Predicting Lignin Removal from Bark-inclusive Mixed-species Wood Chips *via* Ensemble Machine Learning

Hyeon Cheol Kim, Si Young Ha, and Jae-Kyung Yang *

This study investigated the prediction of lignin removal from mixed-species wood chips (oak and pine) containing bark. A two-stage pretreatment consisting of steam explosion (SE) and alkaline (NaOH) treatment was employed, and a dataset was constructed by systematically varying severity factors, bark inclusion, chemical concentration, and pretreatment time. Four predictive models—Random Forest (RF), Extra Trees (ET), Extreme Gradient Boosting (XGBoost), and polynomial regression (PR)—were developed and evaluated using five repeated random 80:20 sample-level splits. XGBoost showed the highest mean test R^2 (0.9848), the lowest mean test RMSE (0.2984), and the highest mean cross-validation R^2 (0.9361), significantly outperforming conventional polynomial regression (R^2 = 0.5145). SHAP analysis identified NaOH concentration as the most influential predictor, followed by severity factor and bark inclusion. Finally, a web-based graphical user interface (GUI) was developed using Streamlit to provide real-time lignin removal predictions. These results showed that ensemble machine learning, particularly XGBoost, can effectively optimize pretreatment processes for heterogeneous biomass feedstocks in industrial biorefinery applications. These results support the use of ensemble machine learning for this heterogeneous feedstock, while indicating sensitivity to the sample-level partition and the need for future condition-level validation.

DOI: 10.15376/biores.21.4.10540-10559

Keywords: Alkaline pretreatment; Extra trees; Lignin removal; Mixed-species wood chips; Steam explosion

Contact information: Department of Environmental Materials Science/Institute of Agriculture and Life Science, Gyeongsang National University, Jinju, 52828, Republic of Korea;
** Corresponding author: jkyang@gnu.ac.kr*

INTRODUCTION

Lignocellulosic biomass, particularly wood biomass, has been recognized as one of the most promising renewable feedstocks for the production of biofuels and biochemicals, owing to its carbon-neutral characteristics, wide availability, and high carbohydrate content (Jung and Yang 2018; Okolie *et al.* 2021; Inyang *et al.* 2022). Wood biomass is primarily composed of cellulose, hemicellulose, and lignin, of which cellulose and hemicellulose — accounting for approximately 60 to 70% of the dry biomass weight — are utilized as the principal substrates for fermentable sugar production (Woźniak *et al.* 2025). However, lignin, a complex aromatic polymer tightly interwoven with the carbohydrate matrix, is considered a major structural barrier to all downstream conversion processes, as carbohydrate accessibility is reduced and hydrolytic catalysts are inhibited by its presence (Djajadi *et al.* 2018; Zhao *et al.* 2026).

Efficient lignin removal is therefore regarded as a critical, though not by itself sufficient, prerequisite for improving the conversion efficiency of lignocellulosic biomass; downstream saccharification efficiency additionally depends on residual lignin chemistry and distribution, cellulose accessibility, residual alkali content, and extractive composition (Chakraborty *et al.* 2024; Chen *et al.* 2024; Liu *et al.* 2025). Lignin is removed from lignocellulosic biomass through a variety of pretreatment methods (Mankar *et al.* 2021). These pretreatment methods are broadly categorized into physical pretreatments (*e.g.*, ball milling, sonication), chemical pretreatments (*e.g.*, alkaline, acid), and physicochemical pretreatments (*e.g.*, steam explosion) (Brodeur *et al.* 2011; Karimi *et al.* 2013).

Among these, steam explosion (SE) has been widely employed as an effective physicochemical pretreatment, for bioenergy and biorefinery applications (rather than for papermaking pulp production) in which the lignocellulosic matrix is disrupted by thermomechanical forces, hemicellulose is primarily solubilized, and lignin structure is modified to increase carbohydrate accessibility for subsequent enzymatic or chemical treatment (Auxenfans *et al.* 2017; Ha *et al.* 2024b).

However, lignin removal achieved by SE alone is limited, and an additional post-treatment step is therefore required to improve overall lignin removal rate (Tomas-Pejo *et al.* 2008). Alkaline treatment can selectively solubilize structurally weakened lignin through saponification of lignin–carbohydrate ester linkages (Takahashi and Koshijima 1988; Yang *et al.* 2002). In this study, NaOH post-treatment was performed under atmospheric-pressure, room-temperature conditions, avoiding the need for a second pressure vessel beyond the steam explosion stage. This mild, non-pressurized step may offer capital-cost and operational-safety advantages for smaller-scale or retrofit biorefinery operations (Cara *et al.* 2006; Min *et al.* 2014). Its selective-delignification concept also parallels recently reported low-dosage, solvent-assisted deconstruction strategies for heterogeneous lignocellulosic feedstocks (Zhang *et al.* 2025).

Extensive research has been conducted on two-stage pretreatment combining steam explosion and chemical treatment, yet the overwhelming majority of studies still use single-species, debarked wood under controlled laboratory conditions (Matsakas *et al.* 2018; Steinbach *et al.* 2020). Extensive research has been conducted on two-stage pretreatment combining steam explosion and chemical treatment, yet most studies use single-species, debarked wood under controlled laboratory conditions (Matsakas *et al.* 2018; Steinbach *et al.* 2020). This study does not target conventional kraft or chemical pulping for papermaking, where hardwood and softwood are generally separated to optimize cooking conditions. Instead, it addresses smaller-scale bioenergy and biorefinery operations, for which dedicated species-sorting and debarking infrastructure may not be economically justified. Mixed-species chips containing bark can therefore represent a commercially relevant feedstock for such operations (Belbo and Talbot 2014; Nurek *et al.* 2019; Hörhammer *et al.* 2018). Bark constitutes approximately 10 to 20% of stem dry mass and has higher lignin and extractive contents than inner wood, increasing resistance to alkaline lignin removal (Dossa *et al.* 2016; Rosdiana *et al.* 2017; Kusiak *et al.* 2022; Xue *et al.* 2024). In our previous work, lignin removal was studied for bark-inclusive oak and pine feedstocks separately (Ha *et al.* 2024a; Ha *et al.* 2025). To the best of our knowledge, systematic modeling of lignin removal from bark-inclusive mixed-species wood feedstocks remains limited.

Under heterogeneous feedstock conditions, the relationship between SE severity factor, bark inclusion, NaOH concentration, treatment time, and lignin removal is nonlinear and multivariate (Kim *et al.* 2025). The dataset combines continuous variables with a binary-coded bark variable, producing an interaction-dependent response surface that is difficult to represent using a conventional polynomial form (López *et al.* 2020). Machine learning provides an approach capable of capturing interactions among nonlinear, multivariate variables and predicting their combined effect on the outcome (Zhong *et al.* 2025). Random Forest (RF), Extra Trees (ET), and Extreme Gradient Boosting (XGBoost) use decision-tree computational structures to partition the feature space. The term "tree" as used in our research (*e.g.,* decision tree, Extra Trees, *etc*.) refers to a decision tree; it has no relation to feedstock. These ensemble methods have shown strong predictive performance for structured nonlinear biomass datasets (Geurts *et al.* 2006; David *et al.* 2025). However, their applicability to lignin-removal prediction for bark-inclusive mixed-species wood chips has not been established.

In this study, a dataset was constructed from steam explosion and alkaline pretreatment experiments using mixed-species oak (*Quercus* spp.)–pine (*Pinus* spp.) wood chips. The experiments were conducted under systematically varied conditions, including severity factors, bark inclusion, NaOH concentration, and pretreatment time. Four predictive models, including RF, ET, XGBoost, and polynomial regression, were developed and evaluated using R^2, RMSE, and MAE. SHAP analysis was subsequently applied to examine the contribution and interaction of the process variables. This model-specific interpretation was aligned with the updated optimum model, whereas the repeated-split analysis was used to compare the robustness of all four models to changes in the train-test partition. The main contribution of this study consists of three related aspects: the use of a bark-inclusive mixed-species oak–pine feedstock, the application of two-stage SE–NaOH pretreatment to this heterogeneous material, and the development of an ensemble machine-learning framework with SHAP interpretation for lignin-removal prediction. To the best of our knowledge, these aspects have not previously been investigated quantitatively in combination. The Streamlit GUI was developed as a supporting tool for applying the trained model and is not presented as a separate core contribution.

EXPERIMENTAL

Wood Chips

Wood chips were purchased from Punglim Co., Ltd., located in Daejeon, Republic of Korea; the chips were not sorted, debarked, or otherwise prepared to papermaking-pulp specifications, consistent with the bioenergy/biorefinery feedstock scenario addressed in this study. Wood chips composed of domestic oak (*Quercus* spp.) and pine (*Pinus* spp.) were prepared at a chip size of approximately 3 cm (W) × 3 cm (L) × 0.5 cm (H). Bark accounted for approximately 10% of the chip volume. The two species were mixed at a 1:1 weight ratio and homogenized manually in a large steel basin prior to being transferred to polyethylene storage bags. Both bark-inclusive and bark-removed preparations were stored in a dark, well-ventilated area until the initiation of experiments. The average initial moisture content of the mixed-species chips was 35.0 ± 0.59% on a wet weight basis, as determined prior to steam explosion treatment.

Steam Explosion Treatment

The wood chip samples were used for steam explosion treatment without prior drying and the average initial moisture content was 35.0 ± 0.59% on a wet weight basis. Steam explosion was performed at a pilot scale (Yurim Hi-Tech Co., Ltd., Daegu, Republic of Korea). The retention time was counted from the point at which steam was introduced into the reactor. Steam explosion treatment was conducted at 225 °C with varying retention times. The Severity Factor (R_0) was calculated using Eq. 1, with temperature and retention time as variables,

$$R_0 = \log\left\{t \cdot \exp\left(\frac{T_r - T_b}{14.752}\right)\right\} \tag{1}$$

where t is the retention time (min), T_r is the retention temperature (°C), and T_b is the base temperature (100 °C). The results are presented in Table 1. Steam explosion-treated samples were sealed in polypropylene bags and stored in a refrigerator at 4 °C until the initiation of experiments.

Table 1. Steam Explosion Condition by Severity Factor Modeling

No.	Severity Factor (R_0)	Retention Temperature (°C)	Retention Time (min)
1	4.25	225	1
2	4.73	225	3
3	4.95	225	5

Alkaline Post-treatment

Lignin was removed from the steam-exploded samples using a sodium hydroxide (NaOH) solution. Aqueous NaOH solutions at concentrations of 0.5%, 1%, and 2% (w/v) were prepared using distilled water. The steam-exploded samples and NaOH solution were mixed at a ratio of 1:20 (w/v) in 300 mL Erlenmeyer flasks and sealed with aluminum foil. The mixtures were allowed to react statically at room temperature for 12 h and 24 h, away from direct sunlight. After the pretreatment, the samples were filtered by gravity filtration using filter paper (Whatman No. 2). The filter residues were washed with distilled water while monitoring the pH with pH paper (AdvanTec, Japan) until a neutral pH of 7 was achieved. The neutralized samples were sealed in polypropylene bags and stored in a refrigerator at 4 °C until further use.

Klason Lignin Content Determination

To determine the lignin content of steam-exploded (SE) and steam-exploded alkaline-pretreated (SEA) samples, the procedure described in ASTM D1106-56 was performed with slight modifications. Briefly, 0.3 g of sample and 3 mL of 72% sulfuric acid (H_2SO_4) solution were added to a glass vial. The mixture was stirred with a glass rod, sealed with aluminum foil, and reacted in a water bath at 30 °C for 60 min. After the reaction, 84 mL of distilled water was added, and the vial was sealed with aluminum foil and autoclaved at 121 °C for 60 min. After completion of the reaction, the mixture was filtered through a 2G3 glass filter, and the weight of the residue was measured to determine the Klason lignin content. The lignin removal (%) was calculated using Eq. 2,

$$\text{Lignin Removal (\%)} = \frac{W_0 - W_1}{W_0} \times 100 \tag{2}$$

where W_0 is the Klason lignin mass (g) of the steam-exploded sample prior to alkaline post-treatment, and W_1 is the Klason lignin mass (g) of the steam-exploded and alkaline post-treated sample. Both values were determined relative to the same initial dry sample mass basis. We note that this metric reflects lignin removal specifically and does not constitute a full mass balance; solid recovery, carbohydrate retention, ash/extractive contribution, and lignin relocation during steam explosion were not independently quantified in this study.

Statistical Analysis and Spearman Correlation Matrix

A statistical analysis was performed on the dataset to determine the range of the target variable and to identify the presence of significant outliers. Based on the data distribution obtained from the statistical analysis, an appropriate correlation method was selected. The Spearman correlation matrix is sufficiently applicable to complex datasets exhibiting highly nonlinear or non-normal distributions (Xiao *et al.* 2016). Accordingly, Spearman correlation analysis was conducted using the four input variables of the model. Prior to the correlation analysis, the bark binary variable was retained as a binary-coded categorical variable (0 = without bark, 1 = with bark), and all continuous variables were used in their original units without normalization. Spearman correlation was performed on condition-averaged data (n = 36) to preserve the statistical independence of observations.

Prediction of Lignin Removal

Machine Learning Algorithm

To benchmark the predictive performance of lignin removal through alkaline pretreatment following steam explosion, three machine learning algorithms and a polynomial regression model were compared. Random Forest (RF) constructs an ensemble of decision trees based on bootstrap samples and averages the outputs to capture complex interactions within the data (Biau and Scornet 2016). Extreme Gradient Boosting (XGBoost) employs gradient boosting with regularization and parallel computing, achieving high predictive accuracy on structured data while effectively handling missing values (Gao *et al.* 2022). Extra Trees (ET) generates extremely randomized trees and computes piecewise multilinear approximations to improve predictive accuracy and mitigate overfitting (Geurts *et al.* 2006). Polynomial regression (PR) was selected as a baseline model representing a conventional prediction approach, against which the performance of the machine learning models was assessed (Ostertagová 2012). This diverse set of algorithms provides a comprehensive evaluation of modeling strategies for pretreatment system design.

Configuration and Optimization of Machine Learning Models

Three machine learning models and a ridge-regularized polynomial regression model were developed using process variables to predict the lignin removal in steam explosion and alkaline pretreatment systems. The severity factor, bark binary (coding the presence of bark as a binary variable), pretreatment time, and chemical concentration were used as training features. Using these features, four models were constructed: PR, RF, ET and XGBoost. The RR, RF, and ET models were implemented using the scikit-learn library, while XGBoost was implemented using its dedicated optimized library (Pedregosa *et al.* 2011). Model development followed a structured and reproducible workflow.

The dataset was randomly split at the sample (replicate) level into training and test subsets at a ratio of 80:20, respectively. Because the split was performed at the sample level, replicate measurements from the same experimental condition could be included in both subsets. The training data were used exclusively for model construction and hyperparameter tuning, whereas the test data were withheld from training and reserved solely for the final evaluation of predictive performance on the held-out samples. Accordingly, this evaluation reflects model performance for sample-level observations drawn from the experimental conditions represented in the dataset rather than generalization to entirely unseen experimental conditions.

Hyperparameter optimization was performed on the training set using a grid search algorithm in conjunction with 5-fold cross-validation (Alibrahim and Ludwig 2021). Predefined combinations of candidate hyperparameters were systematically evaluated, and the combination yielding the highest mean cross-validation performance was selected for each model. The hyperparameter search space for each model is presented in Table 2.

Table 2. Hyperparameter Search Space and Optimal Values Identified *via* GridSearchCV for Each Model

Model	Hyperparameter	Search grid	Optimal value
RF*	n_estimators	[50, 100, 200]	200
	max_depth	[None, 5, 10, 20]	None
	min_samples_split	[2, 5]	2
ET*	n_estimators	[50, 100, 200]	50
	max_depth	[None, 5, 10, 20]	None
	min_samples_split	[2, 5]	2
XGBoost*	n_estimators	*[50, 100, 200]*	200
	max_depth	*[3, 5, 7]*	3
	learning_rate	*[0.01, 0.1, 0.2]*	0.2
PR (Ridge)*	poly__degree	[2, 3]	3
	model__alpha	[0.1, 1.0, 10.0, 100.0]	1.0

* RF: Random Forest
* ET: Extra Trees
* XGBoost: Extreme Gradient Boosting
* PR: Ridge-regularized Polynomial Regression

Model Evaluation

The predictive performance of all models was evaluated using the coefficient of determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), as defined in Eqs. 3, 4, and 5, respectively (Varoquaux and Colliot 2023),

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \tag{5}$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean of the observed values, and n is the number of samples.

Model Interpretation

To interpret the predictive model and evaluate the contribution of each input variable to the lignin removal, SHapley Additive exPlanations (SHAP) analysis was performed on the best-performing model (Albini *et al.* 2022). SHAP values were calculated using the shap library in Python. The results were visualized using a Summary Plot, which illustrates the distribution and magnitude of SHAP values for each input feature, and a Bar Plot, which represents the mean absolute SHAP values to quantify the overall feature importance.

Graphical User Interface (GUI) Development

A web-based Graphical User Interface (GUI) was developed using the Streamlit framework in Python to facilitate practical application of the predictive model (Khorasani *et al.* 2022). The GUI allows users to input process parameters, including severity factor, bark binary, pretreatment time, and NaOH concentration, and obtain the predicted lignin removal percentage in real time. A web-based graphical user interface (GUI) was developed using Streamlit to facilitate application of interpretation in this manuscript focuses on machine learning model as the strongest repeated-split model.

RESULTS AND DISCUSSION

Statistical Characteristics and Spearman Correlation Analysis

The statistical characteristics of the dataset and the Spearman correlation coefficients between the four input variables and the lignin removal are presented in Fig. 1.

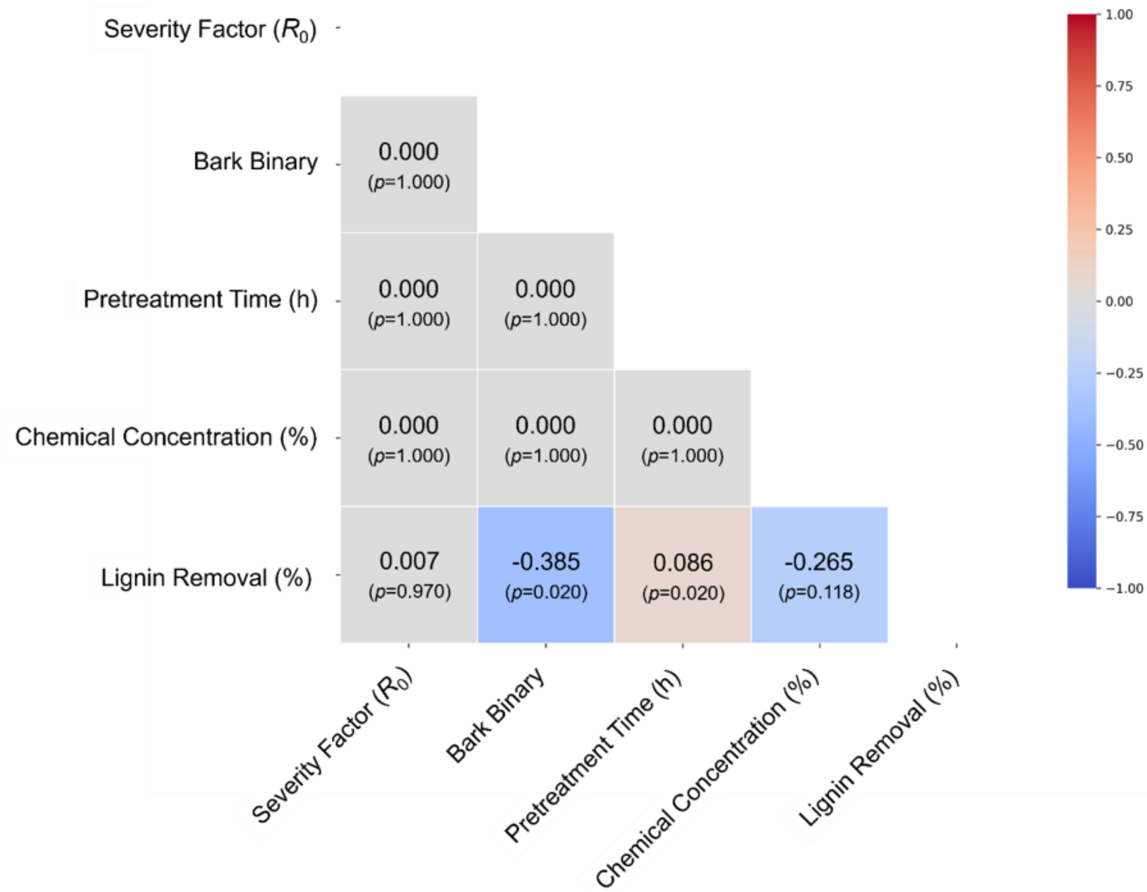

Fig. 1. Spearman correlation analysis between the four input variables (severity factor, bark inclusion, NaOH concentration, and alkaline treatment time) and lignin removal (%)

Spearman correlation analysis was performed on condition-averaged data (n = 36) to ensure statistical independence of observations, given that each experimental condition was conducted in triplicate. A statistically significant correlation was identified between bark binary and lignin removal percentage (r = −0.390, p = 0.019) among the four input variables. This finding suggests that the presence of bark is associated with reduced lignin removal. This is consistent with the substantially higher lignin content and extractive compound concentrations characteristic of bark tissue compared with inner wood, which are known to increase resistance to alkaline lignin removal (Rosdiana *et al.* 2017; Kusiak *et al.* 2022). The remaining three variables – severity factor ($r = 0.013$, $p = 0.940$), pretreatment time ($r = 0.086$, $p = 0.620$), and NaOH concentration ($r = -0.262$, $p = 0.123$) – did not exhibit statistically significant monotonic correlations with lignin removal at the $\alpha = 0.05$ level.

It is noteworthy that the NaOH concentration exhibited a negative correlation with the lignin removal. This finding appears counterintuitive, given that higher alkali concentrations are generally expected to enhance lignin removal (Wang *et al.* 2020; Yang *et al.* 2021). However, an analysis of the experimental data demonstrated that lignin removal percentages under 2% NaOH conditions were lower than those obtained under 0.5% NaOH conditions. This phenomenon is attributed to the combined influence of feedstock heterogeneity and severity factor interactions. The lignin removal value reported here is calculated from Klason lignin content alone and, as such, characterizes delignification specifically rather than overall pretreatment efficiency. A complete assessment would additionally require solid recovery yield, carbohydrate (cellulose/hemicellulose) retention, the contribution of ash and extractives to the Klason residue, and quantification of lignin relocation/redeposition occurring during steam explosion (Li *et al.* 2007; Heikkinen *et al.* 2014). However, these limitations in the current scope reflect the inherent complexity underlying lignin removal mechanisms: under high-severity steam explosion conditions, acid-catalyzed condensation and repolymerization of lignin fragments (increased C–C linkage formation and β-O-4 cleavage) reduce lignin solubility in alkali, and at elevated NaOH concentration the condensed lignin fraction may additionally re-precipitate onto the fiber surface once the system approaches saturation, partially offsetting the saponification of lignin–carbohydrate ester linkages expected at higher alkalinity (Wang *et al.* 2012; Das *et al.* 2019; Bressi *et al.* 2023; Song *et al.* 2026). This mechanism is consistent with the strong severity factor with NaOH concentration interaction identified by SHAP. The absence of statistically significant linear correlations for three of the four input variables does not imply that these variables are unimportant predictors of lignin removal; rather, the relationships between pretreatment input variables and lignin removal are governed primarily by complex multivariate interactions that cannot be adequately captured through univariate statistical approaches. These nonlinear and multivariate interaction patterns were therefore leveraged as training data for ensemble machine learning models capable of learning such complex dependencies. Mass-balance-focused work quantifying solid recovery and carbohydrate retention remains an important direction for future research on this feedstock.

Comparison of Model Predictive Performance

The predictive performances of ET, RF, XGBoost, and PR were evaluated using training and test metrics and 5-fold cross-validation. Table 3 reports the mean ± standard deviation across five repeated random 80:20 sample-level splits, whereas Fig. 2 illustrates predictions from the original split. Although R^2 is a widely used metric, it is not considered a sufficient indicator of how well a model fits the data (Onyutha 2020), and therefore multiple performance metrics were assessed to provide a comprehensive evaluation. XGBoost achieved the highest mean test R^2 (0.9848 ± 0.0135) and the lowest mean test RMSE (0.2984 ± 0.0893), with a test MAE of 0.2286 ± 0.0507. ET achieved a mean test

R^2 of 0.9533 ± 0.0868 and RMSE of 0.3911 ± 0.4730, while providing the lowest mean test MAE of 0.2176 ± 0.1875. RF showed test RMSE, MAE, and R^2 values of 0.7281 ± 0.2556, 0.4629 ± 0.1381, and 0.9193 ± 0.0548, respectively. PR showed substantially lower test performance (RMSE = 1.5514 ± 0.1243, MAE = 1.2215 ± 0.0977, and R^2 = 0.6548 ± 0.1272).

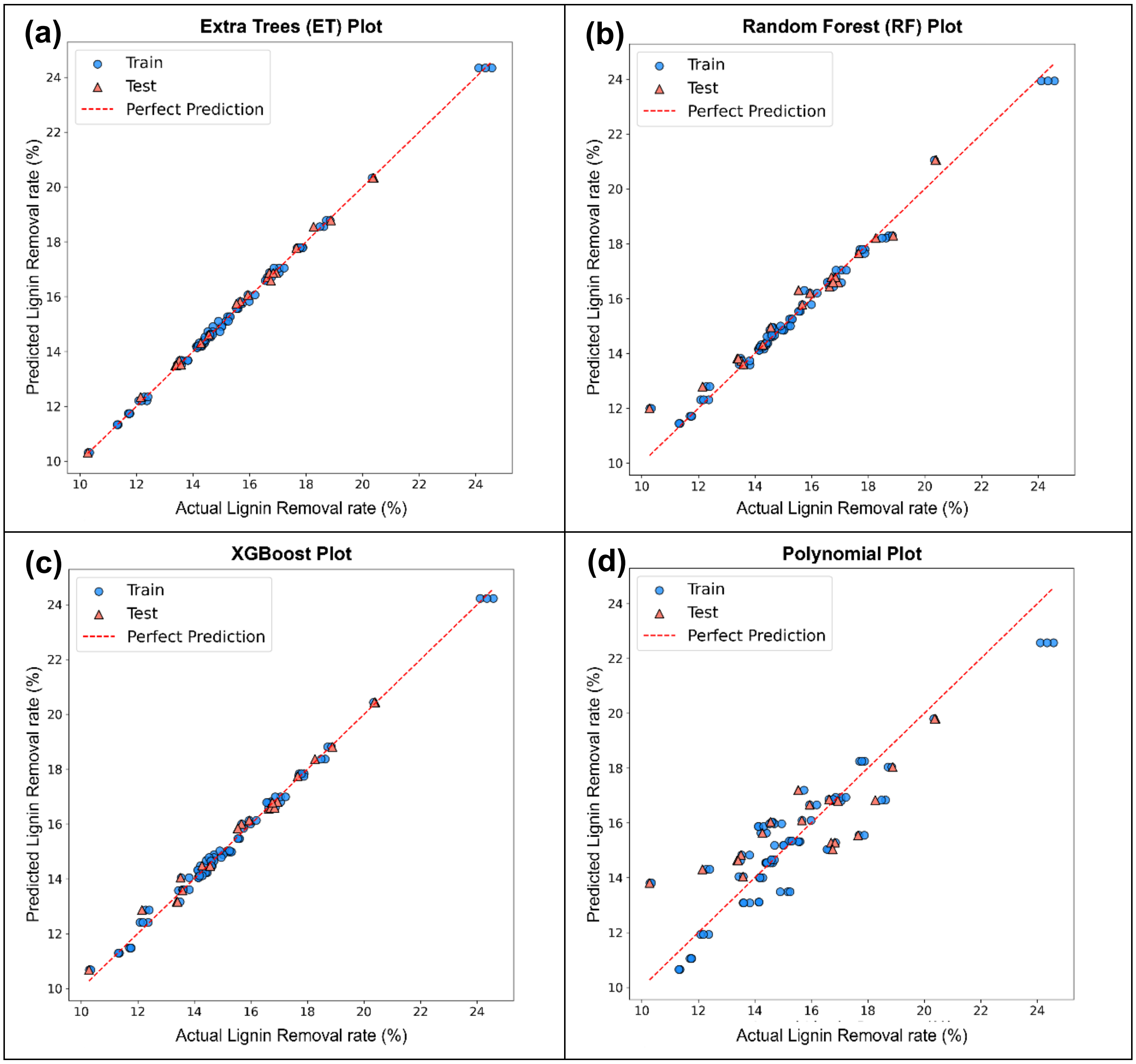

Fig. 2. Comparison of predicted and measured values across train and test sets for ET (a), RF (b), XGBoost (c), and polynomial (d) models

The repeated cross-validation results showed the same overall separation between the ensemble models and PR. XGBoost had the highest mean cross-validation R^2 (0.9361 ± 0.0345), followed by ET (0.8985 ± 0.0813), RF (0.8326 ± 0.0712), and PR (0.5145 ± 0.1347). Although ET remained a strong model and produced the lowest mean test MAE, its comparatively large test RMSE and R^2 standard deviations indicate greater sensitivity to the composition of the training and test subsets. In contrast, XGBoost combined the highest mean test and cross-validation R^2 values with lower variability.

The mean ET test R^2 from the repeated analysis (0.9533 ± 0.0868) was characterized by considerable variability. This result indicates that model performance estimates can be influenced by the selected partition and should not be used alone as evidence of generalization. The repeated analysis reduces reliance on a single favorable split, but all partitions were performed at the sample level. Because replicates from the same experimental condition could occur in both subsets, these results assess robustness to sample-level partitioning and do not demonstrate performance for completely unseen experimental conditions.

The repeated results identified XGBoost as the strongest model in terms of mean test R^2, mean test RMSE, and mean cross-validation R^2. ET provided very high training accuracy and the lowest mean test MAE, but its larger test RMSE and R^2 standard deviations indicate greater sensitivity to the sample-level partition. XGBoost's superiority, attributable to its regularization mechanisms, suggests that under the current dataset size and feature dimensionality, the bagging-based architecture of ET may be prone to marginal overfitting, whereas the boosting approach provides more stable generalization capability (Fatima *et al.* 2023). Future work should include grouped or leave-one-condition-out validation to evaluate transfer to experimental conditions not represented during training.

PR coefficients offer exact and globally consistent interpretation, but this advantage depends on the adequacy of the assumed functional form. The lower mean test R^2 of PR in the repeated analysis (0.6548 ± 0.1272) indicates limited ability to represent the nonlinear and interaction-dependent relationships in this dataset. SHAP analysis of the optimized XGBoost model instead provides local attributions that can be aggregated into global importance, directional summaries, and pairwise interaction rankings. These explanations require more care than a single regression coefficient and remain specific to the fitted XGBoost model. SHAP-based interpretation is therefore complementary to, rather than directly equivalent to, coefficient-based interpretation.

Table 3. Comparison of RMSE, MAE, R^2, and Cross-validation R^2 for ET, RF, XGBoost, and Polynomial Regression Models

Model	Train Data			Test Data			Cross Validation
	RMSE	MAE	R^2	RMSE	MAE	R^2	R^2
ET	0.0947 ±0.0049	0.0724 ±0.0040	0.9986 ±0.0001	0.3911 ±0.4730	0.2176 ±0.1875	0.9533 ±0.0868	0.8985 ±0.0813
RF	0.3114 ±0.0313	0.1796 ±0.0091	0.9848 ±0.0037	0.7281 ±0.2556	0.4629 ±0.1381	0.9193 ±0.0548	0.8326 ±0.0712
XGBoost	0.2166 ±0.0397	0.1709 ±0.0253	0.9926 ±0.0026	0.2984 ±0.0893	0.2286 ±0.0507	0.9848 ±0.0135	0.9361 ±0.0345
Polynomial	1.1391 ±0.0325	0.8696 ±0.0390	0.7980 ±0.0299	1.5514 ±0.1243	1.2215 ±0.0977	0.6548 ±0.1272	0.5145 ±0.1347

Feature Importance Analysis via SHAP

To interpret the predictive behavior of the XGBoost model and quantify the contribution of each input variable to lignin removal rate prediction, SHAP analysis was performed. The SHAP bar plot presents the mean absolute SHAP value for each input variable, providing a quantitative ranking of overall predictive importance (Fig. 4). NaOH concentration ranked first (mean |SHAP| = 1.2211), followed by severity factor (R_0; 0.8979), bark binary (0.7337), and pretreatment time (0.3132).

This ranking confirms that NaOH concentration was the primary determinant of lignin removal prediction and is consistent with the established role of alkali loading in disrupting lignin-carbohydrate ester linkages through saponification (Dziekońska-Kubczak *et al.* 2018; Puss *et al.* 2023).

The SHAP summary plot visualizes the direction and magnitude of each variable's contribution to lignin removal prediction across all observations (Fig. 3). For NaOH concentration, lower feature values were generally associated with positive SHAP values, whereas higher feature values tended to cluster near zero or negative SHAP values. This asymmetric pattern indicates that reduced alkali loading can promote lignin removal under the present experimental conditions, whereas elevated NaOH concentrations may produce condition-dependent inhibitory outcomes. This observation is consistent with the negative Spearman correlation coefficient identified in the univariate analysis ($r = -0.262$; Fig. 1) and suggests that the inhibitory effect of high NaOH concentrations may be attributable to lignin condensation or repolymerization induced under conditions of excess alkali, which renders the lignin matrix less susceptible to alkaline solubilization (Mirahmadi *et al.* 2010; Wang *et al.* 2012). The capacity of the XGBoost model to resolve this nonlinear, condition-dependent relationship represents a meaningful advantage over conventional regression approaches.

Severity factor (R_0) exhibited a negligible univariate Spearman correlation with lignin removal ($r = 0.013$, $p = 0.940$) yet ranked second in overall predictive importance (Fig. 4). As shown in Fig. 3, its SHAP value distribution was diffused, with both high and low feature values scattered across positive and negative SHAP values without a simple monotonic trend. This finding suggests that severity factor does not independently determine lignin removal predictions but rather exerts its predictive contribution through conditional interactions with other variables. Bark binary exhibited a comparatively clear directional pattern, though its relative predictive contribution was lower than that of NaOH concentration and severity factor. Pretreatment time received the lowest importance score and exhibited SHAP values clustered tightly around zero, indicating that within the experimental range of 12 to 24 h, treatment duration exerted a marginal influence on lignin removal (Gao *et al.* 2020).

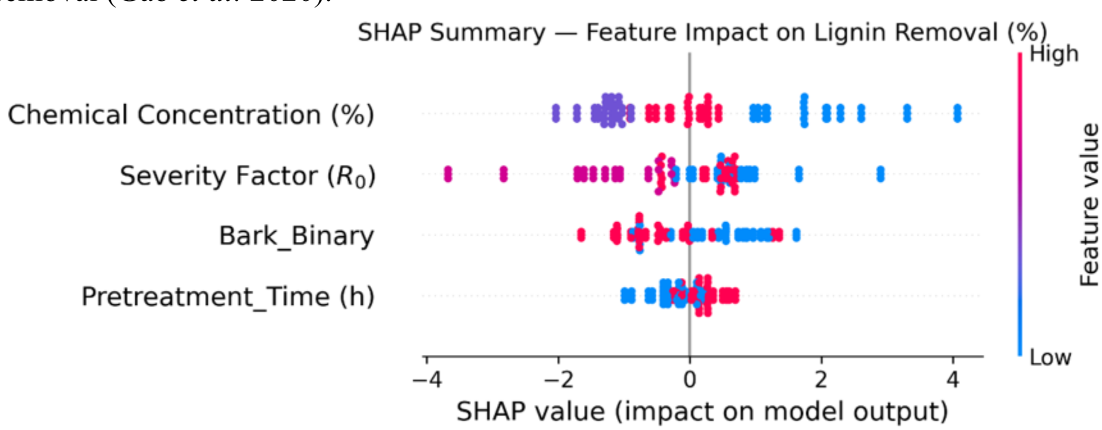

Fig. 3. SHAP summary plot of input variables' contributions to lignin removal prediction by the ET model

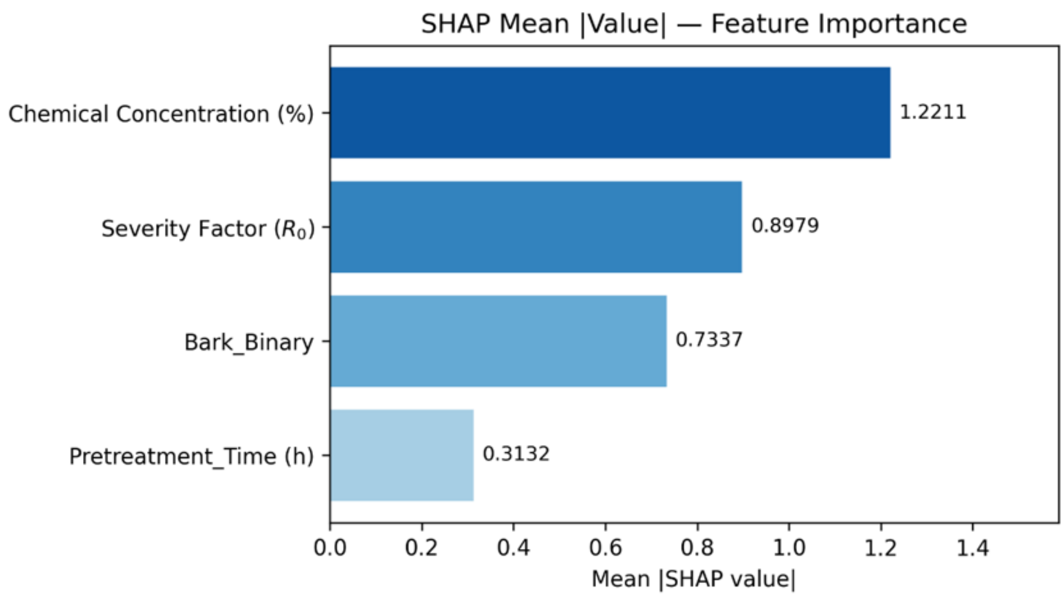

Fig. 4. SHAP bar plot of mean absolute feature importance for lignin removal prediction by the ET model

The pairwise SHAP interaction ranking (Fig. 5) reveals that the predictive structure of the XGBoost model was dominated by interactions involving severity factor. The strongest interaction was between severity factor and bark binary (mean |SHAP interaction value| = 0.4488), followed by severity factor with NaOH concentration (0.3327) and bark binary with NaOH concentration (0.3016). The presence of severity factor in the two strongest interaction pairs helps explain why a variable with virtually no univariate correlation with lignin removal ($r = 0.013$, $p = 0.940$) nonetheless ranked second in overall feature importance.

In the XGBoost model, severity factor does not drive lignin removal on its own; its predictive value lies in how it modulates the effects of feedstock composition and alkali loading. This is mechanistically coherent because thermomechanical and autohydrolysis alterations induced during steam explosion, including partial β-O-4 bond cleavage, lignin relocation, and competing condensation reactions, are known to alter the structural accessibility of lignin before alkaline treatment begins (Li *et al.* 2007; Heikkinen *et al.* 2014; Auxenfans *et al.* 2017b).

The third-ranked interaction, between bark binary and NaOH concentration (0.3016), suggests that the modeled effect of feedstock composition changes with alkali concentration. This pattern is consistent with the condition-dependent behavior discussed above, but it should be interpreted as an association learned by the XGBoost model rather than as direct evidence of a chemical mechanism. Independent experiments are required to test whether the proposed condensation, relocation, or re-adsorption mechanisms explain this interaction.

Taken together, these results indicate that the XGBoost model's internal decision structure is consistent with a chemically plausible picture: lignin removal from a compositionally heterogeneous, mixed-species feedstock is unlikely to be governed by any single variable acting in isolation, but rather by a network of conditional dependencies among severity factor, bark content, and alkali concentration that univariate analysis would miss. We emphasize that SHAP values quantify model-based associations rather than directly proven causal mechanisms; the chemical interpretations offered here should be regarded as hypotheses consistent with the observed patterns, pending confirmation by targeted experiments (*e.g.*, lignin structural characterization across the severity-alkali matrix).

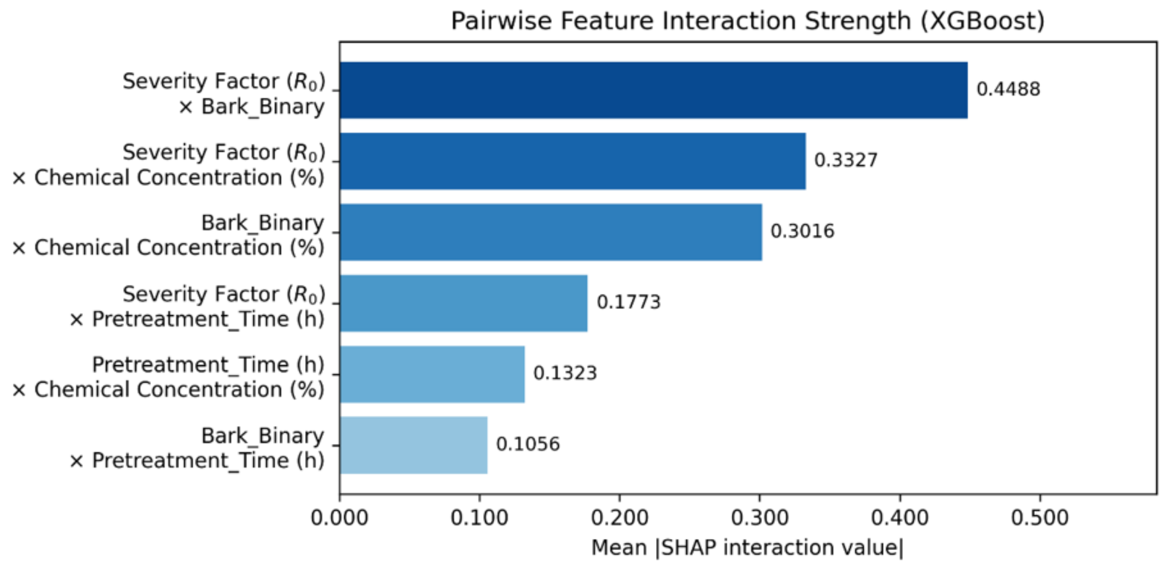

Fig. 5. Pairwise SHAP interaction values indicating the magnitude of feature interaction effects on lignin removal prediction

Web-based GUI for Real-time Lignin Removal Prediction

A web-based graphical user interface (GUI) was developed using Streamlit to facilitate application of the predictive models (Fig. 6). The GUI allows users to enter severity factor, bark inclusion, NaOH concentration, and pretreatment time, select among the evaluated models, and obtain a lignin-removal prediction without programming. Its use should be limited to the feedstock composition and process-variable ranges represented in the present dataset, and it should currently be regarded as a proof-of-concept screening tool rather than a validated industrial process-design system.

The trained ET model and GUI are intended for use without retraining only within the feedstock and process envelope examined in this study: an oak–pine mixture near the tested 1:1 ratio, bark content near 10% of chip volume, and the tested ranges of severity factor, NaOH concentration, and treatment time.

A materially different species mixture, bark fraction, or process range is expected to require model recalibration or retraining because tree-based ensemble models do not reliably extrapolate beyond their training distribution.

In addition, lignin removal was the sole modeled target; downstream enzymatic hydrolysis or fermentation performance was not evaluated and should not be inferred from the GUI output.

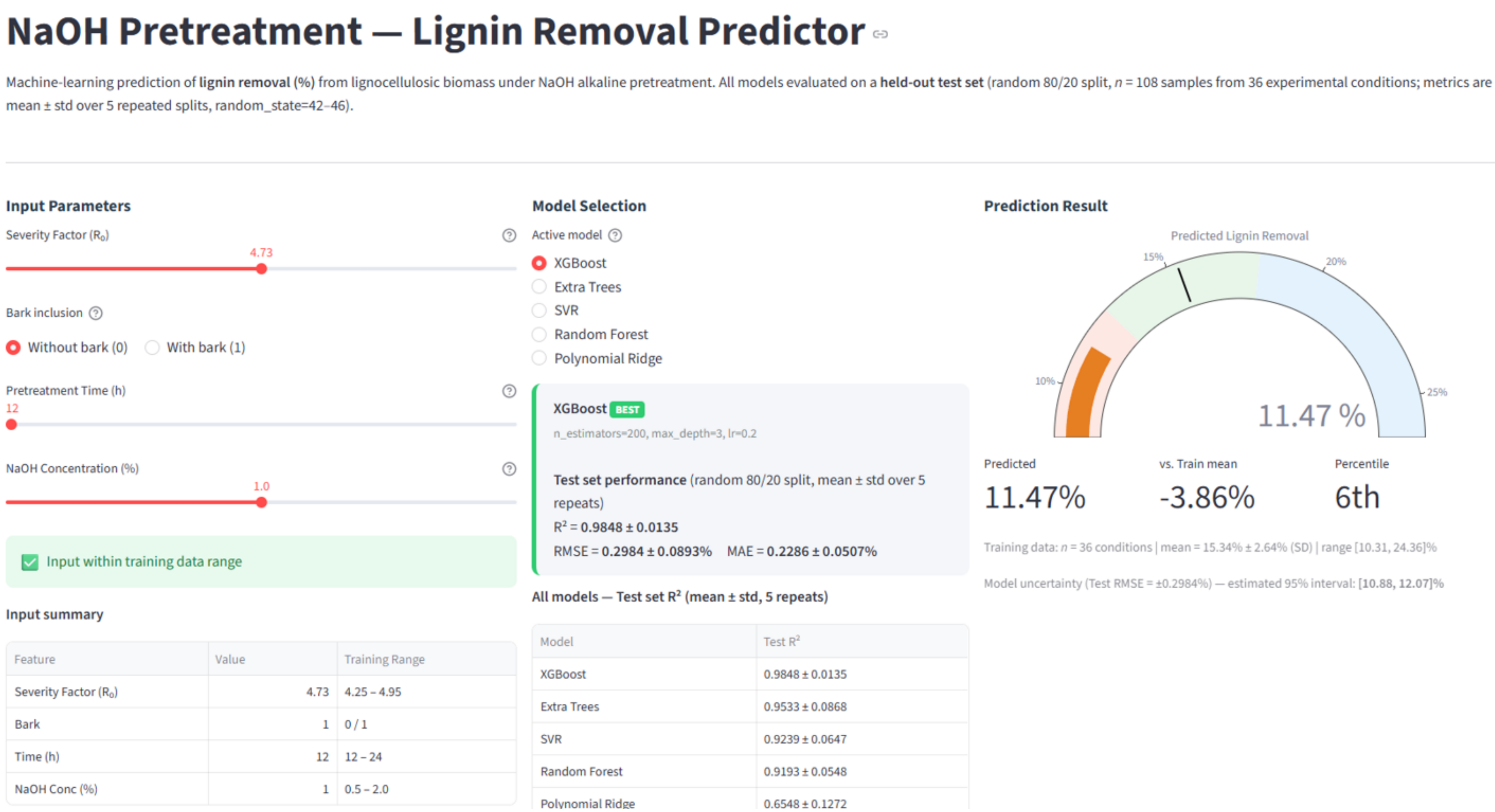

Fig. 6. XGBoost-based lignin removal prediction graphical user interface (GUI)

CONCLUSIONS

1. Across five repeated sample-level splits, XGBoost achieved the highest mean test R^2 (0.9848 ± 0.0135) and cross-validation R^2 (0.9361 ± 0.0345). ET produced the lowest mean test MAE (0.2176 ± 0.1875) but showed greater split-to-split variability. The ensemble models substantially outperformed PR; however, condition-level generalization remains to be evaluated using grouped validation.
2. SHAP analysis identified NaOH concentration as the most influential predictor, while pairwise interaction analysis revealed that severity factor functioned as a conditional modulator, with its two strongest interactions involving bark binary (0.4488) and NaOH concentration (0.3327). These interaction effects explain the failure of univariate Spearman correlations to reflect the true directional contributions of individual variables, demonstrating that lignin removal from heterogeneous biomass is governed by conditional dependencies that univariate approaches cannot capture.
3. The model-selection GUI provides rapid lignin-removal screening only within the tested feedstock and process ranges. It should be regarded as a proof-of-concept tool; different species mixtures, bark contents, or process conditions will require recalibration, retraining, and external validation. Future work should also link delignification with downstream conversion and lignin valorization.

ACKNOWLEDGMENTS

This study was carried out with the support of R&D Program for Forest Science Technology (Project No. "RS-2023-KF002452613821165300003") provided by Korea Forest Service (Korea Forestry Promotion Institute).

Data Availability

The datasets used in this study, together with the model-training code, trained model files, and the Streamlit GUI, will be available at [https://github.com/Hyeon0024/GUI_exe.git]. Data not included in the repository are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

REFERENCE CITED

Albini, E., Long, J., Dervovic, D., and Magazzeni, D. (2022). "Counterfactual Shapley additive explanations," in: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, Seoul, Republic of Korea, pp. 1054-1070. https://doi.org/10.1145/3531146.3533168

Alibrahim, H., and Ludwig, S. A. (2021). "Hyperparameter optimization: Comparing genetic algorithm against grid search and Bayesian optimization," in: *2021 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1551-1559. https://doi.org/10.1109/CEC45853.2021.9504761

Auxenfans, T., Crônier, D., Chabbert, B., and Paës, G. (2017). "Understanding the structural and chemical changes of plant biomass following steam explosion pretreatment," *Biotechnology for Biofuels* 10(1), article 36. https://doi.org/10.1186/s13068-017-0718-z

Belbo, H., and Talbot, B. (2014). "Systems analysis of ten supply chains for whole tree chips," *Forests* 5(9), article 2105. https://doi.org/10.3390/f5092084

Biau, G., and Scornet, E. (2016). "A random forest guided tour," *TEST* 25(2), 197-227. https://doi.org/10.1007/s11749-016-0481-7

Bressi, V., Balu, A. M., Iannazzo, D., and Espro, C. (2023). "Recent advances in the synthesis of carbon dots from renewable biomass by high-efficient hydrothermal and microwave green approaches," *Current Opinion in Green and Sustainable Chemistry* 40, article 100742. https://doi.org/10.1016/j.cogsc.2022.100742

Brodeur, G., Yau, E., Badal, K., Collier, J., Ramachandran, K. B., and Ramakrishnan, S. (2011). "Chemical and physicochemical pretreatment of lignocellulosic biomass: A review," *Enzyme Research* 2011(1), article 787532. https://doi.org/10.4061/2011/787532

Cara, C., Ruiz, E., Ballesteros, I., Negro, M. J., and Castro, E. (2006). "Enhanced enzymatic hydrolysis of olive tree wood by steam explosion and alkaline peroxide delignification," *Process Biochemistry* 41(2), 423-429. https://doi.org/10.1016/j.procbio.2005.07.007

Chakraborty, P., Kumar, R., Chakrabortty, S., Saha, S., Chattaraj, S., Roy, S., Banerjee, A., Tripathy, S. K., Kumar Ghosh, A., and Jeon, B. (2024). "Technological advancements in the pretreatment of lignocellulosic biomass for effective valorization: A review of challenges and prospects," *Journal of Industrial and Engineering Chemistry* 137, 29-60. https://doi.org/10.1016/j.jiec.2024.03.025

Chen, X., Liu, Q., Li, B., Wang, N., Liu, C., Shi, J., and Liu, L. (2024). "Unveiling the potential of novel recyclable deep eutectic solvent pretreatment: Effective separation of lignin from poplar hydrolyzed residue," *International Journal of Biological Macromolecules* 259, article 129354. https://doi.org/10.1016/j.ijbiomac.2024.129354

Das, P., Stoffel, R. B., Area, M. C., and Ragauskas, A. J. (2019). "Effects of one-step alkaline and two-step alkaline/dilute acid and alkaline/steam explosion pretreatments on the structure of isolated pine lignin," *Biomass Bioenergy* 120, 350-358. https://doi.org/10.1016/j.biombioe.2018.11.029

David, A. J., Kannam, A. k., Sam, S., Narasimhan, M. K., and Krishnamurthi, T. (2025). "Optimization of delignification and organic acid synthesis from *Arachis hypogaea* L. shell: A comparison of machine learning approaches and experimental techniques," *Biomass Conversion and Biorefinery* 15(13), 20515-20526. https://doi.org/10.1007/s13399-025-06586-6

Djajadi, D.T., Jensen, M. M., Oliveira, M., Jensen, A., Thygesen, L. G., Pinelo, M., Glasius, M., Jørgensen, H., and Meyer, A. S. (2018). "Lignin from hydrothermally pretreated grass biomass retards enzymatic cellulose degradation by acting as a physical barrier rather than by inducing nonproductive adsorption of enzymes," *Biotechnology for Biofuels* 11(1), article 85. https://doi.org/10.1186/s13068-018-1085-0

Dossa, G. G. O., Paudel, E., Cao, K., Schaefer, D., and Harrison, R. D. (2016). "Factors controlling bark decomposition and its role in wood decomposition in five tropical tree species," *Scientific Reports* 6(1), article 34153. https://doi.org/10.1038/srep34153

Dziekońska-Kubczak, U., Berłowska, J., Dziugan, P., Patelski, P., Balcerek, M., Pielech-

Przybylska, K., Czyżowska, A., and Domański, J. (2018). "Comparison of steam explosion, dilute acid, and alkali pretreatments on enzymatic saccharification and fermentation of hardwood sawdust," *BioResources* 13(3), 6970-6984. https://doi.org/10.15376/biores.13.3.6970-6984

Fatima, S., Hussain, A., Amir, S. B., Awan, M. G. Z., Ahmed, S. H., and Aslam, S. M. H. (2023). "XGBoost and random forest algorithms: An in depth analysis," *Pakistan Journal of Scientific Research* 3(1), 26-31. https://doi.org/10.57041/vol3iss1pp26-31

Gao, F., Yang, F., Tao, Y., Wang, H., and Sun, Q. (2020). "Dilute alkali pretreatment and subsequent enzymatic hydrolysis of amur silvergrass for ethanol production," *Bioresources* 15(3), 4823-4834. https://doi.org/10.15376/biores.15.3.4823-4834

Gao, W., Zhou, L., Liu, S., Guan, Y., Gao, H., and Hui, B. (2022). "Machine learning prediction of lignin content in poplar with raman spectroscopy," *Bioresource Technology* 348, article 126812. https://doi.org/10.1016/j.biortech.2022.126812

Geurts, P., Ernst, D., and Wehenkel, L. (2006). "Extremely randomized trees," *Machine Learning* 63(1), 3-42. https://doi.org/10.1007/s10994-006-6226-1

Ha, S.Y., Jung, J. Y., Kim, H. C., Lim, W. S., and Yang, J. (2024a). "Low-temperature and low-concentration sodium hydroxide pretreatment for enhanced enzyme hydrolysis rate from *Quercus variabilis* blume," *BioResources* 19(2), 2592-2608. https://doi.org/10.15376/biores.19.2.2592-2608

Ha, S. Y., Kim, H. C., Lim, W. S., and Yang, J.-K. (2024b). "Enhanced enzymatic hydrolysis of waste wood by steam explosion and alkaline peroxide delignification," *Journal of Agriculture and Life Science* 58(4), 37-45. https://doi.org/10.14397/jals.2024.58.4.37

Ha, S.Y., Kim, H. C., Lim, W. S., and Yang, J. (2025). "Enhanced enzymatic hydrolysis of pine wood chips by two-step pretreatment combining steam explosion and urea," *BioResources* 20(3), 7293-7304. https://doi.org/10.15376/biores.20.3.7293-7304

Heikkinen, H., Elder, T., Maaheimo, H., Rovio, S., Rahikainen, J., Kruus, K., and Tamminen, T. (2014). "Impact of steam explosion on the wheat straw lignin structure studied by solution-state nuclear magnetic resonance and density functional methods," *Journal of Agricultural and Food Chemistry* 62(43), 10437-10444. https://doi.org/10.1021/jf504622j

Hörhammer, H., Dou, C., Gustafson, R., Suko, A., and Bura, R. (2018). "Removal of non-structural components from poplar whole-tree chips to enhance hydrolysis and fermentation performance," *Biotechnology for Biofuels* 11(1), article 222. https://doi.org/10.1186/s13068-018-1219-4

Inyang, V., Laseinde, O. T., and Kanakana, G. M. (2022). "Techniques and applications of lignocellulose biomass sources as transport fuels and other bioproducts," *International Journal of Low-Carbon Technologies* 17, 900-909. https://doi.org/10.1093/ijlct/ctac068

Jung, J.Y., and Yang, J. (2018). "A two-stage process for increasing the yield of prebiotic-rich extract from *Pinus densiflora*," *Journal of the Korean Wood Science and Technology* 46(4), 380-392. https://doi.org/10.5658/WOOD.2018.46.4.380

Karimi, K., Shafiei, M., and Kumar, R. (2013). "Progress in physical and chemical pretreatment of lignocellulosic biomass," in: *Biofuel Technologies: Recent Developments*, V. K. Gupta and M. G. Tuohy (eds.), Springer, Berlin, Germany, pp. 53-96. https://doi.org/10.1007/978-3-642-34519-7_3

Khorasani, M., Abdou, M., and Fernández, J. H. (2022). "Web application development with streamlit," *Software Development* 498, article 507.

Kim, H.C., Ha, S. Y., and Yang, J. (2025). "Artificial neural network approach for predicting enzymatic hydrolysis of steam exploded pine wood chip in mild alkaline pretreatment." *BioResources* 20(4), 8400-8419. https://doi.org/10.15376/biores.20.4.8400-8419

Kusiak, W.Ł, Majka, J., Zborowska, M., and Ratajczak, I. (2022). "Chemical composition and related properties of lime (*Tilia cordata* mill.) bark and wood as affected by tree growth conditions," *Materials* 15(11). https://doi.org/10.3390/ma15114033

Li, J., Henriksson, G., and Gellerstedt, G. (2007). "Lignin depolymerization/repolymerization and its critical role for delignification of aspen wood by steam explosion," *Bioresource Technology* 98(16), 3061-3068. https://doi.org/10.1016/j.biortech.2006.10.018

Liu, Q., Sun, C., Madadi, M., Zhang, E., Xia, A., Samimi, A., and Sun, F. (2025). "Elucidation of mixed surfactants reducing mass transfer resistance during high-solids enzymatic hydrolysis from lignocellulosic biomass," *Chemical Engineering Journal* 507, article 160332. https://doi.org/10.1016/j.cej.2025.160332

López, R., Fernández, C., Pereira, F. J., Díez, A., Cara, J., Martínez, O., and Sánchez, M. E. (2020). "A comparison between several response surface methodology designs and a neural network model to optimise the oxidation conditions of a lignocellulosic blend," *Biomolecules* 10(5), article 787. https://doi.org/10.3390/biom10050787

Mankar, A.R., Pandey, A., Modak, A., and Pant, K. K. (2021). "Pretreatment of lignocellulosic biomass: A review on recent advances," *Bioresource Technology* 334, article 125235. https://doi.org/10.1016/j.biortech.2021.125235

Matsakas, L., Nitsos, C., Raghavendran, V., Yakimenko, O., Persson, G., Olsson, E., Rova, U., Olsson, L., and Christakopoulos, P. (2018). "A novel hybrid organosolv: Steam explosion method for the efficient fractionation and pretreatment of birch biomass," *Biotechnology for Biofuels* 11(1), article 160. https://doi.org/10.1186/s13068-018-1163-3

Min, D.Y., Jameel, H., Chang, H. M., Lucia, L., Wang, Z. G., and Jin, Y. C. (2014). "The structural changes of lignin and lignin–carbohydrate complexes in corn stover induced by mild sodium hydroxide treatment," *RSC Advances* 4(21), 10845-10850. https://doi.org/10.1039/C3RA47032F

Mirahmadi, K., Kabir, M. M., Jeihanipour, A., Karimi, K., and Taherzadeh, M. J. (2010). "Alkaline pretreatment of spruce and birch to improve bioethanol and biogas production," *BioResources* 5(2), 928-938. https://doi.org/10.15376/biores.5.2.928-938

Nurek, T., Gendek, A., and Roman, K. (2019). "Forest residues as a renewable source of energy: Elemental composition and physical properties," *BioResources* 14(1), 6-20. https://doi.org/10.15376/biores.14.1.6-20

Okolie, J.A., Mukherjee, A., Nanda, S., Dalai, A. K., and Kozinski, J. A. (2021). "Next-generation biofuels and platform biochemicals from lignocellulosic biomass," *International Journal of Energy Research* 45(10), 14145-14169. https://doi.org/10.1002/er.6697

Onyutha, C. (2020). "From R-squared to coefficient of model accuracy for assessing goodness-of-fits," *Geoscientific Model Development Discussions* 2020, 1-25. https://doi.org/10.5194/gmd-2020-51

Ostertagová, E. (2012). "Modelling using polynomial regression," *Procedia Engineering* 48, 500-506. https://doi.org/10.1016/j.proeng.2012.09.545

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). "Scikit-learn: Machine

learning in python," *The Journal of Machine Learning Research* 12, 2825-2830.
Puss, K. K., Loog, M., and Salmar, S. (2023). "Ultrasound enhanced solubilization of forest biorefinery hydrolysis lignin in mild alkaline conditions," *Ultrasonics Sonochemistry* 93, article 106288. https://doi.org/10.1016/j.ultsonch.2022.106288
Rosdiana, N.A., Dumarçay, S., Gérardin, C., Chapuis, H., Santiago-Medina, F. J., Sari, R. K., Syafii, W., Gelhaye, E., Raharivelomanana, P., Mohammed, R., and Gérardin, P. (2017). "Characterization of bark extractives of different industrial Indonesian wood species for potential valorization," *Industrial Crops and Products* 108, 121-127. https://doi.org/10.1016/j.indcrop.2017.06.034
Song, G., Al Azad, S., Hu, W., Madadi, M., Rahman, A., Sun, C., and Sun, F. (2026). "Comparison study and mechanisms insight of $AlCl_3$-catalyzed different organosolv pretreatment of lignocellulose: Enhancing enzymatic hydrolysis, lignin fractionation, and furfural production," *Bioresource Technology* 439, article 133308. https://doi.org/10.1016/j.biortech.2025.133308
Steinbach, D., Kruse, A., Sauer, J. ö, and Storz, J. (2020). "Is steam explosion a promising pretreatment for acid hydrolysis of lignocellulosic biomass?" *Processes* 8(12), article 1626. https://doi.org/10.3390/pr8121626
Takahashi, N., and Koshijima, T. (1988). "Ester linkages between lignin and glucuronoxylan in a lignin-carbohydrate complex from beech (*Fagus crenata*) wood," *Wood Science and Technology* 22(3), 231-241. https://doi.org/10.1007/BF00386018
Tomas-Pejo, E., Oliva, J. M., and Ballesteros, M. (2008). "Realistic approach for full-scale bioethanol production from lignocellulose: A review," *Journal of Scientific and Industrial Research* 67(11), 874-884.
Varoquaux, G., and Colliot, O. (2023). "Evaluating machine learning models and their diagnostic value," in: *Machine Learning for Brain Disorders*, O. Colliot (ed.), Springer, New York, NY, USA, pp. 601-630. https://doi.org/10.1007/978-1-0716-3195-9_20
Wang, K., Jiang, J., Xu, F., and Sun, R. (2012). "Effects of incubation time on the fractionation and characterization of lignin during steam explosion pretreatment," *Industrial and Engineering Chemistry Research* 51(6), 2704-2713. https://doi.org/10.1021/ie2016009
Wang, W., Wang, X., Zhang, Y., Yu, Q., Tan, X., Zhuang, X., and Yuan, Z. (2020). "Effect of sodium hydroxide pretreatment on physicochemical changes and enzymatic hydrolysis of herbaceous and woody lignocelluloses," *Industrial Crops and Products* 145, article 112145. https://doi.org/10.1016/j.indcrop.2020.112145
Woźniak, A., Kuligowski, K., Świerczek, L., and Cenian, A. (2025). "Review of lignocellulosic biomass pretreatment using physical, thermal and chemical methods for higher yields in bioethanol production," *Sustainability* 17(1), article 287. https://doi.org/10.3390/su17010287
Xiao, C., Ye, J., Esteves, R. M., and Rong, C. (2016). "Using spearman's correlation coefficients for exploratory data analysis on big dataset," *Concurrency and Computation: Practice and Experience* 28(14), 3866-3878. https://doi.org/10.1002/cpe.3745
Xue, Y., Yu, C., and Kang, X. (2024). "Quantitative and structural characterization of native lignin in hardwood and softwood bark via solid-state NMR spectroscopy," *Journal of Agricultural and Food Chemistry* 72(32), 18056-18066. https://doi.org/10.1021/acs.jafc.4c03469
Yang, B., Boussaid, A., Mansfield, S. D., Gregg, D. J., and Saddler, J. N. (2002). "Fast

and efficient alkaline peroxide treatment to enhance the enzymatic digestibility of steam-exploded softwood substrates," *Biotechnology and Bioengineering* 77(6), 678-684. https://doi.org/10.1002/bit.10159

Yang, J., Sun, M., Jiao, L., and Dai, H. (2021). "Molecular weight distribution and dissolution behavior of lignin in alkaline solutions," *Polymers* 13(23), article 4166. https://doi.org/10.3390/polym13234166

Zhang, Y., Zhou, Q., Song, G., Madadi, M., Sun, C., Xu, J., Ashori, A., Sun, F., and Lu, X. (2025). "Low dosage ionic liquid-driven mild and selective lignocellulosic deconstruction of corn stover using biphasic co-solvent systems," *International Journal of Biological Macromolecules* 307, article 141810. https://doi.org/10.1016/j.ijbiomac.2025.141810

Zhao, X., Song, G., Cheng, Y., Quan, C., Chen, J., Hashemi, S. S., Yan, H., Sun, C., Madadi, M., and Sun, F. (2026). "Advanced glycerol organosolv pretreatment of sugarcane bagasse for ethanol and lignin-derived co-products with demonstrated environmental benefits," *Green Chemistry* 28(29), 12117-12132. https://doi.org/10.1039/d6gc01884j

Zhong, Z., Gorish, B. M. T., Bai, Y., Abdelmula, W. I. Y., Dang, W., and Zhu, D. (2025). "Predicting lignin removal efficiency in deep eutectic solvent-based biomass fractionation: An explainable machine learning approach," *Green Chemistry* 27(36), 11036-11054. https://doi.org/10.1039/D5GC02370J

Article submitted: April 6, 2026; Peer review completed: July 27, 2026; Revised version received and accepted: August 21, 2026; Published: September 12, 2026.
DOI: 10.15376/biores.21.4.10540-10559