Supporting Information

High-accuracy quantitative analysis of coal by small sample modelling algorithm based Laser Induced Breakdown Spectroscopy

*An Li, Xinyu Zhang, Xianshuang Wang, Yage He, Yunsong Yin, and Ruibin Liu**

I. The detail of experiment samples.

Table S1. The detailed indicators of the samples used in this work.

No.#	Carbon Content (wt.%)	Ash content (wt.%)	Calorific value(MJ/kg)	No.#	Carbon Content (wt.%)	Ash content (wt.%)	Calorific value(MJ/kg)
1	38.97	33.33	21.99	51	40.8	30.19	22.86
2	42.83	27.80	23.50	52	38.5	34.91	20.94
3	40.86	30.13	21.70	53	43.54	26.95	23.89
4	38.64	33.80	21.42	54	43.64	26.57	24.17
5	40.27	32.05	21.83	55	51.16	24.64	25.21
6	40.54	30.99	22.44	56	42.72	29.56	23.09
7	37.68	35.47	21.13	57	38.27	33.84	21.55
8	36.91	35.79	20.83	58	39.93	32.72	21.71
9	37.50	36.93	20.24	59	43.77	27.01	23.98
10	42.76	27.92	23.63	60	44.12	26.07	24.21
11	48.67	25.79	24.48	61	45.84	27.55	23.83
12	40.00	31.08	22.25	62	37.52	35.32	20.96
13	36.42	36.59	20.27	63	38.46	34.43	21.07
14	41.75	30.23	22.65	64	39.79	31.98	22.05
15	39.49	33.56	21.41	65	48.94	26.3	24.46
16	41.73	29.27	22.95	66	41.75	30.3	22.59
17	42.35	28.61	23.33	67	40.52	31.25	22.27
18	37.74	34.69	21.13	68	38.17	34.14	21.18
19	38.37	35.21	20.79	69	43.12	27.46	23.55
20	42.30	28.63	23.21	70	47.41	25.2	24.77
21	40.78	30.34	22.60	71	36.13	37.26	19.94
22	41.96	29.64	22.87	72	42.92	27.44	23.58
23	40.94	31.08	22.63	73	41.43	29.64	22.71
24	41.66	29.83	22.80	74	42.92	27.27	23.85
25	42.60	29.75	22.55	75	36.29	37.08	20.35
26	40.47	31.72	22.32	76	39.23	32.58	21.89
27	37.81	35.60	20.65	77	49.43	18.99	25.8
28	45.08	26.84	24.36	78	48.21	24.6	24.96

29	43.10	27.00	23.93	79	43.59	27.38	23.69
30	38.78	33.37	21.48	80	36.52	36.65	20.58
31	43.52	26.84	23.87	81	52.02	13.54	27.5
32	39.24	32.74	21.72	82	47.47	25.77	24.62
33	41.07	30.01	22.79	83	46.53	25.12	25
34	37.69	35.28	21.04	84	41.31	29.7	22.85
35	38.14	34.37	21.19	85	42.15	29.78	22.72
36	41.51	29.88	22.85	86	42.13	28.96	23.16
37	41.49	29.43	22.92	87	41.75	29.37	22.99
38	39.82	32.35	22.18	88	42.23	28.54	23.24
39	40.56	32.24	21.87	89	37.33	35.89	20.74
40	42.54	28.76	23.22	90	46.23	28.33	24.31
41	38.21	34.18	21.43	91	42.86	28.47	23.23
42	40.37	31.11	22.49	92	40.99	30.1	22.83
43	42.68	29.04	23.23	93	43.05	27.38	23.77
44	55.72	23.70	26.15	94	41.73	29.92	22.98
45	43.04	27.69	23.75	95	36.35	37.33	20.06
46	39.70	32.39	21.98	96	42.65	27.97	23.34
47	39.52	32.44	21.99	97	44.28	25.75	24.3
48	39.34	32.48	22.00	98	49.5	19	25.67
49	41.66	29.24	23.11	99	40.71	32.99	21.65
50	40.89	31.66	21.67	100	42.81	29.89	22.85

II. The validation process in modeling

As shown in Figure S1, six samples(P1~P6) are used as a training sample set. After data pretreatment for each sample, the model training process is implemented based on extracting 75% spectra randomly from each sample. The two data extraction are to form the training data batch $S_1 \sim S_6$ that marked by "Extraction 1" and training data batch $S_1^{'} \sim S_6^{'}$ that marked by "Extraction 2", respectively. Train the model relies on LOOCV (Leave One Out Cross Validation), i.e., each sample will serve as the validation set successively, and the hyperparameter PCs (Principal Components) are adjusted and selected in this process.

In the left of Figure S1, which is defined as model training cycle 1. The model is first trained by spectral data set marked by $S_1 \sim S_5$, S_6 is validation data, and the SE (Square Error) versus PCs (Principal Components) is calculated. Repeats the training process until each sample is served as validation data once, a total of six values of SE obtained ultimately for each PC (principal component). Then calculate the MSE (Mean Square Error) of each PC. The minimum MSE (Min 1, 0.348) is marked by a red circle corresponding to the PC is 2 in CV cycle 1. Using 2 PCs train the model with all spectral data $S_1 \sim S_6$ again, and this model is preparing for the alternative model.

In the right of Figure S2, which is defined as model training cycle 2. The same training samples P1~P6 are used. The difference is that the spectral data set $S_1^{'} \sim S_6^{'}$ are re-extracting randomly from the spectra data set. After model training progress based on LOOCV, the minimum MSE (Min 2, 0.19) is obtained as shown in CV cycle 2. The optimal model is determined ultimately by comparing the root value of Min

1 in “Training cycle 1” and Min 2 in “Training cycle 2”. Finally, the model trained by $S_1^{'}\sim S_6^{'}$ is selected for the optimal model because the Min 2 is less than the Min 1.

Note that the training data extraction and model training cycle is only conducted twice, mainly to illustrate the critical step in the new algorithm. Actually, the data extraction and model training cycle is repeated 1000 times in this work, and the optimal model is selected following the rules mentioned in section 4 of the paper.

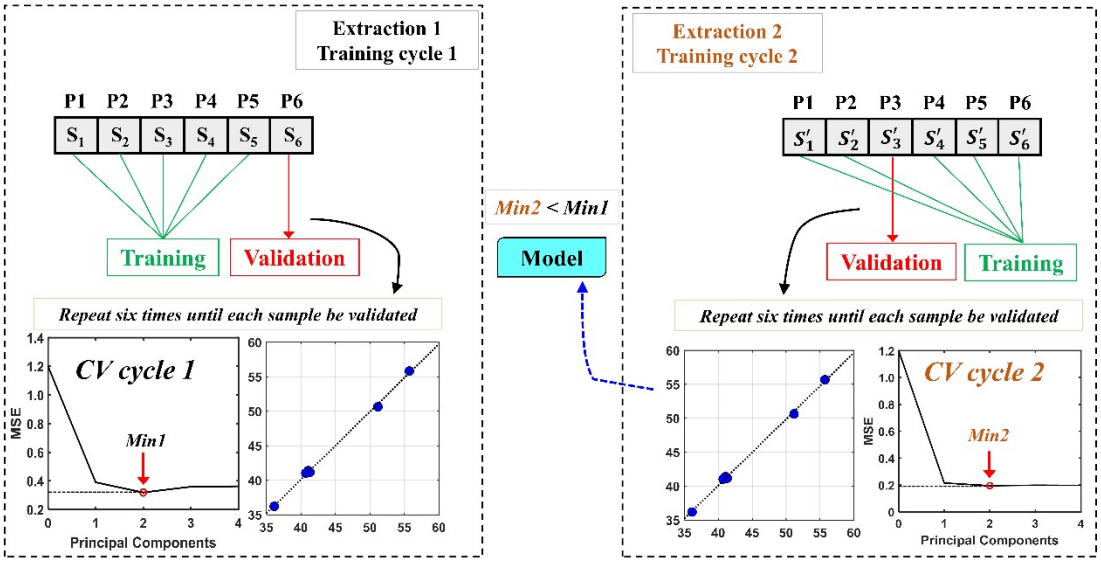
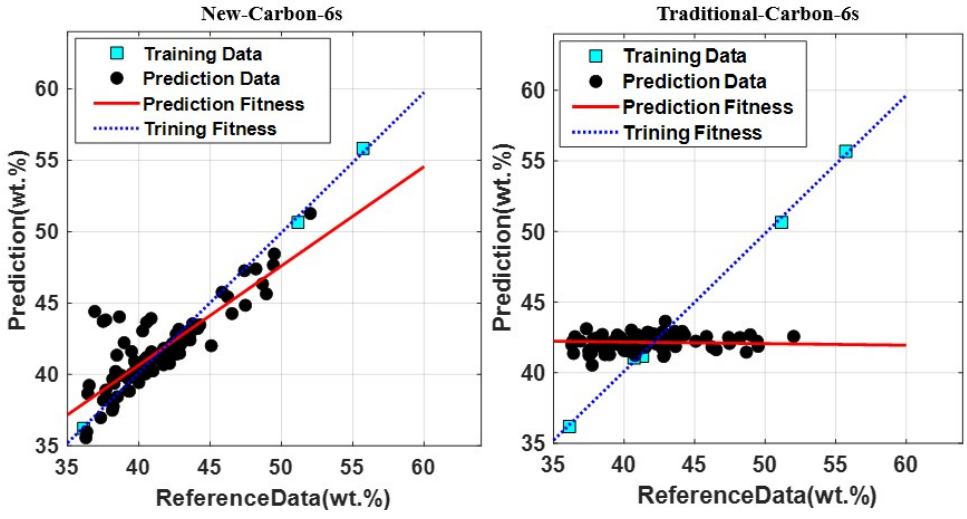
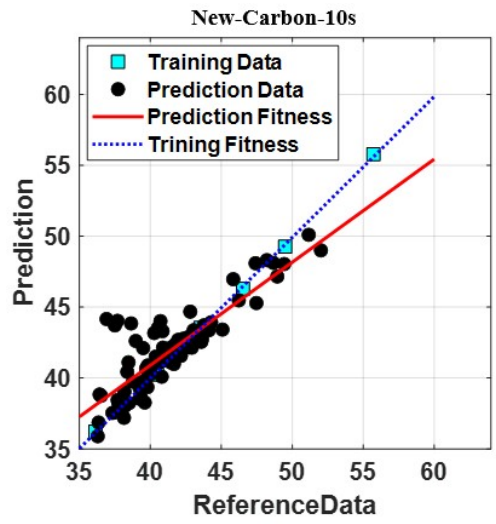


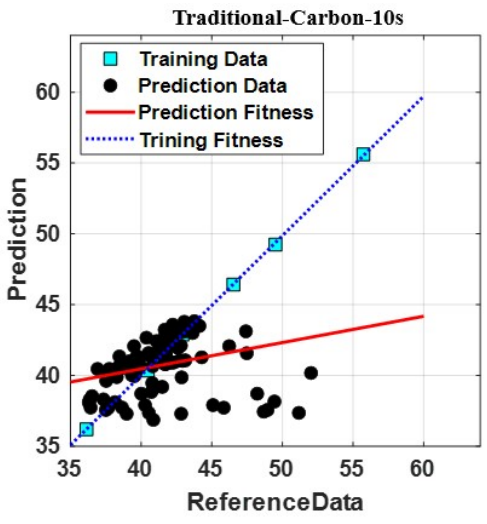
Fig. S1. The model training processes rely on two spectra extractions

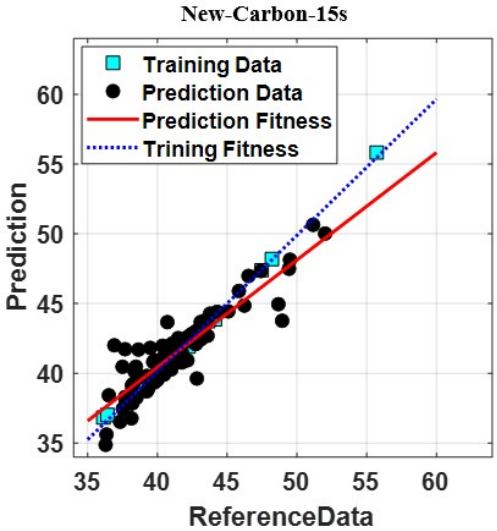
III. Comparing the prediction results of different modeling methods.

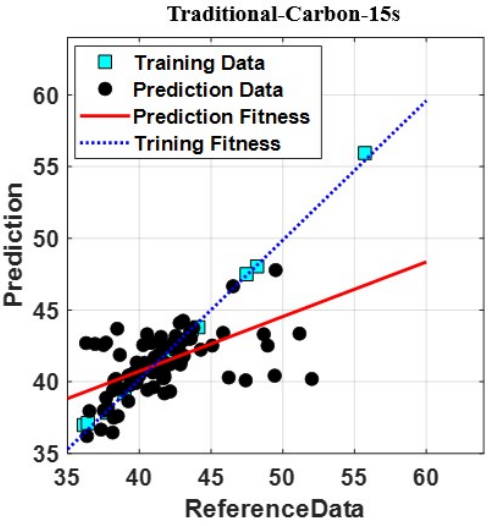
These figures below show the performance comparison between the new model and the traditional model, the title of each figure marks the training information, for example, “New-Carbon-6s” indicating the “Carbon” prediction model is trained by the “New” algorithm with “6 samples”. The model evaluation results such as RMSEP and R_p^2 are listed in Table. S2.

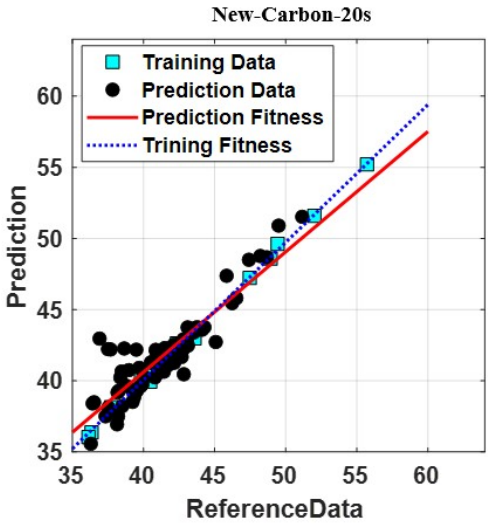


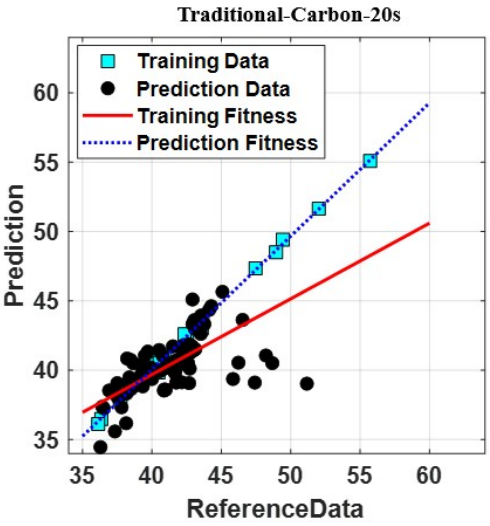

New-Carbon-10s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData
35
40
45
50
55
60

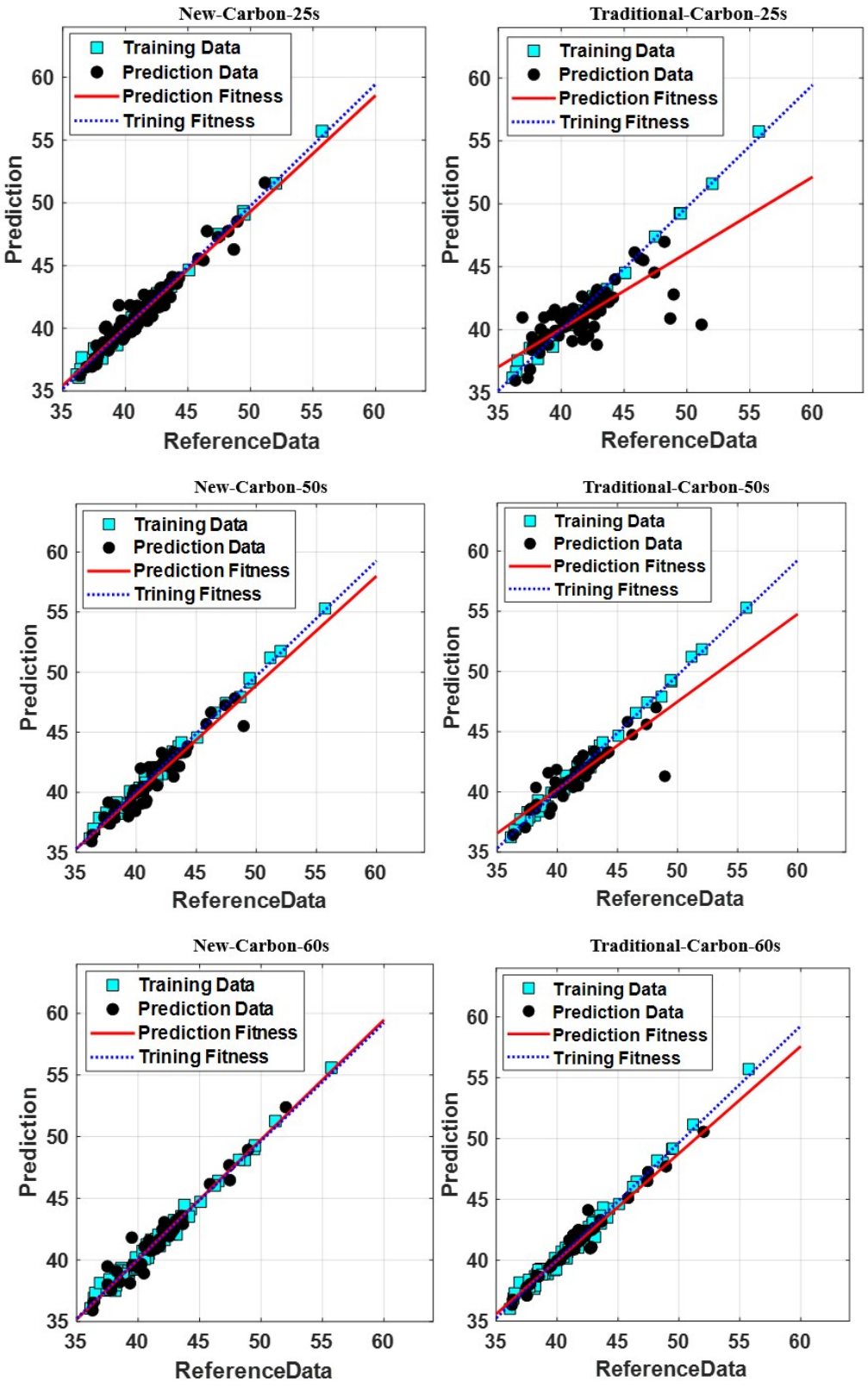

Traditional-Carbon-10s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData
35
40
45
50
55
60


New-Carbon-15s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData
35
40
45
50
55
60


Traditional-Carbon-15s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData
35
40
45
50
55
60


New-Carbon-20s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData
35
40
45
50
55
60


Traditional-Carbon-20s
Training Data
Prediction Data
Training Fitness
Prediction Fitness
Prediction
ReferenceData
35
40
45
50
55
60


New-Carbon-25s
Traditional-Carbon-25s
New-Carbon-50s
Traditional-Carbon-50s
New-Carbon-60s
Traditional-Carbon-60s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData
35
40
45
50
55
60

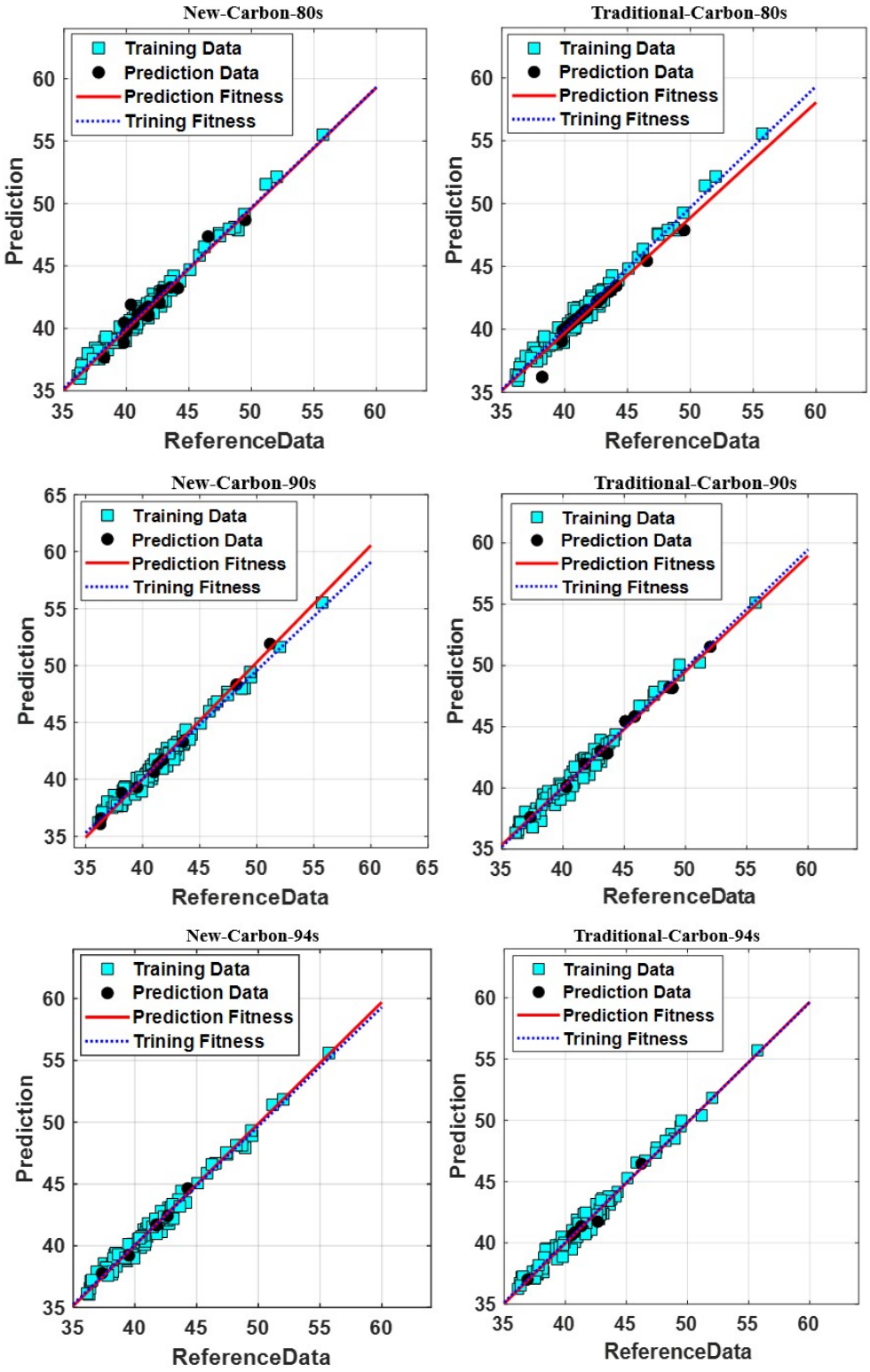
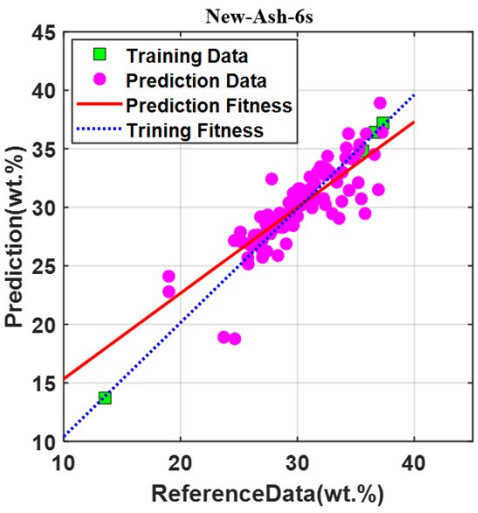
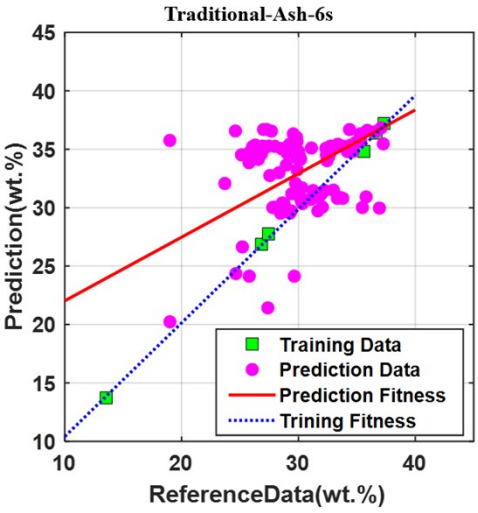
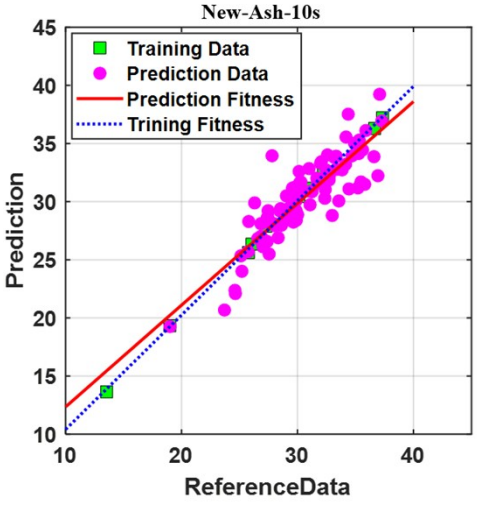
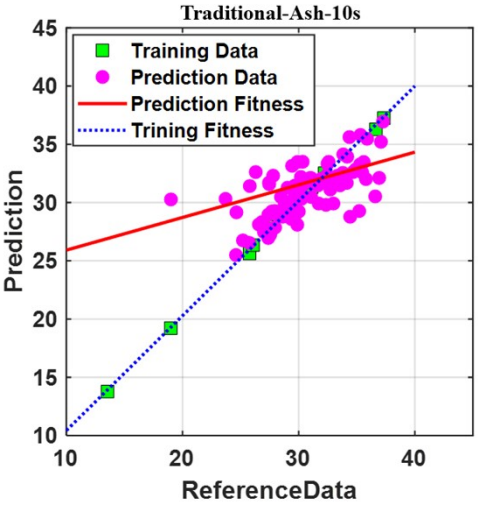


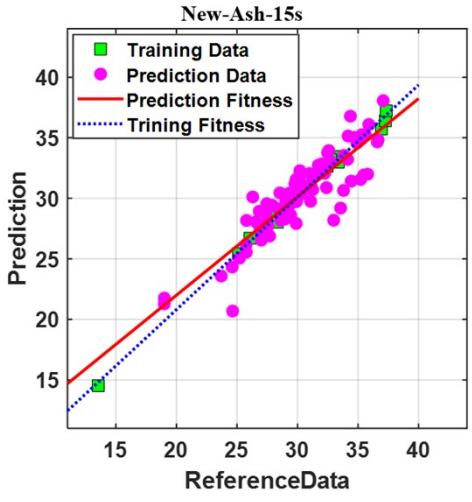
Fig. S2. Utilizes new and traditional algorithms to train the prediction model of carbon content with various training sample sizes between 6 to 94, marked in the title of each picture as 6s to 94s. The figures in the left column are the results of the new model and the figures in the right column are traditional model results.

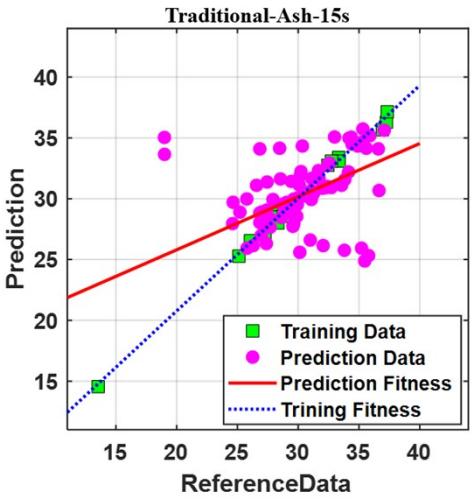

New-Ash-6s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction(wt.%)
ReferenceData(wt.%)

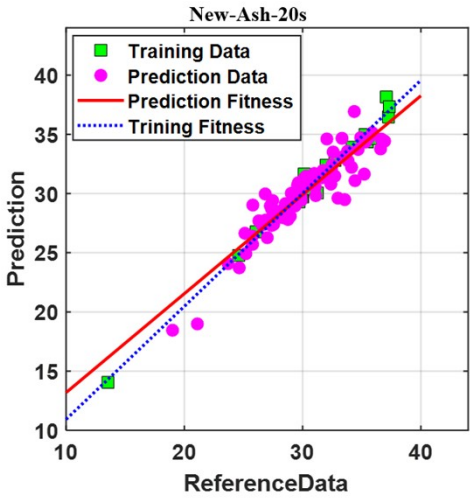

Traditional-Ash-6s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction(wt.%)
ReferenceData(wt.%)

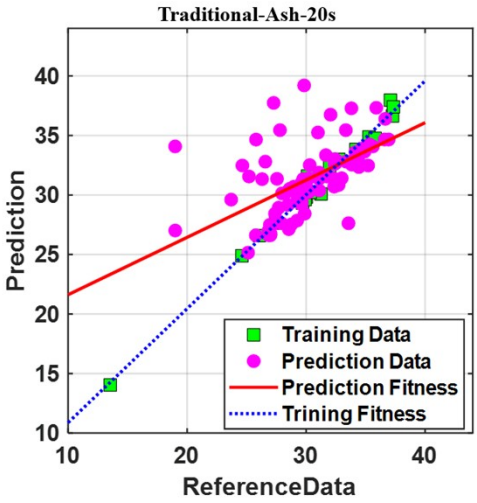

New-Ash-10s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData

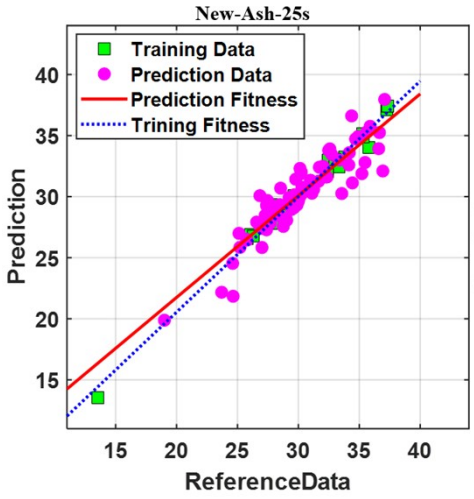

Traditional-Ash-10s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData

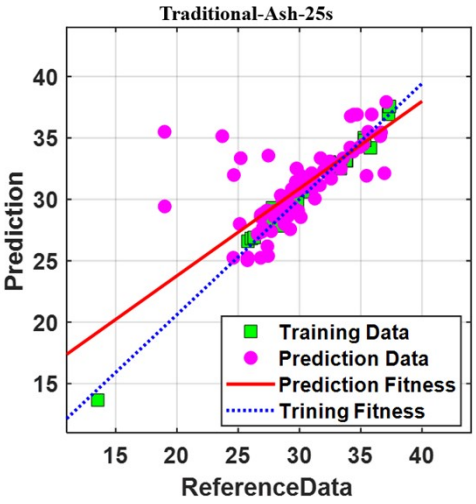

New-Ash-15s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData

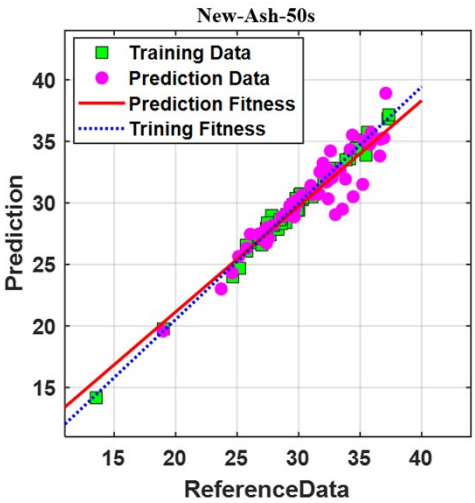

Traditional-Ash-15s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData

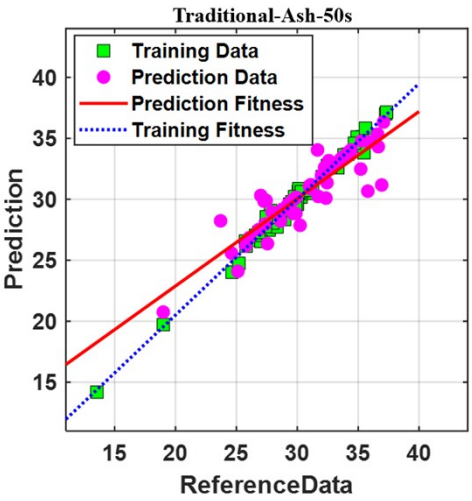

New-Ash-20s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData
10
15
20
25
30
35
40

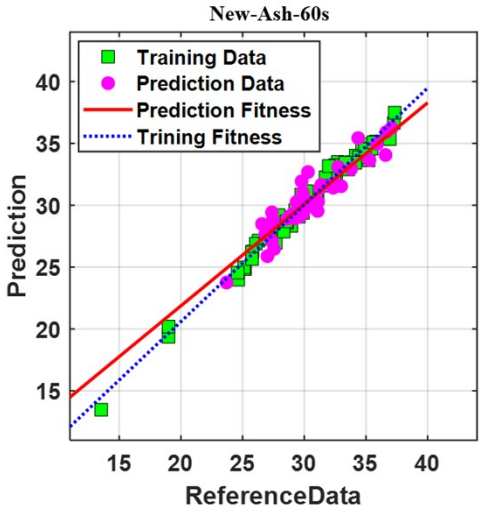

Traditional-Ash-20s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData
10
15
20
25
30
35
40

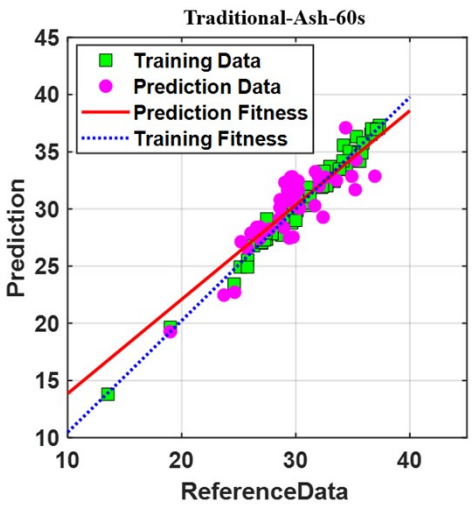

New-Ash-25s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData
15
20
25
30
35
40

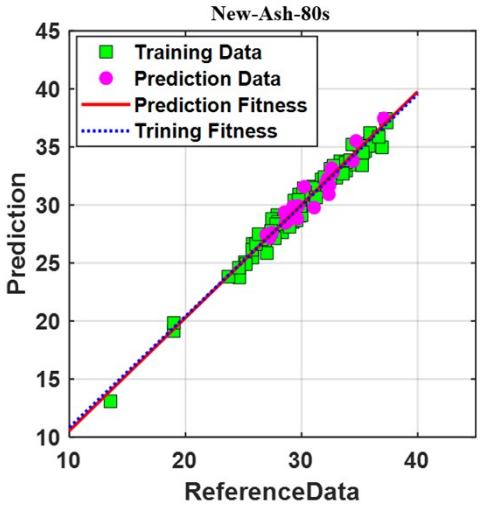

Traditional-Ash-25s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData
15
20
25
30
35
40

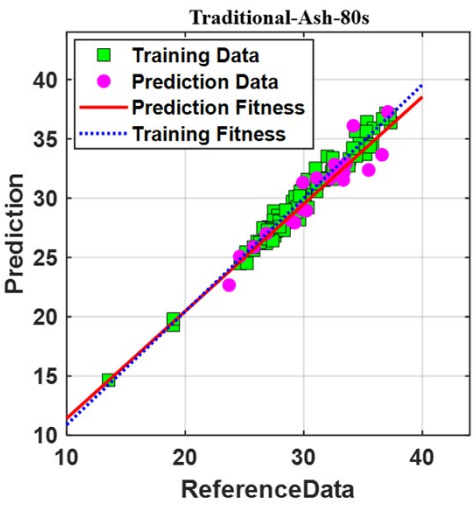

New-Ash-50s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData
15
20
25
30
35
40

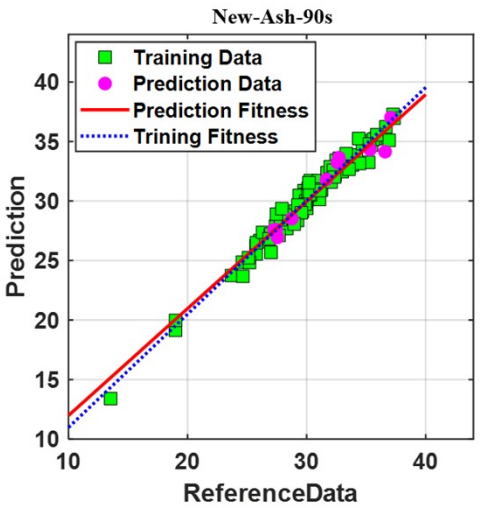

Traditional-Ash-50s
Training Data
Prediction Data
Prediction Fitness
Training Fitness
Prediction
ReferenceData
15
20
25
30
35
40

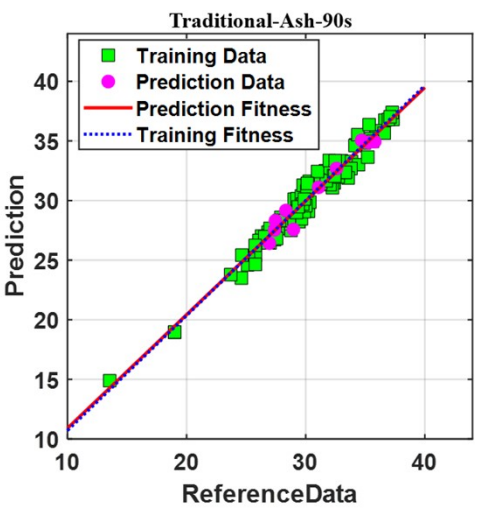

New-Ash-60s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData


Traditional-Ash-60s
Training Data
Prediction Data
Prediction Fitness
Training Fitness
Prediction
ReferenceData


New-Ash-80s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData


Traditional-Ash-80s
Training Data
Prediction Data
Prediction Fitness
Training Fitness
Prediction
ReferenceData


New-Ash-90s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData


Traditional-Ash-90s
Training Data
Prediction Data
Prediction Fitness
Training Fitness
Prediction
ReferenceData

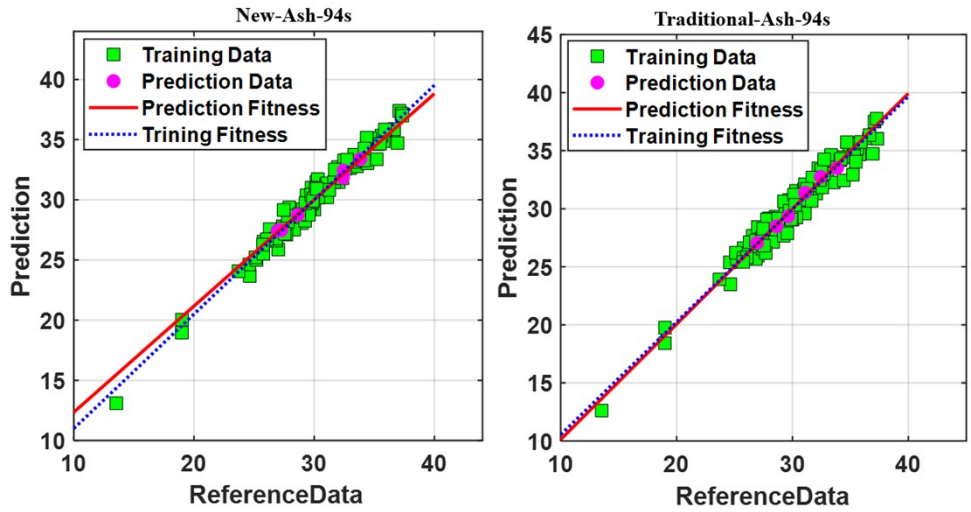
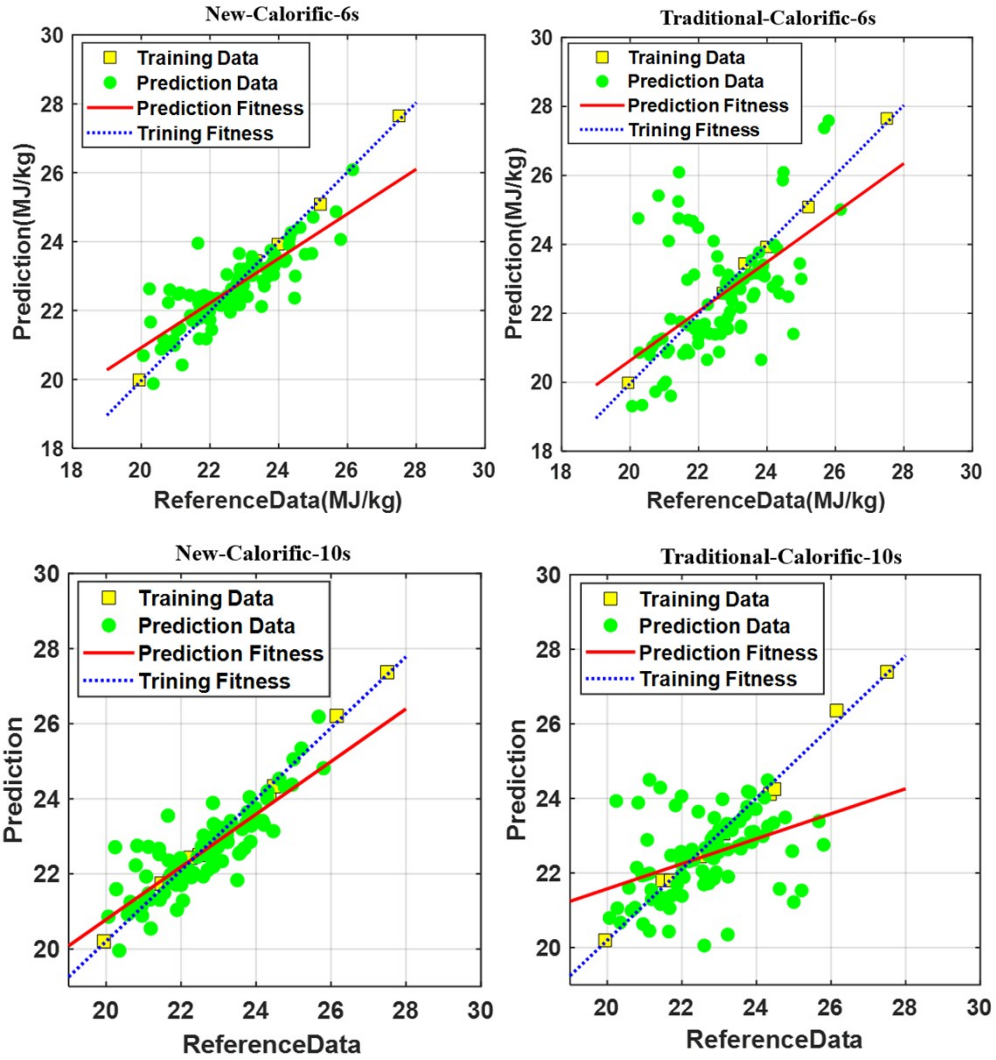
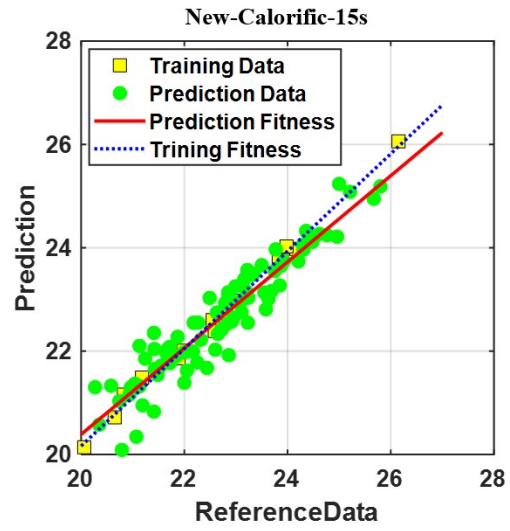
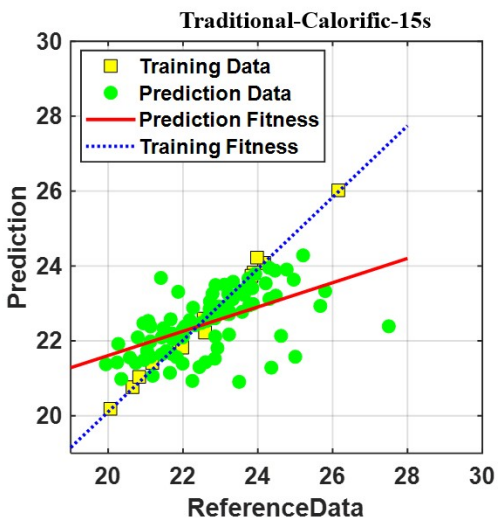
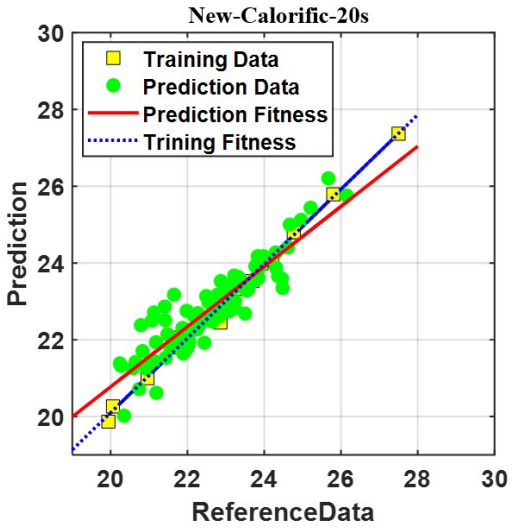


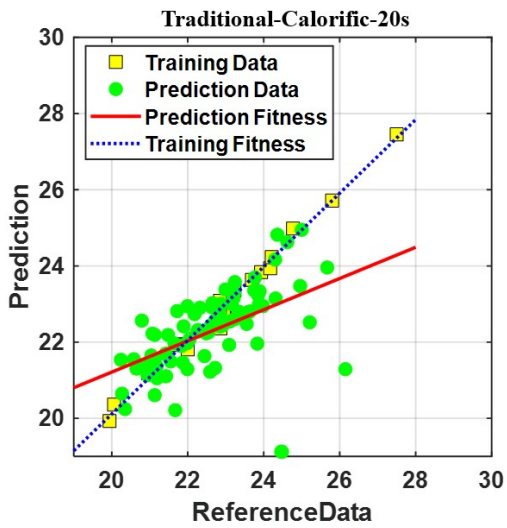
Fig. S3. Utilizes new and traditional algorithms to train the prediction model of ash content with various training sample sizes between 6 to 94, marked in the title of each picture as 6s to 94s. The figures in the left column are the results of the new model and the figures in the right column are traditional model results.

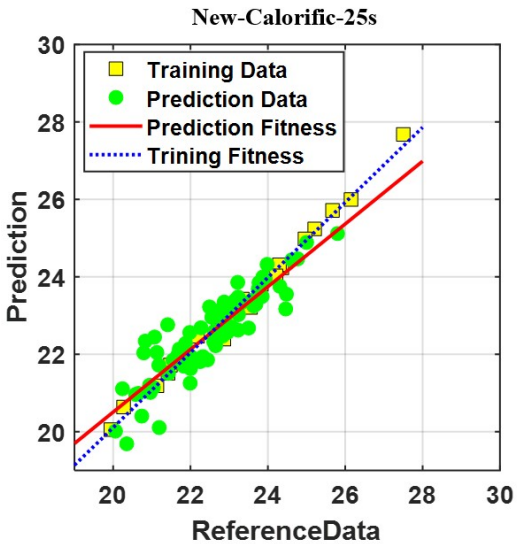


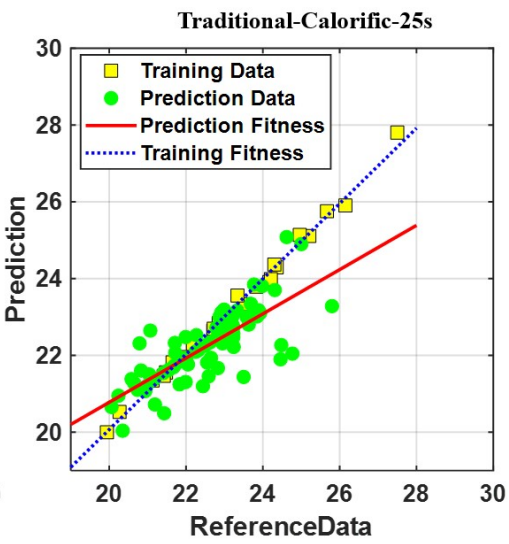

New-Calorific-15s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData
20
22
24
26
28

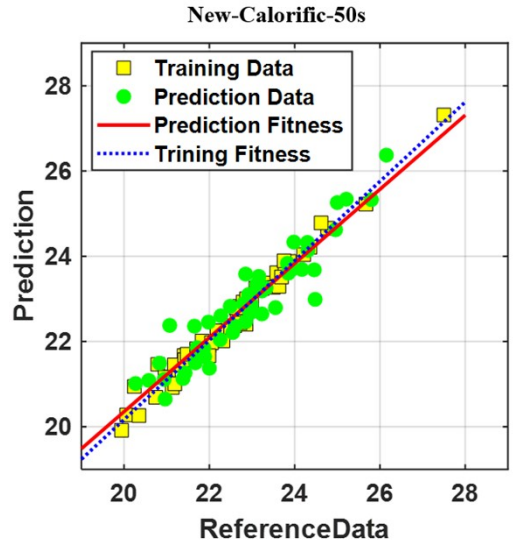

Traditional-Calorific-15s
Training Data
Prediction Data
Prediction Fitness
Training Fitness
Prediction
ReferenceData
20
22
24
26
28
30

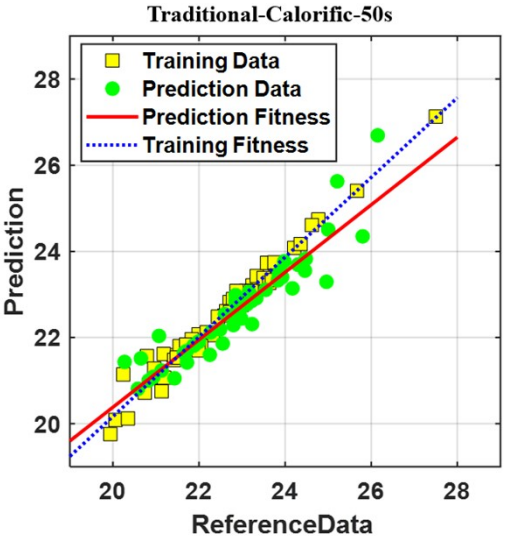

New-Calorific-20s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData
20
22
24
26
28
30

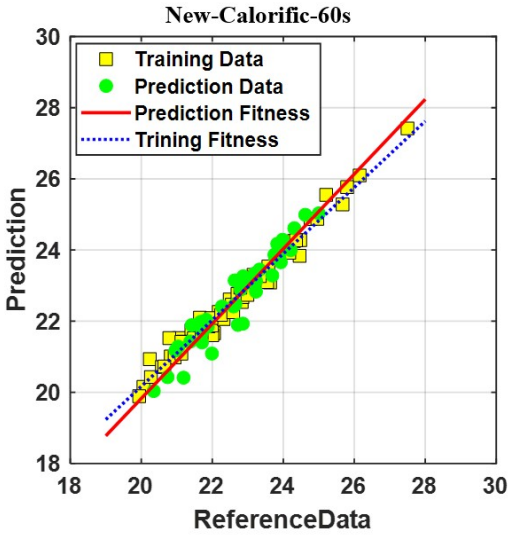

Traditional-Calorific-20s
Training Data
Prediction Data
Prediction Fitness
Training Fitness
Prediction
ReferenceData
20
22
24
26
28
30

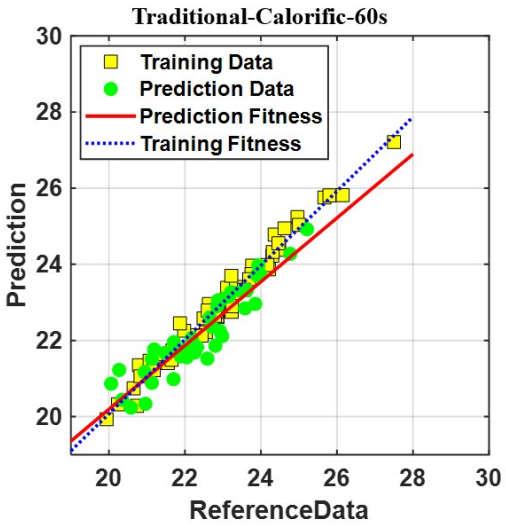

New-Calorific-25s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData
20
22
24
26
28
30

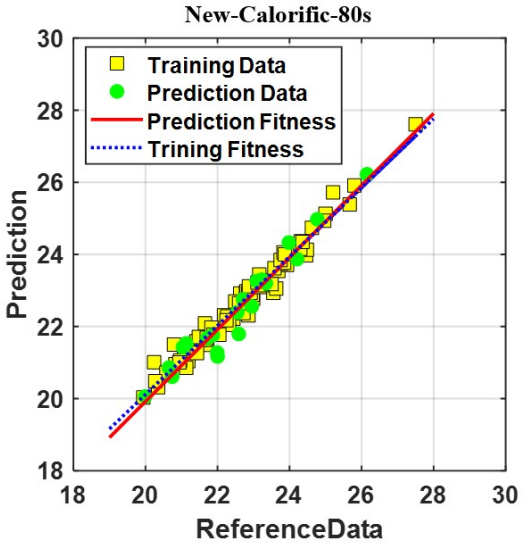

Traditional-Calorific-25s
Training Data
Prediction Data
Prediction Fitness
Training Fitness
Prediction
ReferenceData
20
22
24
26
28
30

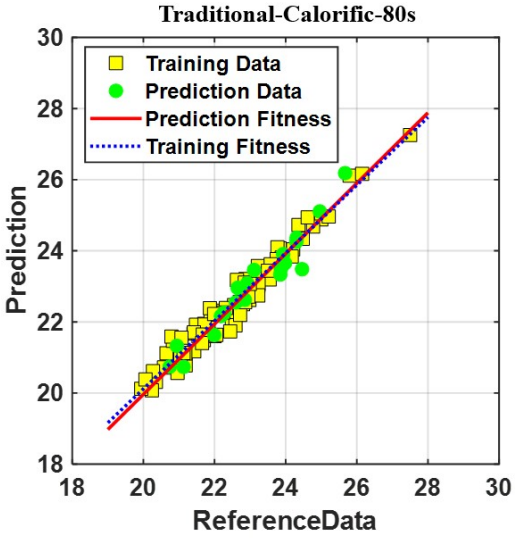

New-Calorific-50s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData
20
22
24
26
28


Traditional-Calorific-50s
Training Data
Prediction Data
Prediction Fitness
Training Fitness
Prediction
ReferenceData
20
22
24
26
28


New-Calorific-60s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData
18
20
22
24
26
28
30


Traditional-Calorific-60s
Training Data
Prediction Data
Prediction Fitness
Training Fitness
Prediction
ReferenceData
20
22
24
26
28
30


New-Calorific-80s
Training Data
Prediction Data
Prediction Fitness
Trining Fitness
Prediction
ReferenceData
18
20
22
24
26
28
30


Traditional-Calorific-80s
Training Data
Prediction Data
Prediction Fitness
Training Fitness
Prediction
ReferenceData
18
20
22
24
26
28
30

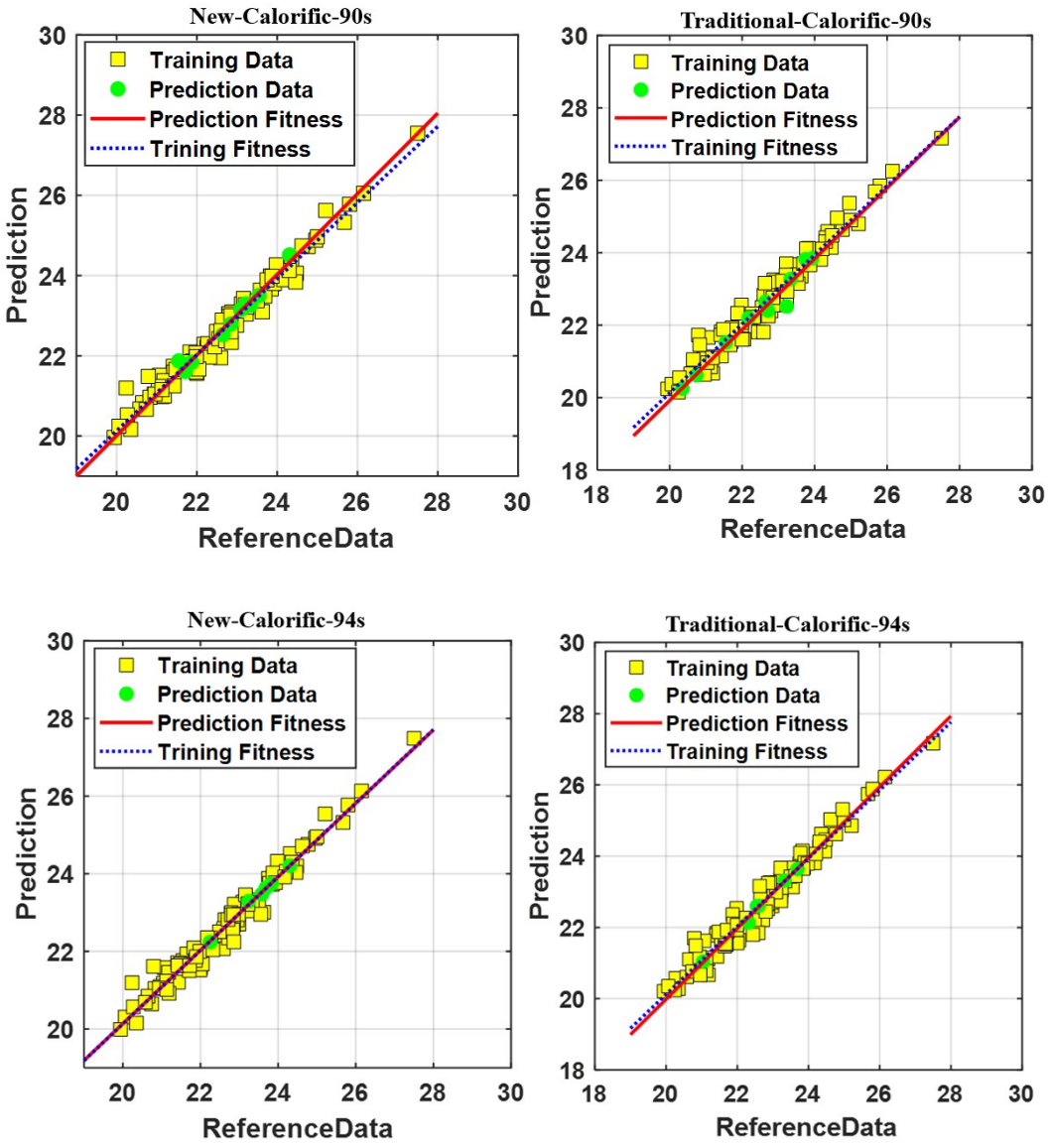


Fig. S4. Utilizes new and traditional algorithms to train the prediction model of calorific value with various training sample sizes between 6 to 94, marked in the title of each picture as 6s to 94s. The figures in the left column are the results of the new model and the figures in the right column are traditional model results.

IV. Conclude the model performance of different coal indicators.

Table S2. The prediction result of carbon, ash and calorific value with different training sample sizes. The total number of samples is 100, thus, the training sample size is 6 (or 10) and the test sample size is 94 (90).

Training Sample size		6		
Indicators		Carbon content	Ash content	Calorific
R_p^2	New	0.74	0.71	0.73
	Traditional	0.00	0.33	0.27
RMSEP	New	1.82	2.15	0.77

	Traditional	3.83	4.46	1.72

Training Sample size		10		
Indicators		Carbon content	Ash content	Calorific
R_p^2	New	0.77	0.80	0.70
	Traditional	0.12	0.25	0.15
RMSEP	New	1.78	1.79	0.71
	Traditional	3.65	4.38	1.47

Training Sample size		15		
Indicators		Carbon content	Ash content	Calorific
R_p^2	New	0.84	0.80	0.80
	Traditional	0.31	0.25	0.34
RMSEP	New	1.46	1.63	0.68
	Traditional	3.02	3.75	1.22

Training Sample size		20		
Indicators		Carbon content	Ash content	Calorific
R_p^2	New	0.84	0.84	0.83
	Traditional	0.56	0.43	0.35
RMSEP	New	1.41	1.38	0.59
	Traditional	2.71	3.49	1.30

Training Sample size		25		
Indicators		Carbon content	Ash content	Calorific
R_p^2	New	0.90	0.84	0.84
	Traditional	0.63	0.58	0.62
RMSEP	New	0.94	1.46	0.53
	Traditional	2.31	2.97	0.97

Training Sample size	50

Indicators		Carbon content	Ash content	Calorific
R_p^2	New	0.93	0.87	0.90
	Traditional	0.78	0.81	0.80
RMSEP	New	0.81	1.43	0.47
	Traditional	2.16	1.99	0.73

Training Sample size		60		
Indicators		Carbon content	Ash content	Calorific
R_p^2	New	0.95	0.87	0.91
	Traditional	0.93	0.71	0.86
RMSEP	New	0.72	1.18	0.37
	Traditional	1.03	1.92	0.49

Training Sample size		80		
Indicators		Carbon content	Ash content	Calorific
R_p^2	New	0.96	0.93	0.95
	Traditional	0.93	0.94	0.93
RMSEP	New	0.62	0.70	0.37
	Traditional	0.64	1.36	0.35

Training Sample size		90		
Indicators		Carbon content	Ash content	Calorific
R_p^2	New	0.99	0.93	0.97
	Traditional	0.99	0.96	0.96
RMSEP	New	0.36	0.94	0.14
	Traditional	0.45	0.97	0.25

Training Sample size		94		
Indicators		Carbon content	Ash content	Calorific
R_p^2	New	0.99	0.99	0.99
	Traditional	0.98	0.98	0.99
RMSEP	New	0.35	0.37	0.08
	Traditional	0.48	0.86	0.08

V. PCA scores of the spectral data before and after data cleansing and preprocessing.

We used the 100 sets of data (used in this work) to show the PCA scores difference before and after data cleansing and preprocessing. As shown in Fig. S5(a), each data point represents a sample in PCA space, and the data discrimination becomes larger after data cleansing and preprocessing, which indicates that the data preprocessing can improve the data consistency of the same sample, and also make the spectral data distinguish of different samples obvious. As shown in Fig. S5(b), the top two principal components (PCs) and top six PCs explained the 55% and 95% variance of spectral data after data preprocessing and cleansing, respectively. In comparison, its explained 45% and 93% variance of spectral data before data preprocessing. In other words, the same number of PCs have a stronger interpretation for spectral data after data preprocessing.

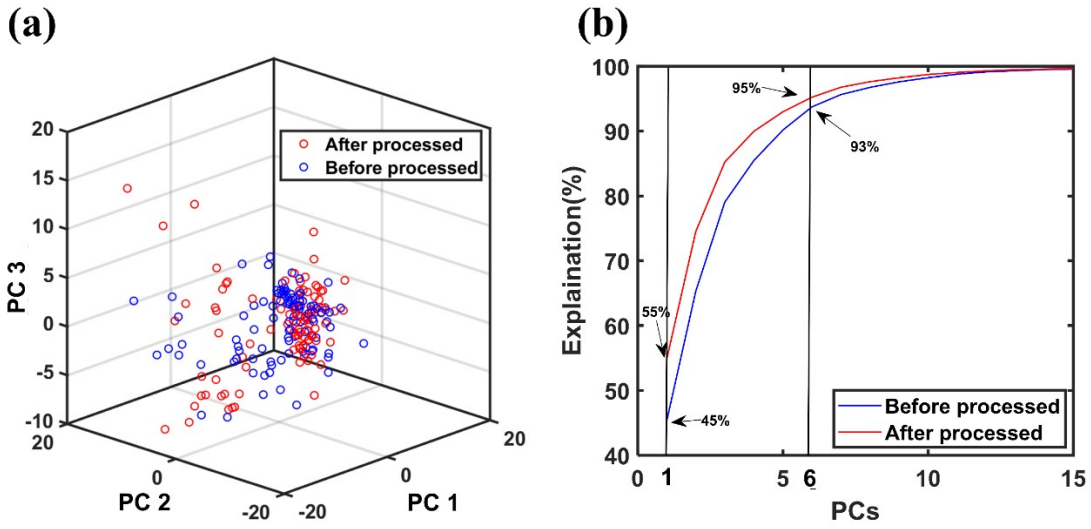


Fig. S5. (a) The scores plot of the top three principal components before and after data processing and cleansing (b) The ability of principal components to explain the variance of origin data.