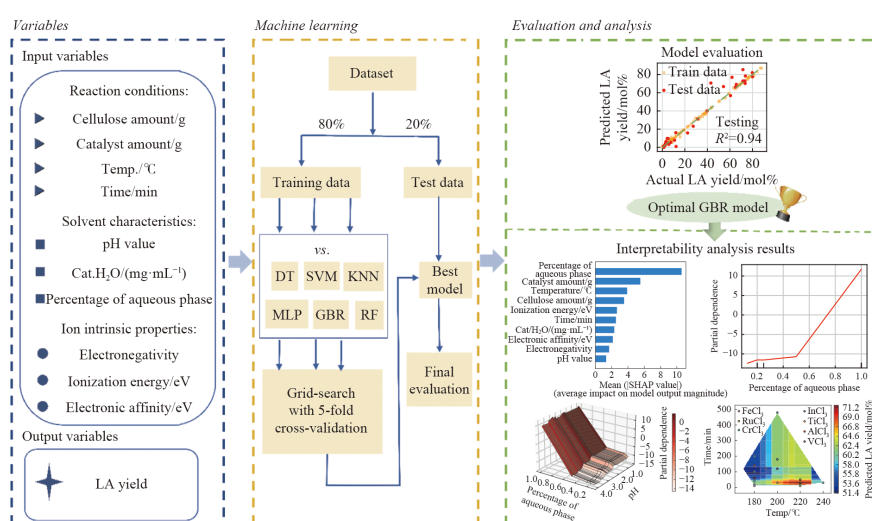


Machine learning-based prediction and optimization of the cellulose conversion process for levulinic acid production

ZHAO Huiting, XIE Yujiao*, XU Dongqian, DONG Fangxu, CUI Hongyou*

School of Chemistry and Chemical Engineering, Shandong University of Technology, Zibo 255000, China

Abstract: Levulinic acid (LA) is a promising platform product with wide industrial applications. In recent years, the efficient conversion of cellulose into LA has become a research hotspot. However, traditional experimental optimization methods are often time-consuming and inefficient, therefore limiting its further upgrading. In this study, we integrated multidimensional data, including reaction conditions, solvent properties, and



physicochemical characteristics of metal salts, to construct a systematic dataset. Six machine learning models, decision tree, Gradient Boosting Regression, K-Nearest Neighbors, multilayer perceptron, random forest, and support vector machine, were developed to predict LA yield. Among them, the gradient boosting regression model achieved the best performance, with a test-set determination coefficient of 0.94 and the lowest root-mean-square error. SHapley Additive exPlanations and partial dependence analysis revealed that water fraction, catalyst dosage, and reaction temperature were the key factors influencing LA formation. By integrating the gradient boosting model with the particle swarm optimization algorithm, RuCl_3 was identified as an efficient catalyst under high-temperature and short-reaction-time conditions. This study demonstrates the potential of applying the machine learning method in cellulose conversion research, and provides a data-driven strategy and theoretical guidance for the efficient and green production of LA.

Key words: machine learning; cellulose conversion; levulinic acid; metal salt catalysts; gradient boosting regression

With the rapid industrialization of modern human society, the energy crisis and environmental pollution have been further aggravated^[1]. To alleviate these issues, biomass, as the only renewable carbon source, has attracted increasing attention^[2]. Cellulose, the most widely distributed component of biomass, can be converted into a series of value-added chemicals and renewable fuels such as 5-hydroxymethylfurfural, levulinic acid (LA), and

polyols^[3]. Among them, LA has been listed as one of the “top 12 value-added platform chemicals” derived from biomass by the U.S. Department of Energy, owing to its multifunctional nature as a fundamental platform molecule. It has been widely employed in the synthesis of high-value chemicals, and further used as pharmaceuticals and solvents^[4–5]. At present, the main routes for cellulose conversion include enzymatic, thermochemical, and

Received: 2025-11-05; Revised: 2025-12-10

Foundation items: Supported by the National Natural Science Foundation of China (22508228, 22378238).

* Corresponding author. E-mail: yujiaoxie@sdtu.edu.cn, cuihy@sdtu.edu.cn.

DOI: 10.1016/S1872-5813(26)60665-2 CSTR: 32157.14.2097-213X.2026.JFCT.0020 Associated Data DOI: 10.57760/sciencedb.j00124.00342

本文的英文电子版由 Elsevier 出版社在 ScienceDirect 上出版 (<http://www.sciencedirect.com/science/journal/18725813>)

chemical methods^[6]. Compared with other routes, chemical approaches offer advantages such as ease of catalyst recovery and high reaction selectivity. Consequently, chemical methods are proposed as promising technology which have been widely studied.

Metal salts, owing to their low cost, high thermal and chemical stability, and strong catalytic activity, have been widely recognized as efficient catalysts for the conversion of cellulose into LA. In many studies, they have obtained relatively high LA yields under suitable conditions. However, the production efficiency of LA is not determined by the catalyst alone; it is strongly influenced by a range of reaction parameters, including solution pH, reaction temperature, reaction duration, catalyst loading, and the composition of the solvent system. Understanding the interplay among these factors is crucial for optimizing LA yield. For instance, catalyst amount has been repeatedly reported as a key parameter. In general, increasing catalyst concentration facilitates the hydrolysis and subsequent transformation of cellulose-derived intermediates, thereby enhancing the LA yield. Nevertheless, once the catalyst concentration surpasses a certain threshold, further increases often fail to improve performance significantly, suggesting that the reaction becomes limited by other kinetic or thermodynamic factors^[7]. Reaction time is another influential variable. Boonyakarn et al.^[8] demonstrated that prolonged reaction times initially promote higher LA yields, but when the duration exceeded approximately 40 min, a rapid decline occurred, likely due to the degradation of LA into by-products such as formic acid, char, or humins. Reaction temperature has also been highlighted as a decisive factor. Within moderate ranges, elevated temperatures accelerate the conversion kinetics and drive cellulose toward LA formation. For example, one study observed continuous yield enhancement between 145 and 175 °C^[9]. Moreover, under specific thermal conditions, extending the reaction time beyond a critical value can shift the product distribution in favor of LA, establishing it as the dominant product^[10]. Solvent systems were also shown to play a critical role. For instance, increasing the water content in the solvent was favorable for LA formation^[11]. Despite these insights, conventional approaches to studying the dependence of LA yield on reaction conditions are still largely empirical. They typically require extensive experimental trials, which are not only time-consuming and labor-intensive but also resource-intensive. More critically, such trial-and-error strategies do not guarantee the identification of the global optimum yield or the true set of ideal operating conditions. This underscores the need for innovative, systematic, and efficient methodologies, potentially integrating experimental design, statistical modeling, and machine learning (ML), to unravel the complex parameter interactions and to accelerate the discovery of optimized conditions for LA production from cellulose.

With the rapid advancement of artificial intelligence, ML has emerged as a powerful tool for addressing complex nonlinear problems in chemical and energy systems. A wide

variety of ML models—such as Decision Trees (DT), Gradient Boosting Regression (GBR), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), Random Forests (RF), and Support Vector Machines (SVM)—have been successfully applied to predict process parameters and optimize reaction conditions in energy-related research^[12–17]. For example, Leng et al.^[18] employed extreme gradient boosting, MLP, SVM and RF to predict the yields of three-phase products and the heating value of bio-oil from biomass pyrolysis. Among these, RF demonstrated the highest prediction accuracy and generalization ability, highlighting its robust predictive capacity. In another study, SVM was found to provide relatively high accuracy and robustness when predicting coke yield from the co-pyrolysis of biomass and waste plastics^[19]. Comparative studies have also been carried out that when SVM, RF and GBR were evaluated for predicting outcomes of biomass microwave pyrolysis, GBR was shown to outperform the others, offering more reliable predictions and valuable insights for experimental planning and optimization^[20]. Similarly, GBR has been applied to predict bio-oil yield, enabling not only accurate outcome forecasting but also the identification of critical influencing factors and optimal conditions for maximizing yield^[21]. In another study on biomass pyrolysis, Shen et al.^[22] applied seven ML models, including artificial neural networks, GBR, DT, RF, KNN, bagging regression, and Lasso regression, to predict the yields of bio-oil, pyrolytic gas, and biochar. The results revealed that all models except Lasso exhibited good predictive capabilities, with GBR performing the best, further underscoring its reliability. Beyond standalone applications, the integration of ML with other methodologies has also shown strong potential. For example, combining artificial neural networks with response surface methodology to optimize the torrefaction process of coal, biomass, and waste enabled the determination of optimal operating conditions and fuel properties, reducing costs and improving the overall efficiency of energy production^[23]. In summary, machine learning has been widely applied in biomass conversion research, fully demonstrating its advantages in handling complex multi-factor systems, reducing experimental costs, and elucidating the relationships between variables and products. The cellulose conversion process is likewise influenced by multiple interacting parameters, such as reaction temperature, reaction time, aqueous-phase fraction, and the type and loading of metal-salt catalysts, that are difficult to comprehensively investigate using traditional experimental approaches. Meanwhile, no studies to date have applied machine learning to predict the reaction outcomes of metal-salt-catalyzed cellulose conversion. Therefore, this study employs machine learning methods to predict and optimize the catalytic conversion of cellulose for the production of LA.

In this work, we constructed a dataset for cellulose conversion to LA by considering the properties of metal salts, solvent characteristics, and reaction conditions. Six ML models—DT, GBR, KNN, MLP, RF and SVM—were

developed and compared. Subsequently, the optimal model was further interpreted using SHAP analysis and partial dependence analysis, which revealed the interactions among key variables and their potential influence on conversion efficiency. Furthermore, by integrating the ML model with particle swarm optimization (PSO), we achieved inverse optimization of input variables, thereby identifying the optimal conditions for maximizing LA yield. This study provides new insights into the cellulose-to-LA conversion process, and establishes a novel data-driven framework to guide future experimental design and process optimization.

1 Models and methods

1.1 Data collection and pre-processing

An initial literature search was conducted using the keywords “cellulose”, “LA” and “metal salts”, followed by screening to identify studies relevant to this work. The input data of the dataset were collected from these published articles, while most of the target data were extracted from figures and tables in the literature using digital visualization tools. After data cleaning and integration, a refined dataset relevant to LA formation was constructed, comprising 10 input variables and 1 target variable. The dataset serves as the basis for subsequent model development and analysis.

This dataset encompasses multiple categories of input variables, including reactant amount, solvent properties, and reaction conditions. The reactant amount includes cellulose amount (g) and catalyst amount (g). Solvent properties consist of pH, Cat./H₂O ratio (mg/mL), and percentage of aqueous phase. Reaction conditions include reaction temperature and reaction time. In addition, electronegativity, ionization energy (eV), and electron affinity (eV) were selected as universal physicochemical descriptors of the metal salts, as these parameters are consistently available across different metal salts and therefore provide a unified basis for comparison. Ultimately, the selected input variables for this study were: cellulose amount (g), catalyst amount (g), Cat./H₂O ratio, pH, percentage of aqueous phase (volume fraction of water in the solvent), temperature, reaction time, electronegativity, ionization energy (eV), and electron affinity (eV), with LA yield defined as the target variable.

Box plots were first employed to provide a statistical description of the dataset, visually representing the quartiles, mean, median, as well as minimum and maximum values of each variable. Subsequently, the Pearson correlation coefficient (PCC) was used to evaluate the correlation between any two variables. The PCC is calculated according to Equation (1):

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}) \sum_{i=1}^n (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

where r denotes the PCC and $\bar{x}$ and $\bar{y}$ represent the mean values of variables x and y respectively. The significance test of the correlation coefficient between any two variables was carried out using Equation (2):

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \quad (2)$$

where N represents the total number of samples.

In this study, PCC was applied to detect the linear relationships among variables. The value of PCC ranges from -1 to 1 , with a larger absolute value indicating a stronger linear correlation between the two variables.

1.2 Machine learning methods

In this study, multiple ML models were employed to predict LA yield, specifically including DT, GBR, KNN, MLP, RF and SVM. These models were implemented using open-source libraries in Python 3.9.12. The dataset was randomly divided into a training set (80%) and a test set (20%). Given the relatively small sample size, hyperparameter tuning was performed using 5-fold cross-validation on the training set, while the independent test set was used for external validation. Consequently, no additional validation set was created to avoid further diluting the available data^[24]. Specifically, DT predicts outputs by recursively partitioning data and generating decision rules, without relying on predefined input-output relationships. Its key hyperparameter is the minimum leaf node size, which can be tuned *via* grid search. Since DT is prone to overfitting, pruning techniques are usually applied for correction. GBR, a boosting method, iteratively trains multiple decision trees to continuously correct prediction errors, thereby forming a strong predictive model. This method minimizes errors using gradient descent and, due to the randomized selection of training data, is less susceptible to overfitting^[25]. KNN is a nonparametric, instance-based approach that predicts outputs by calculating distances between samples and selecting the K nearest neighbors (taking their average in regression). KNN is simple and intuitive but may perform poorly with high-dimensional data. MLP consists of an input layer, one or more hidden layers, and an output layer. By iteratively updating weights through the backpropagation algorithm, it can approximate complex nonlinear relationships. MLP is suitable for modeling high-dimensional nonlinear problems; however, its training process is computationally demanding and may converge to local minima. RF is an ensemble learning method that constructs multiple decision trees using bootstrap aggregating from the training dataset and averages their outputs. This reduces overfitting risks and enhances model robustness. Its main hyperparameters include the number of trees and maximum depth. SVM maps data into a high-dimensional space through kernel functions and seeks the optimal hyperplane for classification or regression. SVM is well-suited for small-sample and nonlinear problems but can be computationally expensive for large

datasets.

1.3 Evaluation indicators

To evaluate the performance of the six models, two statistical indicators were introduced: the coefficient of determination (R^2) and the root mean square error (RMSE). Their calculation formulas are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^N (Y_i^{\text{exp}} - Y_i^{\text{pred}})^2}{\sum_{i=1}^N (Y_i^{\text{exp}} - Y_{\text{ave}}^{\text{exp}})^2} \quad (3)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_i^{\text{exp}} - Y_i^{\text{pred}})^2} \quad (4)$$

where N is the total number of samples, Y_i^{exp} is the experimental value, Y_i^{pred} is the predicted value, and $Y_{\text{ave}}^{\text{exp}}$ represents the mean of all experimental values. When R^2 is close to 1 and RMSE is close to 0, good predictive performance can be achieved.

1.4 ML model-based feature analysis and inverse design

To systematically investigate the influence of each input variable on the target output, this study employed SHapley Additive exPlanations (SHAP) to interpret the predictions of the trained machine learning model. SHAP is a widely adopted technique for explaining machine learning model predictions, with its theoretical foundation rooted in Shapley values from game theory^[26]. By assigning contribution values to each feature, SHAP quantifies the positive or negative impact of individual input variables on the model predictions, thereby significantly enhancing the interpretability and transparency of complex models. Furthermore, to strengthen the practical utility of the model, this study integrated PSO for the inverse design of the cellulose conversion process. In the inverse optimization procedure, PSO first generates an initial population of candidate solutions randomly within the defined ranges of each input variable, with each particle representing a potential solution and evaluated according to the objective function. Subsequently, particles iteratively update their positions based on their personal best and the global best positions, gradually converging to the optimal solution^[27]. By combining the trained machine learning model with the PSO algorithm, the inverse optimization of input variables can be realized that identified the conditions of the maximize LA yield. And furthermore providing a reliable, data-driven guidance for experimental design.

2 Results and discussion

2.1 Data set description and correlation analysis

Figure 1 presents the distribution of input variables in the

dataset using box plots. The ranges of cellulose amount, catalyst amount, and Cat./H₂O ratio are 0.1–2, 0.008–4.513 and 0.79–103.7, with median values of 0.83, 0.158 and 9.74, respectively (Figure 1(a)–(c)). The presence of outliers in these variables can be attributed to variations in cellulose and catalyst loadings adopted across different experimental studies^[28], which reflects the diversity of reaction conditions. For pH and percentage of aqueous phase, the values are mainly distributed within 2.08–3.12 and 0.2–1, with median values of 2.87 and 1 (Figure 1(d) and (e)). These results indicate that cellulose-to-LA conversion generally occurs under acidic conditions, with a single aqueous phase typically employed as the solvent. Similarly, reaction temperature and reaction time are predominantly concentrated in the ranges of 180–220 °C and 30–120 min, with median values of 180 °C and 60 min, respectively (Figure 1(f) and (g)). The distributions of electronegativity, ionization energy, and electron affinity fall mainly within 1.54–1.58, 16.89–21.09 and 5.92–7.32, respectively, reflecting the diversity of metal salts applied in cellulose-to-LA conversion^[29]. Overall, the box plot analysis suggests that the distributions of all variables are well balanced—neither overly concentrated nor excessively dispersed. This indicates that the dataset is appropriate for machine learning-based modeling and predictive analysis.

To further elucidate the linear relationships among the variables, PCC analysis was performed, and the resulting heatmap is presented in Figure 2. The LA yield was positively correlated with several factors, including the percentage of aqueous phase, reaction temperature, cellulose loading, catalyst amount, Cat./H₂O, and electronegativity. Notably, the correlation coefficient between LA yield and the percentage of aqueous phase was 0.79, indicating that a predominantly aqueous environment favors higher LA production^[30]. This observation aligns with the fact that LA formation generally requires a strongly acidic environment, which is more readily established in a pure aqueous system. Similarly, the correlation coefficient with reaction temperature was 0.67, suggesting that elevated temperatures facilitate the conversion of cellulose to LA^[31]. Moderate positive correlations were also observed for cellulose and catalyst loadings, with coefficients of 0.35 and 0.28, respectively, which suggests that increasing the amounts of feedstock and catalyst can moderately enhance LA yield. Conversely, LA yield was negatively correlated with pH, reaction time, ionization energy, and electron affinity, with correlation coefficients of −0.32 and −0.17 for pH and reaction time, and −0.33 and −0.14 for ionization energy and electron affinity, respectively. In general, PCC values greater than 0.8 are regarded as strong correlations; thus, no strong correlations were observed among the variables in this dataset, indicating suitability for ML model development. In addition, PCC was employed primarily to identify preliminary linear relationships among variables, whereas the final interpretation of the underlying influence mechanisms was based on the nonlinear explanations provided by SHAP.

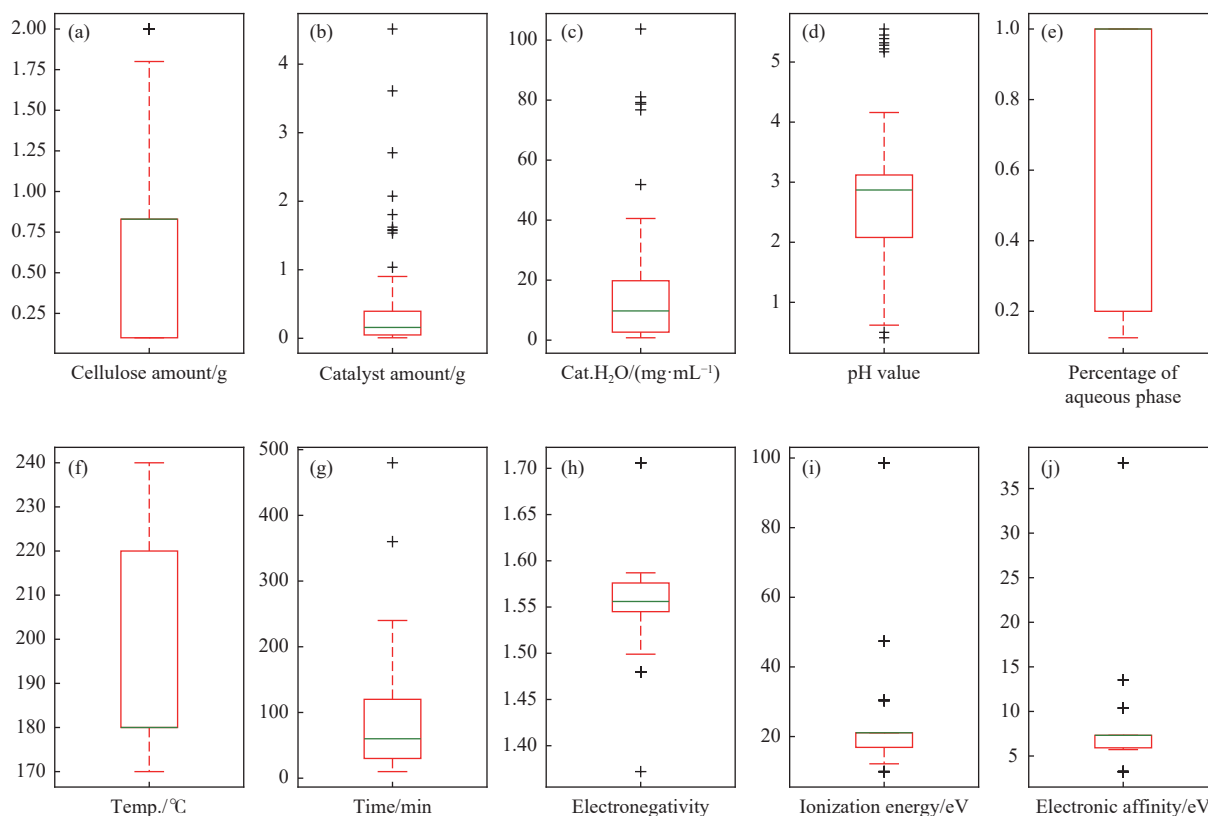


Fig. 1 Box-plot of the dataset

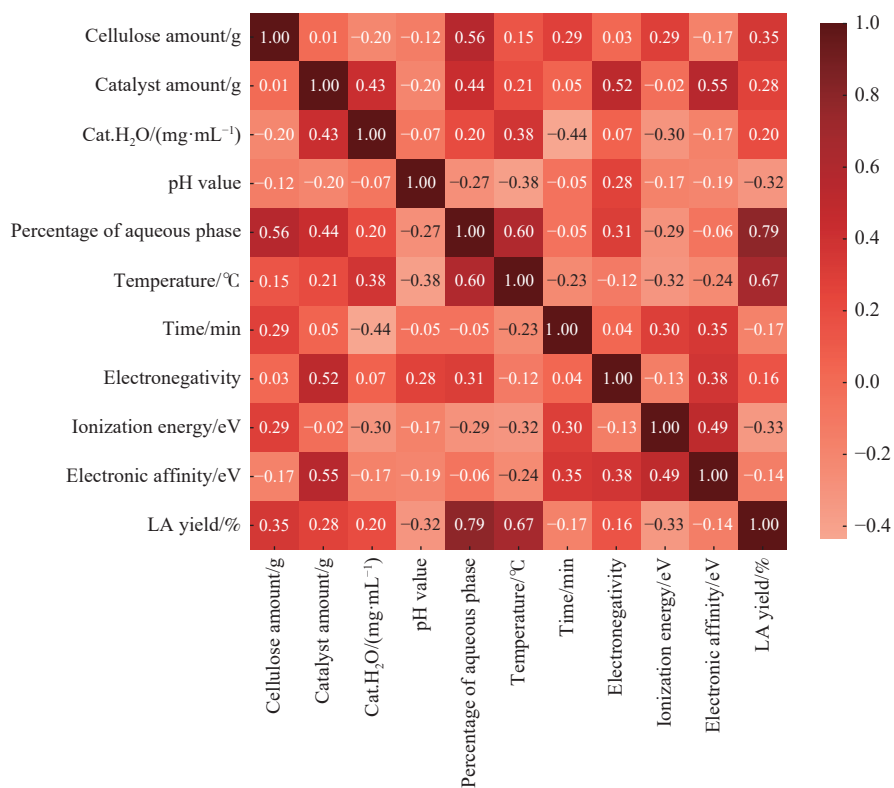


Fig. 2 Pearson correlation coefficient matrix

2.2 Hyper-parameter tuning

Hyperparameter optimization plays a crucial role in the improvement of the predictive performance, robustness, and generalization ability of ML models. The tuning results for the six models are presented in Figure 3. For DT and GBR (Figure 3(a) and (b)), the results showed that DT achieved the best performance at $\text{max_depth} = 12$ and $\text{min_samples_split} = 8$. This suggests that moderately increasing tree

depth enhances the model's ability to capture complex nonlinear relationships, whereas excessive depth may lead to overfitting. By contrast, the performance of the GBR model improved steadily with increasing max_depth but leveled off at a depth of around 6. This indicates that the nonlinear complexity of the reaction system is moderate, and that GBR can model it effectively without requiring excessive tree depth, reflecting an appropriate alignment between model capacity and system complexity.

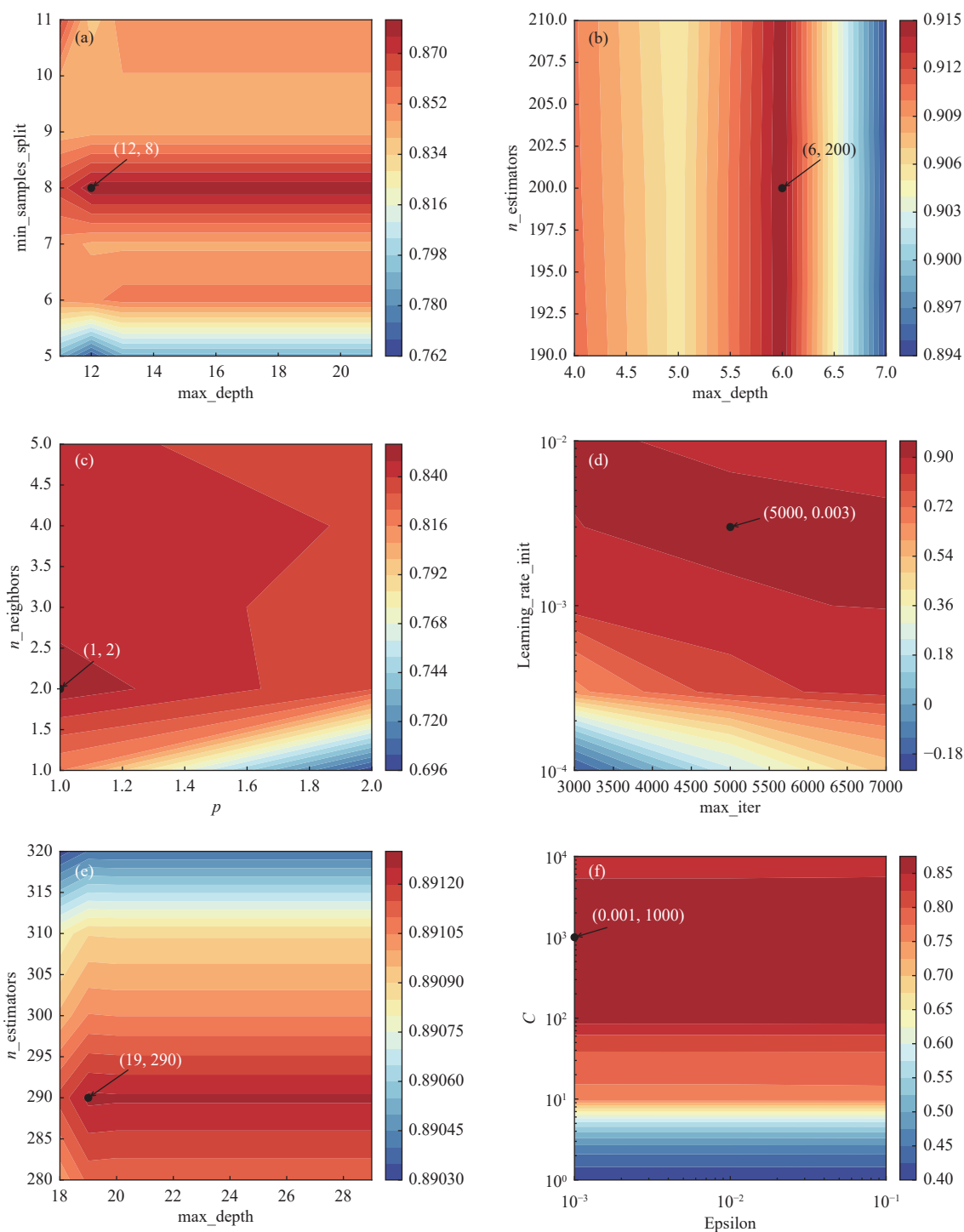


Fig. 3 Hyper-parameter tuning for (a) DT, (b) GBR, (c) KNN, (d) MLP, (e) RF and (f) SVM for LA yield prediction based on dataset

Moreover, the effect of $n_estimators$ on GBR accuracy was relatively limited, reflecting that ensemble models can converge with a moderate number of weak learners^[32]. For KNN and MLP (Figure 3(c) and (d)), KNN performed best with $p = 1$ and $n_neighbors = 2$, suggesting that the data are mainly governed by local patterns and that accurate predictions can be achieved using only a small number of neighboring samples. MLP achieved the highest R^2 at $max_iter = 5000$ and $learning_rate_init = 0.003$, highlighting the importance of sufficient iterative training and appropriate learning rates for convergence and performance improvement in complex nonlinear mapping tasks^[33]. The tuning results for RF and SVM are shown in Figure 3(e) and (f). RF achieved optimal performance at $max_depth = 19$ and $n_estimators = 290$, indicating that moderately increasing both tree depth and ensemble size

can enhance model robustness and generalization ability. For SVM, the average R^2 increased with larger C values, reaching its peak at approximately $C = 1000$. While larger C values reduce the risk of underfitting, excessively large values may lead to overfitting^[34].

2.3 Prediction performance of models

After hyperparameter optimization of the ML models, the dataset was randomly divided into a training set (80 %) and a testing set (20%) to comprehensively evaluate the predictive capability of the six models for LA yield. Figure 4 illustrates the comparison between the experimental values (x -axis) and the predicted values (y -axis) generated by the models. The green dashed line represents the ideal scenario where predictions perfectly match the experimental results.

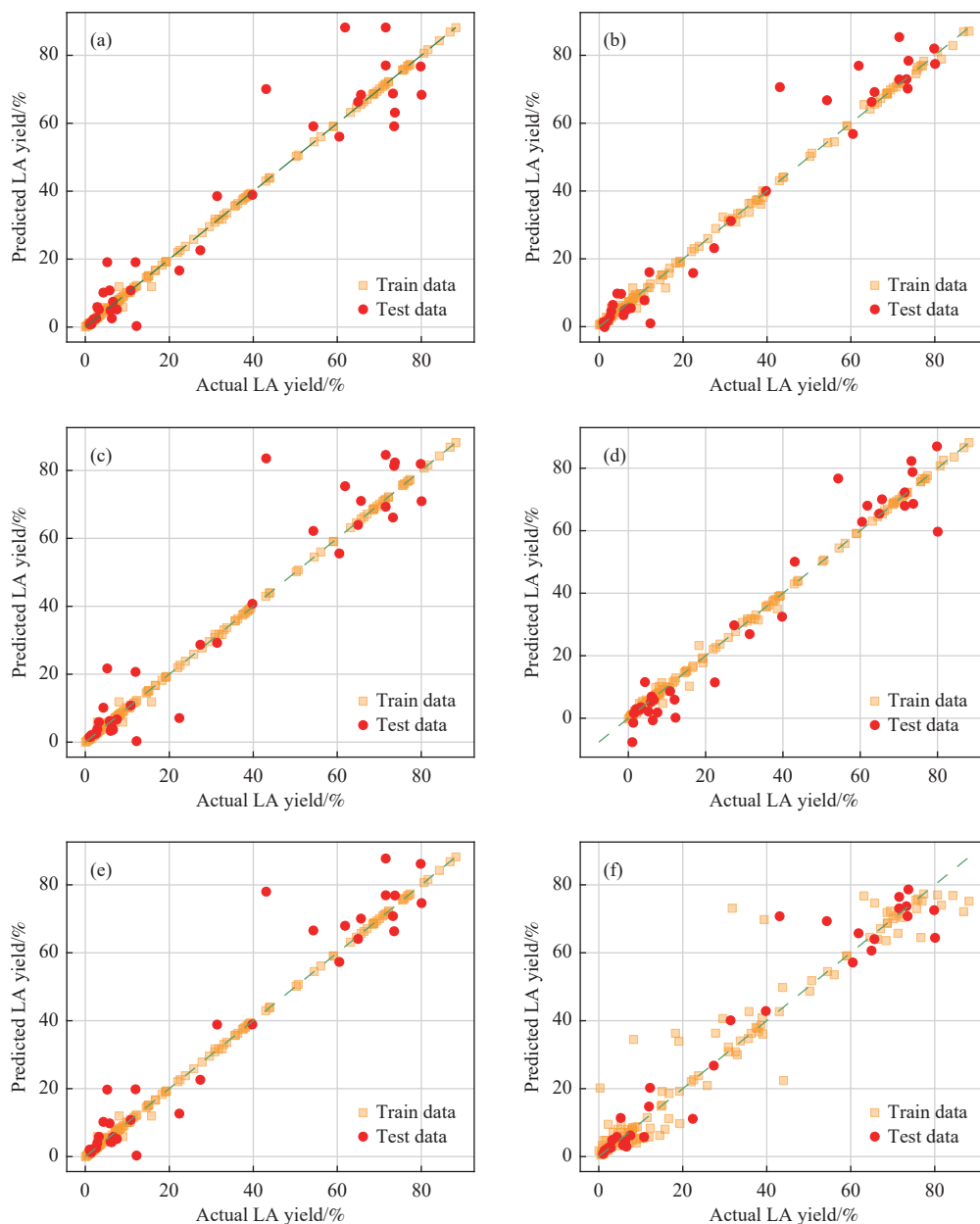


Fig. 4 Optimal model performance of (a) DT, (d) GBR, (c) KNN, (d) MLP, (e) RF and (f) SVM

As observed, the data points in the testing set are more scattered than those in the training set, indicating that a common phenomenon in machine learning—models generally perform better on training data than on unseen testing data^[35–36]. Table 1 summarizes the R^2 and RMSE values for both training and testing datasets. The results show that the SVM model exhibited inferior performance during training compared to the other five models, all of which achieved excellent fits with $R^2 = 0.99$ on the training set. However, performance varied across models in the testing set: DT, KNN, and RF yielded R^2 values of 0.90, 0.89 and 0.92, respectively, whereas GBR and MLP achieved superior predictive accuracy (Testing $R^2 = 0.94$). The RMSE comparison further revealed that GBR slightly outperformed MLP. Overall, all six models successfully predicted LA yield, with performance trends based on R^2 as follows: Training: DT = GBR = KNN = MLP = RF > SVM; Testing: GBR = MLP > SVM > RF > DT > KNN. Notably, the GBR model demonstrated strong robustness across both training and testing evaluations.

2.4 Shapley additional explanation analysis

To further explore the influence of input variables on LA yield, SHAP analysis was performed using the best-performing GBR model, and the results are presented in Figure 5. In the plots, red and blue represent high and low feature values, respectively. SHAP values quantify the

contribution of each feature to the model prediction, with positive values indicating a positive effect and negative values indicating a negative effect on LA yield^[37–38]. The analysis revealed that the three most influential variables were the percentage of aqueous phase, catalyst amount, and reaction temperature. For the percentage of aqueous phase, SHAP values increased progressively as the data points transitioned from blue to red, indicating a strong positive correlation with LA yield. This finding is consistent with the PCC analysis and further emphasizes the critical role of a predominantly aqueous environment in promoting LA production. Reaction temperature exhibited a similar trend, with higher temperatures contributing positively to LA formation. It is noteworthy that the influence of catalyst loading is more complex and does not follow a simple monotonic trend. Both low and high catalyst loadings can lead to relatively high LA yields. At low loadings, an increase in the number of acid sites facilitates the stepwise cleavage and conversion of cellulose and its hydrolysis intermediates, resulting in higher LA yields as the catalyst amount increases. At higher loadings, the higher density of acid sites can rapidly promote the hydrolysis of 5-hydroxymethylfurfural (5-HMF) to LA, thereby reducing the condensation and char formation of 5-HMF at elevated temperatures. Consequently, high yields may appear in distinct loading regions, giving rise to the observed nonlinear relationship. This suggests that although the catalyst plays a crucial role in cellulose conversion, its effectiveness is modulated by the synergistic effects of other reaction parameters.

Table 1 Comparative evaluation of predictive performance of six ML models

Model	Training		Testing	
	R^2	RMSE	R^2	RMSE
DT	0.99	0.64	0.90	9.16
GBR	0.99	0.67	0.94	4.33
KNN	0.99	0.64	0.89	5.88
MLP	0.99	1.01	0.94	5.22
RF	0.99	0.64	0.92	5.51
SVM	0.93	7.11	0.93	4.57

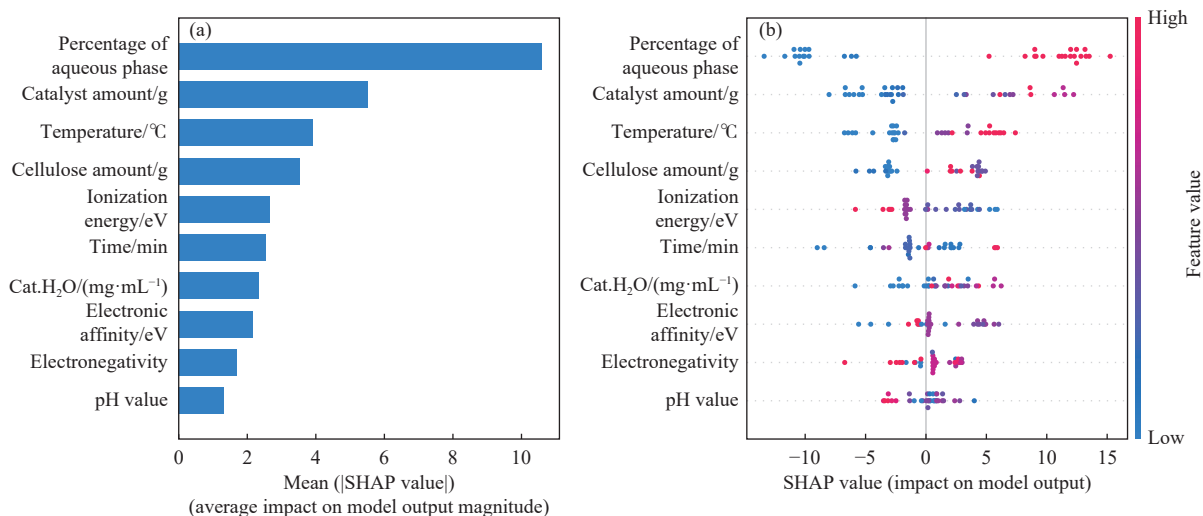


Fig. 5 SHAP value (a) bar plot, (b) beeswarm plot

2.5 Partial dependence plots

To visually elucidate the influence of input variables on the target outcome, partial dependence analysis was employed to investigate both individual and interactive effects on LA yield. Figure 6(a)–(e) presents the partial dependence plot (PDP) of LA yield predicted by the GBR model under its optimal predictive performance. The tick marks along the *x*-axis represent the values of the target features and reflect data density; in regions with sparse data, the trend lines may be less reliable and should be interpreted with caution^[24,39]. As shown in Figure 6(a), LA yield reaches its maximum when the percentage of aqueous phase increases to 1, consistent with previous reports that a single aqueous phase is more favorable for LA production than a biphasic system^[40]. Regarding catalyst amount

(Figure 6(b)), LA yield increases with catalyst amount and peaks at 0.5 g. This observation can be attributed to the higher number of active sites at elevated catalyst loadings, which facilitates the hydrolysis of 5-HMF into LA, ultimately increasing its concentration in the reaction mixture^[41]. Furthermore, LA yield is higher under acidic conditions, reaching a maximum at pH value = 2 (Figure 6(c)). To further explore variable interactions, two-variable partial dependence analysis was conducted. Figure 6(d) illustrates that when the percentage of aqueous phase exceeds 0.6, LA yield is most sensitive to catalyst amount in the 0.5–1 g range. Simultaneously, with pH held constant, LA yield gradually increases as the percentage of aqueous phase surpasses 0.6, indicating a synergistic effect of these two variables on the target yield (Figure 6(e)).

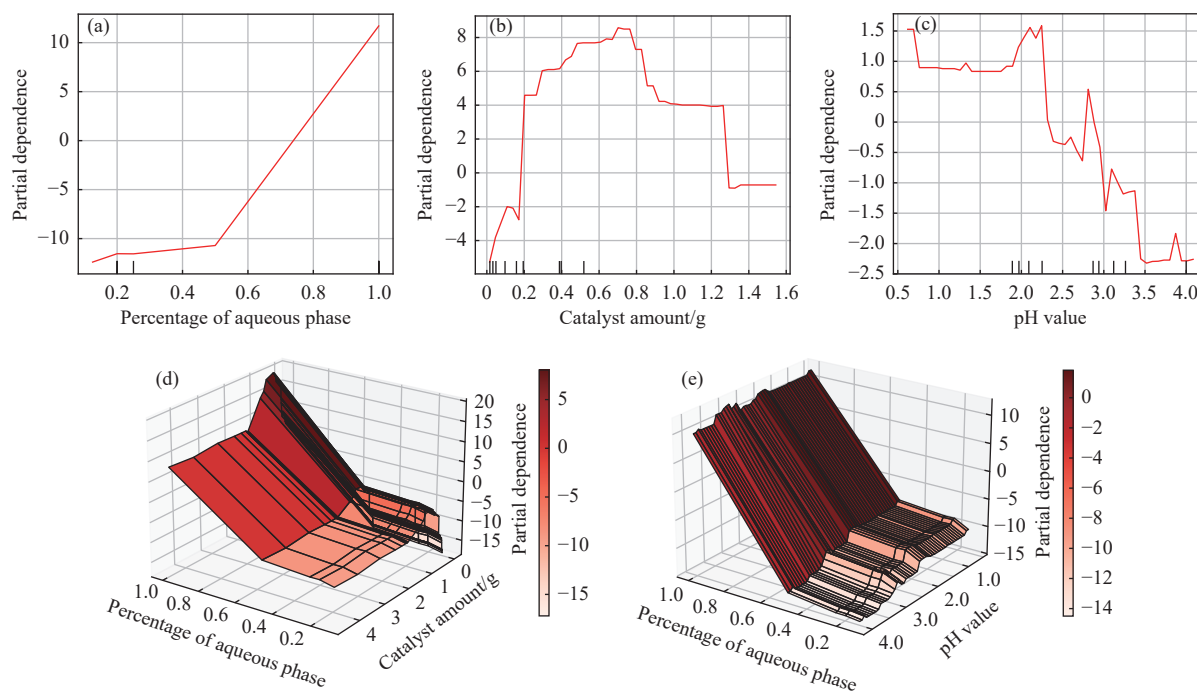


Fig. 6 Partial dependence plots for LA yield prediction: one-way (a)–(c) and two-way (d)–(e) analysis

2.6 Model-based inverse design for guiding experiments

Based on the best-performing predictive model (GBR), contour plots were generated and annotated with different types of metal salts to systematically evaluate their effects on LA yield. To prevent the PSO algorithm from exploring sparsely populated regions of the dataset, the optimization ranges for all input variables were strictly constrained within the actual upper and lower limits reported in the literature, ensuring that the entire search space remained within regions where the model exhibits high confidence. Meanwhile, the tree-based partitioning mechanism of the GBR model further reduces extrapolation errors in sparse regions. Therefore, the results of the optimization and contour analyses are intended primarily to elucidate general

trends rather than to prescribe specific experimental conditions. Specifically, a two-dimensional contour map was first constructed using reaction temperature and aqueous phase ratio as the *x*- and *y*-axes, respectively (Figure 7(a)). Notably, RuCl_3 was distributed within the red high-yield region, followed by CrCl_3 , whereas a considerable portion of the FeCl_3 and AlCl_3 data points were mainly located in the blue low-yield region. This observation can be primarily ascribed to two factors. First, under relatively low-temperature conditions, the stepwise conversion of cellulose *via* sugars and 5-HMF to LA is kinetically constrained, thereby hindering sufficient reaction progress. Second, when the aqueous phase ratio is insufficient, the acid-catalyzed hydrolysis process is suppressed, which not only prevents intermediates from further converting into LA but also facilitates side reactions

such as condensation and polymerization, ultimately leading to a pronounced reduction in the efficiency of target product formation^[42]. To further elucidate the promotional effects of different metal salts on LA formation, a contour plot was constructed using reaction temperature and reaction time as variables (Figure 7(b)). The results show that under conditions of approximately 220 °C and relatively short reaction times, RuCl₃ corresponds to a high-yield region, indicating its ability to promote LA production under high-

temperature, short-duration conditions. Moreover, previous studies have reported that RuCl₃ exhibits a pronounced kinetic advantage in cellulose conversion, achieving an LA yield as high as 87.6% at 220 °C for 30 min^[40]. This finding is fully consistent with the model predictions and underscores the importance of appropriately tuning temperature and time to enhance catalytic selectivity and yield, thereby providing experimental guidance for optimizing metal-salt catalyst systems.

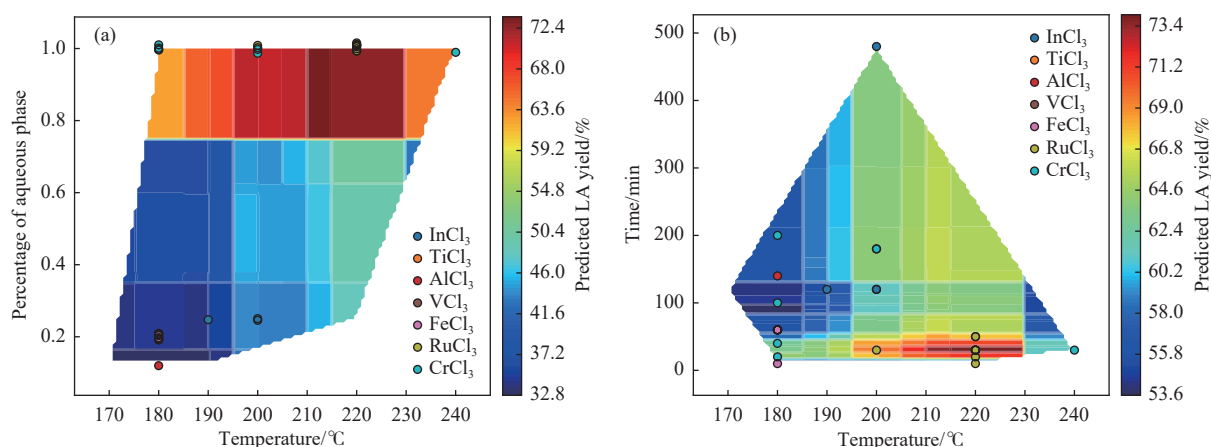


Fig. 7 Optimal results from the GBR model based inverse design of LA production

3 Conclusion

This study established a data-driven framework for predicting and optimizing the conversion of cellulose into LA. By integrating multi-source literature data and developing various ML models, the GBR model was identified as the optimal predictor, enabling accurate estimation of the reaction system. SHAP and partial dependence analyses revealed that the water fraction, catalyst dosage, and reaction temperature exert significant and synergistic influences on LA formation, elucidating the intrinsic regulatory relationships among key variables governing the yield. Furthermore, the integration of the GBR model with the PSO algorithm uncovered coupling effects between the physicochemical properties of metal salts and the reaction conditions. Overall, this study proposes a machine learning-assisted framework for catalytic system design, offering new data-driven insights and theoretical guidance for understanding the reaction mechanism and achieving the efficient and green production of LA from cellulose.

Declaration of competing interest

The authors declare no conflict of interest.

References

- [1] WU H, ZHANG R, ZHAI Y, et al. Solvent effects enable efficient tandem conversion of cellulose and its monosaccharides towards 5 - hydroxymethylfurfural[J]. ChemSusChem, 2023, 16(2): e202201809.
- [2] WANG J, LIU Y, WANG J, et al. Cooperative catalysis of carbon supported zinc salts hybrid for efficient conversion of fructose to ethyl lactate[J]. Chem Eng J, 2023, 468: 143670.
- [3] MA Q, GUAN M, FAN D, et al. Production of high-yield 5-hydroxymethylfurfural from crystalline cellulose via one-pot conversion in molten salt hydrate/acetone and separation[J]. Fuel, 2022, 324: 124678.
- [4] PILEIDIS F D, TITIRICI M M. Levulinic acid biorefineries: New challenges for efficient utilization of biomass[J]. ChemSusChem, 2016, 9(6): 562–582.
- [5] CLIMENT M J, CORMA A, IBORRA S. Conversion of biomass platform molecules into fuel additives and liquid hydrocarbon fuels[J]. Green Chem, 2014, 16(2): 516–547.
- [6] DING D, WANG J, XI J, et al. High-yield production of levulinic acid from cellulose and its upgrading to γ -valerolactone[J]. Green Chem, 2014, 16(8): 3846–3853.
- [7] DE GUZMAN D, DE LEON R. Preliminary optimization and kinetics of SnCl₂-HCl catalyzed hydrothermal conversion of microcrystalline cellulose to levulinic acid[J]. J Renewable Mater, 2021, 9(1): 145–162.
- [8] BOONYAKARN T, WATANIYAKUL P, BOONNOUN P, et al. Enhanced levulinic acid production from cellulose by combined brønsted hydrothermal carbon and lewis acid catalysts[J]. Ind Eng Chem Res, 2019, 58(8): 2697–2703.
- [9] WANG J, CUI H, WANG Y, et al. Efficient catalytic conversion of cellulose to levulinic acid in the biphasic system of molten salt hydrate and methyl isobutyl ketone[J]. Green Chem, 2020, 22(13): 4240–4251.
- [10] WEI C, LIU G, XIE Y, et al. Cellulose dissolution and

- conversion into 5-hydroxymethylfurfural in mixed molten salt hydrate[J]. *Cellulose*, 2023, 30(2): 801–813.
- [11] ZHAO Y, WANG S, LIN H, et al. Influence of a lewis acid and a Brønsted acid on the conversion of microcrystalline cellulose into 5-hydroxymethylfurfural in a single-phase reaction system of water and 1, 2-dimethoxyethane[J]. *RSC Adv*, 2018, 8(13): 7235–7242.
- [12] ZHAO C, LU X, JIANG Z, et al. Prediction of bio-oil yield by machine learning model based on “enhanced data” training[J]. *Renewable Energy*, 2024, 225: 120218.
- [13] ZHOU X, ZHAO J, CHEN M, et al. Effects of hydration parameters on chemical properties of biocrudes based on machine learning and experiments[J]. *Bioresour Technol*, 2022, 350: 126923.
- [14] POTNURI R, SURIAPPARAO D V, SANKAR RAO C, et al. Effect of dry torrefaction pretreatment of the microwave-assisted catalytic pyrolysis of biomass using the machine learning approach[J]. *Renewable Energy*, 2022, 197: 798–809.
- [15] WANG M, XIE Y, GAO Y, et al. Machine learning prediction of higher heating value of biochar based on biomass characteristics and pyrolysis conditions[J]. *Bioresour Technol*, 2024, 395: 130364.
- [16] ZHU X, LIU B, SUN L, et al. Machine learning-assisted exploration for carbon neutrality potential of municipal sludge recycling via hydrothermal carbonization[J]. *Bioresour Technol*, 2023, 369: 128454.
- [17] JIA L, SHAO W, WANG J, et al. Machine learning-aided prediction of bio-BTX and olefins production from zeolite-catalyzed biomass pyrolysis[J]. *Energy*, 2024, 306: 132478.
- [18] LENG E, HE B, CHEN J, et al. Prediction of three-phase product distribution and bio-oil heating value of biomass fast pyrolysis based on machine learning[J]. *Energy*, 2021, 236: 121401.
- [19] POTNURI R, RAO C S, SURYA D V, et al. Utilizing support vector regression modeling to predict pyro product yields from microwave-assisted catalytic co-pyrolysis of biomass and waste plastics[J]. *Energy Convers Manag*, 2023, 292: 117387.
- [20] YANG Y, SHAHBEIK H, SHAFIZADEH A, et al. Biomass microwave pyrolysis characterization by machine learning for sustainable rural biorefineries[J]. *Renewable Energy*, 2022, 201: 70–86.
- [21] SINHA S, SANKAR RAO C, KUMAR A, et al. Exploring and understanding the microwave-assisted pyrolysis of waste lignocellulose biomass using gradient boosting regression machine learning model[J]. *Renewable Energy*, 2024, 231: 120968.
- [22] SHEN J, YAN M, FANG M, et al. Machine learning-based modeling approaches for estimating pyrolysis products of varied biomass and operating conditions[J]. *Bioresour Technol Rep*, 2022, 20: 101285.
- [23] OZONOH M, OBOIRIEN B O, DARAMOLA M O. Optimization of process variables during torrefaction of coal/biomass/waste tyre blends: Application of artificial neural network & response surface methodology[J]. *Biomass Bioenergy*, 2020, 143: 105808.
- [24] ZHU X, LI Y, WANG X. Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions[J]. *Bioresour Technol*, 2019, 288: 121527.
- [25] NIE P, ROCCOTELLI M, FANTI M P, et al. Prediction of home energy consumption based on gradient boosting regression tree[J]. *Energy Rep*, 2021, 7: 1246–1255.
- [26] YAN L, ZHONG S, IGOU T, et al. Development of machine learning models to enhance element-doped g-C₃N₄ photocatalyst for hydrogen production through splitting water[J]. *Int J Hydrog Energy*, 2022, 47(80): 34075–34089.
- [27] GAD A G. Particle swarm optimization algorithm and its applications: A systematic review[J]. *Arch Comput Methods Eng*, 2022, 29(5): 2531–2561.
- [28] PENG L, LIN L, ZHANG J, et al. Catalytic conversion of cellulose to levulinic acid by metal chlorides[J]. *Molecules*, 2010, 15(8): 5258–5272.
- [29] ANTONETTI C, LICURSI D, FULIGNATI S, et al. New frontiers in the catalytic synthesis of levulinic acid: From sugars to raw and waste biomass as starting feedstock[J]. *Catalysts*, 2016, 6(12): 196.
- [30] YAN L, LASKAR D D, LEE S J, et al. Aqueous phase catalytic conversion of agarose to 5-hydroxymethylfurfural by metal chlorides[J]. *RSC Adv*, 2013, 3(46): 24090–24098.
- [31] ZHI Z, LI N, QIAO Y, et al. Kinetic study of levulinic acid production from corn stalk at relatively high temperature using FeCl₃ as catalyst: A simplified model evaluated[J]. *Ind Crops Prod*, 2015, 76: 672–680.
- [32] BOLDINI D, GRISONI F, KUHN D, et al. Practical guidelines for the use of gradient boosting for molecular property prediction[J]. *J Cheminformatics*, 2023, 15(1): 73.
- [33] WANG X, WANG S, LIANG X, et al. Deep reinforcement learning: A survey[J]. *IEEE Trans Neural Netw Learn Syst*, 2024, 35(4): 5064–5078.
- [34] CZARNECKI W M, PODLEWSKA S, BOJARSKI A J. Robust optimization of SVM hyperparameters in the classification of bioactive compounds[J]. *J Cheminformatics*, 2015, 7(1): 38.
- [35] RAINIO O, TEUHO J, KLÉN R. Evaluation metrics and statistical tests for machine learning[J]. *Sci Rep*, 2024, 14(1): 6086.
- [36] LI J, PAN L, SUVARNA M, et al. Fuel properties of hydrochar and pyrochar: Prediction and exploration with machine learning[J]. *Appl Energy*, 2020, 269: 115166.
- [37] LIU H, LIU K, ZHU H, et al. Explainable machine-learning predictions for catalysts in CO₂-assisted propane oxidative dehydrogenation[J]. *RSC Adv*, 2024, 14(11): 7276–7282.
- [38] ONSREE T, TIPPAYAWONG N, PHITHAKKITNUKON S, et al. Interpretable machine-learning model with a collaborative game approach to predict yields and higher heating value of torrefied biomass[J]. *Energy*, 2022, 249: 123676.
- [39] APLEY D W, ZHU J. Visualizing the effects of predictor variables in black box supervised learning models[J]. *J R Stat Soc Ser B Stat Methodol*, 2020, 82(4): 1059–1086.

- [40] YAN L, MA R, WEI H, et al. Ruthenium trichloride catalyzed conversion of cellulose into 5-hydroxymethylfurfural in biphasic system[J]. Bioresour Technol, 2019, 279: 84–91.
- [41] WEI W, WU S. Depolymerization of cellulose into high-value chemicals by using synergy of zinc chloride hydrate and sulfate ion promoted titania catalyst[J]. Bioresour Technol, 2017, 241: 760–766.
- [42] GIRISUTA B, DANON B, MANURUNG R, et al. Experimental and kinetic modelling studies on the acid-catalysed hydrolysis of the water hyacinth plant to levulinic acid[J]. Bioresour Technol, 2008, 99(17): 8367–8375.

基于机器学习的纤维素转化制备乙酰丙酸过程预测与优化

赵慧婷，解玉姣^{*}，徐东倩，董方旭，崔洪友^{*}

(山东理工大学 化学化工学院, 山东 淄博 255000)

摘要: 乙酰丙酸(LA)是一种具有重要应用前景的平台化合物,近年来,纤维素高效转化制备 LA 成为研究热点。然而,传统实验优化方法耗时且效率低,限制了产率提升。本研究整合反应条件、溶剂特性及金属盐理化属性等多维数据,构建系统化数据集,并建立决策树、梯度提升回归、K-近邻、多层感知机、随机森林以及支持向量机六种机器学习模型,用于预测 LA 产率。结果表明,梯度提升回归模型表现最佳,测试集决定系数达 0.94,且均方根误差最低。SHapley 可加性解释与偏依赖分析表明,水相百分比、催化剂用量及反应温度是生成 LA 的关键因素。将梯度提升回归模型和粒子群优化算法结合,发现 RuCl_3 在高温短时条件下表现出优异催化性能。本研究揭示了机器学习在纤维素转化研究中的应用潜力,为 LA 的高效绿色制备提供了新的数据驱动策略与理论指导。

关键词: 机器学习;纤维素转化;乙酰丙酸;金属盐催化剂;梯度提升回归

中图分类号: TQ352.23

文献标识码: A

作者简介:



赵慧婷(2000),硕士研究生,研究方向为纤维素转化及数据驱动优化,1772454712@qq.com。



崔洪友(1968),博士,教授,研究方向为生物质能源与精细化学品清洁化工研究,cuihy@sdut.edu.cn。



解玉姣(1986),博士,副教授,研究方向为生物质高值化利用与化工热力学,yujiaoxie@sdut.edu.cn。