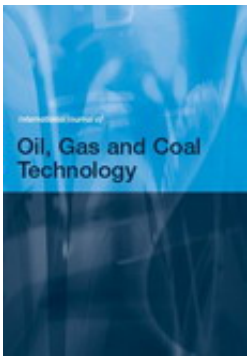


International Journal of Oil, Gas and Coal Technology

ISSN online: 1753-3317 - ISSN print: 1753-3309
https://www.inderscience.com/ijogct

Integration of computational algorithms in enhancing geological modelling and drilling efficiency in coal mine exploration

Xingen Ma, Yubo Song, Qinghai Zou, Ying He, Bo Pan, Sixuan Ye, Guang Yang, Biwen Li, Fuhong Li, Kamel Bou-Hamdan

DOI: 10.1504/IJOGCT.2026.10079168

Article History:

Received:	26 November 2025
Last revised:	20 April 2026
Accepted:	08 May 2026
Published online:	25 June 2026

Integration of computational algorithms in enhancing geological modelling and drilling efficiency in coal mine exploration

Xingen Ma*

Huaneng Coal Technology Research Co., Ltd.,
Beijing 100070, China
and
China University of Mining and Technology,
Xuzhou 22111, China
Email: maxingen.cumtb@foxmail.com.
*Corresponding author

Yubo Song

CCTEG Xi'an Research Institute (Group) Co., Ltd.,
Xi'an 710077, China
Email: songyubo@cctegxian.com

Qinghai Zou

Walnut Valley Coal Mine,
Huaneng Qingyang Coal Power Co., Ltd.,
Qingyang 745300, China
Email: 1596795232@qq.com

Ying He

CCTEG Xi'an Research Institute (Group) Co., Ltd.,
Xi'an 710077, China
Email: heying@cctegxian.com

Bo Pan

Huaneng Coal Technology Research Co., Ltd.,
Beijing 100070, China
and
China University of Mining and Technology,
Xuzhou 22111, China
Email: bo_pan@chng.com.cn

Sixuan Ye

CCTEG Xi'an Research Institute (Group) Co., Ltd.,
Xi'an 710077, China
Email: yesixuan@cctegxian.com

Guang Yang

Huaneng Coal Technology Research Co., Ltd.,
Beijing 100070, China
Email: 2767421435@qq.com

Biwen Li

CCTEG Xi'an Research Institute (Group) Co., Ltd.,
Xi'an 710077, China
Email: libiwen@cctegxian.com

Fuhong Li

Walnut Valley Coal Mine,
Huaneng Qingyang Coal Power Co., Ltd.,
Qingyang 745300, China
Email: 2430569717@qq.com

Kamel Bou-Hamdan

Institute of Geoenergy Engineering,
Heriot-Watt University,
Edinburgh EH14 4AS, UK
Email: k.bouhamdan@hw.ac.uk

Abstract: Coal calorific value calculation is crucial for power planning and mine exploration. Traditional lab techniques and empirical calculations often fail with high variability or limited data. We developed a machine learning framework for small-sample scenarios, combining compositional characteristics, data augmentation, and Bayesian hyperparameter adjustment to improve prediction accuracy. Four regression models (ANN, SVR, decision tree, and LightGBM) were trained on proximate, ultimate, and petrographic features to predict coal calorific value. Among these, the LightGBM model achieved the highest predictive performance with a test (R^2 approximately 0.93) and the lowest error (RMSE approximately 0.19), outperforming the ANN (R^2 approximately 0.91) and other models. Fixed carbon (FC) and volatile matter (VM) were the key predictors of calorific value, aligning with domain knowledge and model interpretability. The improved data-driven approach reliably estimates coal energy content from small samples, enabling evidence-based geological modelling and better drilling decisions for increased exploration efficiency. [Received: January 16, 2026; Accepted: May 8, 2026]

Keywords: calorific value prediction; light gradient boosting machine; LightGBM; artificial neural network; ANN; feature engineering; data augmentation; hyperparameter optimisation; coal characterisation; coal quality prediction.

Reference to this paper should be made as follows: Ma, X., Song, Y., Zou, Q., He, Y., Pan, B., Ye, S., Yang, G., Li, B., Li, F. and Bou-Hamdan, K. (2026) 'Integration of computational algorithms in enhancing geological modelling and drilling efficiency in coal mine exploration', *Int. J. Oil, Gas and Coal Technology*, Vol. 39, No. 6, pp.1–27.

Biographical notes: Xingen Ma is the Deputy Director of the Green Mining Department at Huaneng Coal Technology Research Co., Ltd., Beijing, China. He received his PhD in Geotechnical Engineering from China University of Mining and Technology (Beijing) in 2019. His research interests include rock burst control, mine pressure disaster prevention, roadway rock control, and coal mine dynamic disaster prevention. He published over 30 journal papers and was granted over 60 invention patents. He is a Science and Technology Service Expert of the China National Coal Association, a member of the China Coal Society, and a think tank expert of the Zhongguancun Green Mine Industry Alliance.

Yubo Song is the General Manager and Associate Research Fellow at the Meng-Shaan Business Division of CCTEG Xi'an Research Institute (Group) Co., Ltd. (Xi'an, 710077, China). He received his Master's in Instrument Science and Technology from Xi'an Jiaotong University in 2009. He has 17 years of working experience in coal mine disaster control, and his research interests focus on underground coal mine disaster prevention and control. He has published 10 academic papers in various journals, including the representative publication Research on Borehole Forming Technology of Large-Diameter High-Level Directional Boreholes in Complex Roof Strata of Jincheng Mining Area.

Qinghai Zou is a Chief Engineer at Hetaoyu Coal Mine, Huaneng Qingyang Coal and Electricity Co., Gansu Province, China. He received his Mining Engineering degree from Xi'an University of Science and Technology in 2016. He is engaged in mine management and disaster prevention and control. He received awards including Gansu Coal Industry's Second Prize for Scientific Achievement, China Occupational Safety and Health Association's Third Prize for 'Research on reinforcement and support technology for high-stress soft-rock roadways', and China National Coal Association's First Prize for 'Key technologies for identification of risk sources and hydraulic fracturing prevention and control of rock burst in Ningzheng Mining Area'.

Ying He is the Director and Engineer at the Geological Branch of CCTEG Xi'an Research Institute (Group) Co., Ltd. (Xi'an, 710077, China). She received her Master's in Geological Engineering from Xi'an University of Science and Technology in 2014. She has 12 years of working experience in coal mine disaster control, and her research interests focus on underground coal mine disaster prevention and control. She has published four academic papers in various journals.

Bo Pan is the General Manager and Senior Engineer at Huaneng Coal Technology Research Co., Beijing, China. He received his Bachelor's and Master's in Mining Engineering from China University of Mining and Technology in 2008 and 2017, respectively. He previously served as the Deputy Director of Audit Service Department II at China Huaneng Group Co.,

Director of the Human Resources Department of Huaneng Coal Industry Co., and held technical and management positions at Huaneng Yimin Coal and Electricity Co. His research includes green mining, open-pit mining technology, and coal production management. He published over 20 papers and holds over 40 invention patents.

Sixuan Ye is the Director and Engineer at the Northwest Business Division of CCTEG Xi'an Research Institute (Group) Co., Ltd. (Xi'an, 710077, China). He received his Master's in Geological Engineering from China University of Mining and Technology in 2014. He has 12 years of working experience in coal mine disaster control, and his research interests focus on underground coal mine disaster prevention and control. He has published six academic papers in various journals, including the representative publication research on trajectory prediction of near-horizontal directional boreholes in tunnels based on neural networks.

Guang Yang is the Deputy General Manager and Senior Engineer at Huaneng Coal Technology Research Co., Beijing, China. He received his PhD in Management in Technology Economics and Management from North China Electric Power University. He has experience in energy strategy, coal industry planning, power plant operation management, and coal technology research management. He served as the Deputy Director of Dalate Power Plant, Director of Budget Management Division at Huaneng International Power Development Corporation, and held positions at China Huaneng Group Co. His interests include coal technology innovation, green energy development, coal mine safety, and efficient enterprise operation. He holds over 30 invention patents.

Biwen Li is the Deputy Director and Engineer at the Northwest Business Division of CCTEG Xi'an Research Institute (Group) Co., Ltd. (Xi'an, 710077, China). He received his Master's in Geological Engineering from China University of Mining and Technology in 2021. He has five years of working experience in coal mine disaster control, and his research interests focus on underground coal mine disaster prevention and control. He has published four academic papers in various journals, including the representative publication optimisation of trajectory design method for bedding directional boreholes in complex strata.

Fuhong Li is the Director of the Rock Burst Prevention and Control Department at Hetaoyu Coal Mine, Huaneng Qingyang Coal and Electricity Co., in Gansu Province, China. He received his Mining Engineering degree from Xi'an University of Science and Technology in 2019. He works on rock burst prevention and control and mine construction. His interests include rock burst prevention, mine construction, deep mine disaster control, roadway support, and mine safety management. He received the Third Prize Science and Technology Award of China Occupational Safety and Health Association and the First Prize Science and Technology Progress Award of China National Coal Association.

Kamel Bou-Hamdan holds a PhD and MSc in Petroleum Engineering and a BEng in Electrical Power and Control Engineering. He has held academic appointments at several universities, contributing to teaching, curriculum development, accreditation, and research supervision. His research focuses on subsurface energy systems, enhanced oil recovery, CCUS, hydrogen and CO_2 storage, nanomaterials, and energy-transition applications, with relevance to coalbed methane, unconventional resources, and low-carbon coal technologies.

He has published peer-reviewed research and contributes to international editorial activities, including *Scientific Reports*, *PLOS ONE*, *Symmetry*, and the *International Journal of Coal Science & Technology*.

1 Introduction

Coal is a key component of the global energy mix in many places, particularly where electrical networks continue to rely on thermal generation for stable and economical baseload supply (Gasparotto and Da Boit Martinello, 2021). Because the calorific value of coal determines the amount of energy that can be extracted per unit mass and thus controls both the operational efficiency and economic viability of combustion systems, its accurate determination is crucial to power generation planning and fuel procurement strategies (Liu and Lv, 2020). In addition to its utility in end-use applications, calorific value estimation is important in the exploration and resource appraisal stages. It facilitates drilling decisions, mining development methods, and long-term production scheduling (Bou-Hamdan, 2021; Kumar et al., 2025; Nautiyal and Mishra, 2023). Therefore, there is continuous interest in developing reliable prediction algorithms that can accurately estimate coal energy content from basic compositional and petrographic characteristics (Channiwala and Parikh, 2002; Mesroghli et al., 2009).

Traditional coal characterisation approaches rely on laboratory-based proximal and ultimate analysis, as well as empirical relationships that relate coal's elemental and mineral matter composition to its heating value (Parikh et al., 2005; Speight, 2015). While these standard techniques are well established, their applicability becomes reduced when faced with highly variable coal seams or early-stage exploration wells with limited sample frequency (Mesroghli et al., 2009). Such deterministic methods frequently presume linear correlations and neglect intricate interactions among coal constituents, resulting in significant disparities in quality assessments. When there is significant subsurface heterogeneity, geological and drilling-based evaluations may suffer from interpretational subjectivity and poor predictive capacity (Heriawan and Koike, 2008). As a result, decision makers in coal exploration and production scenarios are increasingly seeking analytical techniques that can capture nonlinear behaviour and deliver more robust predictions when data is scarce (Büyükkanber et al., 2023; Erik and Yilmaz, 2011).

In recent years, machine learning approaches have been widely applied to coal quality forecasting and characterisation tasks because of their capacity to extract patterns from multivariate datasets without imposing limiting functional assumptions (Bui et al., 2019; Chelgani, 2021; Hira et al., 2025). Several learning algorithms, including support vector machines, artificial neural networks (ANNs), random forest methods, and gradient boosting models, have shown promising results in estimating coal rank, ash content, sulphur concentration, and calorific value from laboratory and geochemical measurements (Feng and Mao, 2015; Hira et al., 2025; Khandelwal and Singh, 2010; Mondal et al., 2022). These researches have shown that data-driven frameworks can outperform standard statistical models, particularly when there are complicated connections between coal quality measures (Tan et al., 2015). However, most contributions in this domain have been based on relatively large datasets, with minimal investigation of conditions that reflect the practical restrictions of early exploration

missions where just a few samples are available for model training and validation (Büyükkanber et al., 2023).

There are still large methodological gaps in spite of these improvements. Notably, present contributions seldom use systematic data augmentation strategies to improve learning capacity when samples are limited and often underutilise designed compositional properties [e.g., fuel ratio, fixed carbon (FC), carbon-to-hydrogen ratio] (Akhtar et al., 2017; Büyükkanber et al., 2023). Additionally, rather than being automated by strong Bayesian frameworks that might increase convergence efficiency and prediction robustness, hyperparameter tweaking is often deterministic or manually modified in earlier research (Akiba et al., 2019). These restrictions limit the model's applicability to real-world coal exploration scenarios with limited data availability and significant lithological heterogeneity (Büyükkanber et al., 2023; Heriawan and Koike, 2008).

To increase the prediction accuracy for coal calorific value estimations, this study proposes a data-driven regression system that integrates feature engineering, random interpolation-based data augmentation, and Optuna-driven automatic hyperparameter tuning. Proximate, ultimate, and petrographic data were used to train and evaluate four example machine-learning algorithms: ANN, SVR, DT, and light gradient boosting machine (LightGBM). By demonstrating how an enhanced and optimised machine learning pipeline can precisely predict the calorific value from sparse laboratory data, the developed workflow offers a scalable and computationally efficient tool for resource assessment, drilling decision support, and operational planning in coal exploration settings.

2 Data collection and methodology

2.1 Dataset description and pre-processing

The dataset used in this investigation was obtained from the United States Department of Energy coal characterisation archives, which are managed by the EMS Energy Institute at Pennsylvania State University. The database contains extensive analytical data from 22 coal mines, including proximal and ultimate analyses and petrographic measures. This study examined nine significant input factors, including ash content, sulphur content, carbon content, hydrogen content, volatile matter (VM), vitrinite percentage, vitrinite reflectance, and ASTM rank, with the calorific value serving as the prediction aim. To ensure uniformity in interpretation and conformity with conventional coal assessment processes, all results were reported as dry or mineral matter-free. By isolating the organic fraction of the coal and eliminating the diluting effects of both natural moisture and inorganic material, reporting data on a dry, mineral matter-free (dmmf) basis creates a more consistent basis for evaluating the energy content of coal samples.

Table 1 contains an overview of the variables evaluated in this work, as well as their related definitions and analytical underpinnings, since these indicators are crucial compositional and petrographic descriptors utilised in traditional coal quality evaluation. The category labels that were previously included in the ASTM rank field were encoded using the natural coal rank system, assigning anthracite a value of one and lignite a value of ten. This ensured that the ordinal link between coal ranks was maintained throughout model training while preventing the installation of an artificial numerical structure that may have an impact on model behaviour. Before any preprocessing or augmentation, the

statistical distribution of the raw datasets is shown in Figures 1 and 2 using scaled box plot representations, highlighting the intrinsic variability and central tendency of each feature. The inclusion of these visual summaries emphasises the natural dispersion inherent in the original measurements and guarantees that the ensuing data conditioning processes are consistent with the coal samples' underlying physical features.

Table 1 Description of the modelling data used

Variable	*Units*	*Basis*	*Values*	
Feed inputs			*Training*	*Testing*
ASTM rank	-	-	1–10	2–10
Ash	wt.%	dry	3.9–23.3	5–20
Sulphur (S)	wt.%	dry	0.40–5.53	0.62–4.83
Carbon (C)	wt.%	dmmf	74.1–91.5	74.0–91.8
Hydrogen (H)	wt.%	dmmf	4.1–6.3	4.5–5.8
Volatile matter (V.M.)	wt.%	dmmf	3.9–64.6	19.0–49.2
Vitrinite (Vit.)	wt.%	-	30–95	71–89
Vitrinite reflectance (VRo)	%	-	0.23–5.19	0.25–1.68
Fixed carbon (FC)	wt.%	-	27.4–84.9	39.3–76.0
Carbon to hydrogen ratio (C/H)		-	12.3–22.3	14.0–20.4
Fuel ratio (F.R.)		-	0.42–21.77	0.8–4.0
		Response		
Calorific value (C.V.)	Btu/lb	dmmf	12,269–15,831	12,370–15,980

Figure 1 Box plot diagram for the original DESC training dataset (see online version for colours)

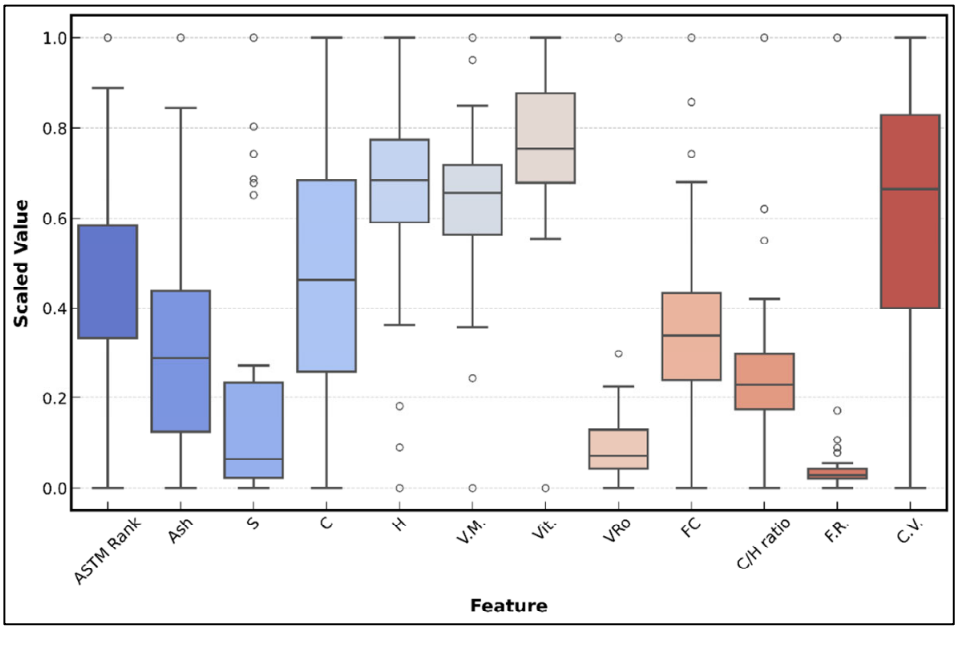
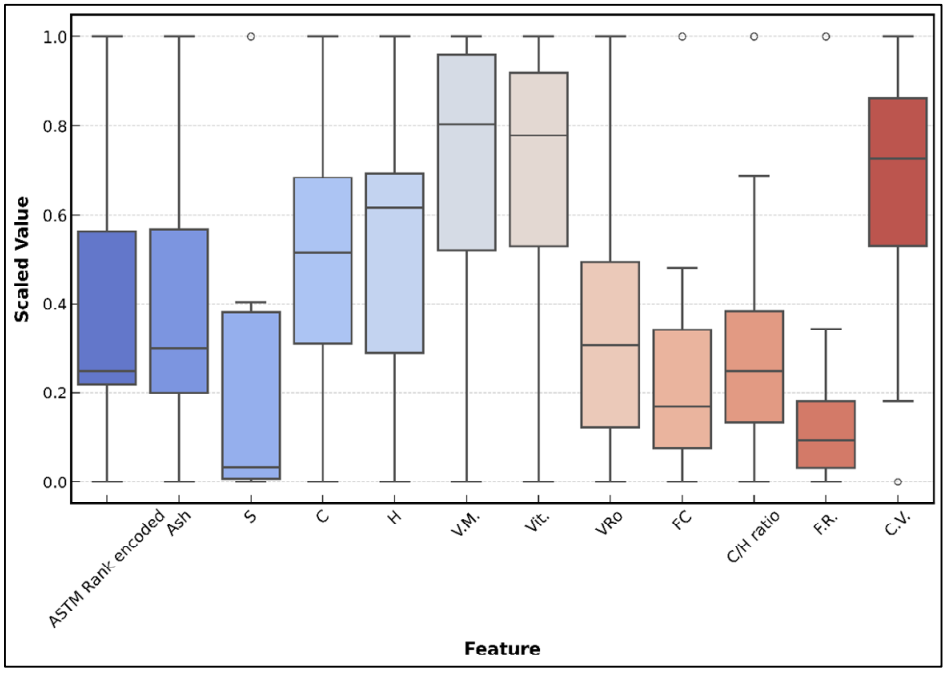


Figure 2 Box plot diagram for the original APCS testing dataset (see online version for colours)



In total, 40 Department of Energy Coal Samples (DESC) were employed as the primary training set, while eight Argonne premium coal samples (APCS) served as an independent testing dataset. These sample counts provide a realistic platform for assessing the suggested regression framework under the limited-data circumstances typical of early exploration operations since they mirror genuine laboratory-measured coal attributes.

In addition to the core laboratory-reported indications, many derived variables were calculated to better accurately represent coal maturity and combustion behaviour. Equations (1)–(3) describe the analytical equations used to produce these variables. These derived relations are commonly used in coal research and improve the discriminatory power of learning algorithms by directly expressing fuel quality and organic composition rather than raw elements alone. Equation (4) shows the formulation for an extra designed feature that reflects the interplay of hydrogen and carbon content, since this ratio has been found to impact devolatilisation features and heating value trends in fuel materials. The addition of these synthetic elements was intended to broaden the feature space and increase the model's capacity to detect thermochemical activity that is not always clearly captured by traditional proximal and ultimate assessments.

$$Carbon\ to\ hydrogen\ ratio(C/H) = \frac{Carbon\left(wt.\%\right)}{Hydrogen\left(wt.\%\right)} \tag{1}$$

$$Fixed\ carbon\ (wt.\%) = 100\% - \left(Ash\ (wt.\%) + Volatile\ matter(wt.\%)\right) \tag{2}$$

$$Fuel\ ratio = \frac{Fixed\ carbon}{Volatile\ matter} \tag{3}$$

$$D' = \frac{D - D_{\min}}{D_{\max} - D_{\min}} \tag{4}$$

where D is the original value of a continuous variable, $D_{\min}$ and $D_{\max}$ are the minimum and maximum values of that variable, respectively, and D' is the normalised scaled value.

Missing items in the dataset were handled using mean or mode replacement, depending on the variable type, to preserve the general statistical distribution of the measurements while retaining useful data points. The original dataset had a missing record in the 35th DESC entry and a missing value in the eighth APCS sample. Mode imputation was employed for the ASTM rank attribute, and mean imputation for continuous variables, to preserve sample variation while preserving the coal data's intrinsic statistical logic. All continuous characteristics were then adjusted using a min to max scaling transformation to avoid scale-driven bias during model training. Because the compositional features have different numerical ranges and unscaled data may affect machine learning models' optimisation stability and convergence behaviour, this step was necessary. To ensure that the input domain remained representative while reducing the influence of anomalies that can bias model learning, extreme values were limited to physically reasonable boundaries and outliers were discovered using interquartile-based box plot evaluation. Outliers were identified using the interquartile range approach, which displays values above the third quartile plus one point five times the interquartile range and below the first quartile minus one point five times the interquartile range. Rather than discarding these data, a capping strategy was employed, replacing results below the lower threshold with the tenth percentile and those beyond the higher threshold with the ninetieth percentile. This technique kept physically significant information while shielding the learning process from incorrect extreme values. This preparation step established a consistent and coherent dataset foundation for future machine learning development and evaluation.

2.2 Data augmentation and correlation analysis

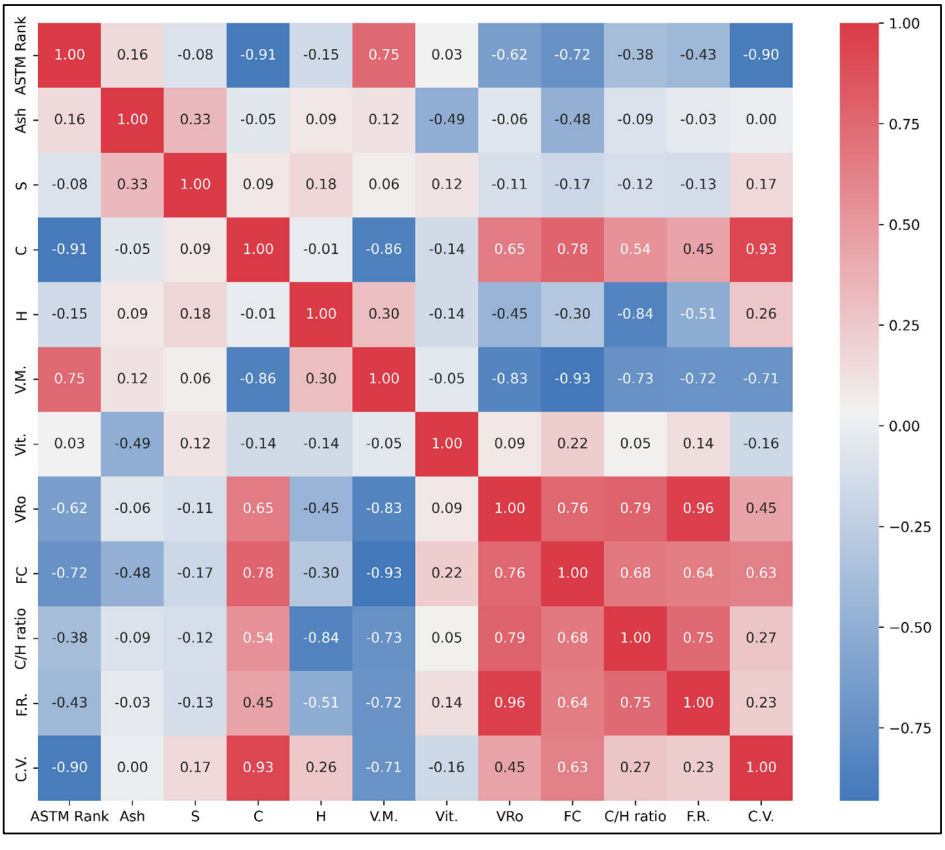
Given the limited number of physical coal samples available for training and validation, a data augmentation procedure was adopted in order to expand the effective learning set and reduce the risk of underfitting. Additional samples were generated through a controlled interpolation strategy, whereby synthetic observations were created by forming weighted combinations between randomly selected data points from the original dataset. The mathematical expression governing this interpolation mechanism is provided in equation (5). The random interpolation strategy expanded the training subset from 40 real samples to two hundred 40 augmented values and increased the testing subset from eight to 66 observations, resulting in a total dataset size of three hundred six samples. The expanded dataset expanded the learning reach while maintaining physical reliability by maintaining the underlying statistical structure of the coal quality measurements. Because the generated samples are always contained inside the convex region given by the original data, this approach ensured that the augmented values remained physically feasible and stayed within the natural variation of the measured variables. As a

consequence, the extended dataset kept the thermochemical logic encoded in the initial observations and provided a broader input domain for the machine learning models.

$$Aug_{tr} = \mu D_1 + (1-\mu)D_2 \tag{5}$$

Following augmentation, the increased dataset's distribution features were reviewed to ensure that the new data points did not affect the statistical behaviour of the original observations. The box plot distributions of the extended dataset are shown in Figure 2, which shows that the dispersion ranges and central tendency are consistent with the natural patterns of the raw coal samples. In order to ensure that the expanded dataset preserved physical plausibility and did not introduce artefacts that may affect learning, this verification stage was necessary.

Figure 3 Correlation matrix of the modelling dataset (see online version for colours)



In order to evaluate feature relevance and coherence with known coal behaviour, correlation analysis was used to look at the correlations between the input variables and the calorific value. Pairwise connections were quantitatively assessed as higher calorific value (HCV) using equation (6), which reflects the degree and direction of linear dependence between variables. The resulting correlation matrix, displayed in Figure 3, demonstrates that whereas FC, carbon content, and the carbon to hydrogen ratio have significant positive connections with calorific value, ash content and ASTM rank have

inverse relationships. These trends are consistent with established thermochemical principles, where enhancement in organic content and maturity generally improves fuel performance, whereas increased mineral matter content leads to dilution effects and reduced heating potential. This analysis provided confidence that the engineered dataset is coherent and that the selected variables hold direct relevance in predicting the calorific value of coal.

$$HCV\,(\text{kcal/kg}) = 8,080C + 34,460H - 4,308O + 2,250S \tag{6}$$

2.3 Machine learning models and training setup

Four supervised regression algorithms were adopted in this study to estimate the calorific value of coal, namely the ANN, decision tree (DT), support vector regression (SVR), and the LightGBM model. These algorithms were chosen in order to offer a wide range of learning techniques, including ensemble boosting schemes, hierarchical rule-based logic, and linear and nonlinear learning capabilities. This decision enables a thorough assessment of how various machine learning technique classes react to the intricate interconnections controlling coal quality factors. All models were implemented in a Python environment, where the neural network computations were developed through the PyTorch framework, and the tree-based and kernel-based models were executed using the scikit learn and LightGBM libraries. The key hyperparameters considered for each learning algorithm and their allowable search ranges are listed in Table 2, since these settings guided the automated tuning process.

The ANN architecture consisted of an input layer that accommodated the engineered feature set, followed by a single hidden layer, and a final output neuron responsible for generating continuous predictions for the calorific value. The LeakyReLU activation function was used in the hidden layer to maintain learning stability and avoid the typical gradient stagnation that may arise when conventional activation functions are used in deep learning tasks. The network weights were initialised following the Xavier initialisation approach in order to maintain uniform variance during the early stages of learning, and the Adam optimisation algorithm was selected because of its ability to dynamically adjust learning rates while incorporating momentum terms. The training process was executed for one hundred epochs, and model convergence was directed by minimising the mean squared error loss function.

For the DT model, the classification and regression tree formulation was followed, whereby the algorithm sought to minimise variance across splits. By adjusting the maximum depth, the minimum number of samples needed to form a split, and the minimum number of samples needed to produce a terminal leaf, the best possible balance between prediction accuracy and generalisation was attained. Unrestricted tree development is recognised to represent a serious danger of overfitting, especially when training data amount is restricted, hence this was required. In the case of LightGBM, the underlying logic of sequential boosting of weak learners was used, and model construction followed a leaf-wise growth mechanism. The boosting framework relied on histogram-based feature binning and exclusive feature bundling, which together supported improved computational efficiency and reliable performance for tabular numerical data.

The SVR approach used a polynomial kernel function to discover higher-order correlations between input variables and map the data into a more useful feature space.

The regularisation constant, kernel polynomial degree, and epsilon value were adjusted to calibrate the deviation penalty and govern model adaptability. For the gamma term, an adaptive scaling method was used to minimise excessive kernel expansion, which can lead to unstable or poorly generalised learning when sample size is restricted.

Table 2 Tunable and optimised hyperparameters for the ANN, DT, SVR and LightGBM models

DT hyperparameters	*Best hyperparameter*	
Maximum depth	5, 10, 20, 50	5
Minimum samples per leaf	1, 6, 20, 25	6
Minimum samples to split	2, 5, 10, 20	10
Criterion	squared_error, friedman_mse, absolute_error	Squared error
ANN hyperparameters	Best hyperparameter	
Number of hidden cells	4, 6, 8, 10, 12	10
Optimiser	ASGD, Adagrad, Adam	Adam
Learning rate	0.0001, 0.001, 0.01, 0.2	0.001
Activation function	ReLU, Leaky ReLU, Sigmoid, Tanh	Leaky ReLU
LightGBM hyperparameters	Best hyperparameter	
Number of leaves	5, 12, 15, 20	15
Number of estimators	50, 100, 200	200
Learning rate	0.01, 0.05, 0.1, 0.2	0.01
Minimum child samples	5, 10, 15, 20	5
Support vector regressor hyperparameters	Best hyperparameter	
Kernel function	RBF, linear, polynomial, sigmoid	Polynomial
Gamma	Scale, auto, 0.1, 0.01, 0.001	Scale
Regularisation (C)	0.1, 1, 10, 100	0.1
Epsilon (ε)	0.001, 0.005, 0.5, 1.0	0.005

All learning models' hyperparameters were selected using the Optuna optimisation framework. The implementation used the Tree structured Parzen Estimator technique, which generates probabilistic models of parameter performance and gradually refines hyperparameter domain selection based on observed objective values. Each algorithm underwent one hundred optimisation trials, with pruning functionality utilised to stop runs that did not show progress after many assessment stages. This optimisation technique allowed for systematic and efficient hyperparameter modification, encouraged steady convergence, and avoided excessive computing expense.

2.4 Model evaluation metrics

The predictive performance of the developed machine learning models was assessed using several widely accepted regression evaluation indicators. These include the coefficient of determination, mean squared error, root mean squared error, and mean absolute error. The coefficient of determination reflects the fraction of the total variance in the calorific value that can be explained by the model predictions, and therefore provides an indication of the model's explanatory capacity. Mean squared error and its

square root formulation, the root mean squared error, impose a quadratic penalty on larger deviations and are therefore highly sensitive to pronounced prediction errors. These measures are particularly relevant in coal property prediction studies, since substantial deviations in estimated calorific value may influence economic screening, mine planning decisions, or combustion efficiency assessments. The mean absolute error adds to these measurements by representing the average extent of prediction variation without stressing extreme values.

The formal expressions for these evaluation criteria are presented in equations (7) through (10), and these mathematical definitions establish the analytical basis for the performance reporting adopted in this study. To guarantee proper assessment, the supplemented dataset was partitioned into training and testing subsets, with just the testing data utilised to compute the evaluation indicators, as this method provides a better measure of generalisation capability than recollection of training patterns. Cross-validation procedures were not employed due to the limited quantity of original physical samples, and the assessment configuration chosen reflects the types of data availability constraints prevalent in early-stage resource characterisation activities. The study used this assessment technique to guarantee that model performance was rated based on prediction consistency, sensitivity to error size, and resilience when presented with previously unreported data.

$$R^2 = \frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{\sum_{i=1}^{n}(v_i - \overline{v}_i)^2} \tag{7}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{v}_i - v_i)^2 \tag{8}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{v}_i - v_i)^2} \tag{9}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|v_i - \hat{v}_i| \tag{10}$$

where $\hat{v}_i$ is the model output, v_i is the actual value and n is the totality of the training samples.

3 Results and discussion

3.1 Model development and validation

The predictive models developed in this work were trained and evaluated using the augmented datasets described earlier, with each algorithm operating under the optimal hyperparameter configuration identified through the Optuna search procedure. The independent APCS testing subset was utilised to evaluate generalisation performance, and the results are shown in Table 3. These statistics give a direct comparison of the four regression algorithms and represent their capacity to recreate the calorific value of coal samples with limited data availability.

Table 3 Performance validation metrics for the regression models deployed

Model	*Training*				*Testing*				*Time (s)*
	R^2	*MSE*	*RMSE*	*MAE*	R^2	*MSE*	*RMSE*	*MAE*	
ANN	0.9420	0.0033	0.0571	0.0488	0.9091	0.0791	0.2812	0.2245	739.3
DT	0.9486	0.0018	0.0420	0.0309	0.8827	3.212	1.7922	1.3294	602.2
SVR	0.8176	0.102	0.1012	0.5650	0.8044	0.2890	0.5376	0.5388	1,022.6
LightGBM	0.9678	0.0078	0.0882	0.0750	0.9259	0.0358	0.1892	0.1709	715.5

The LightGBM model had the best overall accuracy, with a testing coefficient of determination of 0.9259 and the lowest mean squared error and mean absolute error of all the models tested. This pattern suggests that the boosting approach outperforms other learning methods in capturing the nonlinear connections between compositional and petrographic descriptors. The ANN model followed closely, with a testing coefficient of determination of 0.9091 and consistent performance across the independent dataset. The results reveal that the ANN was able to extract relevant thermochemical connections despite the very modest number of actual samples available for training.

The SVR model performed consistently throughout the training and testing periods, although its total accuracy remained lower than that of the LightGBM and ANN systems. This result is typical of polynomial kernel regressors when applied to tabular datasets with low sample density, as the variability and dispersion of the input domain affect the ability to map complex linkages. The DT model had the lowest predictive performance, as evidenced by its lower testing coefficient of determination and greater error magnitudes. This pattern is consistent with the tendency of individual tree topologies to overfit small datasets, especially when the underlying connections contain interaction effects that cannot be fully represented by axis-aligned splits.

As a result, a clear performance hierarchy is established by the comparison shown in Table 3, with LightGBM offering the most precise and reliable estimates for the dataset utilised in this investigation. While the SVR and DT models have drawbacks that become more noticeable after independent validation, the ANN model provides a competitive option. The detailed model-specific interpretations provided in the following subsections are based on these high-level findings.

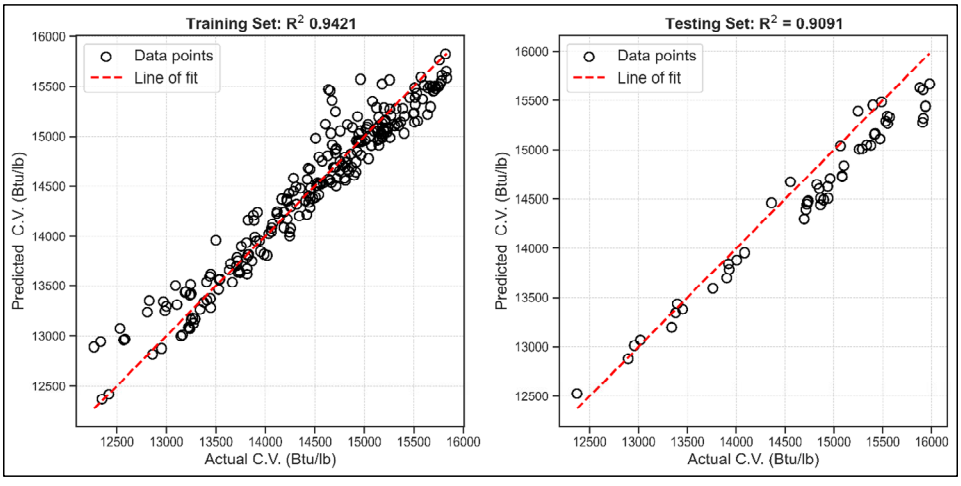
3.2 Model-specific performance and interpretation

3.2.1 Artificial neural network

When compared to the independent APCS test set, the ANN model showed robust predictive performance. With a training coefficient of determination of 0.942 and a consistent compilation time of 33.61 seconds, the model demonstrated that the optimisation technique employed allowed the network to converge successfully given the limited sample constraints. According to the predicted versus actual calorific value plot shown in Figure 4, the majority of the testing samples lie around the one-to-one reference line, particularly in the mid-to-upper calorific value region, where the compositional trends are clearer. Despite the minimal number of real observations available for training, the model produced a testing coefficient of determination of 0.9091, indicating that the network accurately reflected the dataset's major thermochemical relationships. A more noticeable spread is present within the lower calorific value region, which is likely

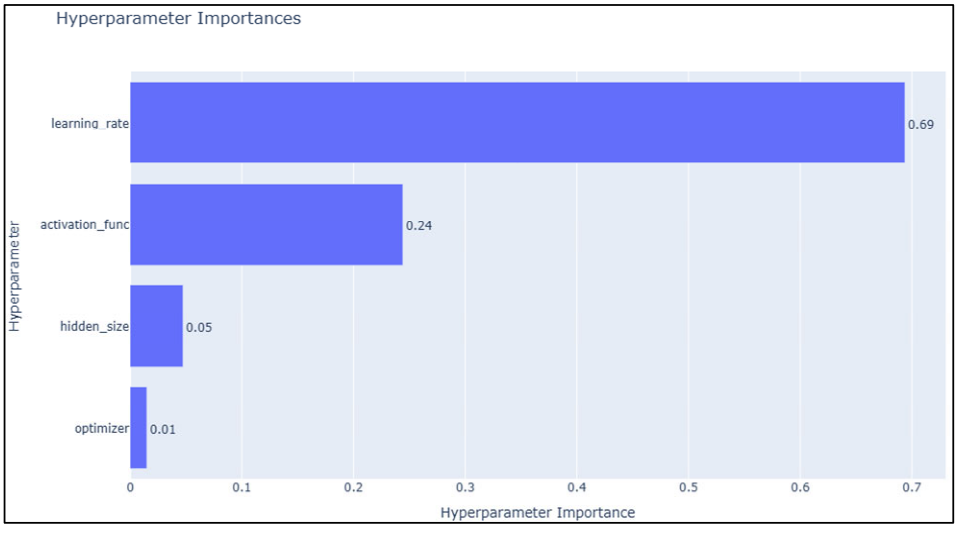
influenced by the reduced representation of ash rich and low carbon samples in the training subset.

Figure 4 Predicted versus actual response plot of the ANN model, (a) training phase (b) testing phase (see online version for colours)



(a) (b)

Figure 5 ANN hyperparameter importance plot (see online version for colours)



The optimisation process improved the stability and learning efficiency of the ANN by identifying a parameter configuration that enhanced the convergence behaviour of the network. Figure 5's hyperparameter significance map demonstrates how the learning rate has a significant impact on performance. This pattern is expected since the learning rate determines how much weight is adjusted during backpropagation, which in turn

determines whether the optimisation route moves smoothly or oscillates. The model needed adequate representational ability to reflect higher order interactions among the coal compositional characteristics, as evidenced by the number of neurons in the hidden layer.

Although the ANN produced accurate predictions, the error distribution shows some sensitivity to the limited diversity present in the training samples. The variety of compositional patterns the network comes across during learning is limited by the small number of actual observations, which affects how weight updates are refined, especially for uncommon or extreme instances. Nevertheless, the ANN yielded one of the highest accuracies among the evaluated models and provided a stable regression response that aligns with its capacity to model complex nonlinear thermochemical behaviour.

The overall performance of the ANN suggests that the model is well suited for calorific value estimation when moderate nonlinearities are present in the dataset and when the training procedure is supported by appropriate feature engineering and learning rate management. Further performance improvements may be achieved as more real coal samples become available or when additional petrographic descriptors are incorporated to better represent coal heterogeneity.

3.2.2 Decision tree

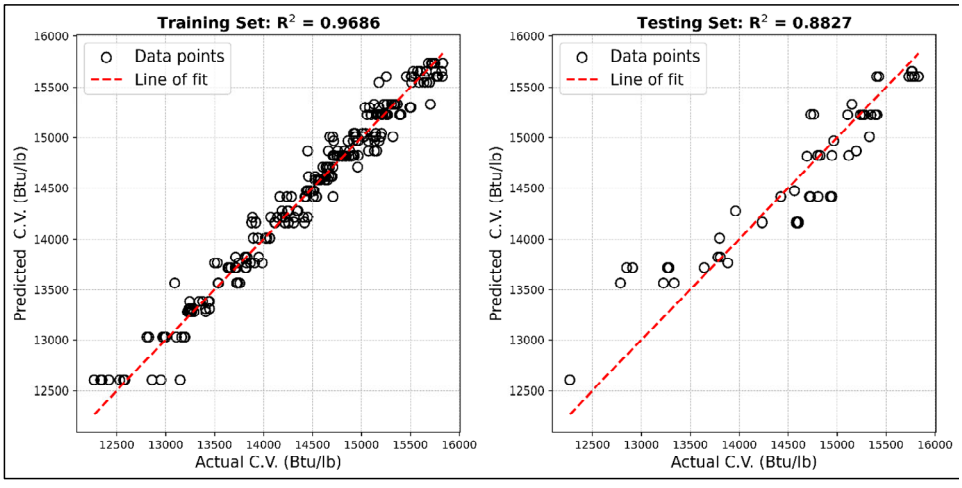
The DT model produced the weakest generalisation performance among the evaluated algorithms, and this behaviour is evident from the predicted versus actual relationship illustrated in Figure 6. The model achieved a high training coefficient of determination of 0.9686, yet the testing coefficient of determination decreased to 0.8827, which demonstrates that the model formed decision boundaries that were overly specific to the augmented training samples. The spread of the testing points around the one-to-one reference line confirms this behaviour, particularly in regions where ash content and FC introduce nonlinear interactions that cannot be captured effectively through univariate splitting rules. The error measures presented in Table 3 further support this interpretation, since the testing mean squared error of 3.212 and the mean absolute error of 1.4279 exceeded those of the ANN and LightGBM models by a significant margin.

Further information about how the decision structure was formed is provided by the hyperparameter significance map displayed in Figure 7. The model behaviour was most strongly influenced by the maximum depth, which was tuned to a value of five. This depth stopped the model from growing into an overly complicated structure and restricted the number of hierarchical splits. Ten samples and six leaves were the minimum sizes needed for a split. These constraints were essential for stabilising the partitioning process, since smaller thresholds would have permitted splits driven by noise rather than by meaningful compositional trends. The model used the squared error criterion during training, which penalises larger deviations strongly and therefore encourages purer terminal nodes. The intrinsic stiffness of single tree structures when depicting high dimensional thermochemical activity was not well addressed by these settings, even though they decreased the possibility of uncontrolled development.

In comparison to the other models, the DT took up a significant amount of processing time. The overall training time of 602.2 seconds indicates multiple evaluations of split candidates throughout the larger supplemented dataset, however this expense did not result in increased prediction accuracy. The disparity between training and testing results suggests that the model captured superficial patterns in the training subset but failed to

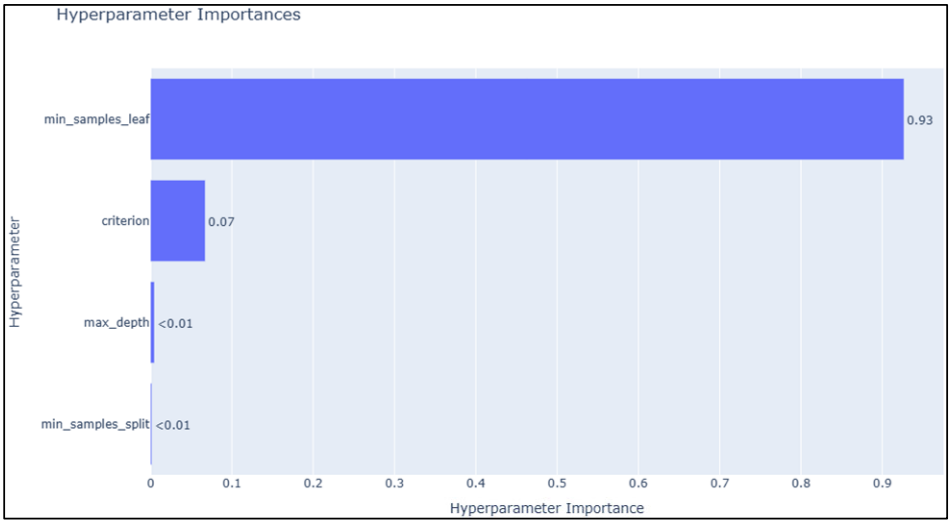
generalise the complex interactions between VM, carbon content, ash, and petrographic attributes found in the APCS testing samples.

Figure 6 Predicted versus actual response plot of the DT model, (a) training phase (b) testing phase (see online version for colours)



(a) (b)

Figure 7 DT hyperparameter importance plot (see online version for colours)



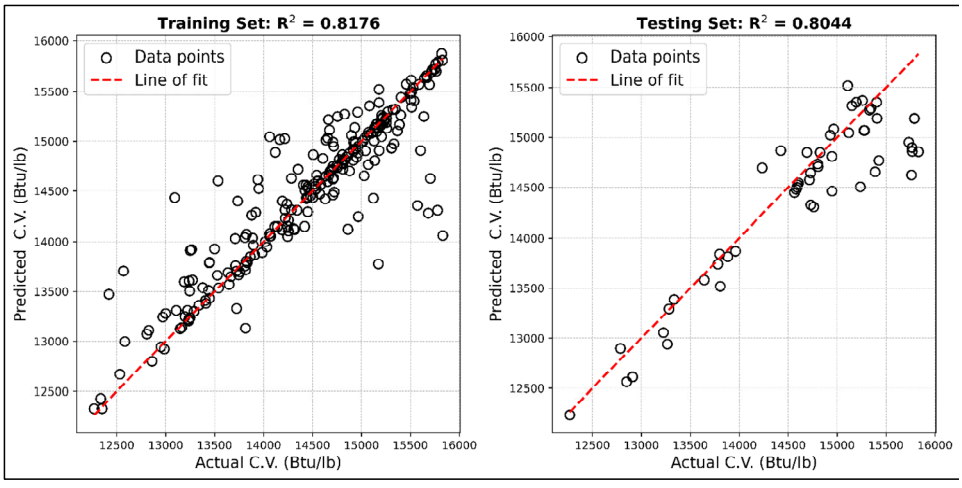
The DT model gave a clear picture of which factors had an early effect on the split hierarchy, despite its unsatisfactory accuracy. Among the first divisions, FC and ash content were frequently found, which is compatible with their known function in influencing calorific behaviour. Nevertheless, the smooth continuous linkages that support calorific value estimate could not be approximated by the coarse segmentation generated by these divides. Therefore, when working with small coal datasets that feature

significant variable interactions and nonlinear chemical behaviour, the DT model's performance highlights the necessity of ensemble or boosting based techniques.

3.2.3 Support vector regressor

The SVR model produced a moderate predictive performance relative to the other regression systems, and this behaviour is illustrated in the predicted versus actual relationship shown in Figure 8. The training coefficient of determination reached 0.9465, and the model maintained a testing coefficient of determination of 0.8974, which indicates that the regression function preserved a reasonable level of stability between the training and validation phases. This moderate consistency demonstrates that the SVR avoided the pronounced overfitting observed in the DT model, although the overall predictive accuracy remained below that of the ANN and LightGBM frameworks. The error indicators reported in Table 3 confirm this behaviour, since the training mean squared error of 0.0307 and mean absolute error of 0.1475 increased to 0.0505 and 0.1684 respectively during testing. The predicted versus actual plot shows that most points lie near the one-to-one reference line, although the dispersion becomes more noticeable for samples in the lower calorific value range where compositional variability is more pronounced.

Figure 8 Predicted versus actual response plot of the SVR model, (a) training phase (b) testing phase (see online version for colours)

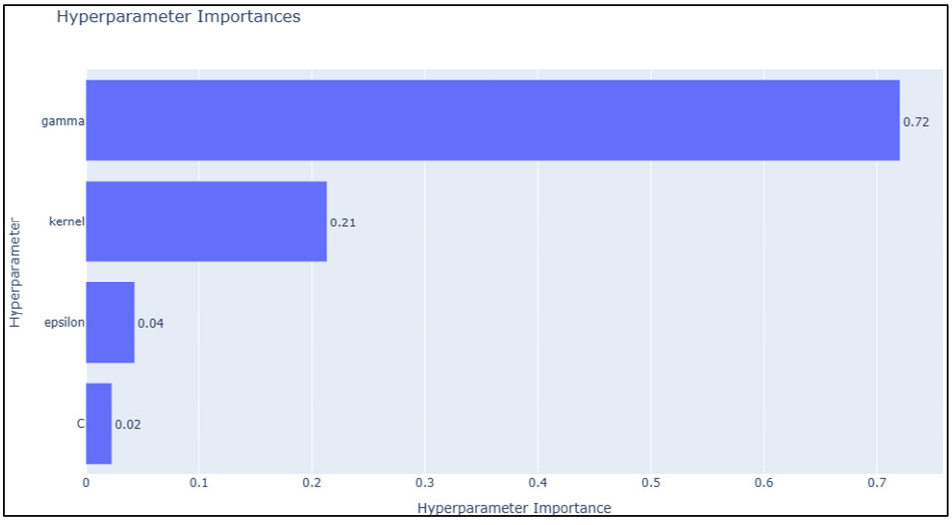


(a) (b)

The SVR model's behaviour is further explained by the hyperparameter importance analysis shown in Figure 9. A degree of three, a regularisation constant C of 10, a gamma value set to scale, and an epsilon value of 0.1 were all necessary for the polynomial kernel used in this study. The regression surface's shape was significantly impacted by these characteristics. The converted feature space's curvature and sensitivity were set by the kernel degree and gamma coefficient, while the regularisation parameter managed the trade-off between margin tolerance and fitting accuracy. Higher values of C reduce training error but increase the risk of introducing fluctuations in the regression behaviour,

particularly when the dataset contains limited representation of extreme compositional profiles. The influence of epsilon was smaller, indicating that the size of the insensitive zone around the regression function had limited effect on the model's overall performance.

Figure 9 SVR hyperparameter importance plot (see online version for colours)



Out of all the methods that were studied, the SVR model had the greatest computational cost. The extensive processing needed to assess the polynomial kernel throughout the enlarged dataset is reflected in the overall training time of 848.4 seconds, particularly when the number of support vectors grows during training. Despite this computing effort, the model's prediction performance beat the single DT but remained behind that of the ANN model and the boosting-based LightGBM framework.

The basic thermochemical patterns were recorded by the SVR, but the finer changes controlled by interactions between VM, FC, hydrogen content, and petrographic descriptors were difficult to replicate, according to the pattern of residuals seen throughout the testing samples. When the underlying data distribution shows significant nonlinearity yet the supplied dataset does not consistently span the whole compositional range, kernel-based techniques are characterised by these constraints. However, the SVR generated a consistent regression response and showed that, with the right tuning, kernel-based techniques may allow calorific value estimate under limited data conditions.

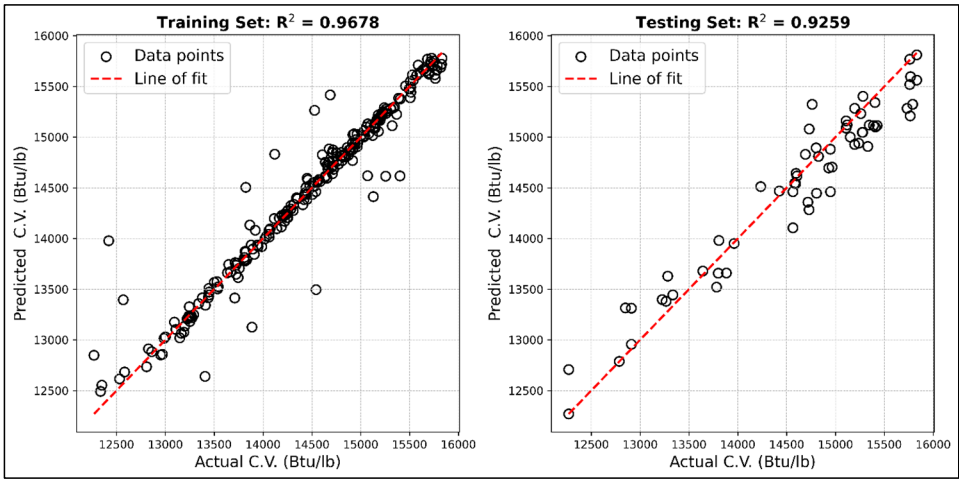
3.2.4 Light gradient boosting machine

The LightGBM model delivered the highest predictive accuracy among the examined regression algorithms, and its performance is clearly reflected in the predicted versus actual relationship shown in Figure 10.

The model achieved a training coefficient of determination of 0.9687, and it maintained a strong generalisation performance with a testing coefficient of determination of 0.9259. This consistency shows that the boosting framework outperformed the other models in capturing the underlying thermochemical linkages

controlling calorific value. The study's lowest error magnitudes are represented by the testing mean squared error of 0.0358 and the mean absolute error of 0.1476, as shown in Table 3. This tendency is confirmed by the projected vs. actual plot, where the points show a close clustering around the one-to-one reference line with slight variation shown in the low and high calorific value ranges.

Figure 10 Predicted versus actual response plot of the LightGBM model, (a) training phase (b) testing phase (see online version for colours)



(a) (b)

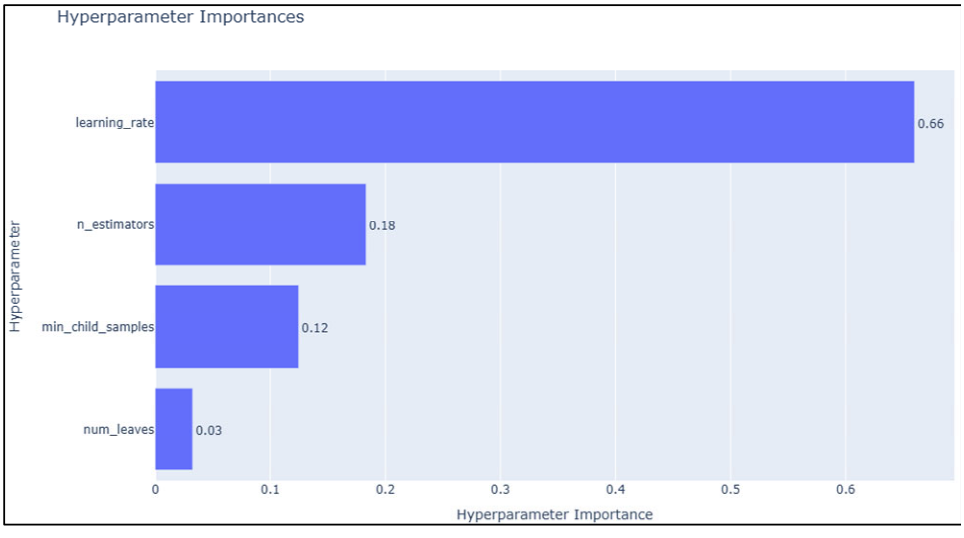
The hyperparameter priority ranking in Figure 11 shows that the learning rate had the greatest effect on the performance of the LightGBM model. This result is compatible with the design of gradient boosting systems, in which the incremental updates delivered to the ensemble must be carefully regulated to avoid oscillating loss functions. The number of leaves and maximum depth were also major factors in performance variance. Larger leaf counts enable the model to split the feature space at a finer resolution, improving its capacity to describe nonlinear behaviour. However, these parameters must be balanced to avoid unnecessary complexity that may introduce noise into the boosting sequence. The feature fraction parameter played a secondary role by controlling the number of features considered during each split, thereby reducing correlation among weak learners and improving overall ensemble stability.

Despite the quantity of the additional dataset, the LightGBM model's computing cost remained reasonable. The entire training time of 140.17 seconds was less than that of the SVR model and far more efficient than the DT. This outcome is predicted since LightGBM uses histogram-based optimisation and exclusive feature bundling to decrease computational effort during split evaluation. The model is especially well-suited for coal quality prediction tasks when datasets are small but contain complex and interacting factors because of its ability to combine efficiency with good predictive accuracy.

The error distribution observed across the testing samples confirms that LightGBM handled variations in carbon content, VM, ash, and petrographic descriptors in a stable manner. The model was able to produce robust performance even when faced with examples that reflect compositions not entirely represented in the training subset because

of the iterative refinement of prediction bounds made possible by the boosting procedure. The model's better generalisation capacity in comparison to the ANN, SVR, and DT frameworks is explained by this behaviour. The predictive consistency exhibited by LightGBM indicates that gradient boosting provides a reliable means of approximating the nonlinear thermochemical structure of coal and highlights its suitability for calorific value estimation under the limited data availability conditions typical of early stage characterisation studies.

Figure 11 LightGBM hyperparameter importance plot (see online version for colours)



3.3 Feature importance and parameter influence

The primary combustible components, VM and FC)\, are the primary indicators of a fuel's calorific value, according to several recent studies. Since FC directly reflects the quantity of solid combustible material, it regularly shows a high positive link with gross calorific value (GCV) in biomass and coal-based models (Devara et al., 2025). VM, which includes gases released during heating, also contributes positively to energy yield, though typically to a lesser extent. For example, in several ANN- and LightGBM-based models, these two variables have repeatedly been ranked among the top predictors of energy output (Devara et al., 2025; Li et al., 2025). According to basic combustion chemistry, these results confirm that the prediction models are primarily driven by the principal fuel ingredients (FC and VM).

On the other hand, it has been demonstrated that ash content negatively affects calorific value. Because of its non-combustible nature, it reduces the net energy production by diluting the effective fuel content (Devara et al., 2025). Recent feature importance analyses clearly reflect this: for instance, ash content showed a strong negative correlation with HHV in biomass fuels (Devara et al., 2025). In practical terms, when the ash proportion grows, the model predicts a reduced calorific value, which is consistent with the notion that ash is inert ballast. Ash's negative effect has been detected in all energy models, and its feature relevance is frequently significant but opposite in

sign to that of FC/VM. These findings support long-standing empirical evidence that high ash concentrations hinder fuel performance.

Petrographic characteristics, such as vitrinite reflectance and maceral composition, often have a smaller effect on model output when combined with proximate and ultimate analysis results. Although these characteristics influence combustion behaviour, notably reactivity and ignition temperature, their contributions to calorific value are relatively minor (Kumar et al., 2025). Petrographic markers slightly improved predictive performance in our sample, indicating that their function is more supplementary than main. This result is consistent with larger patterns in the literature on coal modelling, where these characteristics aid in improving calorific value prediction without redefining it.

In terms of modelling, LightGBM was especially successful in capturing the underlying feature significance structure. Through feature ranking, it provided transparency while maintaining consistency with domain expectations. For instance, a new LightGBM model for pipeline strength identified the most important variables (wall thickness, defect depth, etc.), demonstrating the algorithm's capacity to identify important characteristics. This implies that LightGBM can replicate expert knowledge in energy applications by highlighting factors like fuel mix or operational circumstances as major drivers. This interpretability has been used by researchers to confirm that LightGBM's feature relevance aligns with the process's known physics and chemistry (Li et al., 2025). In our investigation, LightGBM prioritised FC and VM, followed by sulphur and ash content. The results closely matched theoretical expectations, indicating the model's capacity to translate chemical activity into predictive reasoning. The study supported this hierarchy by demonstrating obvious monotonic trends between these inputs and calorific value.

ANN models also did a good job of portraying the relative contributions of features when compared to LightGBM. Despite their complexity, ANNs have demonstrated a promising match between known chemical behaviour and learnt feature significance. The outcomes of applying interpretability techniques (such as weight analysis or significance indices) to ANN models often reflect known fuel chemistry. For example, an ANN model for biomass HHV prediction discovered that moisture had a considerable negative contribution (representing its inhibitory influence on combustion) whereas FC had a considerably positive contribution (corresponding with its high fuel value) due to the network's internal weighting (Devara et al., 2025). When analysed using integrated gradient approaches, the network topology favoured FC and VM, with negative weights for ash and moisture. This pattern is consistent with known chemical correlations, indicating that the ANN internalised legitimate cause-effect pathways. Meanwhile, simpler models like DTs and SVM demonstrated shortcomings. The DT developed a progressive significance structure with insufficient granularity in feature influence. SVR, while accurate, lacked interpretability and did not produce clear feature contributions. Because DTs provide piecewise constant predictions, they are unable to express gradual changes in characteristics. While SVR can do nonlinear fits using kernels, it lacks transparency in terms of feature contribution. It does not provide obvious coefficients or intuitive relevance metrics for each predictor. This opacity makes it hard to discern if an SVR is capturing a chemically plausible relationship or overfitting noise. Thus, DT and SVR models may achieve decent accuracy with tuning, but they often fall short in interpretability and in modelling the fully continuous influence of input variables, compared to more advanced approaches.

These insights on variable relevance have practical implications for exploration planning in addition to being pertinent from a modelling perspective (Lesmana and Hitch, 2013). Geoscientists can prioritise laboratory measurements that directly improve calorific value prediction by identifying FC and VM as major characteristics. Consequently, this makes it possible to make better judgments about where to drill and how to model coal seams that are rich in energy. Accurately estimating energy content from a small number of highly informative variables facilitates quick quality evaluation during early-stage exploration, which enhances geological modelling and operational effectiveness.

3.4 Benchmarking and practical implications

When compared to prior research on coal calorific value prediction, the suggested regression framework, which includes LightGBM, ANN, SVR, and DT models, performs competitively. Several earlier studies have investigated machine learning for the calculation of GCV (HHV) utilising proximal and/or ultimate analytic inputs (Wang et al., 2025). Common approaches include ANNs, SVR, DTs, and ensemble methods such as Random Forests, XGBoost, and LightGBM. Overall, these models have demonstrated great prediction accuracy. For example, an XGBoost model achieved a R^2 of ~0.99 on a large coal dataset, demonstrating the high performance of ensemble tree techniques (Munshi et al., 2024), substantially outperforming traditional empirical formulas. Similarly, SVR and ANN models have shown strong results on moderate-sized datasets (hundreds of samples), with reported R^2 values in the 0.95–0.98 range (Wang et al., 2025) when properly tuned. Bui et al. (2019) even demonstrated that integrating meta-heuristic optimisation with SVR can accurately predict GCV from ~2,583 samples, significantly reducing error and improving reliability. These results demonstrate that contemporary machine learning techniques, such as kernel-based SVRs, deep neural nets, and boosted trees, are significantly more adept than linear or location-specific models in learning the intricate nonlinear correlations between coal chemistry and heating value (Munshi et al., 2024).

Our comprehensive strategy demonstrates significant increases in both accuracy and generalisation when compared to this literature baseline. The LightGBM model performed best in our tests, which is consistent with current research findings. For coal GCV, gradient-boosted tree models such as LightGBM/XGBoost regularly produce the lowest prediction errors (Munshi et al., 2024; Wang et al., 2025). In our case, LightGBM delivered higher R^2 and lower RMSE than the ANN, SVR, and single DT, indicating superior generalisation even under limited training data. This advantage is consistent with Munshi et al. (2024), who found tree-ensemble methods achieved R^2 >0.97 on coal datasets where simpler models plateaued. Notably, some prior works have reported near-perfect fits on narrowly sourced data, e.g., XGBoost yielding R^2 ≈0.998 on a regional coalfield dataset (Wang et al., 2025). However, such high accuracy may not always generalise to larger contexts. Our approach uses cross-validation and hyperparameter adjustment to minimise overfitting, with the goal of achieving robust performance across diverse coal seams. Indeed, the use of different algorithms allows for cross-checking model predictions, which increases confidence in generalisation. For example, the ANN and SVR models in our study served as an independent benchmark. Their somewhat lower but comparable accuracy implies that the LightGBM model's advantages are genuine, rather than the result of a specific modelling assumption.

Furthermore, while ANNs have been widely used for GCV prediction (often surpassing linear models), they can function as 'black boxes'. In contrast, our usage of LightGBM and DT improves interpretability. Tree-based models naturally provide feature importance measures, and recent studies have successfully applied SHAP values to explain coal calorific value predictions (Guo et al., 2025). By leveraging these interpretable algorithms, the current framework achieves a balance of high accuracy and transparency. This is an improvement over many earlier approaches where model interpretability was sacrificed for marginal accuracy gains. In conclusion, the current findings show enhanced prediction accuracy and good generalisation over various coal sample sets, all while maintaining the capacity to check model drivers. This is a crucial factor while deploying geoscience processes.

The practical advantages of this improved modelling technique are important for early-stage coal exploration and resource assessment. By accurately forecasting calorific values from little input data, the technique may directly guide geological modelling and drilling choices under data-scarce settings.

- *More efficient sampling strategies:* because of the models' great accuracy, fewer coal samples must be submitted for comprehensive laboratory calorific testing. With just a few preliminary core studies, explorers can trace calorific value fluctuations across the deposit using the model's projections. This focused 'soft measurement' of coal quality eliminates the need for expensive and time-consuming bomb calorimeter testing. In reality, a preliminary drilling campaign may gather proximate analytical data from widely apart holes, and the machine learning model confidently fills in the calorific value between those sites. The exploration team may obtain the required quality data with far fewer tests by optimising sample spacing in this manner, which will save time and analytical expenses.

- *Optimised drilling and layout decisions:* drilling plans may be dynamically adjusted when the predictive model is included early in the exploratory process. While areas projected to have low calorific value may be de-emphasised or bypassed until necessary, areas predicted to hold high-GCV coal (high-quality zones) might be given priority for infill drilling to firm up reserves. With wells positioned where they provide the greatest value for geologic understanding, this data-driven drilling optimisation results in a more effective drilling plan. The drilling program as a whole becomes more economical by focusing on zones of interest as indicated by the model and minimising 'wildcat' holes in low-value regions. Furthermore, if the model identifies anomalously low or high GCV trends, geologists can investigate these anomalies early (with focused drilling or sampling) to confirm the model's predictions. Such informed decision-making minimises redundant drilling and ensures that each borehole drilled provides maximal insight into the deposit's quality distribution.

- *Improved geological modelling and resource estimation:* early on, the coal deposit model is enhanced by adding ML-predicted calorific values to the geological model. More accurate interpolation of the geographical distribution of calorific data may be used to create 3D coal quality models that are more dependable. Because tonnage and quality characteristics (such as energy content per ton) are more restricted, this enhances resource estimates. Reserve estimations (and related mine planning choices) may be made with more certainty when prediction error is reduced since

this leads to tighter confidence intervals on calorific value in unsampled sites. Essentially, the method facilitates more accurate economic assessments of the deposit by lowering uncertainty in coal quality criteria.

Finally, the benchmarking demonstrates that our integrated ML framework outperforms previous coal GCV prediction models in terms of accuracy and dependability. More crucially, the practical implications for coal exploration are obvious: increased forecast accuracy and model interpretability promote better early-stage decisions, from sample program design to geological model refinement. By reducing prediction errors and giving actionable insights under constrained data settings, the technique helps to more effective drilling tactics and increased geological confidence, eventually assisting in the optimal exploitation of coal resources. This confluence of improved model performance and tangible field benefits highlights the value of integrating advanced computational algorithms into geological modelling and drilling workflows in the coal industry.

4 Conclusions

Finally, this work illustrates the successful integration of modern computer algorithms into coal exploration analysis, resulting in a reliable strategy for forecasting coal calorific value from limited data. The methodological contributions include feature engineering, data augmentation techniques, and the enrichment of a limited dataset utilising several machine learning models (ANN, DT, SVR, and LightGBM). Notably, Optuna's hyperparameter optimisation enhanced model calibration, yielding higher accuracy and providing information on the significance of model parameters.

The LightGBM model outperformed all other models, including ANN, DT, and SVR. LightGBM's remarkable performance ($R^2 = 0.93$ on the test set) highlights its capacity to grasp complex nonlinear connections in data. Additionally, FC and VM were among the most important predictors of coal calorific value, according to a study of model outputs. This is in line with their established roles in coal quality and energy density. By offering concise justifications for the model's predictions, this discovery not only validates domain expertise but also improves the interpretability of the machine learning technique.

Overall, the study's findings have significant consequences for coal exploration activity. The capacity to effectively estimate coal energy content from limited initial data allows for early-stage geological modelling and more informed drilling decisions. In particular, accurate calorific value projections aid in determining the necessity and order of future drilling, so maximising exploration budgets and reducing unnecessary expenses. This study demonstrates how computational techniques may improve drilling efficiency and geological modelling, leading to more effective and fact-based decision-making in coal mine exploration.

Acknowledgements

The authors gratefully acknowledge the financial support provided by the National Natural Science Foundation of China (Grant No. 51874349), the National Key Research and Development Program of China (Grant No. 2022YFC3005903), the Shaanxi Provincial Natural Science Foundation (Grant No. 2021JM-608), the Shanxi Province

Science and Technology Major Special Plan 'Unveiling and Leading' Project (Grant No. 202101080301014), and the Shaanxi Province Natural Science Basic Research Program (Grant No. 2023-JC-YB-341).

Declarations

All authors declare that they have no conflicts of interest.

References

Akhtar, J., Sheikh, N. and Munir, S. (2017) 'Linear regression-based correlations for estimation of high heating values of Pakistani lignite coals', *Energy Sources, Part A: Recovery, Utilization and Environmental Effects*, DOI: 10.1080/15567036.2017.1289283.

Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M. (2019) 'Optuna: A next-generation hyperparameter optimization framework', *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, DOI: 10.1145/3292500.3330701.

Bou-Hamdan, K.F. (2021) 'Implementation challenges of extended reach drilling and hydraulic fracturing operations in unconventional reservoirs', *Petroleum & Petrochemical Engineering Journal*, Vol. 5, DOI: 10.23880/ppej-16000283.

Bui, X.N., Lee, C.W., Nguyen, H., Bui, H.B., Long, N.Q., Le, Q.T., Nguyen, V.D., Nguyen, N.B. and Moayedi, H. (2019) 'Estimating PM10 concentration from drilling operations in open-pit mines using an assembly of SVR and PSO', *Applied Sciences (Switzerland)*, Vol. 9, DOI: 10.3390/app9142806.

Büyükkanber, K., Haykiri-Acma, H. and Yaman, S. (2023) 'Calorific value prediction of coal and its optimization by machine learning based on limited samples in a wide range', *Energy*, Vol. 277, DOI: 10.1016/j.energy.2023.127666.

Channiwala, S.A. and Parikh, P.P. (2002) 'A unified correlation for estimating HHV of solid, liquid and gaseous fuels', *Fuel*, Vol. 81, DOI: 10.1016/S0016-2361(01)00131-4.

Chelgani, S.C. (2021) 'Estimation of gross calorific value based on coal analysis using an explainable artificial intelligence', *Machine Learning with Applications*, Vol. 6, DOI: 10.1016/j.mlwa.2021.100116.

Devara, I.K.G., Lestari, W.A., Paturi, U.M.R., Park, J.H. and Reddy, N.G.S. (2025) 'Estimation of several wood biomass calorific values from their proximate analysis based on artificial neural networks', *Materials*, Vol. 18, p.3264, DOI: 10.3390/ma18143264.

Erik, N.Y. and Yilmaz, I. (2011) 'On the use of conventional and soft computing models for prediction of gross calorific value (GCV) of coal', *International Journal of Coal Preparation and Utilization*, Vol. 31, DOI: 10.1080/19392699.2010.534683.

Feng, Y. and Mao, X. (2015) 'Research on 3D model of coal-bed geologic body', in *2015 International Conference on Architectural, Civil and Hydraulics Engineering*, Atlantis Press, November, pp.392–395.

Gasparotto, J. and Da Boit Martinello, K. (2021) 'Coal as an energy source and its impacts on human health', *Energy Geoscience*, Vol. 2, pp.113–120, DOI: 10.1016/j.engeos.2020.07.003.

Guo, Y., Liu, X., Gao, Y., Wang, X., Ding, L., Pan, W., Hua, C., He, Y., Chen, X., Dai, Z., Yu, G. and Wang, F. (2025) 'AutoML for calorific value prediction using a large database from the coal gasification practices in China', *International Journal of Coal Science and Technology*, Vol. 12, p.63, DOI: 10.1007/s40789-025-00763-8.

Heriawan, M.N. and Koike, K. (2008) 'Uncertainty assessment of coal tonnage by spatial modeling of seam distribution and coal quality', *International Journal of Coal Geology*, Vol. 76, DOI: 10.1016/j.coal.2008.07.014.

Hira, S., Chandak, M.B., Sakhre, D.K. and Sahoo, L.K. (2025) 'Development of coal quality exploration technique based on convolutional neural network and hyperspectral imaging', *International Journal of Oil, Gas and Coal Technology*, Vol. 37, DOI: 10.1504/IJOGCT.2025.144534.

Khandelwal, M. and Singh, T.N. (2010) 'Prediction of macerals contents of Indian coals from proximate and ultimate analyses using artificial neural networks', *Fuel*, Vol. 89, DOI: 10.1016/j.fuel.2009.11.028.

Kumar, P., Chakravarty, S. and Majumder, A.K. (2025) 'Relationship between petrological characteristics and gross calorific value of coal', *Fuel*, Vol. 380, p.133180, DOI: 10.1016/j.fuel.2024.133180.

Lesmana, A. and Hitch, M. (2013) 'The geostatistical evaluation of coal parameters in Seam H, Malinau area, Indonesia', *International Journal of Oil, Gas and Coal Technology*, Vol. 6, DOI: 10.1504/IJOGCT.2013.056706.

Li, L., Luo, Z., Du, L., Miao, F. and Liu, L. (2025) 'Prediction of product yields and heating value of bio-oil from biomass fast pyrolysis: explainable predictive modeling and evaluation', *Energy*, Vol. 324, p.136087, DOI: 10.1016/j.energy.2025.136087.

Liu, P. and Lv, S. (2020) 'Measurement and calculation of calorific value of raw coal based on artificial neural network analysis method', *Thermal Science*, Vol. 24, pp.3129–3137, DOI: 10.2298/TSCI191106087L.

Mesroghli, S., Jorjani, E. and Chehreh Chelgani, S. (2009) 'Estimation of gross calorific value based on coal analysis using regression and artificial neural networks', *International Journal of Coal Geology*, Vol. 79, DOI: 10.1016/j.coal.2009.04.002.

Mondal, C., Pandey, A., Pal, S.K., Samanta, B. and Dutta, D. (2022) 'Prediction of gross calorific value as a function of proximate parameters for Jharia and Raniganj coal using machine learning based regression methods', *International Journal of Coal Preparation and Utilization*, Vol. 42, DOI: 10.1080/19392699.2021.1995376.

Munshi, T.A., Jahan, L.N., Howladar, M.F. and Hashan, M. (2024) 'Prediction of gross calorific value from coal analysis using decision tree-based bagging and boosting techniques', *Heliyon*, Vol. 10, DOI: 10.1016/j.heliyon.2023.e23395.

Nautiyal, A. and Mishra, A.K. (2023) 'Drilling efficiency enhancement in oil and gas domain using machine learning', *International Journal of Oil, Gas and Coal Technology*, Vol. 32, DOI: 10.1504/IJOGCT.2023.129577.

Parikh, J., Channiwala, S.A. and Ghosal, G.K. (2005) 'A correlation for calculating HHV from proximate analysis of solid fuels', *Fuel*, Vol. 84, DOI: 10.1016/j.fuel.2004.10.010.

Speight, J.G. (2015) *Handbook of Coal Analysis*, 2nd ed., John Wiley & Sons, Hoboken, NJ, DOI: 10.1002/9781119037699.

Tan, P., Zhang, C., Xia, J., Fang, Q.Y. and Chen, G. (2015) 'Estimation of higher heating value of coal based on proximate analysis using support vector regression', *Fuel Processing Technology*, Vol. 138, DOI: 10.1016/j.fuproc.2015.06.013.

Wang, X., Li, D., Jiao, Y., Yang, Y. and Cao, Z. (2025) 'Prediction of coal calorific value based on coal quality-derived indicators and support vector regression method', *Energies (Basel)*, Vol. 18, p.5600, DOI: 10.3390/en18215600.