



# Prediction of cellulose micro/nanofiber aspect ratio and yield of nanofibrillation using machine learning techniques

Giovana Signori-Iamin · Alexandre F. Santos ·  
Marcos L. Corazza · Roberto Aguado ·  
Quim Tarrés · Marc Delgado-Aguilar

Received: 5 July 2022 / Accepted: 9 September 2022  
© The Author(s) 2022

**Abstract** Predictive monitoring of two key properties of nanocellulose, aspect ratio and yield of nanofibrillation, would help manufacturers control and optimize production processes, given the uncertainty that still surrounds their influential factors. For that, 20 different types of cellulosic and lignocellulosic micro/nanofibers produced from spruce and pine softwoods, and by different pre-treatment and fibrillation techniques, were used as training and testing datasets aiming at the development and evaluation of three machine learning models. The models used were Random Forests (RF), Linear Regression (LR) and Artificial Neural Networks (ANN), broadening the scope of our previous work (Santos et al. in *Cellulose* 29:5609–5622, 2022. <https://doi.org/10.1007/s10570-022-04631-5>). Performance of these models were evaluated by comparing statistical parameters such as Mean Absolute Percentage Error (MAPE) and  $R^2$ . For the aspect ratio and the yield of nanofibrillation,

inputs were chosen among these easily controlled or measured variables: Total lignin (wt%), Cellulose (wt%), Hemicellulose (wt%), Extractives (wt%), HPH Energy Consumption (kWh/kg), Cationic Demand ( $\mu\text{eq/g}$ ), Transmittance at 600 nm and Consistency index (Ostwald-De Waele's  $k$ ). In both cases, the ANN models trained here provided satisfactory estimates of aspect ratio (MAPE=4.54% and  $R^2=0.96$ ) and the yield of nanofibrillation (MAPE=6.74% and  $R^2=0.98$ ), being able to capture the effect of the applied energy along the fibrillation process. RF and LR models resulted in correlation coefficients of 0.93 and 0.95, respectively, for aspect ratio, while for yield of nanofibrillation the correlation coefficients were 0.87 and 0.92.

**Keywords** Artificial neural networks · Aspect ratio · Yield of nanofibrillation · Lignocellulosic micro/nanofibers · Multiple linear regression · Random forest

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10570-022-04847-5>.

G. Signori-Iamin · A. F. Santos · M. L. Corazza  
Department of Chemical Engineering, Federal University of Paraná, Curitiba, PR CEP 81531-990, Brazil

R. Aguado · Q. Tarrés · M. Delgado-Aguilar (✉)  
LEPAMAP-PRODIS Research Group, University of Girona, C/ Maria Aurèlia Capmany, 61, 17003 Girona, Spain  
e-mail: m.delgado@udg.edu

## Introduction

Nanocellulose consists of a group of cellulose-based materials that have at least one dimension in the nanometric scale and englobes three main types of nanostructures: cellulose nanocrystals (CNCs), cellulose nanofibers (CNFs) and bacterial nanocellulose (BNC) (Abdul Khalil et al. 2014). Taking a closer look at cellulose micro/nanofibers (CMNFs), these are generally

produced by a top-down approach, including a pre-treatment stage of lignocellulosic feedstock followed by a fibrillation step. Pre-treatment methods englobe chemical and enzymatic treatments, oxidation procedures and mechanical techniques. The fibrillation process, on the other hand, is mainly mechanical and it is based on delaminating the pulp, for example, by means of high-pressure homogenization, grinding and others (Klemm et al. 2018; Dai et al. 2019). CMNFs are characterized by the presence of both amorphous and crystalline regions within the cellulose fiber, being both strong and flexible. Because of that, they are able to form an entangled network and, therefore, CMNF applications include emulsification in the food industry, rheology modifiers, reinforcing agent in the packaging and automotive sectors, water treatment and membrane production (Alila et al. 2013; Klemm et al. 2018). If lignin is kept along the whole process (Serra-Parareda et al. 2021a), be it with the aim of valorizing lignocellulosics with high yields, to avoid the use of hazardous chemicals, or to attain less hydrophilic nanocellulose forms, it is convenient to talk about lignocellulosic micro/nanofibers (LCMNFs).

Usually, criteria to discriminate microfibers from nanofibers imply that the diameter of the former is greater than 100 nm (Isogai 2020). This work opts for a pragmatic definition that may or may not coincide with the 100-nm criterion. In dilute suspensions of LCMNFs, the proportion of fibers that are small enough so that Brownian motion prevails over gravitational settling is called “yield of nanofibrillation” (Besbes et al. 2011; Sanchez-Salvador et al. 2021b). Particles that tend to sediment, despite having undergone a destructuring process, constitute the microfibrillated fraction.

In spite of the broad spectrum of applications, large scale production of nanocellulose is challenging and one of the reasons is the lack of easy, rapid and accurate measuring techniques for process monitoring (Serra-Parareda et al. 2021c). Moreover, some methods used for property measurement are time-consuming or tedious, even at laboratory scale. The aspect ratio, an important morphological property of CMNFs that strongly influences various properties regarding final applications, such as mechanical and rheological parameters (Leong et al. 2022), is measured either by gel point methodology or by very high-resolution microscopy. Another relevant parameter,

the yield of nanofibrillation, can be obtained by separation of the dispersed and settling fractions using stages of centrifugation, drying and weighing (Raj et al. 2016; Tarrés et al. 2020b; Sanchez-Salvador et al. 2021a). As noticeable, these techniques involve steps that are quite tiresome for monitoring purposes in the industry. Another relevant aspect is the lack of consistent mathematical models able to describe and capture the complex relations between the features of the CMNFs produced and the characteristics of the raw material and process conditions.

Among the most valuable efforts to benchmark CNF suspensions, the quality index proposed by researchers from LGP2 (Desmaisons et al. 2017) and the correlations of the Arrhenius Lab in Stockholm (Kriechbaum et al. 2018) should be cited. The former relied on optical microscopy, nanosized fraction and turbidity tests, while the latter was mainly based on the porosity and the apparent density of anisotropic CNF foams. Nonetheless, despite the usefulness of *post hoc* quality assessment, accurate monitoring involving variables to be measured on-line, or determined and controlled beforehand, would be an important leap forward in the manufacturing of nanoscale cellulose products. For instance, the strength of nanocellulose-based materials has been predicted, to different levels of satisfaction, by micromechanical models, such as the Cox-Krenchel method (Lee et al. 2014; Mugwagwa and Chimphango 2022) or combinations between the Kelly-Tyson and Pukanszky models (Tarrés et al. 2020a). In this context, machine learning (ML) algorithms constitute an advantageous tool to apply in predicting and controlling the properties of CNF suspensions, films and materials derived thereof, which is a field that has been scarcely explored (Özkan et al. 2019; Torrents-Barrena and Pellicer 2021).

Machine learning is a subfield of artificial intelligence that can be used to predict outputs based on previously presented data without the need of explicit programming, which means, in a way that the computer can learn “by itself” from data and be more accurate with time. ML can be a great tool for getting insights and solving complex problems that require fine-tuning or when existing solutions are not satisfactory (Géron 2019). Some algorithms frequently applied in Machine Learning are Linear Regression (LR), Decision Trees, Random Forests (RF), Artificial Neural Networks (ANN), and Clustering. Putting

a spotlight on the algorithms typically used in Material Science, supervised learning algorithms usually prevail, such as LR, RF and ANN.

When it comes to ML techniques in the cellulose/nanocellulose fields, some work worthy of mention are: Aguado et al. (2016), using a support regression vector approach to predict paper strength from morphological characteristics; Pennells et al. (2022), with a similar approach on nanopaper; Özkan et al. (2019), who evaluated three different algorithms (LR, RF and ANN) in the prediction of mechanical properties of three-component nanocomposite films; Almonti et al. (2019), who made use of ANNs to predict fiber length based on the pulp refining process parameters. Recently, another article of ours provided ANN predictions of the aspect ratio of (L)CMNFs, using pulps from aspen, eucalyptus and spruce to validate the performance of the model with other types of biomass (Santos et al. 2022). This work widens the scope of said attempts, including the yield towards nanostructured cellulose and other alternative methods. The ability to rapidly predict the yield and the morphological properties of nanocellulose based on the joint use of machine learning tools and a set of easily measurable input parameters of the starting pulps is critical for industrial applications. An example of application follows: an on-line series of optical and rheological sensors detect a set of variables that would result in a low yield of nanofibrillation, and thus the cellulose stream is recycled to the fibrillation unit for one more pass. For these reasons, this study focusses on developing techniques that are inexpensive, easy to implement and widely applicable, such as machine learning models combined with selected input variables. All considered, the present work seeks to take one more step towards the effective control and predictive monitoring of the fibrillation of (ligno)cellulosic pulps. Besides ANN, two ML algorithms, LR and RF, were applied for predicting both the yield of nanofibrillation and the aspect ratio of laboratory-produced (L)CMNFs. It should be noted that, as the previous literature understandably focuses on predicting the mechanical properties of the nanocellulose-based end product, the yield of nanofibrillation (a key indicator of the extent of the fiber disruption itself) has been overlooked. Inputs are based on easily obtainable data from the pre-treated fibers and the fibrillation process. After discussing the influence of the most relevant variables, the three models are compared by means of

their performance parameters and they are analyzed regarding the feasibility of their application.

## Experimental

### Dataset materials

The data used for training and testing the models were obtained from previous work of our group (Serra-Parareda et al. 2021c). Therein, four cellulosic and lignocellulosic pulps, following different treatments, were employed to produce mechanical (L)CMNFs with varying nanofibrillation yields. For the sake of comparison, it should be noted that this experimental dataset, explicitly displayed in Table 1, partially overlaps that of our previous work on the ANN-based prediction of the aspect ratio (Santos et al. 2022).

Three of the pulps were obtained from pine (*Pinus radiata*), considering that they had undergone thermomechanical, unbleached kraft and bleached kraft treatments (TMP, UKSP and BKSP respectively), whilst the final one was obtained from spruce (*Picea abies*) that had undergone thermomechanical pulping and peroxide bleaching (BTMP). The pre-treatment step of these pulps consisted in a mechanical refining stage in a Valley beater, designed in compliance to TAPPI T200 (TAPPI 2020), and all four pulps were refined for 60 min. Afterwards, each pulp was passed through a high-pressure homogenizer (HPH), NS1001L PANDA 2 K-GEA (GEA Niro Soavy, Parma, Italy), gradually increasing the number of passes from 3 to 9 and pressure from 300 to 900 bar. This resulted in five HPH levels of energy consumption, which was measured by means of a device from Circutor (Barcelona, Spain), model CVM-C10. Samples were taken from each of the stages, resulting in 5 types of CNFs for each pre-treated pulp and thus totalizing 20 (L)CMNF different samples (Table 1).

The pre-treated pulps were characterized in terms of their chemical composition and crystallinity index (%). Data regarding these parameters, whose extraction and curation is explicit in the aforementioned article (Serra-Parareda et al. 2021c), can be checked in Table 2. Briefly, ash and extractives were determined gravimetrically after combustion at 525 °C and ethanol/benzene extraction, respectively, according to the common TAPPI methods (TAPPI 2020). Structural carbohydrates and total lignin were determined

**Table 1** CMNF and (L)CMNF characteristics as function of the HPH intensity (Serra-Parareda et al. 2021c)

Pulp	HPH intensity (kWh/kg)	CD ( $\mu\text{eq/g}$ )	T at 600 nm (%)	k	n	Aspect ratio	Yield (%)
Pine TMP	5.67	148	2.5	0.222	0.281	51	3.12
	7.73	155	3.1	0.625	0.272	65	4.22
	11.67	170	5.9	0.789	0.270	72	6.74
	14.59	179	7.0	0.889	0.257	74	7.31
	19.72	190	7.4	1.107	0.239	80	8.11
Pine UKSP	5.67	157	4.2	1.461	0.373	63	7.40
	7.73	166	6.5	4.987	0.280	66	9.26
	11.67	177	7.4	6.780	0.276	78	12.41
	14.59	188	9.9	6.638	0.224	90	14.20
	19.72	206	12.5	7.679	0.233	103	17.94
Pine BKSP	5.67	154	5.4	2.178	0.253	140	7.46
	7.73	169	6.8	2.824	0.256	177	8.80
	11.67	188	7.2	3.873	0.247	212	11.57
	14.59	201	9.5	4.816	0.234	214	13.12
	19.72	210	12.7	6.549	0.204	217	15.48
Spruce BTMP	5.67	156	4.1	0.349	0.338	59	2.90
	7.73	173	4.7	0.456	0.324	69	6.00
	11.67	185	7.0	0.616	0.307	68	11.60
	14.59	199	9.7	0.781	0.299	70	15.60
	19.72	214	11.9	1.201	0.278	70	20.60

**Table 2** Chemical composition and morphology of the selected pulps (Serra-Parareda et al. 2021c)

Pulp	Cellulose (wt%)	Hemicellulose (wt%)	Total lignin (wt%)	Extractives (wt%)	Ash (wt%)	Crystallinity Index (%)
Pine TMP	47.9	22.9	27	0.7	1.4	73.3
Pine UKSP	74.2	16.6	8.2	0*	0.8	81.1
Pine BKSP	85.3	8.7	3.9	1.2	0.9	87.0
Spruce BTMP	46.2	22.9	29.4	0.9	0.5	76.1

\*Percentage of extractives was considered 0 for modeling since this component was below the lower limit of detection of the equipment used (0.3%)

by following the analytical procedure NREL/TP-510-42618 (Sluiter et al. 2012), which involves hydrolysis, spectrophotometry, and liquid chromatography. Crystallinity was estimated from the (200) peak of cellulose I in X-ray diffraction patterns (Supplementary Information, Figure S1), applying Segal's empirical method (Segal et al. 1959).

The CNFs produced were characterized by measurement of rheology properties, such as consistency index ( $k$ ) and flow index ( $n$ ), yield of nanofibrillation, aspect ratio, transmittance at 600 nm and

cationic demand. In particular, the yield of nanofibrillation was understood as the fraction of solid material that did not sediment when centrifuging a 0.2% CNF suspension at 3400 g for 20 min. Regarding the aspect ratio, it was estimated *via* the gel point (Varanasi et al. 2013), which was in turn calculated from the sediment height at several concentrations, as described in detail elsewhere (Sanchez-Salvador et al. 2021c). The cationic demand was determined by potentiometric back titration (Serra-Parareda et al. 2021b).

## Machine learning techniques

The ML algorithms were implemented in Python and the environment used for training and testing the models was the open-source platform Jupyter Notebook. The whole dataset consisting in the 20 samples was separated into a training and testing subsets considering 80% and 20% of experimental data, respectively.

The modeling approach here used a multiple input–single output (MISO) structure for each kind of model, considering aspect ratio and yield of nanofibrillation (%) as the output parameter for each model. The initial features chosen to be the inputs were: cellulose (wt%), total lignin (wt%), hemicellulose (wt%), ash (wt%), extractives (wt%), crystallinity index (%), transmittance at 600 nm (T600), cationic demand (CD), consistency index (k), flow index (n) and energy consumption in the high-pressure homogenizer (kWh/kg).

The features to be used in the final models responded to a double criterion. On the one hand, they were selected to be suitably applied at industry. On the other, based on their simplicity and reproducibility. In this sense, some of the parameters were related to pre-treated pulps (cellulose, hemicellulose, lignin, extractives and ash content, and crystallinity), some others to (L)CMNF characteristics (cationic demand, transmittance, consistency index and flow behavior index), and to process (energy consumption during HPH). Although the determination of the chemical composition of the pre-treated pulps requires specific

gravimetric methods and chromatographic analyses, the homogeneity of the material allows their determination less frequently and, indeed, it can be a parameter provided by the pulp supplier. Energy consumption can be directly measured from the fibrillation equipment with the appropriate device and, finally, the selected (L)CMNF characteristics can be directly measured with low sample processing.

Table 3 summarizes the variables and range values investigated for ML modelling. For LR and ANN, input data was normalized since these algorithms are affected by the order of magnitude of the features, as it will be explained in the following sections.

### Linear regression

LR models are based on a single or multiple linear function, relating a dependent output with one or a sum of multiple independent inputs, each of them being multiplied by their specific weight, plus a constant (Géron 2019):

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_n x_n \quad (1)$$

In Eq. 1, the first term  $\hat{y}$  refers to the predicted value, while  $\theta_0$  represents the constant and  $\theta_n$  the weight associated to each input value  $x_n$ .

As aforementioned, applying LR in machine learning requires either normalization or standardization of data. As it can be seen from Eq. 1, if an input is numerically large, even though its weight is not so big, it might still have a strong impact in the

**Table 3** Ranges of the experimental data used for ML modelling (Serra-Parareda et al. 2021c)

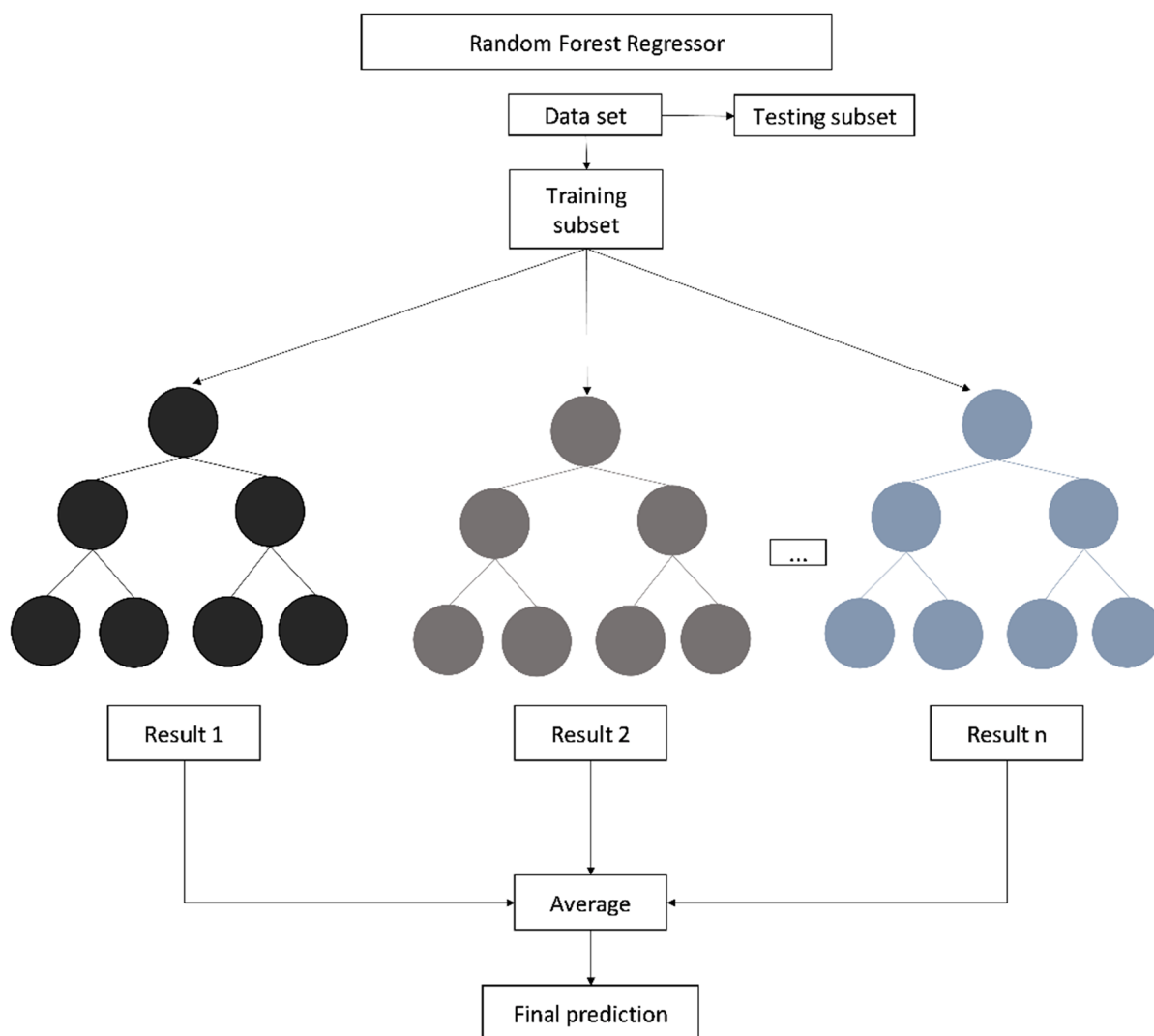
Type of variable	Variable	Range	Units
Pre-treated pulp characteristics (input)	Cellulose content	47.9–85.3	wt%
	Hemicellulose content	8.7–22.9	wt%
	Total lignin content	3.9–29.4	wt%
	Extractives	0–1.2	wt%
	Ash	0.5–1.4	wt%
	Crystallinity Index	73.3–87	%
Process parameters (input)	HPH–Energy consumption	5.67–19.72	kWh/kg
(L)CMNF characteristics (input)	Cationic demand	148–210	µeq/g
	Transmittance at 600 nm	2.5–12.7	%
	Consistency index “K”	0.222–7.679	–
	Flow behavior index “n”	0.204–0.373	–
	Output	Aspect ratio	51–217
	Yield of nanofibrillation	2.90–20.60	%

prediction. Therefore, in order to eliminate this kind of influence, all input values should be in the same range. Further details of the multiple linear regression algorithm can be found elsewhere (Maulud and Abdulazeez 2020).

### Random forest

RF is an algorithm based on a conjunction of multiple Decision Tree models and it is usually trained by the bagging method. It fits a number of Decision Trees, and the final outcome is the average from the results obtained by each individual tree. RF adds extra

randomness while growing the trees and it looks for the best feature among a random subset of them while splitting a node. In this way, RF generally performs better than single Decision Trees and because of the large number of predictors, it does not overfit, giving a more robust performance. Its structure is outlined in Fig. 1. Internal estimators such as generalization error, strength and correlation are used both for monitoring the performance of the model in the face of the increased number of features used in the splitting, and for determining their importance (Breiman 2001). Because of the principal function of RF models, data normalization is not required.



**Fig. 1** Schematic figure of a random forest regressor

## Artificial neural networks

ANNs are systems that resemble biological neurons, being based on small processing units that are intertwined. Each neuron receives input values that have an associated weight, and the network makes use of activation functions, including, e.g., rectified linear unit and sigmoidal functions, to connect between layers of neurons (Haykin 1999; Géron 2019).

Its structure comprises the input layer, the hidden layer and the output layer of neurons. The input layer possesses the same depth as the number of features used in the model. The hidden layer includes as many layers of neurons intended and the depth of each hidden layer is a hyperparameter that can be tuned. Finally, the output layer size is the same as the number of outputs described by the model. ANN makes use of backpropagation and optimization algorithms in order to improve the weights and bias associated. A simplified scheme is depicted in Fig. 2 and more detailed descriptions can be found elsewhere (Haykin 1999; Torrents-Barrena and Pellicer 2021; Santos et al. 2022).

### Model evaluation

ANNs, RF and multiple LR models had their accuracy tested and compared by statistical parameters of customary use: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE) and  $R^2$  score.

MAE and MAPE can be calculated using Eqs. 2 and 3, where  $n$  represents the number of datapoints,  $y_i$  is the actual value for the  $i$ th sample and  $\hat{y}_i$  is the prediction of the  $i$ th sample.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2)$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (3)$$

RMSE is calculated as represented in Eq. 4. The notations are the same for the metrics stated above.

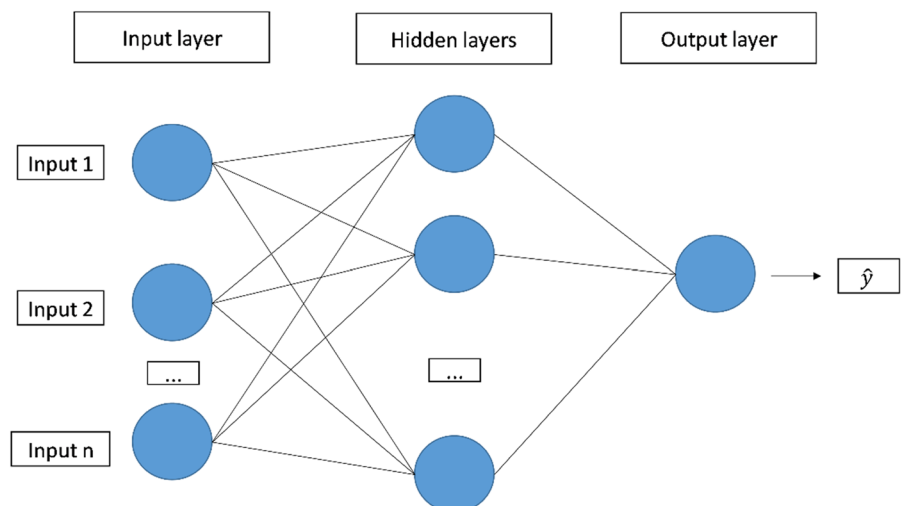
$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (4)$$

The determination coefficient  $R^2$  can be calculated by Eqs. 5 and 6, where the notations are the same as stated above, with exception of  $\bar{y}$ , which represents the mean value of all actual observations.

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

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (6)$$

**Fig. 2** Schematic representation of an artificial neural network



## Results and discussion

### Aspect ratio

#### *On the input parameters*

Due to the number of combinations tested, Table 4 only presents the results for the three best combinations of features in each case. Statistical parameters are shown for the sake of comparison between different groups of parameters and their influence on the performance in the train and test subsets.

As it can be seen from Table 4, modelling with all features was performed for the three algorithms, followed by up to 5-feature models in which mainly chemical components are present in the modelling. This funneling of parameters was done based on the models with subset of features that performed better and on the two guidelines presented in the [Experimental](#) section. Additionally, we also considered the features that were amongst the most important and/or that had higher absolute values of Pearson correlation coefficient to help funneling the options. These features can be seen in Fig. 3; Table 5.

It is possible to observe that cellulose, hemicellulose, total lignin content and crystallinity index appear in both Fig. 3; Table 5 as important features. Since crystallinity index and flow index (n) were not considered as practical parameters, they were not considered to be retained in the final model, even though they were simulated. On the other hand, chemical composition was preserved for initial testing, and later in conjunction to other features, even the ones that did not present at first as the most relevant. HPH energy requirement is an example of parameter that was selected to be maintained in the modelling since it is a process parameter of easy acquirement and provided good performance.

Regarding the performance of the models presented in Table 4, for the random forest models one can observe that when all features were used, the best results were attained. However, they account for 11 inputs, among which there are both easier and harder to measure properties. When model 2 is run, which consisted in five inputs, comparable results are achieved. Other combinations were tried, one of them being model 3, the final subset of features chosen for this algorithm. In this case, we get slightly worse

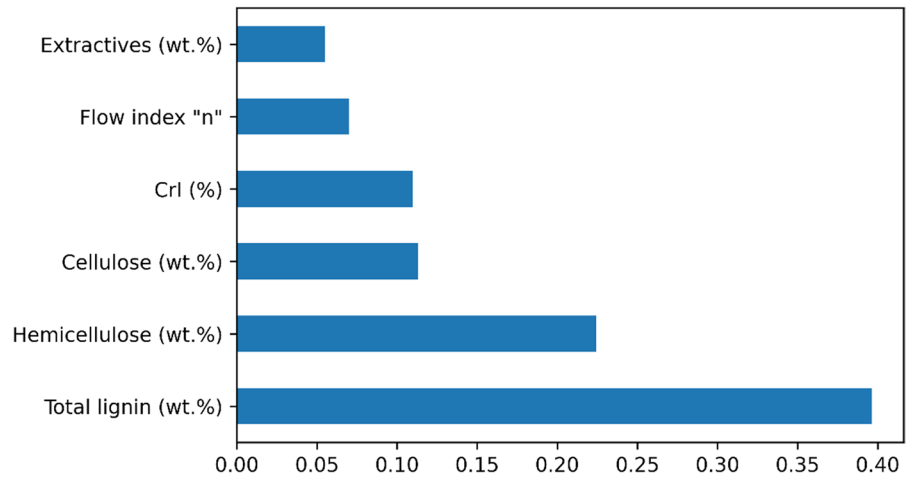
**Table 4** Modelling results for different combinations of features for aspect ratio

Algorithm	Features selected	Type of set	MAE	MAPE	RSME	R <sup>2</sup>
Random forest (17 estimator forest)	Model 1: All 11 features	Train	9.747	8.79	14.025	0.930
		Test	9.813	5.97	16.918	0.927
	Model 2: Features 1–5*	Train	10.019	9.02	14.237	0.928
		Test	8.722	5.37	15.265	0.941
	Model 3: Features 1,2,4*	Train	10.019	9.02	14.237	0.928
		Test	8.723	5.37	15.265	0.941
Linear regression	Model 1: all 11 features	Train	4.755	5.40	5.353	0.990
		Test	8.689	13.13	13.538	0.954
	Model 2: Features 1–5*	Train	7.108	7.25	10.826	0.958
		Test	15.398	16.36	16.673	0.930
	Model 3: Features 1–4*	Train	7.108	7.25	10.826	0.958
		Test	15.398	16.36	16.673	0.930
Artificial neural network	Model 1: all 11 features (11-22-1)	Train	1.576	1.23	3.887	0.995
		Test	6.488	6.64	7.283	0.987
	Model 2: Features 1–5* (5-10-1)	Train	5.760	4.10	11.020	0.957
		Test	8.544	6.29	12.610	0.960
	Model 3: Features 1–4* (4-8-1)	Train	10.209	6.65	19.733	0.861
		Test	16.431	11.27	26.374	0.824

\*Feature 1: Total lignin (wt%); Feature 2: Cellulose (wt%); Feature 3: Hemicellulose (wt%); Feature 4: HPH - Energy consumption (kWh/kg); Feature 5: Extractives (wt%)



**Fig. 3** Six most relevant features according to feature importance in the inherent function for aspect ratio RF modelling



**Table 5** Absolute values for Pearson correlation coefficient between features and aspect ratio (output of interest)

Feature	Pearson correlation coefficient
Hemicellulose content	0.8870
Crystallinity Index	0.8407
Cellulose content	0.7690
Total lignin content	0.7092
Flow index (n)	0.6042
Extractives	0.5489
Consistency index "k"	0.4778
Transmittance at 600 nm	0.4142
Cationic demand	0.3857
HPH– Energy consumption	0.2377
Ash	0.0095

MAE and MAPE, but very similar RSME and  $R^2$  compared to model 1, but exact performance of model 2, indicating that extracts (wt%) and hemicellulose (wt%) were not necessary for good performance. In this case, we had only three features selected which made use of very easy to measure parameters. For the linear regression models, when all features were used as inputs, again very good results are obtained, but this would not be a practical case. When the number of features is diminished to the subset used for model 2, the model still performs well. When the number of features is decreased to the final subset chosen, the exact results are gathered. Finally, for the ANN, model 1 presents very satisfactory results, as already expected for an all-feature simulation. Model

2, which used the same features as model 2 of LR, performed the best among all simulations for neural networks and was chosen as the final one. Model 3 performed worse and was discarded. Other combinations apart from these three presented here were tested, but they were not considered suitable for the application as well, either for low performance, either for requiring not the most interesting inputs.

Therefore, all final models selected were created using easily obtainable parameters as input features. Summarizing them, for LR, they were: cellulose (wt%) ( $x_1$ ), hemicellulose (wt%) ( $x_2$ ), total lignin (wt%) ( $x_3$ ) and energy requirement in the HPH (kWh/kg) ( $x_4$ ). RF employed three input features, the same as LR except hemicellulose (wt%). ANN required the highest number of factors, encompassing total lignin (wt%), cellulose (wt%), hemicellulose (wt%), extractives (wt%) and HPH energy consumption (kWh/kg). The first four belong to the chemical composition of the pulps, but, instead of requiring chromatography or advanced detectors, they can be estimated via the typical TAPPI methods (TAPPI 2020), consisting basically on gravimetric determinations and the use of oven drying, filtration, and cheap reagents. It is worth noting that, although these gravimetric methods were followed in the source work (Serra-Parareda et al. 2021c), they could be, to a certain point, further simplified towards in-line optical measurements in industrial applications. For instance, the conjugated structures of lignin absorb ultraviolet radiation ( $\sim 280$  nm), and they also have unique absorption bands in infrared spectra (Nader et al. 2022).

Additionally, it is possible to observe that the final models chosen for each algorithm encompass a distinct number of features or combination of them. These differences between the number and types of variables among the algorithms occur due to their different principal of function. The parameters required for good modelling will not necessarily be the same among the models, and this discrepancy remains a valid point to use for evaluation and comparison between algorithms. Considering the current knowledge on the effects of chemical composition on nanocellulose properties, the selection of input variables in this regard, such as the weight percentages of lignin and cellulose (common to all models), can be deemed consistent. For instance, lignin has been alleged to be detrimental for aspect ratio (Li et al. 2016), as a lignin-rich cell wall grants lower deformability in the ulterior conversion to nanocellulose. Likewise, a significant influence of the extent of fibrillation, expressed in this case as energy consumption, was expected. Such energy demand is directly proportional to the differential pressure exerted, according to Bernoulli's principle (Serra-Parareda et al. 2022), and thus it is inferred that the more energy supplied, the more nanostructured (ligno)cellulosic fibers

become. The reason lies in the native structure of the fiber itself, which, in a top-down sequence, is made out of fibrils with increasing aspect ratio, ending with the extremely slim elementary fibrils that constitute each microfibril. The lower the number of elementary fibrils that (L)CMNFs encompass, the higher the aspect ratio.

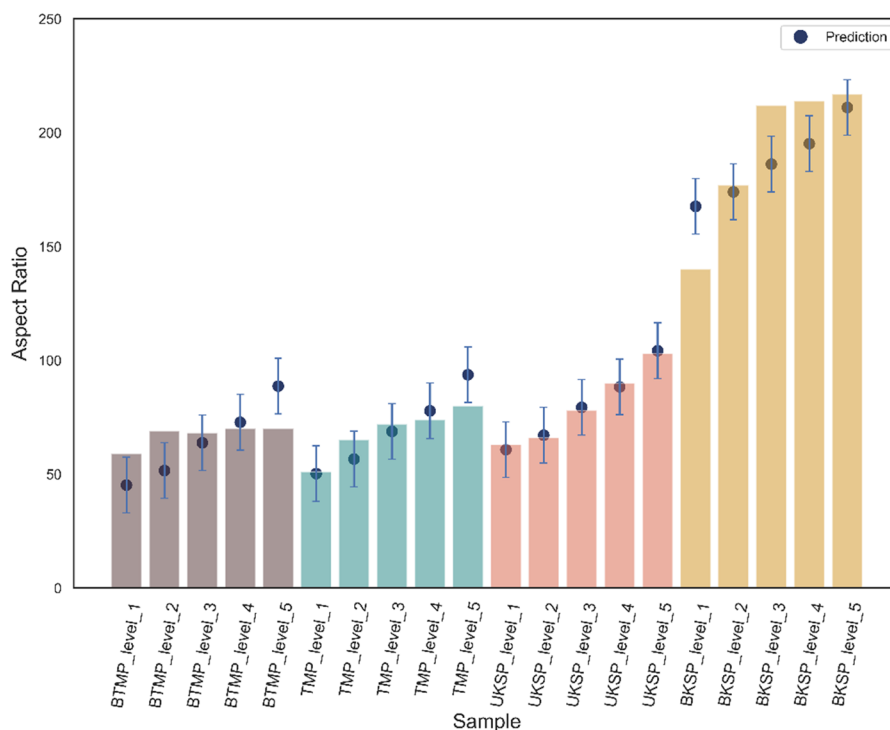
#### Performance of the predictive models

The LR model for the aspect ratio can be seen in Eq. 7. The RF model consisted in a 17-estimator forest. ANN's architecture comprised the aforementioned 5 inputs in the first layer, 10 neurons in the only hidden layer of the system and 1 output—aspect ratio. The activation function used in the hidden layer was the sigmoid function and the linear function was used for generating the final prediction.

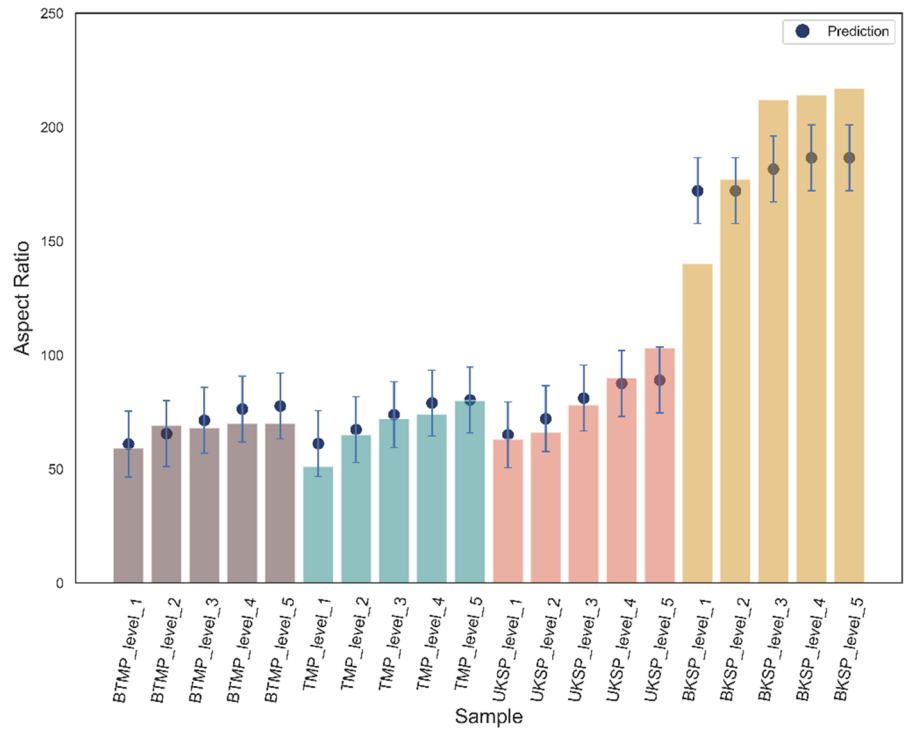
$$\hat{y} = -636.84x_1 - 411.66x_2 - 347.57x_3 + 43.43x_4 + 804.50 \quad (7)$$

Figures 4, 5 and 6 show the target values versus the predicted values for aspect ratio from LR, RF and ANN models, respectively. Bar plots in said figures refer to real values. Cross markers refer to prediction

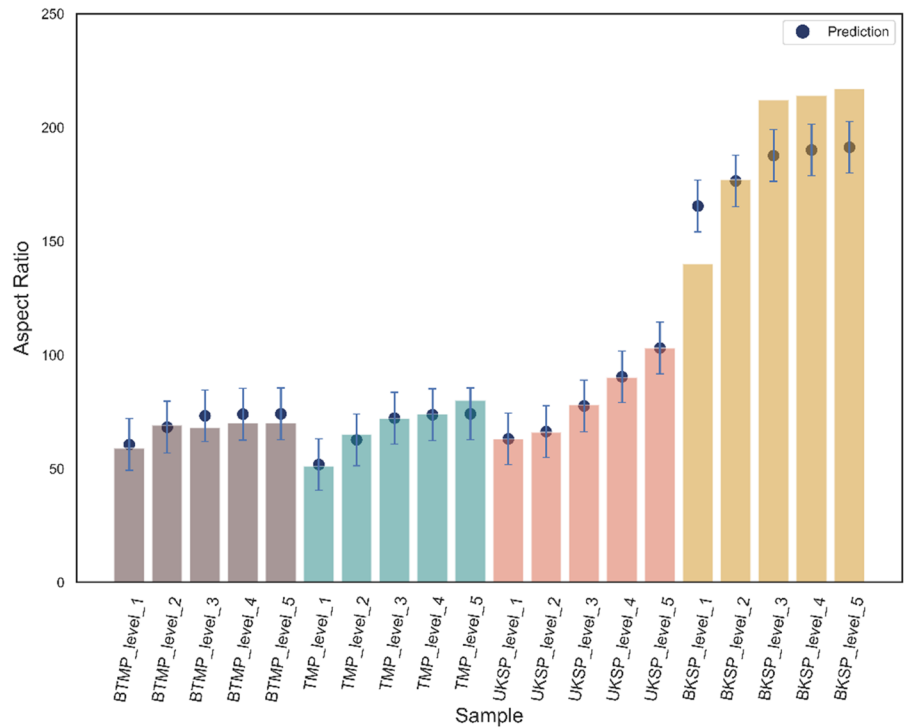
**Fig. 4** Target versus LR prediction values of CMNF aspect ratio



**Fig. 5** Target versus RF prediction values of CMNF aspect ratio



**Fig. 6** Target versus ANN prediction values of CMNF aspect ratio



values. Levels 1 to 5 indicate the increase in the applied energy, accordingly to the five levels of HPH intensity presented in Table 1.

From these plots, it is verified that all algorithms failed to predict accurately the point BKSP\_level\_1 and that all BKSP points are not as precisely met as the other groups of samples were. This may be because they presented very high aspect ratios, and therefore this data point was regarded as an outlier. Nonetheless, ANN model was able to capture the trend that BKSP samples took better than the other models, regarding the energy intensity increase, as noticeable in Fig. 6. Analyzing each model, as expected, LR provided linear relations for each group of samples, which is not quite the best for describing them, with exception of UKSP group. In spite of that, interestingly, it resulted in a better  $R^2$  (0.9508), RMSE (12.221) and MAE (8.766) than RF model. This one was able to catch the non-linearity in the increase of the aspect ratio in each group, but did not quite meet some experimental values. In this way, its MAPE of 8.18% was lower than LR's (9.07%), but the other metrics were worse: MAE of 9.760, RMSE of 14.448 and  $R^2$  of 0.9313. The best model that was able to grasp the trends and had better quality of performance was the ANN model. ANN seized to represent almost perfectly the increasing trends and differences from each group of samples, at the same time it predicted very precise values (with exception of BKSP, as stated above). Metrics for ANN model were: MAE of 6.317, MAPE of 4.54%, RMSE of 11.356 and  $R^2$  of 0.9575. All performance indicators are gathered in Table 6.

Despite the limitations specified above, all models presented an  $R^2$  above 0.9 and mean percent errors lower than 10%. Also, they were able to capture the increase of aspect ratio as the energy intensity applied in the HPH grew, as well as being able to differentiate each group of samples. Perhaps ANN performed better than other models and was able to catch the

nuances of the trends because of the activation function, which was sigmoidal, but at the expense of using more input variables. From Figs. 4, 5 and 6, one can observe that the increase in the aspect ratio is not linear and the behavior of the trend is different for each kind of treated pulp. For instance, BTMP does not show an increase in AR when there is an increase in the energy applied, apart from the BTMP\_level\_1 to BTMP\_level\_2 samples. This particularity is very different from what the other groups of samples show. TMP and UKSP resemble a linear increase, however, BKSP shows a large jump from BKSP\_level\_1 to BKSP\_level\_3 and forth on. ANN is a more complex algorithm, where different hyperparameters can be tuned in order to improve performance. As discussed above, it is able to grasp intricate relationships. Therefore, it is noticeable that, regarding the ability to simulate trends, LR did the poorest job, whilst the RF model performed in average and the ANN model performed very well, being evident that a more elaborate algorithm resulted in a better grasp of the complexity of the problem.

If the performance of these models satisfy nanocellulose manufacturers, they could be used for the predictive monitoring of key properties of the CNF suspension itself and of the end product. For instance, CNFs with high aspect ratio are entangled to a larger extent and, consequently, attain better barrier properties when used to coat packaging paper (Wang et al. 2021).

## Yield of nanofibrillation

### *On the input parameters*

Similarly to the aspect ratio, out of the large number of combinations tested, Table 7 presents the results for the top three combinations of features chosen to predict the nanofibrillation yield. Statistical parameters are displayed for the sake of comparison between distinct groups of parameters and their influence on the performances in the train and test subsets.

As it can be seen from Table 7, in this case, the modelling also started with all the features for the three algorithms. Afterwards, mainly final nanocellulose properties are present in the modelling. This is also due to the funneling of parameters that was done based on the models with a subset of features that performed better and on the two guidelines

**Table 6** Performance metrics for the three ML models regarding aspect ratio prediction

Model	MAE	MAPE (%)	RMSE	$R^2$
RF	9.760	8.18	14.448	0.9313
LR	8.766	9.07	12.221	0.9508
ANN	6.317	4.54	11.356	0.9575

**Table 7** Modelling results for different combinations of features for yield of nanofibrillation

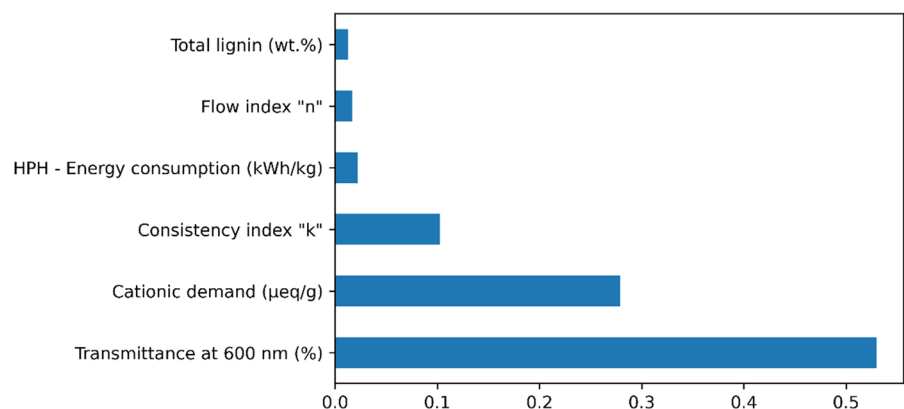
Algorithm	Features selected	Type of set	MAE	MAPE	RSME	R <sup>2</sup>
Random forest (40 estimator forest)	Model 1: All 11 features	Train	0.0124	14.53	0.0188	0.832
		Test	0.0199	43.92	0.0226	0.788
	Model 2: Features 1–5*	Train	0.0120	14.44	0.0177	0.851
		Test	0.0180	38.27	0.0198	0.838
	Model 3: Features 1–3*	Train	0.0136	17.44	0.0172	0.858
		Test	0.0135	34.19	0.0167	0.884
Linear regression	Model 1: all 11 features	Train	0.0097	11.13	0.0110	0.942
		Test	0.0159	42.55	0.0211	0.815
	Model 2: Features 1–5*	Train	0.0111	14.01	0.0129	0.921
		Test	0.0105	27.00	0.0137	0.922
	Model 3: Features 1–4*	Train	0.0122	14.94	0.0157	0.883
		Test	0.0117	23.75	0.0122	0.938
Artificial neural network	Model 1: all 11 features (11-11-1)	Train	0.0056	6.939	0.0068	0.978
		Test	0.0095	16.481	0.0098	0.960
	Model 2: Features 1–5* (5-5-1)	Train	0.0041	5.412	0.0062	0.981
		Test	0.0047	9.190	0.0048	0.990
	Model 3: Features 1–4* (4-4-1)	Train	0.0057	6.035	0.0076	0.972
		Test	0.0059	9.544	0.0073	0.978

\*Feature 1: Cationic demand ( $\mu\text{eq/g}$ ); Feature 2: Transmittance at 600 nm (%); Feature 3: HPH - Energy consumption (kWh/kg); Feature 4: Consistency index "k"; Feature 5: Extractives (wt%)

presented in the [Experimental](#) section. Furthermore, the features that were presented as most important for the RF models and that had higher absolute values of Pearson coefficient for the LR model were taken into consideration. These parameters can be seen in [Fig. 7](#); [Table 8](#).

It is possible to observe that transmittance, cationic demand, HPH energy consumption and consistency index (k) appear in both [Fig. 7](#); [Table 8](#). For

this reason, they were chosen to be tested first and later in conjunction to other features, even the ones that did not present at first as the most relevant. For instance, extractives (wt%) did not appear as one of the most relevant factors in terms of the Pearson correlation coefficient. However, it contributed to the performance of the LR model more significantly than other parameters of the chemical composition, such as ash (wt%) or cellulose (wt%). It is important to

**Fig. 7** Six most relevant features according to feature importance in the inherent function for yield of nanofibrillation RF modelling

**Table 8** Absolute values for Pearson correlation coefficient between features and yield of nanofibrillation (output of interest)

Feature	Pearson correlation coefficient
Transmittance at 600 nm	0.9363
Cationic demand	0.8966
HPH – Energy consumption	0.7703
Consistency index “K”	0.5706
Ash	0.4530
Flow index (n)	0.4168
Crystallinity Index	0.3425
Cellulose content	0.3105
Total lignin content	0.2845
Hemicellulose content	0.2546
Extractives	0.0779

highlight that, in the case of the Pearson correlation coefficient, the values presented regard the individual relevance of each feature with respect to the output of interest. Although the correlation coefficient value for extractives (wt%) was low, when taken into account in conjunction with the other features (transmittance, cationic demand, HPH energy consumption and consistency index) it helped to better fit the response. The improvement in the metrics, mainly in the train subset, can be observed in Table 7, where models 2 and 3 are compared.

From Table 7, one can observe that the best results were achieved also by a three-input model for the random forest, which had better performance when comparing to a model with all the initial features and with a 5-input one. In the case of LR, the all-feature model has a more drastic difference between train and test sets, which is not very good, once this may indicate overfitting. For model 2, the performances of both subsets are both more similar and better when comparing to models 1 and 3, thus it was selected as the chosen one. Finally, for the neural network, it was possible to obtain very good models with all the examples in the table, maintaining the topology scheme of the same number of neurons in the input and hidden layer. Thus, in this case, model 3 (4 input) was chosen as the best one. Other combinations apart from these three presented here were tested, but they were not considered suitable for the application as well, either for low performance, either for requiring not the most interesting inputs.

Therefore, in order to summarize the choices of final models for the yield of nanofibrillation, the final inputs selected were as follows. For LR model: transmittance at 600 nm ( $x_1$ ), cationic demand ( $\mu\text{eq/g}$ ) ( $x_2$ ), consistency index  $k$  ( $x_3$ ), extractives (wt%) ( $x_4$ ), and HPH energy requirement (kWh/kg) ( $x_5$ ). For the ANN model, all inputs above were used, with exception of extractives, totalizing four inputs. RF required even less features, being only three: cationic demand, transmittance and HPH energy supply. Regardless of the number of features used in each model, all of them are low-cost inputs that can be easily measured. In particular, the three factors for the RF model can be roughly quantified in real time, in-line, and by rapid and non-destructive methods: potentiometry with ion-selective electrodes for the cationic demand (although the accurate determination would require a titration), photometry at a single wavelength to detect the loss of intensity of the incident light, and direct reading of electric power. Should the resulting combination of measurements be related to an unacceptably low yield, recycling for further fibrillation would be triggered, automating the common decision in a typical batch process. Extractives can be determined gravimetrically after a solid–liquid extraction with organic solvents, while the consistency index, directly proportional to the apparent viscosity, would require a rheometer (Xu et al. 2016; Serra-Parareda et al. 2021c).

It is noticeable that the majority of these inputs are related to either a process parameter (HPH energy requirement) or final nanocellulose characteristics (cationic demand, transmittance and consistency index). These are consistent with literature as well. A higher differential pressure, as aforementioned for the aspect ratio, causes more disruption, and therefore it generates more nanofibers. This same effect, the fact that the proportion of nano to microfibers increases, is also related to higher transparency, i.e., higher transmittance within the visible light spectrum, since smaller particles scatter light to a lesser degree (Movsesyan et al. 2022). The cationic demand is due to the remaining negatively-charged functional groups in (ligno)cellulosic pulps, be it carboxylates from the hydrolysis of lignin/hemicellulose ester bonds, glucuronic acid units in hemicelluloses, and even polarized hydroxyl groups. In pulps, most of them do not interact with the titrating polyelectrolytes during the determination of the cationic demand, as

they are inside the fiber, but mechanical disruption makes them easily accessible, which is why surface charge is so robustly correlated to the specific surface area (Serra-Parareda et al. 2021b). Regarding the rheological behavior, it is well known that pulps and nanocellulose suspensions, in general, experience thinning with shear ( $n < 1$ ), but, besides that, reducing the particle size has complex effects on the apparent viscosity. On the one hand, the frequency of particle–solvent and particle–particle interactions should increase with decreasing particle size (e.g., by nanofibrillation). On the other hand, the viscosity of nanocolloids is known to increase with increasing particle size (Wang et al. 2011). In any case, what we often observe when increasing nanofibrillation is that the consistency index and the flow behavior index do not follow the same trends (Aguado et al. 2022), and thus the apparent viscosity itself is not as adequately correlated to the yield of nanofibrillation as the consistency index from the Ostwald–De Waele equation.

Aspect ratio and yield share a common mechanical basis, as the progressive breakdown of a fiber into its constituting fibrils implies the generation of thinner particles. Not only do these thinner fibrils have higher aspect ratio, but they are also less prone to gravitational effects given their lower hydrodynamic radius. However, fibrils do not follow a Stokian settling behavior (Kim et al. 2021), and their interlocking and entanglement are ubiquitous when their concentration is high enough. Therefore, the behavior of dilute suspensions is not necessarily correlated to the gel-forming capabilities of highly entangled systems (Sanchez-Salvador et al. 2020). In other words, the aspect ratio and the yield of nanofibrillation are independent responses that, nonetheless, may be affected by some factors in a qualitatively similar way.

#### *Performance of the predictive models*

The multiple LR model can be seen in Eq. 8. The RF model was built as a 40-estimator forest, with a minimum of two samples in each final leaf. ANN's final topology was 4–4–1, meaning 4 inputs in the input layer, 4 neurons in the hidden layer, and 1 output in the output layer. In the hidden layer, the activation function used was Rectified Linear Unit (ReLU) and the output function was the linear function. The difference between each activation function resides in the fact that the linear function is defined by a line,

whereas ReLU, despite its name, is a nonlinear transfer function which is calculated as  $\max(0.0, x)$ , which means that, for negative inputs, it returns a value of zero and when the input is positive, it returns the same value.

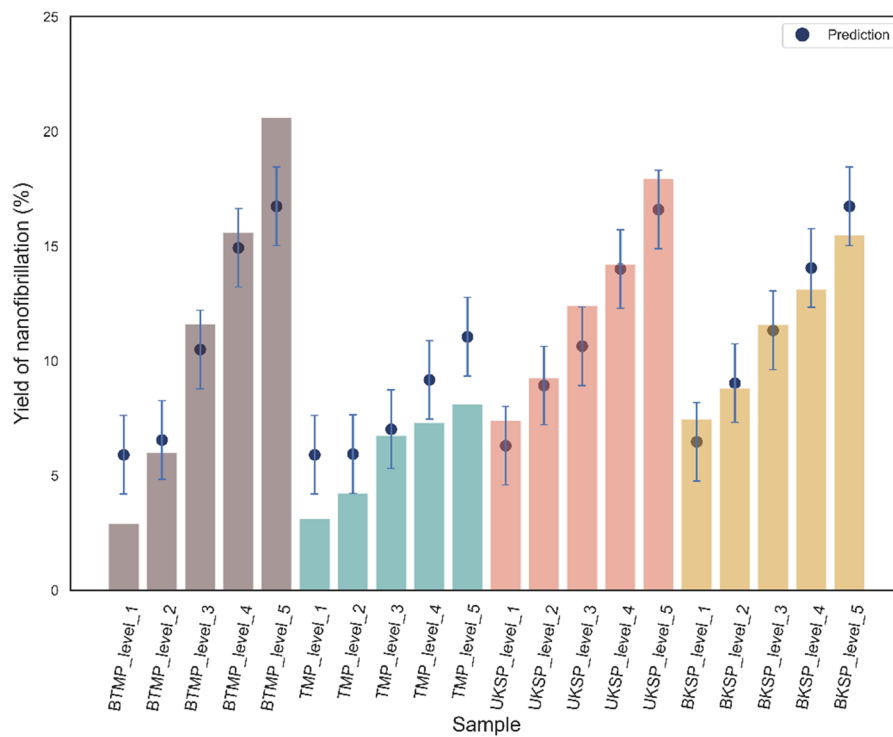
$$\hat{y} = 0.141x_1 + 0.107x_2 - 0.023x_3 - 0.029x_4 - 0.089x_5 + 0.067 \quad (8)$$

Figures 8, 9 and 10 show experimental values against the predicted outputs of RF, LR and ANN, respectively. Similar to the aspect ratio plots, columns in these figures represent actual values, while cross markers refer to predicted values. Levels 1 to 5 indicate the increase in energy consumption (Table 1).

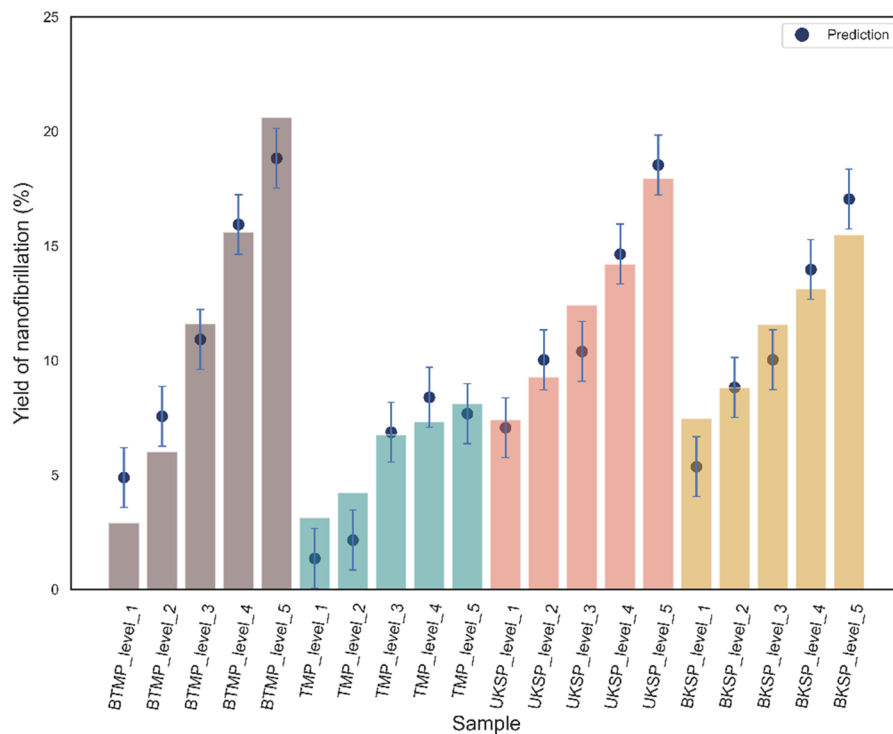
Likewise, as with aspect ratio, RF model performed worse than LR and ANN is the best model among them. Analyzing Fig. 8, it is visible that RF model missed multiple targets, mainly in TMP sample group, which is reflected in its  $R^2$  of 0.8720, the poorest of all three models. MAE, MAPE and RSME also were the biggest ones, being: 1.358, 20.79% and 1.712, as displayed in Table 9. This behavior might be due to a couple of reasons, for instance the robustness of the RF algorithm, or maybe also due to the principle of function of this technique, which consists in averaging the target values of the train samples that fall into a specific range (Géron 2019). Figure 9 shows how LR predicted yields. It meets better the targets, but still not so precisely. Nevertheless, its  $R^2$  is above 0.9, being 0.9255. It predicts better the TMP group than RF model did, although it does not quite fit the trend for points TMP\_level\_3, TMP\_level\_4 and TMP\_level\_5. Other groups were represented better in terms of tendencies related to increase of HPH energy and yield of nanofibrillation (%). MAE, MAPE and RSME for LR can be checked in Table 9. Again, ANN model presented the best results in terms of both performance and capturing of trends. It achieved the highest  $R^2$  of 0.9749 and lowest MAE, MAPE and RSME, being 0.577, 6.74% and 0.758 respectively. From Fig. 10, it can be seen that it almost met all points precisely, as it effectively captured all group trends.

Similar conclusions inferred for aspect ratio can be drawn for nanofibrillation yield. For example, all models were able to grasp, some in a better level than others, the increase in yield when HPH energy increased too. Also, they were able to differentiate each group of samples, considering they belong to

**Fig. 8** Target versus RF prediction values for yield of nanofibrillation (%)

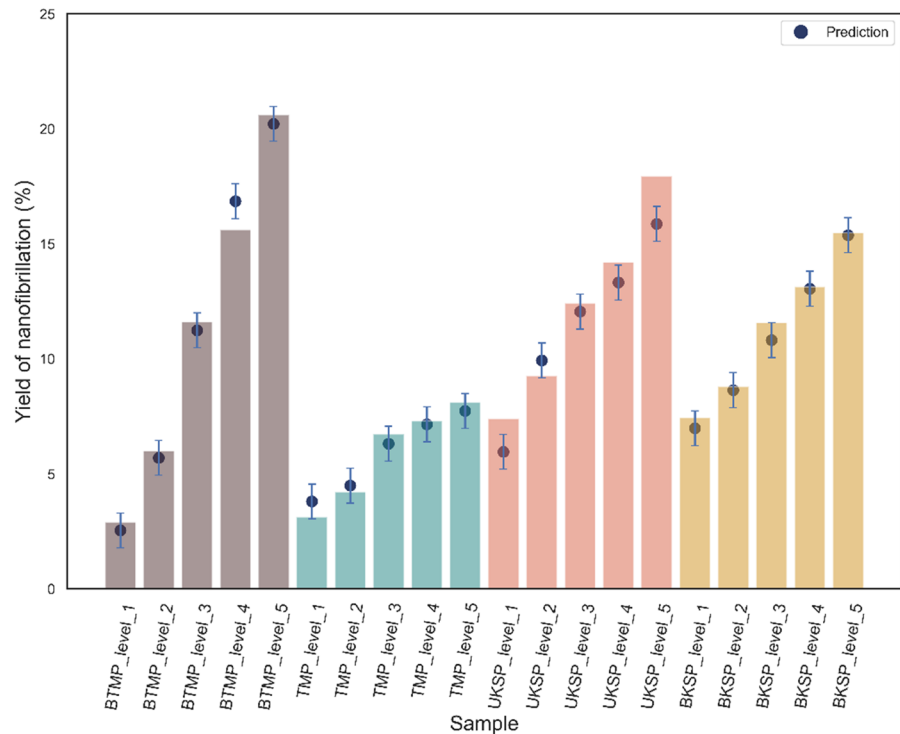


**Fig. 9** Target versus LR prediction values for yield of nanofibrillation (%)





**Fig. 10** Target versus ANN prediction values for yield of nanofibrillation (%)



**Table 9** Performance metrics for the three ML models regarding yield of nanofibrillation prediction

Model	MAE (%)	MAPE (%)	RMSE (%)	R <sup>2</sup>
RF	1.358	20.79	1.712	0.8720
LR	1.101	16.61	1.306	0.9255
ANN	0.577	6.74	0.758	0.9749

different treated pulps. Here, it is possible to state again that an increase in the complexity of the model resulted in a better tendency prediction. But this time, RF model did the poorest job, while LR operated reasonably and ANN performed very well. This same order is respected when it comes to comparing the prediction quality metrics.

Regarding the usefulness of these models, they could be applied in processes where either stable dispersions, homogeneous distribution, and low particle size are required, or in the opposite case, e.g., where nanocelluloses are used as adsorbents and need to be recovered by sedimentation or other means (Ho and Leo 2021). Indirectly, it is known that the yield of nanofibrillation exerts a positive influence on the tensile strength of nanocellulose-based films (Espinosa

et al. 2020) and nanocellulose-reinforced paper sheets (Ehman et al. 2020).

## Conclusions

In the current study, three different machine learning algorithms, namely multiple linear regression (LR), random forests (RF) and artificial neural networks (ANN), were applied in order to predict two important nanocellulose properties: aspect ratio and yield of nanofibrillation. Features required for obtaining satisfactory results were related mainly to the chemical composition of the pre-treated pulps, HPH intensity, and easy-to-measure nanocellulose parameters. The findings of the current work show that one can use different machine learning models to obtain pieces of information on yield and aspect ratio depending on the availability of measurements and sensors. Regarding the particularities of each property, the yield of nanofibrillation was more challenging to model than the aspect ratio of (L) CMNFs. Even though all models showed good performances, ANN outplayed RF and LR models in terms of the correlation coefficient (0.975 for yield,

0.96 for aspect ratio), MAE, MAPE, and RMSE. These results might be due to the greater complexity of the ANN model. However, the random forest approach used two factors less in the case of aspect ratio and one less in the case of yield of nanofibrillation, while still attaining an acceptable prediction power, at least for the morphological property. The overall indications are that machine learning algorithms have great potential in predicting these two important and often overlooked parameters for (L)CMNF suspensions, since very good results were obtained from relatively simple models and easily obtainable and low-cost inputs, not requiring advanced analytical techniques. The models obtained so far are a first step towards the aim of better controlling the production of nanocellulose. A limitation that must be highlighted here is the number of samples, which is quite low. Therefore, future steps involve the inclusion of more data to continuously improve the prediction and generalization capacity of the models. Also, it would be interesting to include in the model different types of pulps, coming from both wood and non-wood feedstock, and other types of pre-treatment.

**Acknowledgments** Authors wish to acknowledge the financial support of the funding agencies listed in the “Funding” section for their support to develop the present work. Authors’ information: Marc Delgado-Aguilar and Quim Tarrés are Serra Hunter Fellows.

**Authors’ contributions** All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by GSI, RA and QT. The first draft of the manuscript was written by GSI and RA, and all authors commented on previous versions of the manuscript. AFS, MLC, and MD-A compiled all comments and prepared the final version of the manuscript, which was read and approved by all the authors.

**Funding** Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. The present work has been funded by the Spanish Ministry of Science and Innovation, projects CON-FUTURO-ES (PID2020-113850RB-C22) and VALORCON-NC (PDC2021-120964-C22). Alexandre Ferreira Santos received funding from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil) for his research stay at University of Girona. The Human Resources Program of the National Agency of Petroleum, Natural Gas and Biofuels (PRH-ANP/Finep) from Brazil has economically supported the grant of Giovana S. Iamin.

**Data availability** Raw data of the present work can be made available upon request.

## Declarations

**Conflict of interest** Authors declare that they have no known competing interests or personal relationships that could have appeared to influence the work reported in this paper.

**Ethical approval and consent to participate** Authors declare that the manuscript is not submitted to any other journal at the time of submission for simultaneous consideration, that the submitted work is original and has not been published elsewhere in any form, that this work is not part of a single study, that results are presented under the principles of honesty, without fabrication falsification or inappropriate data manipulation and that no data, text or theories by others are presented as our own.

**Consent for publication** All authors have revised the last version of the submitted manuscript and we approve its submission.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Abdul Khalil HPS, Davoudpour Y, Islam MN et al (2014) Production and modification of nanofibrillated cellulose using various mechanical processes: a review. *Carbohydr Polym* 99:649–665. <https://doi.org/10.1016/j.carbpol.2013.08.069>
- Aguado R, Moral A, López P et al (2016) Morphological analysis of pulps from orange tree trimmings and its relation to mechanical properties. *Meas J Int Meas Confed*. <https://doi.org/10.1016/j.measurement.2016.06.063>
- Aguado R, Tarrés Q, Pèlach M et al (2022) Micro- and nanofibrillated cellulose from annual plant-sourced fibers: comparison between enzymatic hydrolysis and mechanical refining. *Nanomaterials* 12:1612. <https://doi.org/10.3390/nano12091612>
- Alila S, Besbes I, Vilar MR et al (2013) Non-woody plants as raw materials for production of microfibrillated cellulose

- (MFC): a comparative study. *Ind Crops Prod* 41:250–259. <https://doi.org/10.1016/j.indcrop.2012.04.028>
- Almonti D, Baiocco G, Tagliaferri V, Ucciardello N (2019) Artificial neural network in fibres length prediction for high precision control of cellulose refining. *Materials* 12:3730. <https://doi.org/10.3390/ma12223730>
- Besbes I, Alila S, Boufi S (2011) Nanofibrillated cellulose from TEMPO-oxidized eucalyptus fibres: effect of the carboxyl content. *Carbohydr Polym* 84:975–983. <https://doi.org/10.1016/j.carbpol.2010.12.052>
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32. <https://doi.org/10.1023/A:1010933404324>
- Dai Z, Ottesen V, Deng J et al (2019) A brief review of nanocellulose based hybrid membranes for CO<sub>2</sub> separation. *Fibers* 7:40. <https://doi.org/10.3390/fib7050040>
- Desmaisons J, Boutonnet E, Rueff M et al (2017) A new quality index for benchmarking of different cellulose nanofibrils. *Carbohydr Polym* 174:318–329. <https://doi.org/10.1016/j.carbpol.2017.06.032>
- Ehman NV, Felissia FE, Tarrés Q et al (2020) Effect of nanofiber addition on the physical–mechanical properties of chemimechanical pulp handsheets for packaging. *Cellulose* 27:10811–10823. <https://doi.org/10.1007/s10570-020-03207-5>
- Espinosa E, Rol F, Bras J, Rodríguez A (2020) Use of multifactorial analysis to determine the quality of cellulose nanofibers: effect of nanofibrillation treatment and residual lignin content. *Cellulose*. <https://doi.org/10.1007/s10570-020-03136-3>
- Géron A (2019) Hands-on machine learning with Scikit-learn, Keras, and TensorFlow, 2nd edn. O'Reilly Media, USA
- Haykin S (1999) *Neural networks: a comprehensive foundation*, 2nd edn. Pearson Prentice Hall, Singapore
- Ho NAD, Leo CP (2021) A review on the emerging applications of cellulose, cellulose derivatives and nanocellulose in carbon capture. *Environ Res* 197:111100. <https://doi.org/10.1016/j.envres.2021.111100>
- Isogai A (2020) Cellulose nanofibers: recent progress and future prospects. *J Fiber Sci Technol* 76:310–326. <https://doi.org/10.2115/fiberst.2020-0039>
- Kim JY, Zhou L, Morita N (2021) Study of degradable fibers with and without guar gum as a proppant transport agent using large-scale slot equipment. *SPE J* 26:262–280. <https://doi.org/10.2118/195808-PA>
- Klemm D, Cranston ED, Fischer D et al (2018) Nanocellulose as a natural source for groundbreaking applications in materials science: today's state. *Mater Today* 21:720–748. <https://doi.org/10.1016/j.mattod.2018.02.001>
- Kriechbaum K, Munier P, Apostolopoulou-Kalkavoura V, Lavoine N (2018) Analysis of the porous architecture and properties of anisotropic nanocellulose foams: a novel approach to assess the quality of cellulose nanofibrils (CNFs). *ACS Sustain Chem Eng* 6:11959–11967. <https://doi.org/10.1021/acssuschemeng.8b02278>
- Lee KY, Aitomäki Y, Berglund LA et al (2014) On the use of nanocellulose as reinforcement in polymer matrix composites. *Compos Sci Technol* 105:15–27. <https://doi.org/10.1016/j.compscitech.2014.08.032>
- Leong SL, Tiong SIX, Siva SP et al (2022) Morphological control of cellulose nanocrystals via sulfuric acid hydrolysis based on sustainability considerations: an overview of the governing factors and potential challenges. *J Environ Chem Eng* 10:108145. <https://doi.org/10.1016/j.jece.2022.108145>
- Li Z, Zhang H, Wang X et al (2016) Further understanding the response mechanism of lignin content to bonding properties of lignocellulosic fibers by their deformation behavior. *RSC Adv* 6:109211–109217. <https://doi.org/10.1039/C6RA22457A>
- Maulud D, Abdulazeez AM (2020) A review on linear regression comprehensive in machine learning. *J Appl Sci Technol Trends* 1:140–147. <https://doi.org/10.38094/jastt1457>
- Movsesyan A, Besteiro LV, Wang Z, Govorov AO (2022) Mie sensing with neural networks: recognition of nano-object parameters, the invisibility point, and restricted models. *Adv Theory Simul* 5:2100369. <https://doi.org/10.1002/adts.202100369>
- Mugwagwa LR, Chimphango AFA (2022) Predicting mechanical properties of hemicellulose-based films reinforced with acetylated nanocellulose. *J Polym Res* 29:228. <https://doi.org/10.1007/s10965-022-03092-5>
- Nader S, Brosse N, Khadraoui M et al (2022) A low-cost environmentally friendly approach to isolate lignin containing micro and nanofibrillated cellulose from Eucalyptus globulus bark by steam explosion. *Cellulose* 29:5593–5607. <https://doi.org/10.1007/s10570-022-04632-4>
- Özkan M, Karakoç A, Borghei M et al (2019) Machine learning assisted design of tailor-made nanocellulose films: a combination of experimental and computational studies. *Polym Compos* 40:4013–4022. <https://doi.org/10.1002/polb.25262>
- Pennells J, Heuberger B, Chaléat C, Martin DJ (2022) Assessing cellulose micro/nanofibre morphology using a high throughput fibre analysis device to predict nanopaper performance. *Cellulose* 29:2599–2616. <https://doi.org/10.1007/s10570-021-04405-5>
- Raj P, Mayahi A, Lahtinen P et al (2016) Gel point as a measure of cellulose nanofibre quality and feedstock development with mechanical energy. *Cellulose* 23:3051–3064. <https://doi.org/10.1007/s10570-016-1039-2>
- Sanchez-Salvador JL, Monte MC, Batchelor W et al (2020) Characterizing highly fibrillated nanocellulose by modifying the gel point methodology. *Carbohydr Polym* 227:115340. <https://doi.org/10.1016/j.carbpol.2019.115340>
- Sanchez-Salvador JL, Campano C, Lopez-Exposito P et al (2021) Enhanced morphological characterization of cellulose nano/microfibers through image skeleton analysis. *Nanomaterials*. <https://doi.org/10.3390/nano11082077>
- Sanchez-Salvador JL, Campano C, Negro C et al (2021) Increasing the possibilities of TEMPO-mediated oxidation in the production of cellulose nanofibers by reducing the reaction time and reusing the reaction medium. *Adv Sustain Syst* 5:2000277. <https://doi.org/10.1002/advs.20200277>
- Sanchez-Salvador JL, Monte MC, Negro C et al (2021) Simplification of gel point characterization of cellulose nano and microfibr suspensions. *Cellulose* 28:6995–7006. <https://doi.org/10.1007/s10570-021-04003-5>
- Santos AF, Aguado R, Corazza ML et al (2022) Artificial neural network for aspect ratio prediction of lignocellulosic

- micro/nanofibers. *Cellulose* 29:5609–5622. <https://doi.org/10.1007/s10570-022-04631-5>
- Segal L, Creely JJ, Martin AE, Conrad CM (1959) An empirical method for estimating the degree of crystallinity of native cellulose using the X-ray diffractometer. *Text Res J* 29:786–794. <https://doi.org/10.1177/004051755902901003>
- Serra-Parareda F, Aguado R, Tarrés Q et al (2021a) Chemical-free production of lignocellulosic micro-and nanofibers from high-yield pulps: synergies, performance, and feasibility. *J Clean Prod* 313:127914. <https://doi.org/10.1016/j.jclepro.2021.127914>
- Serra-Parareda F, Aguado R, Tarrés Q et al (2021b) Potentiometric back titration as a robust and simple method for specific surface area estimation of lignocellulosic fibers. *Cellulose* 28:10815–10825. <https://doi.org/10.1007/s10570-021-04250-6>
- Serra-Parareda F, Tarrés Q, Mutjé P et al (2021c) Correlation between rheological measurements and morphological features of lignocellulosic micro/nanofibers from different softwood sources. *Int J Biol Macromol* 187:789–799. <https://doi.org/10.1016/j.ijbiomac.2021.07.195>
- Serra-Parareda F, Aguado R, Arfelis S et al (2022) Techno-economic and environmental evaluation of a market pulp reinforced with micro-/nanofibers as a strengthening agent in packaging paper. *J Clean Prod* 347:131265. <https://doi.org/10.1016/j.jclepro.2022.131265>
- Sluiter A, Hames B, Ruiz R et al (2012) Determination of structural carbohydrates and lignin in biomass
- TAPPI (2020) TAPPI standards, technical information papers, and useful methods. Technical Association of the Pulp & Paper Industry, New York
- Tarrés Q, Oliver-Ortega H, Alcalà M et al (2020) Research on the strengthening advantages on using cellulose nanofibers as polyvinyl alcohol reinforcement. *Polymers*. <https://doi.org/10.3390/POLYM12040974>
- Tarrés Q, Oliver-Ortega H, Boufi S et al (2020) Evaluation of the fibrillation method on lignocellulosic nanofibers production from eucalyptus sawdust: a comparative study between high-pressure homogenization and grinding. *Int J Biol Macromol* 145:1199–1207. <https://doi.org/10.1016/j.IJBIOMAC.2019.10.046>
- Torrents-Barrena J, Pellicer E (2021) Smart cellulose composites: advanced applications and properties prediction using machine learning. In: Reference module in materials science and materials engineering
- Varanasi S, He R, Batchelor W (2013) Estimation of cellulose nanofibre aspect ratio from measurements of fibre suspension gel point. *Cellulose* 20:1885–1896. <https://doi.org/10.1007/s10570-013-9972-9>
- Wang T, Wang X, Luo Z et al (2011) Mechanisms of viscosity increase for nanocolloidal dispersions. *J Nanosci Nanotechnol* 11:3141–3150. <https://doi.org/10.1166/jnn.2011.3613>
- Wang W, Gu F, Deng Z et al (2021) Multilayer surface construction for enhancing barrier properties of cellulose-based packaging. *Carbohydr Polym* 255:117431. <https://doi.org/10.1016/j.carbpol.2020.117431>
- Xu Y, Kuang Y, Salminen P, Chen G (2016) The influence of nano-fibrillated cellulose as a coating component in paper coating. *BioResources* 11:4342–4352

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.