Online Prediction of the Enzymatic Hydrolysis Efficiency of Crop Straw

Xu Fu,[a,1] Fei Yu,[a,1] Huning Zhang,[c] Yunhui Luo,[b,*] and Lihua Zang [a,*]

The extent of removal of lignin and hemicellulose are crucial indicators for evaluating the efficiency of enzymatic hydrolysis of crop straw. Numerous factors influence these two indices. Establishing a quantitative model that correlates these factors with hydrolysis efficiency is essential, as it can guide efficient hydrolysis. In this study, a predictive method for enzymatic hydrolysis efficiency in crop straw was proposed using Grey relational analysis (GRA), Kernel principal component analysis (KPCA), and a least squares support vector machine (LSSVM). The authors collected a dataset from actual production data and developed an efficiency predictive model using GRA for variable selection, KPCA for dimensionality reduction, and LSSVM for model training. This model allows for the direct estimation of the final enzymatic hydrolysis efficiency based on production condition variables, which can include enzyme amount, temperatures, pH, time, agitation, and straw dimensions. Extensive experimental testing validated the effectiveness of the proposed method, resulting in minimal errors, a high degree of fit, and exceptional performance. The methodology described in this study can serve as a foundation for optimising the design of efficient enzymatic hydrolysis production processes for crop straw. Additionally, it offers valuable soft measurements to support efficient control of the enzymatic hydrolysis process.

DOI: 10.15376/biores.19.2.3505-3519

Keywords: Grey relational analysis (GRA); Kernel principal component analysis (KPCA); Least squares support vector machine (LSSVM); Enzymatic hydrolysis of crop straw; Lignin removal; Hemicellulose removal

*Contact information: a: College of Environmental Science and Engineering, Qilu University of Technology (Shandong Academy of Science), Jinan 250353, P. R. China; b: Faculty of Light Industry, Qilu University of Technology (Shandong Academy of Science), Jinan 250353, P. R. China; c: Agricultural Technology Promotion Center in Yongqiao District, Suzhou, 234000, P. R. China; [1] These authors contributed equally to this work; *Corresponding authors: lyh@qlu.edu.cn (Y. Luo); zlh@qlu.edu.cn (L. Zang)*

INTRODUCTION

The pretreatment and utilization of agricultural waste have garnered increasing attention in the evolving landscape of environmental conservation and sustainable development. The academic community widely acknowledges using straw as a common agricultural byproduct (Saravanan *et al*. 2021). When straw is mishandled, it can pose significant environmental threats, such as air pollution, soil acidification, and increased greenhouse gas emissions resulting from open burning (Usmani *et al*. 2021). In contrast, crop straw is rich in lignin, hemicellulose, and other valuable biomass components. These components can be harnessed for bioenergy or bio-based material production upon biodegradation, offering economic benefits. Thus, effective biodegradation and harnessing of crop straw are crucial for advancing sustainable agricultural practices and ensuring environmental protection (Zhao *et al*. 2021).

Various techniques have been employed to degrade crop straw, of which enzymatic hydrolysis is the most prominent. This process uses cellulose and a mixture of enzymes to facilitate the breakdown of polymeric structures into smaller molecular entities, focusing on lignin and hemicellulose in the straw under optimal temperature and pH conditions (Nguyen *et al*. 2020). However, the removal efficiencies of lignin and hemicellulose during the enzymatic hydrolysis of straw are influenced by multiple factors, leading to intricate impacts on the extent of removal. Previous studies have highlighted that pretreatment with potassium ferrate solution augments the efficiency of the enzymatic hydrolysis of corn straw. However, comprehensive studies employing effective mathematical models that quantitatively elucidate the relationship between these determinants and hydrolytic outcomes are scarce. The literature points out that the utilisation of potassium ferrate composite solution as a pretreatment method boosts the enzymatic hydrolysis efficiency of corn straw (Tian *et al*. 2023). Implications of diverse NaOH-ball milling composite pretreatments have been reported (Yang *et al*. 2022). This study examines the effects of various pretreatment methods on the efficacy of straw enzymatic hydrolysis by applying mathematical models to clarify the connection between these influencing factors and enzymatic hydrolysis efficiency (Kumar *et al*. 2022).

Considering the challenges mentioned above, this study proposes a novel approach for predicting the enzymatic hydrolysis efficiency of crop straw by combining grey relational analysis (GRA), kernel principal component analysis (KPCA), and least-squares support vector machine (LSSVM) (Adnana *et al*. 2019). The proposed method utilises GRA to gauge the impact of influencing factors on enzymatic hydrolysis efficiency. Additionally, it identifies pivotal factors, harnesses KPCA for feature extraction, and employs LSSVM to craft a predictive model for the extents of lignin and hemicellulose removal, prioritising efficiency, and precision. This methodology lays the groundwork for refining enzymatic hydrolysis production in crops and provides a sophisticated soft-measurement tool for effectively controlling the enzymatic hydrolysis process (Agrawal *et al.* 2021).

BACKGROUND

The procedure for crop straw enzymatic hydrolysis can typically be divided into three stages: straw pretreatment, three-stage enzymatic hydrolysis, and solid-liquid separation. This is illustrated in Fig. 1. The harvested biomass was thoroughly washed during the straw pretreatment's initial phase, followed by careful drying and pulverisation using a specialised grinder. Subsequently, the crushed straw was transferred to an enzymatic hydrolysis reaction tank, where a specific composite enzyme and water were added and allowed to stand at room temperature for some time. The second stage involves a three-tiered enzymatic hydrolysis process. Three specific combinations of enzymes and additives, such as hydrogen peroxide, were introduced into the reaction tank at specific intervals. The conditions were optimised to boost the enzyme activity and ensure efficient enzymatic hydrolysis (Huang *et al*. 2019). The third stage centres on solid-liquid separation, a pivotal process in which machinery or filtration techniques are employed to distinguish and remove solid residues from the liquid enzymatic hydrolysate. This step ensures the purity of the liquid product and facilitates the subsequent processing and analysis (Zhu *et al*. 2023).

Numerous factors influence the efficiency of enzymatic hydrolysis, including straw length, reaction duration, temperature, pH, and other relevant variables. Identifying the dominant factors and establishing a quantitative link between them and enzymatic hydrolysis efficiency is essential, necessitating robust methodologies (Guo *et al*. 2023).

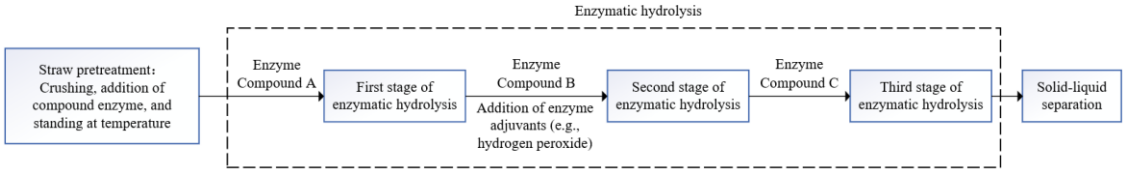

Fig. 1. The enzymatic hydrolysis process of crop straw

METHOD

This study introduces a novel approach, GRA-KPCA-LSSVM, which is designed to predict the extents of removal of lignin and hemicellulose during the enzymatic hydrolysis of crop straw, as illustrated in Fig. 2. During the offline training stage, the LSSVM model was refined using both the GRA variable screening and KPCA dimension reduction techniques applied to the training set (Xiong *et al*. 2018). In the online prediction stage, the test data are selected based on the screening results and averaged before their dimensions are reduced (Adnan Ikram *et al*. 2022). The well-trained LSSVM model offers precise lignin and hemicellulose removal predictions. As depicted in the figure, $X_1, X_2, \cdots, X_m$ denote the initial variables; $X_1^{'}, X_2^{'}, \cdots, X_n^{'}$ signify the selected variables; $P_1, P_2, \cdots, P_z$ correspond to the principal component variables after dimensionality reduction, while ψ_1 and ψ_2 represent the two enzymatic hydrolysis efficiency indicators of the extents of lignin removal and hemicellulose removal, respectively (Liu *et al*. 2022).

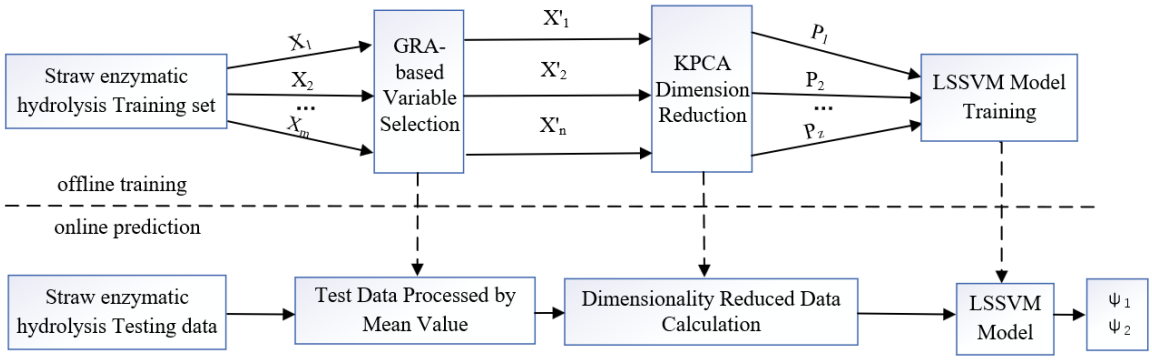

Fig. 2. Prediction of straw enzymatic hydrolysis efficiency based on GRA-KPCA-LSSVM

GRA-based Variable Selection

The GRA method enables quantitative assessment of the interrelationships among different factors within a given system. The fundamental concept behind GRA is to assess the strength of a relationship by evaluating the congruity between the geometric configurations of the reference data column and multiple comparison data columns. Consequently, the higher the degree of similarity, the stronger the correlation (Han *et al*.

2022). Considering the analysis of the lignin removal as an example, the specific steps of variable selection using GRA were as follows:
(1) Input and output sequences were defined. There are m input sequences available, where i represents $\{X_i(k)\}, i = 1,2,\cdots,m$, and m is the number of variables of influencing factors, $k = 1, 2, \cdots, L$, and L is the length of the sequence. The output sequence, denoted as $\{\psi_1(k)\}$, corresponds to the extent of lignin removal. The data sequences were subjected to dimensionless processing by averaging. In this process, each sequence is divided by its respective mean values. For clarity, $\{X_i(k)\}$ and $\{\psi_1(k)\}$ continue to represent the input and output sequences after averaging, respectively (Du 2022).

(2) The correlation coefficient was calculated. The grey correlation coefficient of the i input sequence $\{X_i(k)\}$ and the output sequence $\{\psi_1(k)\}$ at k can be calculated by the following Eq. 1:

$$\zeta_{io}(k) = \frac{\min\limits_i \min\limits_k |\psi_1(k)-X_i(k)| + \max\limits_i \max\limits_k |\psi_1(k)-X_i(k)|}{|\psi_1(k)-X_i(k)| + \rho \cdot \max\limits_i \max\limits_k |\psi_1(k)-X_i(k)|} \quad (1)$$

The resolution coefficient, denoted as $\rho(\rho > 0)$, determines the resolution of the system, with smaller values indicating higher resolution. The value range of ρ is typically confined to (0,1). $\min\limits_i \min\limits_k |\psi_1(k) - X_i(k)|$ represents the minimum discrepancy between the two poles, while $\max\limits_i \max\limits_k |\psi_1(k) - X_i(k)|$ denotes the maximum disparity between the two poles (Antos *et al*. 2022).

(3) The grey correlation coefficient was computed. The numerical value of the correlation degree between the i input sequence $\{X_i(k)\}$ and the output sequence $\{\psi_1(k)\}$ can be expressed as follows:

$$r_i = \frac{1}{L}\sum_{k=1}^{L} \xi_{io}(k) \quad (2)$$

For each of the m input sequences, the grey correlation degrees $r_1, r_2, \ldots, r_m$ were determined.

(4) Reordering was based on the degree of GRA and variable screening. Based on the calculated grey correlation degree, the influencing factor variables $X_1, X_2, \cdots, X_m$ were ranked in descending order. The dominant factor variables with a grey correlation degree greater than a certain threshold η were retained and denoted as $X_1', X_2', \cdots, X_n'$, where $n \leq m$. η ranges from 0.7 to 0.8.

Input Dimension Reduction Based on KPCA

KPCA uses kernel functions to map the original data into a high-dimensional feature space and then performs principal component analysis (Anowar *et al*. 2021). Using the lignin removal analysis as an example, each sample point $\boldsymbol{x}_i$ is an n-dimensional column vector made of $X_1', X_2', \cdots, X_n'$. These N input samples form the input matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N]$. Using a nonlinear mapping function Φ, one can project $\boldsymbol{X}$ into a high-dimensional feature space F, then the transformed representation $\Phi(\boldsymbol{X}) = [\Phi(\boldsymbol{x}_1), \Phi(\boldsymbol{x}_2), \cdots, \Phi(\boldsymbol{x}_N)]$ can be obtained (Kuang *et al*. 2014). If it is assumed that $\boldsymbol{X}$ meets the centralisation requirement in the feature space, meaning $\sum_{i=1}^{N} \Phi(\boldsymbol{x}_i) = \boldsymbol{0}$, then the covariance matrix $\mathbf{C}^{\mathrm{F}}$ in F can be expressed as Eq. 3.:

$$\boldsymbol{C}^{\mathrm{F}} = \frac{1}{N}\Phi(\boldsymbol{X})\Phi(\boldsymbol{X})^{\mathrm{T}} = \frac{1}{N}\sum_{\mathrm{i}=1}^{N}\Phi\,(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_i)^{\mathrm{T}} \quad (3)$$

Matrix $\mathbf{C}^{\mathrm{F}}$ is a square $n \times n$ matrix, and an eigenvector analysis was conducted (Kuang *et al*. 2012). Letting $\lambda_k, \boldsymbol{V}_k$ represent the k-th eigenvalue and the corresponding eigenvector of $\mathbf{C}^{\mathrm{F}}$(where $k = 1,2,\cdots,n$), one obtains:

$$\lambda_k \boldsymbol{V}_k = \mathbf{C}^{\mathrm{F}}\boldsymbol{V}_k \quad (4)$$

Substituting Eq. 3 into Eq. 4 and simplifying, one obtains Eq. 5:

$$\boldsymbol{V}_k = \sum_{i=1}^{N}\Phi(\boldsymbol{x}_i)\frac{\Phi(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{V}_k}{N\lambda_k} \quad (5)$$

The above equation can be further written as:

$$\boldsymbol{V}_k = \sum_{i=1}^{N}\beta_{ki}\,\Phi(\boldsymbol{x}_i) = \Phi(\boldsymbol{X})\boldsymbol{\beta}_k \quad (6)$$

The column vector $\boldsymbol{\beta}_k = [\beta_{k1}, \beta_{k2}, \cdots, \beta_{kN}]^{\mathrm{T}}$ is substituted into Eq. 4 and multiplied by the left with $\Phi(\boldsymbol{X})^{\mathrm{T}}$ to obtain:

$$\lambda_k\Phi(\boldsymbol{X})^{\mathrm{T}}\Phi(\boldsymbol{X})\boldsymbol{\beta}_k = \frac{1}{N}\Phi(\boldsymbol{X})^{\mathrm{T}}\Phi(\boldsymbol{X})\Phi(\boldsymbol{X})^{\mathrm{T}}\Phi(\boldsymbol{X})\boldsymbol{\beta}_k \quad (7)$$

The $N \times N$ dimensional kernel matrix $\boldsymbol{K}$ was introduced, and the value of the i row j column i and row j was computed using the following kernel function (Anowar and Sadaoui 2021),

$$\boldsymbol{K}_{i,j} = \Phi(\boldsymbol{x}_i)^{\mathrm{T}}\Phi\big(\boldsymbol{x}_j\big) = \boldsymbol{\kappa}\big(\boldsymbol{x}_i, \boldsymbol{x}_j\big) \quad (8)$$

where $\kappa(\cdot,\cdot)$ is the kernel function. The RBF kernel function $\kappa\big(\boldsymbol{x}_i, \boldsymbol{x}_j\big) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right)$ can be selected, σ is the kernel width, and the vector norm $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ is the Euclidean distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Substituting $\boldsymbol{K}$ into Eq. 7, one obtains:

$$\lambda_k N\boldsymbol{K}\boldsymbol{\beta}_k = \boldsymbol{K}^2\boldsymbol{\beta}_k \quad (9)$$

This can be simplified to get:

$$\lambda_k N\boldsymbol{\beta}_k = \boldsymbol{K}\boldsymbol{\beta}_k \quad (10)$$

The kernel matrix $\boldsymbol{K}$ in the above equation can be computed using the input sample data, as per Eq. 8. Through solving the eigenvalues and eigenvectors of the kernel matrix $\boldsymbol{K}$, λ_k, $\boldsymbol{\beta}_k$, $k = 1,2,\cdots,n$ can be obtained (Qin 2012).

The authors sorted the eigenvalues in descending order to reduce the data dimensionality and adjusted the corresponding feature vectors accordingly. Next, the kernel principal component was selected based on the cumulative contribution associated with the eigenvalue (the ratio of the variance of the principal component to the total variance of the investigated variables). The cumulative contribution value was determined by adding an eigenvalue's contribution to the preceding eigenvalue's cumulative value. In this study, the authors selected feature vectors corresponding to eigenvalues with a cumulative contribution above 95%, and z principal components were determined to achieve data dimensionality reduction.

For any input vector $\boldsymbol{x}$, one can determine the z principal components as:

$$p_l(\boldsymbol{x}) = \sum_{i=1}^{N}\beta_{ki}\,\kappa(\boldsymbol{x}, \boldsymbol{x}_i) \quad (11)$$

where $l = 1,2,\cdots,z$.

Furthermore, if $\boldsymbol{X}$ does not satisfy the centralisation requirement in feature space F, it is sufficient to substitute $\boldsymbol{K}$ in Eq. 12 with $\boldsymbol{K}'$ computed using the following equation:

$$\boldsymbol{K}' = \boldsymbol{K} - \boldsymbol{I}_N\boldsymbol{K} - \boldsymbol{K}\boldsymbol{I}_N + \boldsymbol{I}_N\boldsymbol{K}\boldsymbol{I}_N \quad (12)$$

where $\boldsymbol{I}_N$ is an $N \times N$ dimensional matrix, and each element is $\frac{1}{N}$.

Training the LSSVM Model

Support vector machine (SVM) is an effective approach that excels at handling small samples and problems that are linearly separable. Building on the SVM, the LSSVM method was designed to address nonlinear problems, offering the advantage of reduced computational complexity. Compared to the SVM, the LSSVM employs distinct optimisation objectives, incorporates equality constraints, and substitutes the original loss function with the sum of squared errors (Tian 2020). To illustrate this, the authors analysed the extent of lignin removal. Here, the z principal components, obtained after GRA variable screening and KPCA dimensionality reduction, serve as the input $\boldsymbol{x}_i = [p_1, p_2, \cdots, p_z]$. The associated output is y_i, represented by ψ_1. This pair, $\{\boldsymbol{x}_i, y_i\}, i = 1,2,\cdots,N$, forms the training set for LSSVM modelling, leading to the construction of the following optimisation problem:

$$\begin{cases} \underset{\boldsymbol{\omega},\xi,b}{\operatorname{argmin}} R(\boldsymbol{\omega},\boldsymbol{\xi}) = \frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega} + \frac{1}{2}c\sum_{i=1}^{N}\xi_i \\ s.t.\ y_i = \boldsymbol{\omega}^{\mathrm{T}}\varphi(\boldsymbol{x}_i) + b + \xi_i, i = 1,2,\cdots,N \end{cases} \quad (13)$$

where $R(\boldsymbol{\omega},\boldsymbol{\xi})$ represents the loss function, $\boldsymbol{\omega}$ denotes the weight parameter, $\boldsymbol{\xi} = [\xi_i], i = 1,2,\ldots,N$, signifies the error variable with ξ_i as its component, b indicates the deviation term, and c >0 serves as the penalty coefficient. To solve the optimisation problem, a Lagrangian function is constructed (Chen and Zhou 2018):

$$L(\boldsymbol{\omega}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega} + \frac{1}{2}c\sum_{i=1}^{N}\xi_i^2 - \sum_{i=1}^{N}\{\alpha_i[\boldsymbol{\omega}^{\mathrm{T}}\varphi(\boldsymbol{x}_i) + b + \xi_i - y_i]\} \quad (14)$$

where the Lagrangian multiplier $\alpha_i > 0$ ($i = 1, 2, \ldots, \mathrm{N}$). The partial derivatives of the Lagrangian function $L(\boldsymbol{\omega}, b, \boldsymbol{\xi}, \boldsymbol{\alpha})$ for $\boldsymbol{\omega}$, b, $\boldsymbol{\xi}$, $\boldsymbol{\alpha}$ are as follows:

$$\begin{cases} \frac{\partial L}{\partial \boldsymbol{\omega}} = \boldsymbol{0} \Rightarrow \boldsymbol{\omega} = \sum_{i=1}^{N}\alpha_i\,\varphi(\boldsymbol{x}_i) \\ \frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N}\alpha_i = 0 \\ \frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \alpha_i = c\xi_i, i = 1,2,\ldots,N \\ \frac{\partial L}{\partial \alpha_i} = 0 \Rightarrow \boldsymbol{\omega}^{\mathrm{T}}\varphi(\boldsymbol{x}_i) + b + \xi_i - y_i = 0, i = 1,2,\ldots,N \end{cases} \quad (15)$$

According to Eq. 15, the following system of linear equations can be derived by eliminating $\boldsymbol{\omega}$ and ξ_i:

$$\begin{bmatrix} 0 & \boldsymbol{1}_N^{\mathrm{T}} \\ \boldsymbol{1}_N & \boldsymbol{\theta} + \frac{1}{c}\boldsymbol{I}_N \end{bmatrix}\begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{y} \end{bmatrix} \quad (16)$$

Here, $\boldsymbol{y} = [y_1, \cdots, y_N]^{\mathrm{T}}$, $\boldsymbol{1}_N = [1, \cdots, 1]^{\mathrm{T}}$, $\boldsymbol{I}_N$ is the identity matrix, $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_N]^{\mathrm{T}}$, $\boldsymbol{\theta}_{i,j} = \varphi(\boldsymbol{x}_i)^{\mathrm{T}}\varphi(\boldsymbol{x}_j), i, j = 1, \cdots, N$. $\boldsymbol{\theta}_{i,j}$ can be calculated by kernel function:

$$\boldsymbol{\theta}_{i,j} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi(\boldsymbol{x}_i)^{\mathrm{T}}\varphi(\boldsymbol{x}_j) \quad (17)$$

For optimal LSSVM performance, the choice of a suitable kernel function, $\kappa(\cdot,\cdot)$, is pivotal. Common kernel functions include the polynomial, RBF (radial basis function), and linear kernels. Given its widespread use in tackling nonlinear problems, the RBF was deemed suitable for this study (Wang and Hu 2015). Therefore, the RBF is selected as follows:

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right) \quad (18)$$

Here, σ is the kernel width, and the vector norm $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ is the Euclidean distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

In summary, using the training data, it is possible to estimate the parameters b and $\boldsymbol{\alpha}$, enabling one to derive the LSSVM regression function model. Consequently, accurate predictions for the new test samples $\boldsymbol{x}$ can be obtained. The result is as follows,

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i \, \kappa(\boldsymbol{x}, \boldsymbol{x}_i) + b \quad (19)$$

where the kernel function $\kappa(\boldsymbol{x}, \boldsymbol{x}_i)$ is calculated according to Eq. 18, and $\boldsymbol{x}_i$ is the training sample vector (Yuan *et al*. 2015).

Prediction of Enzymatic Hydrolysis Efficiency

In the online prediction phase, as shown in Fig. 2, data from *m* influencing variables in the actual project is first collected. The authors selected and averaged the test data based on the screening outcomes from GRA-based Variable Selection. Drawing from the insights in Input Dimension Reduction Based on KPCA and using Eq. 11, these data were employed for dimensionality reduction. Finally, using the trained LSSVM model, the enzymatic hydrolysis efficiency was predicted, particularly the predicted outcomes of the lignin and hemicellulose removal values, as determined by Eq. 19.

TEST VERIFICATION

Process and Data

This study used data from the crop straw enzymatic hydrolysis production process of Zhongnong Jiemei, Ltd. Co., Suzhou (Anhui, China). Through rigorous research and meticulous analysis, 15 influencing factors were selected: the length of crushed straw (X_1), the first stage reaction temperature (X_2), the first stage reaction pH (X_3), the first stage reaction time (X_4), the second stage reaction temperature (X_5), the second stage reaction pH (X_6), the second stage reaction time (X_7), the third stage reaction temperature (X_8), the third stage reaction pH (X_9), the third stage reaction time (X_{10}), the amount of crop straw added (X_{11}), the amount of water added (X_{12}), the volume of the enzymatic hydrolysis tank (X_{13}), the speed of the enzymatic hydrolysis tank (X_{14}), and room temperature (X_{15}).

Additionally, the corresponding extents of lignin removal and hemicellulose removal, *i.e.*, ψ_1 and ψ_2, were recorded. The collected data were organised into training

and testing datasets. The dataset comprised of 200 sample groups. Of these, 160 groups were designated for training, and the remaining 40 groups served as test sets for evaluation.

Evaluating Indicators

To evaluate the efficacy of the model, two evaluation metrics were employed, namely the root mean square error (RMSE) and the coefficient of determination (R^2):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y^{obs}-y^{pred}\right)^2}{n}} \qquad (20)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y^{obs}-y^{pred}\right)^2}{\sum_{i=1}^{n}\left(y^{obs}-\tilde{y}^{obs}\right)^2}$$

The formula uses the variables y^{obs} and y^{pred} , to represent the observed and predicted values, respectively. Additionally, $\tilde{y}^{obs}$ represents the average of all observed values, and n denotes the total number of samples. The RMSE quantifies the difference between the predicted and actual values. The RMSE value typically ranges between 0 and 1, where values closer to 0 indicate high accuracy. In contrast, R^2 measures how well the predicted values fit the actual values. A higher R^2 value suggests a better fit of the model (Wang *et al*. 2022).

RESULTS AND ANALYSIS

Analysis of the Effect

During the training stage, the variables were screened, a dimensionless averaging of the samples was performed, and a grey correlation analysis conducted. For the lignin removal analysis, the resolution coefficient ρ was set to 0.4, and then computed the correlation degree was computed. This resulted in the following correlation degrees: r_1 = 0.642, r_2 = 0.778, r_3 = 0.699, r_4 = 0.747, r_5 = 0.770, r_6 = 0.717, r_7 = 0.751, r_8 = 0.731, r_9 = 0.758, r_{10} = 0.764, r_{11} = 0.672, r_{12} = 0.592, r_{13} = 0.545, and r_{14} = 0.592, and r_{15} = 0.593. Based on the degree of correlation, the influencing factors were arranged in descending order:

$$X_2 > X_5 > X_{10} > X_9 > X_7 > X_4 > X_8 > X_6 > X_3 > X_{11} > X_1 > X_{15} > X_{14} > X_{12} > X_{13}.$$

Using a threshold value of 0.65, twelve influencing factors were selected: X_2, X_5, X_{10}, X_9, X_7, X_4, X_8, X_6, X_3, X_{11}, X_1, and X_{15}.

The data for the influencing factors were first screened using GRA and then subjected to KPCA dimensionality reduction. For instance, Table 1 lists the 12 eigenvalues (in descending order) and their associated and cumulative contribution values, considering the extent of lignin removal. Table 1 shows that, in the training phase, there were four principal components with a cumulative contribution exceeding 95%.

Using the method detailed in Input Dimension Reduction Based on KPCA, the training data were reduced to four principal components and they were used for the LSSVM modelling. The penalty coefficient c and kernel function width σ of the LSSVM model were chosen as 30 and 0.01, respectively. Upon completing the modelling, the authors employed the method described in Prediction of Enzymatic Hydrolysis Efficiency to test and validate the data in the test set.

Table 1. Kernel Principal Component Analysis Results

Serial Number	Eigenvalue	Contributions (%)	Cumulative Contribution (%)
1	8.712	37.301	37.301
2	3.421	27.049	64.351
3	1.413	22.278	86.629
4	1.043	12.877	99.507
5	0.673	0.286	99.793
6	0.452	0.101	99.894
7	0.032	0.033	99.932
8	0.014	0.025	99.957
9	0.004	0.017	99.974
10	0.003	0.015	99.990
11	0.002	0.005	99.996
12	0.003	0.004	100

Figures 3 and 4 depict the results of the model training, prediction, and error analysis for the final removal values of lignin and hemicellulose. The figures show the training and prediction errors, representing the differences between the measured and predicted or actual training values.

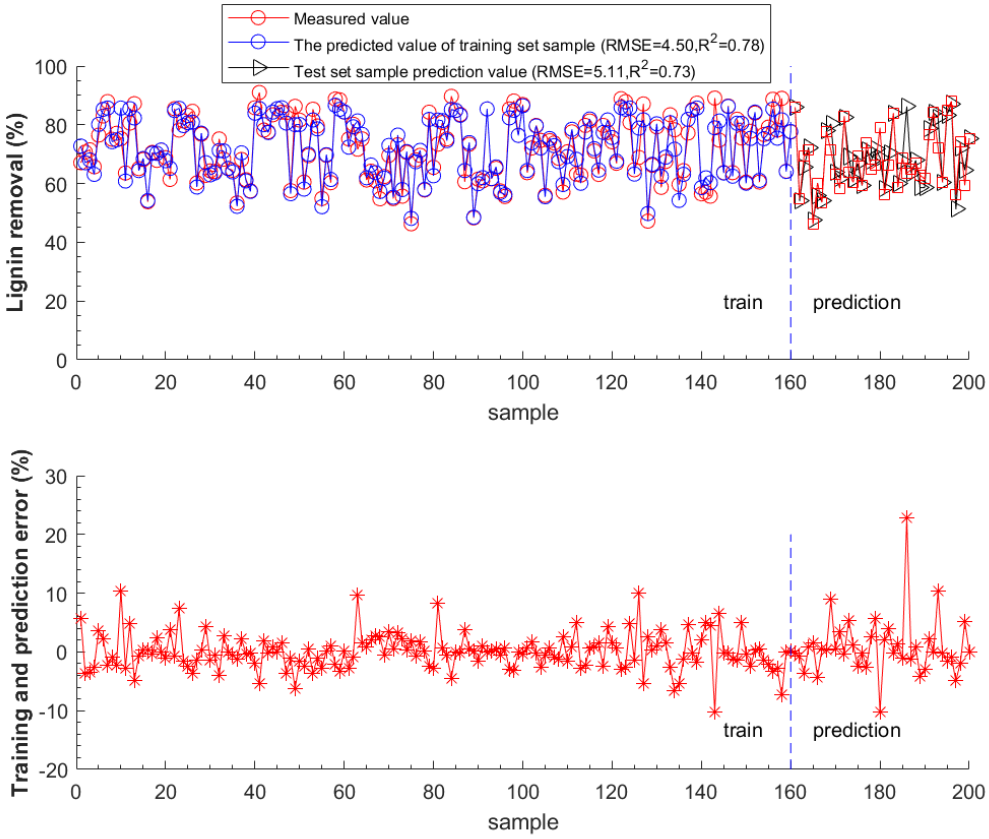

Fig. 3. Lignin removal modelling and prediction results

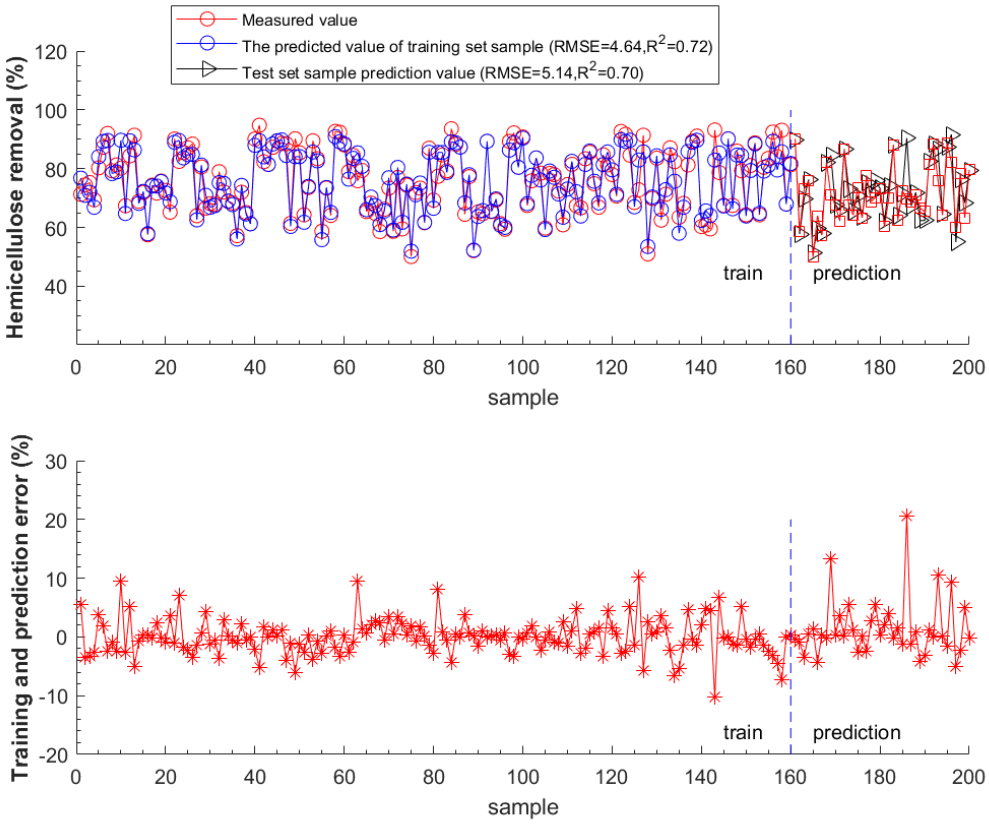

Fig. 4. Hemicellulose removal modelling and prediction results

From Figs. 3 and 4, it was inferred that the method described in this study accurately modeled and predicted the lignin removal. The RMSE of training was 4.50, the fitting degree was 0.78, the RMSE of testing was 5.11, and the fitting degree was 0.73; for the modelling and prediction of hemicellulose removal, the RMSE during training was 4.64, the fitting degree was 0.72, the RMSE during testing was 5.14, and the fitting degree was 0.70. The proposed method demonstrated robust modelling and prediction capabilities for lignin and hemicellulose removal. Additionally, the random selection of actual industrial data for this study led to fewer edge data points in the high-dimensional space of the dataset, enhancing the predictive outcomes at the inference stage.

Comparative Analysis

The authors verified the efficacy of their proposed method through a comparative study that examined various resolution coefficients (ρ), penalty coefficients (c), and kernel function widths (σ). The effectiveness of the GRA variable-screening module in the proposed method (GRA-KPCA-LSSVM) was verified and compared with that of the non-GRA module (KPCA-LSSVM). The results are illustrated in Figs. 5 and 6 in Table 2.

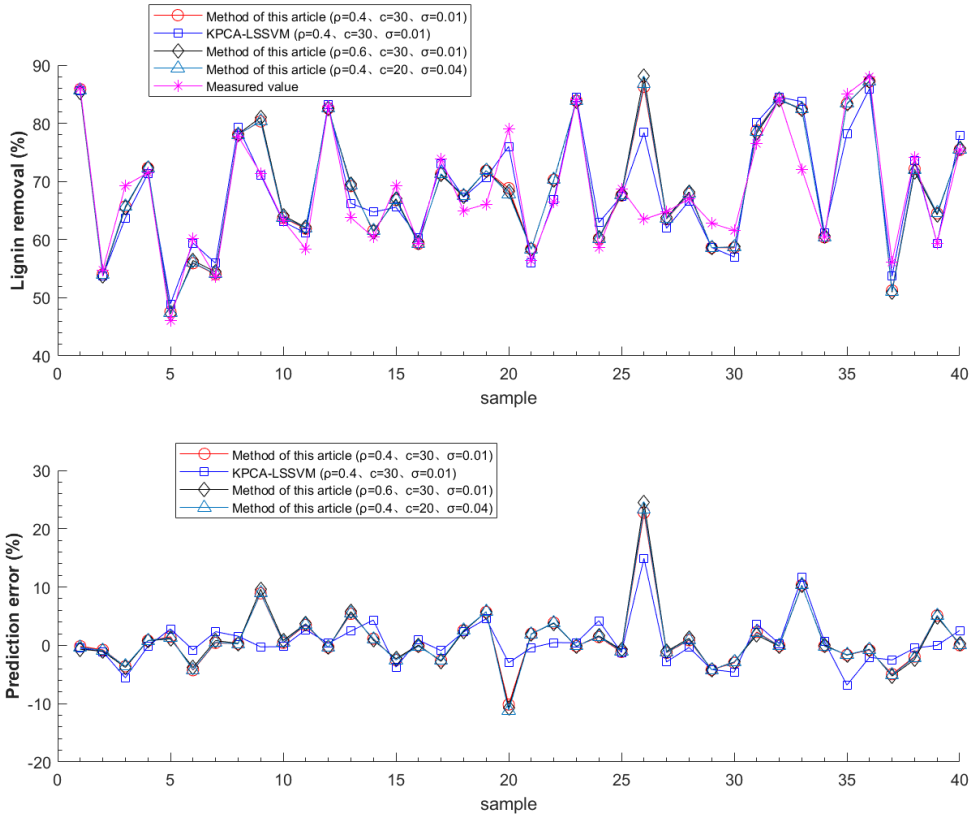

Fig. 5. Prediction results and errors of lignin removal

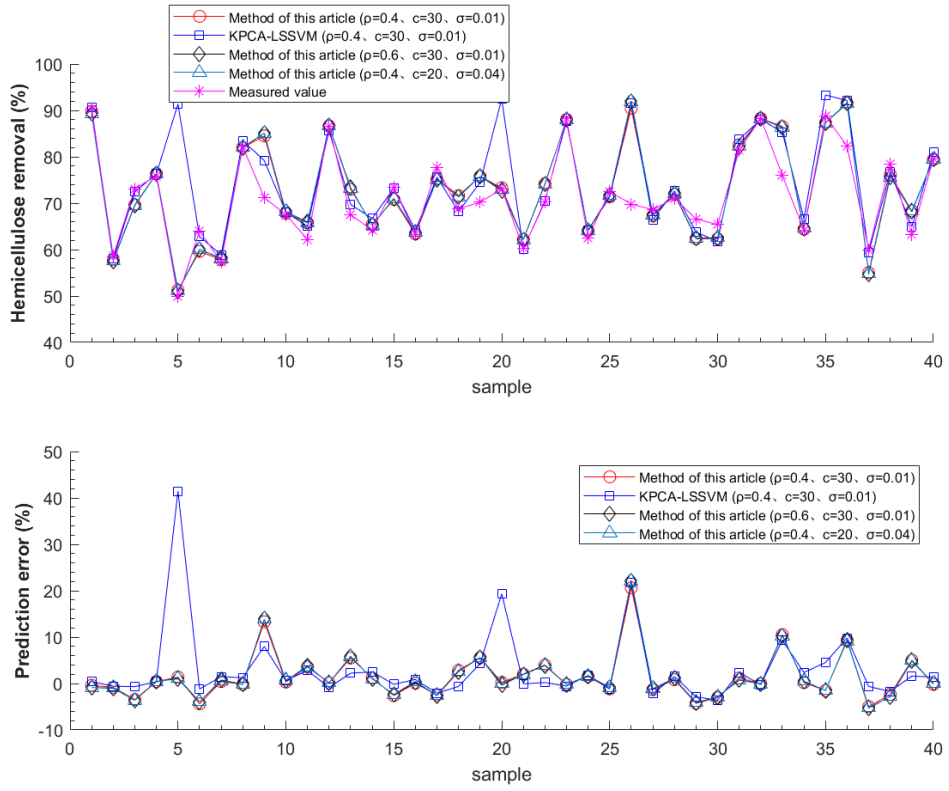

Fig. 6. Prediction results and errors of hemicellulose removal

Table 2. Analysis of Removal Error Results

	Method of this article (ρ = 0.4, c = 30, σ = 0.01)	KPCA-LSSVM (ρ = 0.4, c = 30, σ = 0.01)	Method of this article (ρ = 0.4, c = 20, σ = 0.04)	Method of this article (ρ = 0.6, c = 30, σ = 0.01)
RMSE				
ψ_1 ψ_2	5.11 5.14	6.14 6.29	5.24 5.26	5.39 5.30
R^2				
ψ_1 ψ_2	0.73 0.70	0.67 0.64	0.72 0.69	0.72 0.69

Table 2 shows that the KPCA-LSSVM method provided accurate RMSE predictions for lignin and hemicellulose removal values, measured at 6.14 and 6.29, respectively. High R^2 values of 0.67 and 0.64 were obtained for lignin and hemicellulose removal, respectively. The proposed method (ρ = 0.4, c = 30, σ = 0.01) yielded RMSE values of 5.11 and 5.14 for the predicted lignin and hemicellulose removal, respectively, with corresponding R^2 values of 0.73 and 0.7. The results obtained by this method demonstrate a high level of prediction accuracy, minimal error, and a substantial degree of model fitting. Using GRA for variable selection considerably improved performance in modelling and prediction.

Simultaneously, it is evident that adjusting the ρ value to 0.6 (ρ = 0.6, c = 30, σ = 0.01) within this approach led to an increase in the corresponding RMSE values for predicting lignin and hemicellulose removal, with values of 5.39 and 5.30, respectively. Concurrently, R^2 shows a minor decline to 0.72 and 0.69. Through changing the parameters c and σ (ρ = 0.4, c = 20, σ = 0.04), the corresponding RMSE increased to 5.24 and 5.26, and R^2 decreased to 0.72 and 0.69. The observation mentioned above suggests that variations in ρ, c, and σ will exert a substantial impact on both the accuracy of predictions and the degree of model fitting. The appropriate selection of ρ, c, and σ values can effectively minimise prediction errors and enhance the degree of model fitting. This study used the grid search method to determine the optimal parameters. In addition, advanced optimisation techniques, such as particle swarm optimisation and genetic algorithms, can be employed for further refinement.

CONCLUSIONS

The extents of lignin and hemicellulose removal are crucial for assessing crop straw's efficiency in enzymatic hydrolysis. This study proposes an efficient prediction method for the two straw enzymatic hydrolysis efficiency indicators based on GRA-KPCA-LSSVM.

1. First, a prediction model for the enzymatic hydrolysis efficiency was developed by employing GRA variable screening, KPCA input dimension reduction, and LSSVM model training using actual production data.
2. Subsequently, this model was applied using production condition data to accurately estimate the final enzymatic hydrolysis efficiency in real-time.

3. The effectiveness of this method was validated through numerous experimental tests. Through utilising the acquired model for prediction, the authors observed minimal errors and achieved a high level of fitting accuracy, indicating exceptional performance.

As a broader conclusion, the method introduced in this paper offers an optimised design basis for the efficient enzymatic hydrolysis of crops and provides soft sensor support for effectively controlling the enzymatic hydrolysis process. However, the prediction accuracy of the methodology presented in this paper was constrained by the limited scale of training samples and the absence of sophisticated parameter optimization strategies. Future work should concentrate on assembling more extensive datasets, alongside the utilization of advanced modelling algorithms and refined parameter optimization techniques. Such an approach has potential to amplify the accuracy of predictions and bolster the generalizability of the model.

ACKNOWLEDGMENTS

This work is supported by the major industrial projects to transform old and new growth drivers in Shandong Province "Research and industrial application of citric acid green bio-manufacturing technology".

REFERENCES CITED

Adnana, R. M., Lianga, Z., Heddamb, S., Zounemat-Kermanic, M., and Kisid, O. L. B.-Q. (2019). "Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs," *Journal of Hydrology* 586, article ID 124371. DOI: 10.1016/j.jhydrol.2019.124371

Agrawal, R., Verma, A., Singhania, R. R., Varjani, S., Cheng, D.-D., and Patel, A. K. (2021). "Current understanding of the inhibition factors and their mechanism of action for the lignocellulosic biomass hydrolysis," *Bioresource Technology* 332, article ID 125042. DOI: 10.1016/j.biortech.2021.125042

Anowar, F., and Sadaoui, S. (2021). "Incremental learning framework for real-world fraud detection environment," *Computational Intelligence* 37(1), 635-656. DOI: 10.1111/coin.12434

Anowar, F., Sadaoui, S., and Selim, B. (2021). "Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE)," *Computer Science Review* 40, article ID 100378. DOI: 10.1016/j.cosrev.2021.100378

Antos, J., Kubalcik, M., and Kuritka, I. (2022). "Scalable non-dimensional model predictive control of liquid level in generally shaped tanks using RBF neural network," *International Journal of Control and Automation Systems* 20, 1041-1050. DOI: 10.1007/s12555-020-0904-9

Chen, L., and Zhou, S.-S. (2018). "Sparse algorithm for robust LSSVM in primal space," *Neurocomputing* 275, article ID 2880-2891. DOI: 10.1016/j.neucom.2017.10.011

Du, E. (2022). "Impact of bank research and development on total factor productivity and performance evaluation by RBF network," *The Journal of Supercomputing* 78, 12070-12092. DOI: 10.1007/s11227-022-04358-x

Guo, J.-Y., Yu, H., and Li, Y. (2023). "Related and independent variable fault detection method based on KPCA-SVM," *Journal of Shenzhen University Science and Engineering* 40(1), 14-21. DOI: 10.3724/SP.J.1249.2023.01014

Han, Y.-M., Cao, L., Geng, Z.-Q., Ping, W.-Y., Zuo, X.-Y., Fane, J.-Z., Wan, J., and Lu, G. (2022). "Novel economy and carbon emissions prediction model of different countries or regions in the world for energy optimization using improved residual neural network," *Science of The Total Environment* 860, article ID 160410. DOI: 10.1016/j.scitotenv.2022.160410

Huang, C.-X., Lin, W.-Q., Lai, C.-H., Li, X., Jin, Y.-G., and Yong, Q. (2019). "Coupling the post-extraction process to remove residual lignin and alter the recalcitrant structures for improving the enzymatic digestibility of acid-pretreated bamboo residues," *Bioresource Technology* 285, article ID 12355. DOI: 10.1016/j.biortech.2019.121355

Ikram, R. M. A., Dai, H.-L., Ewees, A., Shiri, J., Kisi, O., and Zounemat-Kermani, M. (2022). "Application of improved version of multiverse optimizer algorithm for modeling solar radiation," *Energy Reports* 8, 12063-12080. DOI: 10.1016/j.egyr.2022.09.015

Kuang, F.-J., Xu, W.-H., Zhang, S.-Y., Wang, Y.-H., and Liu, K.W. (2012). "A novel approach of KPCA and SVM for intrusion detection," *Journal of Computational Information Systems* 8(8), 3237-3244.

Kuang, F.-J., Xua, W.-H., and Zhang, S.-Y. (2014). "A novel hybrid KPCA and SVM with GA model for intrusion detection," *Applied Soft Computing* 18, 178-184. DOI: 10.1016/j.asoc.2014.01.028

Kumar, P., Kumar, V., Adelodun, B., Bedekovic, D., Kos, I., Širic´, I., Alamri, S. A. M., Alrumman, S. A., Eid, E. M., Fayssal, S. A., *et al.* (2022). "Sustainable use of sewage sludge as a casing material for button mushroom (*Agaricus bisporus*) cultivation: Experimental and prediction modeling studies for uptake of metal," *Journal of Fungi* 8(2), article 112. DOI: 10.3390/jof8020112

Liu, Y., Cao, Y., Wang, L., Chen, Z.-S., and Qin, Y. (2022). "Prediction of the durability of high-performance concrete using an integrated RF-LSSVM model," *Construction and Building Materials* 356, article ID 129232. DOI: 10.1016/j.conbuildmat.2022.129232

Nguyen, L. T., Phan, D. P., Sarwar, A., Tran, M. H., Lee, O. K., and Lee, E. Y. (2020). "Valorization of industrial lignin to value-added chemicals by chemical depolymerization and biological conversion," *Industrial Crops & Products* 161, article ID 113219. DOI: 10.1016/j.indcrop.2020.113219

Qin, S.-J. (2012). "Survey on data-driven industrial process monitoring and diagnosis," *Annual Reviews in Control* 36(2), 220-234. DOI: 10.1016/j.arcontrol.2012.09.004

Saravanan, A., Senthil Kumar, P., Jeevanantham S., Karishma, S., and Vo, D. N. (2021). "Recent advances and sustainable development of biofuels production from lignocellulosic biomass," *Bioresource Technol.* 344, article ID 126203. DOI: 10.1016/j.biortech.2021.126203

Tian, X., Cheng, H.-Y., Zhang, H.-B., Ren, Y.-S., Wang, Y.-M., Luo, Y., and Liu, N. (2023). "Mechanism of composite ferrate solution pretreatment for promoting

enzymatic hydrolysis efficiency of corn stover," *Acta Scientiae Circumstantiae* 43(4), 417-426. DOI: 10.13671/j.hjkxxb.2022.0351

Tian, Z.-D. (2020). "Short-term wind speed prediction based on LMD and improved FA optimized combined kernel function LSSVM," *Engineering Applications of Artificial Intelligence* 91, article ID 103573. DOI: 10.1016/j.engappai.2020.103573

Usmani, Z., Sharma, M., Awasthi, A. K., Sivakumar, N., Lukk, T., Pecoraro, L., Thakur, V. K., Roberts, D., Newbold, J., and Gupta, V. K. (2021). "Bioprocessing of waste biomass for sustainable product development and minimizing environmental impact," *Bioresource Technol.* 322, article ID 124548. DOI: 10.1016/j.biortech.2020.124548

Wang, J.-Z., and Hu, J.-M. (2015). "A robust combination approach for short-term wind speed forecasting and analysis – Combination of the ARIMA (autoregressive integrated moving average), ELM (extreme learning machine), SVM (support vector machine) and LSSVM (least square SVM) forecasts using a GPR (Gaussian process regression) model," *Energy* 93, 41-56. DOI: 10.1016/j.energy.2015.08.045

Wang, Y.-K., Tang, H.-M., Huang, J.-S., Wen, T., Ma, J.-W., and Zhang, J.-R. (2022). "A comparative study of different machine learning methods for reservoir landslide displacement prediction," *Engineering Geology* 298, article ID 106544. DOI: 10.1016/j.enggeo.2022.106544

Xiong, J., Wang, T., and Li, R. (2018). "Research on a hybrid LSSVM intelligent algorithm in short term load forecasting," *Cluster Computing* 22(Suppl4), 8271-8278. DOI: 10.1007/s10586-018-1740-z

Yang, J., Yang, X.-Q., and Han, L.-J. (2022). "Effects of different NaOH/ball milling combined pretreatments on the enzymatic hydrolysis of corn stalks," *Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE)* 38(15), 226-233. DOI: 10.11975/j.issn.1002-6819.2022.15.024

Yuan, X.-H., Chen, C., Yuan, Y.-B., Huang, Y.-H., and Tan, Q.-X. (2015). "Short-term wind power prediction based on LSSVM–GSA model," *Energy Conversion and Management* 101, 393-401. DOI: 10.1016/j.enconman.2015.05.065

Zhao, L., Sun, Z. F., Zhang, C. C., Nan, J., Ren, N. Q., Lee, D. J., and Chen, C. (2021). "Advances in pretreatment of lignocellulosic biomass for bioenergy production: Challenges and perspectives," *Bioresource Technol.* 343, article ID 126123. DOI: 10.1016/j.biortech.2021.126123

Zhu, J.-B., Song, W.-L., Chen, X., and Sun, S.-N. (2023). "Integrated process to produce biohydrogen from wheat straw by enzymatic saccharification and dark fermentation," *Science Direct* 48, 11153-11161. DOI: 10.1016/j.ijhydene.2022.05.056

Article submitted: December 7, 2023; Peer review completed: March 22, 2024; Revised version received and accepted: April 6, 2024; Published: April 17, 2024.
DOI: 10.15376/biores.19.2.3505-3519