Fuel Communications 13 (2022) 100078

Contents lists available at ScienceDirect

Fuel Communications

journal homepage: www.sciencedirect.com/journal/fuel-communications

Modeling coking coal indexes by SHAP-XGBoost: Explainable artificial intelligence method

A. Homafar [a], H. Nasiri [b,*], S.Chehreh Chelgani [c,*]

[a] Electrical and Computer Engineering Department, Semnan University, Semnan, Iran
[b] Department of Computer Engineering, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran
[c] Minerals and Metallurgical Engineering, Department of Civil, Environmental and Natural Resources Engineering, Luleå University of Technology, Luleå SE-971 87, Sweden

ARTICLE INFO

Keywords:
Free swelling index
Gieseler plastometer
Coal
Explainable artificial intelligence
Machine learning
Modeling

ABSTRACT

Coking coal is still on the list of critical raw materials in many countries since it is the main element integrated into the blast furnace. While the energy consumption and steelmaking efficiency in the furnace depends on the coke quality, understanding and modeling coking indexes based on their coal parent properties would be a substantial approach for the steelmaking industry. As an innovative approach, this short comminucation has been considered explainable artificial intelligence (XAI) for modeling coal coking indexes (Free Swelling index "FSI" and maximum fluidity "Log (MF)"). XAIs can convert black-box models into human basis systems and develop a significant learning performance and estimation accuracy. SHapley Additive exPlanations (SHAP), as one of the most recently developed XAI models in combination with eXtreme gradient boosting (XGBoost), were used to model coal samples from Illinois, USA. For the first time, FSI and Log (MF) treat as ordinal variables for modeling. Modeling outcomes relieved that SHAP-XGBoost could accurately show interdependency between features, demonstrate the magnitude of their multi relationships, rank them based on their importance, and predict the coking index quite accurately compared with conventional machine learning methods (random forest and support vector regression). These significant results would be opened a new window by applying XAI tools for controlling and modeling complex systems in the energy and fuel sectors.

1. Introduction

Although the steelmaking industry is one of the largest industrial sources of CO_2 emission (~27% of global CO_2 emissions), coking coal is extensively still used in the steel and ironmaking industry as an unsubstitutable ingredient [1–3]. Approximately 0.7 tons of coking coal has to be used for each one-ton steel production. Since steel demand has expressively grown during the last few decades, coking coal has been on many countries' critical raw material list [4]. The coal impurities can be mainly called as "Ash" markedly affect its coking ability [5], and decrease the coke productivity in the blast furnace [6]. It was predicted that by increasing each 1% of coal impurities, the coke productivity decreased by 2–3 wt% [7,8]. Free swelling index (FSI) (ASTM D720) [9] and maximum fluidity "Log (MF)" (gieseler plastometer) (ASTM D2639) [10] are the most known standard coking indexes, which have been widely used for coal coking quality assessments. The FSI as a qualitative factor classified coal samples into three categories: 0–2 (non-coking), <2–4 (medium), and <4–9 (the coking quality increases by rising the FSI). Gieseler plastometer could measure coal plasticity and determine its coke ability based on MF. Hadavandi & Chelgani (2019) indicated that there is a moderate positive correlation between log (MF) and FSI test results (by increasing FSI, log (MF) somehow is also increased) [11]. However, it was documented that various problems such as different particle size distribution of coal samples, frequent calibrating systems, heating rate, weathered samples, and different oxidation variability would limit the reproducibility and representability sociated with these coking index determinations [12]. These limitations would be prioritizing the modeling of coking indexes.

It was well understood that coal rank parameters (moisture, volatile matter, carbon …) could affect their representative coking capability and significantly change it [13]. These effects could be complicated while the heterogeneous structure of coal makes several complex inter-correlations within its components [14,15]. Coal rank parameters can mainly be determined by proximate (ASTM D3172) [16] and

* Corresponding authors.
E-mail address: saeed.chelgani@ltu.se (S.Chehreh Chelgani).

https://doi.org/10.1016/j.jfueco.2022.100078

Available online 18 September 2022

ultimate (ASTM D3176) [17] analyses. Since coking coal producers have to present both coal rank parameters and thermoplastic properties of their products, some investigations have been conducted to explore the possible models for assessing relationships between coal rank properties and their representative coking indexes. These models would be a key principle for the steelmaking industries to provide their desired coal blending, generating a high coke quality as a reductant agent and permeable support.

Therefore, some statistical modeling approaches have been conducted to tackle difficulties associated with coking index determination. Since FSI and log (MF) are qualitative factors, it was documented that common multivariable regression models cannot accurately model them. Thus, random forest [18], support vector regression [11], feed-forward artificial neural network [19,20], neuro-fuzzy inference systems [21], etc., as artificial intelligence (AI) methods (black box models) have been used for modeling them [22]. However, on the one hand, these black box AI and machine learning (ML) methods generally do not provide any insight into the magnitude of relationships among input data [23]. On the other hand, the coking index values of these models have been treated as nominal labels, whereas a closer study of the coal data revealed the fact that the output class is ordinal, and by utilizing the conventional AI and ML methods, some important information would be lost that could potentially improve the model predictability. Therefore, it would be essential to consider a tool highlighting the individual and multivariable correlations of model features for these complicated indexes.

Explainable Artificial Intelligence (XAI) is a recently developed machine learning that could address these shortcomings [24]. XAI models visualize relationships and their magnitude [25] and convert them into interpretable systems [26]. As the most recent XAI development, SHapley Additive exPlanations (SHAP) [27]can provide insight into how black box AI and ML systems make estimations. As an inventive strategy based on the game theory, SHAP assists data scientists with the model development procedure by explaining the decision-making process of the black box models [28]. SHAP can compute the contribution of each feature to the model's output using Shapley values [29], highlight their magnitude, and rank features based on their importance [30].

As an innovative approach, this study will use SHAP to explore interdependencies between various coal properties and their representative coking indexes through their modeling using eXtreme gradient boosting (XGBoost). XGBoost is one of the most recently developed ML models with several advantages over conventional AI and ML models. XGBoost is particularly flexible, can parallel process various learning scenarios, supports regularization, and handles missing data. For the first time, this work is going to examine the SHAP-XGBoost system for modeling coking indexes. As a comparative study, conventional ML models such as Random forest (RF) and support vector regression (SVR) were considered to evaluate the suggested system capability. The outcomes of this work would be potentially suggested the application of SHAP-XGBoost as a powerful AI-based model for online and offline modeling of complex problems within coal and energy processing systems (such as modeling of Hardgrove grindability index (HGI), Gross Calorific value (GCV), vitrinite maximum reflectance (R_{max}), etc.). The detailed list of abbreviations and acronyms used in the paper are shown in Table 1.

Table 1.
List of abbreviations and acronyms used in the paper.

Abbreviation	Definition	Abbreviation	Definition
AI	Artificial Intelligence	ML	Machine Learning
DT	Decision Tree	RF	Random Forest
EML	Ensemble Machine Learning	SHAP	SHapley Additive exPlanations
FSI	Free Swelling Index	SVR	Support Vector Regression
GBDT	Gradient Boosted Decision Tree	XAI	Explainable Artificial Intelligence
Log (MF)	Maximum Fluidity	XGBoost	Extreme Gradient Boosting

2. Materials and methods

2.1. Dataset

Generally, a large database requires constructing a comprehensive soft computing model dealing with a complex problem, which may cause a severe challenge through the computation (challenges like; Lack of knowledge Professionals, Lack of proper understanding of Massive Data, Integrating Data from a Spread of Sources, Confusion while Big Data Tool selection, etc.) [31]. However, the most recent development in ML systems provided this opportunity to use datasets (instead of databases) to generate predictive models [32]. In this investigation for developing an accurate XAI model, a high-dimensional dataset was selected, covering a wide variation of coal and coking properties. A dataset with more than 100 records from Illinois was considered (Table 2) to construct an XAI for FSI and MF prediction. The modeling sequence was based on the diagram illustrated in Fig. 1. All the coal characteristics were determined based on ASTM procedures (ASTM D3172: proximate; ASTM D3176: ultimate; ASTM D720: FSI; ASTM D2639: Gieseler plastometer). Regarding ASTM D3172 and ASTM D3176 for coal proximate and ultimate analyses (respectively), the amount of fixed carbon and oxygen, which may incorporate the bias of other analyzed parts, did not consider as model input features (fixed carbon% = 100 - (moisture + volatile matter + ash), and oxygen% = 100-(carbon + hydrogen + nitrogen + total sulfur)).

2.2. Methodology

2.2.1. SHapley additive exPlanations (SHAP)

The SHapley Additive exPlanations (SHAP) technique facilitates the interpretation of model results by providing a uniform approach [33]. The SHAP value quantifies the effect of each component on model outputs, both for individual observations and the whole dataset. SHAP has an additive characteristic to ensure that the aggregate of all relevant measurements and baseline values adds up to the final output [34]. Linear addition of the input features produces the model's output derived from game theory [35]. The "Shapley value" describes how much of a contribution each characteristic makes [36]. Even the most complicated models may be understood using SHAP's methodology for understanding model predictions [37]. Even though numerous ML-based studies in solid materials have achieved great accuracy in predicting their targets, little attention is paid to the ML models' interpretability. Considerable study quantifies the relevance of features in tree-based models using the decision path, heuristic techniques, or model-agnostic approaches [38]. However, these approaches are frequently impractical and biased for Ensemble Machine Learning (EML) models, particularly those with a strong bias. In order to ensure

Table 2
The statistical description of coal samples and their representative coking indexes.

Features	Symbol	Min	Max	Mean	STD
Moisture (%)	Moist	0.50	18.20	9.35	4.43
Volatile Matter (%)	VM	27.40	48.20	40.03	3.60
Ash (%)	Ash	7.10	23.43	11.90	2.60
Carbon (%)	C	58.35	77.72	70.20	3.04
Hydrogen (%)	H	4.07	5.88	4.99	0.29
Nitrogen (%)	N	0.94	1.84	1.30	0.19
Organic Sulfur (%)	Organic S	0.37	2.82	1.59	0.59
Pyritic Sulfur (%)	Pyritic S	0.29	6.63	2.12	1.01
Sulfate Sulfur (%)	Sulfate S	0.01	0.40	0.05	0.06
Free Swelling Index	FSI	1.00	9.00	4.87	1.56
Maximum Fluidity	Log MF	0.00	4.45	1.87	1.19

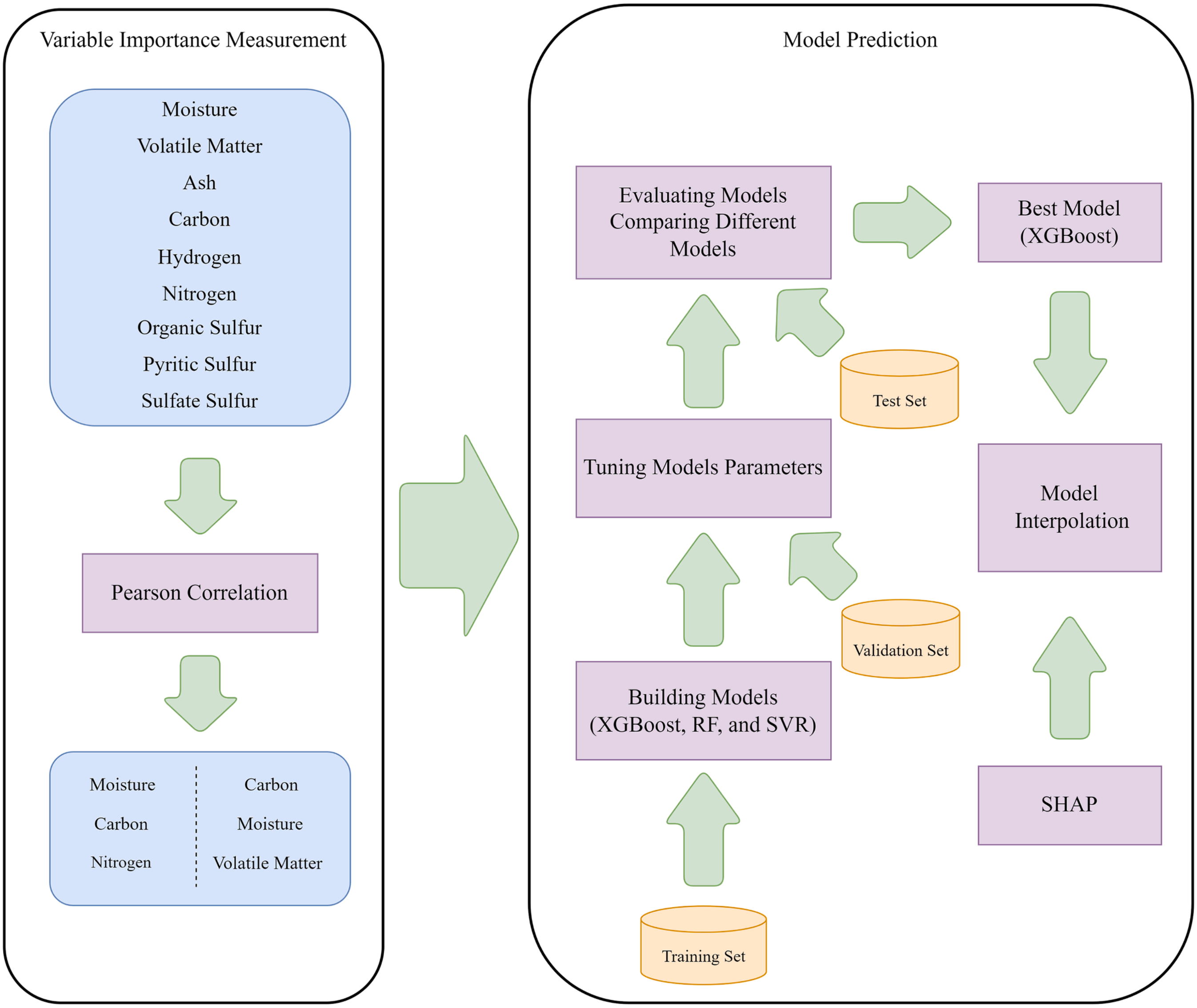

Fig. 1. Modeling sequence for coking coal indexes.

the interpretability of a machine learning model, the output is stated as the linear sum of the model's input features multiplied by the appropriate SHAP values (Eq. (1)).

$$f(x) = \varphi_0 + \sum_{i=1}^{N} \varphi_i X_i' \tag{1}$$

where f denotes the mapping function represented by the machine learning model; N represents the number of input features; φ_0 is the average of all predictions; φ_i is the SHAP value for the i *th* feature; and X_i' denotes the coalition vector for the i th component, which can be calculated from the original input X_i using a mapping function expressed as $X_i = h_x(X_i')$ [39]. Based on hypotheses such as efficiency, dummy, additive, and symmetry, the contribution of each feature (X denotes the assistance of the i th feature) could be determined by Eq. (2).

$$\phi_i = \sum_{S \subseteq N \setminus \{x_i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [V(S \cup \{x_i\}) - V(S)] \tag{2}$$

where S is the subset of N, which does not contain the feature i, and N denotes the entire set of features. The model $V(S \cup \{x_i\})$ is trained using $S \cup \{x_i\}$, but the other model $V(S)$ is trained using S. Both models' predictions are then compared using current input from subset S [40, 41].

2.2.2. Extreme gradient boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a technique developed by Chen and Guestrin in 2016. For classification and regression tasks [42], XGBoost provides a parallel tree boosting extension to gradient boosted decision trees [43,44]. Indeed, it is an enhanced version of the well-established Gradient Boosted Decision Tree (GBDT) algorithm that overcomes its computing restrictions [45]. Nonetheless, it is distinct from the GBDT approach in a way. GBDT employs the first-order Taylor expansion, whereas the XGBoost's loss function uses the second-order Taylor expansion [46,47]. For XGBoost, a sequential Decision Tree (DT) is formed using a technique known as a sequential ensemble approach [48], also known as sequential decision tree construction. Every sample in the dataset is given a weight, determining how likely it is to be picked for further examination by a decision tree. Initially, the weight for each data point is the same, but it varies due to the statistical analysis [49]. Processing large datasets (datasets from different areas: health [50,51], social security [52], earth science [53], …) with significant characteristics and categorizations is also an everyday use for XGBoost. Additionally, this method provides practical and proficient solutions for novel optimization issues [54], particularly when efficiency and accuracy trade-offs are taken into account [55]. XGBoost's objective function is composed of the convex loss function and a regularization term, as given in Eq. (3).

$$Obj(\theta) = L(\theta) + \Omega(\theta) \tag{3}$$

where $L(\cdot)$ is the loss function and $\Omega(\theta) = \gamma T + \frac{1}{2}\lambda \parallel w\|^2$ is a regularization function, controlling the model's complexity [56]. In the

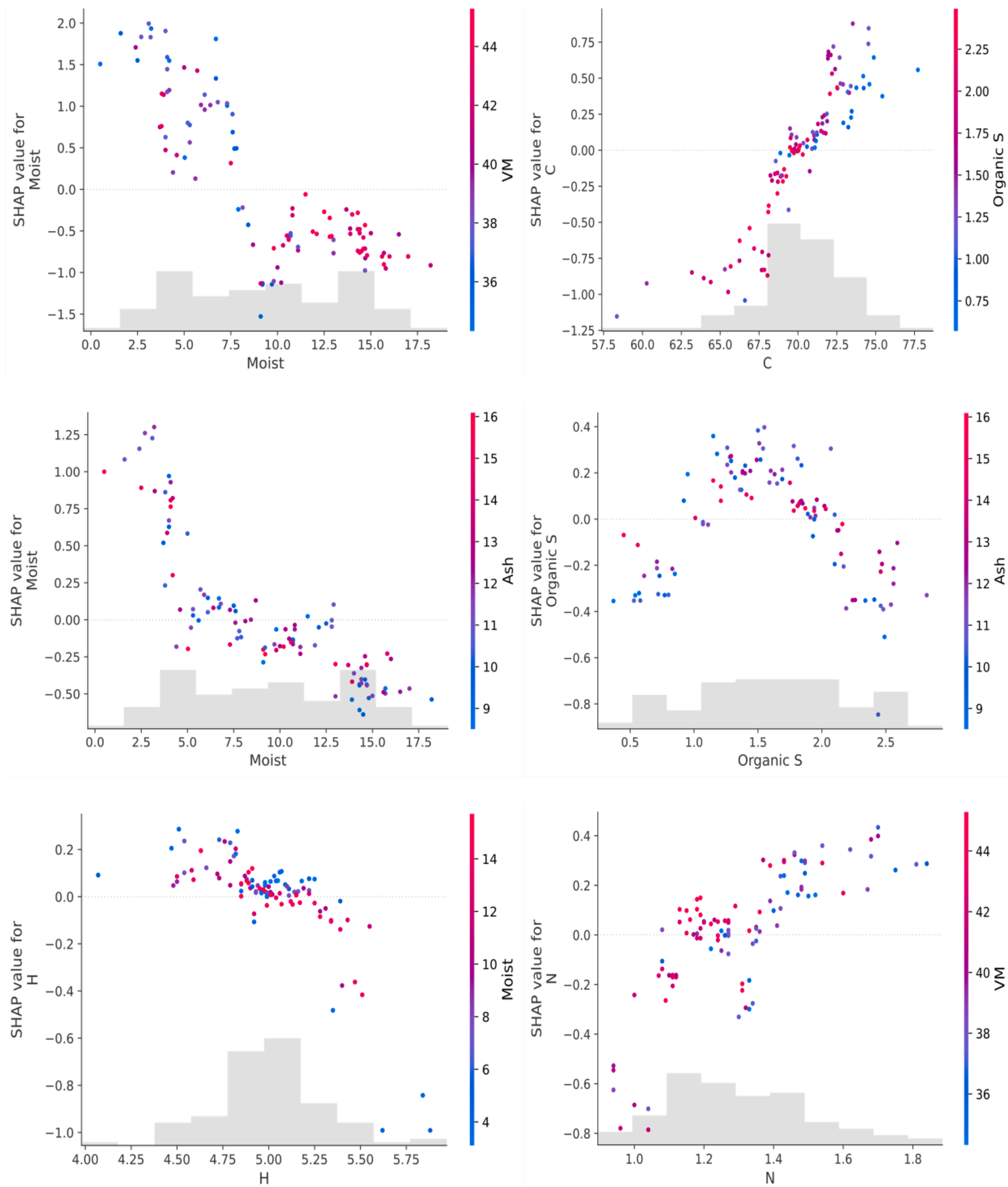

Fig. 2. SHAP feature dependence scatter plots for the XGBoost model to show the complexity of relationships between coal parameters.

regularization function, T represents the number of leaf nodes, and w is the weight of each leaf. γ and λ are regularization parameters that control the penalty associated with T and w [57].

2.2.3. Random forest

Random Forest (RF) is a nonparametric supervised machine learning approach [58]. RF is a mix of Bootstrap aggregation (Bagging) and random variable selection at each node, which is developed by Breiman [59]. RF is an advanced bagging method created based on the Decision Tree (DT) theory [60,61]. The concept behind RF is to employ bootstrap resampling to extract numerous samples from the original data and then create a DT for each bootstrap sample [62,63]. Each DT is created randomly in an RF, and the DTs are utterly independent of one another [64]. Thus, there are many different predictors in an RF, and they are all grown separately. In order to arrive at a final prediction, individual tree projections are combined through the use of averages [65]. Given an input feature vector $x = [x_1, x_2, .., x_n]^T$, the expected output of the RF model $\widehat{\tau}(x)$ could be computed according to Eq. (4).

$$\widehat{\tau}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\tau}_b(x) \tag{4}$$

where Brepresents the total number of trees and $\widehat{\tau}_b(x)$ denotes the estimate given by the b th tree [66]. To conclude, RF models are one of the most powerful supervised machine learning methods available today, as they enable the elimination of irrelevant input features based on their relative relevance [67].

2.2.4. Support vector regression

Support Vector Regression (SVR) is a nonparametric statistical technique developed in 1996 by Drucker and colleagues [42]. Regression using SVR, has been utilized effectively on various engineering challenges [68]. SVR, a revolutionary artificial intelligence system, uses a promising nonlinear kernel-based regression approach to minimize the structural risk principle in a high-dimensional feature space implemented in the SVR model. Using convex optimization methods, SVR transforms nonlinear regression problems into linear regression models [69]. The computational complexity of this approach is not reliant on the dimensions of the input space, which is one of its primary advantages [70]. Furthermore, it has a high level of accuracy in predicting outcomes and broad applicability [59]. The SVR uses Eq. (5) to solve the regression problem.

$$f(x) = \langle w\ \varphi(x)\rangle + b \tag{5}$$

where w denotes the weight of the matrix, $\varphi(x)$ represents the multidimensional space comprising the input vector x, and b is the bias [71].

3. Results and discussions

3.1. SHAP assessment

In this study, the SHAP was applied to the model built by XGBoost. SHAP analyses among proximate-ultimate analyses parameters (coal component) indicate the complexity of relationships between coal parameters (Fig. 2). Exploring multivariate relationships between these parameters and their representative coking indexes (FSI and log (MF)) showed that moisture and carbon contents have the highest effect on the coke capability of coal samples (Fig. 3). Moisture shows a significant negative and carbon a substantial positive correlation with the coking indexes (Fig. 3). Moisture is a coal rank factor since coal's rank decreases as it increases. The negative effect of coal moisture in the blast furnace and coking rate was reported in other investigations [72,73]. In general, through airless heating of coal samples (coke-making procedure), their moisture content is released, leaving a solid residue called coke; thus, high moisture content could reduce the coking rate [14]. Since coke can be considered a macro-porous carbon material, the carbon content level absolutely plays one the most important role in its structure. In general, coke strength and reactivity tremendously should depend on its isotropic carbon content. Therefore, the weighty positive correlation between coking indexes and carbon content would be obvious. These relationships are even observed in other renewable fuels [74–78]. There is a significant agreement between SHAP and Pearson correlation assessments (Fig. 4).

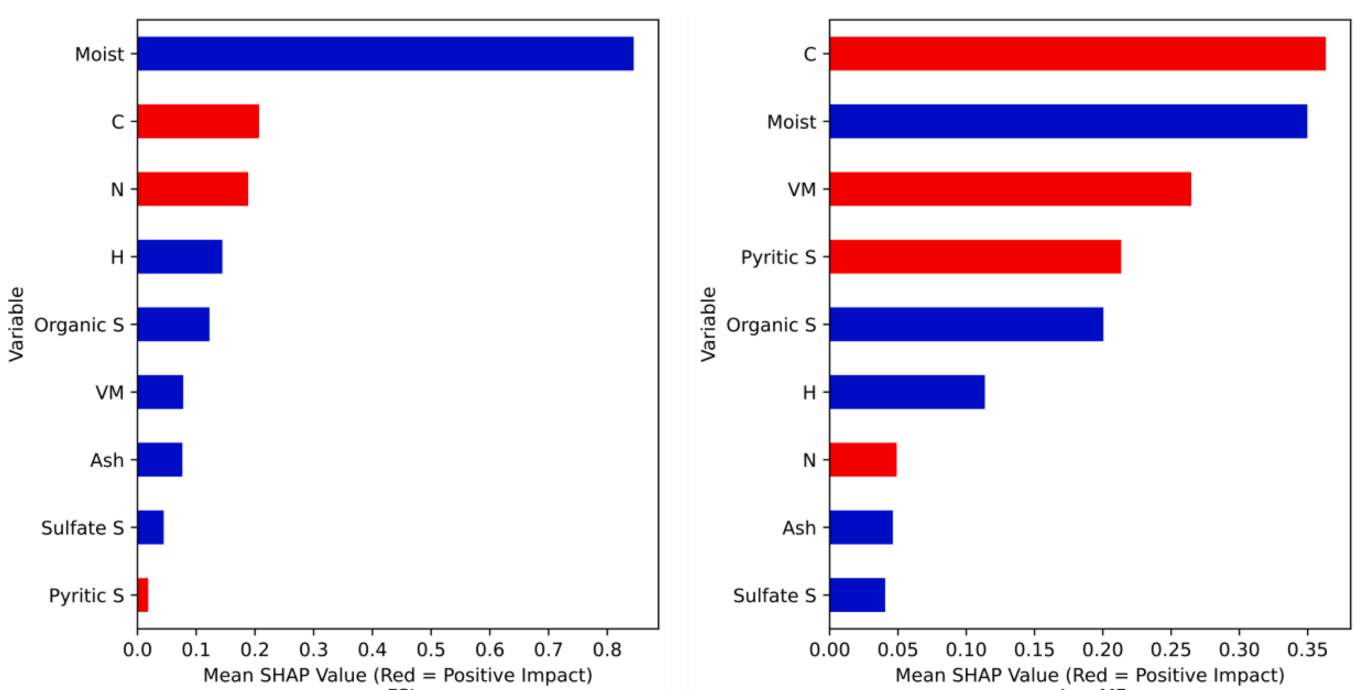

Fig. 3. SHAP feature importance of coal features on the coking indexes for the XGBoost model. Red and blue bars indicate the positive and negative impact of the features on the output.

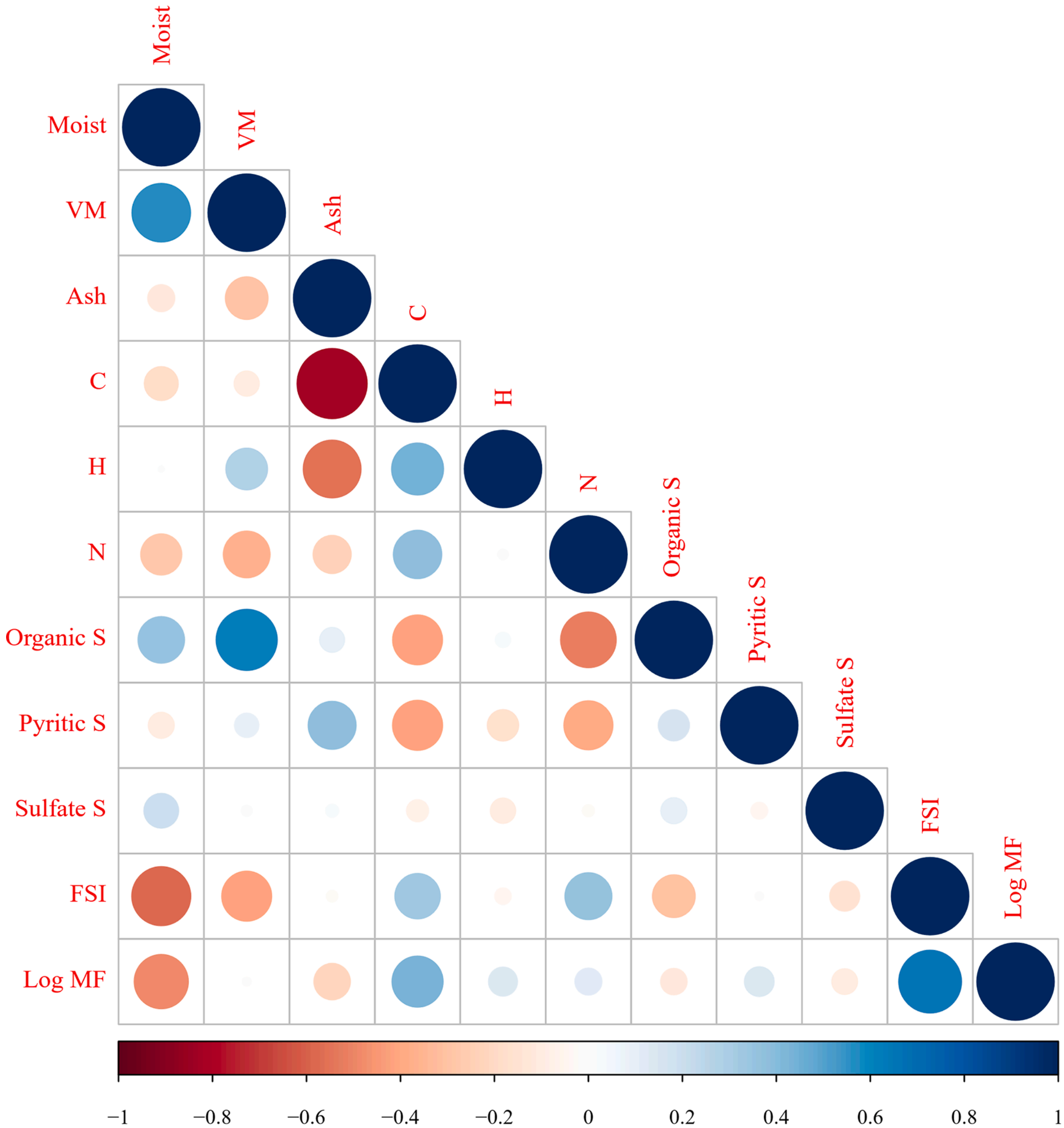

Fig. 4. Linear relationships (Pearson correlation) between coal characteristics and their representative coking indexes.

3.2. Prediction

XGBoost was considered for the perdition of coking index features. From the entire provided dataset, 80% of records were randomly used as the training set, 10% as the validation set, and the remaining 10% as the test set. The XGBoost hyperparameters were selected by try and error approach based on the Grid Search algorithm (Table 3). The XGBoost outcomes (Table 4) indicated that the coking indexes could be accurately estimated by using coal properties. The same records were considered for comparison determinations to develop RF and SVR as conventional ML models. Outcomes (Table 4) highlighted that the XGBoost algorithm could predict the coking indexes quite accurately compared to these two common AI models (Fig. 5). Moreover, to determine whether XGBoost's superiority was statistically significant, a two-tailed Welch's t-test with a significance level $\alpha = 0.05$ was conducted between XGBoost and other methods. Welch's t-test, a nonparametric univariate statistical test, is useful when two samples have unequal variances [79]. As seen in Table 4, in all comparisons, the null hypothesis is rejected based on the tests with a 95% confidence level (p-value < 0.05), giving statistically significant results.

Table 3
The XGBoost parameter settings for predicting coking indexes.

Parameter	Value (Log MF)	Value (FSI)
Base learner	Gradient boosted tree	Gradient boosted tree
Tree construction algorithm	Exact greedy	Exact greedy
Learning objective	Regression with squared loss	Regression with squared loss
Learning rate (η)	0.3128	0.4407
Lagrange multiplier (γ)	0	0
Number of gradients boosted trees	77	13
Maximum depth of trees	3	3
The minimum sum of instance weight (hessian) needed in a child	0	0
L2 regularization term on weights	1	1
The initial prediction score of all instances (global bias)	0.5	0.5
Subsample ratio of the training instances	0.9999	0.9999
Maximum delta step, we allow each leaf output to be	0 (There is no constraint)	0 (There is no constraint)

Table 4
Assessing the results of different AI models for the testing stage.

Method	R-square (FSI) Validation	Test	*p*-value	R- square (Log MF) Validation	Test	*p*-value
XGBoost	0.9133	0.8347	–	0.9227	0.8813	–
Random Forest	0.8226	0.5707	4.33E-15	0.8875	0.8465	3.27E-08
Support Vector Regression	0.7602	0.7159	4.20E-155	0.6036	0.5328	5.54E-164

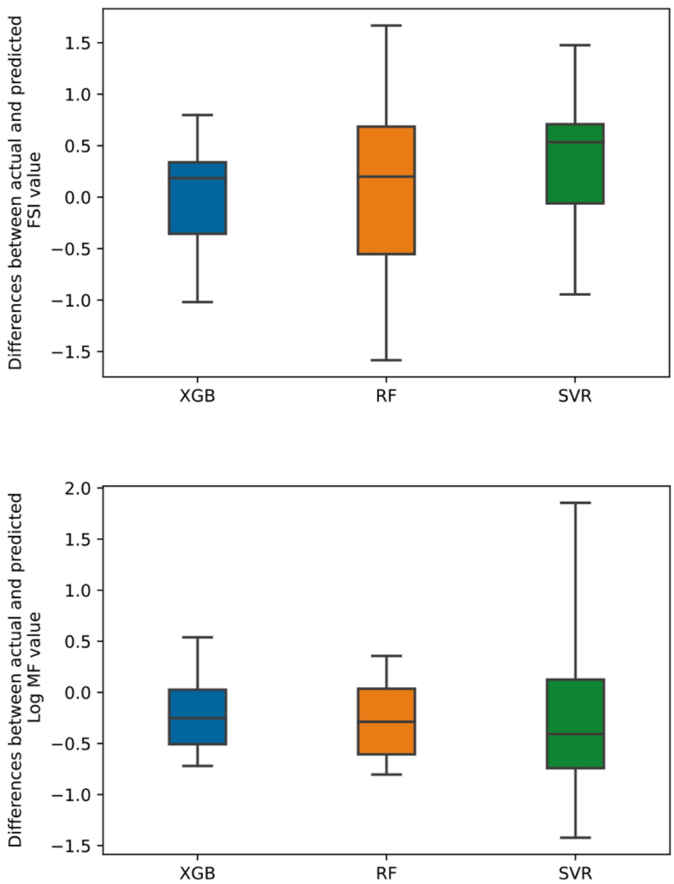

Fig. 5. Differences between actual and predicted coking indexes by various AI models in the testing stage.

It is worth considering that both RF and XGBoost are ensemble models which provide accurate results (SVR is a kernel-based regression). However, XGBoost is a boosting technique requiring less feature engineering, and RF is a bagging model, making XGBoost more adaptable than others [80]. During using XGBoost, the user can customize the objective function. Most SVR modeling shows a significant performance when a high dimensional space is available due to the kernel trick. In terms of training computational cost, XGBoost is cheaper than RF by implementing parallel processing [81], while SVR is one of the most computationally expensive algorithms to train. Although XGBoost is computationally efficient in training, it can be computationally costly in tuning due to its many hyperparameters. It is worth noting that one of the strengths of all three methods is that they make no assumptions about the distribution of the input features [82]. Regarding the bias and variance, RF and XGBoost mainly showed a low bias and variance; however, SVR indicated low bias and high variance [83]. As a substantial XAI system, the high accuracy of the constructed SHAP-XGBoost model indicated that this combination could successfully be applied for developing, modeling, and maintaining complex relationships within the coking and steelmaking industries.

4. Conclusion

Outcomes of this investigation highlighted that by using SHAP as an algorithm to build an explainable artificial intelligence model, complex multivariable correlations within coal properties and their representative coking indexes could be distinguished, and their multivariable correlation magnitude can be converted to the human basis level. SHAP indicated that with the dataset (coal sample's ash content is lower than 25%), moisture and carbon content has the highest importance for predicting coking indexes. There is a positive correlation between carbon content and coking quality, while moisture showed a significant negative correlation. XGBboost, a most recent developed boosting technique, could accurately predict coking indexes by R^2 over 0.9 in the validation and over 0.8 in the testing stages. Comparing the results of conventional machine learning methods (random forest and support vector regression) and SHAP-XGBoost models relieved that this model could provide a higher accuracy (R^2 0.9 vs. 0.8 in the validation, 0.8 vs. 0.6 in the testing step). The success of SHAP-XGBoost in modeling coking indexes could be considered a new window for better understanding complex relationships and predicting complicated factors within energy and fuel processing areas.

Declarations

Availability of data and materials

The data used to support the findings of this study are available from the corresponding author upon request.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

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.

Data availability

Data will be made available on request.

References

[1] Yin C, et al. Strength degradation mechanism of iron coke prepared by mixed coal and Fe2O3. J Anal Appl Pyrolysis 2020;150:104897.
[2] Galván-López E, McDermott J, O'Neill M, Brabazon A. Defining locality as a problem difficulty measure in genetic programming. Genet Program Evolvable Mach 2011;12(4):365–401.
[3] Cai S, et al. A novel method for removing organic sulfur from high-sulfur coal: migration of organic sulfur during microwave treatment with NaOH-H2O2. Fuel 2021;289:119800.
[4] Bobba S, Carrara S, Huisman J, Mathieux F, Pavel C. Critical raw materials for strategic technologies and sectors in the EU : a foresight study. Publications Office 2020. https://doi.org/10.2873/865242.

[5] Ryan BD, Price JT. The predicted coke strength after reaction values of British Columbia coals, with comparisons to international coals. Geological fieldwork 1992:1991–3.
[6] D\iez MA, Alvarez R, Barriocanal C. Coal for metallurgical coke production: predictions of coke quality and future requirements for cokemaking. Int J Coal Geol 2002;50(1–4):389–412.
[7] Huntington HD. Coal properties–measurement and application to cokemaking. Iron Steel Eng 1997;74(11).
[8] Bostick NH, Daws TA. Relationships between data from Rock-Eval pyrolysis and proximate, ultimate, petrographic, and physical analyses of 142 diverse US coal samples. Org Geochem 1994;21(1):35–49.
[9] ASTM D 720, Standard Test Method for Free Swelling Index of Coal. ASTM handbook for coal and coke. 1999. p. 226–30.
[10] ASTM D2639/D2639M-13 standard test method for plastic properties of coal by the constant-torque gieselerplastometer. ASTM International, West Conshohocken; 2013. https://www.astm.org/DATABASE.CART/HISTORICAL/D2639D2639M-13.htm (accessed Dec. 01, 2013).
[11] Hadavandi E, Chelgani SC. Estimation of coking indexes based on parental coal properties by variable importance measurement and boosted-support vector regression method. Measurement 2019;135:306–11.
[12] J.G. Speight, “Handbook of coal analysis,” John Wiley & Sons, 2015, pp. 145–148.
[13] Goscinski JS, Patalsky RM. CSR Control–a Coal Producers Point of View. Ironmaking Conference Proceedings 1990;49:53–74.
[14] Ryan B, Leeder R, Price JT, Gransden JF. The effect of coal preparation on the quality of clean coal and coke. Geological Fieldwork 1999:247–75.
[15] Toroglu I. The effects of ash and maceral composition of Azdavay and Kurucasile (Turkey) coals on coking properties. Energy Sources, Part A 2006;28(3):263–79.
[16] ASTM D3172, standard practice for proximate analysis of coal and coke. West Conshohocken, PA 19428–2959, United State: ASTM International; 2013. p. 1–2.
[17] ASTM D3176, standard practice for ultimate analysis of coal and coke. United States: ASTM International; 2015. p. 1–2.
[18] Matin SS, Chelgani SC. Estimation of coal gross calorific value based on various analyses by random forest method. Fuel 2016;177:274–8.
[19] Golzadeh M, Hadavandi E, Chelgani SC. A new Ensemble based multi-agent system for prediction problems: case study of modeling coal free swelling index. Appl Soft Comput 2018;64:109–25.
[20] Chelgani SC, Hower JC, Hart B. Estimation of free-swelling index based on coal analysis using multivariable regression and artificial neural network. Fuel Process Technol 2011;92(3):349–55.
[21] Khorami MT, Chelgani SC, Hower JC, Jorjani E. Studies of relationships between free swelling index (FSI) and coal quality by regression and adaptive neuro fuzzy inference system. Int J Coal Geol 2011;85(1):65–71.
[22] Chelgani SC, Matin SS. Study the relationship between coal properties with Gieseler plasticity parameters by random forest. Int J Oil Gas Coal Technol 2018;17 (1):113–27.
[23] Chelgani SC, Makaremi S. Explaining the relationship between common coal analyses and Afghan coal parameters using statistical modeling methods. Fuel Process Technol 2013;110:79–85.
[24] Rožanec JM, Fortuna B, Mladenić D. Knowledge graph-based rich and confidentiality preserving Explainable Artificial Intelligence (XAI). Information Fusion 2022;81:91–102.
[25] Ward IR, Wang L, Lu J, Bennamoun M, Dwivedi G, Sanfilippo FM. Explainable artificial intelligence for pharmacovigilance: What features are important when predicting adverse outcomes? Comput Methods Programs Biomed 2021;212: 106415.
[26] Alicioglu G, Sun B. A survey of visual analytics for Explainable Artificial Intelligence methods. Comput Graph 2022;102:502–20.
[27] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30:4765–74.
[28] Antwarg L, Miller RM, Shapira B, Rokach L. Explaining anomalies detected by autoencoders using Shapley Additive Explanations. Expert Syst Appl 2021;186: 115736.
[29] Tideman LEM, et al. Automated biomarker candidate discovery in imaging mass spectrometry data through spatially localized Shapley additive explanations. Anal Chim Acta 2021;1177:338522.
[30] Kannangara KKPM, Zhou W, Ding Z, Hong Z. Investigation of feature contribution to shield tunneling-induced settlement using Shapley additive explanations method. J Rock Mech Geotech Eng 2022.
[31] C. Gaur, “Top 6 Big Data Challenges and Solutions to Overcome,” 2021. https://www.xenonstack.com/insights/big-data-challenges (accessed Aug. 15, 2022).
[32] Choi MY, Ma C. Making a big impact with small datasets using machine-learning approaches. Lancet Rheumatol 2020;2(8):e451–2.
[33] Fatahi R, Nasiri H, Dadfar E, Chehreh Chelgani S. Modeling of energy consumption factors for an industrial cement vertical roller mill by SHAP-XGBoost: a ‘conscious lab’ approach. Sci Rep 2022;12(1):7543. https://doi.org/10.1038/s41598-022-11429-9.
[34] Mao H, et al. Driving safety assessment for ride-hailing drivers. Accid Anal Prev 2021;149. https://doi.org/10.1016/j.aap.2020.105574. Jan.
[35] Ekanayake IU, Meddage DPP, Rathnayake U. A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). Case Stud Construct Mater 2022;16:e01059.
[36] D.-.C. Feng, A.M. Asce, W.-.J. Wang, S. Mangalathu, E. Taciroglu, and M. Asce, “Interpretable XGBoost-SHAP Machine-Learning Model for Shear Strength Prediction of Squat RC Walls,” 2021, doi: 10.1061/(ASCE).
[37] Adland R, Jia H, Lode T, Skontorp J. The value of meteorological data in marine risk assessment. Reliab Eng Syst Saf 2021;209. https://doi.org/10.1016/j.ress.2021.107480. May.
[38] Mangalathu S, Shin H, Choi E, Jeon J-S. Explainable machine learning models for punching shear strength estimation of flat slabs without transverse reinforcement. J Build Eng 2021;39:102300.
[39] Liang M, Chang Z, Wan Z, Gan Y, Schlangen E, Šavija B. Interpretable Ensemble-Machine-Learning models for predicting creep behavior of concrete. Cem Concr Compos 2022;125. https://doi.org/10.1016/j.cemconcomp.2021.104295. Jan.
[40] Park H, Park DY. Comparative analysis on predictability of natural ventilation rate based on machine learning algorithms. Build Environ 2021;195. https://doi.org/10.1016/j.buildenv.2021.107744. May.
[41] Matin SS, Pradhan B. Earthquake-induced building-damage mapping using Explainable AI (XAI). Sensors 2021;21(13):4489.
[42] Nasiri H, Homafar A, Chehreh Chelgani S. Prediction of uniaxial compressive strength and modulus of elasticity for Travertine samples using an explainable artificial intelligence. Result Geophys Sci 2021;8:100034. https://doi.org/10.1016/j.ringps.2021.100034.
[43] Ezzoddin M, Nasiri H, Dorrigiv M. Diagnosis of COVID-19 Cases from Chest X-ray Images Using Deep Neural Network and LightGBM. In: 2022 International Conference on Machine Vision and Image Processing (MVIP); 2022. p. 1–7. https://doi.org/10.1109/MVIP53647.2022.9738760.
[44] Nasiri H, et al. Classification of COVID-19 in Chest X-ray Images Using Fusion of Deep Features and LightGBM. In: 2022 IEEE World AI IoT Congress (AIIoT); 2022. p. 201–6. https://doi.org/10.1109/AIIoT54504.2022.9817375.
[45] Zhang Q, et al. Three-dimensional mineral prospectivity mapping by xgboost modeling: a case study of the Lannigou Gold Deposit, China. Nat Resour Res 2022: 1–22.
[46] Nasiri H, Hasani S. Automated detection of COVID-19 cases from chest X-ray images using deep neural network and XGBoost. Radiography 2022;28(3):732–8. https://doi.org/10.1016/j.radi.2022.03.011.
[47] M.R. Abbasniya, S.A. Sheikholeslamzadeh, H. Nasiri, and S. Emami, “Classification of Breast Tumours Based on Histopathology Images Using Deep Features and Ensemble of Gradient Boosting Methods,” arXiv preprint , 2022.
[48] Chelgani SC. Estimation of gross calorific value based on coal analysis using an explainable artificial intelligence. Mach Learn Appl 2021;6:100116.
[49] Bhati BS, Chugh G, Al-Turjman F, Bhati NS. An improved ensemble based intrusion detection technique using XGBoost. Trans Emerg Telecommun Tech 2021;32(6). https://doi.org/10.1002/ett.4076. Jun.
[50] Wang R, Wang L, Zhang J, He M, Xu J. XGBoost machine learning algorism performed better than regression models in predicting mortality of moderate-to-severe traumatic brain injury. World Neurosurg 2022.
[51] Xian S, Chen K, Cheng Y. Improved seagull optimization algorithm of partition and XGBoost of prediction for fuzzy time series forecasting of COVID-19 daily confirmed. Adv Eng Softw 2022:103212.
[52] Yan Z, Chen H, Dong X, Zhou K, Xu Z. Research on prediction of multi-class theft crimes by an optimized decomposition and fusion method based on XGBoost. Expert Syst Appl 2022;207:117943.
[53] Zhao Z, Duan W, Cai G, Wu M, Liu S. CPT-based fully probabilistic seismic liquefaction potential assessment to reduce uncertainty: integrating XGBoost algorithm with Bayesian theorem. Comput Geotech 2022;149:104868.
[54] Nasiri H, Alavi SA. A novel framework based on deep learning and anova feature selection method for diagnosis of COVID-19 cases from chest x-ray images. Comput Intell Neurosci 2022;2022:4694567. https://doi.org/10.1155/2022/4694567.
[55] Shehadeh A, Alshboul O, al Mamlook RE, Hamedat O. Machine learning models for predicting the residual value of heavy construction equipment: an evaluation of modified decision tree, LightGBM, and XGBoost regression. Autom Constr 2021; 129. https://doi.org/10.1016/j.autcon.2021.103827. Sep.
[56] Hasani S, Nasiri H. COV-ADSX: an automated detection system using X-ray images, deep learning, and XGBoost for COVID-19. Softw Impacts 2022;11:100210. https://doi.org/10.1016/j.simpa.2021.100210.
[57] Zhang W, Wu C, Zhong H, Li Y, Wang L. Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. Geosci Front 2021;12(1):469–77.
[58] Zhang S, et al. Mineral prospectivity mapping based on isolation forest and random forest: Implication for the existence of spatial signature of mineralization in outliers. Nat Resour Res 2021:1–19.
[59] Chehreh Chelgani S, Nasiri H, Tohry A. Modeling of particle sizes for industrial HPGR products by a unique explainable AI tool- A ‘Conscious Lab’ development. Adv Powder Technol 2021;32(11):4141–8. https://doi.org/10.1016/j.apt.2021.09.020.
[60] Zojaji Z, Ebadzadeh MM, Nasiri H. Semantic schema based genetic programming for symbolic regression. Appl Soft Comput 2022;122:108825. https://doi.org/10.1016/j.asoc.2022.108825.
[61] Chehreh Chelgani S. Prediction of specific gravity of Afghan coal based on conventional coal properties by stepwise regression and random forest. Energy Sources Part A 2019:1–12.
[62] Wang J, Zuo R, Xiong Y. Mapping mineral prospectivity via semi-supervised random forest. Nat Resour Res 2020;29(1):189–202.
[63] Bu X, Vahed AT, Ghassa S, Chelgani SC. Modelling of coal flotation responses based on operational conditions by random forest. Int J Oil Gas and Coal Technol 2021; 27(4):457–68.
[64] Hou S, Liu Y, Yang Q. Real-time prediction of rock mass classification based on TBM operation big data and stacking technique of ensemble learning. J Rock Mech Geotech Eng 2022;14(1):123–43. https://doi.org/10.1016/j.jrmge.2021.05.004. Feb.

[65] Han H, Jahed Armaghani D, Tarinejad R, Zhou J, Tahir MM. Random forest and bayesian network techniques for probabilistic prediction of flyrock induced by blasting in quarry sites. Nat Resour Res 2020;29(2):655–67.

[66] Jafrasteh B, Fathianpour N, Suárez A. Comparison of machine learning methods for copper ore grade estimation. Comput Geosci 2018;22(5):1371–88.

[67] Abellán-García J, Guzmán-Guzmán JS. Random forest-based optimization of UHPFRC under ductility requirements for seismic retrofitting applications. Constr Build Mater 2021;285. https://doi.org/10.1016/j.conbuildmat.2021.122869. May.

[68] Ahmad MS, Adnan SM, Zaidi S, Bhargava P. A novel support vector regression (SVR) model for the prediction of splice strength of the unconfined beam specimens. Constr Build Mater 2020;248:118475.

[69] Chelgani SC, Nasiri H, Alidokht M. Interpretable modeling of metallurgical responses for an industrial coal column flotation circuit by XGBoost and SHAP-A 'conscious-lab' development. Int J Min Sci Technol 2021;31(6):1135–44. https://doi.org/10.1016/j.ijmst.2021.10.006.

[70] Paryani S, Neshat A, Pourghasemi HR, Ntona MM, Kazakis N. A novel hybrid of support vector regression and metaheuristic algorithms for groundwater spring potential mapping. Sci Total Environ 2022;807:151055.

[71] Wei J, Dong G, Chen Z. Remaining useful life prediction and state of health diagnosis for lithium-ion batteries using particle filter and support vector regression. IEEE Trans Ind Electron 2018;65(7):5634–43.

[72] Zhang L, Liu W, Men D. Preparation and coking properties of coal maceral concentrates. Int J Min Sci Technol 2014;24(1):93–8.

[73] Gazulla MF, Rodrigo M, Orduña M, Ventura MJ. Determination of organic oxygen in petroleum cokes and coals. Microchem J 2016;126:538–44.

[74] Ayoub M, et al. A comprehensive review on oil extraction and biodiesel production technologies. Sustainability 2021;13(2):788.

[75] Al-Juboori O, Sher F, Hazafa A, Khan MK, Chen GZ. The effect of variable operating parameters for hydrocarbon fuel formation from CO2 by molten salts electrolysis. J CO2 Util 2020;40:101193.

[76] Al-Shara NK, Sher F, Iqbal SZ, Curnick O, Chen GZ. Design and optimization of electrochemical cell potential for hydrogen gas production. J Energy Chem 2021;52:421–7.

[77] Al-Juboori O, Sher F, Khalid U, Niazi MBK, Chen GZ. Electrochemical production of sustainable hydrocarbon fuels from CO2 co-electrolysis in eutectic molten melts. ACS Sustain Chem Eng 2020;8(34):12877–90.

[78] Sher F, Al-Shara NK, Iqbal SZ, Jahan Z, Chen GZ. Enhancing hydrogen production from steam electrolysis in molten hydroxides via selection of non-precious metal electrodes. Int J Hydrog Energy 2020;45(53):28260–71.

[79] Nasiri H, Ebadzadeh MM. MFRFNN: multi-functional recurrent fuzzy neural network for chaotic time series prediction. Neurocomputing 2022;507:292–310. https://doi.org/10.1016/j.neucom.2022.08.032.

[80] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 785–94.

[81] Gajjar A, Kashyap P, Aysu A, Franzon P, Dey S, Cheng C. FAXID: FPGA-Accelerated XGBoost Inference for Data Centers using HLS. In: 2022 IEEE 30th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM); 2022. p. 1–9.

[82] H. Rhys, Machine Learning with R, the tidyverse, and mlr. Simon and Schuster, 2020.

[83] Fan J, et al. Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: a case study in China. Energy Convers Manag 2018;164:102–11.