Contents lists available at ScienceDirect

Environmental Science and Ecotechnology

journal homepage: www.journals.elsevier.com/environmental-science-and-ecotechnology/

Original Research

Data-driven interpretable analysis for polysaccharide yield prediction

Yushi Tian [a, **], Xu Yang [a], Nianhua Chen [a], Chunyan Li [a], Wulin Yang [b, *]

[a] School of Resource and Environment, Northeast Agriculture University, Harbin, 150030, PR China
[b] College of Environmental Sciences and Engineering, Peking University, Beijing, 100871, PR China

ARTICLE INFO

Article history:
Received 9 February 2023
Received in revised form
17 September 2023
Accepted 17 September 2023

Keywords:
Cornstalk
Xylanase
Polysaccharide yield prediction
Machine learning
Model interpretability

ABSTRACT

Cornstalks show promise as a raw material for polysaccharide production through xylanase. Rapid and accurate prediction of polysaccharide yield can facilitate process optimization, eliminating the need for extensive experimentation in actual production to refine reaction conditions, thereby saving time and costs. However, the intricate interplay of enzymatic factors poses challenges in predicting and optimizing polysaccharide yield accurately. Here, we introduce an innovative data-driven approach leveraging multiple artificial intelligence techniques to enhance polysaccharide production. We propose a machine learning framework to identify highly accurate polysaccharide yield prediction modeling methods and uncover optimal enzymatic parameter combinations. Notably, Random Forest (RF) and eXtreme Gradient Boost (XGB) demonstrate robust performance, achieving prediction accuracies of 93.0% and 95.6%, respectively, while an independently developed deep neural network (DNN) model achieves 91.1% accuracy. A feature importance analysis of XGB reveals the enzyme solution volume's dominant role (43.7%), followed by time (20.7%), substrate concentration (15%), temperature (15%), and pH (5.6%). Further interpretability analysis unveils complex parameter interactions and potential optimization strategies. This data-driven approach, incorporating machine learning, deep learning, and interpretable analysis, offers a viable pathway for polysaccharide yield prediction and the potential recovery of various agricultural residues.

© 2023 The Authors. Published by Elsevier B.V. on behalf of Chinese Society for Environmental Sciences, Harbin Institute of Technology, Chinese Research Academy of Environmental Sciences. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Recently, there has been a notable focus on the efficient utilization of cornstalk [1], owing to its substantial annual production [2] and the potential for significant economic and environmental benefits. Polysaccharide is a highly economically worth product obtained from cornstalk by hydrolysis by xylanase [3,4]. Polysaccharide has many possible uses in the biomedical field, such as stimulating the growth of probiotics [5], reducing the risk of colon cancer [5,6], promoting immunomodulatory [7], and protecting skin [8]. The current research on pretreatment techniques [9,10] has improved polysaccharide production efficiency, while the polysaccharide enzymatic production system still suffers from weak operational stability and high cost. A polysaccharide yield prediction model monitoring polysaccharide production and optimizing parameters combination can be developed by exploring interactions among enzymatic parameters. However, many enzymatic reaction parameters, such as temperature and pH, influence the enzymatic process as a biochemical reaction process [11]. Since biochemical reaction processes are usually nonlinear and complex, linear models' prediction accuracy decreases with increasing input parameters. Although multivariate statistical methods can assist [12–14], capturing complex interactions in high-dimensional data is still tricky.

Fortunately, multiple artificial intelligence techniques, such as machine learning (ML) [15–17], deep learning (DL), and interpretability analysis, have made it possible to overcome this difficulty. In recent years, machine learning has benefited the development of biomanufacturing, waste material utilization [18], materials science

Abbreviations: ESV, Enzyme Solution Volume; SC, Substrate Concentration; TEMP, Temperature; TIME, Time; PY, Polysaccharide Yield; LR, Linear Regression; Tree, Decision Tree; RF, Random Forest; XGB, eXtreme Gradient Boost; DNN, Deep Neural Network; SHAP, Shapley additive explanations; CNN, Convolutional Neural Network; LSTM, Long short-term memory; BPNN, Backpropagation neural network; PCC, Pearson's correlation coefficient; Std, standard deviation; ML, Machine Learning; DL, Deep Learning.

[*] Corresponding author.

[**] Corresponding author.

E-mail addresses: tianys@neau.edu.cn (Y. Tian), wulin.yang@pku.edu.cn (W. Yang).

https://doi.org/10.1016/j.ese.2023.100321

2666-4984/© 2023 The Authors. Published by Elsevier B.V. on behalf of Chinese Society for Environmental Sciences, Harbin Institute of Technology, Chinese Research Academy of Environmental Sciences. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

[19–22], drug discovery [23,24], and pollutant treatment [25–28]. Various machine learning models, including linear regression [29], random forests [30–32], support vector machines [33,34], and extreme gradient boost [35,36], have been used in architectural engineering, genetics, materials chemistry, and other disciplines. For example, XGB [37] was used to predict algal biochar yield and its composition with an R^2 of 0.84 on the testing dataset. While these studies demonstrated the applicability of machine learning methods and provided valuable insights through interpretable analysis, the unsatisfactory predictive accuracy and the small number of models explored may have hindered the discovery of better modeling approaches. Apart from the traditional machine learning models, advanced deep learning methods [38,39], including deep neural networks, convolutional neural networks, and recurrent neural networks, show great power [40–42]. For instance, a long short-term memory (LSTM) neural network [43] was established to predict the concentrations of three representative pollutants in the effluent of a practical large-scale constructed wetland with an R^2 above 0.9. A neural network [22] was designed to predict the stiffness and critically resolved shear stress of CoNiCrFeMn alloys with a relative error of 2.77% and 2.17%. These investigations highlighted the potential for leveraging developers' experience to design the structure of deep learning models and improve prediction accuracy. However, constructing deep learning models is considerably more challenging than machine learning models due to the complex model structure, which is challenging to conduct interpretability analysis on deep learning models (Table S1).

The lack of practical experience in utilizing artificial intelligence (AI) for exploring complex polysaccharide production systems prompted this study to introduce a novel data-driven approach (Fig. 1). By applying various AI techniques, this study aimed to enhance the efficiency of xylanase in the polysaccharide production from cornstalk. Based on this approach, this study explored the prediction accuracy of multiple ML models for polysaccharide yield and performed global and local interpretability analysis by Shapley additive explanations (SHAP) on the accurate model. The findings revealed that ensemble machine learning and deep learning neural network models exhibited better prediction accuracy. Additionally, we investigate the interpretability of the polysaccharide yield prediction model, shedding light on potential enzymatic parameter optimization solutions by analyzing the importance and interaction of enzymatic parameter features.

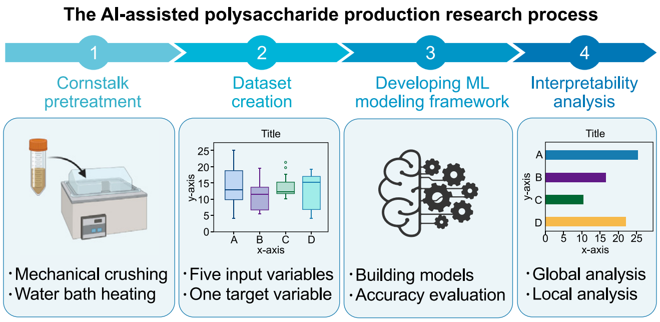

Fig. 1. The workflow diagram in this study. The whole research process began with the cornstalk pretreatment for preparing the suitable cornstalk. Subsequently, a reaction system was constructed for polysaccharide production by digesting the cornstalk with xylanase, obtaining a dataset from the experimental results of this reaction system. The third phase involved data preprocessing, ML model building, and model prediction performance evaluation. To address concerns regarding the black-box nature of most ML models and to mitigate trust risks in real-world applications, the whole process ended with an interpretability analysis of these models using the SHAP library. This diagram was created with BioRender.com.

2. Material and methods

2.1. Cornstalk pretreatment

In preparation for the experiment, the cornstalk underwent a pretreatment process. Initially, it was crushed to a size suitable for passage through a 40-mesh sieve. Then, it was heated in a water bath at 60 °C for 1 h with a 1.5% NaOH solution at the material-to-liquid ratio of 1:10 (w/v). Once the pretreatment was completed, the cornstalk was dried and ready for subsequent experiments.

2.2. Dataset creation

In this research, a dataset was created by gathering data from xylanase hydrolysis experiments conducted on cornstalks. The dataset included several input variables such as temperature (TEMP, °C), pH, time (TIME, min), substrate concentration (SC, g L^{-1}), and enzyme solution volume (ESV, μL). The target variable was the polysaccharide yield (PY, mg) of xylanase enzymatic hydrolysis of cornstalk. A total of 179 data points were collected as the primary dataset. Xylanase hydrolysis polysaccharide yield was measured by 3,5-dinitro salicylic acid (DNS) assays [5], and more detailed configuration methods can be found in the supplementary materials.

2.3. Developing machine learning modeling framework

This section introduces a machine learning modeling framework to construct highly accurate polysaccharide yield prediction models. The strategy involves a gradual increase in the model complexity, transitioning from linear to nonlinear and from low to high integration, thereby achieving enhanced prediction accuracy. We constructed one linear and three nonlinear machine learning models to implement this approach and independently designed a deep learning model using a deep neural network.

2.3.1. Constructing machine learning models

The data preprocessing began with Pearson's correlation coefficient (PCC) analysis. Correlation analysis is commonly used to determine the statistical association between two or more variables and further analyze the association's strength and direction. All input variables in this experiment are continuous, and the linear dependence among variables can be measured using PCC [26] in equation (1):

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\sum_{i=1}^{n}(y_i - \overline{y})}{\sqrt{\Sigma_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\Sigma_{i=1}^{n}(y_i - \overline{y})^2}} \tag{1}$$

where ρ_{xy} refers to the PCC value between two variables. Meanwhile, $\overline{x}$ and $\overline{y}$ are the means of the variable x and the variable y, respectively. ρ_{xy} varies from -1 to 1, where 0 indicates no linear correlation, and a large positive or negative value indicates a strong positive or negative linear correlation [26].

The analysis's second stage involved removing abnormal values, commonly known as outliers. Outliers are data points that significantly deviate from most sample points and may introduce unreasonable characteristics into the dataset. Neglecting these outliers could lead to incorrect conclusions in machine learning modeling scenarios. The most common forms of identifying outliers include graphical methods (e.g., box-line plots, normal distribution plots) and modeling methods (e.g., linear regression, clustering algorithms). In this study, outliers were identified using the box-line plot method. This technique utilizes the data's quantiles to identify points that deviate significantly from the rest and has wide application in academic research. Outliers were identified based on data values that exceeded the upper whisker or less than the lower whisker of the box-line plot, with the upper and lower whiskers set at 1.5 times the quartile deviation. The dataset obtained after removing these outliers was referred to as the sub-dataset.

The third phase involved the application of stratified sampling, which is one of the standard sampling techniques, alongside random sampling, holistic sampling, and systematic sampling. According to the PCC calculation, the input variable ESV is highly correlated with the target variable PY, thus the ideal training and testing sets should include various types of ESV. Considering the sample size of this dataset, the study opted for the stratified sampling method to construct the training and testing sets. Pandas was used to create four ESV types: 0–100 μL for Type 1, 100–200 μL for Type 2, 200–300 μL for Type 3, and 300–400 μL for Type 4. The training set comprised 80% of the sub-dataset samples, while 20% of the data points constituted the testing set which was used for the final evaluation of the developed models.

(1) Linear Regression (LR) Model

The multiple linear regression (LR) model was applied to investigate the relationship among variables. This model utilizes a prediction function based on a linear combination of learned properties [43], as shown in equation (2):

$$f(x) = w_1x_1 + w_2x_2 + \ldots + w_dx_d + b \tag{2}$$

Linear models have many advantages, including their simplified form and ease of modeling. Based on the assumption of linear correlation between target and input variables, the LR model predicts the values of target variables by combining multiple input variables from the study subject into a linear process. Although the LR model is proficient at capturing linear relationships among variables, it lacks the ability to handle nonlinear relationships. Therefore, the regression prediction of the linear model may not be as effective as desired. This study also used the LR model as a control group for nonlinear models.

(2) Decision Tree (Tree) Model

Regarding the four nonlinear models, our initial focus was on constructing the decision tree model, which serves as the base model for the other two tree models (RF and XGB). The decision tree consists of nodes and a directed edge. Each layer corresponds to a sample feature. The nodes include internal and leaf nodes (Fig. S2). Internal nodes represent features or attributes, while leaf nodes correspond to classes. The core principle of the decision tree model is that similar inputs will produce similar outputs with low computational complexity and substantial interpretation advantages. The decision tree model outperforms the linear regression model in interpretation. While constructing this study's decision tree model, we performed hyperparameter tuning using the grid search method.

(3) Random Forest (RF) Model

Although the prediction accuracy of the decision tree model was higher than that of the linear regression model, it might be weaker than that of the ensemble learning model. Ensemble learning methods can combine many decision trees. The algorithms for ensemble learning models include bagging and boosting. The RF model is a typical bagging algorithm that trains a weak learner (base model) by selecting data randomly and with a put-back from the original data (Fig. S3).

(4) eXtreme Gradient Boost (XGB) Model

After completing the construction of the RF model, we constructed the XGB model (Fig. S4), a typical machine-learning model using the boosting algorithm. The core of this algorithm lies in boosting a weak learner to a strong learner. Distinct from the bagging algorithm, there are two primary differences. Firstly, in the bagging algorithm, all weak learners carry equal weight in influencing the result, whereas the boosting algorithm assigns a higher weight to the weak learner, which achieves a more accurate prediction after each training round. Secondly, the boosting algorithm modifies the probability distribution of the training set after each training cycle. This algorithm will increase the weights of samples incorrectly predicted by the weak learner in the previous training cycle while decreasing the weights of correctly predicted samples. Consequently, the process effectively enhances the overall prediction accuracy of the XGB model in subsequent training rounds.

2.3.2. *Constructing deep learning model*

The process began with data standardization, where the objective was to mitigate challenges posed by varying value ranges when feeding data into the neural network. This normalization process entailed subtracting the variable's mean value and dividing it by the standard deviation. The normalization eliminated the influence of magnitude, improved the accuracy of the deep learning model, and allowed faster convergence during gradient descent. In addition, the mean and standard deviation used for standardization calculations of training and testing set data were also calculated from the training set data. The data was partitioned, with 80% randomly selected and labeled as training data, while the remaining 20% labeled as testing data for the final evaluation of the developed model.

The second step involved the development of the deep neural network (DNN) model. To address overfitting, a smaller network structure was adopted (Fig. S5). Nonlinear properties were introduced to the neural network through the activation function, and the Rectified Linear Unit (ReLU) function was chosen for this purpose. ReLU mitigated the issue of gradient disappearance and brought computational simplicity and sparsity to counteract overfitting in the hidden layers. Moreover, L1 regularization was added to hidden layers to limit the complexity of the model. The regularization method added the cost associated with more significant weight values to the model's loss function. In the L1 regularization, the cost added was proportional to the absolute value of the weight coefficients. The final layer of the network was a linear layer with only one neuron and no activation functions. The DNN model was compiled using the optimizer “*adam*".

Subsequently, the method of *k*-fold cross-validation was used to evaluate the DNN model during training. Determining the exact number of epochs required to achieve the lowest validation loss before training the neural network model poses a challenge. This problem can be solved using the Keras library's ModelCheckpoint function, which continuously saves the best model weight coefficients obtained throughout the training process.

2.3.3. *Error metrics for evaluating models*

In general, the loss function of the deep learning model is the mean square error (MSE) [44], and the mean absolute error (MAE) [44] was monitored (equations (3) and (4)). The correlation coefficient (R^2) [43] and root mean square error (RMSE) [43] are used as statistical measures to evaluate the model. Higher values of R^2 and lower values of RMSE indicate greater accuracy of the model, as described in equations (5) and (6).

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

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\overline{y}_i - y_i)^2} \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \quad (6)$$

In equations (3)–(6), $\widehat{y}_i$, y_i, and $\overline{y}_i$ are the predicted, actual, and mean values, respectively.

2.4. Interpretability analysis

Machine learning models are often considered black box models. Despite their high prediction accuracy, understanding the contribution of each input variable to the target variable remains challenging, creating trust risks when deploying these models in real business scenarios. In addition, interpretable analysis helps to enhance insight into the model, aids in model and feature iteration, and helps develop optimized algorithms at a later stage. This study used the SHAP library to perform an interpretability analysis on the developed model to explore the relationship among the variables. Before the widespread adoption of SHAP, researchers used feature importance or partial dependence plot to explain machine learning models. Although feature importance measured the significance of each feature in the dataset, the findings often varied significantly across different models, and interpreting the effect of each feature on individual predictions proved unfeasible. SHAP has three ideal properties, namely local accuracy, missingness, and consistency, interpreting the predicted value of a model as the sum of the attribution values of each input feature [45] (equation (7)). SHAP is the consistent individuation feature imputation method. Suppose consistency does not validate, which indicates that the model does not necessarily rely more on features with high assigned attribution. In that case, comparing the attribution importance between two arbitrary models is impossible. The global and local model analysis through SHAP assists in understanding the black box algorithm, observes the interaction of each input variable, and performs feature importance analysis. Feature importance hierarchical clustering plot was used to perform the global analysis of each feature, while the beeswarm plot and the heatmap were used for the local analysis of the model. These plots provide valuable insights into the prediction process of the model, discover the interactions between the input variables, and provide guidance for model debugging and optimization.

Initially, we conducted SHAP analyses on four distinct machine-learning models: LR, Tree, RF, and XGB. Subsequently, we further delved into a more comprehensive analysis, focusing on the XGB model, which exhibited the highest prediction accuracy among the models under consideration.

$$g(x') = \varnothing_0 + \sum_{j=1}^{M}\varnothing_j x'_j \quad (7)$$

In equation (7), g is the explanatory model, $\varnothing_0$ is the interpretive model constant for each variable, $\varnothing_j$ is the attribution value for each variable and belongs to the real numbers ($\varnothing_j \in \mathrm{R}$), j is any given variable. $x' \in \{0,1\}^M$, M is the number of simplified input variables.

3. Results

3.1. Statistical analysis of model inputs

The descriptive statistical analysis of the input and target variables based on the raw data from the dataset is given in Table S3. All means, and standard deviations of these variables show the distribution pattern of the data collected. Additionally, each variable's minimum and maximum values are presented, offering a clear understanding of the parameter range. To further comprehend the distribution pattern, four quartiles are also provided alongside the minimum and maximum values of the variables.

A color-order plot is Pearson's correlation coefficient (PCC) matric (Fig. 2). Darker colors within the plot indicate stronger linear correlations between variables. By analyzing PCCs between variables, we can determine the input variable that exhibits the highest correlation with the predicted target variable and identify the redundant input variables in the data set. The analysis revealed that all input variables were retained for building the predictive model, as their correlations with one another were not strongly correlated ($-0.5 <$ PCC value < 0.5). In addition, the PCC value of 0.61 for ESV and PY indicated a strong positive correlation (PCC value > 0.6), leading to the selection of ESV as the stratified sampling category.

The box line plot (Fig. S6) presents various statistics, including the median, mean, upper and lower quartiles, and upper and lower bounds for all 179 data points of PY data values. Outliers were identified among these data points if they exceeded 1.5 times the box plot's upper and lower quartile deviation. Subsequently, five outliers were removed based on the box line plot, leaving 174 data points in the sub-dataset utilized by the machine learning models. To ensure sufficient instances of each ESV type were included in the training and testing sets, stratified sampling (Fig. S7) was employed. This sampling method has a more significant potential

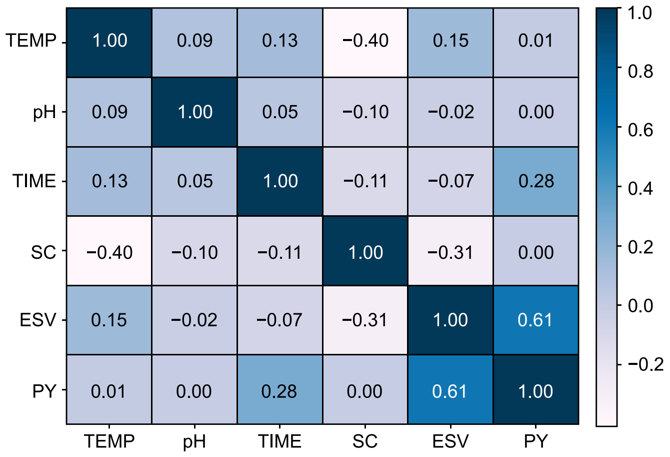

Fig. 2. Pearson's correlation coefficient (PCC) matric. No significant correlations among input variables were observed. A value of 0 indicates no linear correlation, and a high negative or positive value indicates a strong negative or positive linear correlation.

statistical effect and less sample error than simple random sampling.

3.2. Model prediction

3.2.1. Comparison among models with various input variables

The prediction accuracy of the models could be affected due to the absence of any variables. In longitudinal comparison, the models' highest prediction accuracy was consistently achieved when utilizing all five input variables (Table 1). The XGB model had the highest prediction accuracy at 0.956 on the testing set with five input variables, confirming their significance as primary determinants for polysaccharide yield prediction. Moreover, in the horizontal comparison, nonlinear models generally exhibited higher prediction accuracy than the linear model.

3.2.2. Evaluation of models with five input variables

The model prediction plots (Fig. 3) depict the comparison between the predicted and actual data for polysaccharide yield, illustrating the prediction accuracy of these models across different numerical scales. Different types of models have very different prediction results. The R^2 fits for LR, Tree, RF, XGB, and DNN models on the training set were 0.514, 0.979, 0.972, 0.999, and 0.987, respectively. While on the testing data, the R^2 values were 0.514 (LR), 0.879 (Tree), 0.930 (RF), 0.956 (XGB), and 0.911 (DNN). Among these models, XGB outperformed the others, corroborated by its lowest RMSE of only 0.328 (Table S5). Meanwhile, RF, XGB, and DNN had R^2 above 0.9 on the training and testing sets, indicating they were not acute overfitting. The predictive accuracy of LR was poor, primarily due to the predominant nonlinearity among the data in this study. Consequently, the purely linear modeling approach does not yield accurate predictions. Compared to the LR model, the predictive accuracy of the Tree, RF, XGB, and DNN models was greatly enhanced, further confirming the primary nonlinear relationships among the variables.

In this study, the DNN model we designed exhibits better prediction accuracy than the LR and Tree models, though it falls slightly behind the RF and XGB models, suggesting that the deep neural network is suitable for this research scenario. The DNN model's performance did not match that of RF and XGB because of the structured nature of the data. In structured data, some practical features can be artificially constructed (e.g., numerical features of the data in this study), which may be more efficiently processed by machine learning models. With high accuracy, this deep neural network model could be combined with optimization algorithms to find combinations of input variables that increase the target variable (polysaccharide yield) while minimizing research time and production costs. Moreover, as a high-order nonlinear model, the deep neural network model can also be compared with the prediction effects of other nonlinear machine learning models to provide various ideas for finding the optimal polysaccharide prediction model.

The backpropagation neural network (BPNN) model was developed in the previous study [43] for predicting constructed wetland effluent quality, and the DNN model developed in this study is based on the complete neural connection network structure and preprocessing of the dataset. Although both models exhibit strong performance in their respective research scenarios, they differ in terms of the number of hidden layers, the number of neurons, the type of activation function, and the way of data preprocessing. Hence, adapting the deep learning model structure and training approaches to suit specific research scenarios is essential.

For selecting the best polysaccharide yield prediction model and visually comparing the prediction results of different models, we employed the Taylor diagram (Fig. 4). The Taylor diagram displayed these five models' similarities and differences. Tree, RF, DNN, and XGB locations on the Taylor diagram were relatively close, indicating similar prediction performance for these four models. The location of the LR model on the Taylor diagram was far away from the other models, mainly due to the lower R^2 and standard deviation (Std), along with higher RMSE. Therefore, the XGB model has the highest prediction accuracy, closely aligning with the observation point. Meanwhile, the DNN, RF, and Tree models have similar prediction accuracy, while the LR model has the lowest prediction accuracy.

3.3. Influence of input variables on polysaccharide yield

A preliminary SHAP analysis of these models (LR, Tree, RF, and XGB) was conducted to gain insights into their functionality. The feature importance ranking of the input variables (Table 2 and Fig. S8) is almost the same for the four machine learning models, with only the Tree model being slightly different (SC is larger than TIME). The predictive contributions of the individual input variables to the target variable are different in these four models. A more detailed analysis of the XGB model was performed using the SHAP library, as this model has the highest prediction accuracy.

3.3.1. Global analysis

The interpretability analysis on the XGB model (five input variables) started from the SHAP global analysis. The feature importance hierarchical clustering plot (Fig. 5) was used in this section to assess the input variables' contribution to the model prediction results and explore the interaction effects among the input variables. This approach aimed to enhance researchers' better understanding of the predictive mechanism of this model. The results showed that polysaccharide yield was heavily dependent on ESV, but the effect of pH was low. Furthermore, it was realized that ESV, TIME, SC, TEMP, and pH contributed 43.7%, 20.7%, 15%, 15%, and 5.6% to the target prediction, respectively, based on the resolution of the

Table 1
Experimental design and performance of five models for predicting polysaccharide yield.

Parameters	LR		Tree		RF		XGB		DNN	
	Training R^2	Testing R^2	Training R^2	Testing R^2	Training R^2	Testing R^2	Training R^2	Testing R^2	Training R^2	Testing R^2
All	0.514	0.514	0.979	0.879	0.972	0.930	0.999	0.956	0.987	0.911
Except pH	0.514	0.514	0.927	0.846	0.968	0.919	0.997	0.954	0.984	0.888
Except TEMP	0.509	0.502	0.902	0.798	0.915	0.877	0.937	0.792	0.941	0.793
Except SC	0.486	0.451	0.950	0.662	0.934	0.789	0.963	0.760	0.969	0.788
Except TIME	0.453	0.420	0.784	0.723	0.809	0.734	0.832	0.783	0.862	0.731
Except ESV	0.011	−0.026	0.136	0.052	0.542	0.316	0.573	0.104	0.720	0.138
ESV + TIME + SC	0.509	0.504	0.868	0.783	0.884	0.842	0.903	0.819	0.920	0.646
ESV + TEMP + pH	0.436	0.372	0.721	0.378	0.744	0.576	0.763	0.586	0.763	0.559
ESV + SC + TEMP	0.453	0.420	0.783	0.643	0.794	0.717	0.815	0.719	0.838	0.722

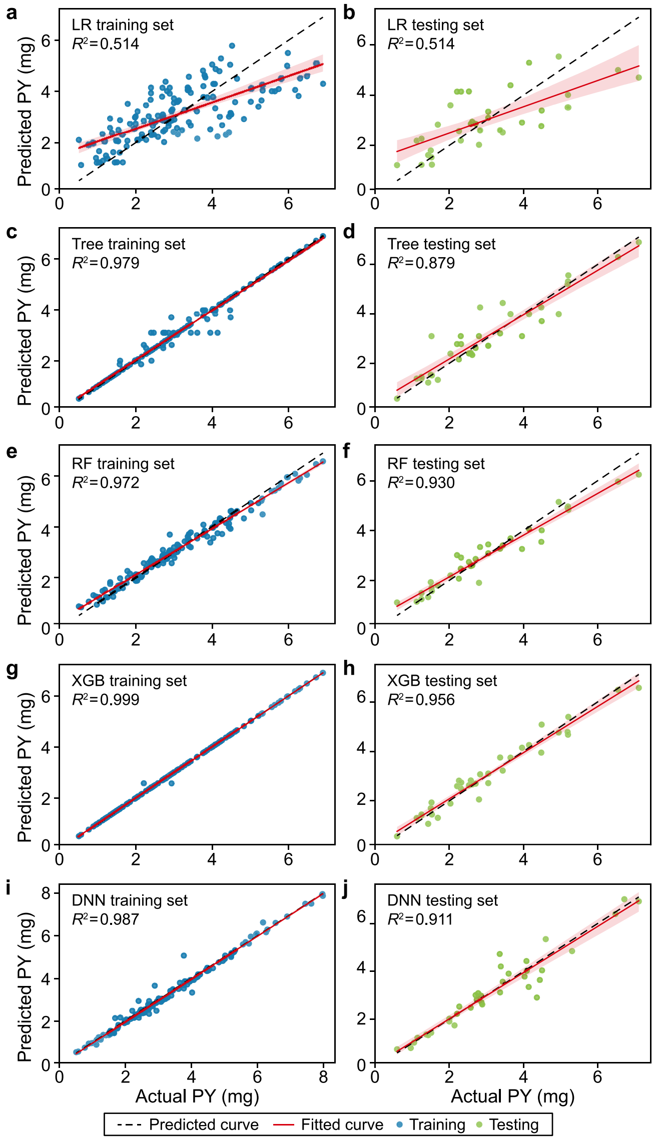

Fig. 3. The machine learning model prediction plots of polysaccharide yield (PY) for: **a**–**b**, training data (**a**) and testing data (**b**) by Linear Regression (LR) model; **c**–**d**, training data (**c**) and testing data (**d**) by Decision Tree (Tree) model; **e**–**f**, training data (**e**) and testing data (**f**) by Random Forest (RF) model; **g**–**h**, training data (**g**) and testing data (**h**) by eXtreme Gradient Boost (XGB) model; **i**–**j**, training data (**i**) and testing data (**j**) by Deep Neural Network (DNN) model. Red shades indicate 95% confidence intervals for the regression lines on the training and testing points. Gray dashed lines represent the line of equality ($y = x$).

black box model. By hierarchical cluster analysis, ESV and TIME were classified as one category of impact factors, while TEMP and pH formed another category. Furthermore, SC formed a large category with TEMP and pH. The input variables in the same category are highly correlated and influenced by variables of other categories, thus reflecting the interaction among input variables.

3.3.2. Local analysis

The SHAP global analysis prioritizes the exploration of feature importance ranking and the interaction effects between input variables while disregarding the impact of nonlinearity and higher-order interactions of individual instances. Consequently, this section provides a more in-depth SHAP local analysis for individual sample predictions using the beeswarm plot (Fig. 6a) and heatmap (Fig. 6b) to provide further clarity on these mechanisms.

The beeswarm plot (Fig. 6a) is a local analysis tool for illustrating the relationship between each variable input value of a single instance and its impact (positive/negative). This plot presents SHAP values on the horizontal axis, while each row corresponds to an input variable. The ESV variable was located at the top of the density scatter plot, which means ESV was the variable with the most significant influence on the reaction system. According to the distribution of the ESV density scatters, the characteristic values of ESV and the corresponding SHAP values show a positive correlation in general. However, the SHAP values of ESV data points with relative typical values (same density scatter color) span a wide range of values. For example, the SHAP values of purple density scatter points take values from about −0.6 to 1, and the SHAP values of red density scatter points take values from about 0 to 2. This indicates that the interaction of other variables strongly influences ESV, implying that the marginal utility of ESV on the response system can be optimized by adjusting other variables. The same pattern also appears for the TIME variable, confirming the conclusion of the feature importance hierarchical clustering plot (Fig. 5) that ESV and TIME are variables of the same category. Increasing the SC characteristic value does not consistently increase its marginal utility, and as evidenced by the scatter plot distribution, tremendous SC values reduce utility. Although extensive SC increase xylanase accessibility to the substrate, they also tend to reach a saturation state, diminishing overall utility. In addition, data points with high TEMP characteristic values (red density scatters) have SHAP values less than 0, showing a negative inhibitory effect on the reaction system. The TEMP data points in the medium temperature state (purple density scatters) illustrate SHAP values greater than 0.5, indicating that the medium temperature facilitates the reaction. Similarly, higher pH values have a negative inhibitory effect, with lower pH values reaching a maximum SHAP value close to 0.4, still well below the maximum SHAP values attainable for the other four variables.

This heatmap (Fig. 6b) revealed the impacts of every input variable on the target variable (polysaccharide yield) for all instances in the entire sub-dataset and the variation of $f(x)$. Many instances of $f(x)$ were below the mean value. The $f(x)$ exhibited a trend generally consistent with the SHAP value of ESV, with a declining and then rising pattern towards the end. This observation suggests that manipulating the remaining four variables when the SHAP value of ESV is less than 0 could potentially lead to $f(x)$ exceeding the sample mean. The analysis combined with the feature importance hierarchical clustering plot (Fig. 5) illustrated that the marginal contribution of the ESV that has the most significant influence on the target variable was only 43.7%. Therefore, a significant space exists to improve the polysaccharide yield by adjusting and optimizing these variables. Moreover, when polysaccharide yield equals, it is possible to explore lower-cost and more efficient combinations of input variables.

4. Discussion

The study proposed a data-driven approach to assist in the recovery of polysaccharide products from cornstalk. This approach facilitates the optimization of lower energy and cost production methods in applications, such as lower temperatures to reduce energy consumption during production and less enzyme solution volume to reduce production costs and carbon emissions. In subsequent studies, machine learning models can be applied to explore more energy-efficient combinations of production conditions that effectively use agricultural waste while protecting the ecological environment. The experimental focus of this paper lies in recovering polysaccharides from hemicellulose of cornstalk, thus

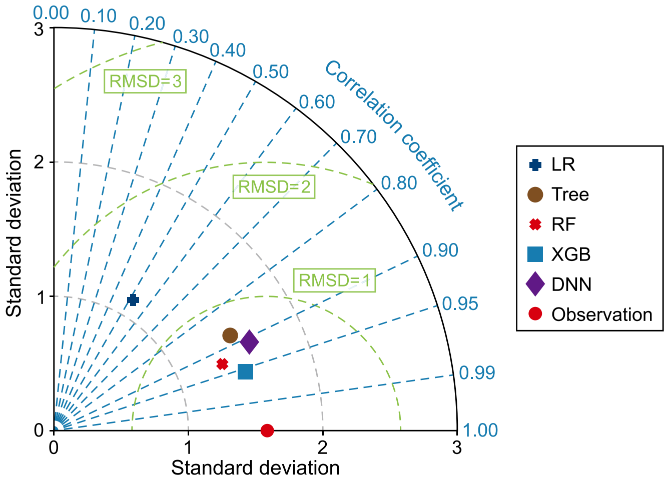

Fig. 4. Taylor diagram comparing the model prediction accuracy. The red point (observation) is the benchmark. The *x* and *y* axes indicate the standard deviation. The quarter-circle arc shows the value of the correlation coefficient. The green arcs indicate the root mean square deviation (RMSD).

Table 2
Interpretability analysis of machine learning models (bar chart).

Input Variable	LR	Tree	RF	XGB
ESV	0.96	0.95	0.94	0.93
TIME	0.34	0.4	0.41	0.44
SC	0.24	0.43	0.37	0.32
TEMP	0.13	0.26	0.2	0.32
pH	0	0.11	0.08	0.12

Note: The values of this table are "mean (|SHAP value|)".

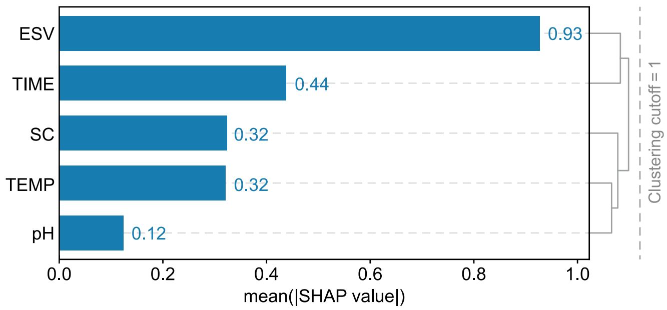

Fig. 5. Feature importance hierarchical clustering plot of all input variables by eXtreme Gradient Boost (XGB) model. The "mean (|SHAP value|)" of the *x*-axis can represent the feature importance of input variables, and the right Y-axis shows the hierarchical clustering results.

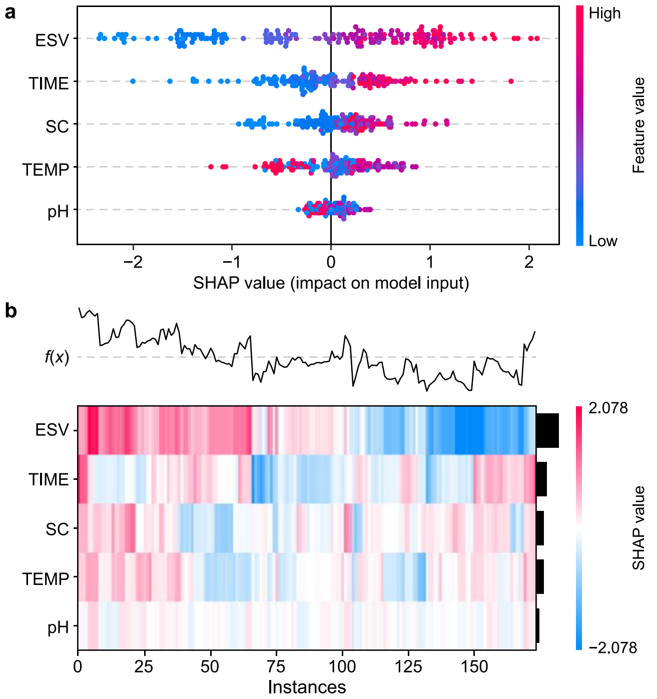

Fig. 6. Local analysis of input variables using Shapley additive explanations (SHAP). **a**, Beeswarm plot. Each dot represents an instance, where the intensity of the red color indicates a higher value of the input variable, while a stronger blue color indicates a lower value. A wider horizontal spread of the instances signifies a more significant impact of this particular input variable. **b**, Heatmap. The function $f(x)$ is the sum of the SHAP values for each instance, representing the level of deviation from the mean; the black histogram on the right of the *y*-axis is the sum of the SHAP values of the variable dimensions. The heatmap represents instances along the horizontal axis and illustrates the influence of each input variable on those instances along the vertical axis. Colors in the heatmap indicate the direction and magnitude of each input variable's impact. The two-dimensional plot of the heatmap showed the SHAP values of every input variable for all instances.

xylanase is chosen as the critical component to construct the data set in the experiment.

The method proposed in this study had an excellent predictive performance for our experimental data, and it was not dependent on our specific experimental approach or the objects chosen (xylanase and cornstalk). This adaptability can be attributed to the similar physical structure of straws from different crops, comprising hemicellulose, cellulose, and lignin, albeit with varying composition percentages. Furthermore, although other species of straw have different structural or compositional ratios compared with cornstalk, such variations do not introduce heterogeneity in data obtained under standard experimental procedures. Various types of xylanases have different preferences for temperature and pH, which mainly affect the biological activity of xylanases. In the feature importance analysis, the predictive importance of temperature and pH on the target products was only 20.6% collectively, implicating that the predictive importance of temperature and pH would increase if the xylanase bioactivity rise but had less impact on the overall experimental conclusion. The individual difference in xylanase did not affect the present method's application. As a result, this modeling approach is cross-scalable and suitable for investigating biomass enzymatic digestion processes with different enzymes and substrates. Additionally, this approach can potentially form an open-source database of related studies.

Deep learning has shown great potential for application in several research fields. Different deep learning models are suitable for different research cases. The LSTM model is suitable for processing time-series data, and the convolutional neural network (CNN) model is suitable for processing image data. Choosing the appropriate deep learning model structure is crucial based on the research problem and the corresponding data type. The DNN model based on deep neural network structure designed in this study achieved accurate prediction for polysaccharide yield. In this study, machine learning has presented great potential for applications in target product yield prediction, feature engineering, production status monitoring, and biomass utilization solution development. Future research opportunities lie in exploring the implementation of machine learning in production anomaly monitoring and solution optimization. To further enhance the practical applications of machine learning, improving the interpretability and applicability analysis of models is crucial. Additionally, the outstanding performance of machine learning in handling large and complex data can offer valuable contributions to addressing the problems and challenges faced in environmental science.

5. Conclusion

This study presented a new data-driven approach to assist in producing polysaccharides from cornstalk. The proposed approach consists of a machine learning modeling framework, followed by an assessment for the prediction accuracy of five models with different input variable combinations. Based on this framework, five polysaccharide yield prediction models were built from linear to nonlinear models to a neural network model. The analysis of model accuracy with different input variables revealed the significance of all five input variables (TEMP, pH, TIME, SC, and ESV) in achieving high prediction accuracy. Omitting any of these variables resulted in decreased prediction performance. Notably, when tested with all five input variables, the RF model achieves a prediction accuracy of 93.0%, while the XGB model attains 95.6%. Especially the independent deep neural network (DNN) model, incorporating five variables, successfully predicted the polysaccharide yield by mining the mapping relationship among the five enzymatic parameters with polysaccharide yield, and its prediction accuracy reached 91.1%.

Subsequently, this approach performed the global and local interpretability analysis using SHAP to explain the predictive mechanism and discover valuable insights related to enzymatic parameters. The model interpretability analysis results showed that the variable ESV had the most significant effect on the enzymatic polysaccharide reaction system, with a marginal utility of 43.7%. In contrast, other variables (TEMP, pH, TIME, and SC) had a combined marginal utility of 56.3% on the reaction system. The interaction analysis of the input variables has presented that ESV was most strongly affected by solid interactions with other input variables, thus providing further evidence that the utility of ESV can be significantly boosted by modulating other variables. The model interpretability analysis provides a reference for adjusting the enzyme reaction system parameters, leading to improved polysaccharide yield and cost reduction.

A new modeling approach was presented in this study for comprehending the multifaceted impact of multiple variables and serving as a guide for improving enzymatic polysaccharide production. The versatility of this approach makes it suitable for the recovery of various agricultural wastes. We expected this technology to open a new energy-saving and efficient utilization pattern for agricultural waste.

CRediT authorship contribution statement

Yushi Tian: Writing - Original Draft, Writing - Review & Editing, Conceptualization, Methodology. **Wulin Yang:** Writing - Review & Editing, Conceptualization, Methodology. **Xu Yang:** Software, Validation, Modeling, Visualization, Writing - Original Draft. **Nianhua Chen:** Data Curation, Writing- Original Draft. **Chunyan Li:** Supervision.

Declaration of competing interest

The authors declare that they have no known competitive interests or personal relationships that might affect the work of this paper.

Acknowledgments

This work was supported by the Academic Core Project of Northeast Agricultural University Scholars Program (20YJ5B01), Heilongjiang Postdoctoral General Fund Project (LBH-Z21110), Key Laboratory of Swine Facilities Engineering, Ministry of Agriculture and Rural Affairs, Northeast Agricultural University 150030, P.R. China.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ese.2023.100321.

References

[1] C. Xu, B. Wu, P. Zhao, Y. Wang, H. Yang, Y. Mi, Y. Zhou, T. Ma, S. Zhang, L. Wu, L. Chen, H. Zang, C. Li, Biological saccharification coupled with anaerobic digestion using corn straw for sustainable methane production, Bioresour. Technol. 367 (2023) 128277, https://doi.org/10.1016/j.biortech.2022.128277.

[2] M.F.S. Khan, M. Akbar, Z. Xu, H. Wang, A review on the role of pretreatment technologies in the hydrolysis of lignocellulosic biomass of corn stover, Biomass Bioenergy 155 (2021) 14, https://doi.org/10.1016/j.biombioe.2021.106276.

[3] Gullon Patricia, Jesus Maria, Martine Gonzalez-Munoz, Van Paula, Henk Gool, Production, refining, structural characterization and fermentability of rice husk xylooligosaccharides, J. Agric. Food Chem. 58 (2010) 3632–3641, https://doi.org/10.1021/jf904508g.

[4] A.K. Samanta, N. Jayapal, A.P. Kolte, S. Senani, M. Sridhar, K.P. Suresh, K.T. Sampath, Enzymatic production of xylooligosaccharides from alkali solubilized xylan of natural grass (Sehima nervosum), Bioresour. Technol. 112 (2012) 199–205, https://doi.org/10.1016/j.biortech.2012.02.036.

[5] J. Bian, F. Peng, X.P. Peng, P. Peng, F. Xu, R.C. Sun, Structural features and antioxidant activity of xylooligosaccharides enzymatically produced from sugarcane bagasse, Bioresour. Technol. 127 (2013) 236–241, https://doi.org/10.1016/j.biortech.2012.09.112.

[6] A.A. Aachary, S.G. Prapulla, Xylooligosaccharides (XOS) as an emerging prebiotic: microbial synthesis, utilization, structural characterization, bioactive properties, and applications, Compr. Rev. Food Sci. Food Saf. 10 (2011) 2–16, https://doi.org/10.1111/j.1541-4337.2010.00135.x.

[7] J.P. Yang, P.H. Summanen, S.M. Henning, M. Hsu, H. Lam, J.J. Huang, C.H. Tseng, S.E. Dowd, S.M. Finegold, D. Heber, Z.P. Li, Xylooligosaccharide supplementation alters gut bacteria in both healthy and prediabetic adults: a pilot study, Front. Physiol. 6 (2015) 11, https://doi.org/10.3389/fphys.2015.00216.

[8] D. Ghosh, A.B. Vir, G. Garnier, A.F. Patti, J. Tanner, Continuous flow production of xylooligosaccharides by enzymatic hydrolysis, Chem. Eng. Sci. 244 (2021) 8, https://doi.org/10.1016/j.ces.2021.116789.

[9] P. Li, C. He, G. Li, P. Ding, M. Lan, Z. Gao, Y. Jiao, Biological pretreatment of corn straw for enhancing degradation efficiency and biogas production, Bioengineered 11 (2020) 251–260, https://doi.org/10.1080/21655979.2020.1733733.

[10] T. Zhang, D. Jiang, H. Zhang, D.J. Lee, Z. Zhang, Q. Zhang, Y. Jing, Y. Zhang, C. Xia, Effects of different pretreatment methods on the structural characteristics, enzymatic saccharification and photo-fermentative bio-hydrogen production performance of corn straw, Bioresour. Technol. 304 (2020) 122999, https://doi.org/10.1016/j.biortech.2020.122999.

[11] O. Akpinar, K. Erdogan, U. Bakir, L. Yilmaz, Comparison of acid and enzymatic hydrolysis of tobacco stalk xylan for preparation of xylooligosaccharides, LWT–Food Sci. Technol. 43 (2010) 119–125, https://doi.org/10.1016/j.lwt.2009.06.025.

[12] X.C. Xie, W. Sun, K.C. Cheung, An advanced PLS approach for Key performance indicator-related prediction and diagnosis in case of outliers, IEEE Trans. Ind. Electron. 63 (2016) 2587–2594, https://doi.org/10.1109/tie.2015.2512221.

[13] S. Yin, S.X. Ding, A. Haghani, H.Y. Hao, P. Zhang, A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process, J. Process Control 22 (2012) 1567–1581, https://doi.org/10.1016/j.jprocont.2012.06.009.

[14] S. Yin, L. Liu, J. Hou, A multivariate statistical combination forecasting method for product quality evaluation, Inf. Sci. 355 (2016) 229–236, https://doi.org/10.1016/j.ins.2016.03.035.

[15] X.Y. Ye, J.E. Zuo, R.H. Li, Y.J. Wang, L.L. Gan, Z.H. Yu, X.Q. Hu, Diagnosis of sewer pipe defects on image recognition of multi-features and support vector machine in a southern Chinese city, Front. Environ. Sci. Eng. 13 (2019) 29–41, https://doi.org/10.1007/s11783-019-1102-y.

[16] T. Ju, M. Lei, G. Guo, J. Xi, Y. Zhang, Y. Xu, Q. Lou, A new prediction method of industrial atmospheric pollutant emission intensity based on pollutant emission standard quantification, Front. Environ. Sci. Eng. 17 (2023) 8, https://doi.org/10.1007/s11783-023-1608-1.

[17] W.J. Lu, W.Z. Huo, H. Gulina, C. Pan, Development of machine learning multicity model for municipal solid waste generation prediction, Front. Environ.

Sci. Eng. 16 (2022) 1–10, https://doi.org/10.1007/s11783-022-1551-6.

[18] J. Li, X.Z. Zhu, Y.A. Li, Y.W. Tong, Y.S. Ok, X.N. Wang, Multi-task prediction and optimization of hydrochar properties from high-moisture municipal solid waste: application of machine learning on waste-to-resource, J. Clean. Prod. 278 (2021) 12, https://doi.org/10.1016/j.jclepro.2020.123928.

[19] K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, Nature 559 (2018) 547–555, https://doi.org/10.1038/s41586-018-0337-2.

[20] V.L. Deringer, M.A. Caro, G. Csanyi, Machine learning interatomic potentials as emerging tools for materials science, Adv. Mater. 31 (2019) 16, https://doi.org/10.1002/adma.201902765.

[21] S.P. Ong, Accelerating materials science with high-throughput computations and machine learning, Comput. Mater. Sci. 161 (2019) 143–150, https://doi.org/10.1016/j.commatsci.2019.01.013.

[22] T. Guo, L.P. Wu, T. Li, Machine learning accelerated, high throughput, multi-objective optimization of multiprincipal element alloys, Small 17 (2021) 8, https://doi.org/10.1002/smll.202102972.

[23] M. Elbadawi, S. Gaisford, A.W. Basit, Advanced machine-learning techniques in drug discovery, Drug Discov. Today 26 (2020) 769–777, https://doi.org/10.1016/j.drudis.2020.12.003.

[24] L. Zhang, J.J. Tan, D. Han, H. Zhu, From machine learning to deep learning: progress in machine intelligence for rational drug discovery, Drug Discov. Today 22 (2017) 1680–1685, https://doi.org/10.1016/j.drudis.2017.08.010.

[25] D. Xiao, F. Fang, J. Zheng, C.C. Pain, I.M. Navon, Machine learning-based rapid response tools for regional air pollution modelling, Atmos. Environ. 199 (2019) 463–473, https://doi.org/10.1016/j.atmosenv.2018.11.051.

[26] X.Z. Yuan, M. Suvarna, S. Low, P.D. Dissanayake, K.B. Lee, J. Li, X.N. Wang, Y.S. Ok, Applied machine learning for prediction of CO_2 adsorption on biomass waste-derived porous carbons, Environ. Sci. Technol. 55 (2021) 11925–11936, https://doi.org/10.1021/acs.est.1c01849.

[27] J. Li, L.J. Pan, M. Suvarna, Y.W. Tong, X.N. Wang, Fuel properties of hydrochar and pyrochar: prediction and exploration with machine learning, Appl. Energy 269 (2020) 10, https://doi.org/10.1016/j.apenergy.2020.115166.

[28] Y.C. Huang, J.Y. Chen, Q.N. Duan, Y.J. Feng, R. Luo, W.J. Wang, F.L. Liu, S.F. Bi, J. Lee, A fast antibiotic detection method for simplified pretreatment through spectra-based machine learning, Front. Environ. Sci. Eng. 16 (2022) 1–12, https://doi.org/10.1007/s11783-021-1472-9.

[29] H. Borna, S. Khalili, A. Zakeri, M. Mard-Soltani, A.R. Akbarzadeh, B. Khalesi, Z. Payandeh, Proposed multi-linear regression model to identify cyclooxygenase-2 selective active pharmaceutical ingredients, J. Pharm. Innov. 17 (2022) 19–25, https://doi.org/10.1007/s12247-020-09482-w.

[30] M.L. Smith, M. Ruffley, A. Espindola, D.C. Tank, J. Sullivan, B.C. Carstens, Demographic model selection using random forests and the site frequency spectrum, Mol. Ecol. 26 (2017) 4562–4573, https://doi.org/10.1111/mec.14223.

[31] C.L. Chen, C.W. Wu, Diagnosing assets impairment by using random forests model, Int. J. Inf. Technol. Decis. Making 11 (2012) 77–102, https://doi.org/10.1142/s0219622012500046.

[32] M.Y. Liu, Y.Z. Huang, J. Hu, J.Y. He, X. Xiao, Algal community structure prediction by machine learning, Env. Sci. Ecotechnol. 14 (2023) 100233, https://doi.org/10.1016/j.ese.2022.100233.

[33] L. Xu, L. Hou, Z.Y. Zhu, Y. Li, J.Q. Liu, T. Lei, X.G. Wu, Mid-term prediction of electrical energy consumption for crude oil pipelines using a hybrid algorithm of support vector machine and genetic algorithm, Energy 222 (2021) 13, https://doi.org/10.1016/j.energy.2021.119955.

[34] W. Jiang, Y.J. Xie, W.X. Li, J.X. Wu, G.C. Long, Prediction of the splitting tensile strength of the bonding interface by combining the support vector machine with the particle swarm optimization algorithm, Eng. Struct. 230 (2021) 10, https://doi.org/10.1016/j.engstruct.2020.111696.

[35] M. Wang, X. Li, M. Lei, L.B. Duan, H.C. Chen, Human health risk identification of petrochemical sites based on extreme gradient boosting, Ecotoxicol. Environ. Saf. 233 (2022) 8, https://doi.org/10.1016/j.ecoenv.2022.113332.

[36] S. Jafari, Z. Shahbazi, Y.C. Byun, S.J. Lee, Lithium-ion battery estimation in online framework using extreme gradient boosting machine learning approach, Mathematics 10 (2022) 17, https://doi.org/10.3390/math10060888.

[37] A. Pathy, S. Meher, P. Balasubramanian, Predicting algal biochar yield using eXtreme Gradient Boosting (XGB) algorithm of machine learning methods, Algal Res. 50 (2020) 102006, https://doi.org/10.1016/j.algal.2020.102006.

[38] J.X. Dong, M.Y. Zhao, Y.S. Liu, Y.S. Su, X.X. Zeng, Deep learning in retrosynthesis planning: datasets, models and tools, Briefings Bioinf. 23 (2022) 15, https://doi.org/10.1093/bib/bbab391.

[39] J. Zhou, O.G. Troyanskaya, Predicting effects of noncoding variants with deep learning-based sequence model, Nat. Methods 12 (2015) 931–934, https://doi.org/10.1038/nmeth.3547.

[40] S. Khan, T. Yairi, A review on the application of deep learning in system health management, Mech. Syst. Signal Process. 107 (2018) 241–265, https://doi.org/10.1016/j.ymssp.2017.11.024.

[41] M.J. Kittlein, M.S. Mora, F.J. Mapelli, A. Austrich, O.E. Gaggiotti, Deep learning and satellite imagery predict genetic diversity and differentiation, Methods Ecol. Evol. 13 (2022) 711–721, https://doi.org/10.1111/2041-210x.13775.

[42] M. Nadif, F. Role, Unsupervised and self-supervised deep learning approaches for biomedical text mining, Brief, Bioinformation 22 (2021) 1592–1602, https://doi.org/10.1093/bib/bbab016.

[43] B.W. Yang, Z.J. Xiao, Q.J. Meng, Y. Yuan, W.Q. Wang, H.Y. Wang, Y.M. Wang, X.C. Feng, Deep learning-based prediction of effluent quality of a constructed wetland, Env. Sci. Ecotechnol. 13 (2023) 100207, https://doi.org/10.1016/j.ese.2022.100207.

[44] R.Z. Xu, J.S. Cao, T. Ye, S.N. Wang, J.Y. Luo, B.J. Ni, F. Fang, Automated machine learning-based prediction of microplastics induced impacts on methane production in anaerobic digestion, Water Res. 223 (2022) 118975, https://doi.org/10.1016/j.watres.2022.118975.

[45] S. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, in: NIPS, Long Beach, CA, USA, 2017, https://doi.org/10.48550/arXiv.1705.07874.