Cellulose (2023) 30:1873–1893
https://doi.org/10.1007/s10570-022-04921-y

ORIGINAL RESEARCH

Bamboo phase quantification using thermogravimetric analysis: deconvolution and machine learning

Fabrício de Campos Vitorino · Michael Nazarkovsky · Arash Azadeh · Camila Martins · Bruno Menezes da Cunha Gomes · Jo Dweck · Romildo Dias Toledo Filho · Holmer Savastano Jr.

Received: 10 May 2022 / Accepted: 27 October 2022 / Published online: 9 December 2022

Abstract The focus of this paper is to provide a fast and reliable quantification method of bamboo's main chemical components. Therefore, thermogravimetric analysis was used to determine holocellulose and lignin content in different bamboo specimens. The influence of nitrogen vs. air atmospheres was investigated on the thermal degradation behavior of *Phyllostachys edulis* (Moso), *Bambusa vulgaris* (BV) and *Iranian Phyllostachys* (IR) bamboos. Due to peaks overlapping, the deconvolution process was carried out to resolve hidden peaks and to allow adequate phase quantification. Also, a set of machine learning (ML) algorithms was applied to predict the composition of the studied bamboos within the 200–500 °C range in their TGA-DTG profiles. The ensembles of the ML models at $R^2 > 0.99$ proved a connection between the features in thermogravimetric curves with two concentrations of the main components, which were preliminarily established by means of chemical extraction from the respective samples.

Keywords Bamboo · Thermogravimetric analysis · Machine learning · Phase quantification

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10570-022-04921-y.

F. de Campos Vitorino · C. Martins · B. M. da Cunha Gomes · J. Dweck · R. D. T. Filho
Federal University of Rio de Janeiro, 149 Athos da Silveira Ramos Av., Rio de Janeiro, RJ 21941-909, Brazil

M. Nazarkovsky (✉)
Chemistry Department, Pontifical Catholic University of Rio de Janeiro, 225 Marquês de São Vicente Str., Rio de Janeiro, RJ 22451-900, Brazil
e-mail: nazarkovsky_m@esp.puc-rio.br

A. Azadeh · H. Savastano Jr.
Faculty of Animal Science and Food Engineering, Department of Biosystems Engineering, University of São Paulo (USP), 225 Duque de Caxias Norte Av., Pirassununga, SP 13635900, Brazil

Introduction

Bamboo is a multi-functional material in its multiple scales with outstanding mechanical and chemical performance. Due to this, it has been used in diverse areas, such as civil engineering, construction material, textile industry, semiconductor materiails (Pandoli et al. 2020), and handicrafts (Valani et al. 2020). Such multi-functionality is a result of a complex chemical and physical composition network. Its main constituents are holocellulose (α-cellulose + hemicellulose) and lignin. In small amounts, there are also starch, free water, volatile extractives and waxes. Cellulose and lignin form a matrix which dictates the mechanical properties of each specific bamboo (Youssefian and Rahbar 2015).

Its components can get varied according to the bamboo species and region harvested. In addition, the holocellulose and lignin is significantly variable

in their chemical constitution depending also on these two factors (Dumitriu 2004; Shen et al. 2013a, b). α-Cellulose, for instance, is a macromolecule containing linear polymerized glucopyranose and its polymerization degree can vary according to its beta 1,4- glycosidic links. Lignin is also a macromolecule polymerized from alcohols of para-hydroxy cinnamic acid (Dumitriu 2004; Shen et al. 2013a). Due to their nature, both of them, when submitted to thermal-degradation get burned to char (Dumitriu 2004; Carrier et al. 2011; Shen et al. 2013b; Zakikhani et al. 2016; Ornaghi et al. 2020). In the TGA analysis, depending on the atmosphere used, char can be decomposed at temperatures above 1000 °C (inert atmosphere) or below 600 °C in the case of oxidative (air) atmosphere (Valani et al.; Shen et al. 2013a, b; Zakikhani et al. 2016).

There are various bamboo species in Brazil but one of them which can be encountered commercially is the *Phyllostachys edulis* or Moso bamboo. The reason to focus more on this bamboo is its proper physical and mechanical characteristics which made it a suitable material for the construction industry as a structural element, laminates and fiber reinforced concrete (Chung and Wang 2018; Kadivar et al. 2019; Wang et al. 2020). In addition, as a matter of comparison, two other bamboo types were analyzed, *Bambusa vulgaris* (BV) and *Iranian Phyllostachys* (IR).

The quantification of the bamboo's main constituents is a time-consuming task. Chemical extraction is the principal technique used for this purpose owing to its reliablity and cost-effectiveness. However, it may take days to complete all analytical procedures (dos Santos Abreu et al. 2006; Michael Buchanan 2007). Other researchers attempt to perform quantifications by means of thermogravimetric analysis (TGA), however, two main problems can arise: peaks overlapping and char formation, which leads to questionable results (Ramiah 1970; Brebu and Vasile 2010a; Carrier et al. 2011; Cao et al. 2019).

Thus, both, afore-mentioned problems can be overcome with the help of the peaks deconvolution and exposure of samples to an oxidative atmosphere (air, for instance) during the analysis. Therefore, in this paper, a stepwise protocol of a deconvolution process is developed to quantify the main components of bamboo. The influence of oxidative (synthetic air) and inert (nitrogen) atmospheres in the course of thermogravimetric analysis was also studied to access the amount of char formation in each analysis. Scanning electron microscopy (SEM) coupled with elemental analysis (EDX) was applied to confirm the presence of char after TGA.

Also, in the present study, thermogravimetric experimental results were analyzed in terms of the 4th paradigm of knowledge, i.e. data science. This is an up-to-date concept which is associated with predictive modeling using available data generated during human activity – in our case, by natural scientists or experimentalists (Agrawal and Choudhary 2016; Himanen et al. 2019; Zhou et al. 2019; Bezerra et al. 2020; Gressling 2020). The tools of data science, in particular, machine learning (ML) algorithms are helpful to develop applications, software, application programming interfaces (APIs) from chemical, instrumental or any technical information to predict or identify the values (regression task) or specify types of the samples/materials from their properties (classification task). Herewith, we took the most specific regions of TGA-DTG curves for each of three bamboo samples to predict their composition by two main components: lignin and holocellulose. Such a strategy to isolate the mentioned regions was prompted by reduction of non-informative flat lapses identical for all three samples of bamboo. It is a necessary stage of data cleaning and preparation. The methods included unsupervised (K-Means Clustering) and supervised (K-Nearest Neighbors—KNN, Decision Tree, Bootstrap Forest, XGBoost) ML methods. This ML study is promising for construction, biology, and materials science in general: using a small amount of a sample to perform the thermal analysis one can receive information about a bamboo's type and the percentage of its main fractions. Due to the non-linearity of the curves, methods like SVM (Support Vector Machine), Naive Bayes, or Linear Regression are impracticable and will result in worse values of the metrics.

With the use of these advanced techniques, deconvolution process and machine learning, this paper brings a fast and reliable method to quantify main bamboo's phase constituents, cellulose and lignin. It also paves the road in to the building of applications to identify different bamboo species by means of thermal analysis and other properties.

Materials and methods

Materials

Three different bamboo types *Phyllostachys edulis* (Moso bamboo), *Bambusa vulgaris* (BV), and *Iranian Phyllostachys* (IR) were used in the present study (plain bamboos). The Moso bamboo culms were provided by 'Takê Cortinas' (São Paulo, Brazil); the BV bamboo culms were obtained from PUC-Rio University (Rio de Janeiro, Brazil); and the IR bamboo samples were received from the Tea Research Center (Lahijan, Iran). All three species were cut between 3 and 5 years of age. After cutting, the bamboo poles were stored vertically in a shadow for 6 months. The samples have been prepared from the middle part of the culm for all three species.

Chemical extraction methods

Chemical extraction methods and quantification of α-cellulose, hemicellulose, and lignin of the Moso bamboo followed recommendations provided in the reference (dos Santos Abreu et al. 2006). In the case of BV and IR bamboos, it was carried out according to standard recommendations of TAPPI.

The extracted holocellulose (α-cellulose + hemicellulose) and lignin were carefully sampled in sealed plastic bags in order to avoid contaminations, and later they were characterized by means of TGA.

Thermogravimetric analysis

Thermogravimetric tests were conducted using the SDT Q600 equipment (TA Instruments). Plain bamboo samples were milled in a knife mill grinder. About 10 mg of sample was poured into the platinum pan to perform the analysis. The analysis conditions were as follows: a heating rate of 10 ºC·min^{-1} from 25 to 1000 ºC. under either 100 mL·min^{-1} of nitrogen or synthetic air flow. Tests were made in triplicates.

Coke quantification

For the calculation of the coke formation under nitrogen atmosphere, the TGA curves were subtracted from synthetic air curves by using OriginPro® software version 2019b, according to Eq. 1.

$$M_{coke,i} = M_{N,i} - M_{Syn,i} \quad (1)$$

where: $M_{coke,i}$ is the mass of coke formed at a given i temperature in nitrogen flow analysis;

$M_{N,i}$ is the mass loss % of bamboo at a given i temperature in nitrogen flow analysis;

$M_{Syn,i}$ é the mass loss % of bamboo at a given i temperature in synthetic airflow.

Deconvolution process algorithm

In order to resolve overlapping peaks in the TGA analysis, a typical deconvolution process had to be carried out to calculate the areas related to the holocellulose and lignin. Deconvolutions were undertaken using OriginPro® software version 2019b. The data range was limited from temperature above 100 °C. At lower temperatures, the mass loss can be related to water and other volatile phases, such as extractives, which are not on the focus of the present study. The baseline was built using the Straight Line built-in option and it was kept fixed during fitting process. In order to reduce overfitting, a minimum number possible of peaks were used to model experimental data using the Gaussian peak function. The maximum number of iterations was set at 500 (tolerance to 1×10^{-6}).

First, deconvolution of the Moso's extracted holocellulose and lignin was performed and its peaks parameter properties, such as the peak center, the peak area, and FWHM were later used as parameters for restrictions during the deconvolution process of plain bamboos' peaks. Restrictions values were set with 10% of tolerated variance to each peak parameter property, except for the peak area. The peak area was allowed to vary from values ≥ 0. The fitting algorithm comprised the following steps:

1. Initially, the fitting process was set to run as a first fitting approximation.
2. Then, fitting peaks that were crossing experimental data peaks must have their restriction limits reduced. First, by increasing or reducing FWHM by more 10%. Second, the peak area was limited to a minimum value that inhibits fitting peak to cross experimental data peak.
3. Fittings were considered acceptable when modeled curve visually best described experimental data and at $R^2 > 0.90$.

In the cases of $R^2 < 0.90$ while using only the fitted peaks found in the extracted holocellulose and lignin process, an extra peak had to be inserted between 345 and 460 °C with no restrictions, except, at non-negative area and FWHM values > 0. After that, the fitting process was restarted from step 1, and acceptable R^2 values were found.

Scanning electron microscopy analysis

The ashes collected after thermogravimetric analysis were studied by scanning electron microscopy (SEM). The ashes were fixed on the equipment support by means of carbon tape without additional coating. Images were obtained with a Quanta 400 equipment (FEI, Brno, Czech Republic) with secondary and backscattered electron detectors. It was operated under low vacuum and at the voltage of 15 kV. Elemental composition was determined by means of an Energy-Dispersive X-ray (EDS) accessory (Bruker Nano GmbH).

Machine learning techniques

In order to predict the composition (lignin or holocellulose content, %) in each bamboo species in consideration and to connect features in TGA-DTG with the chemical extraction data, the set of trained and validated models was proposed: K Nearest Neighbors (KNN, Euclidean distances, $K_{max} = 100$ neighbors for both phases), Decision Tree (307 splits for lignin and 277 splits for holocellulose), Bootstrap Forest (learning rate 0.1, 100 trees, 2 terms per split, the minimal splits per tree—10, the minimal split size—32 at the fixed random seed) and XGBoost (learning rate ≤ 0.1, the maximal depth 8, α—1.3615, λ—1.0618 at 391 iterations). The ensembles of the models were arranged through gradient boosting using sigmoid functions. Given the specific patterns of the thermogravimetric profiles, and characteristic concentrations of two main components (lignin and holocellulose), three samples—BV, IR, Moso—can be classified by K-means clustering. This approach reasons further involvement of supervised machine learning shown below. The entire procedure was coded in JSL within a module of JMP Pro15 (SAS).

Results and discussions

In the following subsections the comparison between TGA analyses carried out under different gas flow atmosphere (inert and oxidative) is presented. Next, is possible to check, the chemical quantification of holocellulose and lignin from different bamboos and also the TGA quantification by means of deconvolution and machine learning.

Char formation during TGA analysis characterization of bamboo

Figure 1 shows the TGA and DTG curves in nitrogen and synthetic air flow for Moso bamboo. It is also possible to see the difference curve between air and nitrogen TGA curves. The difference curve represents the char formation in nitrogen flow analysis.

It can be seen that the atmosphere of thermogravimetric analysis essentially affects the path of bamboo's decomposition. In the case of the oxidative airflow, big macromolecules in the bamboo constituents, such as celluloses and lignin yield char, whereas, in the inert atmosphere, these components will be decomposed above 1000 °C (Brebu and Vasile 2010b; Carrier et al. 2011; Shen et al. 2013b, a; Li et al. 2015). On the other hand, in the oxidative atmosphere, char decomposes at lower temperatures, which is proven by a peak at 468 °C. It is worth

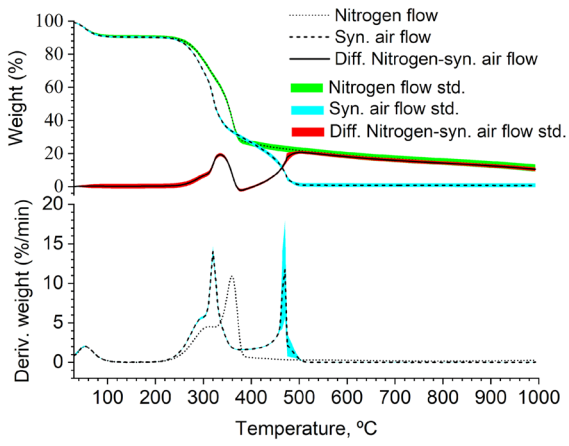

Fig. 1 TGA and DTG results in nitrogen and synthetic airflow of Moso. The colored area is the standard deviation

pointing out, at 1000 °C there is less than 1% mass in the oxidative analysis, which can be attributed to inorganic ashes. While for inert atmosphere analysis it remained 10.2% of mass, which can be attributed to organic char presence plus inorganic ashes.

There are two stages of char formation: first, starting from 240 up to 380 °C and the second one - at 380 °C continuing over 1000 °C. In the first stage, char had the maximal yield at 340 °C (19% by mass) and it was completely decomposed at 380 °C. It can be related to char formed from cellulose or hemicellulose pyrolysis (Cao et al. 2019). In the second stage, the char starts to form at 380 °C, reaching the a maximum yield at 500 °C (20.7% by mass), and then it starts to decrease progressively, although, at 1000 °C the char is still present (10.5% by mass). The second stage can be attributed to the pyrolysis of lignin and it can be associated with its high molecular weight and complex structure (Cao et al. 2019).

Figure 2 and Fig. 3 present SEM images after TGA analysis (1000 °C residue) using syn. air or nitrogen gas flow and, respectively, the representative point EDS analysis of the structures seen in the SEM images.

From Fig. 2, in the oxidative atmosphere, only simple ashes structures remained. The point EDS analysis revealed the composition represented by Mn, Ca, K, Cl, S, P, Si, Mg, Na, O, and C. Carbon may appear due to the carbon tape, used to fix the sample during preparation. That means that all organic compounds were degraded during thermogravimetric analysis and no char was formed.

On the other hand, Fig. 3 illustrates complex structures after TGA analysis in inert atmosphere and that confirms the char residue after 1000 °C. It is composed of a heterogeneous surface and apparent high porosity with circular voids, as can be seen in expanded image. From point EDS analysis it is possible to identify that the carbon peak is too height that overlaps the presence of the other peaks that were seen in oxidative analysis, therefore, confirming its carbonaceous composition.

Taking into account that in the inert atmosphere there is char formation with these complex structures remaining after 1000 °C and in oxidative atmosphere only simple oxide ashes remains, attempts of lignin and cellulose quantifications will be carried out using synthetic air atmosphere. since it guarantees complete thermal degradation of these organic constituents at the set temperature.

Lignocellulose phases quantification of bamboo

Extraction method

Table 1 shows the subject bamboos phase quantification by means of the extraction method. It is noticeable that different extraction methods was executed for Moso bamboo, which allowed the quantification of α-cellulose and hemicellulose. It is possible to see that the Moso bamboo is composed of 37.6% of α-cellulose and 24.6% of hemicellulose in this research. Holocellulose (α-cellulose + hemicellulose) is present in 62.2%, 65.3% and 72.3% of the Moso, BV and IR mass fractions respectively. After holocellulose, lignin is the second main component, its content takes 30.6%, 29.5% and 25.5%, for the same bamboos respectively. The Extractives are present in smaller fractions, reaching 7.2% of the Moso, BV and IR mass respectively. Lastly, the amount of extractives was subtracted from that of holocellulose and lignin which gave respective amounts of 67.0% and 33.0% for Moso; 68.9% and 31.1% for BV; and 73.9% and 26.1%, for IR. Such a normalization was required because the extractives are mostly volatile and their concentrations can significantly vary depending on storage conditions and specimens' age. These two subject components served to be matter for discussions about the thermogravimetric profiles.

TGA method

Figure 4 *a* shows TGA mean results (mean of the triplicate) of Moso's extracted holocellulose and lignin. Regarding to the holocellulose, two sharp peaks of degradation can be assigned: around 320 °C and near 475 °C. It is noticeablethat three peaks appear in the 450 and 525 °C region, however, for each individual replicated curve there is only one peak at different temperature maxima. Hence, the mean results curve only describes this non-uniform behavior of these peaks appearance at different temperatures. On the other hand, lignin degrades continuously, what two main wide peaks speak for - at 300 °C and 525 °C.

The modeleld curves of the Moso's holocellulose and lignin TGA experimental data are represented in Fig. 4 b and c. The peaks deconvolutions were

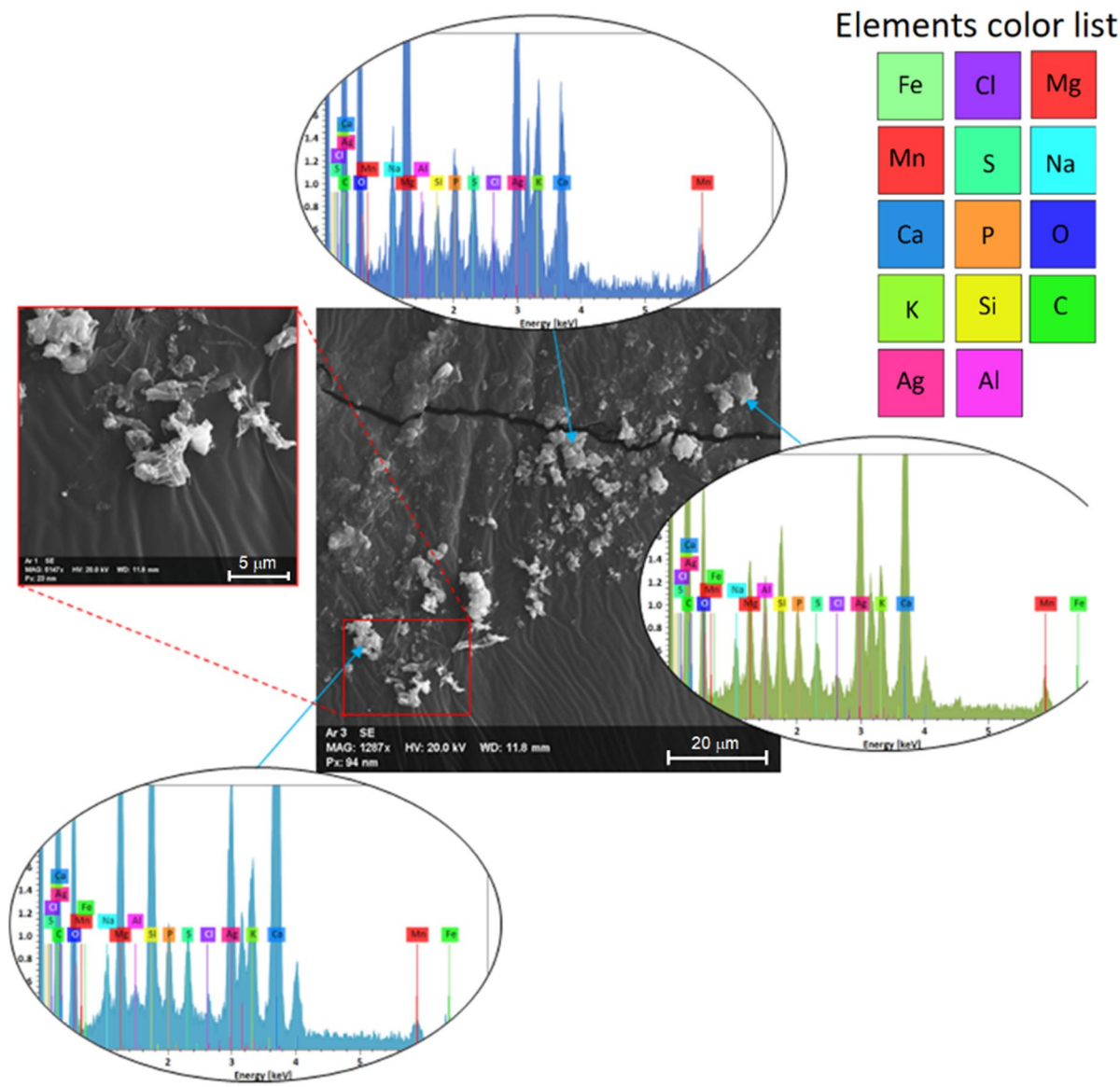

Fig. 2 The SEM images and the point EDS after sythetic airflow analysis of bamboo (a total of 8 points were acquired)

carried out after 100 °C, because at < 100 °C, the mass loss can be addressed to water and more volatil extractive phases which have smaller contribution on mass loss of the samples. It was possible to model holocellulose TGA curve using four peaks with a $R^2 = 0.97$. The residual curve shows small differences between modeled and experimental curves. In the case of lignin, it was possible to describe experimental results with only three peaks ($R^2 = 0.99$). The residual curve also displays a non-significant difference between modeled and experimental profiles. Table 2 summarizes the peak properties of deconvoluted TGA curves for holocellulose (in grey) and lignin (in blue). Thereupon, the peak center and FWHM properties of holocellulose and lignin was used as restrictions to model plain Moso bamboo's TGA curve with a variation tolerance of 10%.

Figure 5a demonstrates TGA profile of plain Moso bamboo. The deconvolution was also conducted after 100 °C for the same reasons mentioned above. The

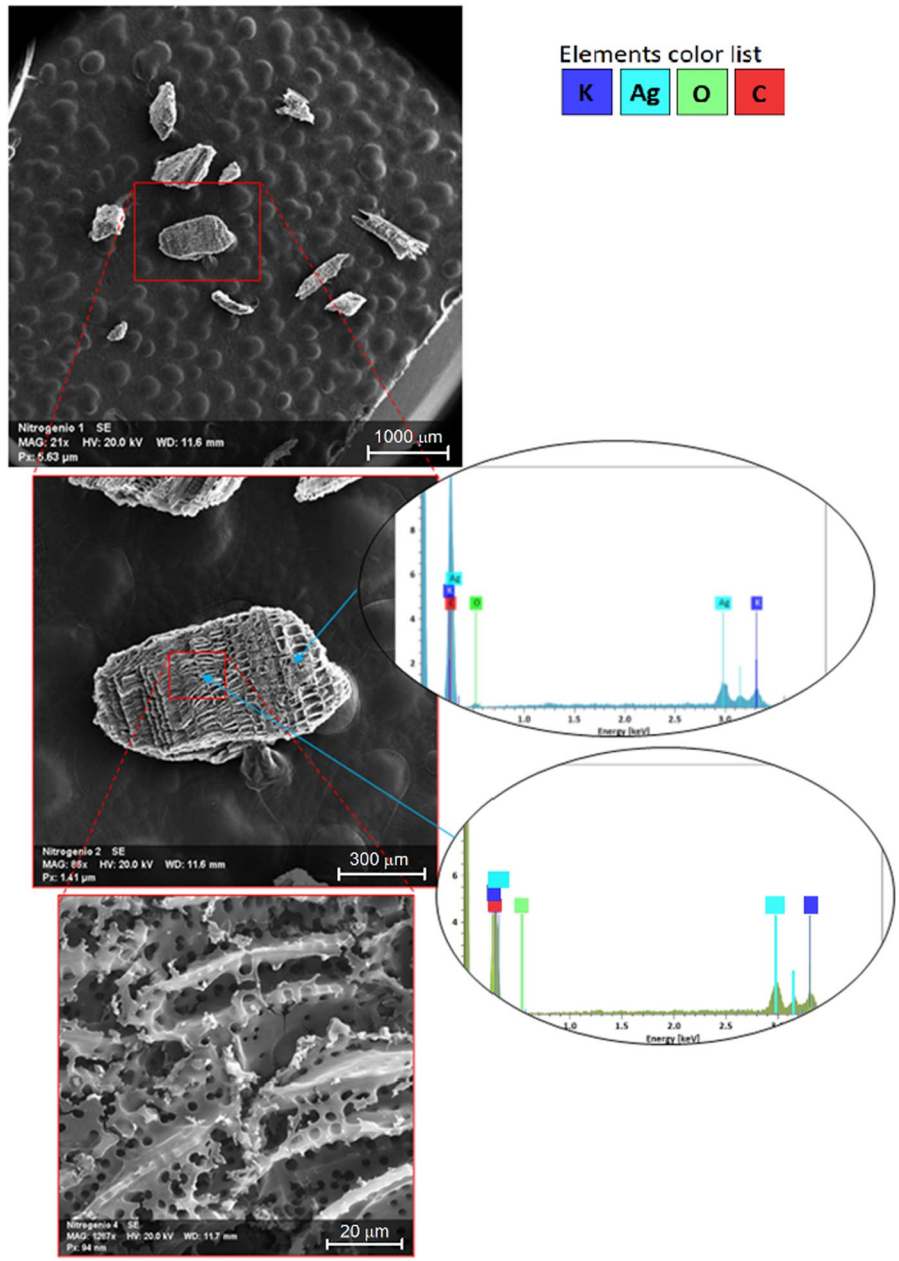

Fig. 3 The SEM images and the point EDS after nitrogen flow analysis of bamboo (a total of 8 points were acquired)

Table 1 Mean values and relative standard deviation (*RSD*) values of lignocellulosic phases composition of bamboos

Bamboo type	Extractives	Lignin	α-Cellulose	Hemicellulose	Holocellulose
Moso	7.2 (5.1)	30.6 (7.7)	37.6 (3.4)	24.6 (3.8)	62.2 (3.3)
BV*	5.5	29.5	–	–	65.3
IR*	2.4	25.5	–	–	72.3

*Single-replicate experiment

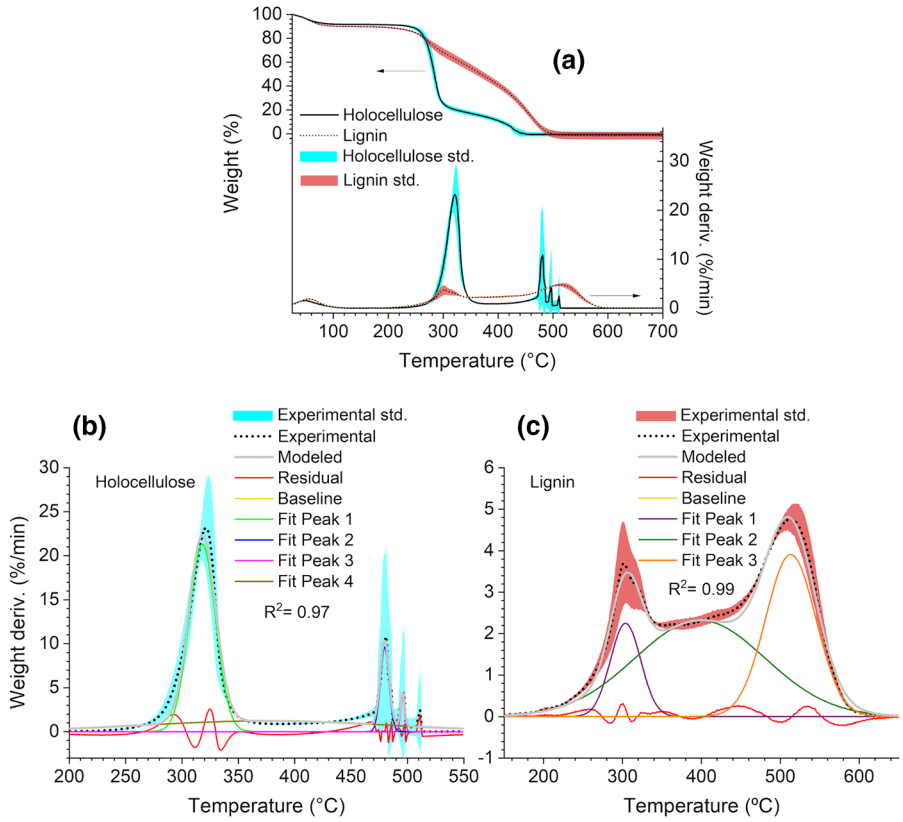

Fig. 4 The TGA analysis of extracted holocellulose and lignin **a**; Deconvoluted mean peaks of holocellulose **b** and lignin **c**. Blue and red areas are standard deviations of experimental data

Table 2 The properties of deconvoluted holocellulose and lignin TGA peaks

	Peak center (°C)	Area (m%)	FWHM (°C)
Holocellulose	318.1	60.7	27.7
	374.4	28.6	255.8
	479.8	9.0	9.1
	495.7	1.7	4.2
Lignin	303.9	12.0	45.1
	398.2	53.3	194.4
	513.2	34.7	74.9

first attempt to model experimental data using only the peaks found for holocellulose and lignin was undertaken. In order to improve the model's performance, due to the lack of the peak an additional peak was inserted in the region of 425–475 °C with no restriction and keeping the non-negative area. Finally, it was possible to describe experimental data with eight peaks ($R^2 = 0.95$). The residual curve revealed the non-significant difference between modeled and experimental data.

Table 2 shows the peak properties of the Moso's deconvoluted TGA profile . It is noticeable that the last peak is a free peak added for better description of the experimental data. Later, the sum of areas was compared to the results from the extraction method and it was possible to assign the last peak to holocellulose. The sum of deconvoluted areas resulted in 61.7% of holocellulose and 38.3% of lignin. As compared to the values found in the extraction method the differences were 7.9% and 16% for holocellulose and lignin, respectively.

After that, by using the identical restriction criteria for the Moso bamboo, attempts to deconvolute the TGA curves of *Bambusa vulgaris* bamboo (BV)

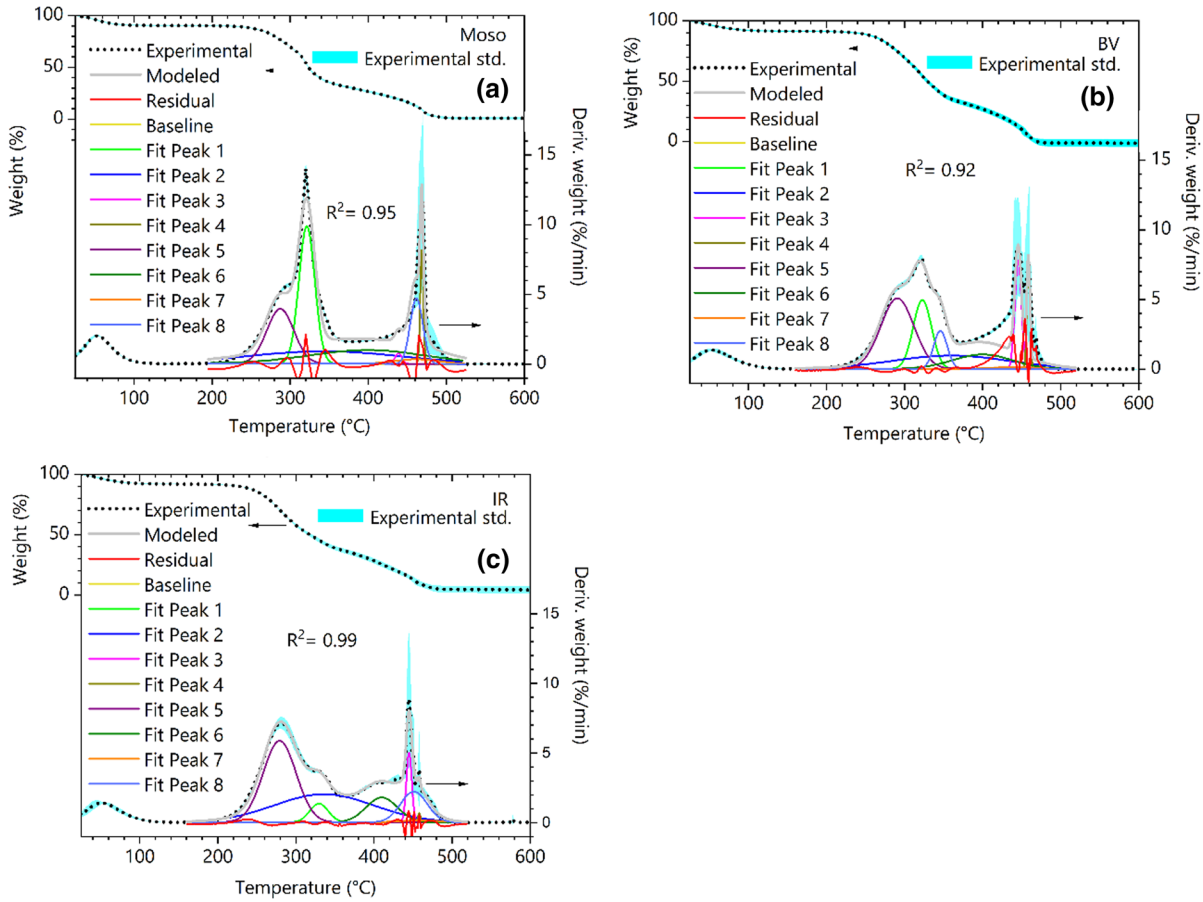

Fig. 5 The TGA and deconvoluted DTG curves of Moso bamboo **b**(*a*); Bambusa Vulgaris, BV (**b**) and Iranian Phyllostachys, IR(**c**) bamboos. Blue areas are corresponding to the standard deviation of experimental data

Table 3 The peak properties of the deconvoluted TGA profiles of Moso, BV and IR bamboos: for holocellulose (gray), for lignin (blue) related peaks and for extra peaks (white)

Moso			BV			IR		
Peak center (°C)	**Area (m%)**	**FWHM (°C)**	**Peak center (°C)**	**Area (m%)**	**FWHM (°C)**	**Peak center (°C)**	**Area (m%)**	**FWHM (°C)**
321.5	26.2	24.0	322.8	16.6	27.7	329.6	4.4	27.9
344.8	20.6	230.0	360.0	20.0	175.4	336.0	35.4	150.0
438.8	0.8	10.0	445.5	9.4	10.0	445.0	4.7	8.1
468.7	4.2	4.6	458.6	4.1	4.7	457.9	0.3	3.8
287.5	17.7	40.6	290.8	30.6	50.0	279.2	33.9	49.6
400.0	18.2	175.0	400.0	11.4	90.0	410.0	9.9	47.2
461.9	2.5	67.0	440.0	1.1	67.0	461.0	1.1	67.0
461.9	9.9	19.1	345.9	6.7	20.3	451.1	10.3	40.0

and Iranian *Phyllostachys bamboo* (IR) were made (Fig. 5b,c).

For the case of the BV bamboo, as well as for Moso, an extra peak was required to describe experimental data, however, it was introduced at approx. 345 °C. Another peak was suggested to be added near 430 °C, but in order to avoid overfitting, it was not used. Even though the residual result was high in the region of 400–480 °C, the modeled curve resulted in a good fit ($R^2 = 0.92$). The properties of deconvoluted BV bamboo peaks are summarized in Table 3. The sum of areas related to holocellulose, including the extra peak, was 56.9% and for lignin – 43.1%.

In the case of the IR bamboo, also an extra peak at 451 °C was added. The final modeled curve was fit at $R^2 = 0.99$. The properties of deconvoluted IR bamboo peaks are summarized in Table 3. The sum of the areas related to holocellulose (gray) together with the extra peak (white) was 55.1% and for lignin (blue) was 44.9%. Compared to the extraction method for the same bamboo, the differences for holocellulose and lignin were 34.1 and 41.9%, respectively.

There is a remarkable mismatch between the results found by TGA and extraction methods. Such a difference can be addressed to the distinct decomposition paths of holocellulose or lignin in each individual type of bamboo. In the present study, it was used the same constituents (extracted holocellulose and lignin) of Moso bamboo to make restrictions and to model BV or IR bamboos in the deconvolution process. However, it is possible to explore and establish an appropriate connection between the TGA profiles and the concentrations of each bamboo specie's extracted constituents by applying the machine learning strategy.

Machine learning techniques

Data split and stratification To optimize the classification process, the most informative ("feature") part of the TGA-DTG curves was taken on oxidative (air) atmosphere: within 200 – 500 °C. This region comprises all necessary features to distinguish the output: either the sample type or the components concentration.

The experimental points were split into training (70%) and validation (30%) sections (Fig. 6) stratified by each phase separately. As the KPIs (key performance indicators), R^2 (coefficient of determination), RMSE (the root-mean-square-error, the difference between the training and validation metric), AAE (average absolute error) were suggested to compare the results. Also, the analysis the over- or underfitting, the difference between the training and validation metric (RMSE) was emphasized for the final conclusion.

Supervised machine learning research Comparing the KNN metrics, the better fit for lignin seems quite clear, its lowest RMSE at K = 1 is twice less than for holocellulose (K = 2). Moreover, from the whole range of the neighbors' distances (the maximal distance 0.173), 97% of neighbors are located within 0.015, whereas 98% for holocellulose occupies a higher distance range – up to 0.03 for holocellulose (the maximal distance 0.192) – Fig. 7, Fig.S1.

The actual by predicted plots favor the lower RMSE for lignin, which is reflected in less dispersed points for each set concentration (Figs. 8,9). Although, correctly predicted values as for lignin and holocellulose have not less than 98% of the population – the respective metrics are the following: $R^2 = 0.9892$, RMSE = 0.2277 and $R^2 = 0.9879$,

Fig. 6 The mosaic plot of training/validation split stratified by lignin or holocellulose phase for three thermogravimetric parameters (derived weight, weight, temperature)

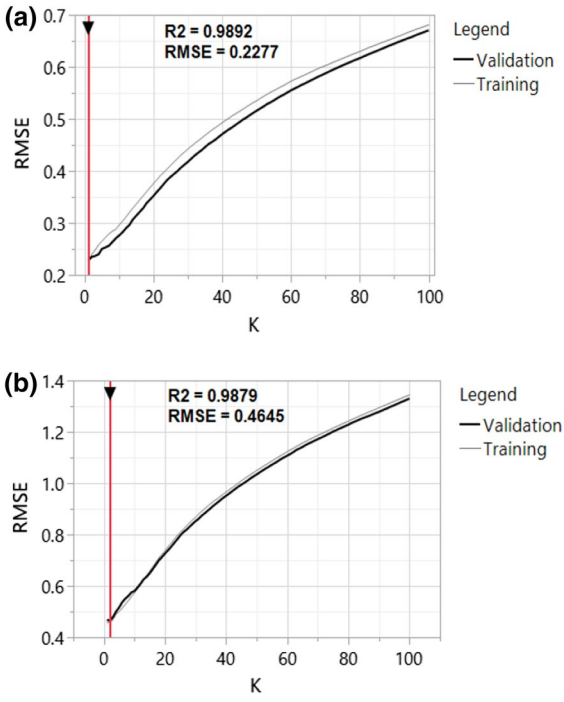

Fig. 7 KNN model selection charts for lignin **a** and holocellulose **b**

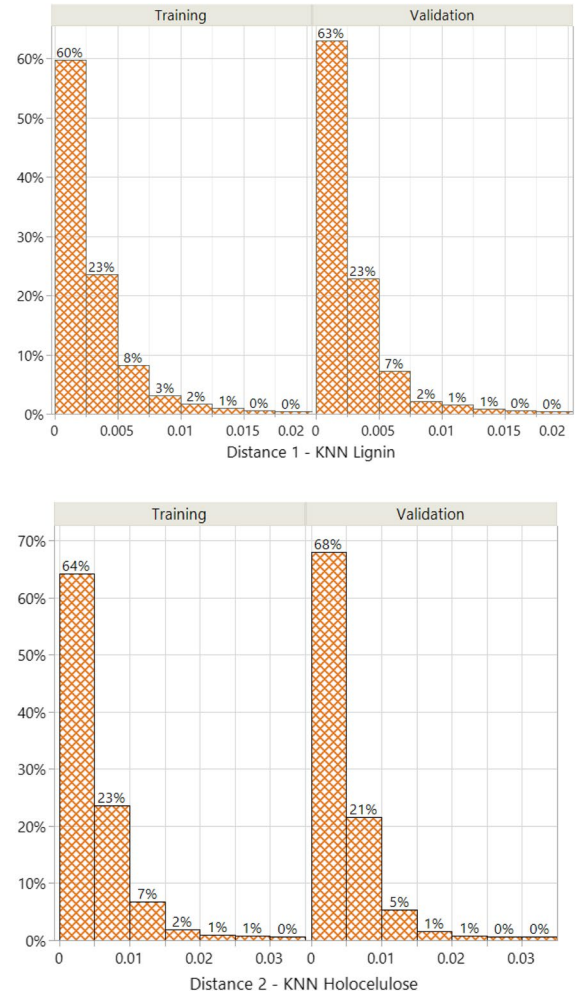

Fig. 8 The neighbors distances distributions for lignin **a** and holocellulose

RMSE = 0.4645. Hence, KNN demonstrates quite strong predictive power for lignin and holocellulose. This is due to easily distinguishable positions of the key experimental points on TGA-DTG. Thus, plotting fragments or a group of points in different positions after thermogravimetric studies under identical conditions, the algorithm allows effectively "recognize" the composition by these two natural components.

Decision Tree and Bootstrap Forest continue with lower RMSEs for lignin rather than for holocellulose despite equally good fits for both models – it is normal because of higher concentration range of holocellulose. Decision Tree has got different numbers of splits for lignin and holocellulose: 307 and 277, respectively (Fig. 10). The main drawback, quite visible from the metrics is the huge difference between RMSEs in validation and training sets (Figure S2, Fig. 11). This may serve as a criterion to set aside this model due to a high risk of overfitting.

Bootstrap Forest exhibits less intensive main patterns, which are compensated, however, by erroneous values at lower densities, than for Decision Tree. Both Decision Tree and Bootstrap Forest take weight, derived weight, and temperature at almost equal proportions which means, in turn, equal power of each variable's contribution to the models in the context of both bamboo phases (Tables S1-4). R^2 does not vary after the number of trees = 20 for lignin or holocellulose (Fig. 12) and, like in the case of Decision Tree, the values of standard deviations for holocellulose are larger than for lignin (Fig. 13). Also, as consisted with the distribution histograms for Bootstrap Forest, the main modes are more spread than for Decision Tree and, hence, the model's exactness may suffer, which will be a subject of discussions in a few paragraphs below (Figure S3).

XGBoost has turned out to be more effective than Decision Tree or Bootstrap Forest – its RMSEs for both bamboo phases are correspondingly lower after

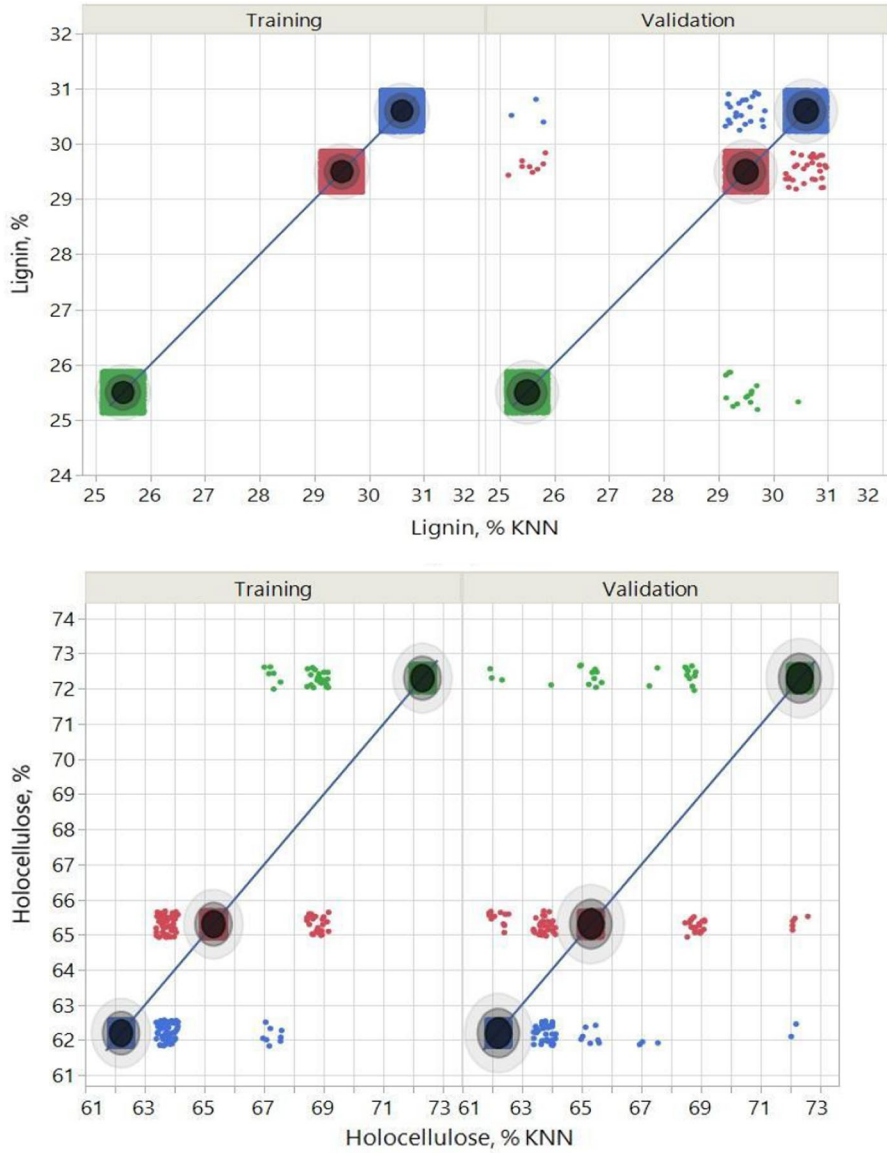

Fig. 9 The actual by KNN-predicted plots for lignin and holocellulose

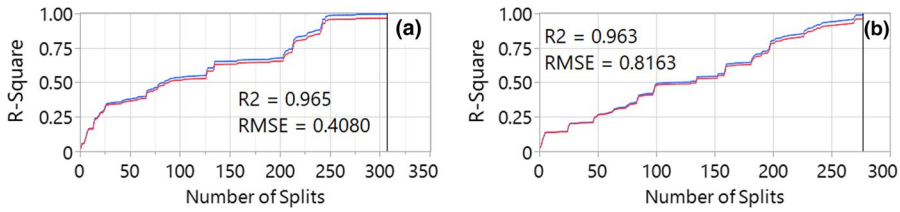

Fig. 10 The split history of Decision Tree for lignin **a** and holocellulose **b**. Blue – training, red – validation

hundreds of iterations (Fig. 14). As opposed to lignin, holocellulose has a larger deviation in the predicted values, however, RMSE in XGBoost is less than in Decision Tree for both phases (Figure S4, Fig. 15) – 0.2549 and 0.5380, respectively. The contribution of weight is 1.6 times higher than another two

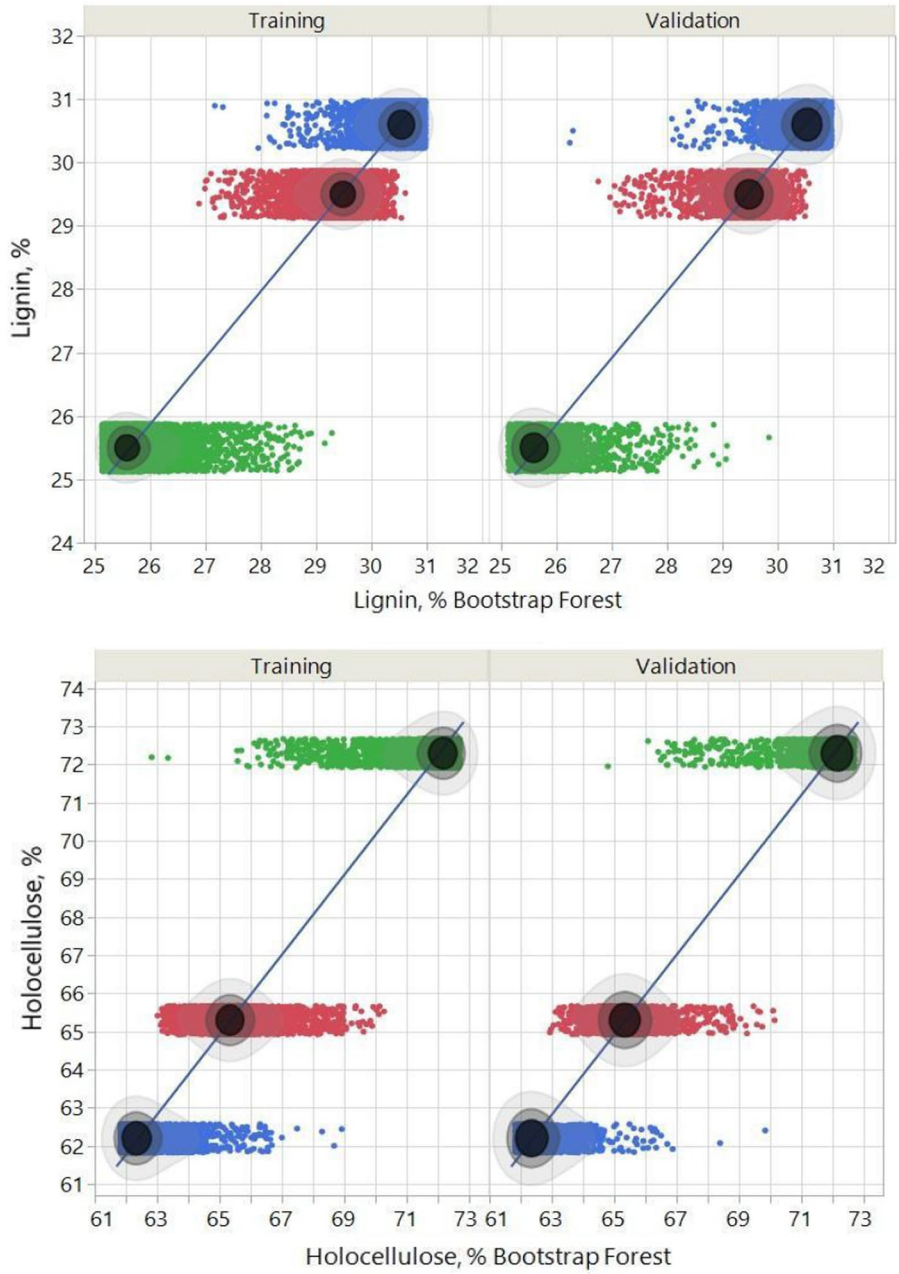

Fig. 11 The actual by Decision Tree-predicted plots for lignin and holocellulose

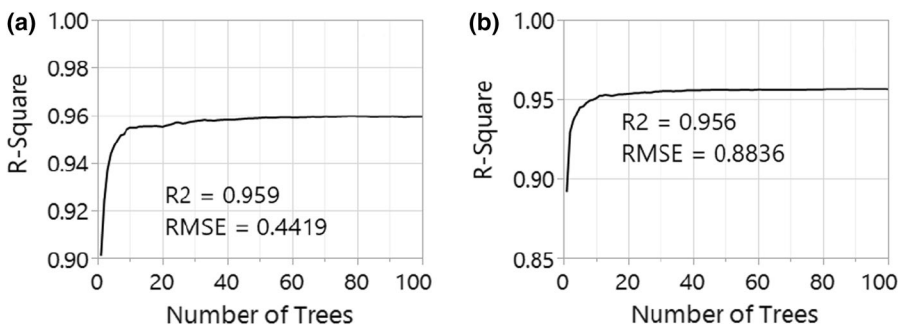

Fig. 12 Progress of R^2 with bootstrapping of the trees for lignin **a** and holocellulose **b**

parameters (temperature and derived weight) – Tables S5, S6. Thus, boosting itself may be a solution to keep the balance between the overfitting and exactness of the model. This is why, we undertook ensemble with a gradient boosting in a neural network from developed single models.

Ensemble of XGBoost and Bootstrap Forest – boosted single-layered Neural Network (50-times boosted of single (for lignin) and three (for holocellulose) sigmoid (hyperbolic tangent) functions, the learning rate = 0.09, the square penalty method at a single tour). Another difference lies in the involved models: XGBoost and Bootstrap Forest – for lignin; XGBoost, KNN, Decision Tree, and Bootstrap Forest – for holocellulose (Figs. 16, 17).

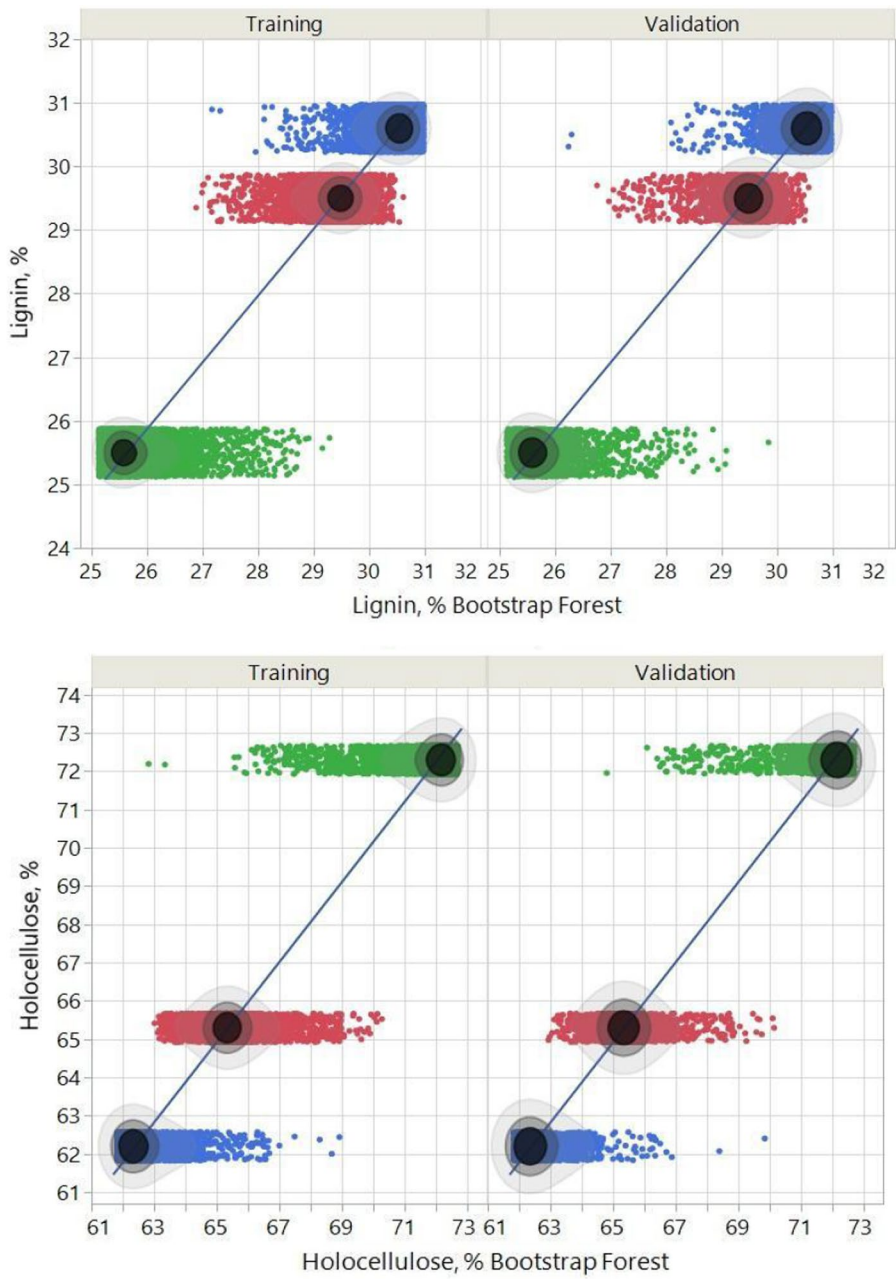

Fig. 13 The actual by Bootstrap Forest-predicted plots for lignin and holocellulose

As a result, the metrics of the boosted ensembles have essentially improved and their lower spread of values (points) is evidenced on the actual by predicted plots for both phases: the standard deviations for the main modes are reduced more than in single models (Figure S5, S6). Summarized metrics—R^2, RMSE, and AAE – for lignin and holocellulose in two charts are shown in Table 4 and Table 5. The ensembles possess a complex structure and may cause difficulties in the interpretation, but in such a delicate case of the natural object like bamboo, enhanced complexity can be justified due to deviative compositions from species to species. For now, the selected ensemble is preferable and must be tested in the future with new samples to refine the predictive performance.

Another step deals with the selection of a model from the over-/underfit point of view. This aspect is critical especially for new data and if, by way of example, a model tends to overfit the existing data (to consider non-significant errors as true experimental points), it may incorrectly "recognize" new points, giving a worse regression fit. Over-/underfitting has been estimated by a simple absolute difference between the RMSEs in validation and training sets – this is typical practice to select the most appropriate model undertaken in data science (Fig. 18). Bootstrap Forest, despite the lowest metrics, tends to the least overfitting, whereas Decision Tree is the most overfitting model. Neural Network as a boosted ensemble of the models is proposed to be the optimal option due to equilibrated low overfitting and the top prediction performance. Interestingly, the prediction for lignin is less overfitted than for holocellulose except by Bootstrap Forest. Thus, the boosted ensemble for both phases can be recommended for deploying and further testing with new data (bamboo samples) to predict the composition by two main components. In future studies, we intend to collect more samples from other species and test boosted models developed in the present research. Also, proposed cluster model may be adjusted over the mean values. With this research we initiate implementation of modern concept to the routine analysis of the bamboo materials and, hence, it can be extended to more species. Clustering itself is a tool for quick classification for different bamboos. At the next phase of the study, it seems possible to recourse to multivariate analysis. It means, mapping of the clusters will be more precise if together with concentrations of lignin and holocellulose some other properties are added. For instance, specific mechanical, chemical, spectral characteristics are relevant as variables: dependence of mechanical

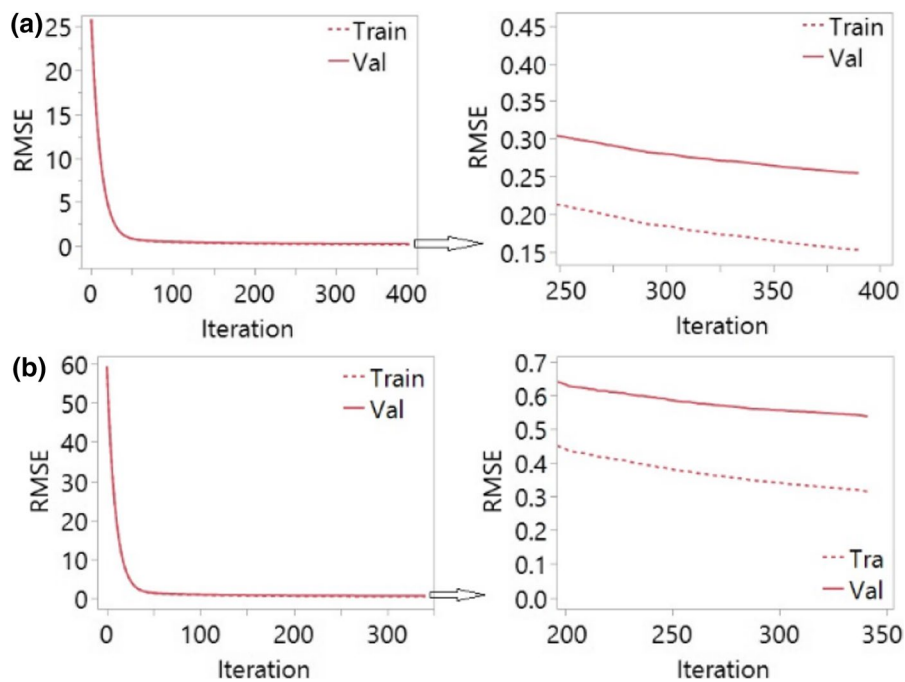

Fig. 14 Progress of RMSE with the number of XGBoost-iterations for **a** lignin and **b** holocellulose

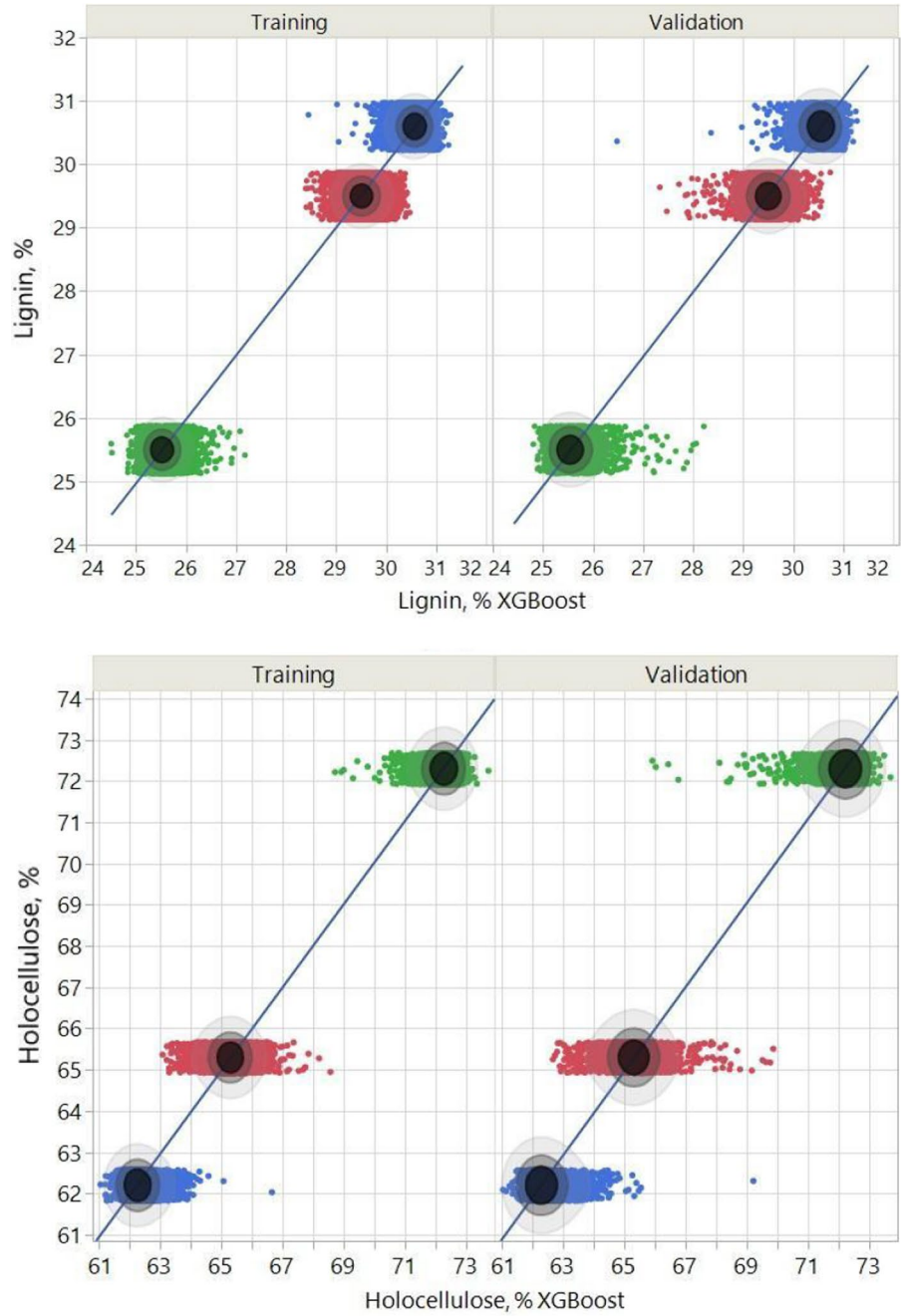

Fig. 15 The actual by XGBoost-predicted plots for lignin and holocellulose

and thermal properties on lignin and cellulose was thoroughly studied in (Richmond et al. 2021). Attempts to perform statistical analysis are recently reported to describe large body of the bamboo species (Biswas et al. 2022). In (Yeh and Yang 2020) the lower temperatures of decomposition at higher content of hemicellulose was also claimed. Application of machine learning can summarize and generalize

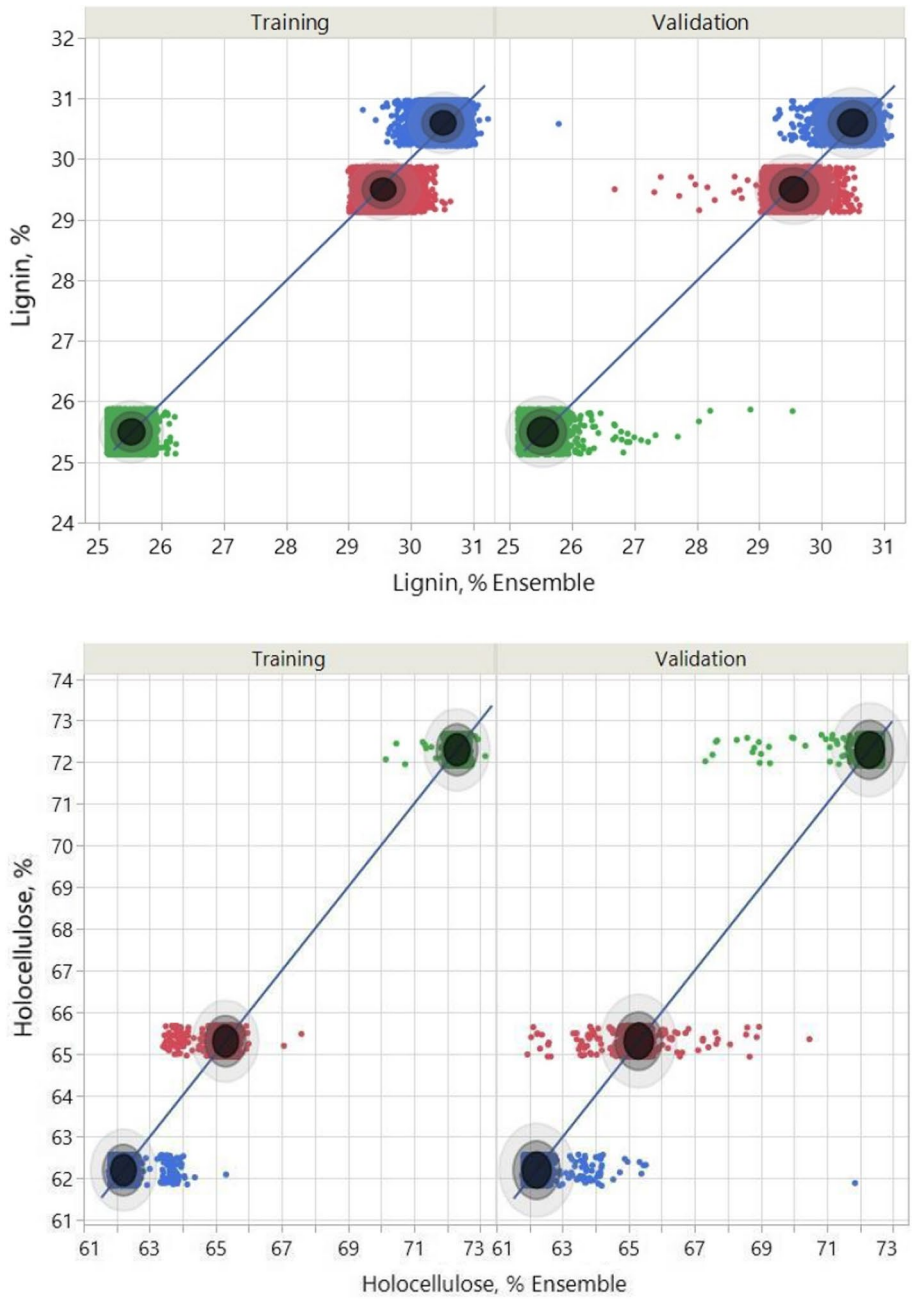

Fig. 16 The actual predicted plot for the Ensemble of models to predict the lignin and holocellulose concentration

the accumulated historical data from the experiments with following prediction of lignin and holocellulose content. Our aim is to extend the variety of the species – accordingly, the subject models can be updated and perform at adequate metrics.

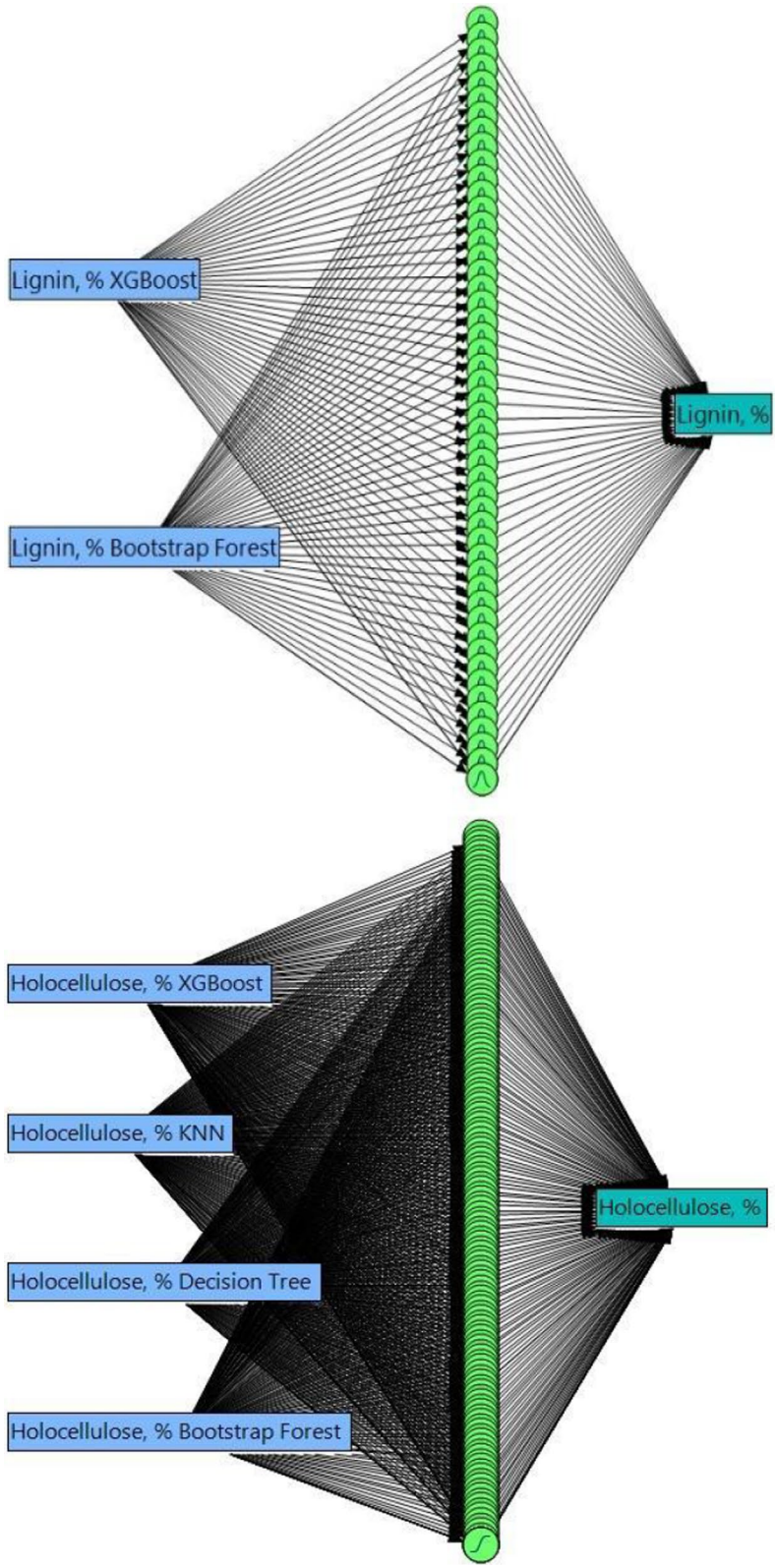

Fig. 17 Boosted neural networks for the Ensemble of Bootstrap Forest and XGBoost to predict the lignin and holocellulose concentration

Table 4 The summary of models performance by training and validation for lignin

Training				Validation			
Model	R^2	RMSE	AAE	Model	R^2	RMSE	AAE
KNN	1.0000	0.0000	0.0000	Ensemble	0.9923	0.1925	0.0986
Ensemble	0.9971	0.1175	0.0730	KNN	0.9892	0.2277	0.0167
Decision Tree	0.9957	0.1440	0.0171	XGBoost	0.9865	0.2549	0.1414
XGBoost	0.9951	0.1530	0.0956	Decision Tree	0.9653	0.4080	0.0579
Bootstrap Forest	0.9663	0.4022	0.1907	Bootstrap Forest	0.9593	0.4419	0.2129

Table 5 The summary of models performance by training and validation for holocellulose

Training				Validation			
Model	R^2	RMSE	AAE	Model	R^2	RMSE	AAE
Ensemble	0.9991	0.1233	0.0178	Ensemble	0.9952	0.2916	0.0488
KNN	0.9971	0.2289	0.0184	KNN	0.9879	0.4645	0.0458
XGBoost	0.9944	0.3152	0.1847	XGBoost	0.9838	0.5380	0.2929
Decision Tree	0.9911	0.3991	0.0632	Decision Tree	0.9627	0.8163	0.1368
Bootstrap Forest	0.9634	0.8084	0.3981	Bootstrap Forest	0.9626	0.8170	0.4008

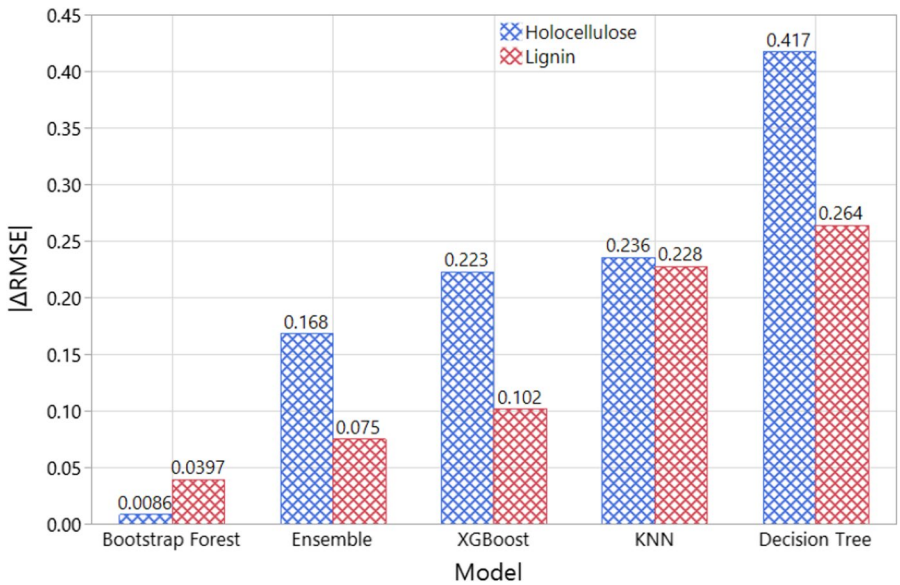

Fig. 18 Thse histogram of the RMSE difference between training and validation by each model for lignin and holocellulose

Conclusions

In this paper, the quantification of bamboo main components, such as holocellulose and lignin was carried out by means of thermogravimetric analysis (TGA). The influence of two types of gas atmosphere, nitrogen and synthetic air, was evaluated, and it has shown that the presence of char up to 1000 °C on inert gas flow. Data analysis protocol was developed to quantify bamboo main components based on Moso thermograms. The quantification of *holocellulose* and lignin phases has demonstrated the satisfactory agreement with conventional chemical extraction method for *Phyllostachys edulis* (Moso), *Bambusa vulgaris* and *Iranian phyllostachys* bamboos when using the deconvolution process. Machine learning models have indicated a connection between the features in thermogravimetric curves with the phase content (lignin and holocellulose), which was established by means of chemical extraction from the respective

samples. These results of the applied machine learning methods serve to develop applications (software) that can be used to predict the phase composition of different bamboo species based on TGA results. The proposed ensembles based on gradient boosting can be deployed after additional tests with more samples and set in production.

Acknowledgments Nazarkovsky is thankful to Dr. David Kirmayer (The Hebrew University of Jerusalem) and Dr. Russ Wolfinger (SAS) for their valuable advice. Nazarkovsky also is thankful to Prof. Volodymyr Zaitsev (Pontifical Catholic University of Rio de Janeiro) for the FAPERJ grant. The UFRJ staffs are also sincerely acknowledged for their kind assistance in the tests.

Author Contributions FdCV. Preparation of the first draft of the manuscript. Materials and samples preparation, data collection, measurements. Writing and arranging the final version of the manuscript. MN. Writing, arranging and editing the manuscript, performing the data analysis and description of the machine learning section. The submission of the manuscript. AA. Materials and samples preparation, data collection, measurements. Writing and arranging the final version of the manuscript. CM. Materials and samples preparation, data collection, measurements. BMdCG. Materials and samples preparation, data collection, measurements. JD. Preparation of the first draft of the manuscript. RDTF. The coordinator of the COPPETEC project. The revision of the manuscript. HS. The coordinator of the FAPESP Grant 2018/25011–9 project. The revision of the manuscript.

Funding All the authors gratefully acknowledge the Brazilian Council for Scientific and Technological Development (CNPq), COPPETEC foundations. Nazarkovsky is thankful to Rio de Janeiro State Research Foundation (FAPERJ, E-26/202.400/2021). Azadeh thanks the financial support from São Paulo Research Foundation (FAPESP Grant 2018/25011–9). Savastano Junior, in particular, is also thankful to Brazilian National Agency CNPq for the financial aid (CNPq 307723/2017–8).

Declarations

Conflict of interests The authors have no relevant financial or non-financial competing interests to disclose.

References

Agrawal A, Choudhary A (2016) Perspective: materials informatics and big data: realization of the "fourth paradigm" of science in materials science. APL Mater 4:053208. https://doi.org/10.1063/1.4946894

Bezerra ETV, Augusto KS, Paciornik S (2020) Discrimination of pores and cracks in iron ore pellets using deep learning neural networks. REM - Int Eng J 73:197–203. https://doi.org/10.1590/0370-44672019730119

Biswas S, Rahaman T, Gupta P et al (2022) Cellulose and lignin profiling in seven, economically important bamboo species of India by anatomical, biochemical, FTIR spectroscopy and thermogravimetric analysis. Biomass Bioenerg 158:106362. https://doi.org/10.1016/j.biombioe.2022.106362

Brebu M, Vasile C (2010a) Thermal degradation of lignin - a review. Cellul Chem Technol 44:353–363

Brebu M, Vasile C (2010b) Thermal degradation of lignin-a review. Cellulose Chem Technol 44(9):353

Cao W, Li J, Martí-Rosselló T, Zhang X (2019) Experimental study on the ignition characteristics of cellulose, hemicellulose, lignin and their mixtures. J Energy Inst 92:1303–1312

Carrier M, Loppinet-Serani A, Denux D et al (2011) Thermogravimetric analysis as a new method to determine the lignocellulosic composition of biomass. Biomass Bioenerg 35:298–307. https://doi.org/10.1016/j.biombioe.2010.08.067

Chung M-J, Wang S-Y (2018) Mechanical properties of oriented bamboo scrimber boards made of *Phyllostachys pubescens* (moso bamboo) from Taiwan and China as a function of density. Holzforschung 72:151–158. https://doi.org/10.1515/hf-2017-0084

dos Santos Abreu H, Monteiro de Carvalho A, Beatriz de Oliveira Monteiro M et al (2006) Métodos de Análise em Química da Madeira. In: Série Técnica Floresta e Ambiente. http://www.if.ufrrj.br/st/ano2006.html. Accessed 13 Jan 2022

Dumitriu S (2004) Polysaccharides: structural diversity and functional versatility. CRC Press, Second

Gressling T (2020) Data science in chemistry. De Gruyter

Himanen L, Geurts A, Foster AS, Rinke P (2019) Data-driven materials science: status, challenges, and perspectives. Adv Sci 6:1900808. https://doi.org/10.1002/advs.201900808

Kadivar M, Gauss C, Mármol G et al (2019) The influence of the initial moisture content on densification process of D. asper bamboo: physical-chemical and bending characterization. Construct Build Mater. https://doi.org/10.1016/J.CONBUILDMAT.2019.116896

Li X, Sun C, Zhou B, He Y (2015) Determination of hemicellulose, cellulose and lignin in moso bamboo by near infrared spectroscopy. Sci Rep. https://doi.org/10.1038/srep17210

Michael Buchanan (2007) TAPPI t204

Ornaghi HL, Ornaghi FG, Neves RM et al (2020) Mechanisms involved in thermal degradation of lignocellulosic fibers: a survey based on chemical composition. Cellulose 27:4949–4961. https://doi.org/10.1007/s10570-020-03132-7

Pandoli OG, Neto RJG, Oliveira NR et al (2020) Ultra-highly conductive hollow channels guided by a bamboo biotemplate for electric and electrochemical devices. J Mater Chem A 8:4030–4039. https://doi.org/10.1039/C9TA13069A

Richmond T, Lods L, Dandurand J et al (2021) Thermal and mechanical performances of bamboo strip. Mater Res Express 8:025502. https://doi.org/10.1088/2053-1591/abe060

Shen D, Hu J, Xiao R et al (2013a) Online evolved gas analysis by thermogravimetric-mass spectroscopy for thermal decomposition of biomass and its components under different atmospheres: part I. Lignin Bioresour Technol 130:449–456. https://doi.org/10.1016/j.biortech.2012.11.081

Shen D, Ye J, Xiao R, Zhang H (2013b) TG-MS analysis for thermal decomposition of cellulose under different atmospheres. Carbohyd Polym 98:514–521. https://doi.org/10.1016/j.carbpol.2013.06.031

TAPPI (2006a) Acid-insoluble lignin in wood and pulp. Test method T 222 om-21

TAPPI (2006b) Solvent extractives of wood and pulp. Test method T 204 cm-17

Ramiah 31 v (1970) Thermogravimetric and differential thermal analysis of cellulose, Hemicellulose, and Lignin

Valani LM, Vitorino FDC, Paiva A, Martins DS (2020) The influence of polymers impregnation on bending behaviour of phyllostachys pubescens (Mosso) bamboo. RILEM-SC2020 ambitioning a sustainable future for built environment: comprehensive strategies for unprecedented challenges. Guimarães, Portugal, pp 1–11

Valani LM, Vitorino FDC, Mar- APDS The influence of polymers impregnation on bending behaviour of phyllostachys pubescens (Mosso) bamboo. 1–11

Wang X, Cheng D, Huang X et al (2020) Effect of high-temperature saturated steam treatment on the physical, chemical, and mechanical properties of moso bamboo. J Wood Sci 66:52. https://doi.org/10.1186/s10086-020-01899-8

Yeh C-H, Yang T-C (2020) Utilization of waste bamboo fibers in thermoplastic composites: influence of the chemical composition and thermal decomposition behavior. Polymers (Basel) 12:636. https://doi.org/10.3390/polym12030636

Youssefian S, Rahbar N (2015) Molecular origin of strength and stiffness in bamboo fibrils. Sci Rep. https://doi.org/10.1038/srep11116

Zakikhani P, Zahari R, Sultan MTH, Majid DL (2016) Thermal degradation of four bamboo species. Bio Res 11:414–425. https://doi.org/10.15376/biores.11.1.414-425

Zhou T, Song Z, Sundmacher K (2019) Big data creates new opportunities for materials research: a review on methods and applications of machine learning for materials design. Engineering 5:1017–1026. https://doi.org/10.1016/j.eng.2019.02.011