GIG Research Institute

JOURNAL OF SUSTAINABLE MINING

Volume 24 | Issue 1 Article 3

2025

Leveraging intrinsic properties for classification of coal seams towards spontaneous combustion proclivity and predicting susceptibility using machine learning: smart and sustainable mining approach

Author(s) ORCID Identifier:
Siddhartha Agarwal: 0000-0001-6883-9660
Pradeep K. Gautam: 0000-0002-1600-7405
Rishabh Dwivedi: 0000-0002-1031-8202
D.C. Panigrahi: 0000-0002-7493-9649
C. Dagli: 0000-0003-4919-1699
A. Singh: 0009-0000-2586-1286

Recommended Citation

Agarwal, Siddhartha; Gautam, Pradeep K.; Zou, Yuhao; Dwivedi, Rishabh; Panigrahi, D.C.; Dagli, C.; and Singh, Atul (2025) "Leveraging intrinsic properties for classification of coal seams towards spontaneous combustion proclivity and predicting susceptibility using machine learning: smart and sustainable mining approach," *Journal of Sustainable Mining*: Vol. 24 : Iss. 1 , Article 3.
Available at: https://doi.org/10.46873/2300-3960.1436

Leveraging intrinsic properties for classification of coal seams towards spontaneous combustion proclivity and predicting susceptibility using machine learning: smart and sustainable mining approach

Abstract

Mine fires and other hazards caused by spontaneous coal combustion are a pervasive and longstanding issue in Jharia coalfields, India. This study proposes a novel approach to classify coal seams based on their propensity to spontaneous combustion using the intrinsic properties of 30 coal samples from different seams. This method eliminates the need for expensive and time-consuming experimental determinations of susceptibility indices (SI) such as crossing point temperature (CPT), critical air blast (CAB), and differential thermal analysis (DTA). All clustering models, viz. hierarchical, k-means, and multidimensional scaling, aptly classify coal seams into three categories: highly risky, medium risky, and low risk in terms of the tendency for spontaneous coal combustion. The results from unsupervised clustering for predicting the fieriness of coal seams match with on-field reports based on the history and nature of seams. The clustering results are also in concurrence with the SI which are generated through lab investigations. Furthermore, three machine learning (ML) algorithms, namely support vector machines (SVM), random forests (RF), and elastic net regression (EN), are used to comprehend the relationship between the coal's intrinsic properties of coal and SI. The actual nature of coal in these seams on the ground confirmed the findings of this study. The proposed methodology has practical implications for mine managers, as it can quickly provide safety risk assessment information to maintain safety and minimize economic losses due to unforeseen incidents.

Keywords

coal, spontaneous combustion, susceptibility indices, machine learning, mine fire

Authors

Siddhartha Agarwal, Pradeep K. Gautam, Yuhao Zou, Rishabh Dwivedi, D.C. Panigrahi, C. Dagli, and Atul Singh

RESEARCH ARTICLE

Leveraging intrinsic properties for classification of coal seams towards spontaneous combustion proclivity and predicting susceptibility using machine learning: Smart and sustainable mining approach

Siddhartha Agarwal [a,*], Pradeep K. Gautam [b], Yuhao Zou [c], Rishabh Dwivedi [d], Durga C. Panigrahi [a], Cihan H. Dagli [e], A. Singh [a]

[a] Department of Mining Engineering, Indian Institute of Technology (ISM), Dhanbad, India
[b] Department of Civil & Environmental Engineering, Colorado School of Mines, Golden, CO, USA
[c] Director of Loss Forecasting and Quantitative Modeling, Sallie Mae, Hockessin, DW, 19707, USA
[d] Department of Earth Sciences, Indian Institute of Technology Bombay, India
[e] Systems Engineering Department, Missouri Science and Technology, Rolla, MO, 65401, USA

Abstract

Mine fires and other hazards caused by spontaneous coal combustion are a pervasive and longstanding issue in Jharia coalfields, India. This study proposes a novel approach to classify coal seams based on their propensity to spontaneous combustion using the intrinsic properties of 30 coal samples from different seams. This method eliminates the need for expensive and time-consuming experimental determinations of susceptibility indices (SI) such as crossing point temperature (CPT), critical air blast (CAB), and differential thermal analysis (DTA). All clustering models, viz. hierarchical, k-means, and multidimensional scaling, aptly classify coal seams into three categories: highly risky, medium risky, and low risk in terms of the tendency for spontaneous coal combustion. The results from unsupervised clustering for predicting the fieriness of coal seams match with on-field reports based on the history and nature of seams. The clustering results are also in concurrence with the SI which are generated through lab investigations. Furthermore, three machine learning (ML) algorithms, namely support vector machines (SVM), random forests (RF), and elastic net regression (EN), are used to comprehend the relationship between the coal's intrinsic properties of coal and SI. The actual nature of coal in these seams on the ground confirmed the findings of this study. The proposed methodology has practical implications for mine managers, as it can quickly provide safety risk assessment information to maintain safety and minimize economic losses due to unforeseen incidents.

Keywords: coal, spontaneous combustion, susceptibility indices, machine learning, mine fire

1. Introduction

In an effort to reduce financial losses and meet the country's substantial energy demands, Coal India Limited (CIL) transitioned from underground mining to open-cast mining, a simpler and more profitable extraction method where coal is mined from the surface [1]. This shift resulted in the stripping and blasting of vast areas of land into pits as deep as 400 feet. These pits now produce over 700 million tons of coal annually. There is, though, the agreement that it was the shift to open-cast mining that exposed the flames to oxygen, which in turn caused the fires to increase in ferocity and break the surface, hampering decades-long efforts to put them out [2]. Poor mining practices can also ignite coal fires that may burn for decades, releasing fly ash and smoke filled with greenhouse gases and toxic chemicals [3]. Additionally, mining activities release coal mine methane, a greenhouse gas that is 20 times more potent than carbon dioxide [4]. For the ambitious targets of meeting India's consumption of

Received 27 December 2023; revised 6 July 2024; accepted 15 July 2024.
Available online 1 January 2025

* Corresponding author.
E-mail address: sagarwal@iitism.ac.in (S. Agarwal).

https://doi.org/10.46873/2300-3960.1436

energy, a methodology has been formulated in the form of a sustainable development cell for environmental mitigation measures [5,6]. A roadblock that stands in the way of achieving this aspiring target set by the Government of India is coal mine fires (CMF), the primary cause of which is spontaneous combustion (SC) or self-oxidation of coal. Previously due to CMF's more than 37 million tons of coal worth billions of dollars have been reckoned worthless due to improper mining [7,8]. SC is caused by the incomplete oxidation of coal, where coal burns by itself without an external heat source with the aid of atmospheric oxygen, as shown in Figure 1 [9]. With SC, the greatest risk is an explosion due to a mine fire and rendering the coal unfit for any profitable use [10]. SC also injects harmful poisonous gases such as carbon monoxide and sulphur oxide into the air and is responsible for roof falls in underground mines due to surface land subsidence [11]. Health and safety risks involve cancer in the lungs and throat, continuous coughing for mine workers, headaches, tuberculosis, and asthma for mine residents [12,13]. Therefore, to mitigate the CMF, it is crucial to study coal seam properties to examine the impact on self-heating susceptibility.

Many different susceptibility indexes (SI) have been prevalent in determining the propensity of coal towards SC. Some of them include critical oxidation temperature study, crossing point temperature (CPT) [14], differential thermal analysis (DTA) [15], and critical air blast (CAB), Wits-Ehac index [16], thermogravimetric analysis (TGA) [17], adiabatic oxidation method [18], and goaf ignition temperature [19]. Pattanaik and others (2011) recognized that the intrinsic properties of coal (such as coal rank, volatile matter, and petrography) are related to two SI, namely, CPT and DTA [20]. Researchers conducted 600 experiments with 50 coal samples covering different geographical locations in India, both involving fiery and non-fiery seams, to determine the relationship of wet oxidation potential (WOP) for SC [21]. The results corroborated with field observations and also matched the crossing point temperature method. Previously, authors conducted a study using an SI as a Russian U-index and its correlation with basic constituents of coal in Jharia Coalfields [22,23]. The relationship between the CPT, moisture, coal rank, and sulphur content has also been examined [24]. Others have explored the statistical relationship between the CPT and the parameters of proximate analysis using linear regression for Indian coals [25]. Pore structure, volatile matter, oxygen content, and fixed carbon content were found to play important roles in spontaneous combustion as well [26]. In China, the analysis of gas patterns to predict spontaneous combustion in one of the coal seams of the Juye coalfield led to the development of an optimized index gas and a six-level warning system [27]. In a study, it was estimated that specific heat capacity, thermal conductivity, moisture content, oxygen concentration, oxidation rate, and gas flow seepage velocity all influence the spontaneous combustion of coal, and combustion risk decreases with higher specific heat capacity and moisture content, but increases with higher thermal conductivity, oxygen concentration, gas seepage velocity, and oxidation rate [28]. In Australian underground coal mining, the risk of spontaneous combustion was assessed using the R70 and CPT methods. A study of 318 samples found a strong correlation between the two methods ($\rho = -0.8875$). The revised risk categories are: low, medium, high, and very high, based on specific R70 and CPT thresholds [29].

All of the above methods that use SI to evaluate the propensity for spontaneous combustion of coal samples are often expensive, time-consuming, and

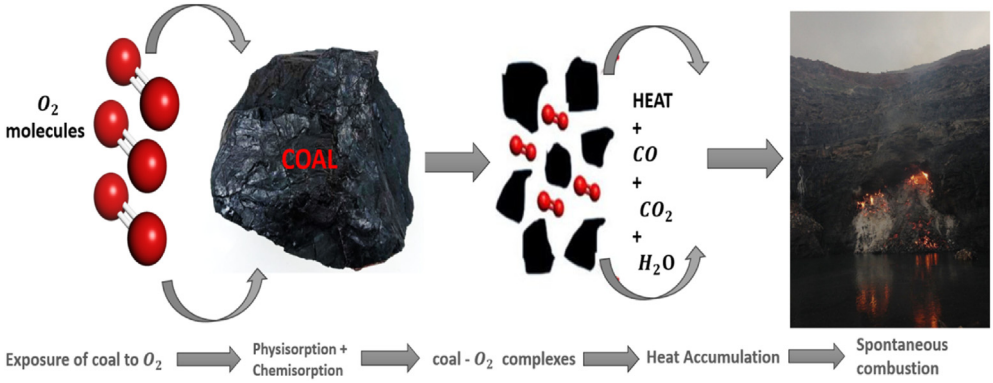

Fig. 1. An overview of the process of spontaneous combustion in coal.

not always deterministic. As a result, these traditional methods have been increasingly replaced by machine learning and AI algorithms, which are faster and require minimal or no apparatus [30,31]. Besides, it helps decrease the time taken to make decisions, which hampers the production rate of the mine.

Several studies undertaken to use AI include the Sparrow Search Algorithm (SSA) and Convolutional Neural Network (CNN) to predict coal spontaneous combustion temperature. This SSA-CNN model was found to be more accurate than other models, accurately predicting combustion temperature with high precision [32]. Based on the 35 coal samples obtained from the Witbank Coalfields, South Africa, the proximate and ultimate analyses of coal samples helped in developing reliable prediction models using artificial neural networks (ANN), neuro-fuzzy inference systems (ANFIS), and multilinear regressions (MLR) [33]. A prediction model based on principal component analysis (PCA), case-based reasoning (CBR), fuzzy clustering (FM), and snake optimization (SO) was proposed in this manuscript in order to accurately predict coal spontaneous combustion hazard grades [34]. A borehole spontaneous combustion prediction model combining the hunger games search (HGS) optimization algorithm and random forest algorithm (RF) was presented [35]. HGS optimized the number of trees and the minimum number of leaf nodes in RF. A study was done in a coal mine in Shanxi using a particle swarm optimization (PSO) algorithm coupled with BP neural network model (BPNN) model [36]. This PSO-BPNN model predicted spontaneous combustion better than only the BPNN model, genetic algorithm (GA-BPNN) model, the SSA-BPNN model, and the marine predators algorithm (MPA-BPNN) model. The application and performance evaluation of artificial neural networks (ANN) and spotted hyena-optimized ANNs (SHO-ANN) were used to develop reliable models for predicting coal spontaneous combustion liability [37]. They show that volatile matter (VM) and oxygen (O) have the greatest influence on Wits-Ehac and FCC (Feng, Chakravorty, Cochrane) liability indices. Researchers have used an improved grey wolf optimization with support vector regression to predict coal spontaneous combustion temperature [38]. Using O_2, CO_2, and C_2H_6 as independent variables, the GA-SVR model is used to calculate CO concentration, which is then utilized to calculate coal temperatures and assess the risk of spontaneous combustion [39]. A study was conducted using five machine learning techniques, such as support vector machine (SVR), ANN, RF, gradient boosting (GB), and extreme gradient boosting (XGB) in the prediction of wet oxidation potential (WOP). XGB showed the highest accuracy and sensitivity, whereas volatile matter was identified as the most influential parameter in assessing fire risk [39]. Many researchers have linear regression to model and compare with ML models for spontaneous combustion liability dependence on ash, volatile matter, carbon, hydrogen, exinite, and inertinite present in coal [40,41]. Moisture and volatile matter were found most useful in assessing coal's susceptibility to spontaneous heating, as they exhibit a stronger link than other intrinsic parameters such as ash content and gross calorific values [42,43]. It has been determined that there is a clear positive or negative correlation between the inherent factors and the risk of spontaneous combustion [44,45].

A systematic search of the literature has revealed that more studies have been done using ANN and susceptibility indices. Besides, no other model has used clustering to categorize the coal seams based on their propensity for spontaneous combustion. This research eliminates the use of calculating susceptibility indices, which require expensive laboratory set-up and are time-consuming. Clustering can reveal natural groupings within the data, such as coal seams with similar properties that might have similar susceptibility to spontaneous combustion. These groupings can provide insights that are not immediately obvious through other analytical methods. This simplification makes it easier to visualize and interpret the data, aiding in better decision-making and communication of findings. This methodology is capable of real-time data processing, and risk assessment addresses the gap in continuous monitoring and timely decision-making in mining operations. Please refer to Figure 2 for an overview of the SC prediction methodology.

The highlights of this study include:

- Sustainable development of the Indian coal sector to ensure increased production and to do so in an environmentally and socially safe manner.
- Understanding the intricate relationship between the intrinsic properties of coal (most important factors) and its propensity to self-combust with the help of atmospheric oxygen and moisture.
- Advancing the use of machine learning techniques and data analytics in the mining sector for real-time analysis of events that are extremely risky and hazardous, such as mine fire.
- Identifying the susceptibility index of coal (a measure of proclivity to self-heat) that is most

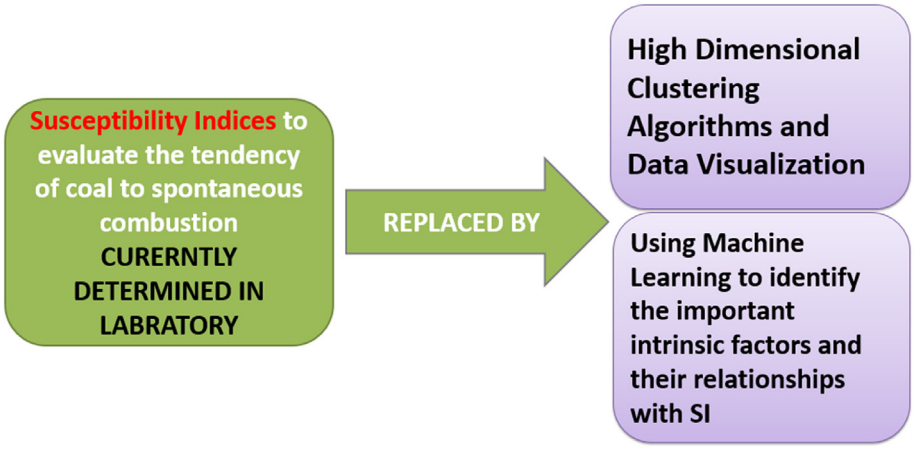

Fig. 2. An overview of the SC prediction methodology.

correlated with its intrinsic properties. Also, recognizing the importance score of each intrinsic property.

This approach also helps in 2D visualization of the results to better interpret what properties of coal are affecting the SC and what the intensity of that effect is. Overall, this research aims to develop an analytical framework that decision-makers can use to predict the SC of coal and classify coal seams based on their level of risk by identifying the importance of intrinsic coal properties. This research can develop models that are scalable and generalizable, applicable to various mining operations and geographic locations, thereby broadening their utility and impact. By delineating coal seams into categories such as highly risky, medium risky, and low risk, the models generated through clustering offer a comprehensive assessment of the likelihood of combustion events.

The results indicate that intrinsic properties are more accurate than standard techniques such as CPT and DTA. By leveraging the framework established in this study, the coal mining sector can efficiently categorize them and improve their mining methodology and production planning. Therefore, this research contributes to integrating artificial intelligence (AI) and machine learning (ML) as essential components of safe and smart mining operations.

Section 2 describes the location of coal samples collected for this study, along with the experimentally determined properties used for analysis. Section 3 illustrates the data pre-processing before applying ML algorithms. Section 4 describes the methodology of clustering the same into varying digress of susceptibility to catch fire, and, finally, section 5 establishes the most important parameters and equations to establish relationships between intrinsic coal properties and various SI.

2. Sampling and experimental property determination

The Jharia coal basin, located in the Dhanbad district, is a significant source of coking coal in India. Jharia Coalfields has been a hotbed for SC and CMF since 1916 and covers a large area of 284.899 sq. km, as shown in Figure 2 [46]. The National Remote Sensing Centre, Hyderabad, India, has confirmed that the surface mine fire in Jharia has now extended from 2.018 sq. km in 2014 to 3.28 sq. km in 2018 [47,48]. Around 1.4 billion metric tonnes of coal at Jharia have now become inaccessible due to CMF. Mine fire occurring in a Jharia coal mine through spontaneous combustion can be seen in Figure 3. Jharia coalfields and principal collieries affected by fires are illustrated on a map in Figure 4 [49]. As the industry moves to deeper mines and excavates

Fig. 3. Coal fires burn early in the morning in January 2014 at a privately owned coal mine in India's Jharkhand state, making air dense with toxic smoke and dust [57].

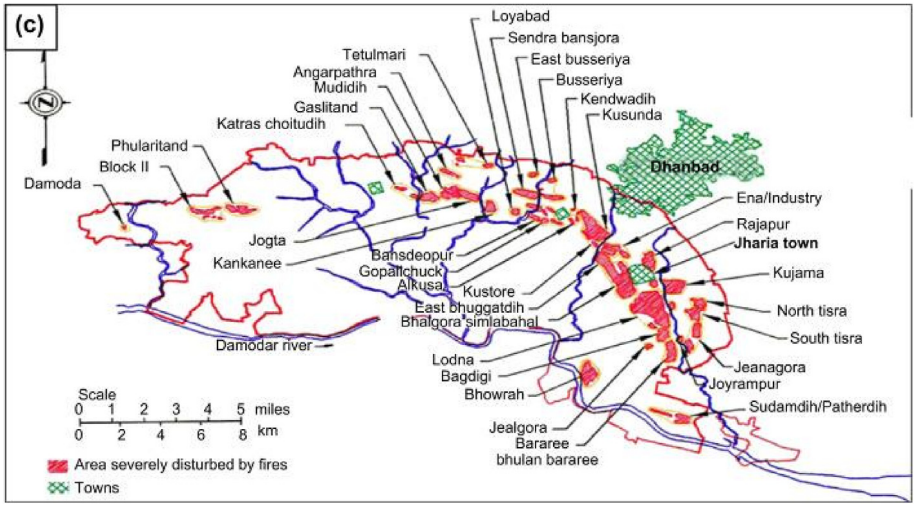

Fig. 4. Plan of the Jharia Coalfield and principal collieries affected by fires (Reproduced from [Michalski et al., 2000]) [46].

lower-rank coking coal, the risk of SC fires and related explosions is expected to increase. Jharia lies in the Gondwana sedimentary basin of the Damodar Valley [50]. The rock formations mainly belong to the Lower Gondwana group of the Permian age. From bottom to the top, they are listed as Talchir, Barakar, Barren measures, and Raniganj formations [51].

A total of 30 coal samples were collected to cover all coal seams in Jharia and a few others that are prone to SC from other coalfields. Table 1 presents information on the seam number, the coal mine area's name, and the operating company for each mine. The intrinsic properties of the coal samples were analyzed using various methods. Proximate analysis, based on the method specified by Indian Standard: 1350, Part I (1969) [52], was used to determine the moisture (M), ash (A), volatile matter (VM), and fixed carbon (FC) content. Ultimate analysis was conducted to determine the carbon (C), hydrogen (H), total sulfur (S), nitrogen (N), and oxygen (O) content as per Indian Standard: 1350, Part IV/Sec 1 (1974) [53], Indian Standard: 1350, Part III (1969) [53], and Indian Standard: 1350 (1975) [53], respectively, by subtracting the addition of the percentage of the remaining from 100 [54]. Physical properties such as porosity were calculated using the bulk volume determination method. The petrographic percentage of vitrinite, inertinite, and exinite, along with visible mineral matter present in coal samples, was determined following standard procedures in both white and fluorescence light (ICCP, 1971; ICCP, 1994; Indian Standard: 9127 [55], 1979 [56,57]. Atomic adsorption spectrophotometric analysis was used to identify the presence and quantity of Ca, Fe, and Mg. However, Ca and Mg compounds increase the thermal stability of coal, reducing the rate of spontaneous combustion and the susceptibility index. All intrinsic property values for the coal seams are detailed in Table 2. The SI

Table 1. List of samples with coal seams, operating collieries, and the subsidiaries of Coal India Ltd.

Sample No.	Coal seams	Operating collieries	Subsidiaries of Coal India Ltd.
1	Seam −0	Bastacola	BCCL
2	Seam-I	Bera	BCCL
3	Seam-II	Dobari	BCCL
4	Seam-III	Bera	BCCL
5	Seam-IV	Sendra BanSora	BCCL
6	Seam-V	Sendra BanSora	BCCL
7	Seam-VI	Godhur	BCCL
8	Seam-VII	Sendra BanSora	BCCL
9	Seam-VIII	Bhowra South	BCCL
10	Seam-IX	Gopalichuk	BCCL
11	Seam-X	Loyabad	BCCL
12	Seam-XI	South Balihari	BCCL
13	Seam-XII	South Balihari	BCCL
14	Seam-XIII	Bararee	BCCL
15	Seam-XIV	Jamadoba	TSL
16	Seam-XV	Kachi Balihari	BCCL
17	Seam-XVI	Bhagaband	BCCL
18	Seam-XVII	Bhowra South	BCCL
19	Seam-XVIII	Moonidih	BCCL
20	Mahuda Bottom	Muriidih	BCCL
21	Laikdih Seam	West Victoria	BCCL
22	Samla Seam	Samla	ECL
23	R−VII	Jhanjhara	ECL
24	Burra Dhemo Seam	Methani	ECL
25	Dakra Seam	Dakra Bukbuka	CCL
26	Hatidhari Seam	Saunda	CCL
27	Seam-III Under round	Chirimiri	SECL
28	Talcher Seam	Deulbera	MCL
29	Seam-IV B	Kampti	WCL
30	60 ft. Seam	Margarita	NECL

Table 2. Intrinsic properties of different coal seams.

Subsidiary seam	Spectrographic		Physical	Proximate				Ultimate						Petrographic
	Ca	Fe	Porosity (%)	Moisture	Ash	Volatile matter	Fixed carbon	C	H	N	S	O	Vitrinite	Exinite
Seam 0	540	7254	1.89	1	15.73	17.36	65.91	88.7	4.71	1.67	1.44	3.48	54.3	2.67
Seam-I	4350	9700	0.95	1.2	21.7	18.9	58.2	90.57	5.18	1.66	1.05	1.54	51.7	1.3
Seam-II	675	6900	2.37	1.3	30.2	15.2	53.3	91.15	5.68	1.58	0.72	0.87	51.62	1
Seam-III	4150	9500	4.5	1	24	18.7	56.3	91.24	5.43	1.9	0.87	0.56	59.5	0.8
Seam-IV	863	10,446	3	0.75	19.85	18.32	61.08	87.68	4.85	1.61	0.77	5.09	45.5	0.7
Seam-V	680	10,816	1.56	0.7	26.2	17.9	55.2	84.84	4.5	1.72	0.66	8.28	55.73	0.3
Seam-VI	4124	806	4.01	1	28.71	15.86	54.43	91.93	4.54	1.7	0.54	1.29	39.6	0.2
Seam-VII	618	3470	5.88	0.6	32.1	23.5	43.8	85.87	4.58	1.71	0.43	6.2	60.53	0.65
Seam-VIII	756	336	2.16	1.6	25.17	23.52	49.71	91.25	5.45	1.99	0.55	0.76	54.8	1.6
Seam-IX	2329	9790	4.65	1.31	21.95	18.97	57.77	82.28	4.43	1.73	50	11.06	48.8	0.7
Seam-X	446	2864	2.57	0.6	16.2	24.7	58.5	87.58	4.94	2.05	0.78	4.65	53	0.8
Seam-XI	4995	2190	4.09	1	16.9	22	60.2	86.58	4.79	1.88	0.72	6.03	77.5	1.9
Seam-XII	1395	4448	2.29	2.5	10.24	21.86	65.4	66.89	4.45	1.76	0.69	6.21	54	0.4
Seam-XIII	104	340	2.15	2.4	8.28	27.83	61.49	84.24	5.04	1.93	1.14	7.65	83.5	1.3
Seam-XIV	1910	13,064	1.1	1.5	15.1	18.1	65.3	87.18	4.99	1.58	0.56	0.05	44.5	0.3
Seam-XV	1576	6198	3.12	1.21	11.21	22.3	65.28	89.74	5.21	228	0.63	2.14	63.3	2.4
Seam-XVI	742	2370	4.42	2.2	9.57	25.7	62.53	85.95	5.26	2.35	0.88	0.05	69.4	0.8
Seam-XVII	2882	376	3.2	2	11.52	29.99	56.49	81.29	5.08	1.77	0.37	11.49	59.6	3.3
Seam-XVIII	3020	11,334	3.58	1.8	13.4	30.81	54.49	79.09	4.98	2.18	0.7	13.05	78.53	1.05
Mahuda Bottom	905	3872	2.72	1.9	16	33.08	49.02	87	4.12	2.35	0.43	5.24	69.4	9.3
Laikdih Seam	308	1314	8.8	0.6	10.73	22.32	66.35	86.75	5.33	1.87	0.47	5.58	71.24	5.01
Samla Seam	3320	2886	9.06	8.43	9.6	24.43	57.54	80.85	5.73	2.48	0.46	10.48	85.42	1.39
R–VII	893	354	14.5	8.11	27.07	27.52	37.3	66.57	5.12	1.8	0.71	25.8	74	9.4
Burra Dhemo Seam	2024	4014	12.5	1.8	14.27	36.13	47.8	80.6	5.99	2.26	0.79	10.36	71.55	10.71
Dakra Seam	1630	2552	18.17	10	18	32.27	39.73	77.71	5.26	1.56	0.64	14.83	77.3	7.1
Hatidhari Seam	728	4734	16.5	10.52	11.91	29.47	48.1	80.44	5.01	2.17	1.1	11.28	80.66	3.61
Seam-III Underground	5515	4284	16.11	7.67	18.88	29.83	43.62	79.8	4.97	0.87	1.05	13.31	37.24	6.28
Talcher Seam	5456	3386	18.02	8.57	9.22	37.46	44.75	80.49	6.53	2.08	1.01	9.89	59.46	8.11
Seam-IV B	4998	4002	21	14.39	12.76	29.31	43.54	79.4	3.73	1.2	0.55	14.75	30.3	12.5
60 ft. Seam	1075	4468	5.05	2.03	6.6	40.02	51.35	82.8	6.75	0.49	1.27	8.69	84.58	9.31

values obtained from the experiments conducted are presented in Table 3.

3. Data pre-processing

3.1. Exploratory data analysis

The present data set has a high number of features relative to the number of samples, and no outliers were detected. Exploratory data analysis (EDA) is an essential tool for understanding the statistical significance of variables in model building. A correlation heat map was also generated in Figure 5.

3.2. Feature selection and normalization

To improve ML results, it's advisable to reduce the number of features using feature selection [58]. Overfitting can occur if the model is trained with too many features, so it's important to use both statistical significance and domain knowledge to identify and remove less important features. In our study, we used the RReliefF algorithm to rank the features based on their importance [59]. It is especially known for its robustness and effectiveness in handling high-dimensional data and datasets with complex feature interactions. Unlike many other feature selection methods that evaluate features individually, RReliefF considers the contribution of each feature in the context of other features, making it capable of capturing complex dependencies. RReliefF is relatively robust to noisy data. It evaluates the importance of a feature based on its ability to distinguish between similar instances, which helps in reducing the impact of noise on the feature selection process. The algorithm assigns a higher score to features that distinguish instances closer to each other. We then applied an ad-hoc cutoff based on statistical or domain knowledge to select a subset of the scored and ranked features. Finally, the authors normalized the dataset as the scale of all variables is different. Different normalization techniques, such as scaling the data to [−1, 1] or [0, 1], or

Table 3. Values of SI for different coal seams.

Subsidiary seam	CAB	CPT	DTA_TEMP
Seam-0	2.47	180	314.28
Seam-I	1.44	172	288.57
Seam-II	1.32	167	295.23
Seam-III	2.36	168	297.14
Seam-IV	1.40	173	301.92
Seam-V	2.89	178	288.57
Seam-VI	1.50	175	298.07
Seam-VII	1.65	169	290.47
Seam-VIII	2.13	165	278.85
Seam-IX	1.65	173	280
Seam-X	1.19	174	284.62
Seam-XI	1.49	169	280.95
Seam-XII	1.03	164	274.29
Seam-XIII	1.32	159	280
Seam-XIV	1.582	179	296.15
Seam-XV	1.64	159	277.14
Seam-XVI	0.99	162	285.71
Seam-XVII	2.09	151	266.66
Seam-XVIII	1.75	163	276.19
Mahuda_Bottom	1.03	155	257.69
Laikdih Seam	1.26	160	269.05
Samla Seam	0.43	152	240
R-VII	0.47	138	200
Burra Seam	0.79	150	200
Dakra Seam	0.63	144	219.23
Hatidhari Seam	0.78	152.5	237.14
Seam-III_Underground	0.47	155	233.33
Talcher Seam	0.33	140	197.14
Seam-IV B	0.16	150	200
60 ft. Seam	0.75	150	240

centering it at $\mu = 0$, were used to improve prediction. The feature selection method RReliefF was used to score the features, as shown in Figure 6, which revealed that Exinite and VM are the most important features.

4. Clustering with high-dimensional data

This section delineates the outcomes of multidimensional clustering applied to coal seams, relying on their intrinsic properties through the utilization of unsupervised learning algorithms [60]. The three methods used are hierarchical clustering (Ward's method) [61,62], k-means [63], and multi-dimensional scaling (MDS) [64]. Fe content is excluded from the clustering process to avoid bias. All three methods yield consistent results regarding the number of clusters and their associated seams.

4.1. Hierarchical clustering

Hierarchical clustering is an unsupervised agglomerative method that is used to understand the clustering structure of data in a broader context. This method builds a cluster hierarchy such that clusters at one level are combined as clusters at the next level. The dendrogram, produced through hierarchical clustering, illustrates the distance between individual data points and clusters, with the x-axis

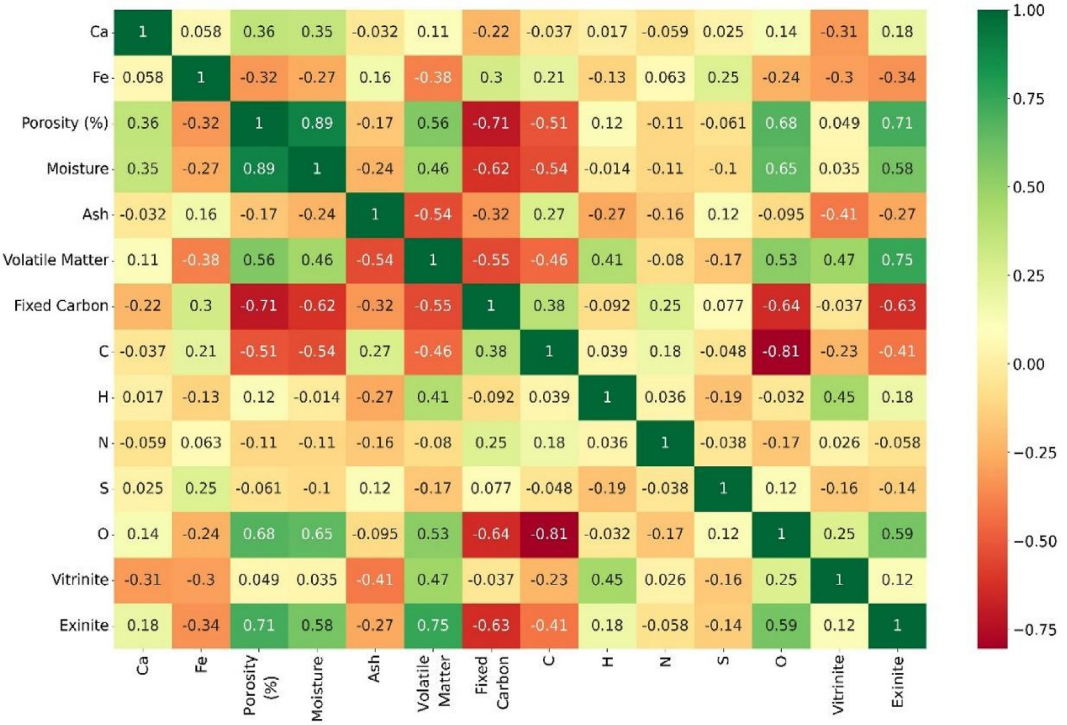

Fig. 5. Correlation score heat map of all input variables.

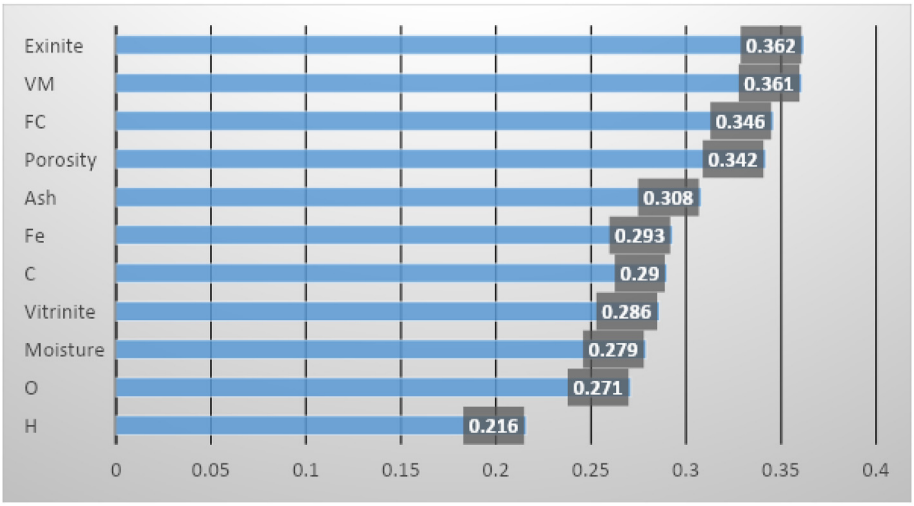

Fig. 6. ReleifF feature scores of input variables for predicting SI.

progressing from right to left. The y-axis lists the sample seam number and CPT to show the assignment of a seam under a cluster and its associated CPT value, as shown in Figures 7 and 8. The black vertical line helps determine the appropriate number of clusters in the dataset. Moving the line to the right increases the number of clusters, while moving it to the left decreases the number of clusters. Three clusters were found to be appropriate since increasing the number of clusters resulted in some clusters having only 2 seams. The blue cluster includes seams such as Talcher and Bhurra Dhemo with CPT values ranging from 138°C to 155°C, rendering it particularly prone to spontaneous combustion. The red cluster includes seams named Mahuda bottom and Samla, etc., with CPT values ranging from 165°C to 180°C, making it the least susceptible to spontaneous combustion. The green cluster has CPT values varying from 150°C to 164°C, indicating intermediate susceptibility to spontaneous combustion. This cluster contains coal seams named Bastacola, Jamadoba, and Godhur. These results are verified by on-field investigations regarding their susceptibility as well as in laboratory experiments. However, clustering itself does not take into account CPT (measured in the lab) but is still able to match the results. K-means clustering is performed next to confirm the results obtained thus far.

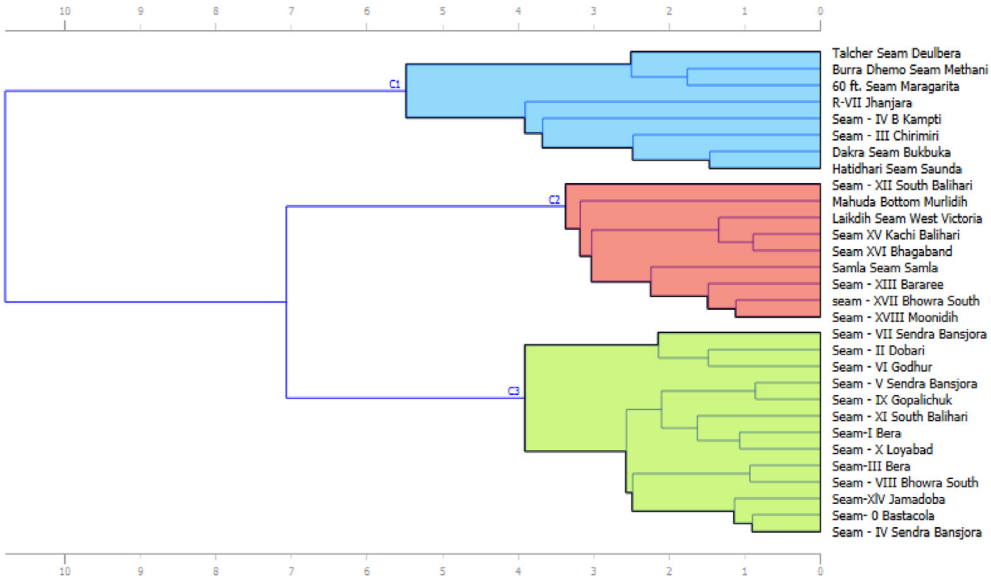

Fig. 7. Dendrogram using the Ward method depicting three major clusters with associated coal seam names.

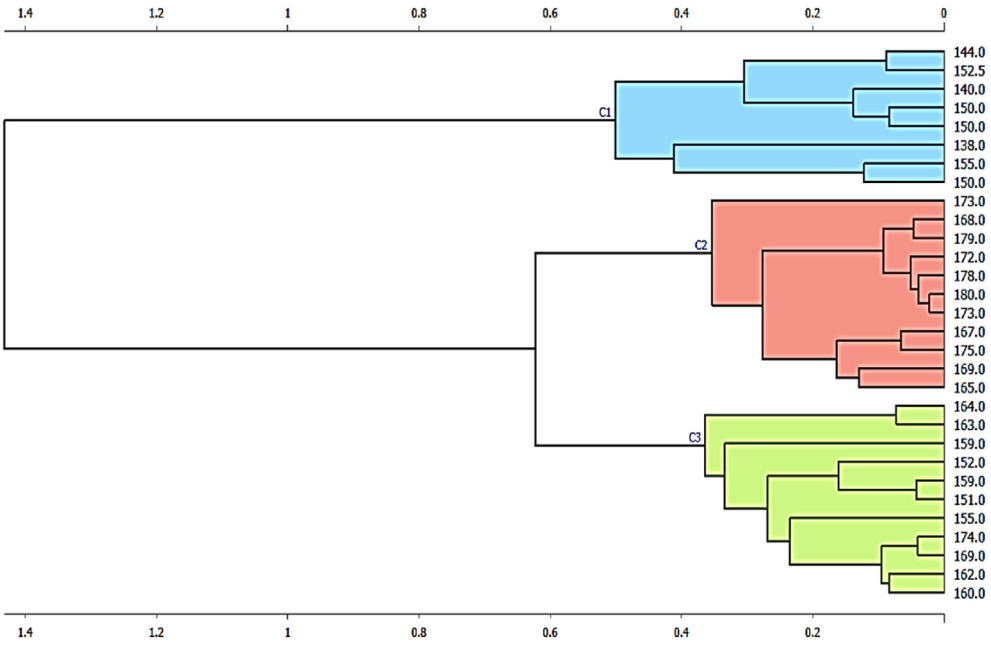

Fig. 8. A dendrogram by Ward method showing three major clusters (with seam CPT values).

4.2. K-means clustering algorithm

The K-means clustering algorithm is another commonly used unsupervised method for clustering. The algorithm necessitates the user to define the number of clusters "k" within the dataset, a parameter that can be determined either through domain knowledge or assessed using criteria like average silhouette width. The average silhouette width is another useful metric for evaluating the quality of clustering results and determining the optimal number of clusters. It measures how similar each data point is to its own cluster compared to other clusters, with a higher silhouette score indicating better-defined clusters [65]. A high positive silhouette score indicates that clusters are well-separated and distinct, meaning each point is more similar to its own cluster than to other clusters. Figure 9 shows the three clusters in the data, with two clusters having nine points each and the third with twelve points, and the score is mentioned on the y-axis. Figure 10 illustrates an elbow plot that is a useful graphical representation to determine the optimal number of clusters in a clustering algorithm like K-means. The idea is to run the clustering algorithm for a range of cluster numbers (k) and for each k, compute the sum of squared distances (inertia) between data points and their corresponding cluster centroids. The best result in terms of the highest average silhouette width obtained by iterating over many numbers of clusters is shown (Figure 9) when k = 3. The results validate the data having 3 major clusters as predicted by hierarchal clustering. Another clustering accuracy criterion is to plot the number of clusters against the average silhouette width that can help to identify the "elbow point", where the rate of decrease sharply falls, indicating the optimal number of clusters. The average silhouette width (in Figure 10) drops sharply when algorithm has an input of 3 and then again when k = 5 clusters while spontaneous combustion subject matter knowledge insists on having no more than 3 to 4 clusters in the data. Hence this too justifies the earlier claim of having 3 clusters in the data.

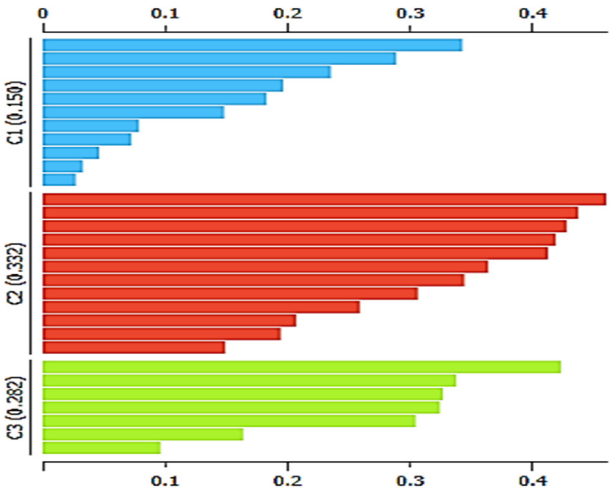

Fig. 9. K-means clustering showing three major clusters with silhouette plot.

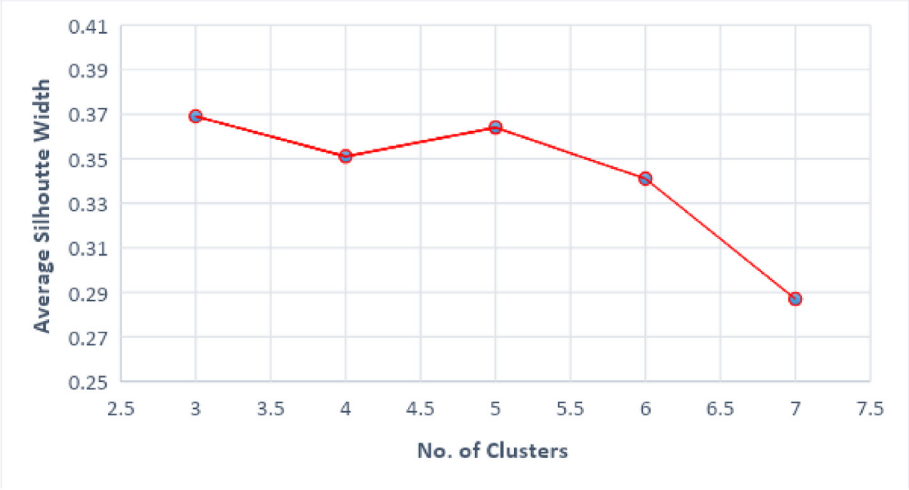

Fig. 10. Elbow plot of average silhouette width against the number of clusters.

4.3. Projection of multi-dimensional clustering results k-means

To better understanding the high-dimensional data, a powerful method is to use two-dimensional projections in conjunction with clustering. The projection directions in this case are intrinsic properties that will assist in providing a better understanding of what properties are affecting the clusters and by how much. Figure 11 presents 3 clusters obtained by k-means and hierarchical clustering along the attributes (intrinsic properties) as axes in a 2-dimensional space. The clusters C1, C2, and C3 are indicated by a circle, cross and inverted triangle, respectively. The dimensions of the clusters are determined by the porosity content in the samples, with larger sizes representing higher porosity. The labels attached to the samples correspond to their ash values, while the coloration is based on CPT. The length of the properties along the axes serves as a gauge for the increasing impact on the samples. The relationship between CPT and both oxygen content and volatile matter is that high oxygen content has a high tendency to chemically bind moisture, thereby rendering the surface, highly susceptible to autogenous heating. Based on this information, it is clearly evident that cluster C2 lies where the influence of moisture, porosity, oxygen, and volatile matter is much higher than that of others and has less ash content than C3. Secondly,

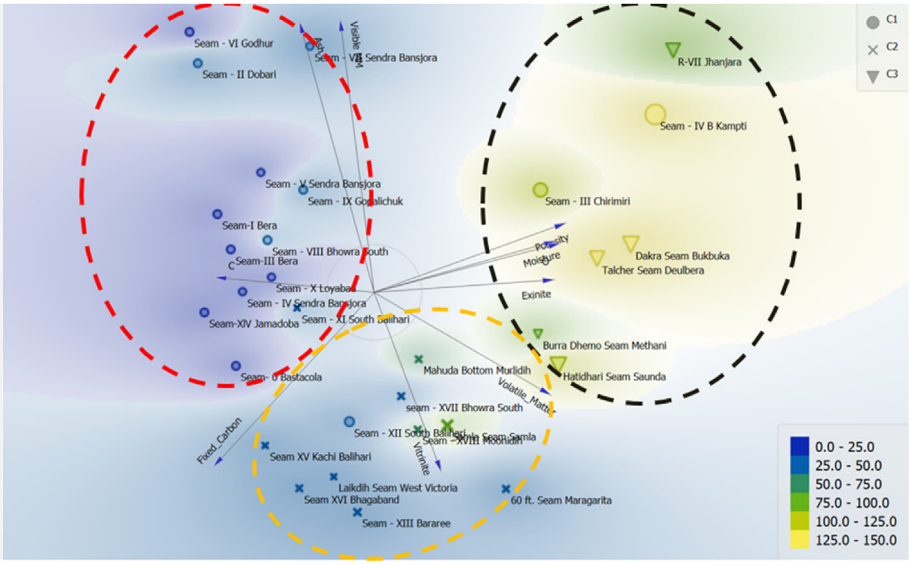

Fig. 11. A 2-D projection of k-means clustering.

the samples in cluster C1 seem to have a higher fixed carbon content, and the ash content is low. This makes C2 the riskiest seams for susceptibility to SC. Cluster C1 represents the riskiest seams for susceptibility to spontaneous combustion because the samples have high volatile matter, less ash content, and higher influence of these intrinsic properties. Finally, C3 has the highest ash and carbon content along with the lowest moisture and volatile matter, hence least prone to SC.

4.4. Visualize multi-dimensional clustering through MDS

MDS is an algorithm that represents the dissimilarities between pairs of objects by reducing the dimensionality of the data in low dimensional space with distances corresponding to dissimilarities. Metric MDS is used for quantitative problems and solves a dissimilarity matrix X, a monotonous function F, and n (n-dimensional data). The algorithm finds the optimal solution for $d_{ij}(X)$ to be the Euclidean distance between rows i and j of X.

$$d_{ij}(X) = \left(\sum_{s=1}^{p} \left(x_{is} - x_{js} \right)^2 \right)^{\frac{1}{2}} \tag{1}$$

$$\sigma^2(X) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} w_{ij} \left(\delta_{ij} - d_{ij}(X) \right)^2 \tag{2}$$

Multi-dimensional scaling (MDS) methods can be used to create a feature-based or spatial representation from similarity data. MDS involves eigen-decomposition of the distance measure as an iterative process that essentially performs triangulation until the points are balanced in a dimensional space. This can be visualized as points connected by springs that are the size of the measured distance. When there is little error in the distances, everything will align correctly. However, if there is a large error, some springs will be under greater tension than others. If you try to take a 3-dimensional object and compress it by flattening it, many of the springs will be stretched or compressed. The basic amount of stretch the system is under is referred to as "stress" or "strain" in MDS terminology, and it serves as a measure of error. Suppose a high-dimensional system can be easily compressed into a lower-dimensional representation (like city layouts on a section of the globe). In that case, it will have low stress and can be well approximated by a 2-dimensional solution. Stress is also a measure of goodness-of-fit and can be used heuristically to assess how appropriate the solution is. The objective is to minimize stress or $\sigma^2(X)$ in the graphical representation, where $d_{ij}(X)$ should be close to the theoretical dissimilarity (δ_{ij}). In MDS, only distances between points matter, and their placement on the map is arbitrary. In Figure 12, the color of the clusters is based on CPT; the large size represents a larger content of volatile matter, and the labels are the content of carbon in the samples. The order of susceptibility seems to be C3 > C4 > C2 > C1.

5. Prediction of susceptibility indices using intrinsic properties using ML techniques

Several spontaneous combustion susceptibility indices (SI) were measured in this study, such as

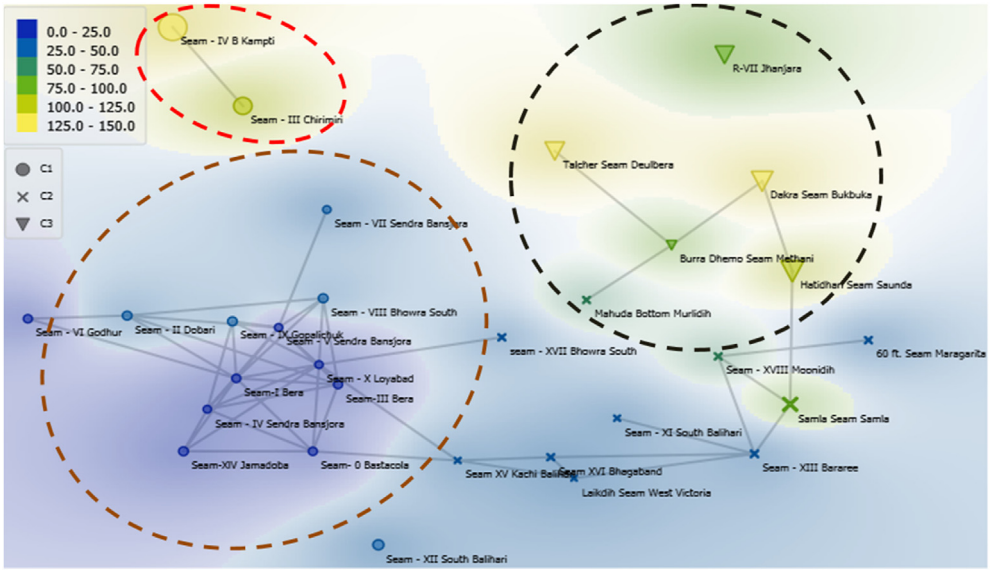

Fig. 12. MDS representation of clustering predicting four clusters (k = 8, resolution = 3, PCA = 6).

crossing point temperature (CPT) [66,67]. A lower CPT value indicates a higher propensity for spontaneous fire hazards. Typically, CPT falls with a rise in the percentage of volatile matter, oxygen, and moisture [68]. Critical air blast (CAB), another index used in this study, measures coal's reactivity to air. The CAB is defined as the minimum rate of air blast required to produce the reactivity of coal to air [69]. The procedure of calculating the CAB of all the coal samples has been previously determined [70]. Differential thermal analysis (DTA), a thermal technique for measuring the heat effects associated with chemical or physical changes in terms of temperature [71,72], was also used. The characteristic or onset temperature obtained through the inflexion point in the thermogram is considered the DTA value.

After tabulating all intrinsic properties, 15 input variables and three susceptibility indexes (SI) were identified as target variables. Three machine learning algorithms were considered for each predictor and output combination to model the relationship between each input variable and SI. The modelling methodology and validation techniques used to select the best model are depicted in Figure 13.

5.1. Model validation

The model's performance is assessed using four metrics: mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and r-squared (R2), as shown in Equations (3–6). To ensure unbiased evaluation, the training and test sets are selected using stratified sampling, with a fixed proportion of data or a fixed number of samples. 3-fold cross-validation is used to evaluate the model's performance, which helps to handle the variance problem of the result set. Grid search with cross-validation is used to find the optimal hyperparameters of the model, resulting in the most accurate predictions. The user defines a range within which each hyper-parameter lies, and grid search builds a combination of models based on different ranges of all hyper-parameters that need tuning. The training data is further split using k-fold cross-validation to find the right combination. Finally, the optimal hyper-parameters are used to test the data from the initial split and compute the performance of the algorithm.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \tag{3}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2 \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2} \tag{5}$$

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{n}(f_i - y_i)^2}{\sum_{i=1}^{n}(\underline{y} - y_i)^2}\right) \tag{6}$$

The following subsections provide the application of the ML algorithms used in this study. These algorithms are effective for small datasets with many features, reducing overfitting and penalizing models with excess features.

5.2. Random forest

The random forest (RF) algorithm is a powerful ensemble learning method that combines multiple decision trees (DT) [73] to create a forest of trees for

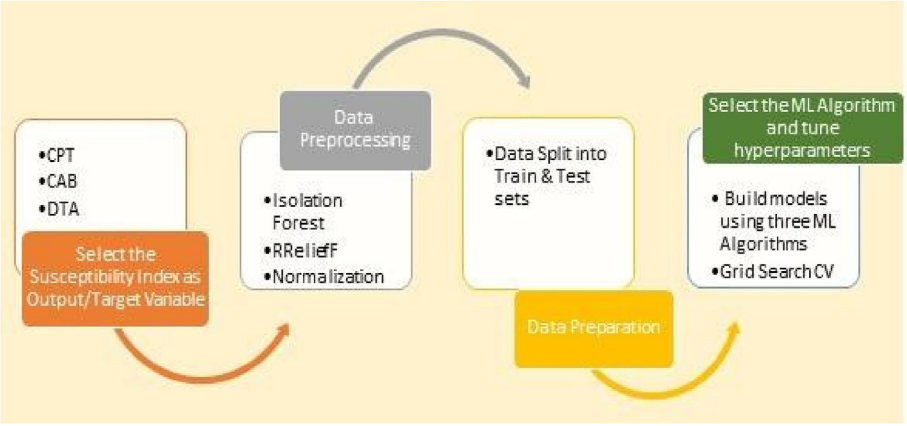

Fig. 13. Methodology for modelling the data for ML regression.

regression and classification tasks [74]. RF is an ensemble learning method that combines multiple decision trees to improve predictive performance, robustness, and generalizability. RF are relatively resilient to outliers and noisy data. Since they aggregate the predictions of multiple trees, the influence of anomalous data points is minimized. Being non-parametric, RF do not assume any underlying distribution for the data. This flexibility allows them to model complex, non-linear relationships between features and the target variable. Each tree is trained on a slightly different set of observations through random sampling with replacement or bootstrapping. The final predictions of the RF are made by averaging the predictions of each tree, as shown in Figure 14. For each subset, a decision tree is built. At each node in the tree, a random subset of features is chosen, and the best split is selected from this subset. This randomness helps to ensure that the trees are diverse and reduces the correlation between them. Once all the trees are built, predictions for a new data point are made by averaging the predictions from all the individual trees. RF uses two key concepts: random sampling of training data points and random subsets of features considered when splitting nodes [75]. The RF hyper-parameters can be tuned using cross-validation or grid search [76]. The following are the user-dependent hyper-parameters used for tuning RF: maximum tree depth, minimum samples required for a node split, number of attributes considered at each split, number of estimators, bootstrap method, and minimum samples in a leaf. A high maximum tree depth improves accuracy in training data but may overfit the data, and the minimum samples before the split determine the minimum number of samples required at each node before making a new split.

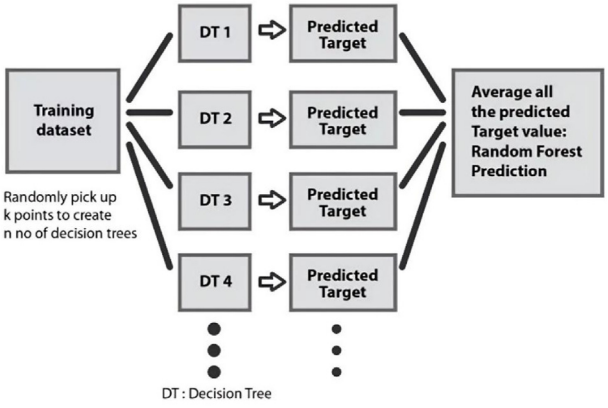

Fig. 14. Methodology of using the random forest in regression problems.

5.3. Support vector machine

Support vector machine (SVM) is a versatile algorithm that can be used for classification or regression tasks, with its distinctive characteristics [77]. It can identify non-linearity in data and produce accurate prediction models [78]. Support vector regression (SVR) aims to minimize the deviation between the predicted and actual values for each training point x. SVR can be mathematically formulated as a convex optimization problem, as shown in Figure 15. The objective of the problem is to find a function $f(x)$ that is as flat as possible while having a maximum deviation of ε from the actual targets for all the training data. The flatness of the

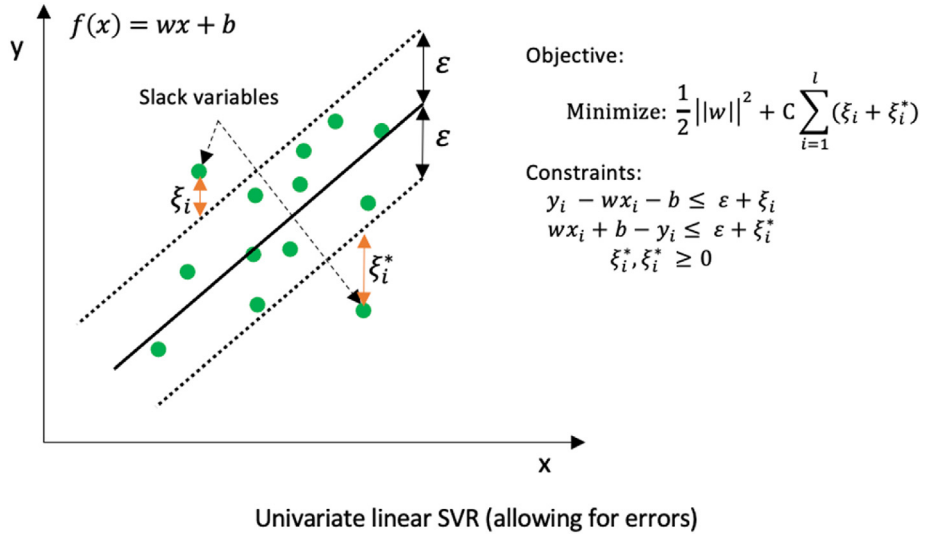

Fig. 15. Methodology for modelling regression problems using SVR.

function implies that it is less sensitive to small changes in the input data, which reduces the risk of overfitting. For linear functions, flatness means having a small value of w in the best-fit function $f(x) = wx + b$. By not penalizing points predicted within a distance of epsilon from the actual value, no training loss is observed (Eq. (3)). The box constraint constant C regulates the penalty imposed on observations that lie outside the epsilon margin (ε), thus avoiding overfitting (regularization). The SVR regularization hyper-parameters include regression cost (C), which controls the trade-off between the flatness of the function and the amount up to which deviations larger than epsilon are tolerated. Higher values of C assign a higher penalty to errors outside the epsilon margin, regression loss (epsilon ε), kernel type, and iteration limit. The kernel specifies the type of function to be used in the algorithm, such as "linear", "poly", "rbf", or "sigmoid" (Eq. (7)).

$$\text{Loss}_{\text{error}} = \{0 \quad \text{if } |y - f(x)| \le \varepsilon \; |y - f(x)| - \varepsilon \quad \text{otherwise} \tag{7}$$

5.4. Elastic net regression

Linear regression models are fundamental tools in statistical analysis and machine learning, often employed to predict an outcome variable based on one or more predictor variables. However, when dealing with high-dimensional data or multicollinearity, standard linear regression may not perform well. To address these issues, regularization techniques such as ridge regression, lasso regression, and elastic net (EN) regression are used. EN can lead to improved predictive accuracy over lasso or ridge alone. The combined penalties help in capturing the true signal more effectively, particularly in scenarios where features are correlated, and some features are not important. Ridge regression uses L2 regularization wherein a penalty is added equal to the square of the magnitude of coefficients to the loss function, to the sums of squared residuals term as shown in Figure 16. Lasso regression adds a penalty equal to the absolute value of the magnitude of coefficients [79]. The elastic net combines the penalties of ridge regression and lasso to create a new method [80]. This allows for an adjustment parameter to control the proportions of L1 and L2 regularizations. One of the primary advantages of EN is its ability to handle multicollinearity among predictor variables. While lasso tends to select one variable from a group of highly correlated variables and ignore the rest, EN can select groups of correlated variables together, which is beneficial

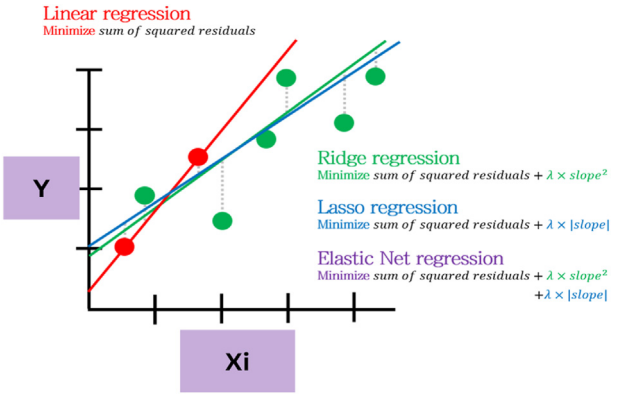

Fig. 16. Types of linear regression model comparison between ridge, lasso and elastic net.

when predictors are correlated. Ridge regression penalizes the sum of squared coefficients (L2 penalty), while lasso regression penalizes the sum of absolute values of the coefficients (L1 penalty). The penalty terms can be tuned via λ to find the best fit for the model. As the value of λ increases, more features are minimized to zero, excluding some features entirely and reducing multicollinearity and model complexity. The alpha parameter finds a compromise between the ridge and lasso, where $\alpha = 0$ corresponds to the ridge and $\alpha = 1$ corresponds to lasso, as shown in Equation (8).

$$L_{\text{elastinet}}(\widehat{\beta}) = \frac{\sum_{i=1}^{n} \left(y_i - x_i^J \widehat{\beta}\right)^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m} \widehat{\beta}_j{}^2 + \alpha \sum_{j=1}^{m} \left|\widehat{\beta}_j\right|\right) \tag{8}$$

6. Prediction of susceptibility indices using intrinsic properties

6.1. Data preparation for modelling

After normalizing the data, it is split into training and testing sets based on a predefined percentage, usually ranging from 70% to 85%, depending on the volume of data. Once the data is split, only the training set is passed to the grid-search algorithm for model parameter tuning [81].

Grid search parameters for modelling RF, SVM, and elastic net are presented in Annexure Table 1. This applies to all exercises of modelling three different SI. For instance, RF has 14,112 candidates based on all combinations, also ten-fold cross-validation makes it a total of 141,120 different models.

6.2. Model to predict CPT

The best hyper-parameters for SVM and random forest models are listed in Annexure Tables 2 and 3, respectively. A hyperbolic tangent function was selected as the kernel for SVM, and six trees were found to be optimal for random forest. One of the decision trees used in random forest to model the relationship is illustrated in Figure 17.

Annexure Table 4 presents the best hyper-parameters selected by EN regression, and Figure 18 shows the values of the regression coefficients. The regularization strength is not too high, and EN has a larger share of ridge than lasso since L1 < L2. Fe, fixed carbon, ash, and carbon have positive values, indicating that higher content of these variables in coal results in higher CPT and lower risk of SC. The most important predictor of CPT is volatile matter, followed by moisture, porosity, vitrinite, exinite, H, and O, all of which have negative coefficient values. The higher their content, the higher the risk of self-heating of coal. None of the input variables is pruned to a coefficient of zero or near zero, thereby indicating their importance in prediction. However, O and ash seem to be the least important variables.

The testing, and ten-fold cross-validation performance measures are given in Tables 4 and 5, respectively. Among all three ML techniques, RF has the highest r-square, lower RMSE, and MSE, while EN has a lower MAE. RMSE penalizes large gaps more harshly than MAE, thus, RF should be used when larger errors are to be avoided, and EN will perform better in case of small data ranges.

Table 4. Testing data performance measures for predicting CPT.

Model CPT Test	MSE	RMSE	MAE	R^2
SVM	0.024	0.154	0.147	0.610
Random forest	0.014	0.117	0.096	0.777
Elastic-net	0.016	0.127	0.095	0.735

6.3. Model to predict DTA_TEMP

According to the RReleifF feature selection method, exinite, and porosity are identified as the most important features for predicting DTA_TEMP, while the intrinsic properties ash, Fe, and H have a comparatively lower impact (Figure 10). The best set of hyperparameters are summarized in Annexure Tables 5–7. The performance metrics for testing and three-fold cross-validation are given in Tables 6 and 7 for SVM, a radial basis function with a degree of 1.5 is selected as the kernel with a high-cost C value of 10. As shown in Figure 19, more trees are required for RF as compared to CPT.

Observing the regularization parameter α, it can be inferred that the extent of regularization is relatively insignificant, as α is a very small value, indicating that regular regression is more applicable

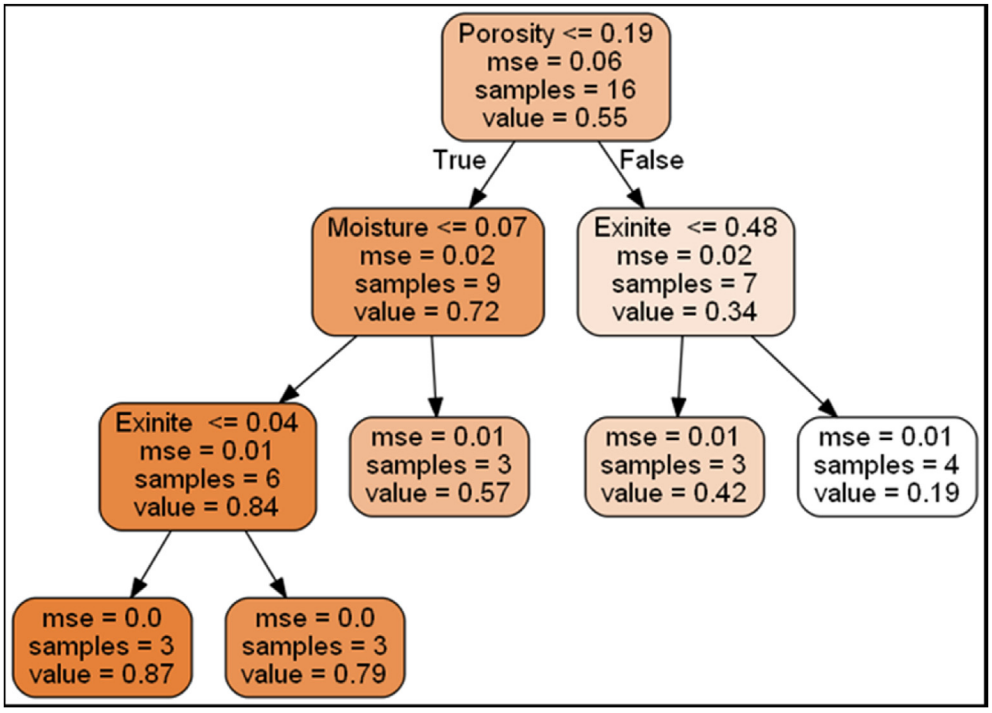

Fig. 17. A single tree from the random forest ensemble employed for CPT estimation.

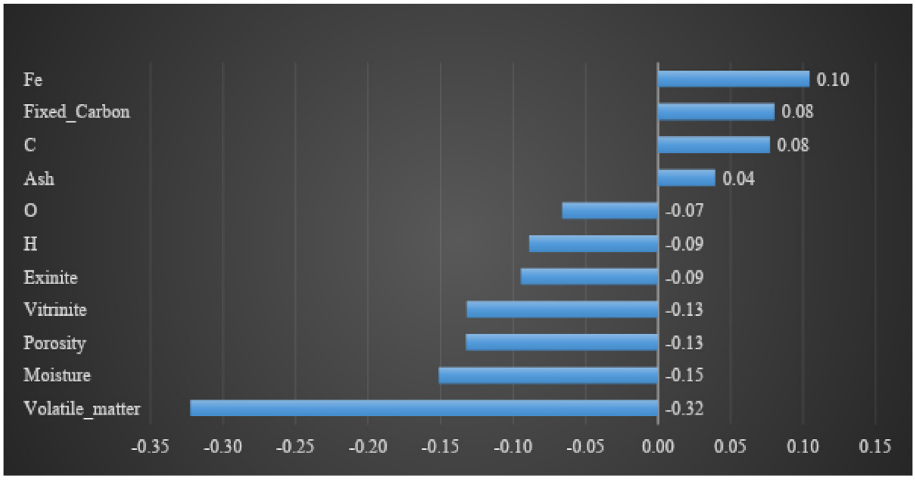

Fig. 18. Elastic net regression coefficients estimate for predicting CPT.

Table 5. Ten-fold cross-validation average performance measures for predicting CPT.

Model CPT Test	MSE	RMSE	MAE	R^2
SVM	0.023	0.153	0.125	0.686
Random forest	0.020	0.142	0.111	0.731
Elastic-net	0.021	0.146	0.111	0.713

Table 6. Testing data performance measures for predicting DTA_TEMP.

Model DTA_TEMP Test	MSE	RMSE	MAE	R^2
SVM	0.003	0.057	0.045	0.794
Random forest	0.004	0.065	0.043	0.729
Elastic-net	0.004	0.061	0.047	0.762

here. Additionally, EN is making some feature selection as the Fe and vitrinite coefficients are close to zero (Figure 20).

All three ML techniques perform equally well in predicting DTA_TEMP. SVM has a higher r-square value compared to the other two techniques, but its MAE is high, indicating that it is not a good predictor for values similar to the given training data, since MAE penalizes small errors.

6.4. Model to predict CAB

Based on the RReleifF feature selection method, exinite, porosity, and volatile matter are found to be the most important features in predicting DTA_TEMP, while O and H have the least impact (Figure 19). Annexure Tables 8–10 provide the best set of hyper-parameters for SVM, RF, and EN, while the performance matrices for testing and three-fold cross-validation are given in Tables 8 and 9 SVM uses a polynomial kernel with a low penalizing cost, while RF gives better results with five trees and a maximum depth of three. EN has a low regularization strength with 80% lasso. Fe, fixed carbon, and vitrinite have zero coefficients in the EN model (Figure 20). However, testing r-square values for all ML techniques used is less than 0.5 for predicting CAB, indicating that it is not linked to the intrinsic properties of coal and is not a good indicator of SC.

Table 7. Ten-fold cross-validation average performance measures for predicting DTA_TEMP.

Model DTA_TEMP 3-fold	MSE	RMSE	MAE	R^2
SVM	0.017	0.130	0.105	0.796
Random forest	0.025	0.160	0.118	0.693
Elastic-net	0.017	0.130	0.084	0.798

7. Discussion of regression results

In this study, RF, SVR, and EN models were used to forecast three different spontaneous combustion indicators (SI) using experimentally measured data. Grid-search cross-validation was used to optimize model parameters and improve the overall goodness of fit. Testing r-square values for predicting DTA_TEMP, CPT, and CAB using three ML techniques are shown in Figure 21, with models sorted in descending order of r-square for each SI. DTA_TEMP and CPT were found to be more closely related to the intrinsic properties of coal than CAB. RF performed better in predicting CPT and CAB, while SVM predicted DTA_TEMP better. Of all compared regression results, RF achieved the highest average overall (DTA & CPT) correlation coefficient (0.88). Analysis revealed DTA and CPT as good predictors of spontaneous combustion and eliminated CAB. The relative importance of each predictor was also determined using RReliefF, RF, and EN to provide a better understanding of which

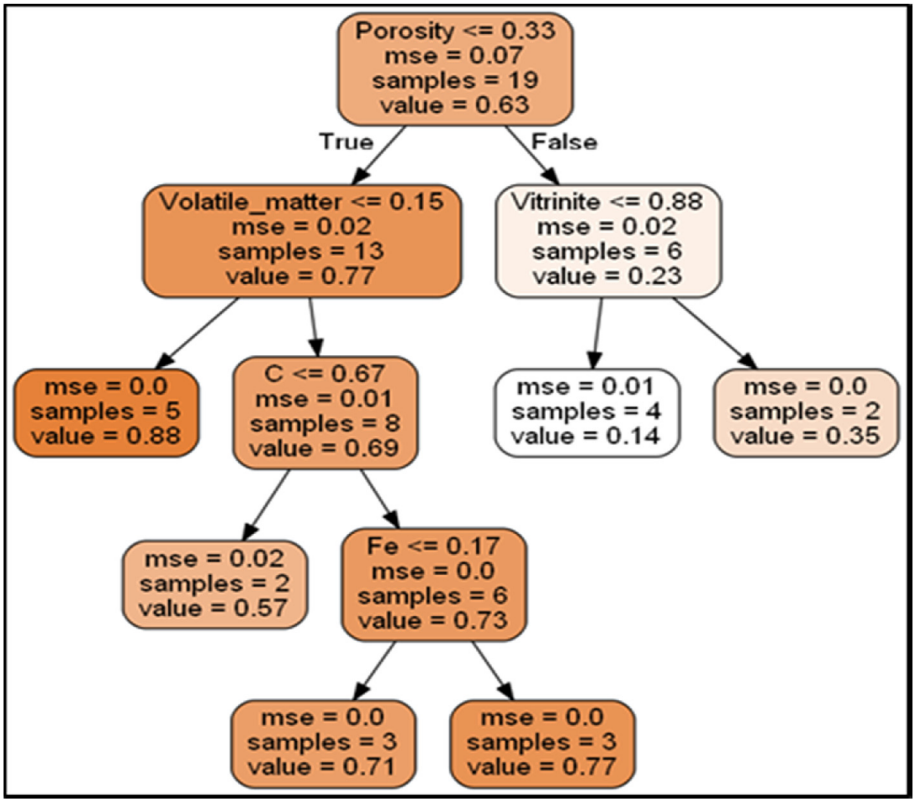

Fig. 19. A random forest tree utilized in the estimation of DTA_TEMP.

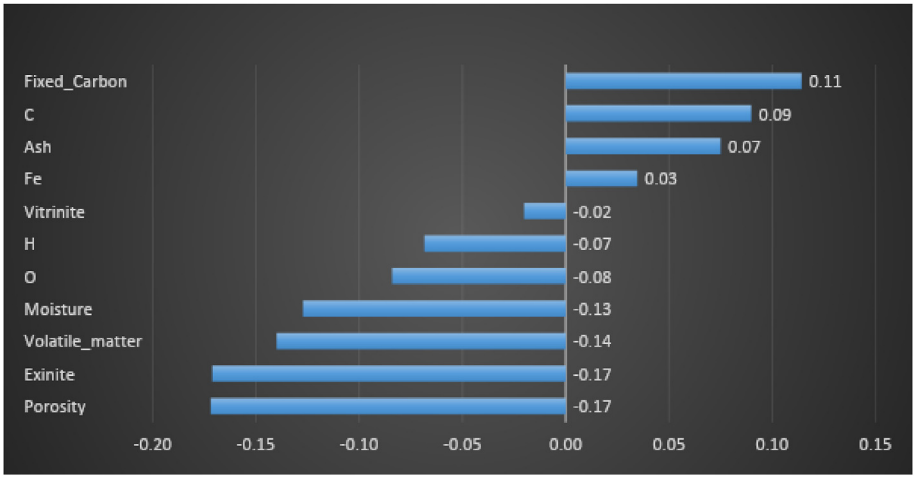

Fig. 20. Elastic net regression coefficients estimate for predicting DTA_TEMP.

Table 8. Testing data performance measures for predicting CAB.

Model CAB Test	MSE	RMSE	MAE	R^2
SVM	0.051	0.226	0.186	0.353
Random forest	0.048	0.219	0.198	0.389
Elastic-net	0.048	0.220	0.200	0.387

Table 9. 10-fold cross-validation data performance measures for predicting CAB.

Model CAB 3 fold	MSE	RMSE	MAE	R^2
SVM	0.039	0.198	0.160	0.423
Random forest	0.040	0.199	0.171	0.415
Elastic-net	0.040	0.201	0.160	0.403

Table 10. Hierarchical and K-means clusters characteristics.

	Clusters/characteristics	C2	C1	C3
1	Spontaneous combustion tendency	Highly combustible	Medium susceptibility	Least susceptible to SC
2	Intrinsic properties	High (moisture, porosity, volatile matter, oxygen)	Less porosity and less ash	High (ash, C, and fixed carbon)
3	Risk assessment	Very risky	Medium risky	Less risky

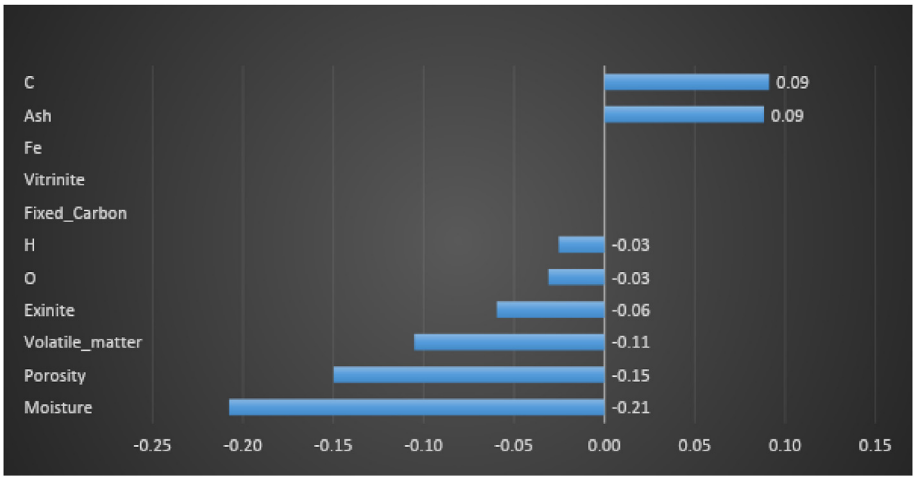

Fig. 21. Elastic net regression coefficients estimate for predicting CAB.

coal properties to focus on. The study suggests that using only 30 data samples might not be optimal for some regression algorithms, and the authors aim to collect more data for comprehensive modelling of Jharia Coalfields in the future. Additionally, not all possible combinations of models with different input variables were developed due to longer training times. The goal was to get a generalized set of predictors or input variables rather than the best possible combination, which may not be realistic given the data size. Hierarchical clustering and k-means methodologies indicate the presence of three clusters in the dataset, delineating high, medium, and low susceptibility to spontaneous combustion risks (Table 10). In contrast, the Louvain method indicates four clusters. All methods are relatively close and not significantly different. As only 30 data samples were used in the study, the authors intend

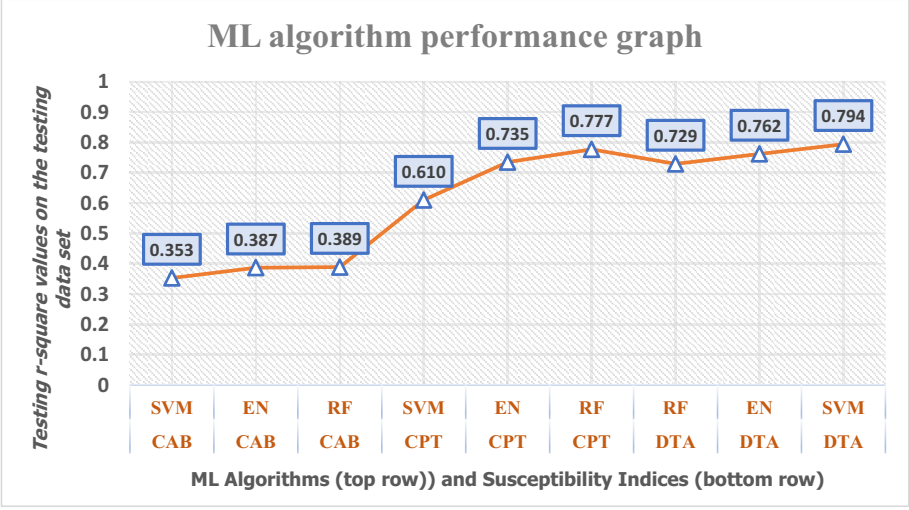

Fig. 22. A comparison of the performance of three ML algorithms to predict DTA_TEMP, CPT, CAB on the testing dataset based on r-square values.

to gather additional data for a more thorough modelling of Jharia Coalfields in future studies).

8. Conclusions

In this study, a new method to determine coal spontaneous combustion hazard grades is proposed. This method is based on the changing concentration of various signature gases during the process of coal spontaneous combustion and includes the establishment of a prediction model for coal spontaneous combustion hazard grades. The findings can be summarized as follows:

i. The Burra Dhemo Seam, Dakra Seam Dakra Bukbuka, Hatidhari Seam Saunda, Kamti Seam IV B, Talcher Seam Deulbera, and Margarita 60 ft. Seams are found to be highly susceptible to spontaneous combustion.
ii. In contrast, Seam XI South Balihari, Seam XV Kachi Balihari, Seam V Sendra Bansora, Seam VI Godhur, Sendra Bansora, Seam Vll, and Seam Vll Bhowra South are among the seams that are least susceptible to spontaneous combustion. These results are consistent with the CPT and DTA results obtained from actual field observations.
iii. CPT and DTA are reliable indicators for the spontaneous combustion propensity of coal, while CAB can be eliminated as a suitable indicator for the same. DTA_TEMP has the highest correlation with the intrinsic properties of coal, followed closely by CPT.
iv. The most important intrinsic properties of coal that affect CPT include volatile matter, moisture, porosity, vitrinite, exinite, H, O (which lower the CPT of coal), and C, FC, Fe (which increase the CPT of coal).
v. On the other hand, DTA_TEMP is better predicted by volatile matter, moisture, porosity, exinite, H, O, C, and FC. There are common parameters between the two, such as volatile matter, moisture, porosity, exinite, H, O, C, and FC.
vi. The highest average r-square on the testing dataset was (0.794) for predicting DTA achieved by SVM and the next highest r-square (0.777) for predicting CPT using EN. CAB had the lowest prediction r-square making it almost uncorrelated with the intrinsic properties of coal (Figure 22).
vii. Optimal hyperparameter configurations suggested for various machine learning techniques can serve as a valuable benchmark for researchers or industry professionals engaged in modeling their datasets [82]. ML techniques can assist in determining the spontaneous heating susceptibility of coal at the exploration stage, thereby helping to develop better coal extraction strategies.

In conclusion, this study presents an innovative approach to identifying and classifying coal seams prone to spontaneous combustion in the Jharia Coalfields, India. By utilizing unsupervised hierarchical, k-means, and multi-dimensional scaling clustering techniques based on the intrinsic properties of 30 coal samples, this method bypasses the need for costly and time-consuming experimental determinations of susceptibility indices such as CPT, CAB, and DTA. Furthermore, machine learning algorithms, including SVR, RF, and EN, effectively elucidate the relationship between coal properties and susceptibility indices. The clustering models successfully categorize coal seams into three risk levels: highly risky, medium risky, and low risk, which were validated by field observations. The authors verified the clustering results with the on-field data, which matches the model results of seams being in different SC susceptibility categories, and that makes it more persuasive. This methodology offers significant practical benefits for mine managers by providing rapid safety risk assessments, thereby enhancing safety and reducing economic losses due to unexpected incidents.

Annexure

https://jsm.gig.eu/cgi/editor.cgi?article=1436&window=additional_files&context=journal-of-sustainable-mining.

Ethical statement

This research complies with all applicable ethical standards and guidelines. By submitting this manuscript, we affirm that the research presented is conducted in accordance with these ethical principles.

Funding body

The research was conducted without financial support from any funding agency, institution, or organization.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] Mukherjee S, Pahari DP. Underground and opencast coal mining methods in India: a comparative assessment. Space and Culture, India 2019;7(1):39–55.

[2] Shao Z, Wang D, Wang Y, Zhong X, Tang X, Hu X. Controlling coal fires using the three-phase foam and water mist techniques in the Anjialing Open Pit Mine, China. Nat Hazards 2015;75:1833–52.

[3] Dontala SP, Reddy TB, Vadde R. Environmental aspects and impacts its mitigation measures of corporate coal mining. Proc Earth Planet Sci 2015;11:2–7.

[4] Singh AK, Kumar J. Fugitive methane emissions from Indian coal mining and handling activities: estimates, mitigation and opportunities for its utilization to generate clean energy. Energy Proc 2016;90:336–48.

[5] Gupta D, Ghersi F, Vishwanathan SS, Garg A. Achieving sustainable development in India along low carbon pathways: macroeconomic assessment. World Dev 2019;123: 104623.

[6] Bava N. Sustainable development in India. In: Sustainable development policy and administration. Routledge; 2017. p. 243–72.

[7] Harriss-White B, Michelutti L, editors. The wild east: criminal political economies in South Asia. UCL Press; 2019.

[8] Kuenzer C, Stracher GB. Geomorphology of coal seam fires. Geomorphology 2012;138(1):209–22.

[9] Künzer C. Remote and in situ mapping of coal fires: case studies from China and India. Coal Peat Fires: A Global Persp Volume 3: Case Stud Coal Fires 2014;3:57–93.

[10] Nimaje DS, Tripathy DP. Characterization of some Indian coals to assess their liability to spontaneous combustion. Fuel 2016;163:139–47.

[11] Chatterjee RS, Thapa S, Singh KB, Varunakumar G, Raju EVR. Detecting, mapping and monitoring of land subsidence in Jharia Coalfield, Jharkhand, India by spaceborne differential interferometric SAR, GPS and precision levelling techniques. J Earth Syst Sci 2015;124:1359–76.

[12] Roy D, Singh G, Seo YC. Coal mine fire effects on carcinogenicity and non-carcinogenicity human health risks. Environ Pollut 2019;254:113091.

[13] Pandey J, Kumar D, Singh VK, Mohalik NK. Environmental and socio-economic impacts of fire in Jharia coalfield, Jharkhand, India: an appraisal. Curr Sci 2016:1639–50.

[14] Saffari A, Sereshki F, Ataei M. The simultaneous effect of moisture and pyrite on coal spontaneous combustion using CPT and R70 test methods. Rudarsko-Geolosko-Naftni Zb 2019;34(3).

[15] Mohalik NK, Mishra D, Ray SK, Varma NK, Khan AM, Sahay N. Laboratory investigation to assess spontaneous combustion/fire during extraction of thick coal seam. J Inst Eng: Series D 2019;100:229–42.

[16] Uludag S. A visit to the research on Wits-Ehac index and its relationship to inherent coal properties for Witbank Coalfied. J S Afr Inst Min Metall 2007;107(10):671–9.

[17] Avila C, Wu T, Lester E. Estimating the spontaneous combustion potential of coals using thermogravimetric analysis. Energy Fuels 2014;28(3):1765–73.

[18] Rambha RV. Spontaneous combustion of coal and correlation with its intrinsic properties using adiabatic oxidation method. In: International conference on emerging trends in engineering (ICETE) emerging trends in smart modelling systems and design. Springer International Publishing; 2020. p. 278–84.

[19] Wang B, Lv Y, Liu C. Research on fire early warning index system of coal mine goaf based on multi-parameter fusion. Sci Rep 2024;14(1):485.

[20] Pattanaik DS, Behera P, Singh B. Spontaneous combustibility characterisation of the chirimiri coals, Koriya district, Chhatisgarh, India. Int J Geosci 2011;2(3):336.

[21] Ray SK, Panigrahi DC, Udayabhanu G, Saxena VK. Assessment of spontaneous heating susceptibility of Indian coals—a new approach. Energy Sources, Part A Recovery, Util Environ Eff 2016;38(1):59–68.

[22] Karmakar NC, Banerjee SP. A comparative study on CPT index, Polish Sz index and Russian U-index of susceptibility of coal to spontaneous combustion. J MGMI 1989;86:109–29.

[23] Panigrahi DC, Ojha A, Saxena NC, Kejriwal BK. A study of coal oxygen interaction by using Russian U-index and its correlation with basic constituents of coal with particular reference to Jharia coalfield. 1997.

[24] Xuyao Q, Wang D, Milke JA, Zhong X. Crossing point temperature of coal. Min Sci Technol 2011;21(2):255–60.

[25] Nimaje DS, Tripathy DP, Nanda SK. Development of regression models for assessing fire risk of some Indian coals. Int J Intell Syst Appl 2013;2:52–8.

[26] Altman DG, Bland JM. Parametric v non-parametric methods for data analysis. BMJ 2009;338.

[27] Wang C, Du Y, Deng Y, Zhang Y, Deng J, Zhao X, et al. Study on spontaneous combustion characteristics and early warning of coal in a deep mine. Fire 2023;6(10):396.

[28] Wu K, Yao Q, Chen Y, Zhao P, Xi C, Zhao Y, et al. Dependence evaluation of factors influencing coal spontaneous ignition. Energy Sci Eng 2023;11(10):3738–50.

[29] Zeng S, D'Hyon S, Widzyk-Capehart E. Data-driven revision of coal spontaneous combustion risk classification for prevailing methods used in Australian mining industry. In: IOP conference series: earth and environmental science. vol. 1295. IOP Publishing; 2024. p. 012002. No. 1.

[30] Mishra A, Gupta SK. Intelligent classification of coal seams using spontaneous combustion susceptibility in IoT paradigm. International Journal of Coal Preparation and Utilization; 2023. p. 1–23.

[31] Said KO, Onifade M, Lawal AI, Githiria JM. An artificial intelligence-based model for the prediction of spontaneous combustion liability of coal based on its proximate analysis. Combust Sci Technol 2021;193(13):2350–67.

[32] Wang K, Li K, Du F, Zhang X, Wang Y, Sun J. Research on prediction model of coal spontaneous combustion temperature based on SSA-CNN. Energy 2024;290:130158.

[33] Said KO, Onifade M, Lawal AI, Githiria JM. Computational intelligence-based models for predicting the spontaneous combustion liability of coal. Int J Coal Prep Utiliz 2022;42(6): 1626–50.

[34] Pei Q, Jia Z, Liu J, Wang Y, Wang J, Zhang Y. Prediction of coal spontaneous combustion hazard grades based on fuzzy clustered case-based reasoning. Fire 2024;7(4):107.

[35] Qi Y, Xue K, Wang W, Cui X, Liang R. Prediction model of borehole spontaneous combustion based on machine learning and its application. Fire 2023;6(9):357.

[36] Wang W, Liang R, Qi Y, Cui X, Liu J. Prediction model of spontaneous combustion risk of extraction borehole based on PSO-BPNN and its application. Sci Rep 2024;14(1):5.

[37] Lawal AI, Onifade M, Abdulsalam J, Aladejare AE, Gbadamosi AR, Said KO. On the performance assessment of ANN and spotted hyena optimized ANN to predict the spontaneous combustion liability of coal. Combust Sci Technol 2022;194(7):1408–32.

[38] Li S, Xu K, Xue G, Liu J, Xu Z. Prediction of coal spontaneous combustion temperature based on improved grey wolf optimizer algorithm and support vector regression. Fuel 2022;324:124670.

[39] Guo Q, Ren W, Lu W. A method for predicting coal temperature using CO with GA-SVR model for early warning of the spontaneous combustion of coal. Combust Sci Technol 2022;194(3):523–38.

[40] Kaymakçi E, Didari V. Relations between coal properties and spontaneous combustion parameters. Turk J Eng Environ Sci 2002;26(1):59–64.

[41] Zhang Q, Li HG, Li H. An improved least squares SVM with adaptive PSO for the prediction of coal spontaneous combustion. Math Biosci Eng 2019;16(4):3169–82.

[42] Shukla US, Mishra DP, Mishra A. Prediction of spontaneous combustion susceptibility of coal seams based on coal

intrinsic properties using various machine learning tools. Environ Sci Pollut Control Ser 2023;30(26):69564–79.
[43] Jena SS. Investigation into spontaneous combustion characteristics of some Indian coals and correlation study with their intrinsic properties (Doctoral dissertation). 2011.
[44] Onifade M, Genc B. Modelling spontaneous combustion liability of carbonaceous materials. Int J Coal Sci Technol 2018; 5:191–212.
[45] Onifade M, Genc B. Comparative analysis of coal and coal-shale intrinsic factors affecting spontaneous combustion. Int J Coal Sci Technol 2018;5:282–94.
[46] Michalski SR, Custer Jr ES, Munsi PL. Investigation of the Jharia coalfield mine fires—India. Am Soc Min Reclam 1997: 211–23.
[47] Shaheen F, Krishna AP, Rathore VS. Delineation of mine fire pockets in Jharia Coalfield, India, using thermal remote sensing. In: Advances in computational intelligence: proceedings of second international conference on computational intelligence 2018. Singapore: Springer; 2020. p. 215–27.
[48] Riyas MJ, Syed TH, Kumar H, Kuenzer C. Detecting and analyzing the evolution of subsidence due to coal fires in Jharia Coalfield, India using Sentinel-1 SAR data. Rem Sens 2021;13(8):1521.
[49] Barik S, Biswaranjan R. Breathing fire on hot coals at Jharia coal fields. The Hindu; 2020, January 5. Retrieved from, https://www.thehindu.com/news/national/other-states/breathing-fire-on-hot-coals-at-jharia-coal-fields/article30479245.ece.
[50] Mukhopadhyay A. Jharia Basin structure and tectonics. In: Developments in structural geology and tectonics. vol. 4. Elsevier; 2019. p. 45–54.
[51] Stracher GB, Prakash A, Sokol EV, editors. Coal and peat fires: a global perspective: volume 1: coal-geology and combustion (Vol. 1). Elsevier.Fox CS. The Jharia Coalfield, (1930). vol. 56. Geological Survey of India Memoirs; 2010. p. 248.
[52] Agarwal S, Singh A, Sharma P. Machine learning based prediction of spontaneous combustion susceptibility of coal using its intrinsic properties: a safe smart and sustainable mining of coal approach. In: Asian mining congress. Cham: Springer Nature Switzerland; 2023, October. p. 145–52.
[53] Indian Standard: 1350. Methods for test of coal and coke. 2006, July.
[54] Indian Standard: 9127. Methods of petrographic analysis of coal. 1979.
[55] Indian Standard: 1979. Specification of high-test pipeline. 1985.
[56] Nimaje DS, Tripathy DP. Characterization of some Indian coals to assess their liability to spontaneous combustion. Fuel 2016;163:139–47.
[57] Aich S, Behera D, Nandi BK, Bhattacharya S. Relationship between proximate analysis parameters and combustion behaviour of high ash Indian coal. Int J Coal Sci Technol 2020;7:766–77.
[58] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3(Mar):1157–82.
[59] Urbanowicz RJ, Meeker M, La Cava W, Olson RS, Moore JH. Relief-based feature selection: introduction and review. J Biomed Inf 2018;85:189–203.
[60] Fan J, Li R. Statistical challenges with high dimensionality: feature selection in knowledge discovery. arXiv preprint math/0602133 2006. https://doi.org/10.48550/arXiv.math/0602133.
[61] Grira N, Crucianu M, Boujemaa N. Unsupervised and semi-supervised clustering: a brief survey. Rev Mach Learn Techn Process Multimedia Content 2004;1:9–16.
[62] Agarwal S, Dagli CH, Pape LE. Computational intelligence based complex adaptive system-of-system architecture evolution strategy. In: Complex systems design & management: proceedings of the sixth international conference on complex systems design & management, CSD&M 2015. Springer International Publishing; 2016. p. 119–32.
[63] Freeman L. Displaying hierarchical clusters. INSNA Connections 1994;17(2):46–52.
[64] Hout MC, Papesh MH, Goldinger SD. Multidimensional scaling. Wiley Interdiscip Revi: Cognit Sci 2013;4(1):93–103.
[65] Agarwal S. Computational intelligence based complex adaptive system-of systems architecture evolution strategy. Doctoral dissertation, Missouri University of Science and Technology); 2015.
[66] Panigrahi DC, Sahu HB. Classification of coal seams with respect to their spontaneous heating susceptibility—a neural network approach. Geotech Geol Eng 2004;22:457–76.
[67] Mohalik NK, Lester E, Lowndes IS. Development of a petrographic technique to assess the spontaneous combustion susceptibility of Indian coals. Int J Coal Prep Utiliz 2020; 40(3):186–209.
[68] Sahu HB, Panigrahi DC, Mishra NM. Assessment of spontaneous heating susceptibility of coal seams by experimental techniques-a comparative study. 2005.
[69] Panigrahi DC, Saxena VK, Udaybhanu G. Research project report: development of handy method of coal categorisation and prediction of spontaneous fire risk in mines. Dep Mining Eng ISM, Dhanbad 1999;1:15–23.
[70] Panigrahi DC, Sahu HB. Development of a new method for the assessment of spontaneous heating susceptibility of coal. IE (I) J MN 2005;85:42–5.
[71] Deng J, Zhao J, Zhang Y, Huang A, Liu X, Zhai X, et al. Thermal analysis of spontaneous combustion behavior of partially oxidized coal. Process Saf Environ Protect 2016;104: 218–24.
[72] Bhoi DS. Assessment of spontaneous combustion of some Indian coals using differential thermal analysis (doctoral dissertation). 2016.
[73] Breiman L. Random forests. Mach Learn 2001;45:5–32.
[74] Kotsiantis SB. Decision trees: a recent overview. Artif Intell Rev 2013;39:261–83.
[75] Biau G. Analysis of a random forests model. J Mach Learn Res 2012;13(1):1063–95.
[76] Scornet E, Biau G, Vert JP. Consistency of random forests. 2015.
[77] Martin M. On-line support vector machine regression. In: European conference on machine learning. Berlin, Heidelberg: Springer Berlin Heidelberg; 2002. p. 282–94.
[78] Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput 2004;14:199–222.
[79] Zou H, Hastie T. Regularization and variable selection via the elastic net. J Roy Stat Soc B Stat Methodol 2005;67(2):301–20.
[80] Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc B Stat Methodol 1996;58(1):267–88.
[81] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13(2).
[82] Agarwal S, Saferpour HR, Dagli CH. Adaptive learning model for predicting negotiation behaviors through hybrid k-means clustering, linear vector quantization and 2-Tuple fuzzy linguistic model. Proc Comput Sci 2014;36:285–92.