Predicting tobacco pyrolysis based on chemical constituents and heating conditions using machine learning approaches

Hao Wei[a], Jiangkuan Xing[a,*], Kun Luo[a], Yuhan Peng[b], Jianren Fan[a], Ke Zhang[c], Hui Wang[b,*]

[a] *State Key Laboratory of Clean Energy Utilization, Zhejiang University, Hangzhou, 310027, China*
[b] *China Tobacco Zhejiang Industrial Co., Ltd.,Hangzhou 310088, P. R. China*
[c] *Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou 450001, P. R. China*

Abstract

Tobacco is a special type of biomass that consists of complex chemical constituents. Currently, only global kinetic models have been developed for tobacco pyrolysis, but accurate kinetics considering the effects of the complex chemical constituents and heating conditions have not been well established. To this end, a general tobacco pyrolysis model was developed based on the complex chemical constituents and heating conditions using machine learning approaches. Specifically, chemical analysis and thermogravimetric analysis (TGA) of 49 tobacco samples under a wide range of heating rates were first conducted by experiments and then used to construct a database for the model development. Subsequently, the constructed database was divided into seen and unseen data-sets for the model development and evaluation. General pyrolysis models for single/multiple heating rates were developed from the seen data-set using an advanced machine learning approach, the Extremely Randomized Trees (Extra-Trees, ET). The performances of mod-

els were further evaluated on the unseen data-set through comparisons with the experimental data. The results showed that after feature selection based on Pearson correlation coefficient and hyper-parameters optimization, the trained models could accurately reproduce the tobacco pyrolysis behaviour on the unseen data with $R^2 > 0.967$ based on a single heating rate and with $R^2 > 0.974$ based on all heating rates. In addition, the predicted derivative thermogravimetry (DTG) profiles were integrated to obtain the TGA profiles, and the results agreed very well with the experimental data ($R^2 > 0.99$).
Keywords: Tobacco, Pyrolysis, Extra-Trees(ET), Machine learning

Nomenclature

TGA	Thermogravimetric analysis
DTG	Derivative thermogravimetry
ANN	Artificial neural networks
RF	Random forest
EC	Empirical correlations
ET	Extremely randomized trees
n_estimators	The number of trees in the forest
max_depth	TGA The maximum depth of each tree
TS	Total sugar
TA	Total alkaloids
RS	Reducing sugar
CL	Chlorine

K	Potassium
N	Nitrogen
CH	Chlorogenic acid
SC	Scopoletin
RU	Rutin
NE	Neochlorogenic acid
CR	Cryptochlorogenic acid
ST	Starch
PH	pH value
HR	Heating rate
T	Temperature
PCC	Pearson correlation coefficients
Mean	The average value
STD	The standard deviation
Min	The minimum value
Max	The maximum value
NRMSE	Normalized root mean square error
RMSE	Root mean square error

1. Introduction

It is a common sense that tobacco use is not good for health [1]. However, there is still a great deal of tobacco production, approximately 6.68 million metric tons in 2019 worldwide [2]. Thus, there is an urgent need to find other uses for tobacco instead of producing cigarettes. The compositions of

tobacco are cellulose, hemicellulose, lignin, and volatile component [3, 4, 5], which are different from the lignocellulosic biomass that is mainly composed of cellulose, hemicellulose, and lignin [6]. Although tobacco, as a special kind of biomass, has different consistent of lignocellulosic biomass. It could also be utilized by thermo-chemical and bio-chemical routes [7]. Among the commonly-used thermo-chemical technologies, such as gasification, combustion, and carbonation [3, 8, 9], pyrolysis is always the initial and fundamental process. Therefore, it is of great significance to comprehensively study the pyrolysis behaviour of tobacco and then accurately model this process for developing healthy and efficient tobacco utilization technologies.

There have been some prior studies reported for tobacco pyrolysis. Experimental methods, such as gas chromatography/mass spectrometry [3, 10], thermogravimetric analyzer coupled with Fourier transform infrared spectrometry [10, 11, 12], thermogravimetric mass spectrometry [3, 10] and macro-thermogravimetric [13], were often used to study the pyrolysis behavior of tobacco, including the effect of oxygen on in-situ evolution of chemical structures during tobacco pyrolysis [3], the pyrolysis characteristics and major pyrolysis products [10], the effect of different tobacco particle sizes on pyrolysis [14], and the formation mechanism of the volatile products [13]. Contrarily, there have been only a few studies on establishing tobacco pyrolysis models. Chen et al. [8] proposed a mathematical model for smouldering cigarettes, including the evaporation of water, the pyrolysis of virgin tobacco and char combustion processes. However, the model only worked for the selected tobacco type, and its capability for various tobacco types was unclear. Encinar et al. [15] established a kinetic model for pyrolysis (including maize, sun-

flowers, grape, and tobacco) from experiments. In this kinetic model, the generation of volatile gases during the pyrolysis was assumed to be dominated by a series of independent first-order parallel reactions. Gao et al. [16] calculated the best linearity of different reaction models with the data from experiments as a reference. However, tobacco pyrolysis involves very complicated heterogeneous reactions, which relates to the tobacco type and heating conditions. Therefore, it is hard to use one reaction or simple linear reactions to represent the whole process. In summary, there is a lack of an accurate kinetic model that considers both the effects of the chemical composition of tobacco and the heating conditions.

The barrier to establishing such a kinetic model is the strong, complicated, and nonlinear relationship between the pyrolysis behavior and the complex chemical composition/heating conditions. Machine learning approaches have been proven to handle complex nonlinear issues competently, as demonstrated in many previous studies of solid fuel thermo-chemical conversion modelling. Abbas et al. [17] proposed an approach based on Artificial Neural Networks (ANN) to predict the volatiles released from coal and biomass. Yildiz et al. [18] studied the co-combustion of the hazelnut husk–lignite coal blends of various compositions by using ANN. Sunphorka et al. [19] modelled the biomass devolatilization process based on the three major components using ANN, and the obtained model showed better performance than conventional analytical method. Except for the ANN, ensemble learning approaches, such as Random Forest (RF), have also been successfully applied in volatile gases yields and compositions prediction [20, 21], bio-oil yield prediction [22], and biomass devolatilization kinetics prediction [23]. Xing et

al. [23] compared Empirical Correlations (EC), ANN, and RF in predicting the kinetic parameters of a single-step kinetic model for biomass pyrolysis, and found that RF gave the best performance. Wei et al. [24] compared RF and Extremely Randomized Trees (ET) in the modelling of biomass and coal co-pyrolysis.

Based on the above backgrounds, the objective of the present study is to establish a general tobacco pyrolysis model based on the chemical constituents and the heating conditions using an advanced machine learning approach, ET. Specifically, in the experimental part, the chemical analysis of 49 kinds of tobaccos were conducted, followed by the thermogravimetric analysis (TGA) of tobaccos under a wide range of heating rates. The derivative thermogravimetry (DTG) results were calculated by differentiating TGA. Subsequently, the chemical analysis results, the heating conditions, and the DTG results are collected to construct the database. Afterwards, the database is divided into train and test data-sets with a stratified sampling method for the model development and evaluation. General pyrolysis models were then developed based on the train data-set using ET. The feature selection and hyper-parameters optimization were conducted to get a better generality and avoid over-fitting. The model performance was further evaluated on the test data-set by comparing with the experimental data.

This work has the following novelty. Firstly, this is the first model directly established the complex relation between tobacco samples' chemical composition and the pyrolysis behaviors. Chemical analysis is used to describe the tobacco sample rather than proximate analysis and ultimate analysis. This is because that there is a strong relation between chemical analysis with the

quality of tobacco, which is very important for the tobacco industry. Secondly, the differential target was used to improve the accuracy, as a small error in TGA will be magnified to DTG by differentiating with a small step. Thirdly, the present work established a database with a large amount range of tobacco samples and heating rates, including 49 different kinds of tobacco samples, and each of them are under a wide range of heating rate (10, 50, 100, 150, 200, 250, 300, 350, and 400 K/min). This database contributes to the high generalization performance of model.

2. Materials and methods

In this section, the chemical analysis methods, the TGA experiment, and the machine learning approaches are introduced in turn.

2.1. Experimental methods

In total 49 tobacco samples were studied in the experiments. The chemical analysis information, including total sugar (TS), total alkaloids (TA), reducing sugar (RS), chlorine (Cl), potassium (K), nitrogen (N), chlorogenic acid (CH), scopoletin (SC), rutin (RU), neochlorogenic acid (NE), cryptochlorogenic acid (CR), starch (ST), and pH value (PH), were measured. The contents of TS, TA, RS, Cl, K, N, and ST were determined by a continuous flow analyzer (Alliance-Futura). The contents of CH, SC, RU, NE, and CR were measured using an Agilent Technologies 1260 High-Performance Liquid Chromatography system equipped with a DAD detector, in which the separation process was performed on a Symmetry C18 column and the detection wavelength was 340 nm. PH was measured by the Mettler-Toledo Seven Compact pH meter. The distribution densities of the chemical analysis of all

the 49 tobacco samples are shown in Fig. 1. The horizontal axis represents the feature distribution interval, and the vertical axis represents the distribution density. The curve in each sub-figure shows the distribution of the chemical analysis value. TS, TA, RS, CL, K, N, CH, SC, RU, NE, CR, ST and PH are within the ranges of 19.22-34.10 %, 1.48-3.27 %, 17.18-30.82 %, 0.07-1.10 %, 1.36-2.89 %, 1.63-2.42 %, 7.75-17.77 %, 0.14-049 %, 4.62-14.53 %, 1.13-2.92 %, 1.81-4.29 %, 3.41-7.17 % and 4.71-5.25, respectively. Table 1 lists some statistical information of the features, including the average value (Mean), the standard deviation (STD), the minimum (Min) and the maximum (Max). Obviously, TS has the largest mean value while SC has the smallest. STD is a measure of the amount of variation of a set of values, which lower STD means the values are closer to the average value. Except for TS, RS, CH and RU, other features are closer to their average value. More details are presented in supplementary materials. In total, there are 13 features used to describe the chemical constituents information of tobaccos, which are different from the traditional features used in biomass pyrolysis modelling such as proximate analysis and ultimate analysis. Chemical analysis is used to describe the tobacco sample rather than proximate analysis and ultimate analysis. This is because that there is a strong relation between chemical analysis with the quality of tobacco, which is very important for the tobacco industry[25].

Same as our previous study [4], all thermogravimetric analysis (TGA) of tobacco samples were conducted by using a *Discovery* thermogravimetric analyzer produced by TA instruments, as shown in Fig. 2, in the following way. Firstly, the tobacco samples were grounded into powders, and the particles

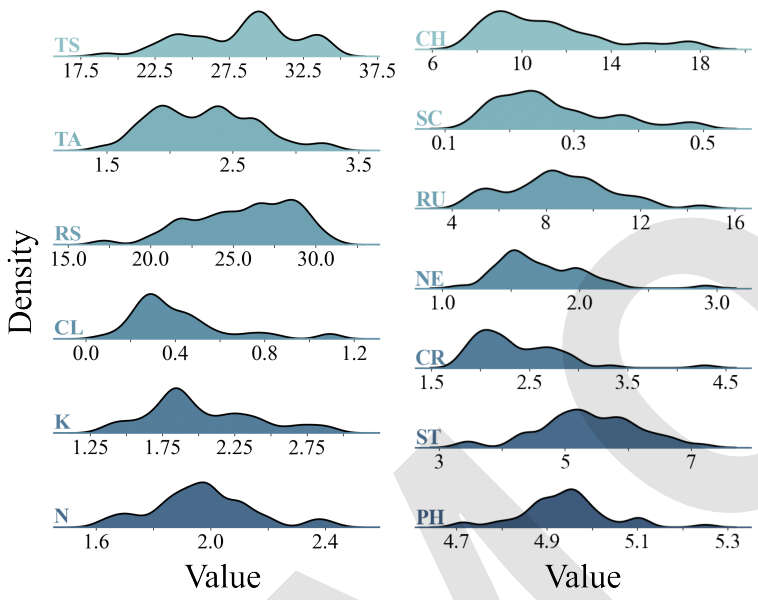

Figure 1: Distribution of tobacco chemical analysis. On the left side: TS, VA, RS, CL, K, N; on the right side: CH, SC, RU, NCH, ICH, ST, PH.

that were trapped on a 60-mesh sieve after passing through a 40-mesh sieve were collected in a sealed valve bag for the subsequent measurements. The size of tobacco particles is within the range of 0.250-0.425 mm. Secondly, 8 mg samples were dehydrated by heating to 373 K at a heating rate of 10 K/min from room temperature and holding for 30 minutes. The weight of samples after dehydration was set as the initial mass of the TGA measurement. Thirdly, the dehydrated samples were then devolatilized at different heating rates (10, 50, 100, 150, 200, 250, 300, 350, and 400 K/min) from 373 K to 1173 K, and their mass loss with the increasing temperature was recorded. Among all tests, the flow rates of the carrier gas (high-purity N_2) and protective gas (high-purity N_2) were set at 50 mL/min and 30 mL/min, respectively. Each test has been repeated three times to ensure the repeata-

Table 1: The sample distribution of the test data-sets.

	Mean	STD	Min	Max
TS	28.31	3.66	19.22	34.10
TA	2.27	0.41	1.48	3.27
RS	25.67	3.10	17.18	30.82
CL	0.41	0.22	0.07	1.10
K	2.02	0.38	1.36	2.89
N	1.97	0.17	1.63	2.42
CH	11.04	2.65	7.75	17.77
SC	0.26	0.09	0.14	0.49
RU	8.54	2.20	4.62	14.53
NE	1.71	0.32	1.13	2.92
CR	2.37	0.46	1.81	4.29
ST	5.38	0.79	3.41	7.17
PH	4.94	0.10	4.71	5.25

bility of the results. Note that keeping the same sampling points during the experiment at both low and high heating rates is hard. Therefore, in order to avoid the uncertainty caused by different sampling points during the experiment, all the TGA results are interpolated into 500 points, and then the corresponding DTG results are calculated by differentiating TGA profiles based on particle temperature with Eq. (1),

$$DTG = \frac{TGA^{(i)} - TGA^{(i-1)}}{T^{(i)} - T^{(i-1)}}, \tag{1}$$

where $TGA^{(i)}$ and $T^{(i)}$ represent the mass loss of tobacco and local particle temperature at point i, respectively. $TG^{(i-1)}$ and $T^{(i-1)}$ represent the mass loss of tobacco and local particle temperature at point $(i-1)$, respectively.

Worth noting that the present work is remarkably different from the pre-

Figure 2: Discovery thermogravimetric analyzer.

vious studies of pyrolysis modelling directly based on TGA profiles. On the one side, empirically, building model follows the experimental way, which first measures TGA by thermogravimetric analyzer then calculates DTG by differentiating TGA using Eq. 1. On the other side, TGA is easier to model than DTG owing to its simpler curve characteristics. However, there could be a risk that some information would be lost if a big temperature step is chosen in the calculation from TGA to DTG. Therefore, a small temperature step is usually preferred. A small difference in TGA will be magnified to DTG by differentiating with a small step. Alternatively, owing to the value of DTG being very small (usually in the range from 0 to -1), when integrating DTG to TGA, the error will be reduced. Moreover, as mentioned before, pyrolysis is the initial and fundamental process of the thermochemical conversion technologies of tobacco, in which the pyrolysis model provides two-way coupling terms between the gas and particle phases in the numerical simulations [26, 27]. Based on all discussion above, for avoiding the possible uncertainties caused by differentiating TGA results and the convenience of

the model implementation in numerical simulations, the present work aims to establish general tobacco pyrolysis models using the DTG results considering the complex chemical constituents and heating conditions.

2.2. Machine learning approaches

The Extremely Randomized Trees (or called as Extra-Trees, ET) method [28] is used to describe the complex nonlinear correlations between the pyrolysis behavior and chemical constituents/heating conditions. The schematic topological of the ET algorithmic can be seen in Fig. 3. This method has been successfully applied for coal and biomass co-pyrolysis modelling in our previous study [24], and therefore is briefly introduced here. Firstly, the whole samples in the train data-set are inputted as learning samples rather than using a bootstrap replica like the traditional methods (such as the random forest). Then the trees will grow by splitting into two nodes as shown in Fig. 3 with a threshold. In the traditional methods, the thresholds are calculated, and then the best one is chosen. For ET, the splitting of each tree is based on a random threshold. When the samples of the node are less than the minimum sample size, the splitting stops. Finally, the predictions of the trees are aggregated by an arithmetic average to yield the final prediction.

3. Results and discussions

In this section, firstly, the training processes, including feature selection and hyper-parameters optimization, are introduced. Then the performance of models on the train and test data-sets, and the further application from the predicted DTG profiles to TGA profiles, are discussed in turn.

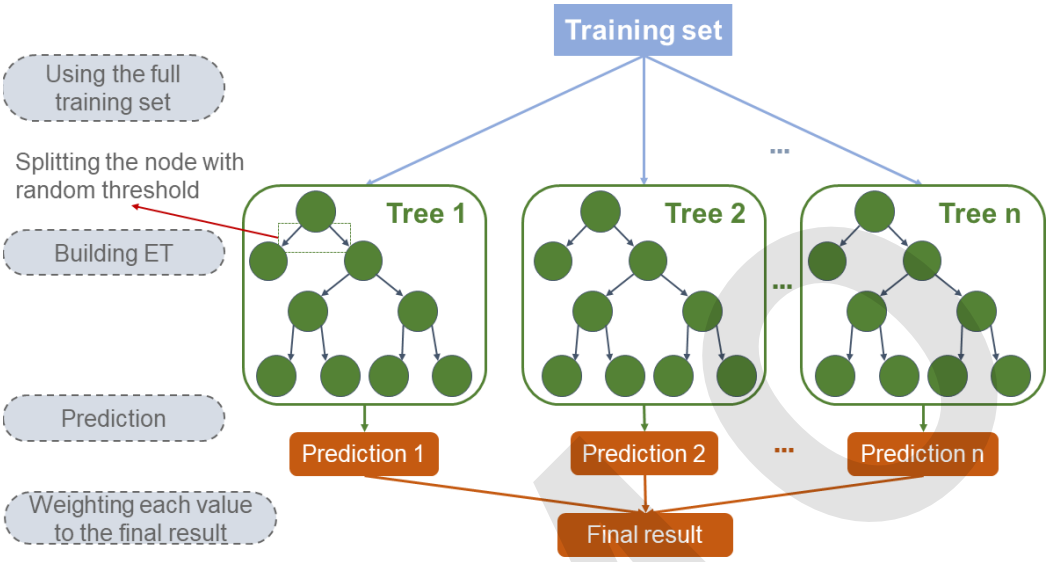

Figure 3: The schematic topological of the ET algorithmic [24].

3.1. Database information

In order to avoid the risk of introducing significant sampling bias when creating the test data-set by a purely random sampling method, herein, the stratified sampling method based on the difference of tobaccos is used[29]. In order to evaluate the difference of tobaccos, a parameter named Normalized Root Mean Square Error (NRMSE) is defined, which can be calculated by Eqs. (2) and (3),

$$RMSE_n = \left(\frac{1}{N} \sum_{i=1}^{N} \left(y_{1,\mathrm{exp}}^{(i)} - y_{\mathrm{n,exp}}^{(i)} \right)^2 \right)^{\frac{1}{2}}, \tag{2}$$

$$NRMSE = \frac{RMSE_n - RMSE_{n,min}}{RMSE_{n,max} - RMSE_{n,min}}, \tag{3}$$

where $y_{1,\mathrm{exp}}^{(i)}$ is the experimental value of tobacco 1 (base sample), and $y_{\mathrm{n,exp}}^{(i)}$ is the experimental value of tobacco n from 2 to 49. The base sample used in

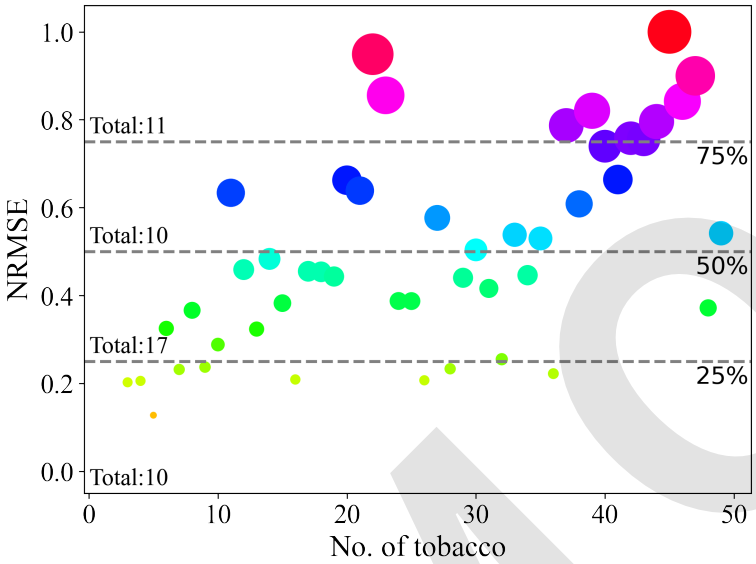

Figure 4: Results of the stratified sampling method, in which the size and color of the scatter vary with the values of NRMSE for a clearer observation.

the calculation is the experimental result of tobacco No.1 under 400 K/min, and the differences between the other 48 samples under 400 K/min with tobacco No.1 are shown in Fig. 4. The reason for choosing 400 K/min is that with the increase of heating rate, pyrolysis becomes more complicated and different [4], and the highest heating rate in the present work is 400 K/min. The different sizes and colors are used for a clearer observation. The size and color of the circles are determined by the NRMSE. The larger size stands for the larger NRMSE. Three grey dotted lines are used to divide the area (NRMSE from 0 to 1) into four equal parts, and the total sample numbers of each part are also shown. All the samples could be categorized into four types using the NRMSE values of 25%, 50%, and 75%. With this treatment, samples in each type are then randomly selected for training at a ratio of

70%, which could reflect the differences of tobacco pyrolysis behaviors and further improve the model performances. The detailed information is shown in Table 2.

Table 2: The sample distribution of the test data-sets.

$NRMSE$	Test	Total
0-25%	2,10,36	3
25-50%	13,15,19,24,31	5
50-75%	12,20,27	3
75-100%	22,37,40,42	4

3.2. Training process

After the train and test data-sets are chosen, the feature selection and hyper-parameters optimization are conducted to train and optimize the model. Worth noting that the present study established models not only based on different heating rates (from 10 K/min to 400 K/min) but also on all heating rates. For convenience, the model built on the individual heating rate is named with the heating rate, and the general model built on all heating rates is named "all".

3.2.1. Feature selection

As mentioned before, there are 13 features to describe the chemical components of tobacco, as called the features of fuel properties, and 2 features, temperature (T) and heating rate (HR), to describe the heating conditions. In order to take a deep sight into the correlations between features of tobaccos, the Pearson correlation coefficients (PCC) [30, 31] are determined

to find the relevance between them. The number in each grid is the PCC calculated by Eqs. (4) and (5),

$$\rho_{(X,Y)} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}, \tag{4}$$

$$\mathrm{cov}(X,Y) = \mathbb{E}\left[(X-\mu_X)\,(Y-\mu_Y)\right], \tag{5}$$

where X and Y are the values of two different features, and $cov(X,Y)$ is the covariance between X and Y. σ_X, σ_Y, μ_X and μ_Y is the standard deviation of X, the standard deviation of Y, the mean of X and the mean of Y, respectively.

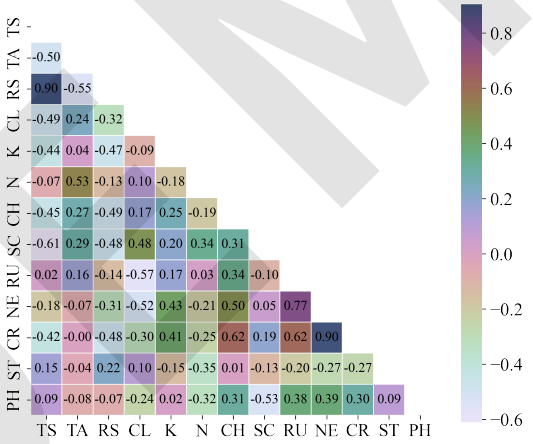

Figure 5: Pearson correlation coefficients between chemical constituents features of tobacco.

The PCC represents the strength and direction of linear relationships between pairs of features. According to Eqs. (4) and (5), The PCC of certain two features measure their linear correlation and has a value from -1 to 1. Positive and negative values of PCC represent the positive and negative

correlations respectively, and the absolute values represent the strength of the correlation (the large values, the stronger correlation). For the training of machine learning models, it is recommended to select input variables that are not highly correlated to decrease the nonlinear degree of the system. For example, TS and RS are highly correlated since their PCC has a value of 0.93. So, the RS is deleted and the TS is kept. CR is also highly correlated with NE, CH, and RU, with PCC values equaling to 0.9, 0.9, 0.62, and 0.62, respectively. So, the NE, CH, and RU are deleted and CR is kept. Finally, the total number of features for the following machine learning modelling is 11, including 9 fuel properties features and 2 heating condition features.

3.2.2. Hyper-parameters optimization

Another important method to avoid over-fitting and reduce computational cost is hyper-parameters optimization. There are several hyper-parameters for ET in *scikit learn* [32], including $n\_estimators$ (the number of trees in the forest) and max_depth (the maximum depth of each tree). Herein, $n\_estimators$ and max_depth are determined by step-wise searching methods. The searching interval is 1 for both the two parameters, and the searching range is 1-150 and 1-30 for $n\_estimators$ and max_depth, respectively. The five-fold cross-validation method is used. This method randomly divides the train data-set into 5 sub-sets of equal size named folds. Each sub-set is used as a validation data-set to test the model, whereas the left 4 sub-sets are used for training [29]. Model performance on the validation data-set is used to determine the optimal hyper-parameters. For brevity, here only hyper-parameters optimization results of the model based on data

of tobacco pyrolysis under all heating rates are shown in Fig. 6, in which the arrows indicate the chosen parameters. It is found that with the increase of $n\_estimators$ and max_depth, the training and validation RMSEs sharply decrease and stabilize at a critical value, which means that the model performance would not be improved with the further increase of $n\_estimators$ and max_depth. Worth noting that there is a zig-zag pattern in the case of $n\_estimators$ as compared to max_depth optimization. This is because that the $n\_estimators$ has a relatively larger influence than that of max_depth in the Extra-tree algorithm. So, when both hyper-parameters begin to increase in the first place, n-estimation shows a zig-zag pattern and becomes smooth later. Therefore, the model with the critical hyper-parameters values (70 and 14 for $n\_estimators$ and max_depth) is chosen and retained in the following discussions.

3.3. Model performance

In this section, the performances of models on both the train and test data-sets are presented and discussed. It's worth noting that all the tobacco on the test data-set is different from the train data-set, so they are totally unseen to the trained models. The $RMSE$ and R^2 calculated by Eqs. (6), (7) and (8) are used to measure the performance of models,

$$RMSE = \left(\frac{1}{N} \sum_{i=1}^{N} \left(y_{\text{exp}}^{(i)} - y_{\text{pred}}^{(i)} \right)^2 \right)^{\frac{1}{2}}, \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(y_{\text{pred}}^{(i)} - y_{\text{exp}}^{(i)} \right)^2}{\sum_{i=1}^{N} \left(y_{\text{exp}}^{(i)} - \hat{y}_{\text{exp}} \right)^2}, \tag{7}$$

$$\hat{y}_{\text{exp}} = \frac{1}{N} \sum_{i=1}^{N} y_{\text{exp}}^{(i)}, \tag{8}$$

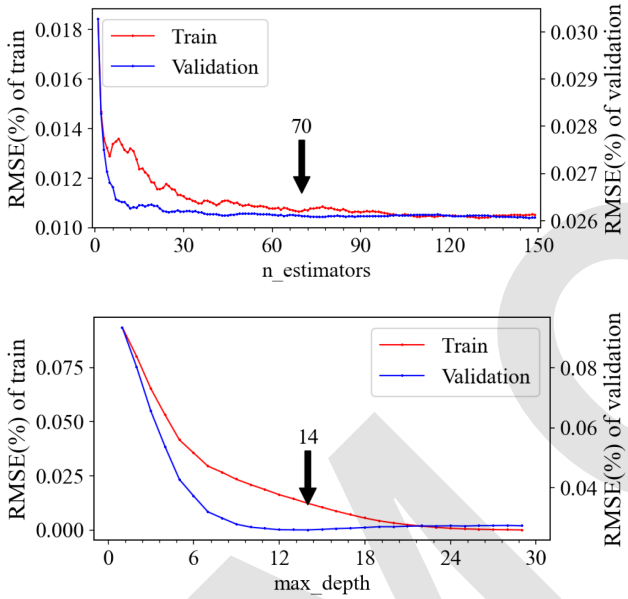

Figure 6: *N_estimators* and *max_depth* optimization results of the trained model based on data from all heating rates.

where $y_{\text{exp}}^{(i)}$ is the experimental value, $y_{\text{pred}}^{(i)}$ is the predicted value and $\hat{y}_{\text{exp}}$ is the average experimental value.

Figure 7 presents models performances on the train and test data-sets under heating rates of 10 K/min, 400 K/min, and all heating rates (including 10 K/min and 400 K/min). In which x axis stands for the true value while y axis stands for the predicted value from model, and different colors are used to present different heating rates (red, blue, and green for heating rate = 10, 400 K/min and all heating rates). In each subplot, the black dotted line is the best fitting line ($R^2 = 1$) and the dark line is the regression line of current data. Obviously, model based on 10 K/min shows the best performance, with R^2 on train test close to 1. While model based on 400 K/min shows

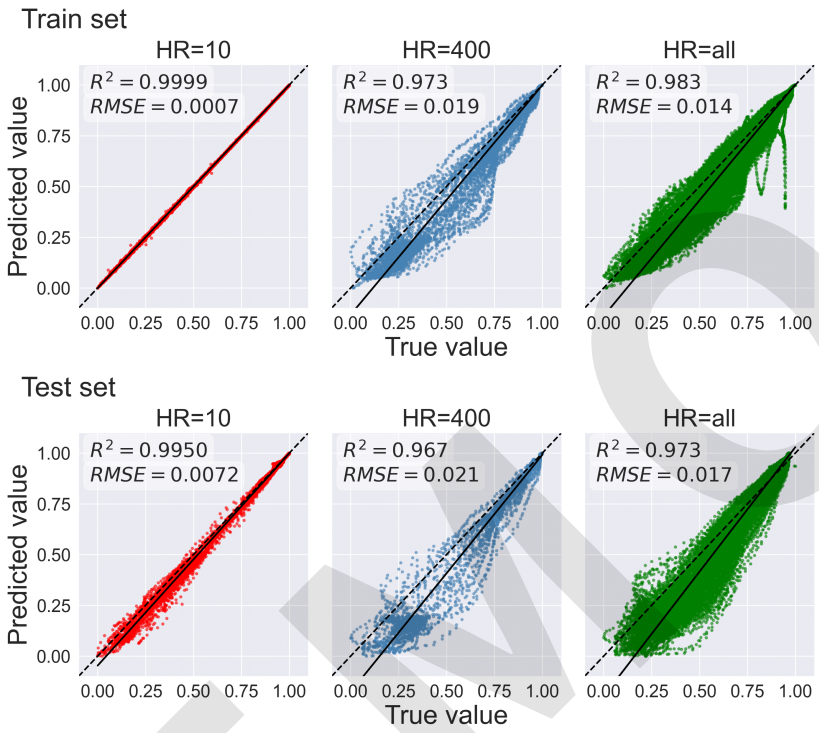

Figure 7: Model performances on the train and test data-sets under heating rates of 10 K/min, 400 K/min, and all heating rates.

the worst performance, which is even worse than model with all heating rate. More details are presented in Table 3, which lists the performance of models on train and test data-sets under different heating rates. It is found that 1) Models are not over-fitted. In all heating rates, the performances on the train data-set are comparable with those on the test data-set. The R^2 and $RMSE$ of predictions on the train data-set are only 0.009 larger and 0.0036 smaller than those on the test data-set at most. 2) Models perform well for both the train and test data-sets. The R^2 and $RMSE$ of the predictions of all the models on the test data-set are larger than 0.967 and smaller than 0.0209,

respectively. The best R^2 even equals to 0.995. 3) Inputting all heating rates to train the model would lead to a good model generality. The performance of "all" is not the best, but it's better than some high heating rate models, which indicates that the data-driven characteristics of the machine learning approaches and more data inputting will improve the model's generality. It's interesting to note that the predictions of the model based on 10 K/min have the best performance on the test data-set with a R^2 value of 0.995, which is much better than those of other models. With the increase of heating rate, the performance of the model decreases. For the train data-set, both R^2 and $RMSE$ decrease with the increase of heating rate, especially from 10 to 200 K/min. The test data-set shows the same trend. This could be attributed to the fact that with the temperature increases, pyrolysis process becomes more complicated with an aggregate phenomenon of the sub-peaks, which was discovered in the work of Mu et al. [4]. According to Mu et al. [4], the aggregate phenomenon under high heating rate is caused by the peak positions of volatile components movement. Because volatile components' pyrolysis is an exothermic reaction, the rate of heat accumulation around tobacco particles will increase accordingly with the increase of heating rate, which promotes the endothermic reaction and inhibits the exothermic reaction. In addition, the peaks in DTG curve are related to the pyrolysis process with different compositions of tobacco. Models for DTG curves with high accuracy can be further used in studying the pyrolysis process of different compositions of tobacco.

For a clearer assessment of model performance, both the best model (heating rate = 10K/min), and the worst model (heating rate=400 K/min) are

Table 3: Model performances on the train and test data-sets under different heating rates.

Heating rate	Train		Test	
(K/min)	R^2	$RMSE$	R^2	$RMSE$
10	0.999	0.0004	0.997	0.0060
50	0.992	0.0009	0.986	0.0119
100	0.985	0.0124	0.981	0.0140
150	0.989	0.0109	0.980	0.0147
200	0.981	0.0147	0.976	0.0165
250	0.973	0.0177	0.969	0.0201
300	0.970	0.0194	0.967	0.0203
350	0.973	0.0187	0.972	0.0109
400	0.970	0.0201	0.968	0.0208
All	0.990	0.0108	0.976	0.0164

further analyzed. Figures 8 and 9 show the comparisons between the DTG profiles predicted with the models and measured in the experiments for various tobacco samples from the test data-set. The red line and blue scatters stand for the predicted results and the experimental data, respectively. The No., R^2 and $RMSE$ are also shown in each subplot. From Figure 8, the predicted values of model based on 10 K/min show great agreement with experimental values, with all R^2s larger than 0.9909. It can be concluded that model on 10 K/min is at very high accuracy. Compared with model on 10 K/min, there is a slightly decrease of accuracy for model based on 400 K/min. Figure 9 demonstrates that predictions are not as good as those of former model, but generally in good agreement with the experimental data. The slight deviations are mainly concentrated in the peak area, where the temperature is within the range of 500 K to 700 K, and in other temperature

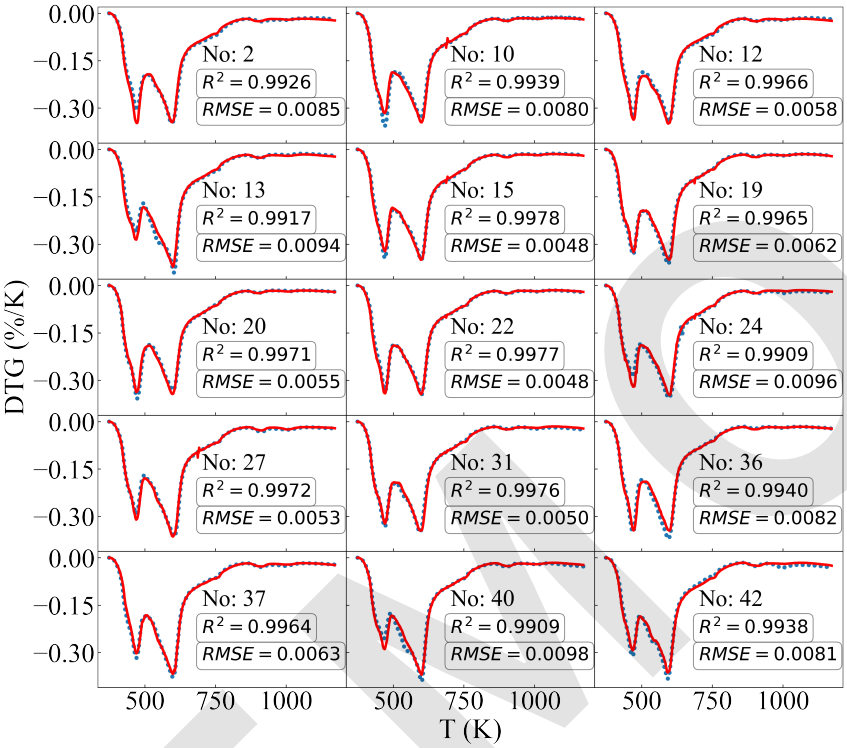

Figure 8: Comparisons between the DTG profiles predicted with the model and measured in the experiments for various tobacco samples at a heating rate of 10 K/min.

ranges, models performed well. In the peak area, the model can predict the peak position temperatures accurately. However, the predicted heights of the peaks are slightly lower than those in experiments.

To check how those deviations in the peak area affect the further calculations of TGA profiles, the predicted DTG at 400 K/min profiles are integrated based on temperature to obtain the TGA profiles. Figure 10 shows the comparisons between TGA profiles obtained from the predicted DTG profiles and measured in the experiments for various tobacco samples. The figure legend is the same as that in Fig. 9. Overall, the predicted TGA

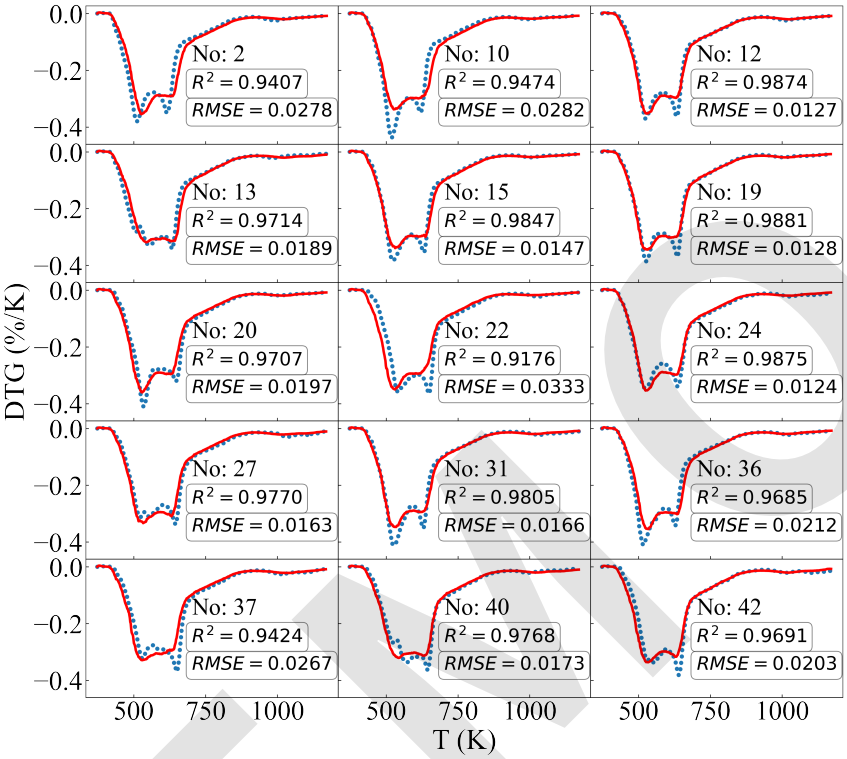

Figure 9: Comparisons between the DTG profiles predicted with the model and measured in the experiments for various tobacco samples at a heating rate of 400 K/min.

profiles are in a better agreement with the experimental data compared with those of the DTG profiles with all the R^2 larger than 0.99. This is because of the small DTG peak values and the slight deviations in the peak height prediction. In addition, the integration process also reduces the deviations statistics as it has been discussed in section 2.1. The peaks in the DTG profiles could be well reproduced with the developed models. Those peaks relate to the pyrolysis of different types of components, including the extractive content, cellulose, hemicellulose and lignin, as described in our previous study [1]. Those components could also be used as the inputs in the model

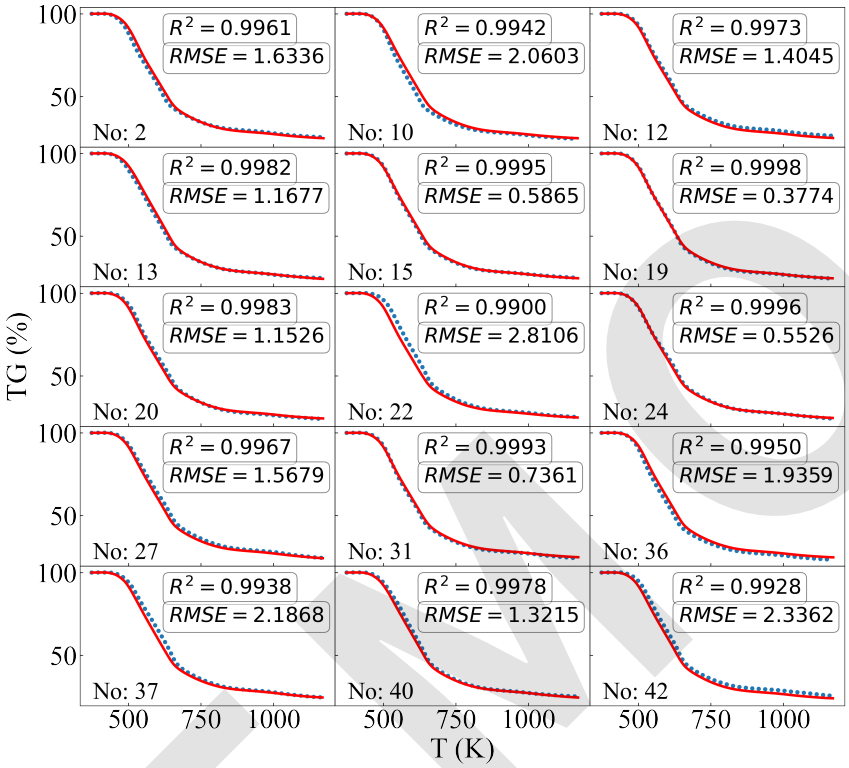

Figure 10: Comparisons between the TGA profiles integrated from the predicted DTG profiles and measured in the experiments for various tobacco samples at a heating rate of 400 K/min.

development, which will be further explored in the future study.

Present work currently has the following limitation. The heating rate range of the experimental equipment used in the present study (The discovery thermogravimetric analyzer produced by TA instruments) can't be over 400 K/min, so only the experiment under 400 K/min can be done. The model performance over 400 K/min is still unclear.

4. Conclusions

In the present study, a general tobacco pyrolysis model was developed based on the chemical constituents and heating conditions using the ET method. Specifically, thermogravimetric and chemical analysis of wide ranges of tobacco types and heating conditions are first conducted by experiments to construct a database for the model development. Subsequently, the constructed database has been divided into train and test data-sets for the model development and evaluation. General devolatilization models were then developed based on the train data-set using ET. The performances of models were further evaluated on the test data-set by comparing with the experimental data. The results showed that after feature selection based on Pearson correlation coefficient and hyper-parameters optimization, the trained model could well reproduce the tobacco pyrolysis behavior on the test data with $R^2 > 0.967$ based on a single heating rate and with $R^2 > 0.974$ based on all heating rates. In addition, the worst predicted DTG profiles were integrated based on particle temperature to obtain the TGA profiles, and the results showed a very well agreement with the experimental data ($R^2 > 0.99$). The present work concluded that the data-driven model for tobacco pyrolysis can achieve a very high accuracy and robustness, which provide a new way of modeling pyrolysis. In addition, setting DTG (differenced TG) as the target and then integrating it back to TG is a more accuracy for data-driven pyrolysis models since a small difference in TGA will be magnified to DTG by differentiating with a small step.

Acknowledgments

The authors are grateful for the support from National Natural and Science Foundation of China (Grant No. 51925603) and the support from Zhejiang University-Zhejiang China Tobacco Joint Laboratory Fund (Grant No. K-20201802).

References

[1] D. Hammond, Health warning messages on tobacco products: a review, Tobacco control 20 (5) (2011) 327–337.

[2] M. Shahbandeh, Topic: Tobacco industry.
URL https://www.statista.com/topics/1593/tobacco/#dossierKeyfigures

[3] Y. Peng, X. Hao, Q. Qi, X. Tang, Y. Mu, L. Zhang, F. Liao, H. Li, Y. Shen, F. Du, et al., The effect of oxygen on in-situ evolution of chemical structures during the autothermal process of tobacco, Journal of Analytical and Applied Pyrolysis 159 (2021) 105321.

[4] Y. Mu, Y. Peng, X. Tang, J. Ren, J. Xing, K. Luo, J. Fan, K. Zhang, Experimental and kinetic studies on tobacco pyrolysis under a wide range of heating rates, ACS Omega (2021) acsomega.1c06122.

[5] Y. J. Sung, Y. B. Seo, Thermogravimetric study on stem biomass of nicotiana tabacum, Thermochimica Acta 486 (1-2) (2009) 1–4.

[6] S. Wang, G. Dai, H. Yang, Z. Luo, Lignocellulosic biomass pyrolysis mechanism: a state-of-the-art review, Progress in energy and combustion science 62 (2017) 33–86.

[7] S. Wang, G. Dai, H. Yang, Z. Luo, Lignocellulosic biomass pyrolysis mechanism: A state-of-the-art review, Progress in Energy and Combustion Science 62 (2017) 33–86.

[8] P. Chen, A mathematical model of cigarette smoldering process, Beiträge Zur Tabakforschung International/Contributions to Tobacco Research 20 (4) (2002) 265–271.

[9] J. Xing, Y. Bai, C. Zhao, Z. Gao, H. Wang, Numerical studies of coal devolatilization characteristics with gas temperature fluctuation, Energy & Fuels 32 (8) (2018) 8760–8767.

[10] B. Liu, Y.-M. Li, S.-B. Wu, Y.-H. Li, S.-S. Deng, Z.-L. Xia, Pyrolysis characteristic of tobacco stem studied by py-gc/ms, tg-ftir, and tg-ms, BioResources 8 (1) (2013) 220–230.

[11] R. R. Baker, L. J. Bishop, The pyrolysis of tobacco ingredients, Journal of analytical and applied pyrolysis 71 (1) (2004) 223–311.

[12] E. Calabuig, N. Juárez-Serrano, A. Marcilla, Tg-ftir study of evolved gas in the decomposition of different types of tobacco. effect of the addition of sba-15, Thermochimica Acta 671 (2019) 209–219.

[13] F. Barontini, A. Tugnoli, V. Cozzani, J. Tetteh, M. Jarriault, I. Zinovik, Volatile products formed in the thermal decomposition of a tobacco substrate, Industrial & Engineering Chemistry Research 52 (42) (2013) 14984–14997.

[14] G. Guo, C. Liu, Y. Wang, S. Xie, K. Zhang, L. Chen, W. Zhu, M. Ding, Comparative investigation on thermal degradation of flue-cured tobacco

with different particle sizes by a macro-thermogravimetric analyzer and their apparent kinetics based on distributed activation energy model, Journal of Thermal Analysis and Calorimetry 138 (5) (2019) 3375–3388.

[15] J. M. Encinar, F. J. Beltrán, J. F. González, M. J. Moreno, Pyrolysis of maize, sunflower, grape and tobacco residues, Journal of Chemical Technology & Biotechnology: International Research in Process, Environmental AND Clean Technology 70 (4) (1997) 400–410.

[16] W. Gao, K. Chen, Z. Xiang, F. Yang, J. Zeng, J. Li, R. Yang, G. Rao, H. Tao, Kinetic study on pyrolysis of tobacco residues from the cigarette industry, Industrial Crops and Products 44 (2013) 152–157.

[17] T. Abbas, M. Awais, F. Lockwood, An artificial intelligence treatment of devolatilization for pulverized coal and biomass in co-fired flames, Combustion and flame 132 (3) (2003) 305–318.

[18] Z. Yıldız, H. Uzun, S. Ceylan, Y. Topcu, Application of artificial neural networks to co-combustion of hazelnut husk–lignite coal blends, Bioresource technology 200 (2016) 42–47.

[19] S. Sunphorka, B. Chalermsinsuwan, P. Piumsomboon, Artificial neural network model for the prediction of kinetic parameters of biomass pyrolysis from its constituents, Fuel 193 (2017) 142–158.

[20] Q. Tang, Y. Chen, H. Yang, M. Liu, H. Xiao, S. Wang, H. Chen, S. R. Naqvi, Machine learning prediction of pyrolytic gas yield and compositions with feature reduction methods: Effects of pyrolysis conditions and biomass characteristics, Bioresource Technology 339 (2021) 125581.

[21] J. Xing, K. Luo, H. Wang, T. Jin, J. Fan, Novel sensitivity study for biomass directional devolatilization by random forest models, Energy & Fuels 34 (7) (2020) 8414–8423.

[22] Z. Ullah, S. R. Naqvi, W. Farooq, H. Yang, S. Wang, D.-V. N. Vo, et al., A comparative study of machine learning methods for bio-oil yield prediction–a genetic algorithm-based features selection, Bioresource Technology 335 (2021) 125292.

[23] J. Xing, H. Wang, K. Luo, S. Wang, Y. Bai, J. Fan, Predictive single-step kinetic model of biomass devolatilization for cfd applications: A comparison study of empirical correlations (ec), artificial neural networks (ann) and random forest (rf), Renewable Energy 136 (2019) 104–114.

[24] H. Wei, K. Luo, J. Xing, J. Fan, Predicting co-pyrolysis of coal and biomass using machine learning approaches, Fuel 310 (2022) 122248.

[25] H. Wei, Y. Peng, H. Huang, J. Fan, J. Xing, K. Luo, J. Fan, L. Dai, Toba-cpd: An extended chemical percolation devolatilization model for tobacco pyrolysis, ACS Omega 7 (41) (2022) 36776–36785. arXiv:https://doi.org/10.1021/acsomega.2c05098, doi:10.1021/acsomega.2c05098.
URL https://doi.org/10.1021/acsomega.2c05098

[26] J. Xing, K. Luo, H. Wang, Z. Gao, T. Jin, J. Fan, Comparative study on different treatments of coal devolatilization for pulverized coal combustion simulation, Energy & Fuels 34 (3) (2020) 3816–3827.

[27] C. Zhao, K. Luo, R. Cai, J. Xing, Z. Gao, J. Fan, Large eddy simulations and analysis of no emission characteristics in a laboratory pulverized coal flame, Fuel 279 (2020) 118316.

[28] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Machine learning 63 (1) (2006) 3–42.

[29] A. Géron, Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems, O'Reilly Media, 2019.

[30] Pearson correlation coefficient (2021).

[31] K. Pearson, Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs, Proceedings of the royal society of london 60 (359-367) (1897) 489–498.

[32] sklearn.ensemble.ExtraTreesRegressor.
URL `https://scikit-learn/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html`