Fuel 280 (2020) 118676

Contents lists available at ScienceDirect

Fuel

journal homepage: www.elsevier.com/locate/fuel

Full Length Article

Reflectance spectroscopy based rapid determination of coal quality parameters

Nafisa Begum[a,*], Abhik Maiti[a], Debashish Chakravarty[a], Bhabani Sankar Das[b]

[a] Department of Mining Engineering, Indian Institute of Technology Kharagpur, 721302, India
[b] Department of Agricultural and Food Engineering, Indian Institute of Technology Kharagpur, 721302, India

ARTICLE INFO

Keywords:
Diffuse reflectance spectroscopy
Coal quality analysis
Partial least square regression
Random forest
Extreme gradient boosting

ABSTRACT

In this work, the reflectance spectroscopy of 212 coal samples of different origins was investigated across the Vis-NIR-SWIR range (wavelength: 350–2500 nm) to estimate their ash, moisture, volatile matter, fixed carbon content and gross calorific value (GCV). Several mathematical pre-treatments were applied to each spectrum for improving the signal-to-noise ratio. Partial-least-square (PLS), random forest (RF), and extreme gradient boosting (XGBoost) based regression methods were used to capture the relationships between coal quality parameters with corresponding spectral responses. The predictive models were generated by taking a combination of a set of differently pre-processed spectra with the above-mentioned regression methods to obtain the optimal prediction performance. The results show that spectral pre-processing improves the prediction accuracy of a model. Excessive pre-processing, however, could reduce the model accuracy due to the loss of information. RF regression model works best for estimating moisture and fixed carbon content, while XGBoost shows the best result for ash content and GCV, and PLS models provide the most accurate prediction for volatile matter content.

1. Introduction

Coal is a heterogeneous aggregate of organic and inorganic materials. It mainly contains carbon, hydrogen, oxygen, and little amount of sulphur and nitrogen. Characterization of coal involves several standard methods that can generate very accurate results. However, one of the major drawbacks associated with these techniques is they are laborious and time-consuming processes [1]. Therefore, there is a requirement for alternative techniques for the characterization of coal, which can provide reliable results in a relatively easy and swift manner. A variety of optical spectroscopic techniques have gained attention in recent times because these techniques can provide information on the physico-chemical composition of the target from only a single measurement [2,3].

Application of infrared transmission and absorption spectroscopy has its long history in coal characterization (e.g., study of functional group, maceral composition, oxidation, rank) [4-8]. However, this method involves the preparation of KBr-coal pellets, which is time-consuming. Further, it involves light scattering at the KBr-coal interface, which leads to a shift in the spectral baseline in high-frequency region [9]. In last few decades, diffuse reflectance spectroscopy (DRS) has gained interest for quantitative analysis of coal, primarily because of the fact that the data acquisition process is simpler, and no sample pre-treatment is required. The Beer-Lambert law states that the absorbance is linearly related to the concentrations of chemical components present in the target material [10]. Researchers have done chemometric modelling using this law to estimate the quality parameters of coal [1,11-15]. It should be mentioned that most of these analyses were performed by taking the spectral responses in short-wave infrared range (SWIR), where the wavelength ranges from 1000 to 2500 nm. There is, therefore the further scope of utilizing spectral characteristics from visible to near-infrared (VNIR) range (wavelength: 350–1000 nm) to obtain additional information related to the coal characterization. In this context, hyperspectral remote sensing-based spectral reflectance covering a wavelength of 350–2500 nm can be examined to improve the prediction accuracy of the coal properties.

Compositionally coal is very complex and heterogeneous in nature, and a specific spectral band cannot be assigned to quantify a particular coal property. For that, the entire spectrum should be considered in the chemometric analysis [16], which involves a huge number of variables. Thus, multivariate statistical methods are used for the prediction of quality parameters of coal. In most of the related research work, linear regression models such as PLS regression, principal component analysis (PCA), and multiple linear regression (MLR) have been used to estimate coal properties [1,6,12-18].

In the present work, attempts have been made to estimate some

* Corresponding author.
E-mail address: nafisa.geo@gmail.com (N. Begum).

https://doi.org/10.1016/j.fuel.2020.118676
Received 27 March 2020; Received in revised form 22 May 2020; Accepted 7 July 2020

Abbreviations

DRS	Diffuse reflectance spectroscopy
GCV	Gross calorific value
MLR	Multiple linear regression
MSC	Multiplicative scatter correction
nm	Nanometre
PCA	Principle component analysis
PLS	Partial least square
R	Reflectance
RF	Random forest
RMSET	Root mean square error of test set
Vis-NIR-SWIR	Visible-near infrared-short wave infrared
VNIR	Visible to near infrared
XGBoost	Extreme gradient boosting

important quality parameters (viz. ash, moisture, volatile matter, fixed carbon content, and GCV) of Indian coal which is known for its high ash content using diffuse reflectance spectroscopy in Vis-NIR-SWIR (350–2500 nm) range. The coal properties predicted from reflectance spectra have been compared with the measured data obtained from laboratory experiments to check its accuracy. In DRS, the detector captures the reflectance (R) value at each wavelength interval, and the absorbance value is computed from the logarithmic of 1/R [i.e., Log(1/R)]. This calculated absorbance value is not always the true representation of the Beer-Lambert absorbance since the path length of the electromagnetic wave through the sample is influenced by both the absorbance and scattering phenomenon [19]. This gives rise to non-linearity in absorbance-concentration relationship due to the additive and multiplicative effect. To overcome such physical effects, various mathematical pre-treatments have been carried out to the spectra before applying different multivariate statistical methods. However, spectral pre-processing is not always capable of removing all the non-linearities because it considers almost constant scattering across the wavelength interval, and it may also cause the removal of some important chemical information [20]. Therefore, different mathematical pre-treatments and their combinations were applied to the spectra to reduce the nonlinearity before performing the chemometric analysis. An effort has also been made to optimize the application of different spectral pre-processing methods. A schematic representation of the process flow chart of the present investigation is shown in Fig. 1. The PLS, RF, and XGBoost based regression models are used in order to capture both the linear and non-linear relationships of coal quality parameters with its spectral properties. At the best of our knowledge, for the first time, RF and XGBoost have been used for chemometric modelling of coal quality estimation.

2. Methodology

2.1. Sample description

A total of 212 coal samples (Table 1) were collected from geographically widely distributed Indian coal basins (Fig. 2) of different geological ages and coal ranks. Majority (~90%) of the samples were collected from open cast mines (both active and abundant) and fewer (~10%) from underground mines. The samples were collected from coal seams in such a way that they become representative of the seam as much as possible. Effort was made to assess both vertical and horizontal variations in the seam by collecting the samples through channel sampling method. The coal samples were collected from the states of Assam, West Bengal, Jharkhand, Gujarat and Jammu & Kashmir. A chronologically widely distributed coal basins ranging from Permian to Oligocene age were selected for this study to understand the spectral behaviour of coal of different rank, grade, and geochemistry. The rank of coal ranges from lignite to semi-anthracite as listed in Table 1.

2.2. Experimental details

All coal samples were pulverized to carry out different coal quality analyses and recording of reflectance spectra. Proximate analysis of the samples was carried out following ASTM D3172 standard [22] to determine the ash, moisture, volatile matter, and fixed carbon contents. The gross calorific value (GCV) was determined using Bomb calorimeter (Model: Parr 660) following ASTM D5865 standard [25].

Reflectance spectra of coal samples were recorded using ASD FieldSpec® spectroradiometer. The spectra were recorded in a dark room with room temperature ranging from 26 − 29 degreesC and a humidity level of 65 – 70%. The spectral reflectance was recorded

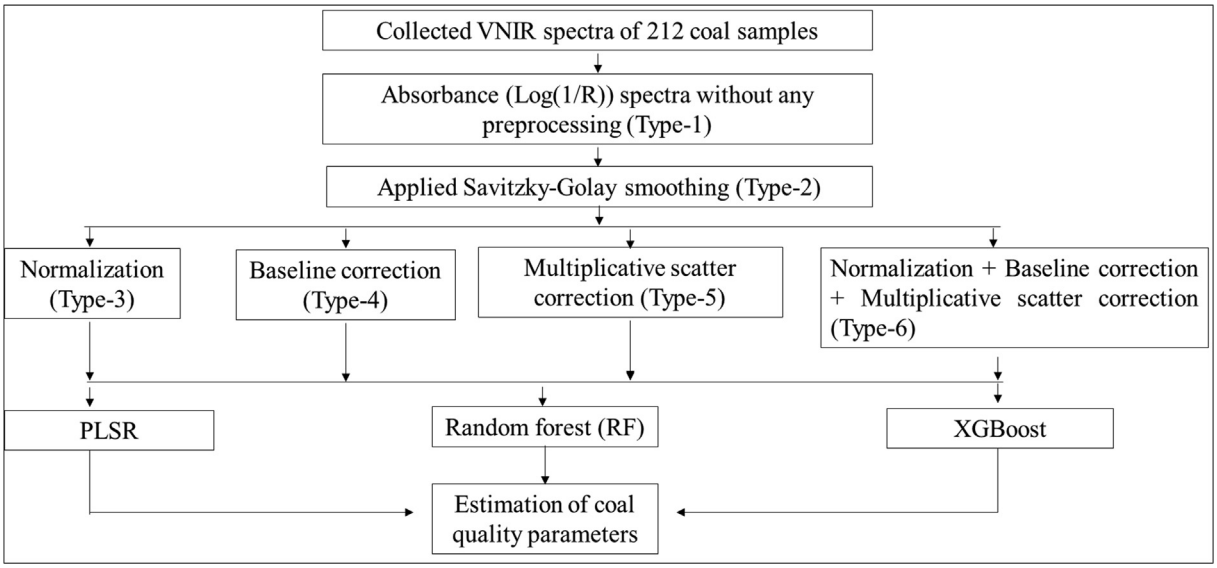

Fig. 1. Schematic representation of the process flow chart for estimation of coal properties from VNIR spectral data used in the present investigation.

Table 1
Details of the coal samples collected for the present investigation.

Sl. no.	Area	Coal bearing formation	Geological age	No. of coal mines	No. of coal samples collected	Rank of coal
1	Raniganj coal field, West Bengal	Raniganj Formation	Late Permian	6	14	High volatile B bituminous to medium volatile bituminous coal.
2	Jharia coal field, Jharkhand	Barakar Formation	Middle Permian	21	123	High volatile A bituminous to Low volatile bituminous coal.
3	Makum coal field, Assam	Tikak-Parbat Formation	Oligocene	3	28	Mainly high volatile A bituminous coal.
4	Gujarat	Tadkeshwar Formation Khadsaliya Formation	Oligocene to Lower Eocene Eocene	4	23	Lignite A to sub-bituminous C coal.
5	Jammu and Kashmir	Laki Formation Murree Series	Lower to Middle Eocene age Eocene	2	24	Mainly low volatile bituminous to semi-anthracite coal.

within the range of 350–2500 nm at a spectral resolution of 3 nm up to 700 nm and 10 nm in farer wavelength (up to 2500 nm) [23]. In DRS, the quantitative information of a sample is mostly affected by the sensor-source geometry and sample particle size [24]. Therefore, we have performed the spectral measurement taking a uniform grain size (less than 74 μm) [17] and a fixed sensor-source geometry. It is also reported that the coal quality parameters (e.g. volatile matter, fixed carbon and ash content) can vary with the variation in particle size [38]. The lower limit of particle size was therefore limited to 63 μm in the present study in order to maintain a relatively uniform particle size distribution. The sensor was positioned at nadir (42 cm above the sample), providing a field of view (FOV) of 27.07 cm^2. The pulverized coal samples were taken into a flat container and placed within the FOV of the sensor. Every spectrum is an average of 30 scans. After capturing each spectrum, the sample container was rotated by 90 degrees resulting in four spectra for one sample. Average of the four spectra were taken as the representative spectra for each sample.

2.3. Data processing methods

2.3.1. Spectral pre-processing

The spectral data were captured in the reflectance (R) mode and transformed into absorbance [Log(1/R)] spectra for this analysis. The initial reflectance values from 350 to 399 nm were removed because of the instrumental noise. This was followed by application of several mathematical pre-treatments to minimize the extraneous effect on the spectra, which can subsequently improve the predictive models or quantitative analysis [20]. The pre-processing technique is broadly divided into two categories: a) correction of particle scattering effect and b) spectral derivatives (which involves smoothing of the spectra in order to reduce the signal to noise ratio). However, the use of several pre-processing techniques may lead to the loss of some valuable information [20]. Thus, it is imperative to optimize the numbers of pre-processing techniques. This could be achieved by comparing the prediction performance of the models after applying different pre-processing techniques. In this study, we have considered one spectral derivative process viz. Savitzy-Golay smoothing and three different scatter correction methods viz. normalization, baseline correction, and multiplicative scatter correction (MSC). These are the most efficient and commonly used spectral pre-processing methods to remove non-linearities [14,16]. As shown above, in Fig. 1, the spectra have been divided into the following six types:

a) Type-1: No pre-treatments were applied, and raw absorption spectra had been used for the prediction modelling.
b) Type-2: Savitzky-Golay smoothing was applied to the absorption spectra taking a window of nine points and second-order polynomial.
c) Type-3: Range normalization was applied to Type-2 spectra.
d) Type-4: Baseline correction was applied to Type-2 spectra.
e) Type-5: MSC algorithms were applied to Type-2 spectra.
f) Type-6: All three scatter correction methods were applied to Type-2 spectra (normalisation followed by baseline correction and finally MSC).

2.3.2. Data modelling

a) Data description: The dataset consists of 212 coal reflectance spectra in VNIR (400–2500 nm) range. Each row consists of spectro-radiometer readings at 1 nm interval from 400 to 2500 nm. Each column has corresponding 2101 sets of absorbance [Log (1/R)] values of coal samples followed by the properties (ash, moisture, volatile matter, fixed carbon and GCV) of coal samples. The descriptive statistics of these properties obtained from laboratory measurements following ASTM standards are described in Table 2. The box-plots of each of the coal properties are shown in Fig. 3. In the following sections, attempts have been made to compare the measured coal properties with

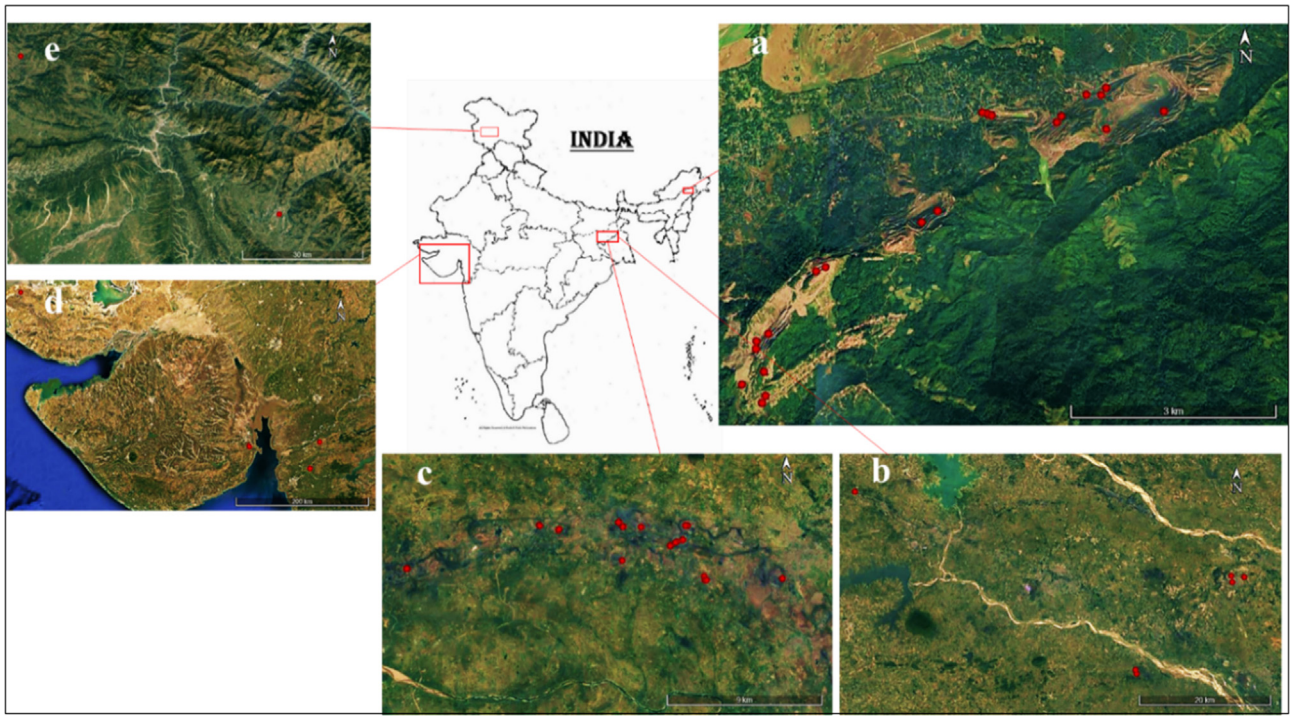

Fig. 2. Map of India [21] showing areas where coal samples were collected. It comprises of the maps (taken from Google Earth) of coal mines in (a) Assam, (b) West Bengal, (c) Jharkhand, (d) Gujarat and (e) Jammu & Kashmir.

Table 2
Statistical results of coal properties obtained from standard laboratory methods, as discussed above.

Coal properties	Proximate analysis (as-received basis)				GCV (MJ/kg)
	Ash (wt.%)	Moisture (wt.%)	Volatile matter (wt. %)	Fixed carbon (wt. %)	
Count	212	212	212	212	212
Mean	19.50	3.30	21.15	56.07	25.90
St. Deviation	11.94	6.00	9.74	11.02	5.00
Minimum	0.90	0.02	0.54	15.86	12.20
Maximum	54.77	33.07	47.86	78.21	35.60

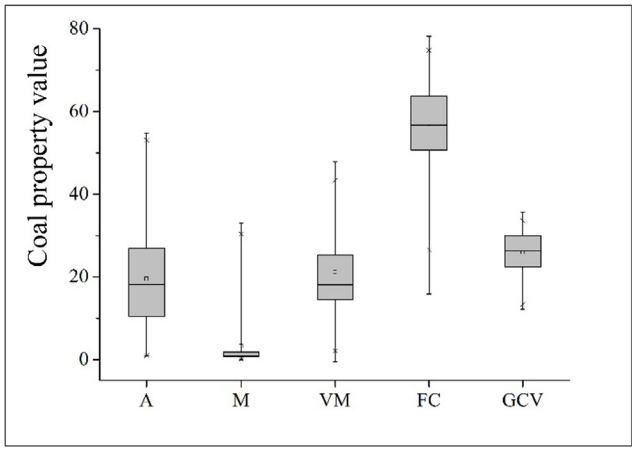

Fig. 3. Data distribution of ash (A), moisture (M), volatile matter (VM), fixed carbon (FC) and gross calorific value (GCV) of the coal samples investigated in the present study.

that predicted from the reflectance spectroscopy using different statistical methods.

b) Outlier treatment and data normalization: Outlier removal was done on the dataset based on the Z-score of each of the coal properties. The Z-score is the signed number of standard deviations by which the value of a data point deviates from the mean value of the respective property, and it is calculated from:

$$Z_{Score(n)} = (Y_n - \bar{Y})/\sigma \quad (1)$$

where Y_n is the n^{th} data point, $\bar{Y}$ is the mean and σ is the standard deviation. The coal properties having Z-score greater than + 3 or lower than −3 were removed as outliers to improve the model accuracy [26]. Post outlier treatment, the entire dataset was normalized to ensure the values of all the columns across the 212 data points have a range between 0 and 1. The formula for normalizing the dataset for i^{th} column is given below:

$$\hat{a}[:,i] = \frac{a[:,i] - \min(a[:,i])}{\max(a[:,i]) - \min(a[:,i])} \quad (2)$$

After outlier removal and normalization, 205 coal sample data were split (80:20 ratio) into training (164) and test set (41) for further analysis. Each of the regression models (PLS, RF and XGBoost) was run 30 times on the training set for each property to remove randomisation bias. The predictions were made on the test set of 41 samples each time. Among these 30 predictions per model of each property, the best result is considered. Details of each statistical model are discussed below.

2.3.3. Model selection

2.3.3.1. Partial least square (PLS) regression. Theory: Given a set of independent variables (X) and a set of dependent variables (Y), the goal of the regression is to predict Y from X. However, when the number of predictors (X) are large compared to the number of training samples, linear regression fails due to multi-collinearity. To work-around this issue, PCA technique projects X to a lower-dimensional space U, which captures the variance of X using Singular Value Decomposition. Then, the U vectors are used for predicting Y.

In contrast, PLS regression projects both X and Y to lower dimensional spaces(T&U) by finding the latent features of X which are relevant to finding$Y$. In PLS, X and Y are simultaneously decomposed so that the resultant latent variables can explain the variance of X and Y as much as possible. Hence, PLS can be conceptualised as a generalization of PCA. The algorithm of PLS can be mathematically described as:

$$X = TP^T \& Y = UQ^T \tag{3}$$

After the decomposition of X and Y to latent space T and U, regression is performed between T&U [27].

The variation of root mean square error of a test set (RMSET) with number of latent variables is shown in Fig. 4 for different coal properties. It could be seen that as we vary the latent variable numbers, the RMSE of the properties changes. In this study, we experimented with several values of latent variables like (2, 5, 10, 15, 20, 30, and 40). We ran a grid search over the latent variable numbers, by running the PLS models with different latent variables and chose the one where the RMSE of the 5-fold cross-validation set was lowest. We use that 'best model' to predict the test set.

2.3.3.2. Random forest regression. Theory: The concept of random forest regression comes from a well-known ensemble learning algorithm called bagging or bootstrap aggregation. In bagging, n machine learning models are trained on n splits of a dataset$D$. Given that the splits are randomised and independent to each other, the n models capture different components of the variance of dataset splits. When these n models are combined in a greedy way, the resultant aggregated model (i.e., the random forest regressor) becomes powerful enough to minimize the variance of the prediction, thereby producing an improved prediction.

A random forest regression model consists of a set of randomized regression trees $\{m(x, \theta_m, D_n), m \geq 1\}$ where D_n is the n numbers of training datasets and θ_m are the respective parameters of the regression trees trained on D_n. The n numbers of random regression trees are trained on n independent splits of the dataset$D$. Then, the output of the random forest regressor $r_n(X)$ is estimated by combining the prediction of each of the regression trees using the Monte-Carlo method [28]:

$$\widehat{Y}_{RF} = r_n(X) = E_\theta[r_n(X, \theta, D_n)] \tag{4}$$

Where$D_n = \{(X_1, Y_1), (X_2, Y_2), \cdots\cdots,(X_n, Y_n)\}$

The randomized variable is used for *determining* the splits of each of the regression trees in random forest regressions. More randomized splits of individual regression trees ensure more versatile random forest regressor at predicting a dependent variable. Conceptually, the random forest regression works in the following way:

a) Creating n (say 100) random sub-samples of the dataset with replacement.
b) Training uncorrelated decision trees (parameterised by) on each of the$n$ samples.
c) During prediction, the test data is predicted using each of the$n$ decision trees. The random forest prediction for the dataset becomes the average value of the predictions of each of the$n$ decision tree regressors.

As the dataset contained 2101 sets of independent variables for predicting coal properties, the independent variables were transformed into a latent space using PCA. 20 principle components obtained from PCA were taken for predicting the coal properties.

The optimal set of hyperparameters was selected using grid-search on a 5 fold cross-validation setup in the training samples. The parameters used for grid-search are shown below. The model best performing on the 5-fold cross-validation setup was used for predicting the coal properties in the test set.

Param grid = {'n estimators': (1000, 3000), 'max_depth': (10, 50, 200, 500),'min_samples_split': (2, 5),'max_features': (0.1, 0.3, 0.7, 0.99)}

2.3.3.3. Xgboost regression. Theory: While random forest uses bagging to capture the variance of the dataset, XGBoost is based on another popular machine learning technique for data modelling, known as gradient boosting. As described earlier, in random forest, the decision trees are built parallely on different splits of the dataset. However, in boosting-based algorithms like XGBoost, the decision trees are built sequentially to minimize the residual error modelled by the previous tree.

Accodrding to Chen et al. [29], if $\widehat{y_i^{(t)}}$ is the prediction of i^{th} instance at t^{th} iteration of a XGBoost regression model $f(x)$ the objective function of XGBoost regressor becomes:

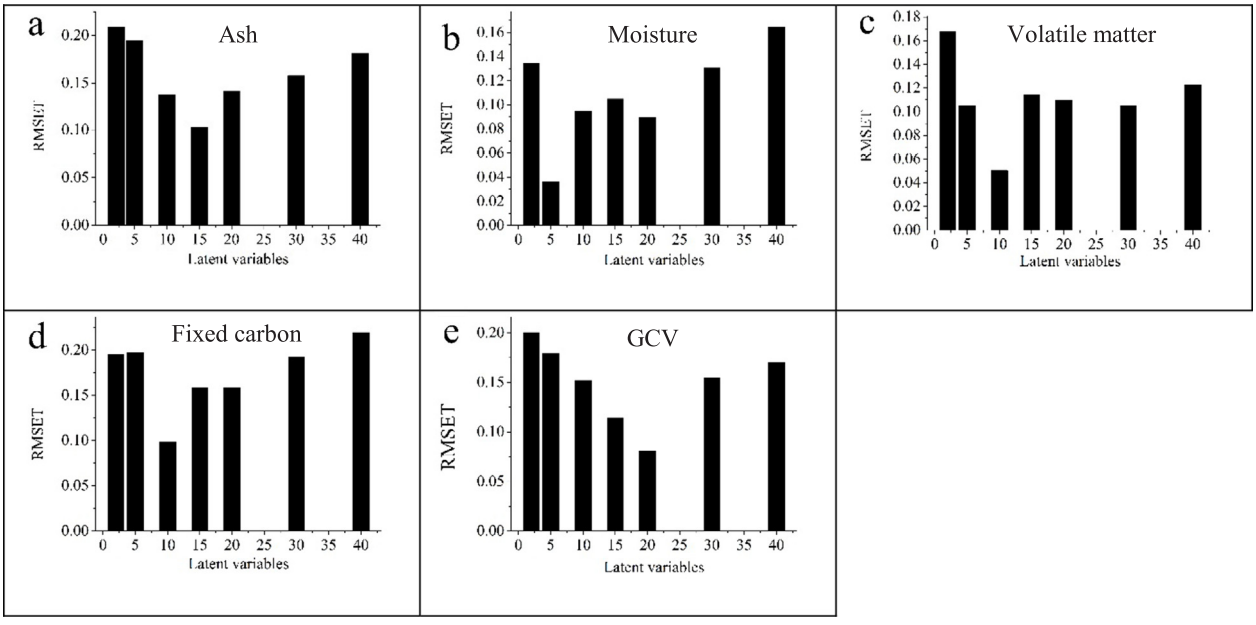

Fig. 4. Variation of RMSET with number of latent variables for different coal properties–(a) ash, (b) moisture, (c) volatile matter, (d) fixed carbon and (e) GCV.

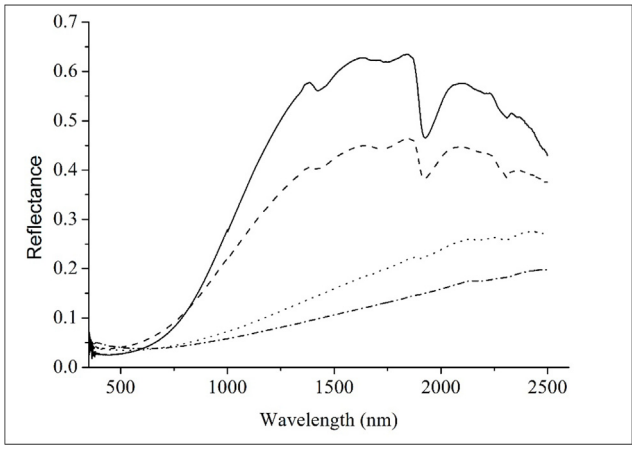

Fig. 5. Vis-NIR-SWIR diffuse reflectance spectra of coal of different rank. At the top a. lignite, followed by b. subbituminous, c. bituminos and d. semi-anthracite at the bottom.

$$L^{(t)} = l(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)) + \Omega \tag{5}$$

where $f_t(x_i)$ is the t^{th} iteration decision tree, Ω is a regularization term which prevents individual decision tree from over-complicating by avoiding over fitting and $l(y, \hat{y})$ is defined as:

$$l(y, \hat{y}) = (y - \hat{y})^2 \tag{6}$$

The XGBoost regression works in the following manner [30]:

a) An initial model $f_0(x)$ is defined to predict the target variable y. The residual of this model is defined by $\{y - f_0(x)\}$.
b) A new model $f_1(x)$ is modelled to fit the residuals from previous step.
c) After repeating the steps *a* and *b* for convergence, the model $f(x)$ is selected as a linear combination of $f_t(x)$ when $t \in T$, T being the number of iterations of steps *a* and *b*.

This 'boosted model' $f(x)$ is the resultant XGBoost regressor which minimizes the error of the predictions sequentially, resulting in a model better equipped to capture the variance of the dataset. In the present study, after splitting the dataset into training and test sets, the dataset containing 2101 features were transformed into a latent space using PCA and 20 PC were taken for predicting coal properties. The XGBoost algorithm was optimised using grid-search on a 5-fold cross-validation setup of training dataset; the hyperparameters of the model are shown below.

param_grid = {'max_depth': (2, 5), 'n_estimators': (2000, 3500)}

To prevent over fitting, weak learners were used for building the individual decision trees with *max_depth* value being 2–5. The weak learners ensure that XGBoost does not over-fit the training dataset, as the learners are individually not capable of learning very complex features. The model best performing on 5-fold cross-validation setup was used for prediction of coal properties from the test data.

2.3.4. Data modelling metrics

2.3.4.1. Model optimisation. All the models (PLS, RF, and XGB) mentioned above were trained on the dataset to learn the relationships between spectral [Log (1/R)] data and coal properties. In each model, the training set was split into 5 parts; four of them were used to train the models and one was used to validate the same. Grid search was used for selecting the hyperparameters using 5-fold cross-validation. The model which performed best on validation split was considered for estimation of coal properties.

2.3.4.2. Loss function and evaluation metrics. The models were trained using RMSE (Eqn-7) loss function and the test-set error was measured using RMSET (Eqn-8). Performance of the models were measured based on RMSET and percentage error [1,12-14]. Since a data set with lower mean value always gives smaller RMSE and vice versa, biasness may arise in the model output. The percentage error (Eqn-9) was therefore used here to remove the biasness.

$$RMSE = \sqrt{\frac{\sum_{i \in trainingset} (y - \hat{y})^2}{N_{tr}}} \tag{7}$$

$$RMSET = \sqrt{\frac{\sum_{i \in testset} (y - \hat{y})^2}{N_{te}}} \tag{8}$$

where, N_{tr} is number of training samples and N_{te} is number of test data and $y, \hat{y}$ are respectively the measured and predicted values.

$$\text{PercentageError} = 100 * \frac{RMSET}{Meanvalue} \tag{9}$$

Table 3
RMSET and percentage of error for estimation of coal properties using different statistical models.

	Predictive model	Ash%		Moisture%		Volatile matter%		Fixed carbon		GCV (MJ/kg)	
		RMSET	Percentage of error	RMSET	Percentage of error	RMSET	Percentage of error	RMSET	Percentage of error	RMSET	Percentage of error
Type-1	PLS	0.1282	36.76	0.0477	39.88	0.0788	18.61	0.0986	23.29	0.0811	13.58
	RF	0.1194	34.23	0.0372	31.1	0.0667	15.75	0.0931	21.99	0.1066	17.86
	XGBoost	0.1171	33.58	0.0452	37.79	0.0748	17.67	0.1080	25.51	0.1121	18.78
Type-2	PLS	0.1168	33.49	0.0500	41.8	0.0508	12	0.1287	22.13	0.1122	18.79
	RF	0.1079	30.94	0.0352	29.43	0.0573	13.53	0.1136	19.53	0.1018	17.05
	XGBoost	0.0925	26.52	0.0506	42.3	0.0622	14.69	0.0955	16.42	0.0890	14.91
Type-3	PLS	0.1024	29.36	0.0537	44.89	***0.0505***	***11.93***	0.1014	17.43	0.0901	15.09
	RF	0.0976	27.98	***0.0163***	***13.66***	0.0508	12.01	***0.0850***	***14.61***	0.1065	17.84
	XGBoost	***0.0898***	***25.75***	0.0429	35.86	0.0599	14.15	0.0969	16.66	***0.0733***	***12.28***
Type-4	PLS	0.1175	33.69	0.0378	31.6	0.0665	15.71	0.0982	16.88	0.1040	17.42
	RF	0.1105	31.68	0.0237	19.81	0.0586	13.84	0.0983	16.9	0.0992	16.62
	XGBoost	0.0940	26.95	0.0216	18.06	0.0563	13.3	0.0972	16.71	0.0947	15.86
Type-5	PLS	0.0907	26.01	0.0286	23.91	0.0541	12.78	0.1292	22.21	0.0930	15.58
	RF	0.1031	29.56	0.0290	24.24	0.0621	14.67	0.0936	16.09	0.0951	15.93
	XGBoost	0.0912	26.15	0.0377	31.52	0.0511	12.07	0.0981	16.87	0.0871	14.59
Type-6	PLS	0.1080	30.97	0.0362	30.26	0.0565	13.34	0.1127	19.38	0.0903	15.13
	RF	0.0997	28.59	0.0515	43.01	0.0609	14.39	0.1086	18.67	0.1027	17.21
	XGBoost	0.0946	27.12	0.0325	27.17	0.0579	13.67	0.0960	16.51	0.0918	15.38

3. Results and discussion

3.1. Spectral properties of coal

Fig. 5 shows the reflectance spectra of coal of four broad ranks viz., lignite, subbituminous, bituminous and semi anthracite. The absorption of coal in Vis-NIR-SWIR range is due to absorbance and or overtones of several functional groups (C–H, O–H, N–H mainly) [16]. The lower rank coal shows a board absorption in the visible range and then a relatively sharp increase in the reflectance value in the NIR range. In SWIR range, lower rank coal shows a very high reflectance value. On the other hand, the bituminous and semi anthracite coal exhibits relatively lower reflectance in the NIR-SWIR range. Coal (lignite and subbituminous) shows absorption band near 1400 and 1900 nm which is mainly attributed to the free, bonded and/or absorbed water [31-33]. Weaker absorption band near 1700 and 2300 nm is observed. First order overtones and combination of aliphatic C–H stretching causes weak absorption near 1700 nm [34]. While, absorption near 2300 nm is attributed to the presence of organic combination and overtone bands and/or clay-OH absorption band [35]. Overall, the reflectance value of coal and the intensity of absorption band decreases with increase in coal rank. The reflectance spectra of higher rank coal are more flatter and absorption featureless than the lower rank coal. This is due to the fact that with increase in coal rank the degree of aromatization increases which results in shift in absorption edge of aromatic molecule to the higher wavelength [36].

3.2. Comparison of prediction performance of different types of spectral pre-processing methods

Table 3 summarizes the RMSET and percentage of error for estimation of coal properties using different models and Fig. 6 shows the RMSET of each model from best to worst. It is evident that the pre-processed spectra (Type-2 to Type-6) exhibit better prediction performance than the raw absorption spectra (Type-1) by reducing the effect of non-linearities. It signifies that spectral pre-processing can improve the prediction ability of the model. Type-2 however shows second lowest prediction ability, possibly due to the adverse effect of particle scattering, as no particle scattering pre-treatments were applied to it. In all other spectral types (Type-3 to Type-6) different particle scatter correction were applied to Type-2 instead of Type-1 because of the better performance of Type-2 over Type-1. Savitzsky-Golay/normalization (Type-3) provides best result over Savitzsky-Golay/baseline correction (Type-4), Savitzsky-Golay/MSC (Type-5) and Savitzsky-Golay/normalisation/baseline correction/MSC (Type-6). This is possibly because normalization of the spectra makes it independent of sample weight and thereby improves the correlation between the spectra and coal properties [14]. However, error values of Type-6 indicate that overdo of pre-processing could reduce the model accuracy as it causes loss of some valuable chemical information from the spectra [20].

3.3. Comparison of prediction performance of different regression models

Ash: XGBoost regression performs best for estimation of ash content of coal with RMSET of 0.09 and percentage of error 25.57. Ash content shows higher error percentage as compared to the other four coal properties. This might be because ash represents the mineral matter in coal and in general, minerals show lower sensitivity in infrared region [37].

Moisture: RF regression predicts the moisture content of the coal samples better than other models. All types of coal shows an absorption band near 1900 nm and coal of lower rank shows an additional absorption band near 1400 nm. Absorption near 1400 and 1900 nm are attributed to the presence of water molecule [31-33]. VNIR spectra can readily detect OH molecule. Thus, the moisture content of coal is estimated well with RMSET 0.0163 and percentage error 13.66.

VM: Volatile matter represent the organic matter in coal, mainly composed of different types of hydro-carbons and this could absorb more infrared wave [12]. Thus, it could be estimated well from the spectral responses and PLS regression algorithm performs best for predicting volatile matter with RMSET 0.0505 and percentage of error of 11.93.

FC: Fixed carbon is the ash free non-volatile matter in coal and it can be estimated well from the reflectance spectroscopy. RF regression gives lowest error value for fixed carbon content in coal with a RMSET of 0.08 and percentage of error of 14.61 respectively.

GCV: GCV is largely dependent on the fixed carbon content of the coal. Therefore, it could also be estimated well from spectral data and XGBoost gives lowest error compared to the other regression models. The lowest RMSET and percentage error for GCV are 0.07 and 12.28 respectively.

The measured versus predicted values for each coal properties obtained from the best performing model is shown in Fig. 7(a–e) with their respective residuals plots. The random patters of the residuals indicate a good fit of the models to the dataset. In case of moisture content (Fig. 7b) it can be seen that both the measured vs. predicted plot and the residual plot is clustered in lower value. This is because the higher rank coal (i.e., high volatile A bituminous to semi-anthracite) which comprises 89% of the data, has < 2% moisture content whereas

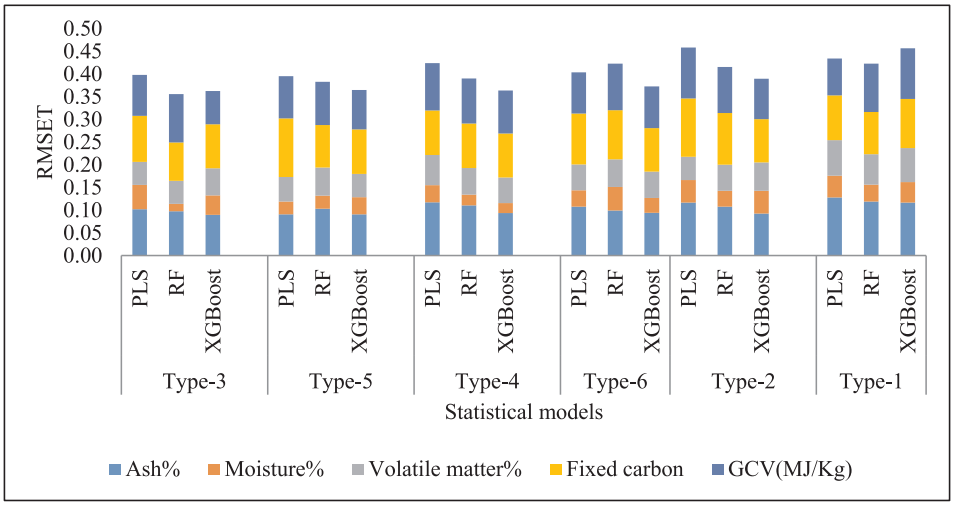

Fig. 6. Root-mean-square error of different models for different type of absorption spectra and different coal properties.

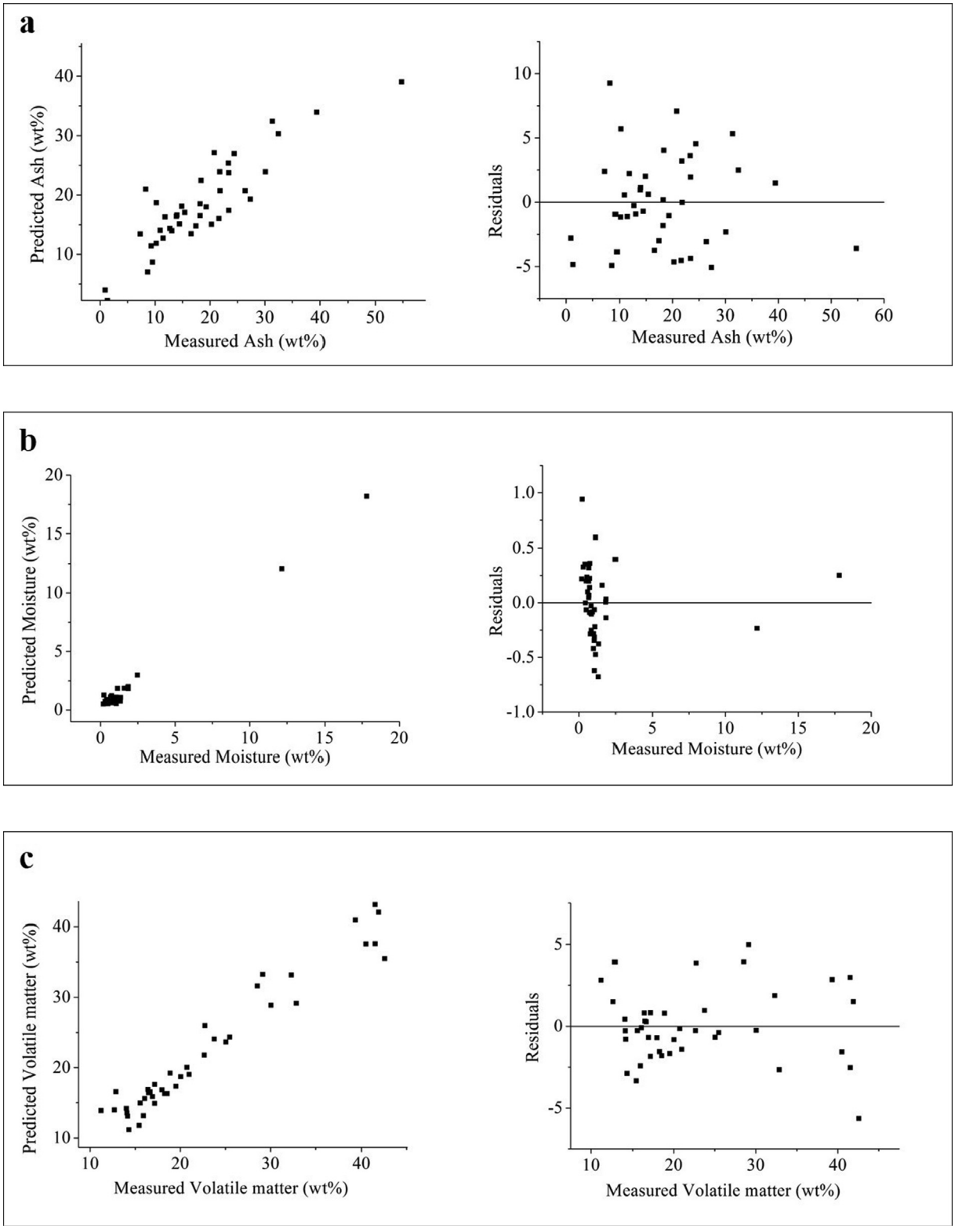

Fig. 7. Predicted versus measured coal quality plots obtained from the best performing models with their corresponding residuals. a) ash content, b) moisture content, c) volatile content, d) fixed carbon, and e) GCV.

the lower rank coal (lignite B to subbituminous C) contains 8.33 to 33.07% moisture. Thus, the data itself is not uniformly distributed and resulted in unbalanced X axis residual plot.

4. Conclusion

Rapid determination of some essential coal properties (ash, moisture, volatile matter, fixed carbon, and GCV) based on its Vis-NIR-SWIR spectrum can be done with a satisfactory accuracy level. This spectral analysis of coal properties is advantageous over the standard

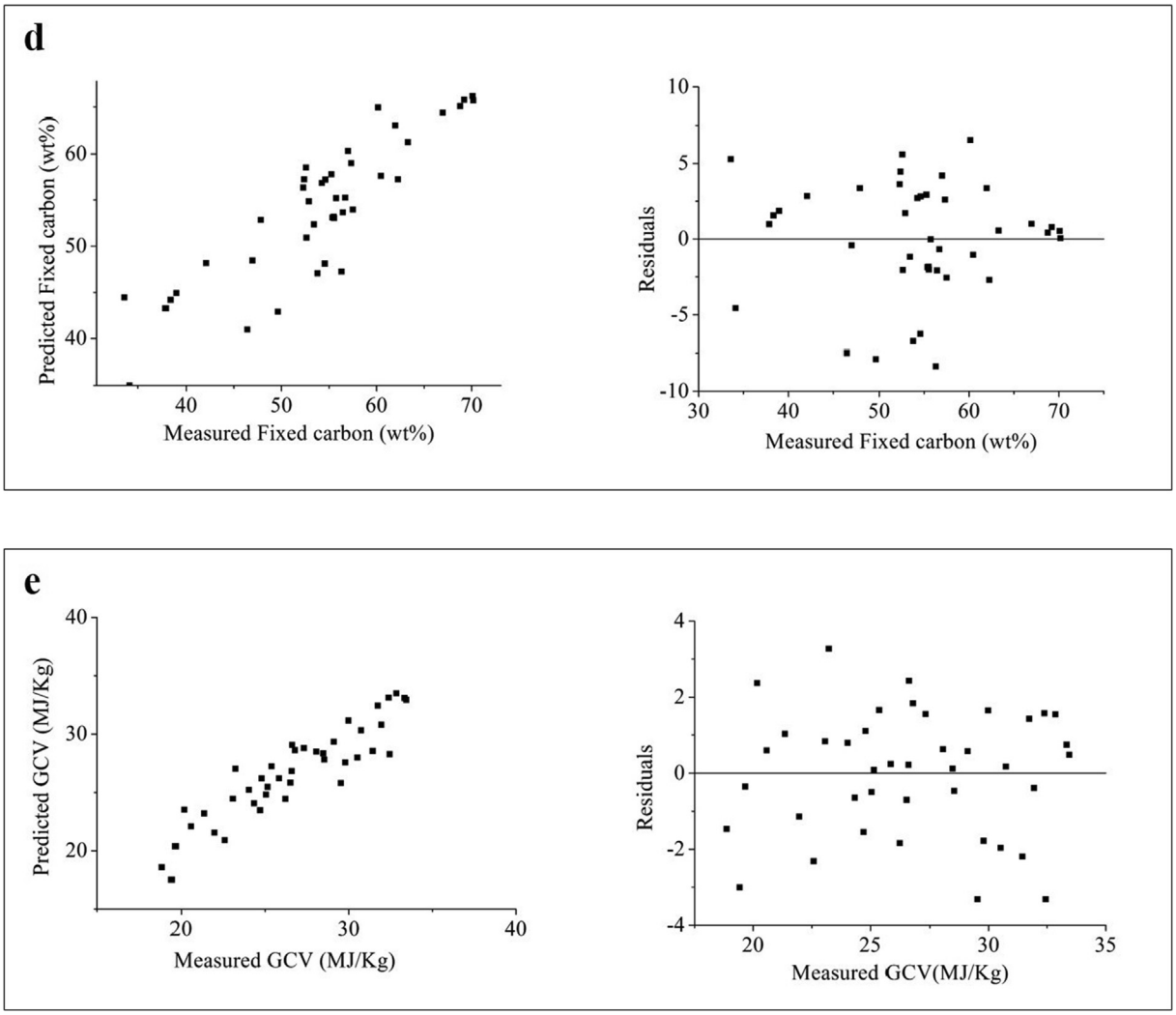

Fig. 7. (*continued*)

analytical methods, because the former can provide the physicochemical information of coal almost 20 times faster at a cost roughly 1/10th of the latter. It is evident that the application of different spectral pretreatments improves the model accuracy. However, it is imperative to optimize the spectral pre-processing methods since over application can increase the model complexity and decrease the prediction accuracy. Savitzky-Golay smoothing followed by normalization (Type-3) has been found to provide the best result. PLS, RF, and XGboost based regression models were used to estimate the coal properties. While the PLS model gives higher accuracy for estimation of volatile matter, RF provides the lowest error value for moisture and fixed carbon contents; XGBoost based model gives the best result for ash content and GCV of coal. In the present investigation, DRS has been found to predict the volatile matter and GCV with high accuracy followed by moisture, fixed carbon and ash content. In future, we will consider a homogenous cluster of coal to improve the performance accuracy of the predictive models.

It should however, be mentioned that the present study has been carried out in laboratory conditions. Practical service conditions are more severe, and some additional factors such as, bulk quantity of coal, variation in coal grades and sizes, environmental conditions, etc. could influence the prediction accuracy of the method. Therefore, all these parameters should be considered before implementation of this method into a larger scale. In future, the authors intend to carry out an in-line investigation of coal quality parameters by placing the sensor over a conveyor belt to assess the applicability of the method.

Declaration of Competing Interest

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.

Acknowledgement

The authors are thankful to the Big Data Initiatives Division, Department of Science and Technology, India for the financial support (Grant Number: BDIB/01/23/2014-HSRS). The authors are also thankful to Mr. Nukala Nagarjuna Reddy from Department of Agriculture and Food Engineering, IIT Kharagpur for his help in carrying out the spectrometer readings.

References

[1] Kim DW, Lee JM, Kim JS. Application of near infrared diffuse reflectance spectroscopy for on-line measurement of coal properties. Korean J Chem Eng 2009;26:489–95. https://doi.org/10.1007/s11814-009-0083-0.
[2] Friedel RA, Retcofsky HL JAQ USBM-640.pdf 1967.
[3] Cannon CG, Sutherland GBBM. The infra-red absorption spectra of coals and coal extracts. Trans Faraday Soc 1945;41:279–88. https://doi.org/10.1039/tf9454100279.
[4] Mastalerz M, Bustin RM. Electron microprobe and micro-FTIR analyses applied to maceral chemistry. Int J Coal Geol 1993;24:333–45. https://doi.org/10.1016/0166-5162(93)90018-6.
[5] Iglesias MJ, Del Río JC, Laggoun-Défarge F, Cuesta MJ, Suárez-Ruiz I. Control of the

chemical structure of perhydrous coals; FTIR and Py-GC/MS investigation. J Anal Appl Pyrol 2002;62:1–34. https://doi.org/10.1016/S0165-2370(00)00209-6.
[6] Bona MT, Andrés JM. Reflection and transmission mid-infrared spectroscopy for rapid determination of coal properties by multivariate analysis. Talanta 2008;74:998–1007. https://doi.org/10.1016/j.talanta.2007.08.016.
[7] Chen Y, Mastalerz M, Schimmelmann A. Characterization of chemical functional groups in macerals across different coal ranks via micro-FTIR spectroscopy. Int J Coal Geol 2012;104:22–33. https://doi.org/10.1016/j.coal.2012.09.001.
[8] He X, Liu X, Nie B, Song D. FTIR and Raman spectroscopy characterization of functional groups in various rank coals. Fuel 2017;206:555–63. https://doi.org/10.1016/j.fuel.2017.05.101.
[9] Yang CQ, Simms JR. Comparison of photoacoustic, diffuse reflectance and transmission infrared spectroscopy for the study of carbon fibres. Fuel 1995;74:543–8. https://doi.org/10.1016/0016-2361(95)98357-K.
[10] Swinehart DF. The Beer-Lambert law. J Chem Educ 1962;39:333–5. https://doi.org/10.1021/ed039p333.
[11] Tuan B, Xiao D, Mao Y, He D. Coal analysis based on visible-infrared spectroscopy and a deep neural network. Infrared Phys Technol 2018;93:34–40. https://doi.org/10.1016/j.infrared.2018.07.013.
[12] Wang Y, Yang M, Wei G, Hu R, Luo Z, Li G. Improved PLS regression based on SVM classification for rapid analysis of coal properties by near-infrared reflectance spectroscopy. Sensors Actuators, B Chem 2014;193:723–9. https://doi.org/10.1016/j.snb.2013.12.028.
[13] Bona MT, Andrés JM. Coal analysis by diffuse reflectance near-infrared spectroscopy: hierarchical cluster and linear discriminant analysis. Talanta 2007;72:1423–31. https://doi.org/10.1016/j.talanta.2007.01.050.
[14] Andrés JM, Bona MT. ASTM clustering for improving coal analysis by near-infrared spectroscopy. Talanta 2006;70:711–9. https://doi.org/10.1016/j.talanta.2006.05.034.
[15] Begum N, Chakravarty D, Das BS. Estimation of gross calorific value of bituminous coal using various coal properties and reflectance spectra estimation of gross calorific value of bituminous coal using various coal properties and reflectance spectra. Int J Coal Prep Util 2019:1–7. https://doi.org/10.1080/19392699.2019.1621301.
[16] Andrés JM, Bona MT. Analysis of coal by diffuse reflectance near-infrared spectroscopy. Anal Chim Acta 2005;535:123–32. https://doi.org/10.1016/j.aca.2004.12.007.
[17] Carlos E, Alciaturi ME, Escobar VR. Prediction of coal properties by derivative DRIFT spectroscopy. Fuel 1996;75:491–9. https://doi.org/10.1016/0016-2361(95)00246-4.
[18] Cloutis EA. Quantitative characterization of coal properties using bidirectional diffuse reflectance spectroscopy. Fuel 2003;82:2239–54. https://doi.org/10.1016/S0016-2361(03)00209-6.
[19] Gobrecht A, Bendoula R, Roger JM, Bellon-Maurel V. Combining linear polarization spectroscopy and the representative layer theory to measure the beer-lambert law absorbance of highly scattering materials. Anal Chim Acta 2015;853:486–94. https://doi.org/10.1016/j.aca.2014.10.014.
[20] Van Den BF, Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. Trends Anal Chem 2009;28. https://doi.org/10.1016/j.trac.2009.07.007.
[21] Blank_map_of_Hidalgo, (n.d.). http://www.websbages.com/blank-map-of-india-pdf/. [accessed 13 March 2019].
[22] Standard a. D3172. Standard Practice for Proximate Analysis of Coal and Coke. Annu B ASTM Stand 2002:1–2. https://doi.org/10.1520/D3172-07A.2.
[23] Danner M, Locherer M, Hank T, Richter K. Spectral Sampling with the ASD FIELDSPEC 4. EnMAP Consort – GFZ Data Serv 2015. http://gfzpublic.gfz-potsdam.de/pubman/item/escidoc:1388298/component/escidoc:1388299/EnMAP_FieldGuide_ASD_2015_008.pdf [accessed 25 January 2017].
[24] Christy AA, Dahl B, Kvalheim OM. Structural features of resins, asphaltenes and kerogen studied by diffuse reflectance infrared spectroscopy. Fuel 1989;68:430–5. https://doi.org/10.1016/0016-2361(89)90263-9.
[25] Standard Test Method for Gross Calorific Value of Coal and Coke 1. October 2003;14:1–11. https://doi.org/10.1520/D5865-13.2.
[26] Garcia F. Tests to identify outliers in data series. Pontif Cathol Univ Rio Janeiro. 2012:1–16. http://habcam.whoi.edu/HabCamData/HAB/processed/OutlierMethods_external.pdf [accessed 15 January 2019].
[27] Ng KS. A Simple Explanation of Partial Least Squares 2013:1–10. https://doi.org/10.1.1.352.4447.
[28] Pereira BM, Pereira RG, Wise R, Sugrue G, Zakrison TL, Dorigatti AE, et al. The role of point-of-care ultrasound in intra-abdominal hypertension management. J Mach Learn Res 2017;49:373–81. https://doi.org/10.5603/AIT.a2017.0074.
[29] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proc ACM SIGKDD Int Conf Knowl Discov Data Min 2016;13-17-August-2016:785–94. https://doi.org/10.1145/2939672.2939785.
[30] Sundaram RB D. Resource, An End-to-End Guide to Understand the Math behind XGBoost. https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/, 2018 [accessed 12 December 2018].
[31] Sgavetti M, Pompilio L, Meli S. Reflectance spectroscopy (0.3-2.5 μm) at various scales for bulk-rock identification. Geosphere 2006;2:142–60. https://doi.org/10.1130/GES00039.1.
[32] Lugassi R, Ben-Dor E, Eshel G. Reflectance spectroscopy of soils post-heating-Assessing thermal alterations in soil minerals. Geoderma 2014;213:268–79. https://doi.org/10.1016/j.geoderma.2013.08.014.
[33] Longhi I, Sgavetti M, Chiari R, Mazzoli C. Spectral analysis and classi cation of metamorphic rocks from laboratory re ectance spectra in the 0. 4–2.5 mm interval: a tool for hyperspectral data interpretation. Int J Remote Sens 2001;22:3763–82. https://doi.org/10.1080/01431160010006980.
[34] Cloutis EA. Spectral reflectance properties of hydrocarbons: remote-sensing implications. Science 1989;245:165–8. https://doi.org/10.1126/science.245.4914.165.
[35] Clark RN, King TVV, Klejwa M, Swayze GA, Vergo N. High spectral resolution reflectance spectroscopy of minerals. J Geophys Res 1990;95. https://doi.org/10.1029/jb095ib08p12653.
[36] Ito O. Diffuse reflectance spectra of coals in the UV-visible and near-IR regions. Energy Fuels 1992;6:662–5. https://doi.org/10.1021/ef00035a019.
[37] Román Y, Cabanzo R, Enrique J, Mejía-ospino E. FTIR-PAS coupled to partial least squares for prediction of ash content, volatile matter, fixed carbon and calorific value of coal. Fuel 2018;226:536–44. https://doi.org/10.1016/j.fuel.2018.04.040.
[38] Yu D, Xu M, Sui J, Liu X, Yu Y, Cao Q. Effect of coal particle size on the proximate composition and combustion properties. Thermochim Acta 2005;439(1–2):103–9. https://doi.org/10.1016/j.tca.2005.09.005.