

Enhancing Coal Seam Detection through Data-Driven Borehole Log Interpretation

*Vu Hong Duong, Nguyen Minh Hoa**

Hanoi University of Mining and Geology, Vietnam

Received May 5, 2025; Accepted August 28, 2025

Abstract

Coal energy remains one of the most vital fossil fuel sources used worldwide. Coal has played a significant role in driving the economic and industrial growth of many countries across the globe. To optimise efficiency in its extraction, exploring and identifying coal deposits through geological and geophysical methods is essential, with well-logging geophysics standing out as a key technique. The analysis and interpretation of well-logging geophysical data provide precise insights critical for developing reliable geological and stratigraphic models, which are fundamental for ensuring safe and efficient coal mining operations. Globally, scientists have increasingly turned to artificial intelligence and machine learning to unlock the full potential of extensive datasets for coal assessment, exploration, and discovery. In a comprehensive study, researchers applied various machine learning techniques to address the challenge of determining the depth and thickness of coal seams using well-logging geophysical data from multiple wells in a coal mine in Quang Ninh province, Vietnam. When validated against sample data from the studied site, the machine learning models achieved an accuracy of up to 89%. These initial results underscore the transformative potential of emerging technologies, offering significant time savings and improved accuracy in assessing coal seam depth and thickness from well-logging geophysical data.

Keywords: Coal mining; Well-logging data; Machine learning; Data-driven; Quang Ninh province.

1. Introduction

Although renewable energy sources are gaining prominence worldwide, coal continues to play a vital role in the energy sector, particularly in developing nations where it provides a reliable and accessible electricity supply. According to the International Energy Agency (IEA), coal accounted for approximately 36% of global electricity production in 2022. Beyond its role in power generation, coal is essential in industries such as steel and cement manufacturing. Accurate identification of coal seams is crucial for optimizing mining operations, improving resource assessment, and enhancing mine planning, while also contributing to socio-economic benefits such as employment opportunities and regional development.

Inaccurate identification or oversight of coal seams can lead to suboptimal resource extraction, operational inefficiencies, and missed opportunities for reserve utilization. Conversely, precise seam delineation enhances mining accuracy and guides effective production strategies. Well logging, which measures properties such as gamma ray intensity, bulk density, and electrical resistivity, acts as a primary method for distinguishing coal from surrounding lithologies like shale or sandstone [1-2]. Traditional approaches, often reliant on subjective interpretation and empirical relationships [3], encounter challenges when dealing with noisy data or large-scale datasets. Machine learning (ML) techniques, however, offer substantial advancements in automation, accuracy, and pattern recognition for subsurface geological analysis [4], leveraging coal's distinct properties, such as low gamma radiation and density [5-6].

Key well logging measurements, including gamma ray logs, density logs (typically 1.2–1.5 g/cm³ for coal), and resistivity logs, provide essential insights for coal seam identification [8-9]. Historically, manual thresholding and cross-plot analysis, guided by geological expertise, were

used [10-11], but these methods are labour-intensive and prone to errors. The complexity of heterogeneous datasets and overlapping petrophysical signatures between coal and rocks like organic-rich shales further complicates accurate interpretation.

To address these limitations, ML algorithms such as Support Vector Machines (SVM), Random Forests (RF), and Convolutional Neural Networks (CNNs) have been utilised for automated coal seam detection, improving classification accuracy and extracting essential parameters like seam thickness and quality [12-13]. For example, Wang et al. demonstrated that ML-based methods significantly enhanced seam thickness predictions from well logs compared to manual techniques [14]. The emergence of deep learning (DL), driven by better computational capabilities and access to larger datasets [15], has further advanced coal seam classification. Research by Kumar *et al.* [16] and Banerjee *et al.* [17] showed that DL frameworks outperform traditional methods, while Yu *et al.* [18] successfully employed CNNs to detect subtle patterns in well log data.

A growing trend involves combining well logging data with seismic attributes to enhance the spatial characterization of coal seams, overcoming the limitations of log-based analysis alone. Ding *et al.* [19] demonstrated that integrating multimodal data offers a more comprehensive understanding of coal-bearing formations. Despite these advances, challenges persist. ML models can struggle with noisy or incomplete well log data, deep learning methods require extensive labelled datasets and substantial computational power, and model generalization across different coal basins is often restricted [20]. Additionally, an overreliance on standard logs like gamma ray may miss valuable insights from other measurements, such as neutron porosity or acoustic logs.

To address these challenges, this study proposes a robust ML framework that integrates Random Forests, SVMs, and K-Nearest Neighbours (KNN) to enhance classification accuracy and reliability. The aim is to develop an automated, scalable solution that improves operational precision, minimizes misclassification, and supports effective coal resource management across diverse geological contexts.

2. Database

2.1. Overview of data

The dataset used in this research consists of well-logging data gathered from three wells at a coal mine in Quang Ninh province, Vietnam. It includes measurements of various parameters, such as natural gamma, resistivity, calliper, and density. Common logs like gamma ray, resistivity, density, and calliper play a crucial role in distinguishing coal seams from surrounding lithologies, including shale, sandstone, and siltstone. These tools not only help identify the location and thickness of coal seams but also provide information on coal quality, supporting mine planning and resource assessment efforts.

The gamma ray log measures natural gamma radiation emitted by radioactive elements like uranium, thorium, and potassium present within rock formations. For identifying coal seams, this log is especially useful because coal, a sedimentary rock rich in organic material, generally has low levels of radioactive minerals, leading to noticeably lower gamma ray readings compared to lithologies such as shale or sandstone [5]. These gamma-ray signatures help establish empirical relationships to distinguish coal-bearing layers from surrounding non-coal strata [7]. However, organic-rich shales can sometimes produce similar Gamma Ray responses, which necessitates combining this data with other logs to enhance identification accuracy.

The Resistivity log measures a rock's resistance to electrical current flow, influenced by factors such as moisture content, porosity, and mineral composition. Coal generally shows higher resistivity due to its low water content and insulating properties, in contrast to water-saturated rocks like sandstone or shale. Besides aiding in coal seam detection, Resistivity logs are useful for evaluating coal bed methane reservoirs, which are a key part of energy resource assessment. When combined with Gamma Ray data, Resistivity logs offer a complementary tool for lithological differentiation, especially in geologically complex areas.

The density log, which measures bulk density in g/cm^3 through gamma-ray scattering, is another essential tool for coal seam identification. Coal generally has a lower density compared to shale or sandstone, making it a reliable indicator for finding coal seams. When used alongside other logs, density measurements not only help locate coal seams but also allow for the assessment of coal quality through variations in density related to ash content and porosity [8].

Analysis of the sample data shows three main lithological categories within the wells: sand, shale, and coal seams, with layer thicknesses ranging from a few centimetres to several metres (Figure 1). The arrangement of rock layers across the two wells is highly complex, characterised by interbedded strata and mainly thin layers. This indicates that, despite their proximity within the same mining area, the depth intervals of lithological layers vary greatly between wells, making it difficult to predict the lithological sequence in a new well without reference data.

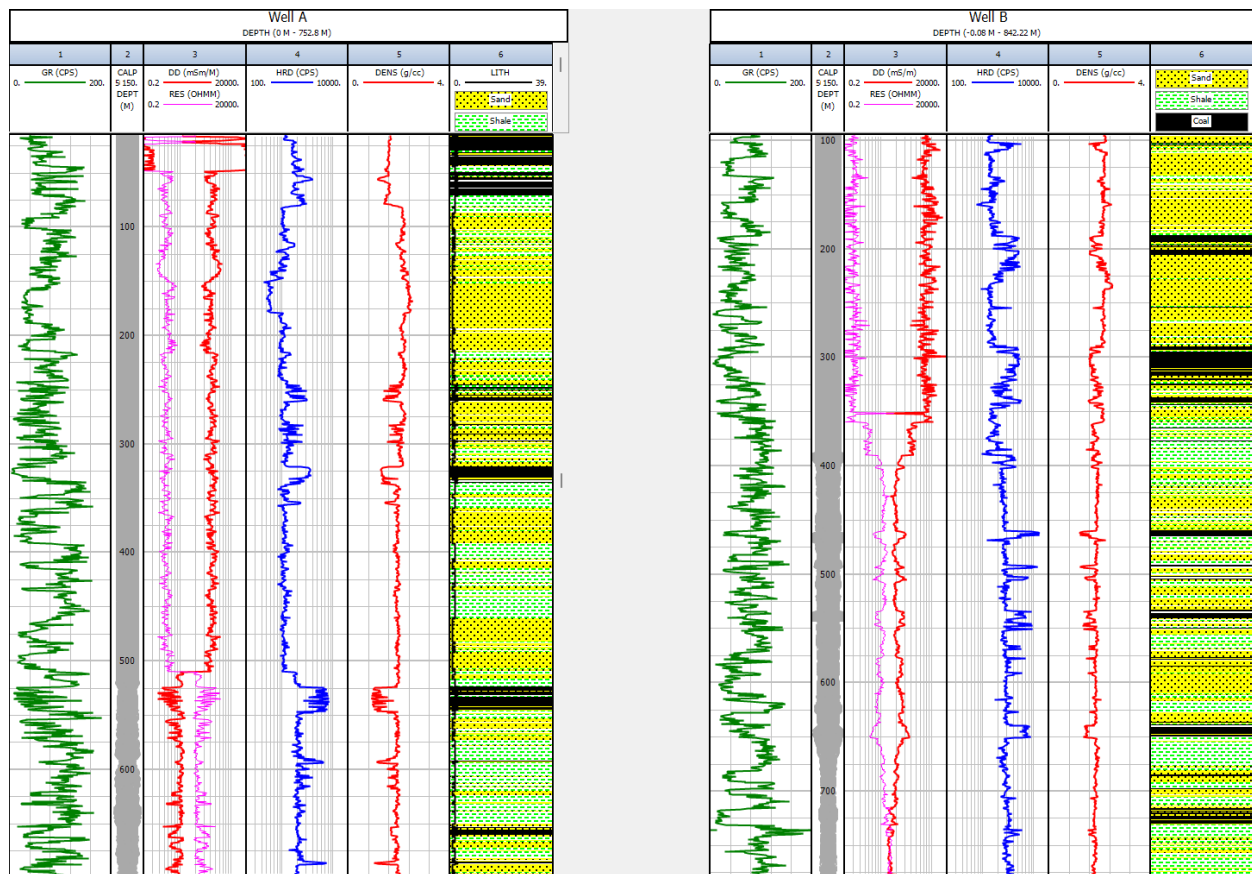


Figure 1. Well logging data of well A and B in study area.

2.2. Detect and eliminate outliers from the database

In geophysical data analysis, particularly within the coal industry, well logging datasets are often subject to noise, measurement errors, and skewed distributions. Parameters such as gamma ray intensity, resistivity, and density frequently show non-normal characteristics due to geological variability, borehole conditions, or equipment issues. Traditional statistical methods, like the Z-score approach, which rely on the mean and standard deviation, are less effective in these situations because they are sensitive to extreme values and assume a Gaussian distribution. Robust statistics offers a more dependable alternative by using estimators such as the median and the Interquartile Range (IQR) to identify and eliminate outliers, thus maintaining data integrity for subsequent geological interpretations [25]. This method employs the IQR to detect outliers by calculating the lower and upper bounds as follows: $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$, where $Q1$ and $Q3$ are the first and third quartiles respectively, and

IQR is the difference between Q3 and Q1. Data points outside these bounds are marked as outliers and removed from further analysis.

This method is reliable for non-normal distributions and offers practical simplicity and flexibility, making it ideal for automated pre-processing in coal exploration. It surpasses parametric approaches in datasets with heavy-tailed distributions or considerable noise, effectively removing outliers—such as physically impossible negative density values—while maintaining the dataset's main trends. As a result, in this study, the authors used the IQR method to exclude outliers, as shown in Figure 2.

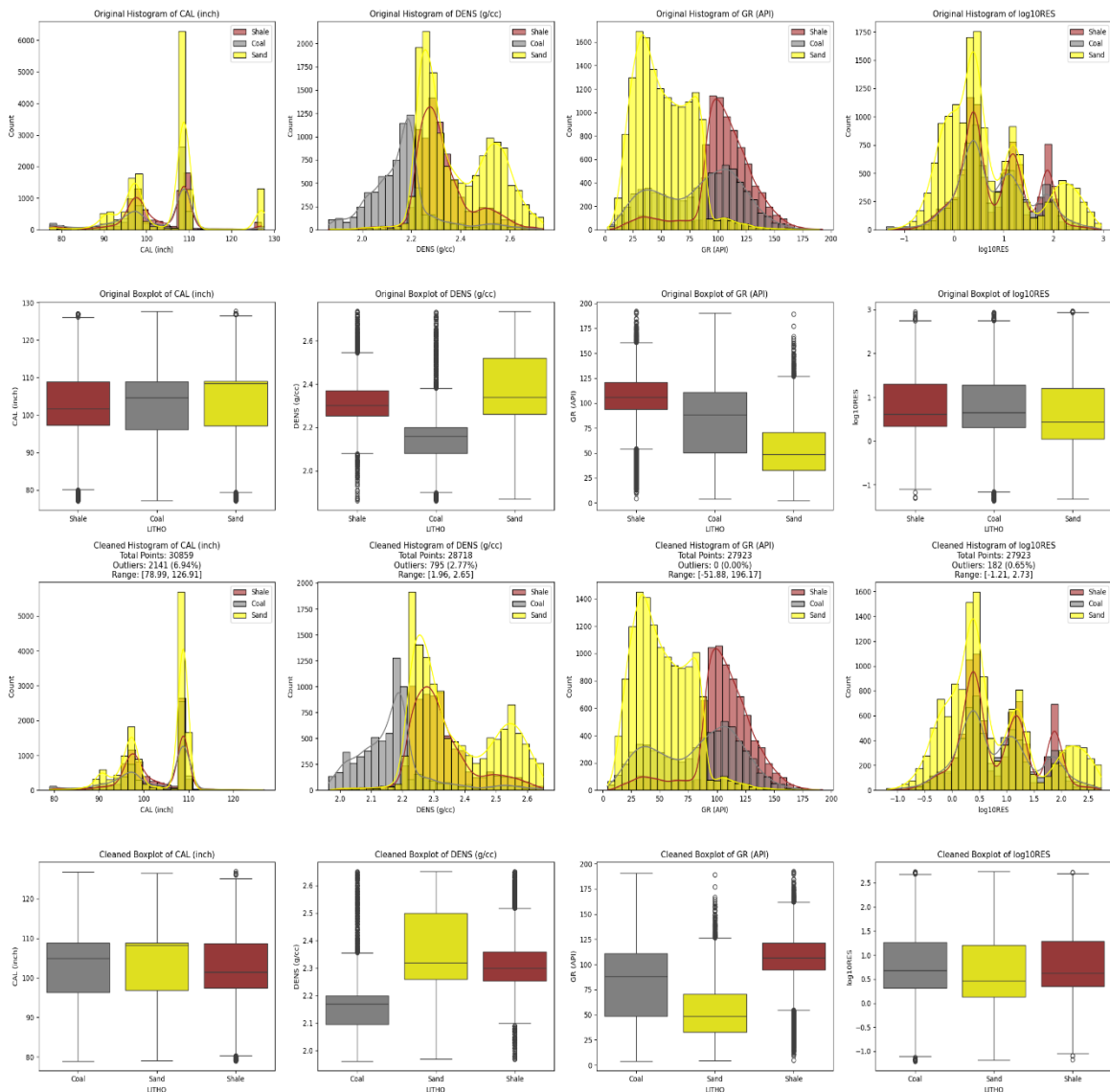


Figure 2. Original and cleaned datasets.

3. Methodology and results

The primary aim of this study is to investigate and assess the performance of machine learning methods in addressing the challenge of lithological layer prediction for wells A and B in the X coal mine, utilizing well-logging parameters such as natural gamma (GR), density gamma (DENS), resistivity (RES), and calliper (CAL), alongside the vertical depth distribution of lithological categories (sand, shale, and coal). This task is framed as a classification problem, well-suited for supervised machine learning approaches. To tackle this objective, the

authors implemented four distinct algorithms: K-Nearest Neighbors (KNN) and Random Forest (RF), each leveraging unique theoretical foundations and tailored to specific analytical contexts.

Model performance was assessed using the following evaluation metrics. Based on the confusion matrix (True Positive - TP, True Negative - TN, False Positive - FP, False Negative - FN), key metrics include:

$$\text{Precision}_i = (\text{TP}_i) / (\text{TP}_i + \text{FP}_i)$$

$$\text{Recall}_i = (\text{TP}_i) / (\text{TP}_i + \text{FN}_i)$$

$$\text{F1}_i = 2 (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

where: TP_i : True positives, i.e., data points correctly classified as belonging to class "i"; FP_i : False positives, i.e., data points incorrectly classified as class "i"; FN_i : False negatives, i.e., data points of class "i" incorrectly classified as belonging to other classes.

The classification results indicate that the Random Forest (RF) model surpasses both the K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) algorithms, particularly in coal seam identification (Table 1). Analysis of the confusion matrices for the training and testing datasets reveals that RF consistently delivers superior precision, recall, and F1-scores across all lithology classes, with a marked enhancement in detecting coal layers.

Table 1. Performance comparison between 3 models

Model	Litho	Training data			Testing data		
		Precision	Recall	F1 score	Precision	Recall	F1 score
KNN	Coal	0.8031	0.8806	0.8401	0.7937	0.8775	0.8335
	Sand	0.9393	0.8852	0.9115	0.9415	0.8722	0.9055
	Shale	0.8535	0.8816	0.8673	0.8230	0.8733	0.8474
RF	Coal	0.8191	0.8913	0.8537	0.8207	0.8903	0.8541
	Sand	0.9472	0.8952	0.9205	0.9504	0.8888	0.9186
	Shale	0.8669	0.8961	0.8813	0.8479	0.8969	0.8717
SVM	Coal	0.8165	0.6139	0.7008	0.8216	0.6185	0.7057
	Sand	0.8315	0.9392	0.8820	0.8253	0.9423	0.8799
	Shale	0.8358	0.8224	0.8291	0.8384	0.8067	0.8222

During the training phase, the Random Forest (RF) algorithm shows a notably lower misclassification rate compared to K-Nearest Neighbours (KNN) and Support Vector Machine (SVM). For coal seam classification, RF attains an F1-score of 0.85, surpassing KNN (0.84) and SVM (0.70), which indicates better precision and recall. This performance benefit comes from RF's ensemble learning method, which combines many decision trees to better capture complex decision boundaries and manage high-dimensional data effectively. The aggregation process inherent in RF decreases sensitivity to noise and reduces overfitting. Conversely, KNN and SVM's reliance on distance-based metrics makes them more vulnerable to issues related to feature scaling and overlapping class distributions, especially in transitional zones between shale and coal, raising their misclassification rates. On the testing dataset, RF continues to demonstrate strong generalisation capabilities, achieving a recall of 0.89 for coal layers, compared to 0.87 for KNN and 0.62 for SVM. These results highlight RF's reliability and effectiveness in geological classification tasks, establishing it as a preferred model for coal seam identification and resource estimation.

To evaluate the real-world usefulness of the classification models, a blind test was performed with data from a previously untested well (Well C) within the study area. This dataset was set aside during model development, guaranteeing an unbiased assessment of how the models perform in practice. Blind testing is an essential step in validating machine learning models, as it demonstrates their ability to generalise beyond the training data and detects potential overfitting by highlighting differences between training and real-world performance. Additionally, it provides a neutral evaluation by removing prior data exposure and assesses model robustness under different data distributions—an important factor for field deployment. The results of the blind test for Well C, shown in Figure 3, confirm the superiority of the RF model, which achieved an F1-score of 0.736 in coal seam detection.

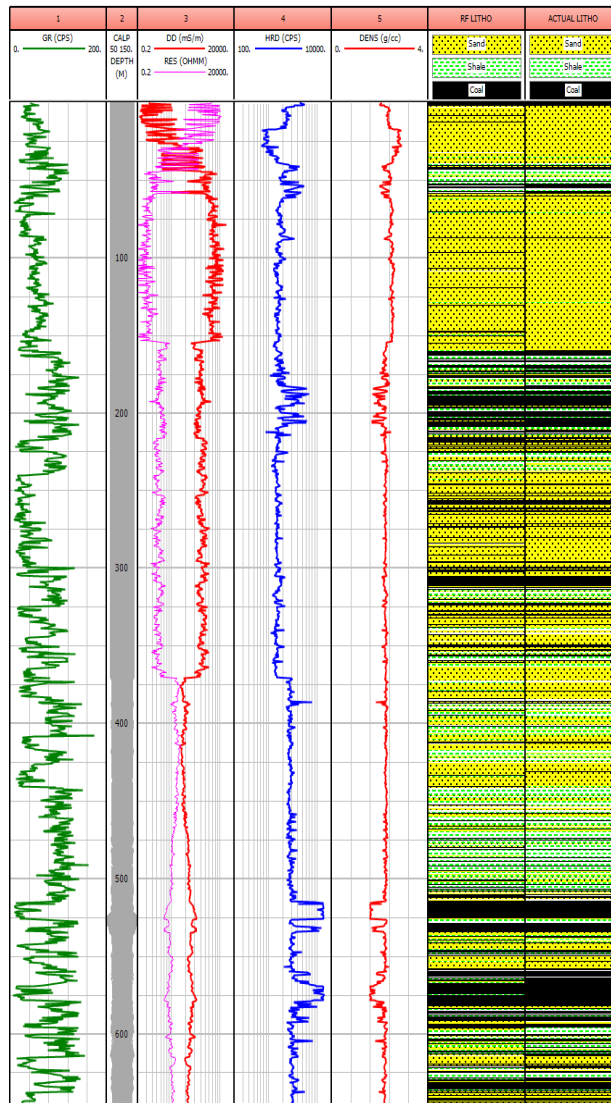


Figure 3. Blind test for Well C.

4. Conclusion and recommendations

This study demonstrates the superior abilities of machine learning-based lithology classification compared to traditional coal seam identification methods. Conventional techniques, such as manual interpretation of well logs or reliance on empirical thresholds, often encounter challenges like inconsistency, dependency on subjective expertise, and difficulty in identifying complex lithological boundaries. Conversely, the Random Forest (RF), K-Nearest Neighbours

(KNN), and Support Vector Machine (SVM) models developed here provide a more accurate, automated, and scalable approach for classifying sand, shale, and coal formations.

Using Grid Search for optimisation, the RF model was found to be the most effective, showing exceptional accuracy and dependability compared to other methods. Unlike traditional techniques, which struggle in complex geological environments, RF employs an ensemble of decision trees to identify subtle patterns within well log data, boosting both predictive accuracy and robustness. A blind test on data from a previously unseen well (Well C) confirmed the RF model's ability to achieve high classification accuracy, demonstrating its strong generalisation to new datasets.

In summary, this study emphasises the transformative potential of machine learning in coal seam identification, offering an objective, efficient, and accurate alternative to traditional techniques. Future research will aim to expand the dataset, add more well-log parameters, and explore deep learning methods further to improve classification performance and flexibility across different geological environments.

Acknowledgements: *This research is funded by Hanoi University of Mining and Geology under grant number T25-07.*

References

- [1] Gordon HW, Jr., Kehn TM, Carter MD, Culbertson WC. Coal resource classification system of the U.S. Geological Survey. Geological Survey Circular 891. U.S. Geological Survey, 2016.
- [2] Teshome AM, Berhanu G, Hadush G, Fanta M, Borko A. Potential Assessment of Coal Deposits in Gamo Zone, Kucha Woreda, Ethiopia. *Petroleum and Coal*, 2024; 66(2): 632-650.
- [3] Hatherly P. Overview on the application of geophysics in coal mining. *International Journal of Coal Geology*, 2013, 114: 74-84.
- [4] Zhou B, O'Brien G. Improving coal quality estimation through multiple geophysical log analysis. *International Journal of Coal Geology*, 2016; 167: 75-92.
- [5] Zhang N, Liu C. Radiation characteristics of natural gamma-ray from coal and gangue for recognition in top coal caving. *Scientific Reports*, 2018; 8: Article 190.
- [6] Bemgba B. Nyakuma, Nasirudeen M. Baba, Aliyu Jauro, Segun A. Akinyemi, Mohammed S. Isyaka, John V. Anyam, Olagoke A. Oladokun. Compositional, Structural, Mineralogical and Thermal Characterisation of Duduguru Coal from Benue Trough Nigeria for Potential Energy Recovery and Utilization. *Petroleum and Coal*, 2024; 66(3): 871-878.
- [7] Jalil S, Rashid M. Analysis of Natural Radioactivity in Coal and Ashes from a Coal Fired Power Plant. *Chemical Engineering Transactions*, 2015; 45: 1549-1554.
- [8] Shao X, Sun Y, Sun J, Tang D, Xu H, Dong X, Lü Y. Log interpretation for coal petrologic parameters: A case study of Hancheng mining area, Central China. *Petroleum Exploration and Development*, 2013; 40(5): 599-605.
- [9] Srinaiah J, Raju D, Udayalaxmi G, Ramadass G. Application of well logging techniques for identification of coal seams: A case study of Auranga Coalfield, Latehar District, Jharkhand State, India. *Journal of Geology & Geophysics*, 2018; 7(1): Article 322.
- [10] Ren P, Xu H, Tang D, Li Y, Sun C, Tao S, Li S, Xin F, Cao L. The identification of coal texture in different rank coal reservoirs by using geophysical logging data in northwest Guizhou, China: Investigation by principal component analysis. *Fuel*, 2018; 230: 258-265.
- [11] Mclean C. Pseudo proximate analysis: Method using wireline logs to estimate components of coal bearing rock matrix without control data. Master's thesis, University of the Western Cape 2015, Department of Earth Science, 2015.
- [12] Zhong R. , Johnson Raymond L., Zhongwei C. Using Machine Learning Methods To Identify Coal Pay Zones from Drilling and Logging-While-Drilling (LWD) Data. *SPE J.*, 2020, vol. 25, pp.1241-1258.
- [13] Yegireddi S, Bhaskar GU. Identification of coal seam strata from geophysical logs of borehole using adaptive neuro-fuzzy inference system. *Journal of Applied Geophysics*, 2009; 67(1): 9-13.
- [14] Wang Z, Cai Y, Liu D, Qiu F, Sun F, Zhou Y. Intelligent classification of coal structure using multinomial logistic regression, random forest and fully connected neural network with multisource geophysical logging data. *International Journal of Coal Geology*, 2023; 268: 104208.
- [15] Shi J, Zeng L, Dong S, Wang J, Zhang Y. Identification of coal structures using geophysical logging data in Qinshui Basin, China: Investigation by kernel Fisher discriminant analysis. *International Journal of Coal Geology*, 2020; 217: 103314.

- [16] Kumar T., Seelam NK, Rao GS. Lithology prediction from well log data using machine learning techniques: A case study from Talcher coalfield, Eastern India. *Journal of Applied Geophysics*, 2022; 199: 104605.
- [17] Banerjee A, Mukherjee B, Sain K. Machine learning assisted model based petrographic classification: A case study from Bokaro coal field. *Acta Geodaetica et Geophysica*, 2024; 59(3): 463-490.
- [18] Shi Y, Liao J, Gan L, Tang R. Lithofacies prediction from well log data based on deep learning: A case study from Southern Sichuan, China. *Applied Sciences*, 2024; 14(18): 8195.
- [19] Ding T, Wu Y, Wang L, Nie Z, Zhang L. A case study comparing methods for coal thickness identification in complex geological conditions. *Applied Sciences*, 2024; 14(22): 10381.
- [20] Banerjee A, Mukherjee B, Sain K. Machine learning assisted model based petrographic classification: A case study from Bokaro coal field. *Acta Geodaetica et Geophysica*, 2024; 59(3): 463-490.

To whom correspondence should be addressed: Dr. Nguyen Minh Hoa, Hanoi University of Mining and Geology, Vietnam, E-mail: nguyenminhhoa@hug.edu.vn