSO (i.e., when the Co loading (X_1) is 1.2508, the Co/SiO_2 and HAP mass ratio (X_2) is 1.4273, the ethanol concentration (X_3) is 0.9026, the total catalyst mass (X_4) is 399.88, and the reaction temperature (X_5) was 430.36) the highest C4 olefin yield (Y) of 5547.3526 was obtained, which was improved by nearly 24.02% compared with the highest C4 olefin yield of 4472.81 obtained from 109 manual experiments.

Table 7. Optimization results of the three algorithms and manual experimental results.

	GXGB-SSA	GXGB-GWO	GXGB-PSO	ME
Running time（*unit: s*）	**2.21**	4.36	2.69	--
Number of convergences	**27**	61	139	--
C4 olefin yield	**5547.35**	5285.74	5211.68	4472.81
Co loading	1.2508	0.9574	1.1425	1
Co/SiO_2 and HAP mass ratio	1.4273	0.8317	1.6237	1
Ethanol concentration	0.9026	0.8162	0.3372	0.9
Total mass of catalyst	399.88	370.76	399.01	400
Reaction temperature	430.36	424.13	413.25	400

4. Discussion

4.1. Investigation of the effect of each reaction condition on C4 olefin based on SHAP

The SHAP value was combined with GXGB to investigate the effect of ethanol reaction conditions on the yield of C4 olefin, and the constraints of the optimized decision variables were adjusted

according to the results of the analysis. The SHAP feature variable importance diagram is shown in Figure 5， and the summary diagram of the SHAP feature analysis is shown in Figure 6.

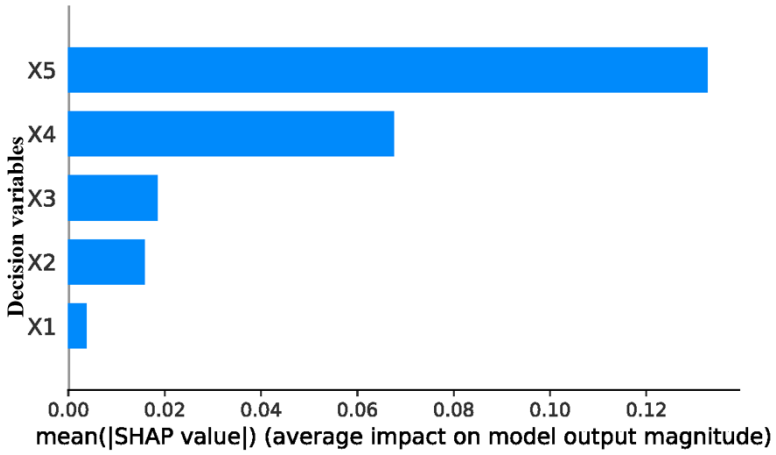

Figure 5. SHAP feature variable importance diagram.

In Figure 5, the X-axis represents the SHAP value of the feature variable, and the larger the value is, the greater the influence of the feature variable. Then, in descending order, the degree of influence of each decision variable on the yield of C4 olefin was the reaction temperature (X_5), total catalyst mass (X_4), ethanol concentration (X_3), the Co/SiO_2 and HAP mass ratio (X_2), and Co loading (X_1).

In Figure 6, the points of each feature represent the feature samples in the corresponding dataset, with the color change from blue to red indicating the value of the sample feature from small to large, and the positive or negative SHAP value indicating the positive or negative correlation of the feature with the target feature, respectively.

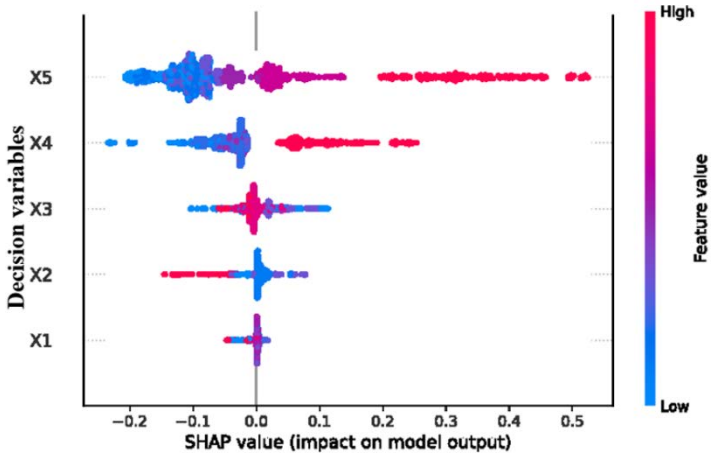

Figure 6. Summary diagram of SHAP feature analysis.

For the reaction temperature (X_5), which has the greatest influence, the samples with larger eigenvalues have a positive SHAP value, indicating that it acts as a positive contributor to the target characteristic and increases the target value of the C4 olefin yield. The feature points near the bottom half of the variable bars are mainly concentrated in the region with negative SHAP values, indicating that when the reaction temperature (X_5) decreases, the target value of the C4 olefin yield also decreases. Similarly, the total catalyst mass (X_4), which is the second most influential catalyst, also positively

contributes to the C4 olefin yield. For the Co/SiO_2 and HAP mass ratio (X_2) and Co loading (X_1), which are the third and fourth most influential catalysts, the SHAP value of the sample with the larger characteristic value is negative, indicating that it has a reverse inhibitory effect on the target characteristic and reduces the target value of C4 olefin yield. For the least influential Co loading (X_1), the positive and negative effects on the C4 olefin yield were not significant.

The magnitude and positive and negative effects of each decision variable on C4 olefin yield are able to be obtained in the SHAP characteristic analysis summary diagram; however, to explore the dynamic process of the effect of each decision variable on C4 olefin yield, it is necessary to view the SHAP characteristic analysis dependence diagram, as shown in Figures 7, 8, 9, 10, and 11.

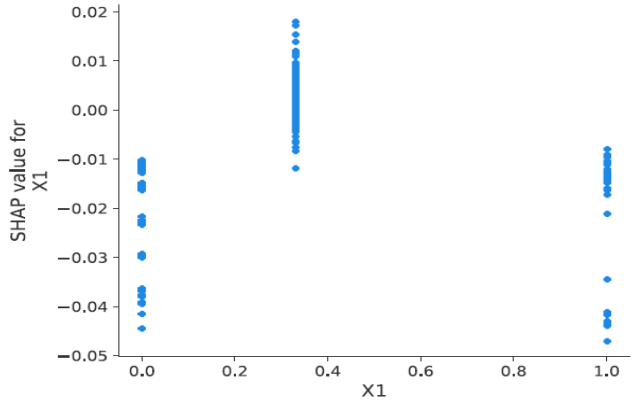

Figure 7. SHAP feature analysis dependence plot for Co loading (X_1).

In Figure 7, the Y-axis represents the SHAP value of the feature variables. The points of each feature represent the feature samples in the corresponding data set. Additionally, the samples with larger feature values have positive SHAP values, indicating that they play a positive contributing role to the target features. From Figure 9, it can be seen that when the Co loading (X_1) takes the value of 1, there exists a positive value of its SHAP value (i.e., the contribution to the C4 olefin yield is positive).

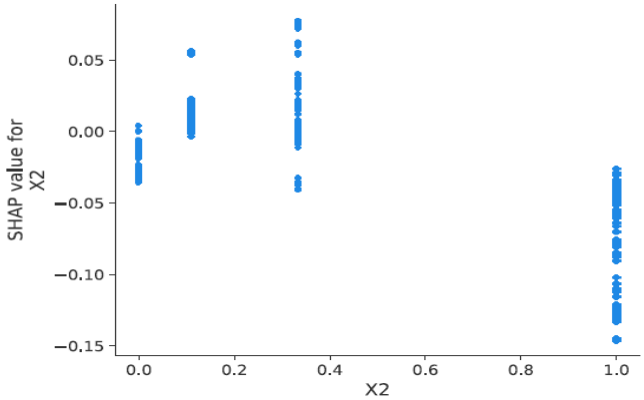

Figure 8. SHAP characteristic analysis dependence plot for Co/SiO_2 and HAP mass ratio (X_2).

From Figure 8, it can be seen that when the Co/SiO_2 and HAP charge ratio (X_2) takes the value of either 1 or 2, there is a positive value of SHAP value (i.e., the contribution to C4 olefin yield is positive).

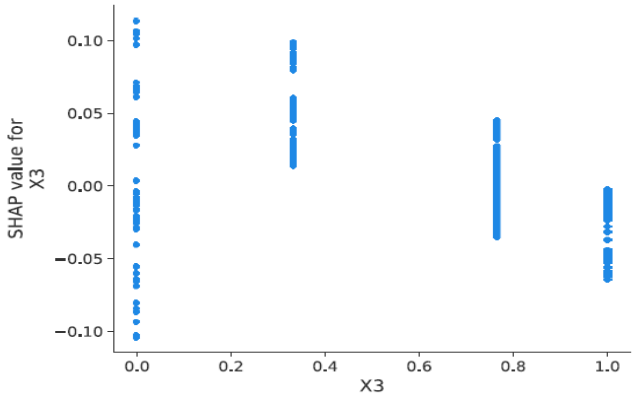

Figure 9. SHAP characteristic analysis dependence plot for ethanol concentration (X_3).

From Figure 9, it can be seen that when the ethanol concentration (X_3) is taken as either 0.3, 0.9, or 1.68, there is a positive value of its SHAP value (i.e., the contribution to the C4 olefin yield is positive).

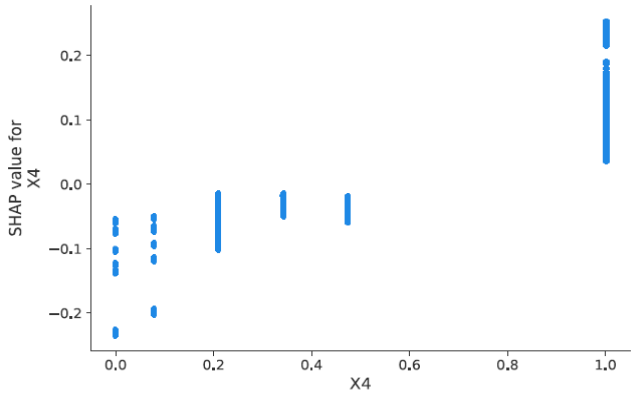

Figure 10. SHAP characteristic analysis dependence plot for total catalyst mass (X_4).

From Figure 10, it can be seen that when the total catalyst mass (X_4) is taken as 400, there is a positive value of SHAP value (i.e., the contribution to the C4 olefin yield is positive).

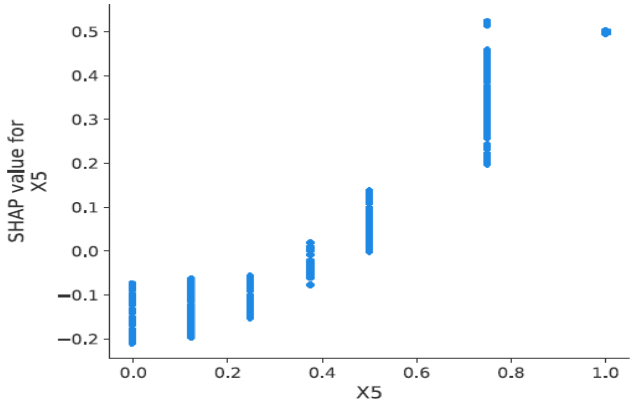

Figure 11. SHAP characteristic analysis dependence plot for reaction temperature (X_5).

From Figure 11, it can be seen that when the reaction temperature (X_5) is taken as either 400 or

450, there exists a positive value of its SHAP value (i.e., the contribution to the C4 olefin yield is positive).

Therefore, based on the SHAP characteristic analysis dependence diagram, the possible range of values for each decision variable when the contribution to C4 olefin yield is positive can be derived, and then the adjusted range of values for the decision variables is shown in Table 8.

Table 8. The range of values of the adjusted decision variables.

Decision variables	Constraints
Co loading(X_1)	$1 \leq X_1 \leq 2$
Co/SiO_2 and HAP mass ratio (X_2)	$1 \leq X_2 \leq 2$
Ethanol concentration (X_3)	$0.3 \leq X_4 \leq 1.68$
Total mass of catalyst (X_4)	$X_3 = 400$
Reaction temperature (X_5)	$400 \leq X_5 \leq 450$

The range of values of the adjusted decision variables in Table 8 are used as the constraint of GXGB-SSA for the optimization search again, and the iterative process is shown in Figure 12.

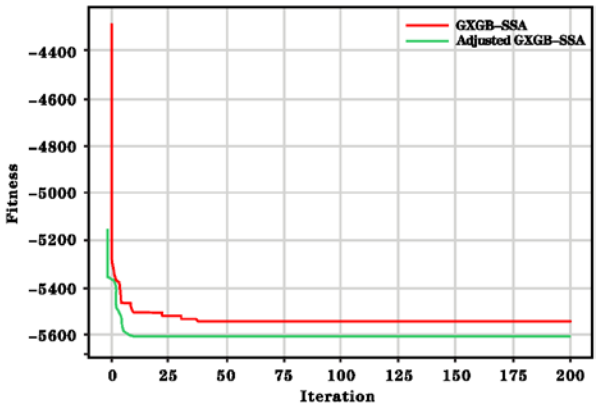

Figure 12. Iterative process of GXGB-SSA after adjusting the constraints.

As can be seen in Figure 12, with adjusted constraints, the GXGB-SSA is already in a converged state after 200 iterations, and its convergence speed and optimized results for the optimization of C4 olefin production conditions are improved compared with the GXGB-SSA without adjusted constraints. In this paper, the results of this optimization are compared with those of GXGB-SSA without adjustment of constraints, as shown in Table 9.

As can be seen from Table 9, the convergence rate and optimization results of GXGB-SSA with reduced constraints are improved compared to the optimization results without reduced constraints (i.e., when the Co loading (X_1) is 1.1248, the Co/SiO_2 and HAP mass ratio (X_2) is 1.8402, the ethanol concentration (X_3) is 0.8992, the total catalyst mass (X_4) is 400.00, and a reaction temperature (X_5) of 420.37) resulted in a higher C4 olefin yield of 5611.46, which was nearly 25.46% higher than the highest C4 olefin yield of 4472.81 in the experimental data.

Table 9. Optimization results of GXGB-SSA with adjusted constraints.

	Adjustive GXGB-SSA	GXGB-SSA
Running time (unit: s)	**2.16**	2.21
Number of convergences	**12**	27
C4 olefin yield	**5611.46**	5547.35
Co loading	1.1248	1.2508
Co/SiO_2 and HAP mass ratio	1.8402	1.4273
Ethanol concentration	0.9026	0.8626
Total mass of catalyst	400.00	399.88
Reaction temperature	420.37	430.36

5. Conclusions

In order to improve the efficiency of the ethanol dehydration reaction to produce C4 olefin, a GXGB-SSA hybrid model is developed in this paper to obtain the combination of ethanol reaction conditions that makes the highest yield of C4 olefin. First, the objective function of C4 olefin to be optimized was established by using GXGB module with the C4 olefin yield as the output variable and ethanol reaction conditions as the input variable. Second, the SSA module was used to optimize the objective function to obtain the combination of ethanol reaction conditions that makes the C4 olefin yield as high as possible. Finally, the SHAP value was used to investigate the effect of each ethanol reaction condition on the C4 olefin yield. The SHAP value was used to investigate the effect of ethanol reaction conditions on the C4 olefin yield, and the constraints of the decision variables involved in the optimization were adjusted according to the analysis results. The team will continue to investigate the machine learning algorithm in order to build a better optimization model and solve more complex problems.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. F. Yu, Z. J. Li, Y. L. An, P. Gao, L. S. Zhong, Y. H. Sun, Research progress on direct preparation of low carbon olefins by catalytic conversion of syngas, *J. Fuel Chem.*, **44** (2016), 14. https://doi.org/10.3969/j.issn.0253-2409.2016.07.005
2. Z. B. Guo, Y. Y. Zhang, X. Feng, Separation and purification of C_4 to C_6 hydrocarbons by metal-organic framework, *J. Chem-NY.*, **78** (2020),10. https://doi.org/10.6023/A20030081
3. H. T. Wang, G. Z. Qi, X. H. Li, W. Li, Coupling catalytic cracking of C4 olefins with methanol conversion to olefins over SAPO-34 catalyst, *Chem. React. Eng. Process.*, **2** (2013), 47–53. https://doi.org/CNKI:SUN:HXFY.0.2013-02-007

4. X. H. Gao, Y. L. Wang, S. P. Lu, J. L. Zhang, S. B. Fan, T. S. Zhao, Microwave hydrothermal preparation of Fe-Zr catalysts and their performance in CO hydrogenation to olefins, *J. Fuel Chem.*, **42** (2014), 219–224. https://doi.org/10.3969/j.issn.0253-2409.2014.02.014
5. B. H. Liu, W. J. Wang, Research progress on catalysts for the synthesis of low carbon olefins by carbon dioxide hydrogenation, *Anhui Chem. Industry*, **45** (2019), 5. https://doi.org/CNKI:SUN:AHHG.0.2019-05-004
6. J. Wang, S. Q. Zhong, Z. K. Xie, C. F. Zhang, Analysis of ethanol dehydration to olefin process, *Nat. Gas Chem. C1 Chem. Chem. Eng.*, **33** (2008), 27–32. https://doi.org/10.3969/j.issn.1001-9219.2008.05.007
7. J. H. Zhu, J. Yu, Z. K. Duan, Z. G. Zhang, G. W. Xu, Preparation of low carbon olefins by Fe_2O_3/HZSM-5 catalyzed conversion of ethanol, *J. Process Eng.*, **1** (2010), 6. https://doi.org/CNKI:SUN:HGYJ.0.2010-01-019
8. Y. C. Liu, Y. B. Xu, J. Y. Lu, ZSM-5-catalyzed ethanol to oligocene, *Adv. Chem.*, **22** (2010), 6. https://doi.org/CNKI:SUN:HXJZ.0.2010-04-026
9. W. Xia, J. G. Wang, C. Qian, Y. X. Huang, C. Ma, Y. Fan, et al., Comprehensive experiments on the preparation of low carbon olefins from ethanol catalyzed by Zr/ZSM-5 molecular sieve, *Exper. Technol. Manag.*, **38** (2021), 149–153. https://doi.org/10.16791/j.cnki.sjg.2021.08.032
10. G. S. Tang, X. W. Xiu, M. T. Wang, W. X. Chang, Effect of catalyst and temperature on the preparation of C4 olefins, *Labor. Res. Explor.*, **41** (2022), 28–32. https://doi.org/10.19927/j.cnki.syyt.2022.11.006
11. S. Y. Li, Advances in ant colony optimization algorithms and their applications, *Comput. Measur. Control*, **11** (2003), 911–913. https://doi.org/10.3321/j.issn:1671-4598.2003.12.001
12. J. Yang, S. S. Yang, Z. Y. Duan, L. Z. Zhu, Development and application of chemical experimental design and optimization methods, *Guangdong Chem. Indust.*, **37** (2010), 2. https://doi.org/10.3969/j.issn.1007-1865.2010.10.035
13. W. Fang, Population intelligence algorithm and its research in digital filter optimization design, *Wuxi Jiangnan University*. https://doi.org/10.7666/d.y1399305
14. Y. Deng, Q. Y. Jiang, Z. K. Cao, J. Shi, H. Zhou, An improved particle swarm algorithm for chemical process optimization, *Comput. Appl. Chem.*, **28** (2011), 4. https://doi.org/10.3969/j.issn.1001-4160.2011.06.020
15. D. Y. Diao, Z. F. Cao, Improved multi-objective particle swarm optimization algorithm for intermittent steaming process, *Comput. Appl.*, **32** (2012), 4. https://doi.org/CNKI:SUN:JSJY.0.2012-S2-017
16. S. Fan, W. Zhong, H. Cheng, Q. Feng, Novel control vector parameterization method with differential evolution algorithm and its application in dynamic optimization of chemical processes, *Chinese J. Chem. Eng.*, **21** (2013), 64–71. https://doi.org/10.1016/S1004-9541(13)60442-5
17. X. F. Wang, X. H. Wang, H. Z. Du, P. Wang, Fuzzy rough set and support vector machine-based fault diagnosis for chemical processes, *Control Decision Making*, **30** (2015), 4. https://doi.org/10.13195/j.kzyjc.2014.0246
18. Y. W. Wang, L. Zhang, Simulation and optimization of response surface method for synthesis of methyl chloroacetate by reaction distillation next door tower process, *Nat. Gas Chem. C1 Chem. Chem. Eng.*, **45** (2020), 7. https://doi.org/CNKI:SUN:TRQH.0.2020-03-025

19. P. Gao, S. Liu, S. H. Cheng, F. S. Ouyang, M. Y. Zhao, Optimization of gasoline octane loss in S Zorb unit based on BP neural network and genetic algorithm, *Petrol. Refin. Chem.*, **52** (2021), 8. https://doi.org/10.3969/j.issn.1005-2399.2021.07.020
20. J. Chen, J. F. Tang, X. M. Jin, Y. H. Hua, J. Chu, Y. Wang, et al., Pilot experiments and simulation optimization of decarbonization process parameters of alcohol-amine method, *JPT*, **33** (2017), 9. https://doi.org/10.3969/j.issn.1001-8719.2017.05.020
21. T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2016), 785–794. https://doi.org/10.1145/2939672.2939785
22. D. Zhang, L. Qian, B. Mao, C. Huang, Y. Si, A data-driven design for fault detection of wind turbines using random forests and XGboost, *IEEE Access*, **6** (2018), 21020–21031. https://doi.org/10.1109/ACCESS.2018.2818678
23. J. Q. Shi, J. H. Zhang, Load forecasting method based on multi-model fusion Stacking integrated learning approach, *Chinese J. Electr. Eng.*, **39** (2019), 10. https://doi.org/10.13334/j.0258-8013.pcsee.181510
24. H. Jiang, Z. He, G. Ye, H. Zhang, Network intrusion detection based on PSO-XGBoost model, *IEEE Access*, **8** (2020), 58392–58401. https://doi.org/10.1109/ACCESS.2020.2982418
25. A. Ogunleye, Q. G. Wang, XGBoost model for chronic kidney disease diagnosis, *IEEE/ACM Transact. Computat. Biol. Bioinform.*, **17** (2019), 2131–2140. https://doi.org/10.1109/TCBB.2019.2911071
26. B. S. Bhati, G. Chugh, F. AlmilUrjman, N. S. Bhati, An improved ensemble-based intrusion detection technique using XGBoost, *Transact. Emerg. Telecommun. Technol.*, **1** (2020). https://doi.org/10.1002/ett.4076
27. Y. Liang, D. X. Niu, M. Q. Ye, W. C. Hong, Short-term load forecasting based on wavelet transform and least squares support vector machine optimized by improved cuckoo search, *Energies*, **9** (2016), 827. https://doi.org/10.3390/en9100827
28. G. Y. Liu，C. Shu，Z. W. Liang，B. H. Peng，L. F. Cheng, A modified sparrow search algorithm with application in 3D route planning for UAV, *Sensors*, **21** (2021), 1224. https://doi.org/10.3390/s21041224
29. C. Zhang, S. Ding, A stochastic configuration network based on chaotic sparrow search algorithm, *Knowledge-Based Syst.*, **220** (2021), 106924. https://doi.org/10.1016/j.knosys.2021.106924
30. Q. H. Mao, Q. Zhang, An improved sparrow algorithm incorporating Corsi variation and backward learning, *Comput. Sci. Explor.*, **15** (2021), 10. https://doi.org/10.3778/j.issn.1673-9418.2010032
31. Y. Zhu, N. Yousefi, Optimal parameter identification of PEMFC stacks using adaptive sparrow search algorithm, *Int. J. Hydrogen Energ.*, **46** (2021). https://doi.org/10.1016/j.ijhydene.2020.12.107
32. P. Wang, Y. Zhang, H. Yang, Research on economic optimization of microgrid cluster based on chaos sparrow search algorithm, *Comput. Int. Neurosci.*, **3** (2021), 1–18. https://doi.org/10.1155/2021/5556780
33. S. Lundberg, S. I. Lee, A unified approach to interpreting model predictions, *Adv. Neural Inform. Process. Syst.*, (2017), 30. https://doi.org/10.48550/arXiv.1705.07874
34. A. B. Parsa, A .Movahedi, H. Taghipour, S. Derrible, A. K. Mohammadian, Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis, *Accident Anal. Prev.*, **136** (2020), 105405. https://doi.org/10.1016/j.aap.2019.105405

35. Y. D. Gu, H. Liu, Improved load forecasting based on optimal feature combination for limit gradient lifting, *Comput. Appl. Res.*, **38** (2021), 2767–2772. https://doi.org/10.19734/j.issn.1001-3695.2020.12.0552
36. Y. Luo, F. Wang, W. L. Ye, Explainable prediction model for acute kidney injury based on XGBoost and SHAP, *J. Electron. Inf. Technol.*, **44** (2022), 12. https://doi.org/10.11999/JEIT210931
37. S. C. Chelgani, H. Nasiri, M. Alidokht, Interpretable modeling of metallurgical reactions in industrial coal pillar flotation circuits via XGBoost and SHAP - a "Conscious Laboratory" development, *J. Min. Sci. Technol. English Edition*, **31** (2021), 10. https://doi.org/10.3969/j.issn.2095-2686.2021.06.016
38. H. Yang, E. Li, Y. F. Cai, J. P. Li, G. X. Yuan, The extraction of early warning features for the predicting financial distress based on XGboost model and shap framework, *Int. J. Finan. Eng.*, **8** (2021), 2141004. https://doi.org/10.1142/S2424786321410048
39. J. Xue, B. Shen, A novel swarm intelligence optimization approach: sparrow search algorithm, *Syst. Sci. Control Eng.*, **8** (2020), 22–34. https://doi.org/10.1080/21642583.2019.1708830
40. J. Yuan, Z. Zhao, Y. Liu, B. He, Y. Gao, DMPPT control of photovoltaic microgrid based on improved sparrow search algorithm, *IEEE Access*, **9** (2021), 16623–16629. https://doi.org/10.1109/ACCESS.2021.3052960

AIMS Press