Journal Pre-proof

Fast characterization of biomass and waste by infrared spectra and machine learning models

Junyu Tao, Rui Liang, Jian Li, Beibei Yan, Guanyi Chen, Zhanjun Cheng, Wanqing Li, Fawei Lin, Lian Hou

PII: S0304-3894(19)31677-2

DOI: https://doi.org/10.1016/j.jhazmat.2019.121723

Reference: HAZMAT 121723

To appear in: *Journal of Hazardous Materials*

Received Date: 3 October 2019

Revised Date: 7 November 2019

Accepted Date: 19 November 2019

Please cite this article as: Tao J, Liang R, Li J, Yan B, Chen G, Cheng Z, Li W, Lin F, Hou L, Fast characterization of biomass and waste by infrared spectra and machine learning models, *Journal of Hazardous Materials* (2019), doi: https://doi.org/10.1016/j.jhazmat.2019.121723

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2019 Published by Elsevier.

Journal Pre-proof

Fast characterization of biomass and waste by infrared spectra and machine learning models

Junyu Tao[a], Rui Liang[a], Jian Li[a], Beibei Yan[a]* yanbeibei@tju.edu.cn, Guanyi Chen[b,a,c]* chen@tju.edu.cn

, Zhanjun Cheng[a,c], Wanqing Li[a], Fawei Lin[a], Lian Hou[a]

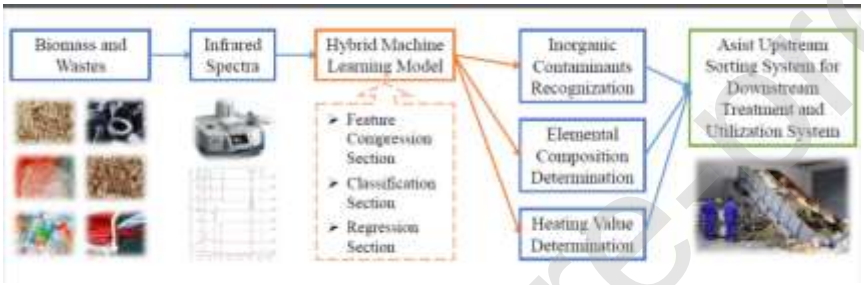
[a] *School of Environmental Science and Engineering, Tianjin University, Tianjin 300350, China*

[b] *School of Science, Tibet University, Lhasa 850012, China*

[c] *Tianjin Key Lab of Biomass Wastes Utilization/Tianjin Engineering Research Center of Bio Gas/Oil Technology, Tianjin 300072, China*

* Corresponding author: Beibei Yan, Guanyi Chen; Tel./fax: +86 22 87402075; Email address:;

Graphical abstract



Highlights

- A new method is proposed to characterize chemical features of biomass and waste.
- The method is based on infrared spectroscopy and machine learning models.
- The method is fast and the optimal accuracy could reach as high as 95.54%.
- The robustness of this method is validated, and its properties are discussed.
- The method can enhance sorting of biomass and waste for downstream utilization.

Abstract

Heterogeneity is a most serious obstacle for treatment and utilization of biomass and waste (BW). This paper proposed a fast characterization method based on infrared spectroscopy and machine learning models, thus to roughly predict the elemental composition and heating value of BW. The fast characterization results could be used to sort different BW components by their suitable downstream utilization techniques. The infrared spectra based hybrid model contained a feature compression section to extract core information from raw infrared spectra, a classification section to distinguish inorganic dilution, and a regression section to generate the elemental composition and heating value results. By parameters optimization, the accuracy of this

Journal Pre-proof

hybrid model reached 95.54%, 85.53%, 92.40%, and 92.49% for C content, H content, O content, and low heating value prediction, respectively. The robustness analysis was conducted by completely rearranging the training and test sets, and it further validated the hypothesis that the infrared spectra contains enough qualifying and quantifying information to characterize these properties of BW. Compared with previous literature, the C-H, C-O, and O-H correlations in BW were also well kept in the predicted results. This work is hoped to enhance upstream sorting system design for treatment and utilization of BW.

Keywords: biomass and waste; elemental composition; heating value; infrared spectra; machine learning.

1. Introduction

Biomass and waste (BW) are renewable and carbon neutral alternative energy resources[1]. Direct[2, 3] and indirect[4-7] combustion of BW is promising to recover energy and solve its environmental problems. While heterogeneity is a serious obstacle for utilization of BW. Inorganic contaminants such as glass and metals can cause damages in the conveying systems or result in contaminations of the fuel products[8, 9]. In respect of the organic fraction, BW components with different elemental compositions and heating values are diversified in fuel quality, and thus preferred to be treated separately[10-12]. To solve the heterogeneity problem of BW, a number of sorting methods have been developed based on computer vision[13-19], spectroscopy[20-26], and other techniques (e.g. X-ray[27] and sonar[28]).

Generally speaking, computer vision, X-ray, and sonar techniques sort different BW components in a similar way as human does. They use cameras in different forms to obtain dimensional and graphic features (people see objects and get pictures in their minds), and then match these features with the database (people use memories), thus to generate the final sorting result. There are two drawbacks for these vision based techniques. Firstly, highly twisted or fragmented samples are hard to be recognized by the shape based classification algorithm. Secondly, samples which in the similar shape may be made from completely different materials (e.g. plastic bottles and glass bottles), and they are likely to be misclassified to the same group. Therefore, these vision based techniques are usually applied in rather simple tasks, such as searching for plastic bottles in trash (without glass bottles)[13, 14, 16, 17], classifying well-shaped samples[15, 19], and sorting relatively monotonous types of samples[18, 29, 30].

Infrared (IR) spectroscopy is an optical technique that detects molecular bond vibrations and rotations upon absorption of infrared light. Compared with vision based techniques, spectroscopy method is capable to distinguish different samples by their materials regardless of their shapes. Currently, spectroscopy method has been used to distinguish different kinds of plastics (namely polyethylene terephthalate, polyvinyl chloride, polypropylene, etc.)[20, 22-25] and recognize glass[26] in trash. The strong classification capability towards different chemicals can help spectroscopy method

Journal Pre-proof

finish more exquisite work than vision based techniques. While in actual BW components, there could be numerous chemicals, and sorting these BW components by their specific chemical composition is an extremely complicated work. For this reason, there are few publications shedding light on general sorting work of BW components.

As mentioned above, current BW sorting techniques have different problems, a fast and effective sorting method for downstream treatment and utilization is thus needed. While in this field, the potential of quantitative capability of IR spectroscopy seems to have been underestimated. As we know, IR spectra contain qualitative information (such as peak position) corresponding to functional groups identification, as well as quantitative information (such as peak area and peak height) corresponding to relative amount of function groups. Types and contents of function groups in some degree determine the elemental composition of organic matters, and elemental composition is highly corresponding to heating value. Therefore, the hypothesis is, IR spectra may contain sufficient information to predict elemental composition and heating value of BW. In fact, it has been reported that IR spectra is capable to predict octane number, water content, total acid number and phenolic content of liquid fuels[31-34]. However, relevant research for elemental composition and heating value prediction of BW has rarely been reported. This is very valuable because, for utilization of BW, the heating value and elemental composition strongly impact its energy potential and suitable utilization methods[35-37]. For example, high heating value feedstocks could be directly combusted to generate heat, while low heating value feedstocks may be better converted to other fuel products first before being combusted. Meanwhile, different thermochemical processes may favor different carbon/hydrogen or carbon/oxygen ratios[38]. As a result, sorting different components in BW by their elemental compositions and heating values could lead to significant economic and environmental benefits.

Accordingly, a sorting system for BW treatment and utilization is proposed and shown in **Figure 1**. In this scheme, inorganic contaminants are first eliminated from rough BW. After that, the combustible organic components are sorted by their elemental compositions and heating values, thus to obtain refined BW for different uses.

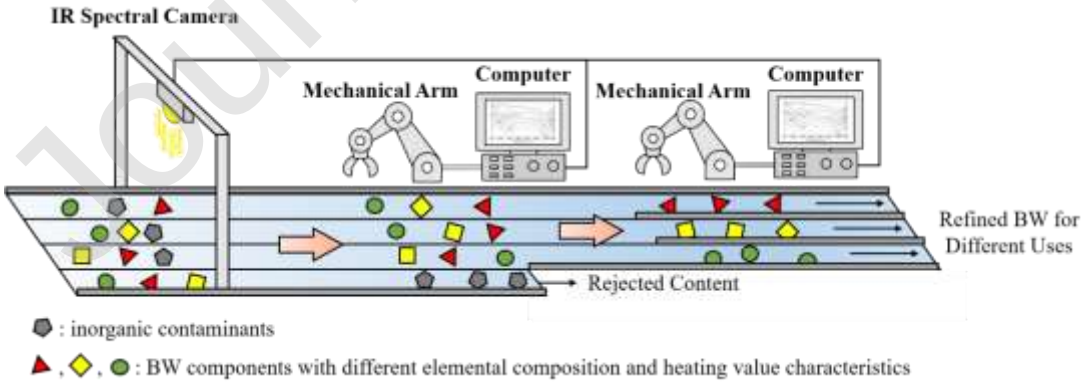


Figure 1 The scheme of an upstream sorting system for RDF utilization.

Journal Pre-proof

This research aims to propose an IR spectra based model for inorganic contaminant recognization and rough ultimate analysis of BW, thus to validate the hypothesis mentioned above. Concerning the requirement of sorting BW components, a hybrid machine learning model which contains (a) a feature compression section to extract core information from raw IR spectra, (b) a classification section to recognize inorganic contaminant, and (c) a regression section to generate the rough ultimate analysis results, was proposed and evaluated in this work. IR spectra of various BW samples were collected, the parameters of each section in the hybrid model were optimized, and the some properties of the hybrid model were discussed. This work is hoped to significantly enhance treatment and energy recovery of different kinds of BW.

2. Methods and materials

2.1 Materials

18 kinds of biomass, 8 kinds of organic waste and 4 kinds of inorganic material were used as spectroscopy materials. For each kind of spectroscopy material, 4 samples were collected for spectroscopy test. All samples were collected from northeastern China. Among these 4 samples, one was included in testing set, and the other three were included in training set. It made a training set of 90 samples and a testing set of 30 samples. The training set and test set were not arranged randomly because, with a limitation of sample size, manually arranging the training set and test set was supposed to cover more kinds of BW during establishment of the predicting model, and thus better enhance the model's reproducibility.

The C, H, and O composition (wt.%) and low heating value (LHV, MJ/kg) results came from *Phyllis2*, a database for BW (https://phyllis.nl/). A list of all kinds of samples and their characterization results is shown in **Table 1**.

Table 1 List of all kinds of samples and their characterization results.

No.	Sample	C content	H content	O content	LHV
		wt.%	wt.%	wt.%	MJ/kg
1	lilac flower	42.79	5.44	34.64	15.52
2	lilac leave	42.79	5.44	34.64	15.52
3	maple leave	49.89	6.09	43.27	17.53
4	malus spectabilis flower	42.79	5.44	34.64	15.52
5	malus spectabilis leave	49.89	6.09	43.27	17.53
6	willow leave	45.86	5.47	38.28	17.00
7	holly leave	43.04	5.37	38.12	15.74
8	poplar sawdust	41.19	5.03	37.17	14.94
9	dry pine needle	48.21	6.57	43.72	18.69
10	pine bark	51.21	5.51	36.35	18.97
11	pine branch	45.08	5.62	39.64	16.99
12	fresh pine needle	48.21	6.57	43.72	18.69
13	chlorella powder	49.98	6.84	27.34	20.63

Journal Pre-proof

14	sakura flower	42.79	5.44	34.64	15.52
15	sakura leave	36.64	4.30	31.80	12.91
16	corn straw	41.18	4.96	38.41	14.30
17	corn leave	43.98	5.39	38.85	15.68
18	bamboo leave	42.68	5.32	42.79	16.41
19	nylon cloth	39.00	5.00	26.00	17.50
20	kitchen waste	45.80	5.05	41.05	17.04
21	polyvinyl chloride	40.03	5.09	0.65	19.95
22	latex	31.55	6.99	3.27	31.55
23	polyethylene	86.00	14.00	0.00	37.44
24	rubber	64.40	10.62	5.64	30.15
25	sewage sludge	30.63	4.41	18.03	12.40
26	paper	49.14	0.61	43.03	18.39
27	glass*	0.00	0.00	0.00	0.00
28	aluminum foil*	0.00	0.00	0.00	0.00
29	stainless steel*	0.00	0.00	0.00	0.00
30	iron*	0.00	0.00	0.00	0.00

*：Samples with asterisk mark refer to inorganic samples.

2.2 Apparatus and ATR-IR measurement

IR spectra were collected in a ThermoFisher, IS 50 spectrometer equipped with an attenuated total reflectance (ATR) unit and a DTGS detector. ATR is an enhanced technique to obtain IR spectra without complex sample preparation procedures. Its convenient feature has made it one of the most commonly used IR spectroscopy techniques in recent years[39, 40]. The scanned wavenumber range was 400 - 4000 cm^{-1} and the resolution was set to 0.482 cm^{-1}, thus to obtain absorbance data on 7469 wavenumbers. Air background spectra were collected before sample spectra. For each test the sample was scanned 32 times and the averaged result was used for following analysis. No special pretreatment was conducted to the samples. The bulk or powder samples were put directly on the sample holder of the spectrometer and then measured.

2.3 Model establishment

The scheme of the hybrid predicting model used in this study is shown in **Figure 2**. Feature compression sections and a classification section were included in this scheme because, in our previous attempts, lack of feature compression section could cause serious underfitting problem[41], and lack of classification section could cause serious misjudgment problem for inorganic dilution samples. Relevant results are included in Figure S1 and S2.

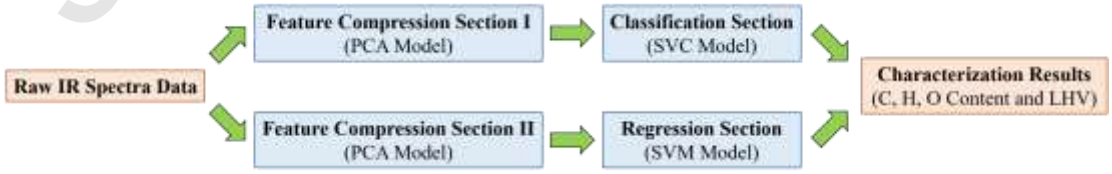


Figure 2 Scheme of the hybrid predicting model.

Journal Pre-proof

Feature compression of raw IR spectra was conducted by principal component analysis (PCA) method[42]. PCA gives a series of principal components (PCs) to represent the original data, so that the original data can be dimensionally reduced with most information reserved in these PCs. In this work, each PC is a linear combination of absorbance values under different wavenumbers. The formation of each PC is expressed in Equation (1), where PCm is the m th PC number; n is the dimension of the original data (which was 7469 in this work, because the IR data were absorbance values under 7469 different wavenumbers); $l_{m,n}$ is the loading coefficient of the n th dimension in the m th PC; a_n is the value (absorbance in this work) of n th dimension from the original data. The importance of each PC is determined by explained variance ratio (EVR). Higher EVR implies that more information from the original data is kept in this PC, and all PCs are numbered by descending order of EVR. The sum of EVR of all PCs is equal to or less than 1.

$$\mathrm{PCm}=\sum_1^n l_{m,n} \times a_n \quad (1)$$

Support vector machine (SVM) algorithms[43] were adopted to get the final characterization results. Support vector classification (SVC) algorithm was used to establish the classification model which could distinct combustible organic samples and inorganic samples. Support vector regression (SVR) algorithm was used to establish the regression model which could give the final heating value and elemental composition results of combustible organic samples. SVC and SVR both belong to SVM, and SVM is recognized as one of the most efficient algorithm for processing data with relatively small sample size and high dimension[44-47]. In general, the principle of SVM is to find a hyperplane which has the longest "distance" to the "nearest" data point (SVC), or a hyperplane which has the shortest "distance" to the "farthest" data point (SVR). The "distance" in SVM is represented by kernels. Commonly used kernels include linear, poly and rbf (referring to linear kernel, polynomial kernel and radial basis kernel, respectively). The mathematical definitions of these kernels are shown in Equation (2)-(4), where x and y are data vectors; $\mathrm{k}(\mathrm{x},\mathrm{y})$ is the kernel of data x and data y; a, c, d, and γ are constant parameters in these kernels.

linear: $$\mathrm{k}(\mathrm{x},\mathrm{y})=x^T y + c \quad (2)$$

poly: $$\mathrm{k}(\mathrm{x},\mathrm{y})=(ax^T y + c)^d \quad (3)$$

rbf: $$\mathrm{k}(\mathrm{x},\mathrm{y})=\exp(-\gamma\|x - y\|^2) \quad (4)$$

PCA, SVC and SVR were conducted by Scikit-learn v0.21.2 package in Python 3.7.3 programming environment. PC numbers within 20 were researched for PCA. Kernels including linear, poly and rbf in SVM were investigated. The detailed Python codes were uploaded to GitHub (https://github.com/tjutjy/atr-ml).

2.4 Data analysis

The performance of classification model was evaluated by accuracy, precision, recall and F1 score:

$$\mathrm{Accuracy}=\frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}\times 100\% \quad (5)$$

Journal Pre-proof

where TP is the count of true positive samples (the value of actual class is 1 and the value of predicted class is also 1), TN is the count of true negative samples (the value of actual class is 0 and the value of predicted class is also 0), FP is the count of false positive samples (the value of actual class is 0 but the value of predicted class is 1), FN is the count of false negative samples (the value of actual class is 1 but the value of predicted class is 0).

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}}\times 100\% \tag{6}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FP}}\times 100\% \tag{7}$$

$$\text{F1}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}}\times 100\% \tag{8}$$

The performance of regression model was evaluated by mean relative error (MRE):

$$\text{MRE}=\frac{\frac{1}{n}\times\sum_{i=1}^{n}|x_i-y_i|}{|\bar{y}|}\times 100\% \tag{9}$$

where n is the number of evaluated samples, x_i is the predicted value of sample i, y_i is the actual value of sample i, $\bar{y}$ is the mean of the actual values of all evaluated samples.

The correlation of investigated predicting subjects was also evaluated by Pearson correlation coefficient (PCC):

$$\text{PCC}_{x,y}=\frac{\text{cov}(x,y)}{\sigma_x\sigma_y} \tag{10}$$

where cov(x,y) is the covariance of data set x and y, σ_x and σ_y are standard deviations of data set x and y.

3. Results and discussion

3.1 Feature compression section

Journal Pre-proof

Journal Pre-proof

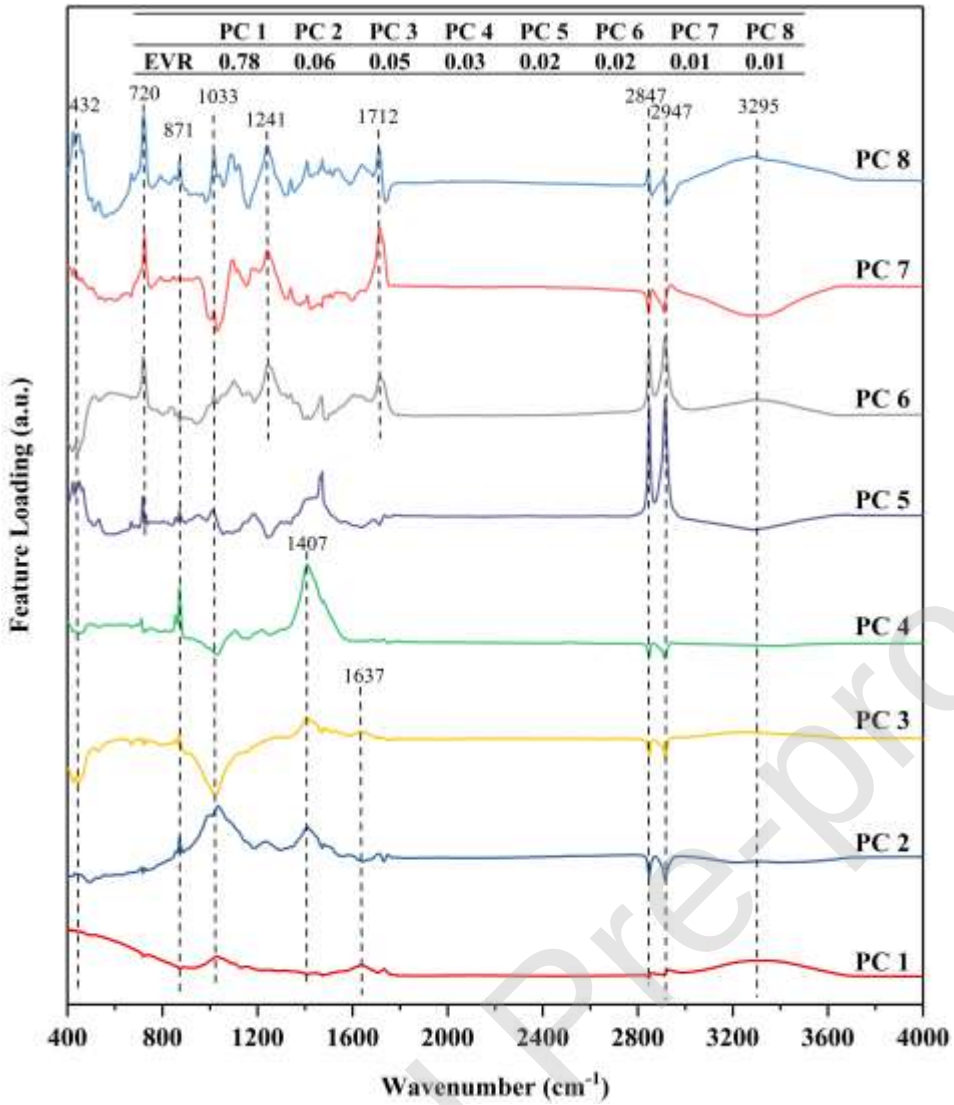


Figure 3. Feature loadings of top 8 PCs in different wavenumbers.

The IR spectra samples obtained in this research had a relatively high feature dimension and relatively small sample size, which is likely to lead to serious underfitting problem. Therefore, PCA algorithm was used as feature compression section to reduce the feature dimension of the samples. In this process, comprehension about the PCs is important to discuss the working mechanism of the feature compression section. As shown in **Figure 3**, the top 8 PCs had a total EVR of 0.98, which implies that most information of the IR spectra could be reserved in the 8 PCs. Therefore, the loadings of the top 8 PCs were further discussed.

According to **Figure 3**, EVR of 0.78 was obtained by PC1, implying that most distinctive information in IR spectra was extracted from the high loading regions of PC1. In feature loading pattern of PC 1, there were four high loading districts. Two broad districts were found in wavenumber range of 400-820 cm^{-1} and 3000-3700 cm^{-1}. And two sharp districts were found around wavenumber of 1033 cm^{-1} and 1637 cm^{-1}. With respect to PC 2-4, the high loading districts were found around wavenumber of

Journal Pre-proof

432 cm^{-1}, 1033 cm^{-1}, 1407 cm^{-1}, 2847 cm^{-1}, and 2947 cm^{-1}. PC2, PC3, and PC4 were also with relatively high EVRs, and PC 1-4 shared a similar loading peak in wavenumber of 1407 cm^{-1}, which was not observed in PC 5-8. In comparison with PC 1-4, a loading peak in wavenumber of 720 cm^{-1} was observed in PC 5-8, and loading peaks of 1241 cm^{-1} and 1712cm^{-1} were observed in PC 6-8. These unique loading peaks in PCs with lower EVR are likely to contain more detailed qualifying or quantifying information extracted from IR spectra. The function groups that these wavenumber ranges might belong to has been summarized in **Table 2**. While it's important to note that compared with regular IR analysis, this fast characterization method doesn't rely on identification of IR peaks. The roles that different PCs played in classification section and regression section were further discussed in following sections.

Table 2 A list of function groups that IR wavenumber ranges might belong to[34, 48].

Wavenumber range (cm^{-1})	Function group
400-1000	fingerprint region of various function groups
around 1033	C-O stretch, C-C stretch
around 1241	C-O stretch
around 1407	C-H bend, O-H bend
around 1637	C=O stretch, C=C stretch
around 2847	C-H stretch of CH_3
around 2947	C-H stretch of CH_2
3000-3700	O-H stretch

3.2 Classification section

Journal Pre-proof

Journal Pre-proof

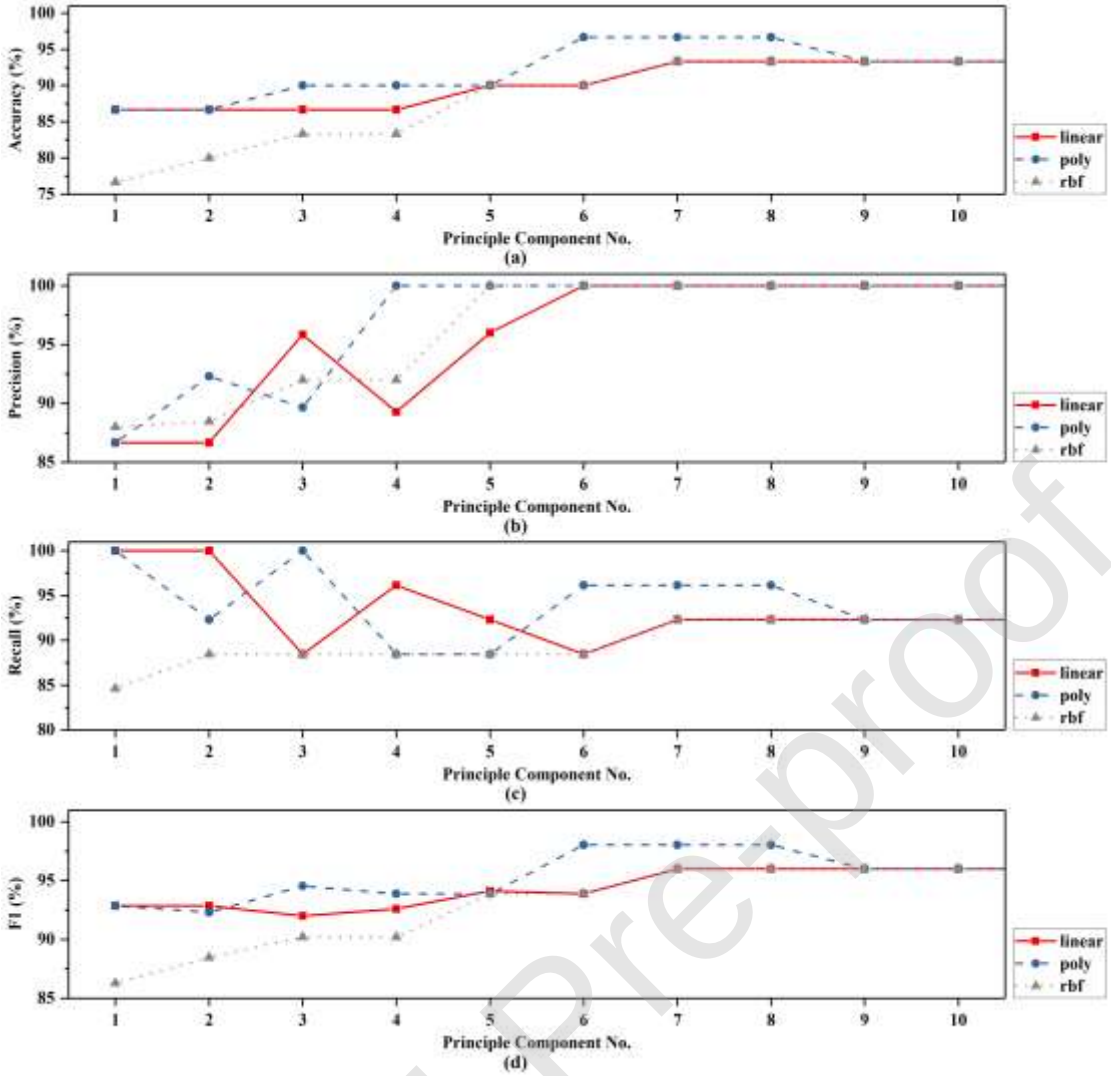


Figure 4. Effect of principal component number on performance of the classification section. The performance of the classification section is evaluated by (a) accuracy, (b) evaluated by precision, (c) evaluated by recall and (d) evaluated by F1 score.

In actual treatment and utilization of BW, there is likely to be inorganic dilution such as metal and glass. Therefore, it's important to recognize these inorganic contaminants so that they would not be sent for further characterizations. A classification section using SVC model was thus applied and evaluated with different PC number in feature compression process of IR spectra. Relevant results are shown in **Figure 4**. According to precision results, when PC number reached 4, 5, 6 for poly, rbf, linear kernels, the precision reached 100%. It means all samples classified as BW were exactly BW. According to recall results, when PC number was relatively low, the SVC models with linear and poly kernels were more likely to find all BW samples. While referring to the precision results, many inorganic samples were also wrongly classified to combustible BW under this condition. According to the accuracy and F1 score results, the optimal kernel was poly and the optimal PC number range was 6 - 8. Under this condition, the feature reservation and data structure simplification functions of PCA

Journal Pre-proof

model were best balanced. In this optimized classification model where the PC number was set at 7 and the kernel was set as poly, only sample 24 was misclassified.

Meanwhile, it was observed that when the PC number was as low as 1, the accuracy and F1 score of the SVC model could still reach 92% and 86%, respectively. It suggests that PC 1 was likely to have kept most classification information from the IR spectra. That is to say that, IR pattern in range of 400-820 cm^{-1}, 3000-3700 cm^{-1}, around 1033 cm^{-1}, and around 1637 cm^{-1} were essential for classification between BW and inorganic dilution. The accuracy of the SVC model increased rapidly in PC number range of 2-5. Considering the wavenumber loadings shown in **Figure 3**, IR pectra data in wavenumber around 1407 cm^{-1} were likely to be auxiliary for the classification work.

3.3 Regression section

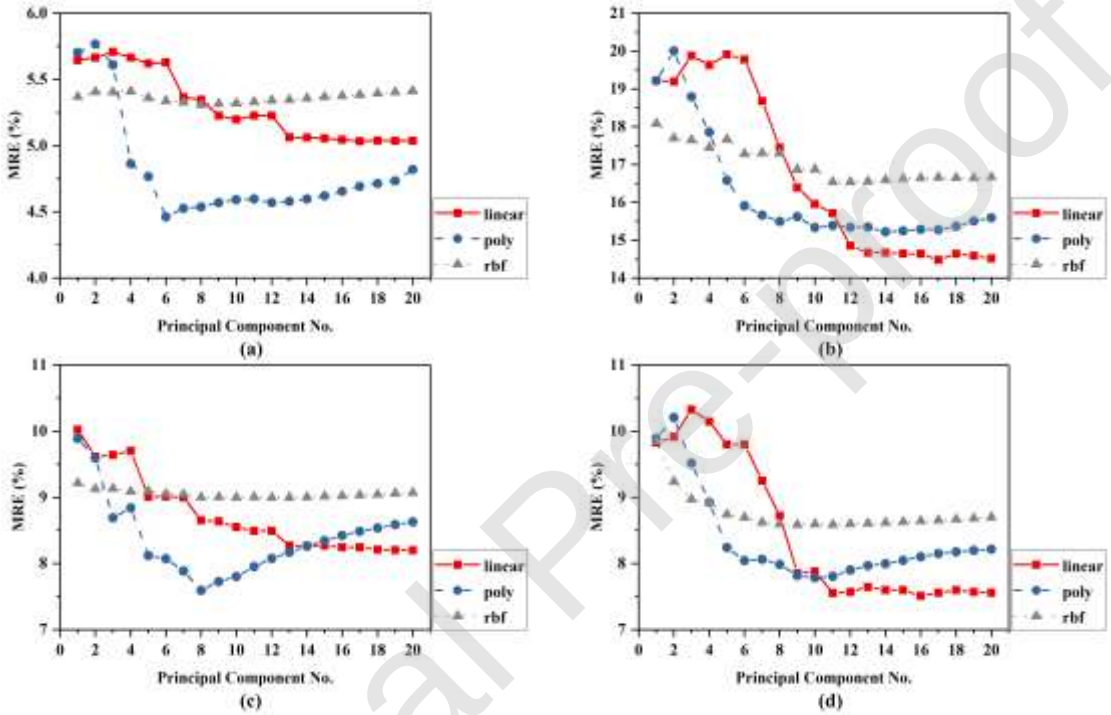


Figure 5. Effect of principal component number on performance of the regression section for predicting a) C content, b) H content, c) O content, and d) LHV.

SVR models were used in regression section to generate the final C content, H content, O content, and LHV results. The models were trained by combustible BW samples in the training set, and the performance of the regression section under different PC numbers is shown in **Figure 5**. When PC numbers were below 8, almost all curves showed a significant decreasing trend. This implies that the top 8 PCs contained most quantifying data for these characterizations, and the PCA models were gaining more information from raw IR spectra as PC numbers were increasing in this range. In respect of O content characterization, the MRE result gradually increased after the PC number was higher than 8, which indicates that the growing PC numbers led to more complex SVR model structure, and thus intensified the underfitting problem during model training process.

Journal Pre-proof

In the case of C content and O content characterization, the optimal kernel was poly. The optimal MRE results for C content and O content were 4.46% and 7.80%, which were obtained by PC number of 6 and 8, respectively. For H content and LHV characterizations, the optimal kernel was still linear. Their optimal MRE results were 14.48% and 7.51%, which were obtained when PC numbers are 17 and 16, respectively. While it was observed that for H content and LHV characterizations, the optimized MRE results obtained with linear kernel were just slightly lower than that with poly kernel. Therefore, it could be concluded that poly kernel showed rather stable and effective performance towards all the four characterization subjects. In contrary, although rbf kernel showed the best performance in low PC number range, its improvement with increasing PC number was negligible.

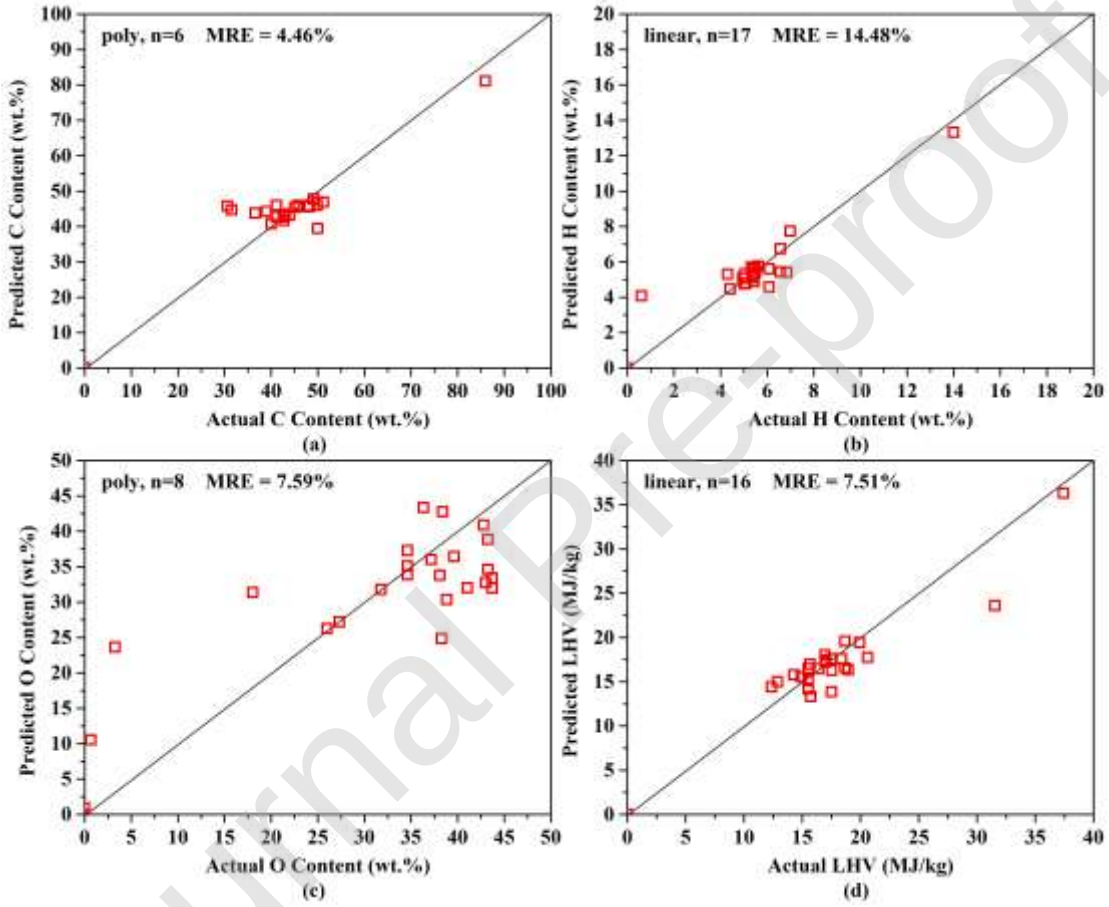


Figure 6. Parity plots of the combined classification section and regression section for predicting a) C content, b) H content, c) O content, and d) LHV.

Figure 6 shows Parity plots of the combined classification section and regression section for predicting a) C content, b) H content, c) O content, and d) LHV. The combination was achieved by multiplying the classification results (0 if classified as inorganic dilution, and 1 if classified as combustible BW) and regression results together. As shown in the parity plots, the classification section ensured that the predicting results for inorganic dilution would be exactly zero. According to **Figure 6** a), b), and c), when there were unique test samples with extremely high actual values,

Journal Pre-proof

the regression section could still precisely predict the characterization results. The MRE ranged from 4.46% to 14.48% for the four characterization subjects. This result may not be as accurate as traditional characterization methods (e.g. by using an elemental analyzer and a calorimeter), but the measuring process is significantly faster, and the accuracy is supposed to be acceptable for downstream sorting of BW. The error of the prediction results might be attributed to that the IR spectra contains much but not all information required for elemental composition and heating value prediction. Another reason could be that the models were not good enough. In this case, further model enhancement methods would include sample size expansion, model type selection, model parameters optimization, etc.[49, 50].

3.4 Robustness validation of the characterization method

Table 3 Performance of predicting models trained by original sample sets and modified sample sets.

Training and Test Sets	Subject	C Content	H Content	O Content	LHV
Original	**Min MRE (%)**	**4.46**	**14.47**	**7.6**	**7.51**
	Optimized Principal Component No. for SVR	6	17	8	16
	Optimized Kernel for SVR	poly	linear	poly	linear
Modified	**Min MRE (%)**	**8.09**	**14.11**	**11.47**	**4.77**
	Optimized Principal Component No. for SVR	7	9	16	8
	Optimized Kernel for SVR	poly	poly	poly	poly

As mentioned above, to guarantee the reproducibility of the predicting model, all types of samples were used in both training set and test set. While in this case, it's possible that the predicting models just overfitted the original training set thus to obtain positive predicting results. While in contrary, it's preferred that the mechanism of this characterization method could work universally. Therefore, a modified training set and test set were used to train the same hybrid predicting model. The predicting performance was evaluated thus to validate the robustness of this characterization method. The modified test set contained samples with the highest C content, H content, O content, or LHV. All types of samples which were used in the modified test set were eliminated from the original training set, thus to obtain a modified training set. It could be seen that the modified test set contained all of the 'strangest' samples, and there was no chance for the predicting model to get trained by the similar samples in the modified test set. The details about parameter optimization process with the modified sample sets are included in **Figure S3**, and the robustness validation results are shown in **Table 3**.

According to **Table 3**, the optimal SVR kernels for C content, H content, O content, and LHV were all poly. It's in accordance with former results with original sample sets

Journal Pre-proof

that poly kernel showed stable and effective performance towards all the four characterization subjects. The optimal PC numbers for C content, H content, and LHV were in the range of 7-9. The optimal PC number for O content was 16, while according to **Figure S3**, its MREs in PC number of 7 and above were very close with that in PC number of 16. As mentioned above, a PC number up to 8 could reserve almost all information from the IR spectra in original sample sets. Therefore, it could be concluded that the characteristic of the feature compression section with modified sample sets was in accordance with that with original sample sets.

In respect of accuracy of the regression section, the predicting results on modified samples sets were in the same MRE level with that on original sample sets. The MRE results for H content and LHV prediction were even lower than on original sample sets. As a result, this predicting model did not simply overfit the samples in the training set. It could work effectively with very unique samples that was not greatly different with samples in training set. In other words, the hypothesis that IR spectra contain enough qualifying and quantifying information to generate rough ultimate analysis and heating value results was validated.

3.5 Comparison of elemental correlations in actual and predicted results

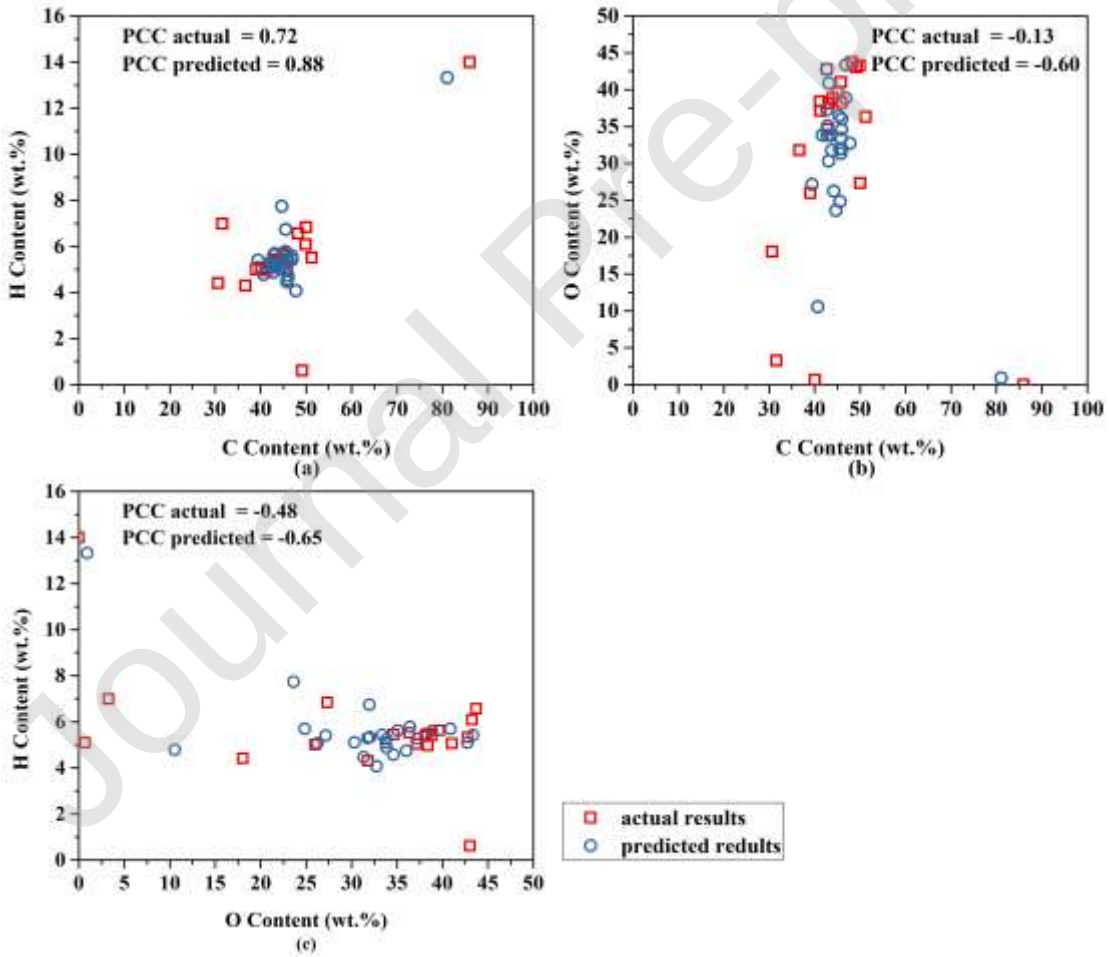


Figure 7. Actual and predicted correlations of (a) C content and H Content, (b) C

Journal Pre-proof

content and O content, (c) O content and H content.

It was reported that there are strong correlations among different elemental contents (e.g. C-H correlation, C-O correlation, O-H correlation) in biomass [37]. In order to see if these correlations could be reserved by this characterization method, actual and predicted correlation of C-H, C-O, and O-H contents were investigated. The predicted results were obtained by the optimized predicting models. The inorganic samples were excluded for no significant meaning. The correlation analysis results are shown in **Figure 7**.

According to **Figure 7**, the PPC for actual C-H, C-O, and O-H correlation were 0.72, -0.13, and -0.48. In comparison, the PPC for these subjects reported by Vassilev et al. [37] were 0.31, -0.88, and -0.49, respectively. The positive or negative relationships were in line with research of Vassilev et al., while except for O-H relation, the absolute PPC values for other pairs were different in varied degrees. This difference was reasonable because solid waste samples were also included in this research, while research of Vassilev et al. mainly concerned biomass samples. As for PPC for predicted correlations, that for C-H, C-O, and O-H showed the same positive or negative relationships with actual PCC. Another founding was that, for C-H, C-O, and O-H correlation, their absolute predicted PPC were all greater than absolute actual PCC. It means these correlations were strengthened during the predicting process, which might be attributed to similar regression processes from the IR spectra.

4. Conclusions

In this work, a fast characterization method for inorganic dilution recognization, heating value determination, and rough ultimate analysis of BW was proposed. The feature compression section extracted the qualifying and quantifying information from the IR spectra by composing a series of PCs with different wavenumber loadings. IR pattern in wavenumber range of 400-820 cm^{-1}, 3000-3700 cm^{-1}, around 1033 cm^{-1}, and around 1637 cm^{-1} were most essential for this characterization method. Poly kernel was proved to be stable and effective as SVM kernel for both classification section and regression section. By parameters optimization, the accuracy of the characterization method reached 95.54%, 85.53%, 92.40%, and 92.49% for C content, H content, O content, and LHV predicting, respectively. The robustness analysis validated that the IR spectra contain enough information to characterize fuel properties of BW samples. The correlation analysis indicated that elemental correlations of C-H, C-O, and O-H were maintained by this characterization method and were in accordance with previous research.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Journal Pre-proof

Acknowledgement

This research was financially supported by the National Natural Science Foundation of China (51676138, 51878557), the National Key R&D Program of China (2016YFE0201800), and the Tianjin Science and Technology Project (18YFJLCG00090). Discussions with Bob Weber from the Institute for Integrated Catalyst at Pacific Northwest National Laboratory, USA are appreciated.

Journal Pre-proof

Journal Pre-proof

References

[1] E. Iakovou, A. Karagiannidis, D. Vlachos, A. Toka, A. Malamakis, Waste biomass-to-energy supply chain management: a critical synthesis, Waste Manage., 30 (2010) 1860-1870.

[2] L. Zhen, W. Han-qing, Z. Yue-yun, L. Jian-wen, Z. Xiao-dong, H. Guang-wei, Effect of microwave chlorine depleted pyrolyzate on the combustion characteristics of refuse derived fuel derived from package waste, Waste Manage., 82 (2018).

[3] W. Teng, H. Haobo, Y. Yang, R. Hao, L. Jinping, X. Yongjie, Combustion behavior of refuse-derived fuel produced from sewage sludge and rice husk/wood sawdust using thermogravimetric and mass spectrometric analyses, J. Cleaner Prod., 222 (2019) 1-11.

[4] N. Catarina, A. Octávio, L. Andrei, V. Cândida, G. Margarida, Torrefaction and carbonization of refuse derived fuel: Char characterization and evaluation of gaseous and liquid emissions, Bioresour. Technol., 285 (2019) 121325-121333.

[5] Š. Patrik, H. Juma, H. Jakub, S. Pavol, Š. Ivan, Catalytic gasification of refuse-derived fuel in a two-stage laboratory scale pyrolysis/gasification unit with catalyst based on clay minerals, Waste Manage., 85 (2019) 1-10.

[6] C.A. Salman, M. Naqvi, E. Thorin, J. Yan, Impact of retrofitting existing combined heat and power plant with polygeneration of biomethane: A comparative techno-economic analysis of integrating different gasifiers, Energy Convers. Manage., 152 (2017) 250-265.

[7] M. Minutillo, A. Perna, D. Di Bona, Modelling and performance analysis of an integrated plasma gasification combined cycle (IPGCC) power plant, Energy Convers. Manage., 50 (2009) 2837-2842.

[8] B. Krüger, A. Mrotzek, S. Wirtz, Separation of harmful impurities from refuse derived fuels (RDF) by a fluidized bed, Waste Manage., 34 (2014) 390-401.

[9] R. Luciano, Infiesta, R.N. Cassius, Ferreira, G. Alam, Trovó, L. Valério, Borges, R. Solidônio, Carvalho, Design of an industrial solid waste processing line to produce refuse-derived fuel, J. Environ. Manage., 236 (2019) 715-719.

[10] A. Gallardo, M. Carlos, M.D. Bovea, F.J. Colomer, F. Albarrán, Analysis of refuse-derived fuel from the municipal solid waste reject fraction and its compliance with quality standards, J. Cleaner Prod., 83 (2014) 118-125.

[11] E.S. Myrin, P.-E. Persson, S. Jansson, The influence of food waste on dioxin formation during incineration of refuse-derived fuels, Fuel, 132 (2014) 165-169.

[12] C. Chiemchaisri, B. Charnnok, C. Visvanathan, Recovery of plastic wastes from dumpsite as refuse-derived fuel and its utilization in small gasification system, Bioresour. Technol., 101 (2010) 1522-1527.

[13] M.A. Zulkifley, M.M. Mustafa, A. Hussain, A. Mustapha, S. Ramli, Robust identification of polyethylene terephthalate (PET) plastics through bayesian decision, PLoS One, 9 (2014) e114518.

[14] H. Xiangyu, H. Zaixing, Z. Shuyou, Z. Xinyue, A novel vision-based PET bottle recycling facility, Measurement Science & Technology, 28 (2017) 025601-025609.

[15] Y. Chu, C. Huang, X. Xie, B. Tan, S. Kamal, X. Xiong, Multilayer hybrid deep-learning method for waste classification and recycling, Computational Intelligence and Neuroscience, 2018 (2018)

Journal Pre-proof

5060857.

[16] K. Özkan, S. Ergin, Ş. Işık, İ. Işıklı, A new classification scheme of plastic wastes based upon recycling labels, Waste Manage., 35 (2015) 29-35.

[17] Z. Wang, B. Peng, Y. Huang, G. Sun, Classification for plastic bottles recycling based on image recognition, Waste Manage., 88 (2019) 170-181.

[18] A. Shaukat, Y. Gao, J.A. Kuo, B.A. Bowen, P.E. Mort, Visual classification of waste material for nuclear decommissioning, Rob. Auton. Syst., 75 (2016) 365-378.

[19] C. Vrancken, P. Longhurst, S. Wagland, Deep learning in material recovery: Development of method to create training database, Expert Syst. Appl., 125 (2019) 268-280.

[20] G. Bonifazi, G. Capobianco, S. Serranti, A hierarchical classification approach for recognition of low-density (LDPE) and high-density polyethylene (HDPE) in mixed plastic waste based on short-wave infrared (SWIR) hyperspectral imaging, Spectrochim. Acta, Part A, 198 (2018) 115-122.

[21] S.-B. Roh, S.-B. Park, S.-K. Oh, E.-K. Park, W.Z. Choi, Development of intelligent sorting system realized with the aid of laser-induced breakdown spectroscopy and hybrid preprocessing algorithm-based radial basis function neural networks for recycling black plastic wastes, J. Mater. Cycles Waste Manage., 20 (2018) 1934-1949.

[22] M. Moroni, A. Mei, A. Leonardi, E. Lupo, F. Marca, PET and PVC separation with hyperspectral imagery, Sensors, 15 (2015) 2205-2227.

[23] S. Serranti, A. Gargiulo, G. Bonifazi, Classification of polyolefins from building and construction waste using NIR hyperspectral imaging system, Resour. Conserv. Recycl., 61 (2012) 52-58.

[24] Y. Zheng, J. Bai, J. Xu, X. Li, Y. Zhang, A discrimination model in waste plastics sorting using NIR hyperspectral imaging system, Waste Manage., 72 (2018) 87-98.

[25] W. Van Den Broek, D. Wienke, W. Melssen, L. Buydens, Plastic material identification with spectroscopic near infrared imaging and artificial neural networks, Anal. Chim. Acta, 361 (1998) 161-176.

[26] O. Winn, K.T. Sivaram, I. Aslanidou, J. Skvaril, K. Kyprianidis, Near-infrared spectral measurements and multivariate analysis for predicting glass contamination of refuse-derived fuel, Energy Procedia, 142 (2017) 943-949.

[27] J. Duvillier, M. Dierick, J. Dhaene, D. Van Loo, B. Masschaele, R. Geurts, L. Van Hoorebeke, M.N. Boone, Inline multi-material identification via dual energy radiographic measurements, NDT & E International, 94 (2018) 120-125.

[28] E. Lopez-Caudana, O. Quiroz, A. Rodríguez, L. Yépez, D. Ibarra, Classification of materials by acoustic signal processing in real time for NAO robots, Int. J. Adv. Rob. Syst., 14 (2017) 1729881417714996.

[29] V. Chaloupková, T. Ivanova, O. Ekrt, A. Kabutey, D. Herák, Determination of particle size and distribution through image-based macroscopic analysis of the structure of biomass briquettes, Energies, 11 (2018) 331-323.

[30] X. Wang, W. Yang, Z. Li, A fast image segmentation algorithm for detection of pseudo-foreign fibers in lint cotton, Computers & Electrical Engineering, 46 (2015) 500-510.

[31] S.R. Daly, K.E. Niemeyer, W.J. Cannella, C.L. Hagen, Predicting fuel research octane number using Fourier-transform infrared absorption spectra of neat hydrocarbons, Fuel, 183 (2016) 359-365.

Journal Pre-proof

[32] J.H. Al-Fahemi, N.A. Albis, E.A. Gad, QSPR models for octane number prediction, Journal of Theoretical Chemistry, 2014 (2014).

[33] J.J. Kelly, C.H. Barlow, T.M. Jinguji, J.B. Callis, Prediction of gasoline octane numbers from near-infrared spectral features in the range 660-1215 nm, Analytical Chemistry, 61 (1989) 313-320.

[34] R.B. Madsen, K. Anastasakis, P. Biller, M. Glasius, Rapid determination of water, total acid number, and phenolic content in bio-crude from hydrothermal liquefaction of biomass using FT-IR, Energy & fuels, 32 (2018) 7660-7669.

[35] M. Sharifzadeh, M. Sadeqzadeh, M. Guo, T.N. Borhani, N.M. Konda, M.C. Garcia, L. Wang, J. Hallett, N. Shah, The multi-scale challenges of biomass fast pyrolysis and bio-oil upgrading: Review of the state of art and future research directions, Prog. Energy Combust. Sci., 71 (2019) 1-80.

[36] T.M. Dabros, M.Z. Stummann, M. Høj, P.A. Jensen, J.-D. Grunwaldt, J. Gabrielsen, P.M. Mortensen, A.D. Jensen, Transportation fuels from biomass fast pyrolysis, catalytic hydrodeoxygenation, and catalytic fast hydropyrolysis, Prog. Energy Combust. Sci., 68 (2018) 268-309.

[37] S.V. Vassilev, D. Baxter, L.K. Andersen, C.G. Vassileva, An overview of the chemical composition of biomass, Fuel, 89 (2010) 913-933.

[38] M. Gong, W. Zhu, Y. Fan, H. Zhang, Y. Su, Influence of the reactant carbon–hydrogen–oxygen composition on the key products of the direct gasification of dewatered sewage sludge in supercritical water, Bioresour. Technol., 208 (2016) 81-86.

[39] D.-W. Sun, Infrared spectroscopy for food quality analysis and control, Academic Press, 2009.

[40] V. Singh, F. Bux, Y.C. Sharma, A low cost one pot synthesis of biodiesel from waste frying oil (WFO) using a novel material, β-potassium dizirconate (β-K2Zr2O5), Appl. Energy, 172 (2016) 23-33.

[41] X. Guyon, J.-f. Yao, On the Underfitting and Overfitting Sets of Models Chosen by Order Selection Criteria, Journal of Multivariate Analysis, 70 (1999) 221-249.

[42] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometrics Intellig. Lab. Syst., 2 (1987) 37-52.

[43] M. Zhao, J. Shi, C. Lin, Optimization of integrated energy management for a dual-motor coaxial coupling propulsion electric city bus, Appl. Energy, 243 (2019) 21-34.

[44] E. Parhizkar, H. Saeedzadeh, F. Ahmadi, M. Ghazali, A. Sakhteman, Partial least squares-least squares-support vector machine modeling of ATR-IR as a spectrophotometric method for detection and determination of iron in pharmaceutical formulations, Iranian journal of pharmaceutical research: IJPR, 18 (2019) 72.

[45] S.F. Hussain, A novel robust kernel for classifying high-dimensional data using Support Vector Machines, Expert Syst. Appl., 131 (2019) 116-131.

[46] M.A. Patil, P. Tagade, K.S. Hariharan, S.M. Kolake, T. Song, T. Yeo, S. Doo, A novel multistage Support Vector Machine based approach for Li ion battery remaining useful life estimation, Appl. Energy, 159 (2015) 285-297.

[47] Y.Y. Chia, L.H. Lee, N. Shafiabady, D. Isa, A load predictive energy management system for supercapacitor-battery hybrid energy storage system in solar application using the Support Vector Machine, Appl. Energy, 137 (2015) 588-602.

[48] T.E. Barber, N.L. Ayala, J.M. Storey, G.L. Powell, W.D. Brosey, N.R. Smyrl, Infrared absorption spectroscopy, Environmental Instrumentation and Analysis Handbook; Down, R.; Lehr, JH

Journal Pre-proof

Eds, (2005) 87-117.

[49] B. Fulkerson, Machine learning, neural and statistical classification, Taylor & Francis, 1995.

[50] E. Alpaydin, Introduction to machine learning, MIT press, 2014.

Journal Pre-proof