Quantitative Analysis of UV-Visible Spectroscopy and Machine Learning Coupled for Alkali-Soluble Lignin in Steam-Exploded Biomass

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

Lignin is one of the most abundant biopolymers in lignocellulosic biomass, yet its efficient quantification remains a significant challenge for biorefineries due to the time-consuming nature and limitations of traditional wet-chemical analysis methods. This study aimed to develop a rapid and accurate approach for quantifying lignin concentration in alkali extracts of steam-exploded woody biomass by integrating UV-visible spectroscopy with machine learning algorithms as a practical complement to the conventional Klason lignin assay (modified ASTM D1106-56). UV-visible spectral data were collected and subjected to outlier removal using the Isolation Forest algorithm, followed by various preprocessing techniques and feature selection *via* the SelectKBest algorithm to optimize inputs for four regression models: Extra Trees, Random Forest, XGBoost, and Support Vector Regression. The combination of Baseline Correction and Standard Normal Variate (SNV) was the optimal preprocessing method, while the selection of the top 150 characteristic wavelengths effectively maximized information retention. Among the models evaluated, the Extra Trees (ET) regressor exhibited superior generalization capability and stability, achieving a test coefficient of determination (R^2) of 0.803 and a Mean Absolute Percentage Error (MAPE) of 4.0%, significantly outperforming SVR and XGBoost, which suffered from overfitting and underfitting, respectively.

DOI: 10.15376/biores.21.2.4436-4456

Keywords: Alkali-soluble lignin; UV-visible spectroscopy; Machine learning; Extra Trees regression; Steam explosion; Feature selection

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

The increasing demand for fossil fuel resources and growing environmental concerns have driven significant interest in utilizing lignocellulosic biomass as a renewable source for value-added chemicals, fuels, and materials (Isikgor and Becer 2015). Lignin, a natural phenolic polymer that constitutes 15 to 25% of the dry weight of lignocellulosic biomass, represents one of the most abundant biopolymers in the biosphere (Watkins *et al.* 2014). Despite being the most abundant aromatic compound in nature, lignin remains underutilized from a chemical perspective (Bourbiaux *et al.* 2021; Weiland *et al.* 2021). The accurate quantification of such biopolymers is therefore a critical prerequisite for their effective valorization, as demonstrated by recent advances in the characterization and functional application of other biomass-derived biopolymers, including agarose-based antibacterial films developed from algal sources (Ahmad *et al.* 2025)

Currently, substantial quantities of lignin are generated as byproducts in

various industries. The pulp and paper industry processes lignocellulosic biomass containing approximately 30% to 45% lignin, while the sugar industry handles biomass with 27 to 30% lignin content (Asim *et al.* 2020; Mateo *et al.* 2025). Similarly, biorefineries produce lignin as residual waste during the manufacture of value-added products, including furfural and ethanol (Qiao *et al.* 2022; Madadi *et al.* 2023). The pulp and paper industry alone generates 50 to 70 million tons of lignin annually, and the expansion of the biofuels industry is projected to contribute an additional 62 million tons per year (Tribot *et al.* 2019). Researchers estimate that global lignin production could reach 250 million tons by 2030, highlighting the critical need for effective lignin valorization strategies (Xu *et al.* 2019).

Lignin is a complex three-dimensional biopolymer composed primarily of phenylpropane derivatives. Based on methoxy group substitution patterns, lignin can be categorized into three primary structural units: syringyl (S) units, guaiacyl (G) units, and p-hydroxyphenyl (H) units (Creteanu *et al.* 2024). The depolymerization of lignin through hydrodeoxygenation yields various monomeric compounds with significant economic value. These lignin-derived monomers include methoxylated phenols (such as guaiacol and syringol), phenolic aldehydes and acids (including vanillin and vanillic acid), and hydroxycinnamic acid derivatives (such as p-coumaric acid) (Martínková *et al.* 2023; Kumari and Vinu 2025).

Lignin quantification methods can be broadly classified into direct and indirect approaches. Direct quantification methods, such as the Klason lignin method and the acid lithium bromide trihydrate (ALBTH) method, isolate and quantify lignin gravimetrically (Li *et al.* 2016; Lu *et al.* 2021). However, measuring the content of Klason lignin requires the use of highly concentrated acids (such as 72% sulfuric acid) and other chemical reagents, along with a lengthy process that typically requires 48 to 72 hours to complete. Experimental data may also be subject to error, depending on the skill of the experimenter.

To address these constraints, recent studies have actively applied machine learning (ML) to predict biomass composition, predominantly utilizing Near-Infrared or Fourier Transform Infrared spectroscopy on solid samples. While these solid-state methods are non-destructive, they often suffer from surface heterogeneity and light scattering, necessitating complex physical preprocessing. In contrast, indirect approaches using UV-visible spectroscopy on alkali extracts offer superior homogeneity and practicality. Although Lee *et al.* (2013) demonstrated the feasibility of using UV-visible spectroscopy for quantifying solubilized lignin, significant challenges exist due to spectral interference; for instance, lignin absorption bands overlap substantially with those of furfural, a degradation product of hemicellulose (Kline *et al.* 2010). Furthermore, hardwood biomass exhibits significant compositional variability depending on species and processing conditions, which directly affects the spectroscopic and chemical properties of the derived lignin fractions (Wijitkosum and Sriburi 2023). Consequently, univariate analyses used in previous studies were susceptible to these interferences. To mitigate such spectral interferences and physical variations, mathematical preprocessing techniques are indispensable. Methods such as baseline correction and Standard Normal Variate (SNV), are widely employed to effectively remove non-chemical signal variations caused by light scattering and instrumental drift (Yan 2025). Through enhancing spectral clarity and minimizing noise, these preprocessing steps play a crucial role in enabling advanced ML algorithms to robustly deconvolute interferences, thereby improving robustness and predictive accuracy (Bilak *et al.* 2025; Cui *et al.* 2025). Therefore, this study integrates UV-visible spectroscopy with advanced ML algorithms to achieve precise quantification.

Machine learning (ML) approaches have recently emerged as powerful tools

for the quantitative analysis of complex materials (Chan *et al.* 2022). Unlike traditional chemometric methods, ML algorithms can effectively model non-linear relationships between spectral data and chemical properties (Ouyang *et al.* 2019). The versatility of these data-driven approaches has been demonstrated across diverse biomass analytical platforms; for instance, ML models integrating remote sensing data have been successfully applied to estimate aboveground biomass and carbon stocks in forest ecosystems (Pungpa *et al.* 2025). Among various algorithms, ensemble learning methods, such as Random Forest and Extra Trees, and gradient boosting techniques have demonstrated exceptional capability in handling high-dimensional spectroscopic data (Jafarzadeh *et al.* 2021). These models are particularly robust against noise and collinearity, which are common challenges in UV-visible spectral analysis of complex biomass hydrolysates (Yue *et al.* 2018).

In this study, a rapid and accurate approach for quantifying lignin concentration in alkali extracts of steam-exploded woody biomass was developed by combining UV/visible spectroscopic analysis with spectral preprocessing and machine learning algorithms. UV/vis spectral data was collected to predict alkali-soluble lignin content. To enhance prediction accuracy while minimizing noise interference, spectral noise was reduced using preprocessing techniques (such as baseline correction and SNV), and characteristic wavelengths were identified using the SelectKBest algorithm. Finally, four regression models—Extra Trees (ET), Random Forest (RF), XGBoost, and Support Vector Regression (SVR)—were evaluated to determine the optimal method for predicting alkali-soluble lignin content.

EXPERIMENTAL

Materials

For steam explosion, the oak and pine mixed samples were dried and stored at a particle size of approximately 2.5 cm (W) × 2.5 cm (L) × 0.5 cm (H). Then, 10 kg of biomass was placed in a 100-L batch reactor (Yulim Hightech, Daegu, Korea) and steam exploded at 225 °C for 1, 3, and 5 min. The steam exploded samples (SES) were refrigerated at 4 ℃ in zip-lock bags until analysis.

Extraction Using an Alkaline Solution

Lignin extraction was performed using 1% to 2% alkaline (NaOH, KOH) solutions, where 5 g of the steam explosion sample was placed in a conical flask with extraction solution (100 mL). The flasks were placed at room temperature, capped with aluminum and extracted for 12 h to 24 h. The extracted solution was filtered using grade 2 Whatman filter paper. The filtration was used for UV-visible analysis, and the residual solids were used to determine the residual lignin content after alkaline extraction.

Analysis of Alkali-soluble Lignin

The quantitative analysis of alkali-extracted lignin was performed using a modified version of the ASTM D1106-56 (1979) standard method. Specifically, the 300 mg sample used for hydrolysis refers to dried alkali-extracted solids or raw biomass. The amount of lignin solubilized during the alkali treatment was calculated by measuring the lignin content of both the raw biomass and the post-extraction residues. In this two-step method, 300 mg of lignin extracted in step 1 was reacted with 3 mL of 72% H_2SO_4 solution at 30 °C for 1 h. In step 2, the H_2SO_4 solution was diluted to 12% using distilled water and autoclaved at 121 °C for 1 h. The hydrolyzed sample was cooled and filtered through Whatman filter paper (No. 1573). The samples were

thoroughly washed with hot distilled water until a neutral pH was reached and then dried at 105 °C to a constant weight. The total amount of lignin was calculated as acid-insoluble lignin (Klason lignin). Ash correction was not included.

The concentration of alkali-soluble lignin was determined based on the mass balance of acid-insoluble lignin before and after extraction. Specifically, the mass of lignin solubilized into the liquid phase was calculated by subtracting the mass of residual lignin in the solid residue from the total initial lignin mass in the raw biomass. The final concentration in mg/mL was obtained by dividing this solubilized mass by the volume of the extraction solvent, using Eq. 1,

$$Alkali - solublelignin(mg/mL) = \frac{[(W_{raw} \times L_{raw}) - (W_{res} \times L_{res})] \times 10}{V} \quad (1)$$

where W_{raw} and W_{res} are the dry weights (g) of the raw biomass and the alkali-extracted residue, respectively. L_{raw} and L_{res} denote the acid-insoluble lignin content (%) of the raw biomass and the residue, determined by the Klason method. V is the volume of the alkaline solution used for extraction (mL). The factor 10 is used to convert the mass unit from grams to milligrams while accounting for the percentage.

UV-Visible Analysis of Alkaline Extracts

UV-visible absorbance spectra of the alkali lignin extracts were acquired using a microplate reader (SpectraMax 190, Molecular Devices, USA). Spectroscopic measurements were conducted in 96-well microplates with a sample volume of 200 µL per well. Spectra were recorded in the wavelength range of 200 to 800 nm with a 1 nm scanning interval. To correct for solvent absorption, fresh alkaline solutions corresponding to the extraction solvent (*e.g.*, 1% or 2% NaOH) were used as blanks for baseline correction. This ensured that the obtained spectra represented only the solubilized lignin. Each of the 72 extract samples was scanned six times, resulting in a comprehensive dataset of 432 UV-visible spectra. Each replicate measurement involved an independent sample injection into the microplate well, which introduced minor but measurable variation in absorbance arising from pipetting differences, subtle concentration gradients, and instrumental baseline fluctuations between scans (Johnson *et al.* 2023; Udo *et al.* 2024). Therefore, the six replicate spectra obtained from the same extract were not strictly identical, reflecting realistic measurement variability inherent to the analytical procedure.

Preprocessing of UV-Visible Spectra

Prior to mathematical preprocessing, the Isolation Forest algorithm was employed to identify and remove spectral outliers from the training spectral matrix (Song *et al.* 2021). By setting the contamination parameter to 0.05, approximately 5% of the training spectra exhibiting anomalous profiles were identified and excluded. It should be noted that outlier detection was performed based solely on the multivariate spectral feature space, not on the reference lignin concentration values. Visual inspection of the removed spectrum revealed abnormal baseline fluctuations and sudden intensity spikes that did not match the overall spectral pattern (Fig. A1). The reference lignin concentrations of the removed samples were subsequently verified and confirmed to fall within the normal range of the dataset, indicating that the detected anomalies originated from spectroscopic measurement errors rather than meaningful chemical or process variability. Therefore, the removal of these outliers was necessary to improve model training accuracy and ensure data integrity.

Following outlier removal, various preprocessing techniques were evaluated to enhance spectral features and minimize noise. These included derivatives, Standard

Normal Variate (SNV), and baseline correction (Grisanti *et al.* 2018; Li *et al.* 2020). The first and second derivatives were calculated using the Savitzky-Golay algorithm to resolve overlapping peaks and remove baseline offsets (Zimmermann and Kohler 2013). The SNV was applied to correct for light scattering effects by normalizing each spectrum to zero mean and unit variance. Baseline correction was performed using asymmetric least squares smoothing (parameter lambda = 10^5) to remove baseline drift while preserving chemical features.

Spectral Feature Selection using SelectKBest

Following spectral preprocessing, feature selection was performed using the SelectKBest algorithm from the scikit-learn package (Abraham *et al.* 2014) to identify the most informative wavelengths for predicting lignin content. SelectKBest ranks features (wavelengths) based on their correlation with the target variable (lignin content) using univariate statistical tests and retains only the top k features with the highest scores. The SelectKBest algorithm was selected for its distinct advantages in spectroscopic data analysis. It offers computational efficiency by evaluating each feature independently, making it particularly suitable for high-dimensional spectral datasets containing hundreds of wavelength variables. Second, it enhances robustness against overfitting by reducing model complexity through the elimination of redundant and irrelevant wavelengths. Third, SelectKBest facilitates model interpretability by identifying specific absorption bands that are most strongly associated with lignin content, thereby providing chemical insights into the chromophore structures responsible for UV-visible absorption. Feature selection was performed exclusively on the training set prior to the cross-validation process, ensuring that the test set data had no influence on the feature selection procedure. The selected feature subset was subsequently applied to the test set without refitting, thereby preventing any form of data leakage from the test set into the model development pipeline.

Following Eq. 2, the F-regression score was employed as the scoring function, which measures the linear dependency between each wavelength variable and the lignin content *via* analysis of variance (Ayikpa *et al.* 2025),

$$F = \frac{SS_{regression}/df_{regression}}{SS_{residual}/df_{residual}} \quad (2)$$

where SS represents the sum of squares and df denotes the degrees of freedom.

Development and Comparative Analysis of Machine Learning Models

To predict the alkali-extracted lignin content from the preprocessed UV-visible spectra, four machine learning regression algorithms were evaluated: Extra Trees (ET), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Support Vector Regression (SVR). These algorithms represent diverse approaches for capturing the non-linear relationships between spectral features and lignin content. The Extra Trees algorithm is an ensemble learning method that constructs multiple decision trees using random node splitting criteria (Galelli and Castelletti 2013). Unlike Random Forest, it introduces additional randomness by selecting split thresholds randomly rather than optimizing them. This approach reduces variance and computational costs, with the final prediction obtained by averaging the predictions of all trees. Random Forest utilizes bootstrap aggregation (bagging) of decision trees trained on random subsets of samples and features (Lee *et al.* 2019)). This dual randomization strategy effectively mitigates overfitting while maintaining high predictive accuracy for the complex non-linear relationships typical in spectroscopic data. XGBoost is an optimized gradient boosting framework that sequentially builds decision trees to correct the residual errors

of previous iterations (Bentéjac *et al.* 2019). It minimizes a regularized objective function that simultaneously considers a loss function and a regularization term to control model complexity. Support Vector Regression establishes a linear regression in a transformed space by mapping input features into a high-dimensional space using a kernel function (Xie *et al.* 2018). In this study, the Radial Basis Function (RBF) kernel was employed, as it is effective in capturing non-linear patterns in spectroscopic data by calculating the similarity between samples using kernel coefficients and Euclidean distance.

Model Training and Evaluation

The total dataset of 432 spectra was randomly partitioned into a training set (80%) and a test set (20%) using stratified sampling to ensure a representative distribution of the lignin content range in both subsets. Hyperparameter optimization for all models was performed *via* 5-fold cross-validation on the training set using RandomizedSearchCV, evaluating 100 random parameter combinations for each algorithm. Model performance was assessed using statistical metrics calculated for both the cross-validation and independent test sets. The coefficient of determination (R^2) quantifies the proportion of variance in lignin content explained by the model, where a value closer to 1 indicates superior explanatory power, following Eq. 3,

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

The Mean Absolute Percentage Error (MAPE) evaluates relative prediction accuracy by expressing prediction errors as a percentage of the observed values, making it useful for assessing consistent model performance regardless of the absolute lignin content (Eq. 4),

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \quad (4)$$

The Root Mean Square Error (RMSE) is the square root of the sum of the squares of the errors between the predicted and actual values. It can be used to assess the accuracy of a model by comparing the difference between the model's predicted lignin content and the actual measured value (Eq. 5),

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}i)^2} \quad (5)$$

where y_i and $\hat{y}_i$ represent the observed and predicted lignin contents, respectively, $\hat{y}_i$ is the mean of the observed values, and n is the number of samples. The optimal model was selected based on the combination of the highest R^2 and the lowest RMSE, MAE, and MAPE on the independent test set. All machine learning analyses were implemented in Python 3.8 using the scikit-learn (v1.0.2), XGBoost (v1.5.0), and NumPy (v1.21.2) libraries.

RESULTS AND DISCUSSION

Sample Partitioning Results and Descriptive Statistics

A total of 432 spectral data samples were randomly partitioned into a training set (n = 345, 80%) and a test set (n = 87, 20%) using stratified sampling to ensure equal representation of lignin content distributions. Ash correction was not applied in this study. Previous studies have reported that the ash content of steam-exploded oak–pine mixed biomass is generally low (<2%) (Chandrasekaran *et al.* 2012; Wang *et al.* 2011).

Nevertheless, the absence of ash correction represents a minor limitation that may slightly overestimate Klason lignin values.

As summarized in Table 1, the average lignin content was found to be approximately 0.33 mg/mL for both the training and test sets, with similar standard deviations of 0.054 and 0.048, respectively. Because the range of lignin content in the training set fully encompasses that of the test set, it confirms that the data splitting strategy was appropriate. This statistical similarity ensures that the training set serves as a reliable and representative basis for model development, minimizing potential bias during the learning process.

Table 1. Basic Statistics of Lignin Content in the Training and Test Datasets

	Lignin Content	
Basic statistics	Train data	Test data
No. of samples	345	87
Content range (mg/mL)	0.235 to 0.537	0.241 to 0.537
Average (mg/mL)	0.3305	0.3328
Standard deviation (mg/mL)	0.054	0.048

Evaluation of Preprocessing Methods

Prior to applying any mathematical preprocessing, the integrity of the raw spectral dataset was first ensured using the Isolation Forest algorithm. Through applying this unsupervised learning method with a contamination parameter of 0.05, 18 data exhibiting anomalous spectral profiles were identified as outliers within the training set and were subsequently excluded. The refined dataset was then utilized for model training. Removing these outliers at the initial raw data stage is critical because extreme values caused by instrumental errors or sample heterogeneity can distort the global statistical parameters (such as mean and variance) used in subsequent normalization steps such as SNV.

The visual impact of the preprocessing techniques on the spectral features is illustrated in Fig. 1. The raw spectra (Fig. 1A) exhibited significant baseline variations and scattering effects, which obscured the chemical information. However, the application of Baseline Correction followed by Standard Normal Variate (SNV) (Fig. 1D) effectively aligned the spectra by correcting the baseline drift and normalizing the intensity variations. This transformation revealed distinct absorption features that were previously hidden by physical noise.

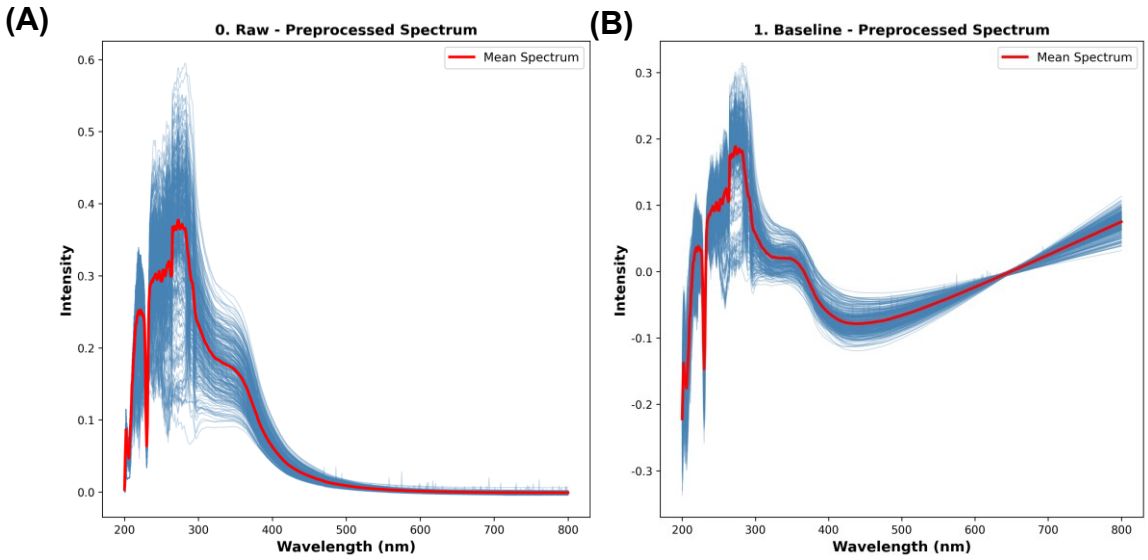
To quantitatively validate these visual improvements, four different preprocessing methods were evaluated using the Extra Trees (ET) regressor (Table 2). The Baseline + SNV combination achieved the highest predictive accuracy with a test R^2 of 0.74, outperforming the raw spectra (R^2 = 0.73). More importantly, the stability of the model was confirmed by analyzing the performance gap between cross-validation (CV) and the independent test set. While the raw data model showed a notable discrepancy between CV and test scores, the Baseline + SNV method demonstrated a significantly narrower gap. This minimized difference indicates enhanced model generalization and robustness, ensuring that the model is not biased towards specific data subsets. Based on its superior accuracy and stability, the Baseline + SNV method was selected as the optimal preprocessing step for the subsequent feature selection.

Table 2. Comparison of Prediction Performance (R^2 and MAPE) for Lignin Quantification Using the ET Model with Various Spectral Preprocessing Methods

Pretreatment Methods	Arity (K)	Train Data		Cross Validation		Test Data	
		MAPE	R^2	MAPE	R^2	MAPE	R^2
RAW	601	0.77	0.98	6.15	0.54	4.75	0.73
Bas*	601	0.71	0.98	6.48	0.48	6.89	0.55
Bas+ 1st*	601	0.69	0.99	6.48	0.50	5.78	0.58
Bas + 2nd*	601	0.59	0.99	7.08	0.55	6.59	0.60
Bas + SNV*	601	0.68	0.99	6.01	0.66	5.60	0.74

*Bas: baseline pretreatment
*Bas + 1st: baseline + first derivative
*Bas + 2st: baseline + second derivative
*Bas + SNV: baseline + standard normal variate

Previous studies have consistently demonstrated that mathematical preprocessing is essential for spectroscopic analysis of heterogeneous solid samples, such as lignocellulosic biomass, to mitigate physical interferences including light scattering and path length variations. Rinnan *et al.* (2009) highlighted in their comprehensive review that preprocessing techniques like SNV and baseline correction are critical for improving the predictive performance of multivariate calibration models. Specifically, Barnes *et al.* (1989), who introduced the SNV transformation, established that this method effectively removes the multiplicative interferences caused by particle size differences and scattering, thereby isolating the chemical absorption features.



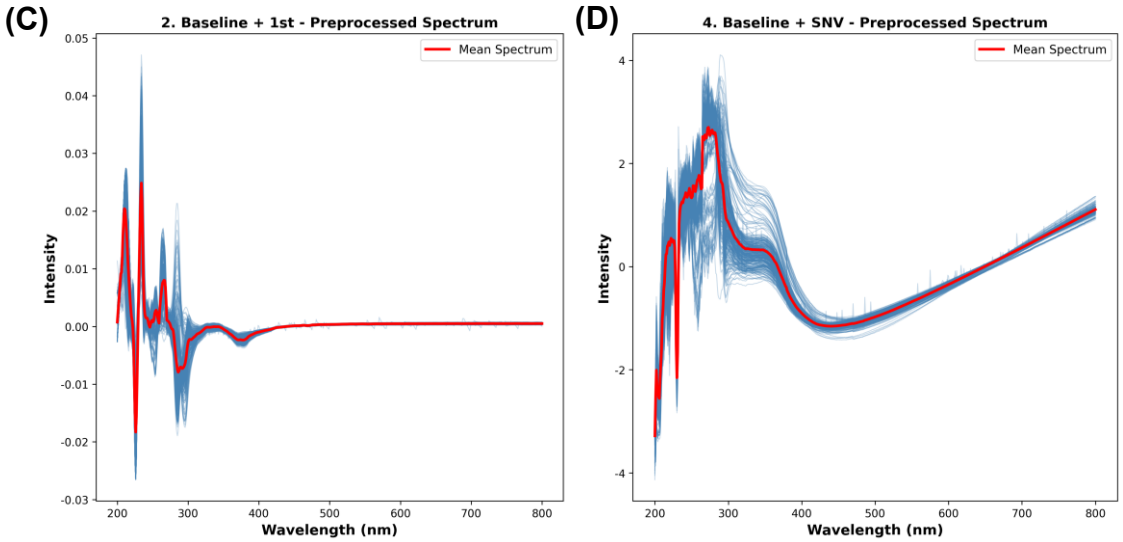


Fig. 1. UV-visible absorption spectra of alkali-extracted lignin: (A) raw spectra, (B) baseline-corrected spectra, (C) first derivative spectra, and (D) spectra with baseline correction and SNV

Feature Selection using SelectKBest

Following the optimization of the preprocessing method, feature selection was performed using the SelectKBest algorithm to identify the most informative wavelengths for lignin quantification. This step is crucial for reducing model complexity and eliminating redundant spectral variables that do not contribute to predictive accuracy. To determine the optimal feature subset, the ET model was trained with varying numbers of selected features (K), ranging from 50 to 300. Table 3 summarizes the variations in model performance metrics (R^2 and MAPE) corresponding to the number of features.

The analysis revealed a distinct optimal point at K = 150, where the model achieved its highest predictive accuracy with a test R^2 of 0.80 and the lowest MAPE of 4.01%. As illustrated in Fig. 2(A), these selected wavelengths were primarily concentrated in specific regions of the UV-visible spectrum, likely corresponding to the characteristic absorption bands of lignin-derived compounds. The lignin component of wood is generally known to absorb UV light at short wavelengths in the range of 295 to 400 nm. The selected wavelengths were confirmed to include the standard quantification wavelength range for alkali lignin (256 to 281 nm, centered at 280 nm), which has been reported to exhibit strong absorbance for alkali-solubilized lignin (Lee *et al.* 2013). The 200 to 230 nm wavelength region is presumed to be attributable to the n→π* transitions of phenolic groups, as the aromatic rings, methoxy groups, and conjugated double bonds characteristic of lignin structure have been reported to contribute to ultraviolet absorption in this region. Furthermore, the 320 to 400 nm wavelength region is suggested to be influenced by the abundant phenolic hydroxyl groups (-OH), methoxy groups, carbonyl groups, and other functional groups present in lignin (Duy *et al.* 2024). Wavelengths above 400 nm fall within the visible light region, where the dark brown coloration of lignin extracts is a characteristic feature. This coloration is primarily governed by selective absorption in the violet range (407 to 423 nm) and the orange-to-red range (630 to 705 nm), which are considered to be among the key determinants of lignin color and are therefore presumed to contribute to the selected wavelength features (Cheng *et al.* 2013). As lignin is a complex molecular assembly possessing diverse functional groups and chromophores that exist in varying proportions depending on the extraction method and raw material, the precise

mechanisms governing its chromogenic properties have not yet been fully elucidated (Ajao *et al.* 2018).

When fewer features (K = 50) were retained, the model performance dropped significantly (R^2 = 0.76), indicating that essential spectral information required for accurate quantification was lost (underfitting).

Conversely, increasing the number of features beyond 150 (*e.g.*, K = 200, 250, or 300) resulted in a gradual decline in performance, with test R^2 values decreasing to the 0.77 to 0.78 range and MAPE increasing to over 5%. This degradation suggests that including excessive wavelengths introduced irrelevant noise or collinear variables, which hindered the model's generalization capability. Therefore, the top 150 wavelengths were identified as the optimal feature set, effectively balancing information retention and noise reduction. This optimized subset was subsequently used for the final comparative analysis of machine learning models.

These findings are consistent with previous studies on spectroscopic analysis of lignocellulosic biomass, which emphasize the importance of wavelength selection. For instance, Li *et al.* (2022) demonstrated that identifying and utilizing specific absorption bands related to lignin chromophores, rather than the entire spectral range, significantly enhances model interpretability and predictive accuracy. Through eliminating redundant variables that contribute to multicollinearity, feature selection mitigates the 'curse of dimensionality' and prevents the model from fitting to noise (Rajendran *et al.* 2022). Consequently, the observation that the K = 150 subset outperformed the full spectrum supports the conclusion that optimal feature selection is a critical step for developing robust machine learning models with high generalization capability (Fig. 2(B)).

However, the prediction scatter observed in the test data plot of Fig. 2(B) may be partially attributable to the presence of pseudo-lignin generated during steam explosion pretreatment at 225°C for treatment durations of 1, 3, and 5 minutes, with pseudo-lignin formation expected to intensify with increasing treatment duration. As reported by Aarum *et al.* (2019), under the high temperature and pressure conditions of steam explosion, hemicellulose and cellulose undergo degradation, generating byproducts such as hydroxy-methyl-furfural (HMF) and furfural (FF). These species can subsequently undergo condensation reactions to form pseudo-lignin. As pseudo-lignin is acid-insoluble, it co-precipitates with native lignin during Klason analysis, potentially leading to overestimation of the reference lignin values. Furthermore, the UV-vis absorbing chromophores of pseudo lignin overlap with those of native alkali-soluble lignin, potentially introducing systematic spectral interference into the absorbance profiles of the alkali-extracted fractions (Bikova and Treimanis 2004). These characteristics are a contributing factor to the prediction variability observed in the current dataset, and represent an important limitation to be addressed in future studies.

Table 3. Variation in Predictive Performance (RMSE, MAPE and R^2) of the ET Model According to the Number of Selected Spectral Features (K)

Pretreatment Methods	Arity (K)	Train Data			Cross Validation			Test Data		
		RMSE	MAPE	R^2	RMSE	MAPE	R^2	RMSE	MAPE	R^2
Bas + SNV	50	0.006	1.121	0.981	0.026	5.47	0.684	0.024	5.14	0.761
	100	0.006	0.932	0.983	0.028	5.42	0.713	0.021	4.98	0.802
	150	0.005	0.803	0.991	0.026	5.31	0.758	0.021	5.03	0.803
	200	0.005	0.762	0.992	0.027	5.34	0.740	0.022	5.18	0.782
	250	0.005	0.735	0.991	0.028	5.63	0.702	0.023	5.38	0.771
	300	0.005	0.721	0.991	0.028	5.66	0.702	0.023	5.36	0.771

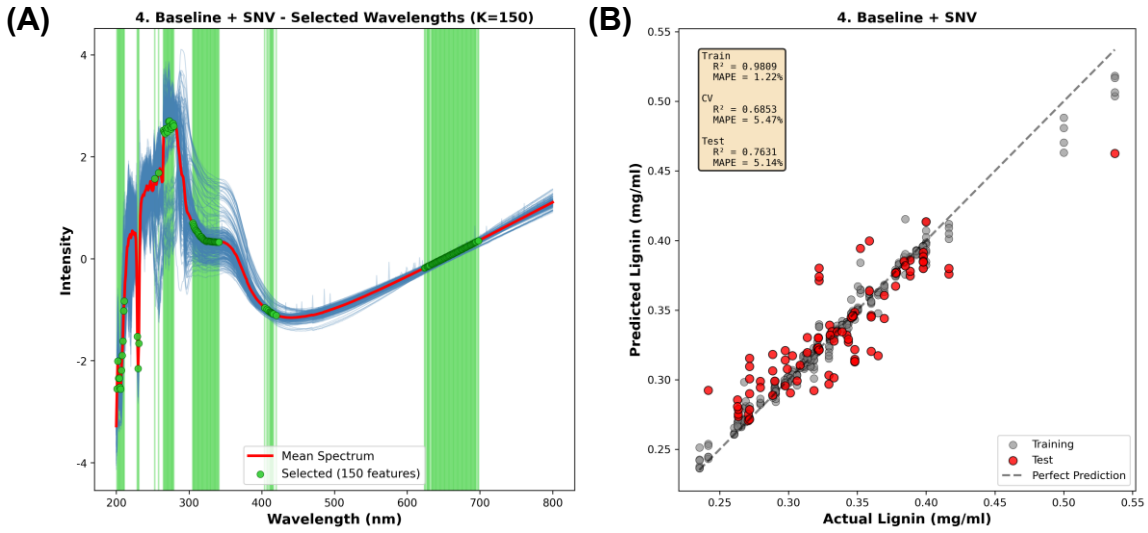


Fig. 2. Visualization of (A) characteristic spectral features identified for lignin quantification using the F-regression score (K = 150) and (B) the corresponding prediction accuracy of the ET regressor for training, cross-validation, and test datasets

Comparison of Machine Learning Models

To identify the optimal regression strategy for lignin quantification, four distinct machine learning algorithms—Extra Trees (ET), Random Forest (RF), XGBoost (XGB), and Support Vector Regression (SVR)—were rigorously evaluated using the preprocessed spectral dataset (K=150, Baseline + SNV). To further investigate model-dependent preprocessing effects, the performance of all four models was compared between the Raw and Baseline + SNV preprocessed datasets (Table 4, Fig. 4). The results demonstrated that Baseline + SNV preprocessing consistently improved predictive performance across all models compared to Raw data, confirming that the selected spectral preprocessing approach is not specific to a single algorithm but provides universal performance enhancement regardless of the machine learning method employed. Among the models evaluated, the ET regressor exhibited the highest generalization capability under the Baseline + SNV condition, achieving a test R^2 of 0.803, RMSE of 0.021 mg/mL, and MAPE of 5.01%. The performance improvement from Raw to Baseline + SNV was most pronounced for the ET model, where test R^2 increased from 0.754 to 0.803, indicating that the ET algorithm benefited most from the enhanced spectral quality provided by the combined preprocessing approach. Notably, the ET model also demonstrated strong consistency between cross-validation and test set performance under Baseline + SNV (CV R^2 = 0.757 vs. test R^2 = 0.803, CV

RMSE = 0.020 vs. test RMSE = 0.021 mg/mL), suggesting reliable generalization without significant overfitting.

While the ET model demonstrated superior performance, it should be noted that the performance differences between models may partly reflect the hyperparameter search space rather than inherent algorithmic limitations alone. As detailed in Table A1 (Appendix), hyperparameter optimization was conducted *via* RandomizedSearchCV across predefined search ranges for all four models. The optimal hyperparameters revealed distinct characteristics for each model: the ET model achieved the best generalization with max_depth of 12 and max_features of 'sqrt', while the RF model showed a relatively low optimal n_estimators of 67, suggesting that further ensemble diversity may not have been fully explored. The relatively shallow optimal max_depth of XGBoost (3) suggests that the model struggled to capture the complexity of the spectral data within the given search space, while the SVR's optimal C value (16.05) indicates moderate regularization strength.

This finding corroborates previous chemometric studies reported by Geurts *et al.* (2006) and recent applications in biomass analysis, which suggest that extremely randomized trees offer superior resistance to noise and multicollinearity compared to standard bagging or boosting methods (Zhang *et al.* 2020; Wang *et al.* 2023). Notably, the predictive reliability achieved in this study outperformed many conventional solid-state spectroscopic approaches. While previous studies utilizing FTIR or NIR spectroscopy often struggle with surface heterogeneity and light scattering (Zhang *et al.* 2020), recent reviews emphasize that liquid-state analysis, when combined with appropriate baseline correction, offers superior homogeneity, thereby enhancing the robustness of machine learning predictions (Yan 2025).

Furthermore, this study distinguishes itself by specifically targeting the alkali-soluble lignin fraction derived from steam-exploded biomass. This fraction represents the high-value "technical lignin" directly available for downstream valorization (Mateo *et al.* 2025). Unlike prior UV-vis applications that relied on simple univariate regression and were susceptible to interference from degradation byproducts such as furfural (Kline *et al.* 2010), the current ML-integrated approach successfully deconvolved these spectral overlaps. The Random Forest (RF) model also displayed satisfactory predictive potential, ranking second with a test R^2 of 0.724 and MAPE of 6.023% under the Baseline + SNV condition. Conversely, the boosting and kernel-based algorithms failed to provide reliable predictions for this specific dataset. The XGBoost model exhibited signs of underfitting, characterized by a low training R^2 of 0.565 and poor test performance (R^2 = 0.417; MAPE = 8.874%), indicating an inability to adequately capture the complex non-linear dependencies present in the spectral data. Furthermore, the SVR model suffered from severe overfitting despite a respectable training accuracy (R^2 = 0.794), as its predictive power collapsed on the test set (R^2 = 0.437; MAPE = 8.833%). This substantial gap between training and test performance implies that the SVR model likely memorized noise within the training data rather than learning the underlying chemical features. Consequently, the Extra Trees algorithm was established as the most robust and accurate tool for the rapid quantification of alkali-soluble lignin in this study.

However, limitations exist in generalizing the model. The present model was developed exclusively on steam-exploded oak–pine mixed biomass, which restricts direct generalization to other feedstock types. Agricultural residues and pure hardwood or softwood species differ in lignin monomer composition and spectral characteristics and would therefore require dedicated training datasets. Future studies should explore transfer learning or domain adaptation strategies to extend the applicability of this framework to a broader range of biomass feedstocks.

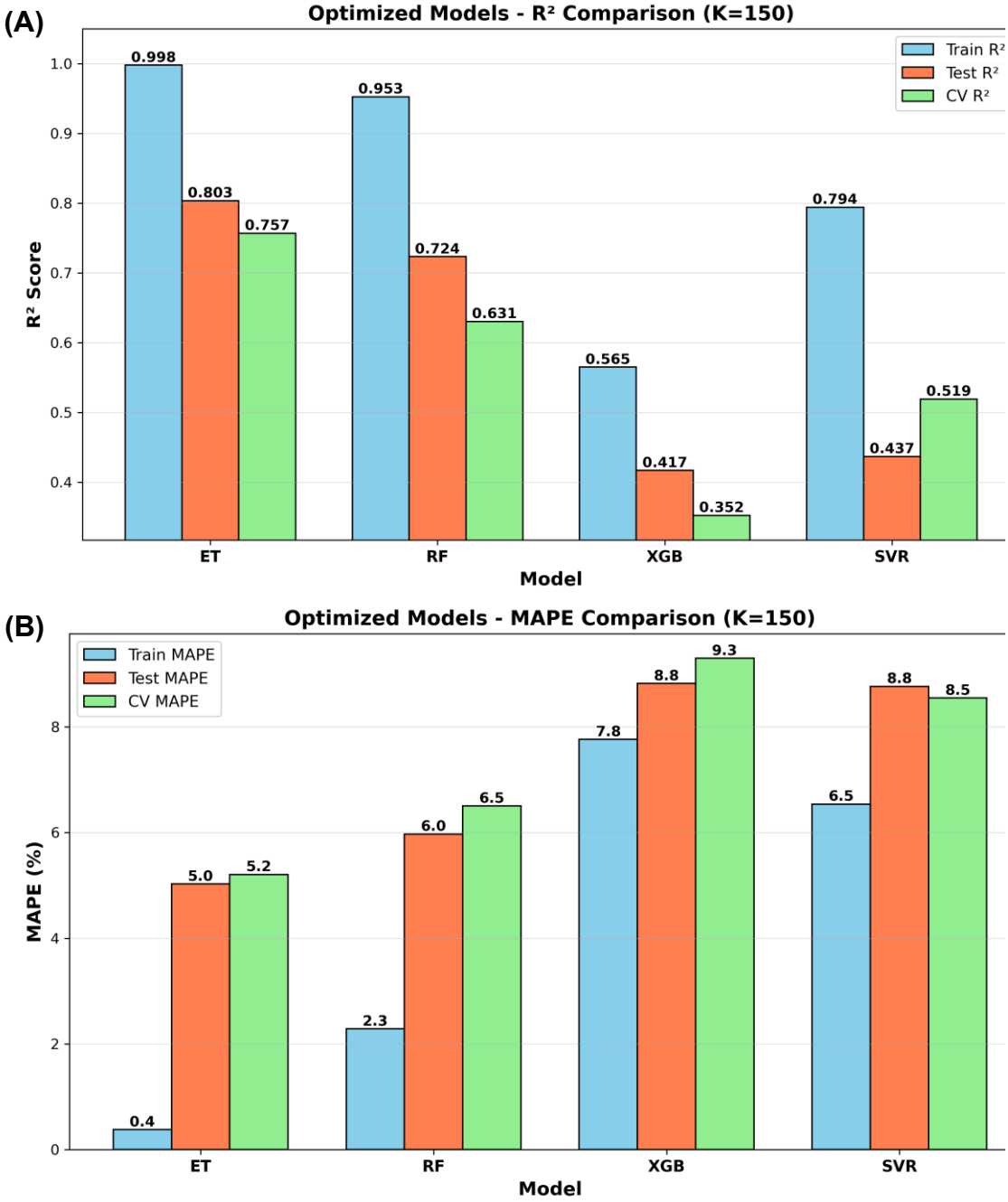


Fig. 3. Performance comparison of four machine learning models (ET, RF, XGB, and SVR) trained with the top 150 spectral features. The metrics shown are R^2 (A) and MAPE (B) for training, cross-validation, and testing datasets.

Table 4. Comparison of Prediction Performance (RMSE, MAPE, and R^2) for Lignin Quantification across Four Machine Learning Models with and without Baseline + SNV Preprocessing

Pretreatment Methods	Model	Train Data			Cross Validation			Test Data		
		RMSE	MAPE	R^2	RMSE	MAPE	R^2	RMSE	MAPE	R^2
RAW	ET	0.002	*4.224*	0.998	0.031	*4.860*	0.631	0.024	*4.132*	0.754
	RF	0.014	*5.115*	0.926	0.036	*5.652*	0.505	0.029	*5.318*	0.646
	XGB	0.007	*6.823*	0.424	*0.036*	*7.245*	0.423	0.033	*6.596*	0.531
	SVR	0.014	*6.148*	0.923	*0.032*	*5.456*	0.692	0.033	*6.753*	0.515
Bas + SNV	ET	0.002	0.430	0.998	0.02	5.259	0.757	0.021	5.010	0.803
	RF	0.004	2.323	0.953	*0.025*	6.542	0.631	*0.026*	6.023	0.724
	XGB	*0.028*	7.832	0.565	*0.038*	9.323	0.352	*0.033*	8.874	0.417
	SVR	*0.022*	6.565	0.794	*0.03*	8.521	0.519	*0.03*	8.833	0.437

In addition, it should be noted that the developed framework has been validated only at the laboratory scale, which represents an inherent limitation with respect to direct industrial deployment. Transitioning to an industrial scale would necessitate additional model training using data acquired from UV-vis sensors deployed under actual process conditions. Furthermore, various environmental variables encountered in industrial settings — including optical surface fouling, temperature fluctuations, and entrained particulates — may introduce spectral distortions not represented in the current calibration dataset, and therefore further model validation and recalibration under industrially relevant conditions will be required in future studies

CONCLUSIONS

1. This study successfully demonstrated the potential of combining UV-visible spectroscopy with machine learning algorithms to quantify alkali-soluble lignin in steam-exploded biomass. The proposed framework offers a rapid, non-destructive, and chemical-free alternative to traditional gravimetric analysis, addressing the time-consuming nature of wet-chemical methods.
2. Effective data processing strategies were critical for model accuracy. The application of the Isolation Forest algorithm successfully identified and removed spectral outliers (approx. 5%) from the raw data. Subsequently, the combination of Baseline Correction and Standard Normal Variate (SNV) was identified as the optimal preprocessing method, significantly enhancing signal quality by mitigating light scattering and instrumental drift.
3. The Extra Trees (ET) regressor, which was trained using the top 150 spectral features selected by the SelectKBest algorithm, demonstrated superior predictive performance, remarkable stability, and resistance to overfitting. This performance was superior to that of Random Forest, XGBoost, and SVR. Under the optimal baseline + SNV preprocessing condition, the ET model achieved a test R^2 of 0.803, an RMSE of 0.021 mg/mL, and a MAPE of 5.01%. There was strong concordance between the cross-validation and test set metrics (CV R^2 = 0.757 and CV RMSE = 0.020 mg/mL), which confirms its consistent generalization capability. This level of precision is sufficient for real-time component monitoring in lab-scale processes.

4. However, the current MAPE of ~5% may not meet the more stringent accuracy thresholds required for commercial transactions or formal product grading. Therefore, independent external validation is recommended prior to such applications. To maintain long-term reliability across varying biomass feedstocks and processing conditions, periodic recalibration against the Klason lignin wet-chemical reference method or equivalent ASTM-standardized procedures is also advised.

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

All datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Use of Generative AI (If Applicable)

The authors acknowledge the use of AI assistance during manuscript preparation. This includes English translation, formatting of the Experimental section, formatting of references, and language refinement. All AI-generated content was thoroughly reviewed and verified by the authors to ensure accuracy and adherence to scientific standards.

REFERENCES CITED

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. (2014). "Machine learning for neuroimaging with scikit-learn," *Frontiers in Neuroinformatics* 8, article 14. https://doi.org/10.3389/fninf.2014.00014

Ahmad, A., Zainuddin, R., Saksono, B., Anita, S. H., Zulfiana, D., Ermawar, R. A., Arfah, R., Natsir, H., Karim, H., Irmawati, Ramli, S. R., and Karim, A. (2025). "Agarose-based antibacterial films from *Gracilaria* sp.: Isolation, characterization, and metal nanoparticle incorporation," *Emerging Science Journal* 9(5), 2331-2349. https://doi.org/10.28991/ESJ-2025-09-05-03

Ajao, O., Jeaidi, J., Benali, M., Restrepo, A. M., El Mehdi, N., and Boumghar, Y. (2018). "Quantification and variability analysis of lignin optical properties for colour-dependent industrial applications," *Molecules* 23(2), article 377. https://doi.org/10.3390/molecules23020377

Asim, A. M., Uroos, M., and Muhammad, N. (2020). "Extraction of lignin and quantitative sugar release from biomass using efficient and cost-effective pyridinium protic ionic liquids," *RSC Advances* 10, 44003-44014. https://doi.org/10.1039/d0ra09098k

ASTM D1106-56 (1979). "Standard test method for lignin in wood," American Society for Testing and Materials, West Conshohocken, PA, USA.

Ayikpa, K. J., Mamadou, D., Ballo, A. B., and Gouton, P. (2025). "Intelligent selection of spectral bands from high-precision spectroradiometer measurements for optimizing cocoa bean classification," *Electronics* 14(10), article 976. https://doi.org/10.3390/electronics14101976

Barnes, R. J., Dhanoa, M. S., and Lister, S. J. (1989). "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra," *Applied Spectroscopy* 43(5), 772-777. https://doi.org/10.1366/000370289420220

Bentéjac, C., Csörgo, A., and Martínez-Muñoz, G. (2019). "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review* 54, 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Bilak, Y., Reblian, A., Buchuk, R., and Fedorka, P. (2025). "Development of a combined neural network model for effective spectroscopic analysis," *Eastern-European Journal of Enterprise Technologies* 1(4), 41-51. https://doi.org/10.15587/1729-4061.2025.322627

Bourbiaux, D., Pu, J., Rataboul, F., Djakovitch, L., Geantet, C., and Laurenti, D. (2021). "Reductive or oxidative catalytic lignin depolymerization: An overview of recent advances," *Catalysis Today* 373, 24-37. https://doi.org/10.1016/j.cattod.2021.03.027

Chan, C. H., Sun, M., and Huang, B. (2022). "Application of machine learning for advanced material prediction and design," *EcoMat* 4(4), article e12194. https://doi.org/10.1002/eom2.12194

Chandrasekaran, S. R., Hopke, P. K., Rector, L., Allen, G., and Lin, L. (2012). "Chemical composition of wood chips and wood pellets," *Energy Fuels* 26(8), 4932-4937. DOI: 10.1021/ef300884k

Cheng, J., Dai, Q., Sun, D., Zeng, X., Liu, D., and Pu, H. (2013). "Applications of non-destructive spectroscopic techniques for fish quality and safety evaluation and inspection," *Trends in Food Science and Technology* 34(1), 18-31. DOI: 10.1016/j.tifs.2013.08.005

Creteanu, A., Lungu, C., and Lungu, M. (2024). "Lignin: An adaptable biodegradable polymer used in different formulation processes," *Pharmaceuticals* 17(10), article 1406. https://doi.org/10.3390/ph17101406

Cui, J., Chen, X., and Zhao, Y. (2025). "Beyond traditional airPLS: Improved baseline removal in SERS with parameter-focused optimization and prediction," *Analytical Chemistry* 97(30), 16211-16218. https://doi.org/10.1021/acs.analchem.5c01253

Duy, N.V., Tsygankov, P. Y., and Menshutina, N. V. (2024). "Facile lignin extraction and application as natural UV blockers in cosmetic formulations," *ChemEngineering* 8(4), article 69. DOI: 10.3390/chemengineering8040069

Galelli, S., and Castelletti, A. (2013). "Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling," *Hydrology and Earth System Sciences* 17(7), 2669-2684. https://doi.org/10.5194/hess-17-2669-2013

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

Grisanti, E., Totska, M., Huber, S., Calderon, C. K., Hohmann, M., Lingenfelser, D., and Otto, M. (2018). "Dynamic localized SNV, peak SNV, and partial peak SNV: Novel standardization methods for preprocessing of spectroscopic data used in predictive modeling," *Journal of Spectroscopy* 2018, article 5037572. https://doi.org/10.1155/2018/5037572

Isikgor, F., and Becer, R. (2015). "Lignocellulosic biomass: A sustainable platform for the production of bio-based chemicals and polymers," *Polymer Chemistry* 6, 4497-4559. https://doi.org/10.1039/c5py00263j

Jafarzadeh, H., Mahdianpari, M., Gill, E., Mohammadimanesh, F., and Homayouni, S. (2021). "Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral and PolSAR data: A comparative evaluation," *Remote Sensing* 13, article 4405. https://doi.org/10.3390/rs13214405

Johnson, J. B., Mani, J. S., and Naiker, M. (2023). "Microplate methods for measuring phenolic content and antioxidant capacity in chickpea: Impact of shaking," *Engineering Proceedings* 48(1)DOI: 10.3390/CSAC2023-15167

Kline, L., Hayes, D., Womac, A., and Labbe, N. (2010). "Simplified determination of lignin content in hard and soft woods via UV-spectrophotometric analysis of biomass dissolved in ionic liquids," *BioResources* 5(3), 1366-1383. https://doi.org/10.15376/biores.5.3.1366-1383

Kumari, P., and Vinu, R. (2025). "Sono-fenton-assisted depolymerization of lignin to valuable phenolic compounds," *ACS Omega* 10, 28944-28955. https://doi.org/10.1021/acsomega.5c00431

Lee, R. A., Bédard, C., Berberi, V., Beauchet, R., and Lavoie, J. (2013). "UV–Vis as quantification tool for solubilized lignin following a single-shot steam process," *Bioresource Technology* 144, 658-663. DOI: 10.1016/j.biortech.2013.06.045

Lee, T., Ullah, A., and Wang, R. (2019). "Bootstrap aggregating and random forest," in: *Macroeconomic Forecasting in the Era of Big Data* 52, 389-429. https://doi.org/10.1007/978-3-030-31150-6_13

Li, N., Pan, X., and Alexander, J. (2016). "A facile and fast method for quantitating lignin in lignocellulosic biomass using acidic lithium bromide trihydrate (ALBTH)," *Green Chemistry* 18, 5367-5376. https://doi.org/10.1039/c6gc01090c

Li, Y., Pan, T., Li, H., and Chen, S. (2020). "Non-invasive quality analysis of thawed tuna using near infrared spectroscopy with baseline correction," *Journal of Food Process Engineering* 43, article 13445. https://doi.org/10.1111/jfpe.13445

Li, Y., Zhao, S., Li, Y., Ragauskas, A. J., Song, X., and Li, K. (2022). "Revealing the relationship between molecular weight of lignin and its color, UV-protecting property," *International Journal of Biological Macromolecules* 223, 1287-1296. https://doi.org/10.1016/j.ijbiomac.2022.11.067

Lu, F., Wang, C., Chen, M., Yue, F., and Ralph, J. (2021). "A facile spectroscopic method for measuring lignin content in lignocellulosic biomass," *Green Chemistry* 23, 5106-5112. https://doi.org/10.1039/d1gc01507a

Madadi, M., Elsayed, M., Sun, F., Wang, J., Karimi, K., Song, G., Tabatabaei, M., and Aghbashlo, M. (2023). "Sustainable lignocellulose fractionation by integrating p-toluenesulfonic acid/pentanol pretreatment with mannitol for efficient production of glucose, native-like lignin, and furfural," *Bioresource Technology* 371, article 128591. https://doi.org/10.1016/j.biortech.2023.128591

Martínková, L., Grulich, M., Pátek, M., Křístková, B., and Winkler, M. (2023). "Bio-based valorization of lignin-derived phenolic compounds: A review," *Biomolecules* 13(5), article 717. https://doi.org/10.3390/biom13050717

Mateo, S., Fabbrizi, G., and Moya, A. (2025). "Lignin from plant-based agro-industrial biowastes: From extraction to sustainable applications," *Polymers* 17(7), article 952. https://doi.org/10.3390/polym17070952

Ouyang, T., Wang, C., Yu, Z., Stach, R., Mizaikoff, B., Liedberg, B., Huang, G., and Wang, Q. (2019). "Quantitative analysis of gas phase IR spectra based on extreme learning machine regression model," *Sensors* 19(24), article 5535. https://doi.org/10.3390/s19245535

Pungpa, S., Jaroensutasinee, K., Jaroensutasinee, M., Srisang, W., Chumkiew, S., and Sparrow, E. B. (2025). "Integrating satellite and UAV imagery for mangrove aboveground biomass and carbon stock modeling," *Journal of Human, Earth, and Future* 6(2), 474-487. DOI: 10.28991/HEF-2025-06-02-014

Qiao, H., Han, M., Ouyang, S., Zheng, Z., and Ouyang, J. (2022). "An integrated lignocellulose biorefinery process: Two-step sequential treatment with formic acid for efficiently producing ethanol and furfural from corn cobs," *Renewable Energy* 191, 775-784. https://doi.org/10.1016/j.renene.2022.04.027

Rajendran, T., Balakrishnan, C., Yamini, B., Srilakshmi, C. H., Maheswari, B., Nalini, M., and Siva Subramanian, R. (2024). "Optimizing prediction accuracy in high-dimensional data: Comparative analysis of feature selection methods with naive bayes algorithm," *International Journal of Electronics and Communication Engineering* 11(3), 41-52. https://doi.org/10.14445/23488549/ijece-v11i3p105

Rinnan, Å., van den Berg, F., and Engelsen, S. B. (2009). "Review of the most common pre-processing techniques for near-infrared spectra," *TrAC Trends in Analytical Chemistry* 28(10), 1201-1222. https://doi.org/10.1016/j.trac.2009.07.007

Song, X., Aryal, S., Ting, K., Liu, Z., and He, B. (2021). "Spectral–spatial anomaly detection of hyperspectral data based on improved isolation forest," *IEEE Transactions on Geoscience and Remote Sensing* 60, 1-16. https://doi.org/10.1109/tgrs.2021.3104998

Tribot, A., Amer, G., Alio, M. A., Baynast, H., Delattre, C., Pons, A., Mathias, J., Callois, J., Vial, C., Michaud, P., *et al.* (2019). "Wood-lignin: Supply, extraction processes and use as bio-based material," *European Polymer Journal* 112, 228-240. https://doi.org/10.1016/j.eurpolymj.2019.01.007

Udo, E. F., Idowu, O. J., and Bada, A. T. (2024). "Influence of successive manual pipetting of multiple samples on precision and reliability of glucose test results," *Sokoto Journal of Medical Laboratory Science* 9(2), article 327. DOI: 10.4314/sokjmls.v9i2.38

Wang, H., Srinivasan, R., Yu, F., Steele, P., Li, Q., and Mitchell, B. (2011). "Effect of acid, alkali, and steam explosion pretreatments on characteristics of bio-oil produced from pinewood," *Energy Fuels* 25(8), 3758-3764. DOI: 10.1021/ef2004909

Wang, Z., Mu, L., Miao, H., Shang, Y., Yin, H., and Dong, M. (2023). "An innovative application of machine learning in prediction of the syngas properties of biomass chemical looping gasification based on extra trees regression algorithm," *Energy* 275, article 127438. https://doi.org/10.1016/j.energy.2023.127438

Watkins, D., Nuruddin, M., Hosur, M., Tcherbi-Narteh, A., and Jeelani, S. (2014). "Extraction and characterization of lignin from different biomass resources," *Journal of Materials Research and Technology* 4, 26-32. https://doi.org/10.1016/j.jmrt.2014.10.009

Weiland, F., Kohlstedt, M., and Wittmann, C. (2021). "Guiding stars to the field of dreams: Metabolically engineered pathways and microbial platforms for a sustainable lignin-based industry," *Metabolic Engineering* 71, 13-41. https://doi.org/10.1016/j.ymben.2021.11.011

Wijitkosum, S., and Sriburi, T. (2025). "Effects of hardwood biomass variability on biochar properties: Insights from wood waste utilization," *Journal of Human, Earth, and Future* 6(1), 12-26. DOI: 10.28991/HEF-2025-06-01-02

Xie, M., Wang, D., and Xie, L. (2018). "A feature-weighted SVR method based on kernel space feature," *Algorithms* 11(5), article 62. https://doi.org/10.3390/a11050062

Xu, Z., Lei, P., Zhai, R., Wen, Z., and Jin, M. (2019). "Recent advances in lignin valorization with bacterial cultures: Microorganisms, metabolic pathways, and bio-products," *Biotechnology for Biofuels* 12, article 32. https://doi.org/10.1186/s13068-019-1376-0

Yan, C. (2025). "A review on spectral data preprocessing techniques for machine learning and quantitative analysis," *iScience* 28(7), article 112759. https://doi.org/10.1016/j.isci.2025.112759

Yue, J., Feng, H., Yang, G., and Li, Z. (2018). "A comparison of regression techniques for estimation of above-ground winter wheat biomass using near-surface spectroscopy," *Remote Sensing* 10(1), article 66. https://doi.org/10.3390/rs10010066

Zhang, Y., Jun, Liang, S., Li, X., and Li, M. (2020). "An evaluation of eight machine learning regression algorithms for forest aboveground biomass estimation from multiple satellite data products," *Remote Sensing* 12, article 4015. https://doi.org/10.3390/rs12244015

Zimmermann, B., and Kohler, A. (2013). "Optimizing Savitzky–Golay parameters for improving spectral resolution and quantification in infrared spectroscopy," *Applied Spectroscopy* 67, 892-902. https://doi.org/10.1366/12-06723

Article submitted: December 2, 2025; Peer review completed: February 7, 2026; Revised version received: March 8, 2026; Accepted: March 9, 2026; Published: April 1, 2026.
DOI: 10.15376/biores.21.2.4436-4456

APPENDIX

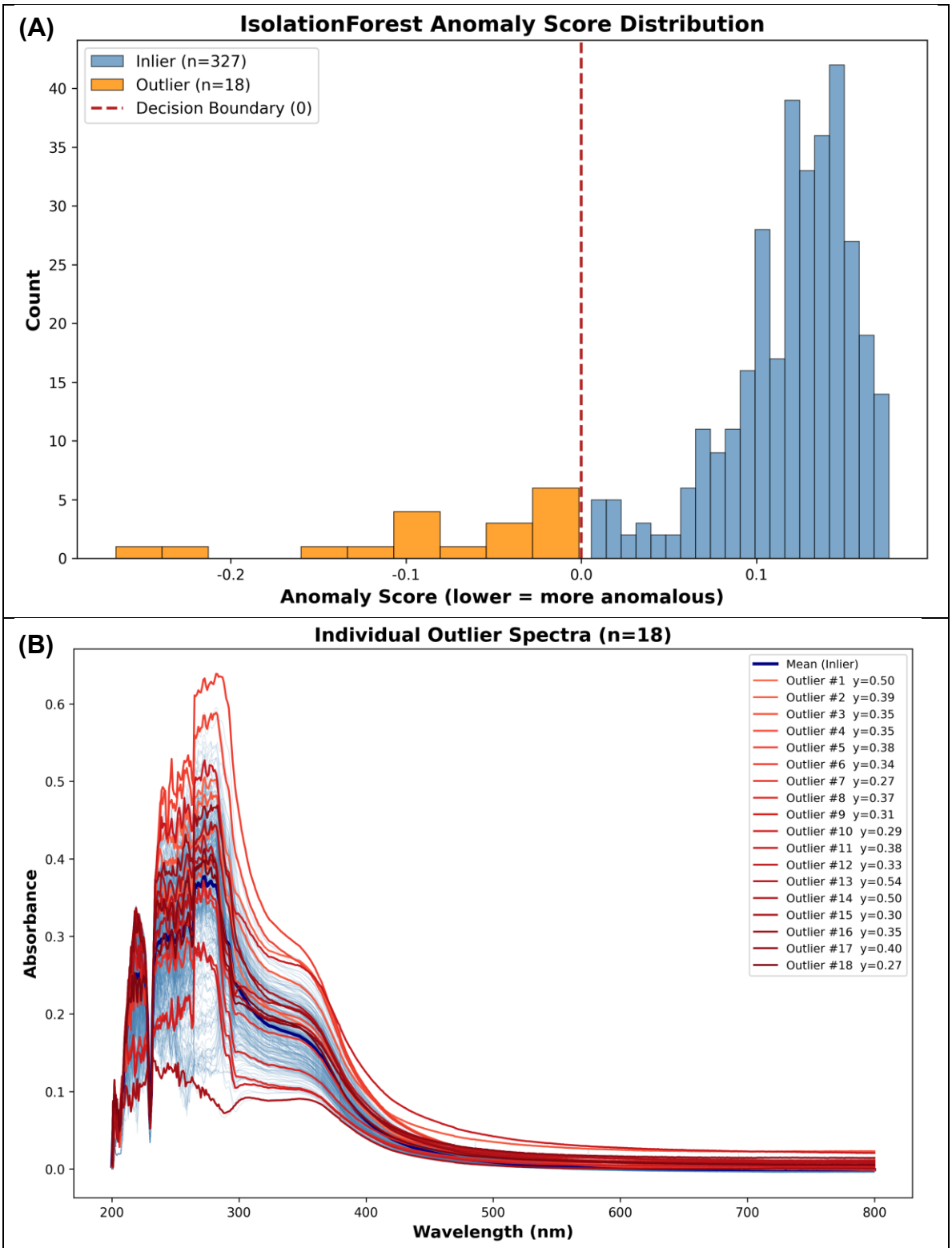


Fig. A1. Isolation Forest outlier detection results: (A) anomaly score distribution with decision boundary and (B) UV-visible spectra of identified outliers (n = 18) compared to the mean inlier spectrum

Table A1. Detailed Hyperparameter Search Space and Optimized Results for Machine Learning Models

Model	Hyperparameter	Search Space / Type	Optimized Value
ET (Extra Trees)	n_estimators	[50, 300] (Integer)	143
	max_depth	[5, 20] (Integer)	12
	min_samples_split	[2, 10] (Integer)	3
	min_samples_leaf	[1, 5] (Integer)	1
	max_features	{sqrt, log2, None}	sqrt'
RF (Random Forest)	n_estimators	[50, 300] (Integer)	67
	max_depth	[5, 20] (Integer)	13
	min_samples_split	[2, 10] (Integer)	3
	min_samples_leaf	[1, 5] (Integer)	1
XGB (XGBoost)	n_estimators	[50, 300] (Integer)	195
	max_depth	[3, 10] (Integer)	3
	learning_rate	[0.01, 0.3] (Log-float)	0.11436
	subsample	[0.5, 1.0] (Float)	0.64996
	colsample_bytree	[0.5, 1.0] (Float)	0.68917
SVR (Support Vector)	C	[0.1, 100] (Log-float)	16.05391
	gamma	{scale, auto}	auto'
	epsilon	[0.001, 0.1] (Log-float)	0.00171
	kernel	{rbf, linear, poly}	rbf'