Journal of Wood Science

ORIGINAL ARTICLE

Open Access

Feature importance measures from random forest regressor using near-infrared spectra for predicting carbonization characteristics of kraft lignin-derived hydrochar

Sung-Wook Hwang[1], Hyunwoo Chung[2], Taekyeong Lee[1], Jungkyu Kim[3], YunJin Kim[3], Jong-Chan Kim[3], Hyo Won Kwak[1,2,3], In-Gyu Choi[1,2,3] and Hwanmyeong Yeo[1,2,3*]

Abstract

This study investigated the feature importance of near-infrared spectra from random forest regression models constructed to predict the carbonization characteristics of hydrochars produced by hydrothermal carbonization of kraft lignin. The model achieved high coefficients of determination of 0.989, 0.988, and 0.985 with root mean square errors of 0.254, 0.003, and 0.008 when predicting the carbon content, atomic O/C ratio, and H/C ratio, respectively. The random forest models outperformed the multilayer perceptron models for all predictions. In the feature importance analysis, the spectral regions at 1600–1800 nm, the first overtone of C–H stretching vibrations, and 2000–2300 nm, the combination bands, were highly important for predicting the carbon content and O/C predictions, whereas the region at 1250–1711 nm contributed to predicting H/C. The random forest models trained with the high-importance regions achieved better prediction performances than those trained with the entire spectral range, demonstrating the usefulness of the feature importance yielded by the random forest and the feasibility of selective application of the spectral data.

Keywords Feature importance measures, Hydrochar, Hydrothermal carbonization, Lignin, Near-infrared spectroscopy, Random forest

Introduction

To combat global warming and the resulting climate change, the Intergovernmental Panel on Climate Change (IPCC) has adopted a special report on global temperature increase of 1.5 ℃ [1]. In response, countries around the globe are establishing carbon–neutral strategies to reduce CO_2 emissions, including the use and development of sustainable and renewable sources of energy. Wood is a representative natural resource that is both renewable and sustainable. Lignin, one of the main elements in wood is an aromatic component of lignocellulosic biomass and is carbon rich. The pulp industry mass produces kraft lignin as a waste [2]. Lignin is a bioresource known for maximizing resource efficiency and converting waste into valuable resources, helping in achieving carbon neutrality.

Hydrothermal carbonization (HTC) is a thermochemical process that converts organic compounds, such as lignin, into structured carbon materials called hydrochars

*Correspondence:
Hwanmyeong Yeo
hyeo@snu.ac.kr
[1] Research Institute of Agriculture and Life Sciences, Seoul National University, 1 Gwanak-Ro, Gwanak-Gu, Seoul 08826, Republic of Korea
[2] Department of Forest Sciences, College of Agriculture and Life Sciences, Seoul National University, 1 Gwanak-Ro, Gwanak-Gu, Seoul 08826, Republic of Korea
[3] Department of Agriculture, Forestry and Bioresources, College of Agriculture and Life Sciences, Seoul National University, 1 Gwanak-Ro, Gwanak-Gu, Seoul 08826, Republic of Korea

with relatively mild temperatures of 130–250 ℃ [3, 4]. HTC can convert biomass into carbonaceous solids without an energy-intensive drying process [5]. The HTC yield of lignin is higher than that of cellulose and hemicellulose because of the thermally stable phenolic structures of lignin [6, 7].

Lignin-derived hydrochar can be used in various applications, including fuels, batteries, polymer composites, adsorbents, and electrochemical devices [8–11]. The carbonization characteristics of hydrochar are typically determined using elemental analysis. However, it has recently been reported that rapid and non-destructive evaluation is possible using near-infrared (NIR) spectroscopy with multivariate analysis [12]. Numerous studies have demonstrated that NIR spectroscopy is suitable for capturing the characteristics of biological materials [13–20]. Nevertheless, the evaluation of the spectral bands contributing to predictions by the models has been limited to estimation using indirect methods.

Random forest (RF) is a prediction model and machine learning technique that incorporates the importance of input variables for prediction [21]. RF is a framework for aggregating predictions from multiple tree models, yielding quantified information about the features and their contribution to the prediction. This nature of tree models makes them an ideal choice for studies using biological data [22–25].

This study established RF regression models trained with NIR spectral data to predict the carbonization characteristics of hydrochars produced by the hydrothermal carbonization of kraft lignin. Furthermore, the NIR spectral regions were classified according to their feature importance as computed by the RF models. The performance evaluation of the RF models trained with each selected spectral region verified the practical usefulness of the feature importance computed by the RF.

Materials and methods

Samples and hydrothermal carbonization

Kraft lignin, a raw material for hydrochars, was provided by a domestic pulp manufacturer (Moorim P&P, Ulsan, Korea). Lignin was produced as a byproduct of an industrial-scale pulping process to manufacture bleached hardwood pulp using an alkaline white liquor consisting of sodium hydroxide, chlorine dioxide, and sodium sulfide.

For HTC, 5.6 g of lignin powder was mixed with 140 mL of distilled water to prepare suspensions with a solid-to-liquid ratio of 1:25. A glass liner containing the suspension was placed in a reaction vessel and heated in a heating mantle at target temperatures of 150, 175, 200, 225, and 250 ℃ for retention times of 1, 2, 3, and 5 h, respectively. Retention time is the duration at which the target temperature is maintained. After heating, the reaction vessel was allowed to cool naturally to room temperature (13.4–23.1 ℃). The solid residues generated from the HTC of lignin were vacuum filtered, oven-dried, and pulverized to produce powdered hydrochars.

Elemental analysis

Elemental analysis was performed to investigate the elemental compositions of the hydrochars. The weight percentages of carbon (C), hydrogen (H), nitrogen (N), and sulfur (S) were measured using an elemental analyzer (Flash EA 1112, Thermo Electron Corp., Waltham, MA, USA). The oxygen (O) weight percentage was estimated to be $100 - (C + H + N + S)$. The C wt%, O/C, and H/C ratios were calculated as indicators for evaluating the carbonization characteristics of the hydrochars.

Spectral dataset

NIR spectral data were used as input variables to build regression models for predicting the carbonization characteristics of the hydrochars. NIR reflectance spectra were collected from the hydrochars using an NIR spectrometer (NIR Quest, Ocean Insight, Orlando, FL, USA) with a reflection probe and tungsten halogen lamp. The optical resolution of the spectrometer was 6.6 nm, and the spectra were collected in the wavelength range of 870–2500 nm.

All spectra were second derivatized by Savitzky–Golay smoothing to 13 points with the fifth-order function [26]. A wavelength range of 1250–2300 nm, excluding noisy regions in the original range, was selected and used to build the prediction models. The selected spectral region corresponded to 165 input variables. Hwang et al. [12] demonstrated the effectiveness of second-derivative transformation and spectral selection in predicting the carbonization characteristics of hydrochars.

Three NIR spectra were acquired for each HTC condition, including the control sample. The dataset thus consisted of 63 spectra. The dataset was independently divided into training and test sets at a ratio of 8–2 and used for the construction and evaluation of the prediction model.

Regression models

Random forest regressor

The RF model for regression [21], an ensemble learning technique, was used to predict the carbonization characteristics of the hydrochars. Ensemble learning combines predictions from multiple models to produce more accurate results than a single model. This study

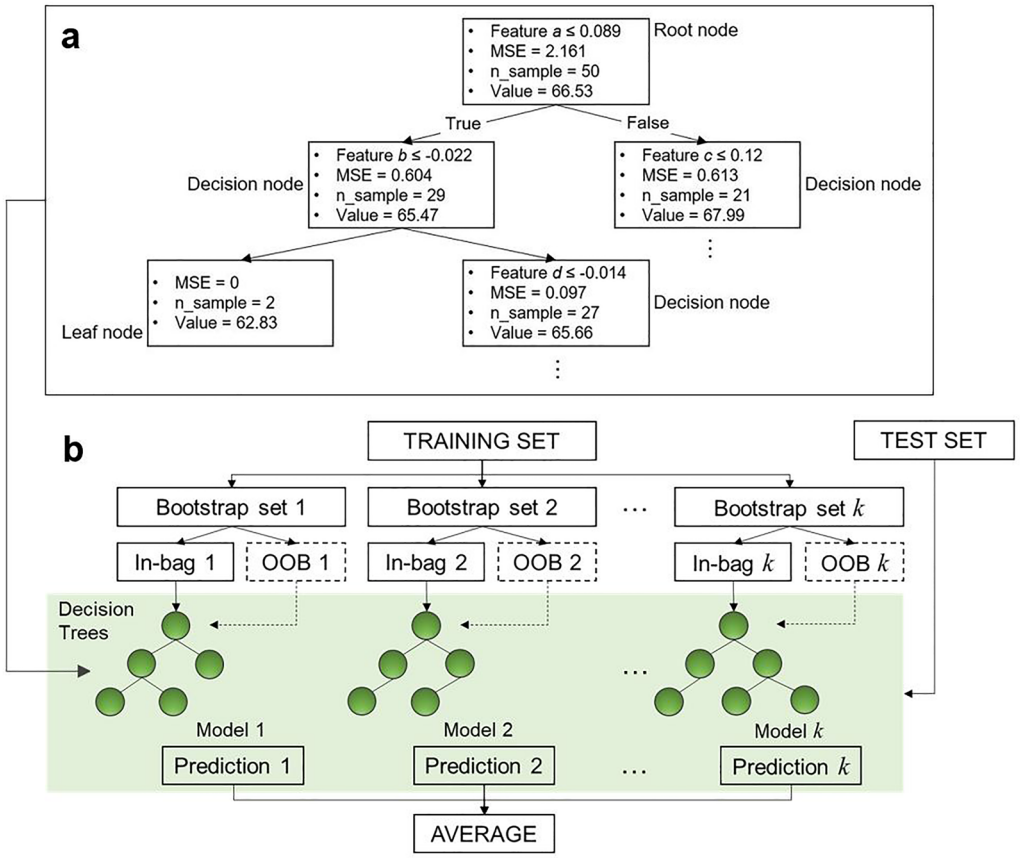

Fig. 1 Schematic diagrams of decision tree (**a**) and random forest (**b**) models for regression. MSE, mean square error; n_sample, number of samples in a node; OOB, out-of-bag samples

used decision trees (DT) for regression [27] as the base learner to construct an RF model.

DT is a simple model that predicts the result by performing a split based on the predictor (input variable) that reduces the mean squared error (MSE) the most. As shown in Fig. 1a, when predicting the output from the input variables, DT starts from a single node (root node) and creates branches (decision nodes) based on the input variable (feature) with the smallest MSE (Eq. 1).

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y}\right)^2, \tag{1}$$

where y and $\widehat{y}$ are the measured and predicted values of the samples in a node, respectively, and n is the number of samples in a node. This process is repeated at each decision node to create a tree-shaped decision structure. A node that cannot branch further owing to a non-decreasing MSE is called a leaf node, and the average value of the samples in that node becomes a candidate for the prediction. When unseen data are input into the completed DT model, the data move according to predetermined branching criteria. The value of the leaf node where the data finally arrived was used as the predicted value of the DT.

A schematic of the RF model is illustrated in Fig. 1b. RF is a combination of multiple DTs and improves prediction performance by averaging the predictions of multiple DTs and controls overfitting, a chronic weakness of DTs. To increase the randomness of the DTs, the RF model builds them using random sampling without using all input variables. This process is performed to create independent trees. In addition, RF generates bootstrap sets from the training set using random sampling with replacement. In this process, approximately 2/3 (in-bag samples) of the training data are used for DT training. The remaining 1/3 (out-of-bag samples) of the data are used to validate the tree model, similar to the threefold cross-validation. The probability that a sample is not selected from m data points during random sampling with replacement is $(m-1)/m$. If this is repeated m times, the out-of-bag (OOB) probability is approximately 36.8%, according to Eq. (2).

$$\mathrm{OOB} = \lim_{m\to\infty}\left(\frac{m-1}{m}\right)^m = \lim_{m\to\infty}\left(1-\frac{1}{m}\right)^m = e^{-1} \approx 0.368. \quad (2)$$

The RF model was built with multiple DTs generated by learning the in-bag samples, and the final prediction of the model for new data was output as the average value of the predictions of the DTs.

In this study, all DTs constituting the RF were based on the classification and regression tree (CART) [27] algorithm and were independently generated without pruning. For n_feature, which are input variables for DT generation, the square root ('sqrt'), binary logarithm ('log2'), and one-third ('1/3') of all spectral points were tested. An n_tree from 10 to 300 was tested, and the optimal n_feature and n_tree were determined based on the minimum OOB error via grid searches. The coefficient of determination (R^2) and root mean square error (RMSE) were used as evaluation metrics for the model performance.

Multilayer perceptron regressor

Multilayer perceptron (MLP) models were employed to compare the prediction performance with RF models. The MLP regressor learns data using backpropagation, with no activation function in the output layer. The squared error was used as the loss function, and the model was optimized using the stochastic gradient descent-based optimizers SGD and Adam. The various network architectures listed in Table 1 were tested with logarithmic learning rates ranging from 0.0001 to 0.1 to optimize the network configuration of the MLP. The maximum number of iterations was set to 3000. Grid searches with threefold cross-validation optimized the hyperparameters of network architectures, optimizers, and learning rates.

The prediction performances of the RF and MLP regression models established in this study were compared with those of the reported partial least squares (PLS) regression models, which are a chemometric approach combined with NIR spectroscopy and multivariate analysis [12].

Feature importance measures

The spectral feature importance was measured based on the mean decrease in impurity (MDI) [28] to identify the NIR regions that contribute to predicting the carbonization characteristics of hydrochars. Tree-based models provide information about the contribution of the input variables used in prediction, called feature importance.

Table 1 Network architectures of MLP regression models tested

2-layer MLP					3-layer MLP						
Input layer	/	Hidden layer	/	Output layer	Input layer	/	Hidden layers			/	Output layer
							1st		2nd		
165	/	64 128 256 512 1024	/	1	165	/	64	/	64 128 256 512 1024	/	1
							128	/	64 128 256 512 1024		
							256	/	64 128 256 512 1024		
							512	/	64 128 256 512 1024		
							1024	/	64 128 256 512 1024		

The feature importance of an ensemble model, such as RF, is also an ensemble of the feature importance of its base models. In this study, variance reduction based on MSE, a criterion for branching nodes, was used to measure the feature importance of the DTs.

When an upper node (parent node) branches into two lower nodes (child nodes) by feature *i*, the importance of the feature is defined as the difference between the MSE of the parent node and the sum of the MSEs of the child nodes, which is called information gain (Eq. 3).

$$G_{Pj} = w_{Pj}M_{Pj} - w_{Lj}M_{Lj} - w_{Rj}M_{Rj}, \tag{3}$$

where G_{Pj} is the information gain of node *j*. *Pj* is the parent node *j*, and *Lj* and *Rj* are the left and right child nodes branched from *Pj*, respectively. *w* is the weight of the node and is the number of samples in the node relative to the total number of samples. *M* is the mean squared error of the node. The importance of feature *i* in a decision tree can be calculated as follows:

$$I(f_i)_{\mathrm{DT}} = \frac{\sum_{j:\text{node}j \text{ splits on} f_i} G_j}{\sum_{a\in\text{all nodes}} G_a}, \tag{4}$$

where $I(f_i)_{\mathrm{DT}}$ is the importance of feature *i* in the tree model and G_j is the information gain of node *j* branched by feature *i*. The feature importance of the RF model is computed as an ensemble of the importance of all the DTs in the RF. Before the ensemble, the importance of each feature was normalized using Eq. (5).

$$I(f_i)_{\mathrm{norm}} = \frac{I(f_i)}{\sum_{j\in\text{all fetures}} I(f_j)}. \tag{5}$$

Then, the final importance of each feature in the RF model was averaged over all DTs, as follows:

$$I(f_i)_{\mathrm{RF}} = \frac{\sum_{j:\text{all trees}} I(f_j)_{\mathrm{norm}}}{N_T}, \tag{6}$$

where $I(f_i)_{\mathrm{RF}}$ is the importance of feature *i* in the RF model and N_T is the total number of DTs in the RF.

Results and discussion

Elemental analysis

The elemental compositions of the hydrochars are listed in Table 2. The carbon content (C wt%) of the samples increased as the temperature and retention time for HTC, i.e., the reaction severity, increased (Fig. 2a). The C wt% of the control sample was 62.83 wt%, which increased to 69.37% after HTC at 250 ℃ for 5 h. The C wt% values of

Table 2 Elemental composition of hydrochars produced by hydrothermal carbonization of kraft lignin

Sample		Elemental composition (wt%)					O/C	H/C
Temp (℃)	Time (h)	C	H	O[a]	N	S		
Control		62.83	5.79	29.25	0.39	1.74	0.35	1.10
150	1	65.07	5.65	28.02	0.31	0.95	0.32	1.03
	2	65.68	5.68	27.34	0.33	0.97	0.31	1.03
	3	65.38	5.65	27.48	0.31	1.18	0.32	1.03
	5	65.46	5.66	27.50	0.32	1.07	0.32	1.03
175	1	65.36	5.66	27.23	0.46	1.29	0.31	1.03
	2	65.76	5.65	26.95	0.40	1.24	0.31	1.02
	3	65.73	5.64	27.02	0.41	1.20	0.31	1.02
	5	65.82	5.63	26.93	0.41	1.21	0.31	1.02
200	1	65.73	5.52	27.06	0.41	1.28	0.31	1.00
	2	66.19	5.46	26.51	0.43	1.41	0.30	0.98
	3	66.25	5.55	26.53	0.42	1.24	0.30	1.00
	5	67.02	5.58	25.73	0.43	1.23	0.29	0.99
225	1	67.24	5.58	25.62	0.43	1.13	0.29	0.99
	2	67.43	5.58	25.44	0.42	1.13	0.28	0.99
	3	67.44	5.48	25.56	0.41	1.11	0.28	0.97
	5	68.11	5.51	24.82	0.43	1.12	0.27	0.96
250	1	68.23	5.48	24.73	0.45	1.12	0.27	0.96
	2	68.64	5.54	24.43	0.47	0.92	0.27	0.96
	3	69.12	5.47	24.00	0.49	0.92	0.26	0.94
	5	69.37	5.38	23.88	0.49	0.87	0.26	0.92

[a] O (wt%) = 100 – (C + H + N + S) (wt%)

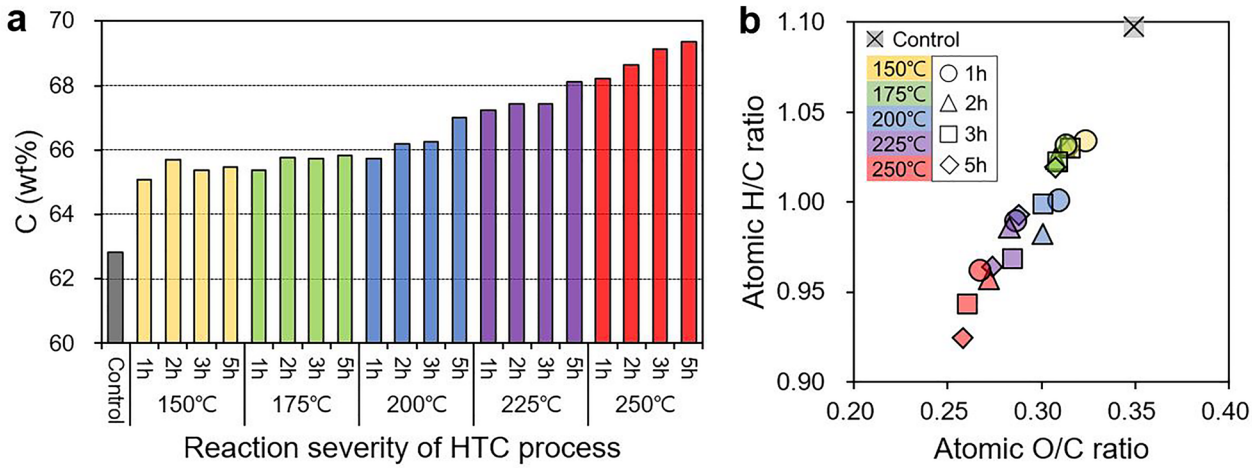

Fig. 2 Changes in the carbon content of lignin samples with increasing reaction severity in hydrothermal carbonization process (**a**), and the van Krevelen diagram for hydrochars (**b**)

hydrochars produced at 150 and 175 ℃ were similar at approximately 65%. However, above 200 ℃, the content increased linearly with an increase in HTC temperature and time. The increase in the C wt% was attributed to dehydration and decarboxylation during carbonization [4, 29]. Simultaneously, the atomic oxygen/carbon (O/C) ratio and atomic hydrogen/carbon (H/C) ratio gradually decreased (Fig. 2b), which was mainly due to chemical dehydration [30], suggesting that carbon-intensive materials were produced from HTC. The van Krevelen diagram (Fig. 2b) shows that lignin, whose elemental composition is similar to that of lignite brown coal, was converted to brown coal via HTC [7]. The reduced S wt% of the hydrochars was due to their dissolution during HTC [4].

Prediction of carbonization characteristics

RF models

Figure 3 shows the change in OOB errors with the addition of each regression tree during RF training to predict the carbonization characteristics of the hydrochars. At the beginning of the tree addition, the OOB errors were reduced, and the minima were measured for less than 50 trees. The following errors recovered slightly and remained at a similar level from 100 trees or more. Similar trends were observed for the C wt%, O/C, and H/C predictions. The optimal number of features for tree generation was 'log2' for C wt% prediction and 'sqrt' for O/C and H/C predictions.

The prediction results of the RF models for the C wt%, O/C, and H/C of the hydrochars are presented in Fig. 4 and Table 3. The scatter plots (Fig. 4) show that the training and test sets had similar trends. The RF models accurately predicted the carbonization performance of the hydrochars (Table 3). For the C wt% prediction, the model achieved the R^2 value of 0.989 on the test set from 43 regression trees built using the 'log2' (7 features) of all input variables. For O/C and H/C predictions, R^2 values of 0.988 and 0.985 were obtained from 43 and 16 trees (n_tree), respectively, with the number of NIR spectral points (n_feature) of 'sqrt' (13 features).

The O/C ratio indicates polarity and is related to the adsorbability of the material [31]. Both O/C and H/C ratios are indicators of the stability of the carbonaceous materials [32]. The lower the ratios the materials have, the more stable, inert, and resistant to decomposition, as they resemble the characteristics of graphite [33]. Therefore, the RF models established in this study have the potential to be applied for predicting the carbonization characteristics and evaluating the quality of hydrochars. In addition, the use of NIR spectroscopy supports rapid and non-destructive predictions.

Performance comparison

Table 4 shows the comparison of the performance of the RF model with that of other regression models in predicting the carbonization characteristics of the hydrochars. Because the DT models for regression yielded good predictions enough, their collaboration RF did not improve the performance except for O/C prediction. However, RF was found to be robust against bias and overfitting, whereas a single DT was vulnerable [34]. The RF models were comparable to the PLS regression models [12], a chemometric approach, and produced better predictions than MLP models. Although MLP is applicable to various non-linear problems, it performed poorer than the other tested models. The relatively low performance

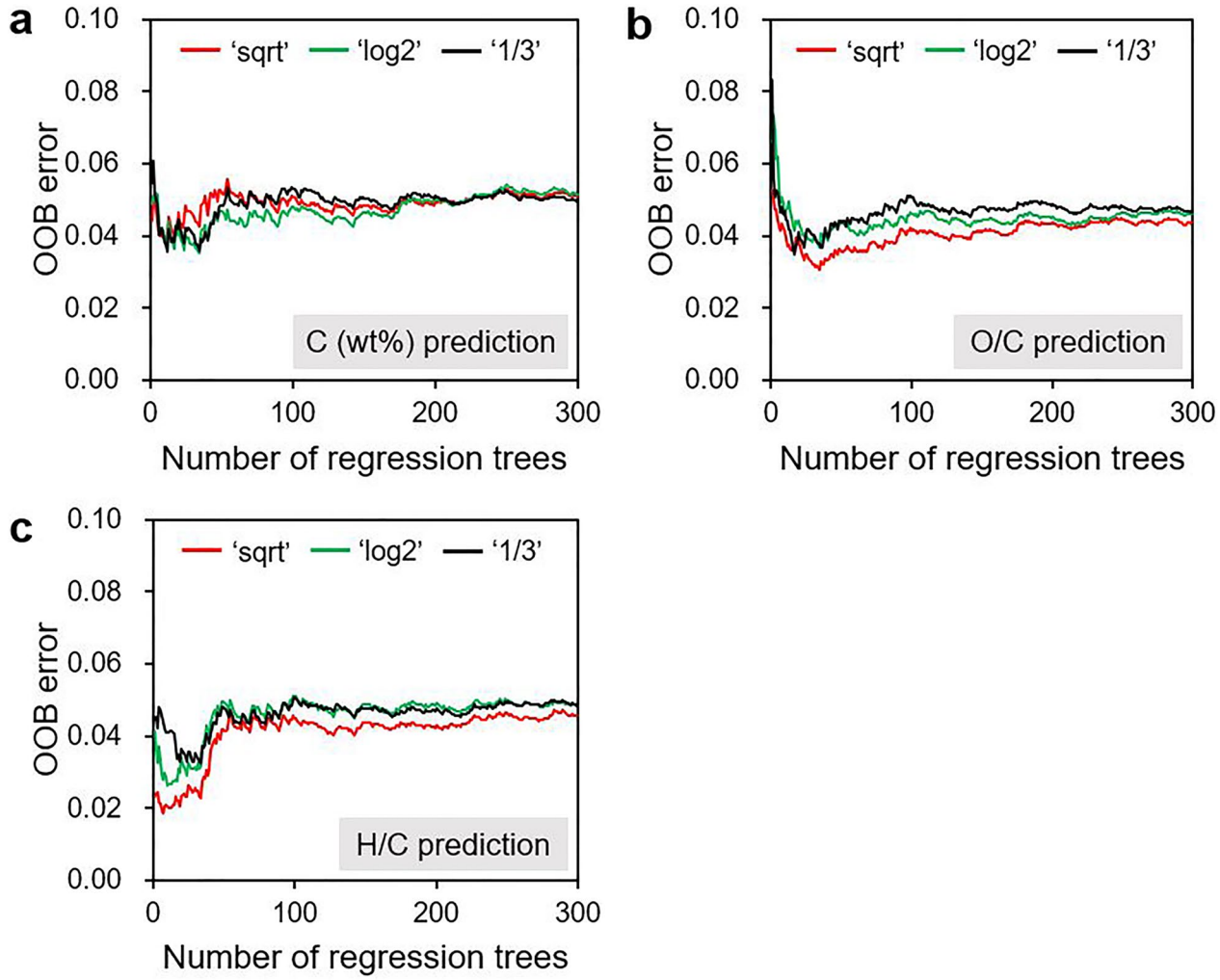

Fig. 3 Changes in out-of-bag error rates with increasing number of regression trees when predicting carbon content (**a**), O/C (**b**), and H/C (**c**)

of MLPs was attributed to the small scale of the NIR spectral dataset. These results support the methodological validity of RF regression combined with NIR spectroscopy to predict the carbonization characteristics of hydrochars.

Spectral feature importance

Mean decrease in impurity-based importance

The feature importance of the RF models for predicting the carbonization characteristics of hydrochars was computed based on the total reduction in the MSE. The second-derivative NIR spectra in the range of 1250–2300 nm of the control and hydrochar samples and their corresponding importance are shown in Fig. 5.

In the C wt% and O/C predictions, the spectral regions with high importance were 1600–1800 and 2000–2300 nm, respectively. The first region was dominated by the first overtone of C–H stretching vibrations [35]. However, the band at 1684 nm may have originated from a combination of carboxyl groups [36]. The decrease in the intensity of the peaks at 1684 nm with increasing temperature can be attributed to decarboxylation by HTC (Fig. 5). This is because removing the carboxyl groups increases the C wt% and decreases the O/C ratio [37]. Consequently, the band at 1684 nm yielded the highest importance for the C wt% and O/C predictions (Fig. 5a, b). Hwang et al. [12] estimated that the band at 1449 nm, assigned to the phenolic group, strongly influenced the prediction of the carbonization characteristics. However, in this study, the RF models suggested that the band had a low contribution in predicting C wt% and O/C with its low feature importance values. The second region comprises the combination bands, and the bands at 2132 nm (coupling of C–H and C=C stretching vibrations) and 2267 nm (coupling of O–H and C–O stretching vibrations) were assigned to lignin [36, 37]. The latter

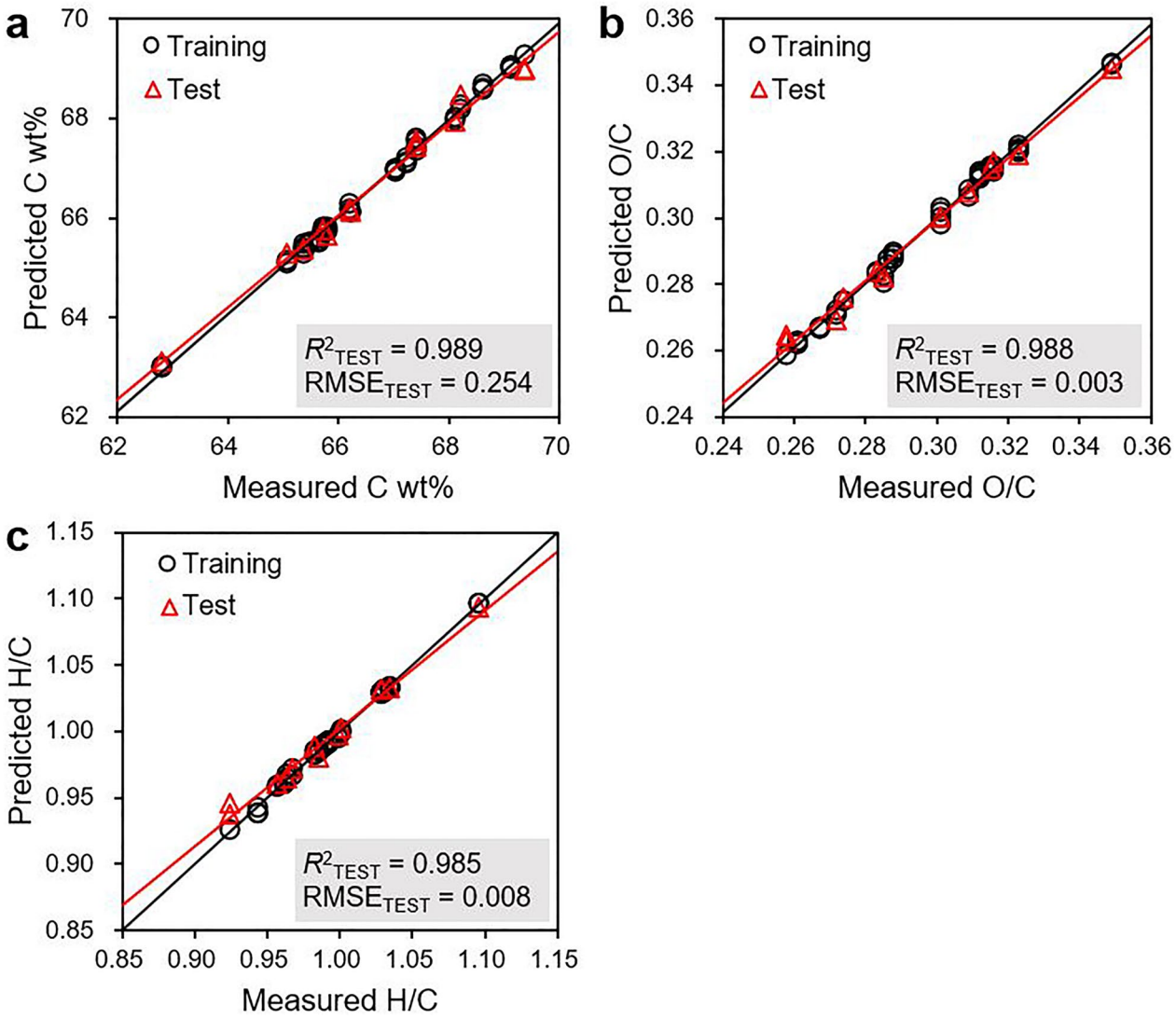

Fig. 4 Scatter plots for measured values of carbon content (**a**), O/C (**b**), and H/C (**c**) of hydrochars and their predicted values by RF models. R^2_{TEST}, coefficient of determination on test set data; $RMSE_{TEST}$, root mean square error on test set data

Table 3 Performance of random forest models in predicting carbon content, O/C, and H/C predictions

Output variable	RF parameters		Training set		Test set	
	n_feature	n_tree	R^2	RMSE	R^2	RMSE
C (wt%)	'log2' = 7	43	0.997	0.083	0.989	0.254
O/C	'sqrt' = 13	43	0.996	0.001	0.988	0.003
H/C	'sqrt' = 13	16	0.998	0.002	0.985	0.008

n_feature, number of features; n_tree, number of decision trees; all, all features (input variables); sqrt, square root of n_feature; R^2, coefficient of determination; RMSE, root mean square error

has been observed in the second-derivative spectrum of milled hardwood lignin [37].

The feature importance computed in the H/C prediction was different from the others. The importance of the band at 1449 nm, which has low importance for C wt% and O/C predictions, was highest for H/C prediction. This band was assigned to the first overtone of the O–H stretching vibration of the phenolic groups of lignin [38].

Table 4 Prediction performance for carbonization characteristics of random forest regression models and comparison with other regression models

Output variable	RF		DT		MLP		PLS [12]	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
C (wt%)	0.989	0.254	0.983	0.229	0.969	0.357	0.976	0.246
O/C	0.993	0.003	0.963	0.005	0.946	0.007	0.964	0.006
H/C	0.985	0.008	0.984	0.006	0.908	0.022	0.984	0.004

RF, random forest; DT, decision tree for regression, MLP, multilayer perceptron; R^2, coefficient of determination; RMSE, root mean square error

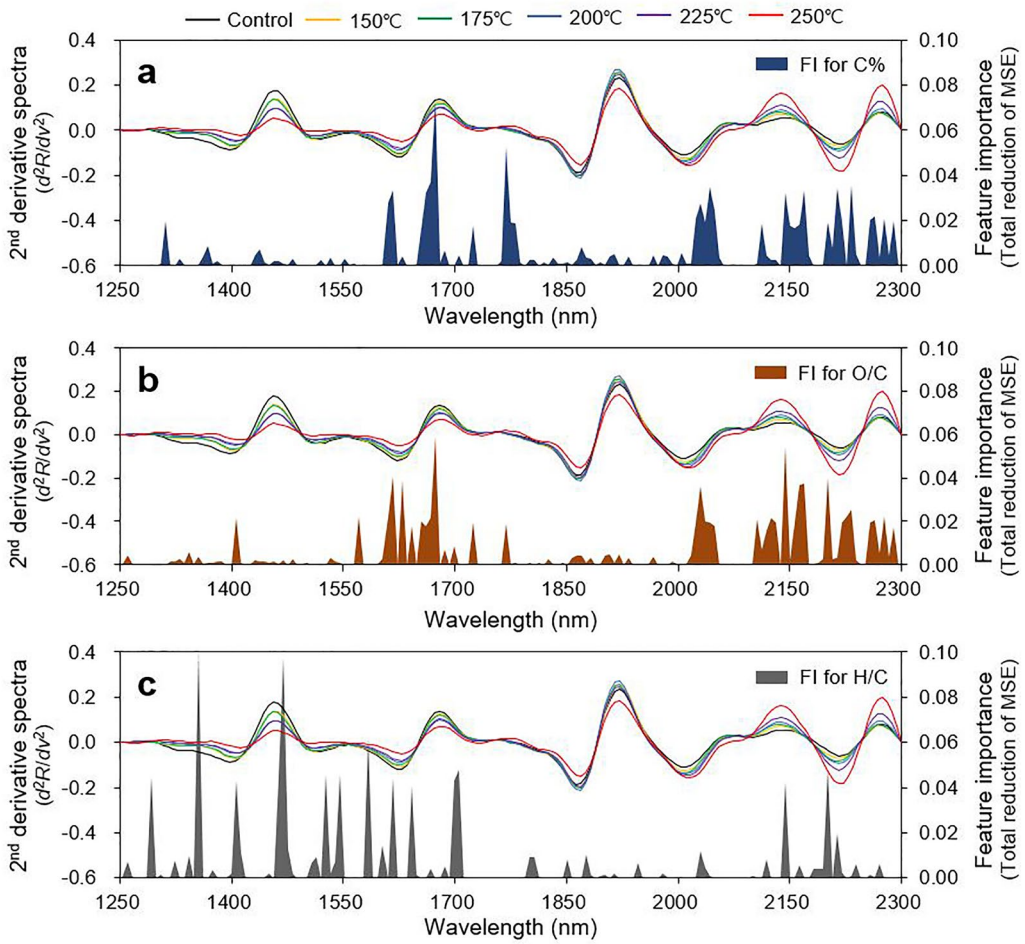

Fig. 5 Second-derivative NIR spectra of hydrochars and spectral feature importance of random forest regression models for predicting carbon content (**a**), O/C (**b**), and H/C ratios (**c**). FI, feature importance; MSE, mean square error

In contrast, the bands at 1684 and 2267 nm, which contributed to the C wt% and O/C predictions, respectively, had low importance for H/C prediction. For the spectral region of 1800–1999 nm, assigned to components regrading cellulose and water, low importance values were observed for all predictions.

Feature selection

The spectral regions were classified into high and low importance based on the feature importance values, and the RF models trained with the data from each region were reconstructed. The RF models trained with the high-importance spectral regions outperformed those trained with the low-importance regions and the full spectral range for all predictions. In addition, the models

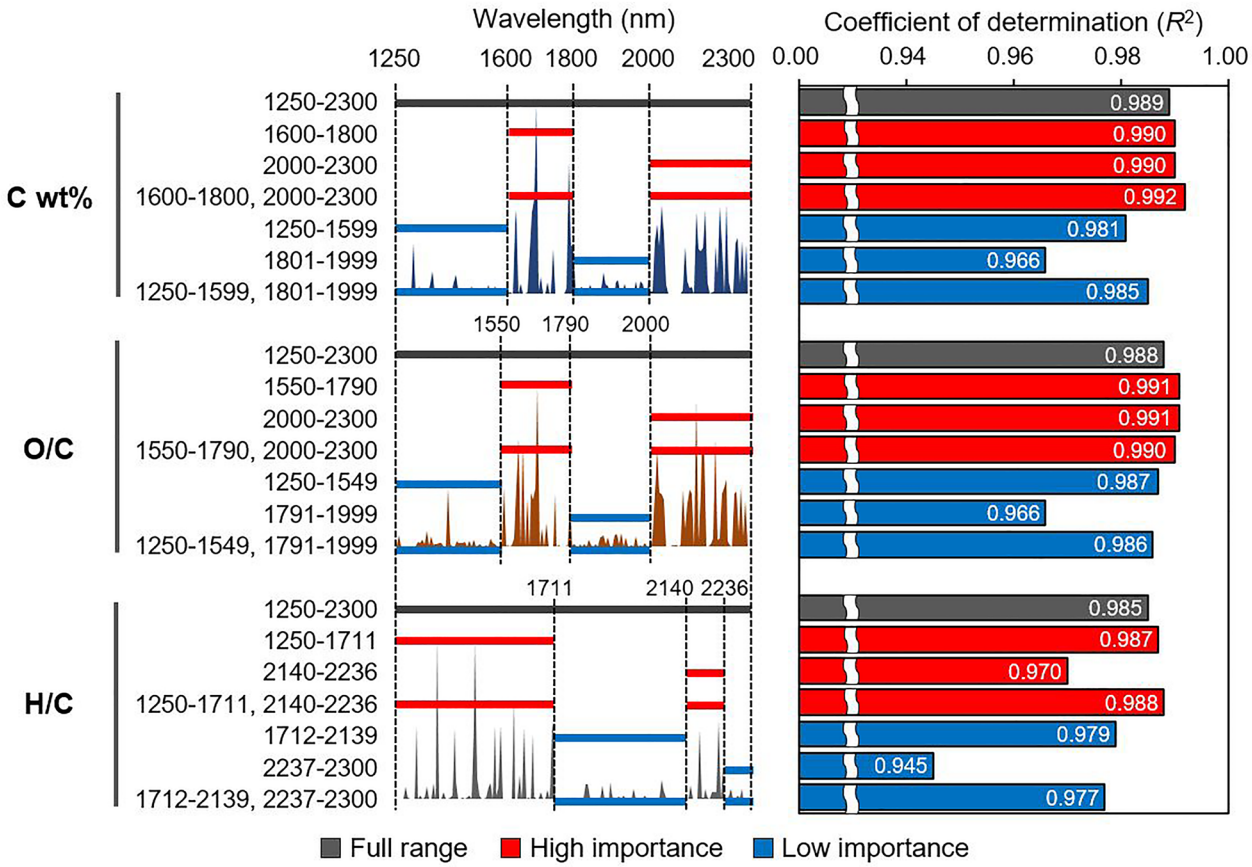

Fig. 6 Selection of NIR spectral regions by feature importance and comparison of prediction performance of RF models trained with each selected region

trained with a combination of high-importance regions yielded the best performance in this study (Fig. 6).

Although the number of features available for model building was lower owing to partial selection of the spectral region, the improvement in model performance proves that the feature importance computed from RF is reliable for predicting the carbonization characteristics of the hydrochars. Furthermore, these results suggest that the selection of custom spectral regions according to the output variables may be a better strategy for prediction.

Conclusions

RF regression models combined with NIR spectroscopy predicted the carbonization characteristics of lignin-derived hydrochars with R^2 values above 0.98. The MDI-based feature importance computed from RF models indicated that the spectral regions influencing C wt% and O/C predictions differed from those for H/C. The high-importance regions helped improve model performance. These results suggest that the selective application of spectral regions, depending on the prediction target, might be a better strategy for prediction. In rapid and non-destructive prediction using NIR spectroscopy, the ensemble method of tree models is a promising technique and is comparable to chemometrics.

Abbreviations

CART	Classification and regression tree
DT	Decision tree
HTC	Hydrothermal carbonization
IPCC	Intergovernmental Panel on Climate Change
log2	Binary logarithm of all features
MDI	Mean decrease in impurity
MLP	Multilayer perceptron
MSE	Mean square error
NIR	Near-infrared
OOB	Out-of-bag
PLS	Partial least squares

R^2	Coefficient of determination
RF	Random forest
RMSE	Root mean square error
sqrt	Square root of all features

Acknowledgements
The authors would like to thank Editage (www.editage.co.kr) for English language editing.

Author contributions
SH was a major contributor to the study and wrote the manuscript. HC, TL, JK, YK, and JCK contributed to sample preparation and data analysis. HK, IC, and HY conceived the original ideas. HY supervised the study. All the authors have read and approved the final manuscript.

Funding
This study was supported by the Korea Forestry Promotion Institute through the R&D Program for Forest Science Technology funded by the Korea Forest Service (Project No. 2020215D10-2122-AC01).

Availability of data and materials
The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Competing interests
The authors declare no competing interests.

Received: 19 August 2022 Accepted: 11 December 2022
Published online: 05 January 2023

References

1. Masson-Delmotte V, Zhai P, Pörtner HO, Roberts D, Skea J, Shukla PR, Pirani A, Moufouma-Okia W, Péan C, Pidcock R, Connors S, Matthews JBR, Chen Y, Zhou X, Gomis M, Lonnoy E, Maycock T, Tignor M, Waterfield T (2018) Global Warming of 1.5°C in an IPCC Special Report on the Impacts of Global Warming of 1.5°C. Intergovernmental Panel on Climate Change
2. Atta-Obeng E, Dawson-Andoh B, Seehra MS, Geddam U, Poston J, Leisen J (2017) Physico-chemical characterization of carbons produced from technical lignin by sub-critical hydrothermal carbonization. Biomass Bioenerg 107:172–181. https://doi.org/10.1016/j.biombioe.2017.09.023
3. Borrero-López AM, Masson E, Celzard A, Fierro V (2018) Modelling the reactions of cellulose, hemicellulose, and lignin submitted to hydrothermal treatment. Ind Crops Prod 124:919–930. https://doi.org/10.1016/j.indcrop.2018.08.045
4. Davies G, El Sheikh A, Collett C, Yakub I, McGregor J (2021) Catalytic carbon materials from biomass. In: Sadjadi S (ed) Emerging carbon materials for catalysis. Elsevier, Amsterdam
5. Yoganandham ST, Sathyamoorthy G, Renuka RR (2020) Emerging extraction techniques: hydrothermal processing. In: Torres MD, Kraan S, Dominguez H (eds) Sustainable seaweed technologies. Elsevier, Amsterdam
6. Kang S, Li X, Fan J, Chang J (2012) Characterization of hydrochars produced by hydrothermal carbonization of lignin, cellulose, D-xylose, and wood meal. Ind Eng Chem Res 51:9023–9031. https://doi.org/10.1021/ie300565d
7. Wikberg H, Ohra-aho T, Pileidis F, Titirici M (2015) Structural and morphological changes in kraft lignin during hydrothermal carbonization. ACS Sustain Chem Eng 3:2737–2745. https://doi.org/10.1021/acssuschemeng.5b00925
8. Feng Q, Chen F, Wu H (2011) Preparation and characterization of a temperature-sensitive lignin-based hydrogel. Bioresour 6:4942–4952
9. Aro T, Fatehi P (2017) Production and application of lignosulfonates and sulfonated lignin. Chemsuschem 10:1861–1877. https://doi.org/10.1002/cssc.201700082
10. Luo H, Mahdi M, Abu-Omar M (2017) Chemicals from lignin. In: Abraham MA (ed) Encyclopedia of sustainable technologies. Elsevier, Amsterdam
11. Puziy AM, Poddubnaya OI, Sevastyanova O (2020) Carbon materials from technical lignins: recent advances. In: Serrano L, Luque R, Sels B (eds) Lignin chemistry. Topics in current chemistry collections. Springer, Cham
12. Hwang SW, Hwang UT, Jo K, Lee T, Park J, Kim JC, Kwak HY, Choi IG, Yeo H (2021) NIR-chemometric approaches for evaluating carbonization characteristics of hydrothermally carbonized lignin. Sci Rep 11:16979. https://doi.org/10.1038/s41598-021-96461-x
13. Raymond CA, Schimleck LR (2002) Development of near infrared reflectance analysis calibrations for estimating genetic parameters for cellulose content in *Eucalyptus globulus*. Can J For Res 32:170–176. https://doi.org/10.1139/x01-174
14. Via BK, Shupe TF, Groom LH, Stine M, So C (2003) Multivariate modelling of density, strength and stiffness from near infrared spectra for mature, juvenile and pith wood of longleaf pine (*Pinus palustris*). J Near Infrared Spectrosc 11:365–378. https://doi.org/10.1255/jnirs.388
15. Tsuchikawa S (2007) A review of recent near infrared research for wood and paper. Appl Spectrosc Rev 42:43–71. https://doi.org/10.1080/05704920601036707
16. Reza MT, Becker W, Sachsenheimer K, Mumme J (2014) Hydrothermal carbonization (HTC): near infrared spectroscopy and partial least-squares regression for determination of selective components in HTC solid and liquid products derived from maize silage. Bioresour Technol 161:91–101. https://doi.org/10.1016/j.biortech.2014.03.008
17. Horikawa Y, Imai T, Takada R, Watanabe T, Takabe K, Kobayashi Y, Sugiyama J (2011) Near-infrared chemometric approach to exhaustive analysis of rice straw pretreated for bioethanol conversion. Appl Biochem Biotechnol 164:194–203. https://doi.org/10.1007/s12010-010-9127-5
18. Horikawa Y, Mizuno-Tazuru S, Sugiyama J (2015) Near-infrared spectroscopy as a potential method for identification of anatomically similar Japanese diploxylons. J Wood Sci 61:251–261. https://doi.org/10.1007/s10086-015-1462-2
19. Hwang SW, Horikawa Y, Lee WH, Sugiyama J (2016) Identification of Pinus species related to historic architecture in Korea using NIR chemometric approaches. J Wood Sci 62:156–167. https://doi.org/10.1007/s10086-016-1540-0
20. Yang SY, Park Y, Chung H, Kim H, Park SY, Choi IG, Kwon O, Cho KC, Yeo H (2017) Partial least squares analysis on near-infrared absorbance spectra by air-dried specific gravity of major domestic softwood species. J Korean Wood Sci Technol 45:399–408. https://doi.org/10.5658/WOOD.2017.45.4.399
21. Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
22. Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecol 88:2783–2792. https://doi.org/10.1890/07-0539.1
23. Finch K, Espinoza E, Jones FA, Cronn R (2017) Source identification of western Oregon Douglas-fir wood cores using mass spectrometry and random forest classification. Appl Plant Sci 5:1600158. https://doi.org/10.3732/apps.1600158
24. Brieuc MSO, Waters CD, Drinan DP, Naish KA (2018) A practical introduction to Random Forest for genetic association studies in ecology and evolution. Mol Ecol Resour 18:755–766. https://doi.org/10.1111/1755-0998.12773
25. Hwang SW, Kobayashi K, Sugiyama J (2020) Evaluation of a model using local features and a codebook for wood identification. IOP Conf Ser Earth Environ Sci 415:012029. https://doi.org/10.1088/1755-1315/415/1/012029
26. Savitzky A, Golay MJE (1964) Smoothing and differentiation of data by simplified least squares procedures. Anal Chem 36:1627–1639
27. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Routledge, New York
28. Louppe G, Wehenkel L, Sutera A, Geurts P (2013) Understanding variable importances in forests of randomized trees. In: Burges CJ, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Proceedings of the 26th international conference on neural information processing system, vol. 1. Curran Associates Inc., New York, pp. 431–439.
29. Berge ND, Ro KS, Mao J, Flora JRV, Chappell MA, Bae S (2011) Hydrothermal carbonization of municipal waste streams. Environ Sci Technol 45:5696–5703. https://doi.org/10.1021/es2004528

30. Funke A, Ziegler F (2020) Hydrothermal carbonization of biomass: a summary and discussion of chemical mechanisms for process engineering. Bioprod Biorefin 4:160–177. https://doi.org/10.1002/bbb.198
31. Bakshi S, Banik C, Laird DA (2020) Estimating the organic oxygen content of biochar. Sci Rep 10:13082. https://doi.org/10.1038/s41598-020-69798-y
32. International biochar initiative (2015) Standardized product definition and product testing guidelines for biochar that is used in soil. https://www.biochar-international.org/wp-content/uploads/2018/04/IBI_Biochar_Standards_V2.1_Final.pdf. Accessed 19 Aug 2022.
33. Budai A, Zimmerman AR, Cowie AL, Webber JBW, Singh BP, Glaser B, Masiello CA, Andersson D, Shields F, Lehmann J, Camps Arbestain M, Williams M, Sohi S, Joseph S (2013) Biochar carbon stability test method: an assessment of methods to determine biochar carbon stability. Int Biochar Initiat. https://www.biochar-international.org/wp-content/uploads/2018/06/IBI_Report_Biochar_Stability_Test_Method_Final.pdf. Accessed 28 Nov 2022.
34. Bramer M (2007) Avoiding overfitting of decision trees. In: Bramer M (ed) Principles of data mining. Undergraduate topics in computer science. Springer, London
35. Michell AJ, Schimleck L (1996) NIR spectroscopy of woods from *Eucalyptus globulus*. Appita J 49:23–26
36. Schwanninger M, Rodrigues JC, Fackler K (2011) A review of band assignments in near infrared spectra of wood and wood components. J Near Infrared Spectrosc 19:287–308
37. Kirtania K (2018) Thermochemical conversion processes for waste biorefinery. In: Bhaskar T, Pandey A, Mohan SV, Lee DJ, Khanal SK (eds) Waste biorefinery. Potential and perspectives. Elsevier, Amsterdam
38. Fackler K, Schwanninger M (2010) Polysaccharide degradation and lignin modification during brown rot of spruce wood: a polarised Fourier transform near infrared study. J Near Infrared Spectrosc 18:403–416

Publisher's Note