

UDC 550.832

<http://doi.org/10.18799/24131830/2026/6/5224>

Scientific paper



Machine learning-based analysis of borehole geophysical logs in coal mining

Vu Hong Duong¹, Nguyen Tien Hung^{1✉}, Nguyen Minh Hoa²,
Nguyen The Vinh¹, Nguyen Xuan Duy³

¹ Hanoi University of Mining and Geology, Hanoi, Vietnam

² Office of the Government, Hanoi, Vietnam

³ Vietnam Petroleum Institute, Hanoi, Vietnam

✉ nguyentienhung.dk@humg.edu.vn

Abstract. Relevance. Coal remains a crucial fossil fuel source globally, significantly contributing to economic and industrial development. Efficient coal extraction relies heavily on accurate identification and evaluation of coal deposits, where geophysical methods, particularly well-logging, play a pivotal role. **Aim.** To enhance the accuracy and efficiency of determining coal seam depth and thickness by applying advanced machine learning techniques to well-logging geophysical data. This approach seeks to optimize coal exploration by automating data analysis, reducing human error, and improving the precision of subsurface assessments critical for efficient resource extraction. **Objects.** Well-logging geophysical data collected from multiple boreholes in a coal mine in Quang Ninh province, Vietnam. This region, a hub for coal production, provides a rich dataset for testing and refining machine learning models tailored to real-world mining conditions. **Methods.** A variety of machine learning algorithms, including regression and classification techniques, were utilized to process and analyze the well-logging data. These models were trained and validated using labeled datasets derived from borehole measurements, enabling accurate predictions of coal seam depth and thickness with enhanced precision and reliability. **Results.** The machine learning models achieved an impressive predictive accuracy of up to 89% when compared to actual sample data. These findings demonstrate the transformative potential of integrating artificial intelligence into geophysical analysis. By offering improved reliability, time efficiency, and accuracy in coal seam characterization, this approach supports more sustainable and cost-effective coal extraction practices, paving the way for broader adoption in the mining industry.

Keywords: coal mining, well-logging data, Machine Learning, Data-driven, Quang Ninh province

Acknowledgements: This research was funded by Hanoi University of Mining and Geology under grant no. T25-07.

For citation: Vu Hong Duong, Nguyen Tien Hung, Nguyen Minh Hoa, Nguyen The Vinh. Machine learning-based analysis of borehole geophysical logs in coal mining. *Bulletin of the Tomsk Polytechnic University. Geo Assets Engineering*, 2026, vol. 337, no. 6, pp. 44–55. <http://doi.org/10.18799/24131830/2026/6/5224>

Conflict of interest: none.

УДК 550.832
<http://doi.org/10.18799/24131830/2026/6/5224>
Научная статья

Анализ данных геофизических исследований скважин на основе машинного обучения в угольной добыче

Бу Хонг Зыонг¹, Нгуен Тиен Хунг^{1✉}, Нгуен Минь Хоа²,
Нгуен Тхе Винь¹, Нгуен Суан Зуи³

¹ Ханойский университет горного дела и геологии, Вьетнам, г. Ханой

² Аппарат Правительства Вьетнама, Вьетнам, г. Ханой

³ Вьетнамский нефтяной институт, Вьетнам, г. Ханой

✉ nguentienhung.dk@humg.edu.vn

Аннотация. Актуальность. Уголь остается важнейшим источником ископаемого топлива во всем мире, внося значительный вклад в экономическое и промышленное развитие. Эффективная добыча угля в значительной степени зависит от точной идентификации и оценки угольных месторождений, где геофизические методы, в частности каротажа скважин, играют ключевую роль. **Цель:** повысить точность и эффективность определения глубины и толщины угольного пласта путем применения передовых методов машинного обучения к геофизическим данным каротажа скважин. Этот подход направлен на оптимизацию разведки угля путем автоматизации анализа данных, снижения человеческих ошибок и повышения точности оценок недр, критически важных для эффективной добычи ресурсов. **Объекты:** данные геофизических исследований скважин, собранные из нескольких скважин на угольной шахте в провинции Куангнинь, Вьетнам. Этот регион, являющийся центром добычи угля, предоставляет богатый набор данных для тестирования и уточнения моделей машинного обучения, адаптированных к реальным условиям добычи. **Методы.** Для обработки и анализа данных каротажа скважин использовались различные алгоритмы машинного обучения, включая методы регрессии и классификации. Эти модели были обучены и проверены с использованием маркированных наборов данных, полученных из измерений скважин, что позволяет делать точные прогнозы глубины и толщины угольного пласта с повышенной точностью и надежностью. **Результаты.** Модели машинного обучения достигли впечатляющей точности прогнозирования до 89 % по сравнению с фактическими данными образцов. Эти результаты демонстрируют преобразующий потенциал интеграции искусственного интеллекта в геофизический анализ. Предлагая повышенную надежность, эффективность времени и точность в характеристике угольного пласта, этот подход поддерживает более устойчивые и экономически эффективные методы добычи угля, прокладывая путь для более широкого внедрения в горнодобывающей промышленности.

Ключевые слова: добыча угля, данные геофизических исследований скважин, машинное обучение, анализ данных, провинция Куангнинь

Благодарности: Исследование финансируется Ханойским университетом горного дела и геологии по гранту № T25-07.

Для цитирования: Бу Хонг Зыонг, Нгуен Тиен Хунг, Нгуен Минь Хоа, Нгуен Тхе Винь, Нгуен Суан Зуи. Анализ данных геофизических исследований скважин на основе машинного обучения в угольной добыче. *Известия Томского политехнического университета. Инжиниринг георесурсов*, 2026, Т. 337, № 6, С. 44–55. <http://doi.org/10.18799/24131830/2026/6/5224>

Конфликт интересов: отсутствует.

Introduction

The use and exploitation of coal continue to be essential processes in the global energy market in spite of the gradual shift to renewable sources of energy. Coal continues to be part of energy generation, especially in the developing world, where it is a source of cheap and stable electricity that can sustain industrialization and economic development. According to the International Energy Agency (IEA), in 2022, coal accounted for about 36% of the global electricity generation, indicating its continued significance. Beyond electricity gen-

eration, coal is also a vital raw material for several industries, such as steel manufacturing, where metallurgical coal is indispensable for the production of coke, and in the manufacture of cement, where it is used both as a fuel and as a chemical feedstock. Accurate identification and mining of coal seams can greatly affect the efficiency and affordability of such industrial processes, and thus advanced exploration methods are crucial. Furthermore, coal exploration and mining have extensive socio-economic consequences. In regions where coal is abundant, mining operations offer job opportu-

nities, consolidate local economies, and boost national income through exports. Coal seam identification and classification in subsurface geology is a basic component of coal exploration and exploitation, supplying a key support for resource estimation, mine planning, and economic feasibility studies.

Incorrect classification or omission of coal seams can result in inferior resource recovery, high operating costs, or missed reserves; whereas accurate delineation allows for efficient mining methods and better production strategies. In this scenario, well-logging – a collection of geophysical methods that directly measure physical properties like gamma ray intensity, density, and resistivity in boreholes – has become an essential tool for discriminating coal seams from the enveloping lithologies, including shale, sandstone, and siltstone [1, 2]. Interpretation of well log data has traditionally relied on subjective observation and empirical correlations with some parameters such as low gamma ray values and reduced density with coal [3, 4]. Yet, these traditional approaches are usually plagued by subjectivity in interpretation, difficulty in dealing with noisy or complicated data sets, and lack of effectiveness when extended to large volumes of data from numerous wells. The advent of machine learning (ML) methods presents a groundbreaking opportunity to advance the effectiveness, precision, and automation associated with coal seam classification and interpretation. The innovation contradicts the limitations inherent in conventional approaches and enables the recognition of new patterns applicable to subsurface geology [5, 6]. Coal, being a sedimentary rock with certain physical properties, generally has low natural gamma radiation levels because of its limited content of radioactive minerals, along with a lower density in comparison to the surrounding non-coal rocks [7]. It is these properties that are the foundation for lithological discrimination in well-logging records.

Well-logging is a geophysical method that logs several measurements indicative of the physical properties of formations in the subsurface. Natural radioactivity, for example, is measured by gamma ray logs and tends to be low in coal because of its organic nature, making it a significant marker for identifying coal seams [8]. Density logs, recording bulk density, also distinguish coal (typically between 1.2 and 1.5 g/cm³) from denser adjacent rocks like sandstone or shale [9]. Resistivity logs, on the other hand, give data about electrical conductivity, with usually high values in coal due to its relatively low water content compared to other lithological types. The combination of these logs gives a composite dataset that corresponds to the lithological and petrophysical characteristics of coal-bearing sequences [10, 11]. Conventionally, coal seam identification was achieved through manual thresholding and cross-plotting of log responses assisted by geological experience [12, 13].

Yet these methods are time-consuming, prone to error, and less successful for application with highly variable or noisy datasets. Moreover, heterogeneity of the subsurface, measurement noise, and overlap of signatures between coal and certain non-coal lithologies (e.g., organic-rich shales) render manual interpretation problematic. The limitations of traditional approaches have spurred the adoption of machine learning techniques, which excel at handling complex, multivariate datasets. Moreover, the increasing availability of high-resolution logging data from modern tools necessitates advanced analytical frameworks capable of processing multidimensional datasets efficiently. Machine learning, with its ability to discern patterns, model non-linear relationships, and adapt to diverse data inputs, offers a robust solution to these challenges. By integrating ML algorithms – ranging from supervised methods like Support Vector Machines (SVM) and Random Forests (RF) to deep learning approaches such as Convolutional Neural Networks (CNNs) – researchers and practitioners can automate coal seam detection, improve classification precision, and extract quantitative parameters such as seam thickness and quality from well logs. Early applications of ML to coal seam identification focused on supervised learning algorithms [14–16]. For instance, Guo et al. employed machine learning to predict coal seam thickness from well logs, demonstrating improved accuracy over manual methods [17]. Furthermore, the integration of ML with well-logging data facilitates real-time decision-making, reduces human bias, and enables scalability across diverse geological settings, from the coal basins of India and China to those in Australia and the United States.

Recent years have witnessed a surge in the use of advanced ML and deep learning (DL) methods for coal seam classification, driven by the increasing computational power and availability of large datasets [18, 19]. Deep learning, in particular, has shown promise in extracting intricate patterns from well-logging data. Kumar et al. introduced a deep learning framework for coal seam identification, leveraging the ability of neural networks and other machine learning techniques to model non-linear relationships across multiple log types [20]. Similarly, Maxwell et al. applied a DL-based approach to classify coal and non-coal intervals, achieving superior performance compared to traditional methods [21]. CNNs, adept at processing sequential or spatial data, have also been explored, with Yu et al. demonstrating their effectiveness in identifying coal seams from well logs by capturing subtle variations in log responses [22].

An unexpected yet significant trend in recent literature is the integration of well-logging data with seismic attributes to enhance coal seam characterization. Ding et al. demonstrated that combining well logs with seismic data improves the spatial mapping of coal

seams, overcoming the limitations of log-based analysis alone, such as restricted lateral resolution [23]. This hybrid approach, underscores the potential of multi-modal data fusion in advancing subsurface interpretation, offering a more holistic view of coal-bearing formations.

Despite progress in applying ML to coal seam identification, challenges remain, including handling noisy or incomplete well-logging data, ensuring model generalization across diverse coal basins, and leveraging underutilized log types, such as neutron porosity or acoustic logs. Established ML methods, such as RF, SVM, and K-Nearest Neighbors (KNN), are widely used in geophysical data classification but often struggle with these issues. Meanwhile, DL models require significant computational resources and large labeled datasets, which are frequently unavailable in coal exploration contexts [24]. This study addresses these gaps by developing an optimized ML framework for coal seam classification using well-logging data, focusing on constructing robust predictive models with the potential for future development into a ready-to-use software tool. By integrating RF, SVM, and KNN, the proposed framework emphasizes model optimization through hyperparameter tuning, feature selection, and computational efficiency to enhance classification accuracy, robustness, and adaptability to varied geological settings. This framework lays the groundwork for sustainable coal resource management by providing reliable predictive models, with prospects for deployment as a practical tool for subsurface interpretation.

Data overview

Following the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology [25], the dataset for this study was systematically collected, organized, and prepared to support coal seam classification using ML. The dataset comprises well-logging data from three boreholes in a coal mine located in Quang Ninh province, Vietnam. These data were acquired through geophysical logging operations conducted by deploying specialized tools into the boreholes to measure physical properties of subsurface formations. The dataset includes four primary predictors: Natural Gamma (measured in API units), Resistivity (in ohm-meters), Density (in g/cm^3), and Caliper (in centimeters), each capturing distinct geological characteristics critical for distinguishing coal seams from surrounding lithologies such as shale and sandstone.

The well-logging data were collected by a geophysical service provider during exploration activities in 2023. Logging tools were lowered into the boreholes to record continuous measurements at regular depth intervals (0.1 meter resolution). The data were extracted from three wells, designated as Well 288, Well 269, and Well 302, within the same mining area. Each well

log was accompanied by geological core samples, which were used to label lithological categories (coal, shale, and sandstone) for supervised ML model training. The acquisition ensured high-resolution data, though challenges such as borehole irregularities and tool calibration variations introduced noise in some measurements.

The dataset was organized into a structured, non-relational format, stored as comma-separated value (CSV) files to facilitate processing and analysis. Each CSV file corresponds to a single well, with rows representing depth intervals and columns capturing the four predictors (Gamma Ray, Resistivity, Density, Caliper) alongside a target variable indicating lithology (coal, shale, or sandstone). The dataset contains a total of 12450 records across the three wells, with Well 288 contributing 7529 records, Well 269 (8424 records), and Well 302 (7518 records). Each record represents a depth interval of 0.1 meters, resulting in a dataset size of approximately 1.5 MB per well. The data were discretized using a uniform depth sampling rate, ensuring consistency across wells. The dataset is stored on a local server with restricted access, available only to project researchers under a secure protocol to maintain data integrity and confidentiality.

The dataset exhibits variability due to the complex geological structure of the Quang Ninh coal mine, characterized by interbedded strata with layer thicknesses ranging from a few centimeters to several meters. The predictors were selected based on their established relevance in coal seam identification:

- Gamma Ray: reflects natural radioactivity, typically lower in coal (10–30 API) compared to shale (80–150 API) or sandstone (40–100 API);
- Resistivity: indicates electrical resistance, with coal showing higher values (50–200 ohm-m) due to low water content;
- Density: measures bulk density, with coal ranging from 1.2–1.5 g/cm^3 , significantly lower than shale (2.2–2.7 g/cm^3) or sandstone (2.0–2.6 g/cm^3);
- Caliper: records borehole diameter, used to assess well integrity and correct for measurement artifacts.

Preprocessing followed the CRISP-DM data preparation phase to ensure data quality for ML model training. Missing values were imputed using linear interpolation for continuous predictors, preserving the depth-dependent trends. Outliers, identified using the interquartile range method, were capped to reduce the impact of extreme values. The predictors were normalized to a [0, 1] range to ensure compatibility with ML algorithms like SVM and KNN. The lithology labels, derived from core sample analysis, were encoded as categorical variables (0 for coal, 1 for shale, 2 for sandstone). The dataset is managed under a controlled access regime, stored on a secure server with role-based access limited to authorized researchers.

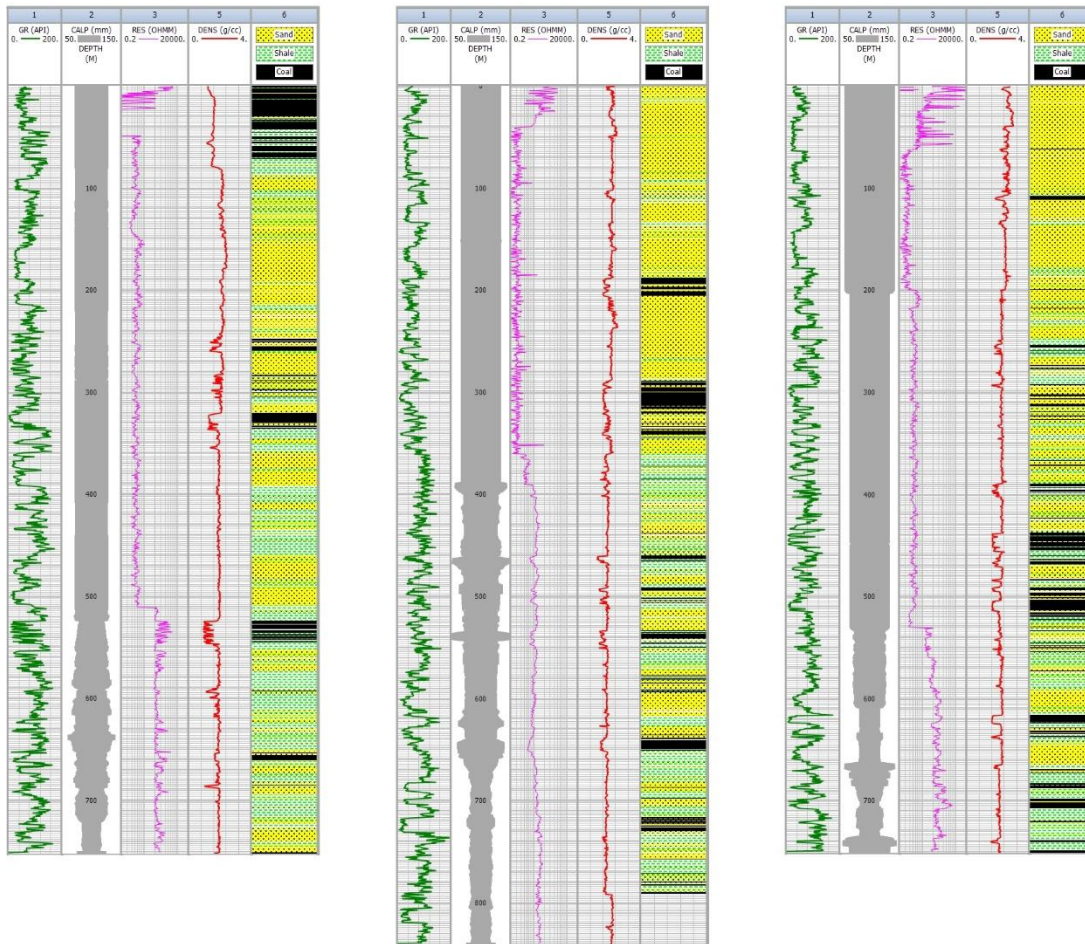


Fig. 1. Well logging data of three wells in study area

Рис. 1. Данные геофизических исследований трех скважин на исследуемой территории

Data retrieval is performed via scripted queries using Python, with Pandas and NumPy libraries for efficient handling. The dataset is version-controlled to track modifications during preprocessing and analysis, ensuring reproducibility. This structured dataset, combined with rigorous preprocessing, provides a robust foundation for developing and evaluating ML models for coal seam classification, as described in subsequent sections.

Based on sample data from three wells, the rock layers here are classified into three main lithological types: sand (48.5%), shale (28%), and coal, which occupy a relatively small proportion (23.5%), with thicknesses varying from a few centimeters to several meters (Fig. 1). The arrangement of rock layers in the three wells appears notably intricate, characterized by interwoven strata and generally thin layer thicknesses. Consequently, it can be inferred that, despite being situated within the same mining area, the depth intervals of the rock layers differ significantly between the wells, complicating the prediction of the lithological sequence in a new well without reference to sample data.

Outlier detection and data preparation

In geophysical data analysis, particularly for coal seam identification, well-logging datasets are prone to noise, measurement errors, and non-normal distributions due to geological heterogeneities, borehole conditions, or equipment malfunctions. These factors can introduce outliers, such as physically implausible negative density values or erratic gamma ray readings, which may distort ML model performance. To address this, a robust outlier detection method based on the Median Absolute Deviation (MAD) was employed as part of the data preparation phase within the CRISP-DM framework. This approach ensures data integrity for subsequent ML-based lithology classification, where the target variable is the categorical lithology type (coal, shale, or sandstone).

The dataset, described previously, comprises four predictors – Natural Gamma, Resistivity, Density and Caliper – derived from well-logging data across three boreholes in Quang Ninh, Vietnam. These predictors are linked to the target variable, which represents lithological categories encoded as integers (0 for coal, 1 for shale, 2 for sandstone). The goal is to predict lithology

based on the physical properties captured by these predictors, leveraging their distinct signatures: low Gamma Ray and Density values, high Resistivity, and stable Caliper measurements typically characterize coal seams compared to shale or sandstone.

Outlier detection utilized the Modified Z-score method, defined as equation (1) [26]:

$$M_i = \frac{0.6745 \cdot (x_i - \bar{x})}{MAD}, \quad (1)$$

where x_i denotes individual data points, and the constant 0.6745 aligns MAD with the standard deviation of a normal distribution. Data points with an absolute Modified Z-score exceeding a predefined threshold typically 3 or 3.5 are flagged as outliers and excluded from further analysis. This threshold was chosen to balance sensitivity to extreme values with preservation of natural geological variability. Approximately 7% of the data were identified as outliers, primarily in Density and Caliper log due to tool malfunctions or borehole irregularities.

The outlier detection was integrated into a broader modeling workflow:

1. Data Understanding: analysis of predictor distributions revealed non-normal behavior, with heavy-tailed distributions in Resistivity and Gamma Ray, justifying the use of robust statistics;
2. Data Preparation: beyond outlier removal, missing values (7% of records) were imputed using linear interpolation, and predictors were normalized to a [0, 1] range to ensure compatibility with ML algorithms. The categorical target variable was balanced using oversampling to address the uneven distribution (coal: 23.5%, shale: 28%, sandstone: 48.5%);
3. Modeling: cleaned dataset, with outliers removed, was used to train ML models to predict lithology, leveraging the relationship between predictor signatures and lithological classes;
4. Evaluation: model performance was assessed using metrics such as accuracy and F1-score, with results presented in Methodology and result section.

This robust approach to outlier detection and data preparation enhances the reliability of the dataset for lithology prediction, mitigating the impact of noise and ensuring that the ML models capture meaningful geological patterns. The results of outlier removal are illustrated in Fig. 2, demonstrating improved data quality for subsequent analysis.

Statistical analysis plays a crucial role in this study by providing a comprehensive understanding of the dataset before model development. It helps identify key trends, detect outliers, and assess the distribution of essential features such as DENS, GR, HRD, and RES. By analyzing minimum, maximum, standard deviation, and percentiles (P10, P50, P90), we ensure data consistency and reliability (Table 1).

Methodology

The objective set by the authors is to try applying and evaluating the effectiveness of machine learning techniques in solving the problem of predicting lithological layers for three wells at the X coal mine based on well-logging parameters, including GR, DENS, RES, CAL, along with the distribution information of lithological groups (sand, shale, and coal) by vertical depth. This task can be categorized as a classification problem – a common type effectively handled by supervised machine learning techniques. To solve the task, the authors applied three prominent algorithms: KNN, RF and SVM, each grounded in unique theoretical frameworks and suited to specific applications. The study was conducted using Python 3.9 on a workstation with an Intel Core i7-11700 CPU, 32 GB RAM, and an NVIDIA RTX 3060 GPU. Libraries included Pandas (v1.3.5) for data handling, NumPy (v1.21.0) for computations, Scikit-learn (v1.0.2) for ML implementation, and Matplotlib (v3.4.3) for visualization. The dataset, stored as CSV files, was processed in a Jupyter Notebook environment, with scripts version-controlled using Git for reproducibility. To evaluate model performance, we partitioned the data using a 75/25 ratio for training and testing, respectively. This 25% hold-out set provided an unbiased validation of the models ability to generalize to new data without overfitting.

Modeling methods

K-Nearest Neighbors [27] is a non-parametric, instance-based learning algorithm that classifies data points based on the majority class among their nearest neighbors in a feature space. In the context of well logging, features such as gamma ray, resistivity, and porosity measurements are used to define the multidimensional space. KNN operates by calculating the distance (typically Euclidean) between a test sample and all training samples.

Table 1. Detailed statistics of the well-logging dataset

Таблица 1. Подробная статистика набора данных ГИС

Parameter Параметр	Min	Max	Mode	Std Dev	P10	P50	P90
CAL (mm)	78.99	126.86	108.99	7.418925	92.78	106.71	109.19
DENS (g/cc)	1.962	2.651	2.25	0.146041	2.14	2.28	2.54
GR (API)	4.14	192.16	100	36.4642	27.451	73.75	122.66
log10RES	-1.2006	2.73412	-0.2596	0.77029	-0.1249	0.542825	1.95041

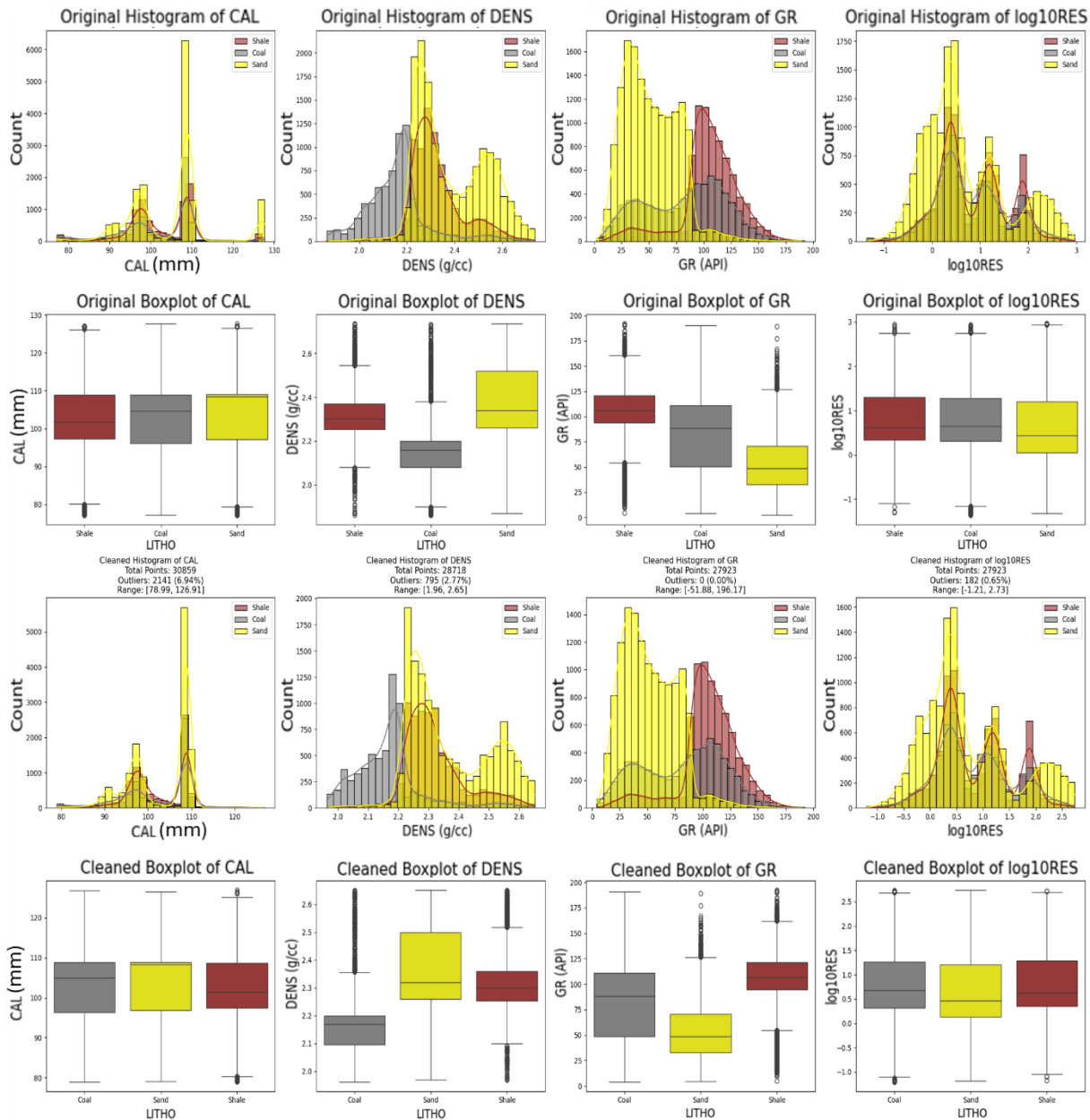


Fig. 2. Original and cleaned datasets

Рис. 2. Исходные и очищенные наборы данных

It then assigns the class most common among the k closest neighbors. Its simplicity makes it intuitive and easy to implement, requiring no explicit training phase. However, KNN performance heavily depends on the choice of k and the quality of the feature space, and it can be computationally intensive for large datasets, such as those derived from extensive well logs, due to its reliance on distance calculations during prediction.

Random Forest [27] in contrast, is an ensemble learning method that constructs multiple decision trees during the training phase and outputs the class that is the mode of the individual trees predictions. Each tree is built on a random subset of the data and features,

introducing diversity and reducing overfitting a common issue in single decision trees. RF can effectively handle high-dimensional data and capture complex relationships between variables like density, neutron porosity, and sonic velocity. Its robustness to noise and missing values, often encountered in geological datasets, makes it particularly suitable. However, RF computational complexity increases with the number of trees, potentially limiting scalability for very large datasets.

Support Vector Machine [27] is a supervised learning algorithm that seeks an optimal hyperplane to separate classes in a high-dimensional space, maximizing the margin between the nearest points (support vectors)

of different classes. SVM transforms the input features into a higher-dimensional space using kernel functions (e.g., radial basis function). SVM can handles non-linearly separable data, such as overlapping lithological signatures. This capability makes SVM highly effective for complex classification tasks where traditional linear methods may struggle.

Table 2. Optimum set of parameters for models

Таблица 2. Оптимальный набор параметров для моделей

Model/Модель	Hyper parameter/Гиперпараметр
KNN	NumNeighbors: 11 Distance: 'cityblock'
RF	NumTrees: 100 MinLeafSize: 5 MaxNumSplits: 100
SVM	Kernel: 'rbf' C: 1.0 Gamma: 'scale' Class Weight: 'balanced'

During the development of classification models, model optimization (fine-tuning) was conducted by adjusting hyperparameters throughout the training to identify the best-performing model aligned with the predefined objective function. The Grid Search algorithm was applied to identify the optimal model by evaluating the predictive performance of different models as hyperparameters were adjusted. The final optimal hyperparameters for each model are listed in Table 2. 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 are defined in equations (2)–(4) [28]:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}, \quad (2)$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}, \quad (3)$$

$$\text{F1}_i = 2 \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (4)$$

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.

Results

A comparative evaluation of three machine learning models – KNN, RF, and SVM – for coal seam classification on Train and Test datasets reveals distinct performance differences in Precision, Recall, and F1-score (Fig. 3). On the Train dataset, RF outperforms with a Precision of 89.42%, Recall of 87.77%, and F1-score of 88.52%, surpassing KNN (88.36%, 86.61%, 87.39%) and SVM (82.79%, 79.18%, 80.39%). Similarly, on the Test dataset, RF maintains its lead with a Precision of 89.20%, Recall of 87.30%, and F1 Score of 88.14%, compared to KNN (87.43%, 87.27%, 86.22%) and SVM (82.84%, 78.92%, 80.26%).

To elucidate these differences, RF algorithmic strengths are notable. RF achieves a lower misclassification rate, particularly in coal layer identification, with an F1 Score of 0.85 compared to KNN 0.84 and SVM 0.70 on the Train dataset. Its ability to aggregate multiple decision trees effectively handles complex decision boundaries and high-dimensional data, reducing overfitting. In contrast, KNN and SVM, reliant on distance metrics, are more susceptible to feature scaling issues and overlapping class distributions, leading to higher errors in transitional zones between shale and coal. On the Test dataset, RF Recall for coal layers reaches 0.89, compared to 0.87 for KNN and 0.62 for SVM, underscoring RF superior generalization and reliability for geological applications.

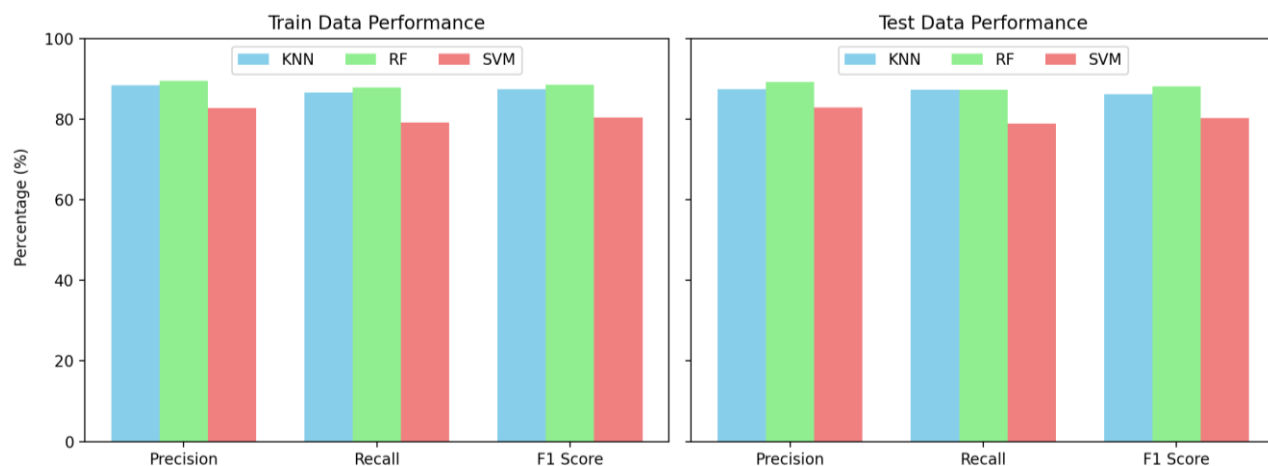


Fig. 3. Model performance comparison

Рис. 3. Сравнение производительности моделей

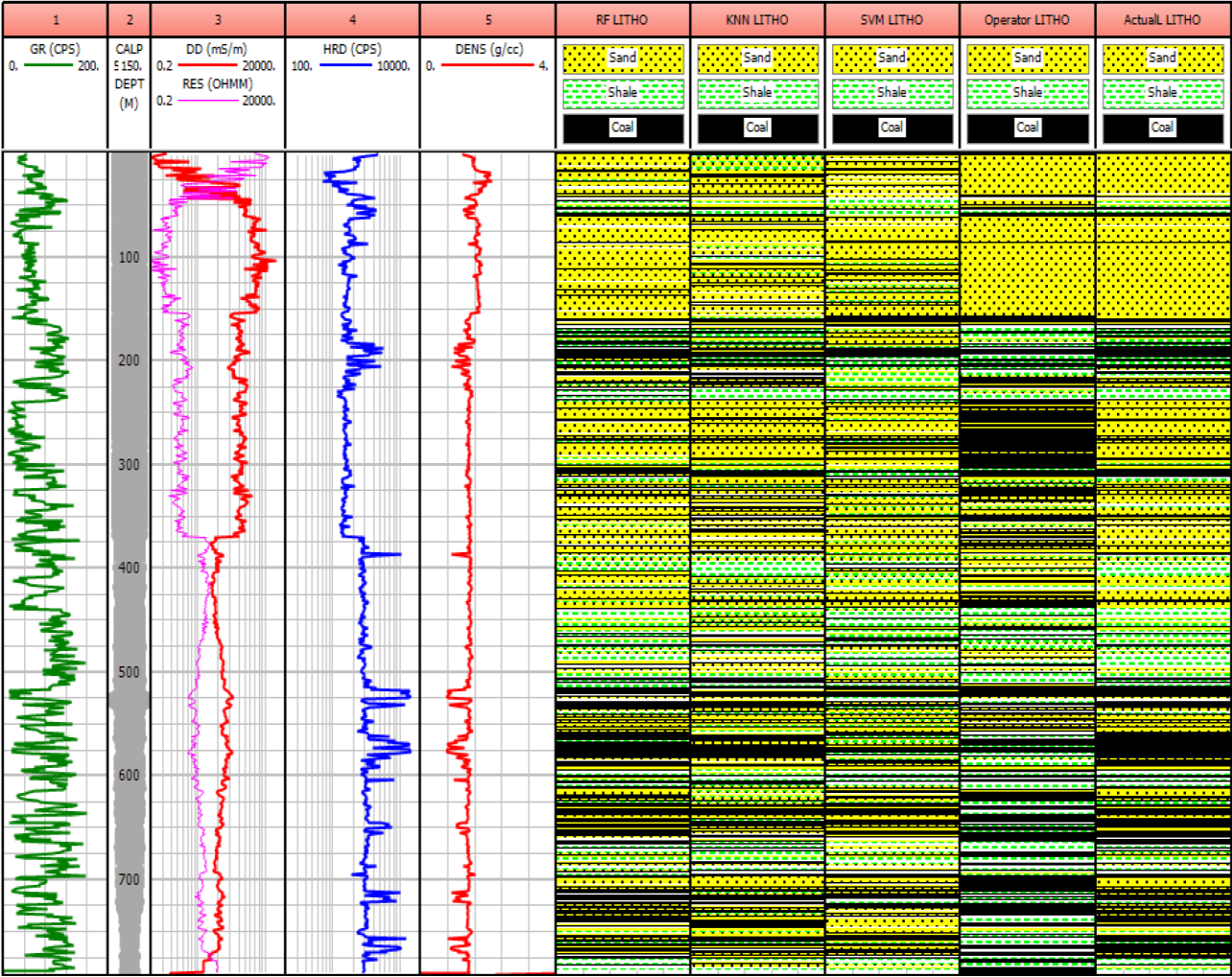


Fig. 4. Blind test for Well X
Рис. 4. Слепой тест для скважины X

The final evaluation of the classification model practical utility involves a Blind Test, utilizing data from a new well (Well X) within the study area (Fig. 4). This well data, previously unused in model development, ensures an unbiased assessment of real-world performance. Blind testing [29] is essential for validating machine learning models, confirming their ability to generalize to unseen data rather than overfitting to the training set. It detects overfitting by revealing discrepancies between training and real-world performance and provides an impartial evaluation by eliminating bias from prior data exposure. Furthermore, it assesses the model reliability under varying data distributions, a critical factor for practical applications.

Fig. 5 presents the "Model Performance Metrics" table, which illustrates the performance of geological prediction models (KNN Lithology, Operator Lithology, RF Lithology, and SVM Lithology) compared to the actual data (Actual Lithology) from the Well X

dataset. The table in Fig. 5 includes evaluation metrics such as Accuracy, Precision, Recall, and F1-score for three main geological classes: coal, sand, and shale. Among the models, RF Lithology stands out with the highest Accuracy of 0.830, outperforming KNN (0.750), Operator (0.628), and SVM (0.710). It also achieves high and balanced Precision, Recall, and F1-score values – particularly for coal (0.861), sand (0.769), and shale (0.832) – indicating strong classification capability. In contrast, the Operator Lithology model shows the lowest performance, with an Accuracy of 0.628 and inconsistent Precision and Recall, suggesting instability possibly due to manual methods or suboptimal input data. SVM achieves moderate performance with an Accuracy of 0.710 and F1-scores ranging from 0.652 to 0.752. KNN, with an Accuracy of 0.750, demonstrates greater stability, especially for the coal class (F1-score 0.790).

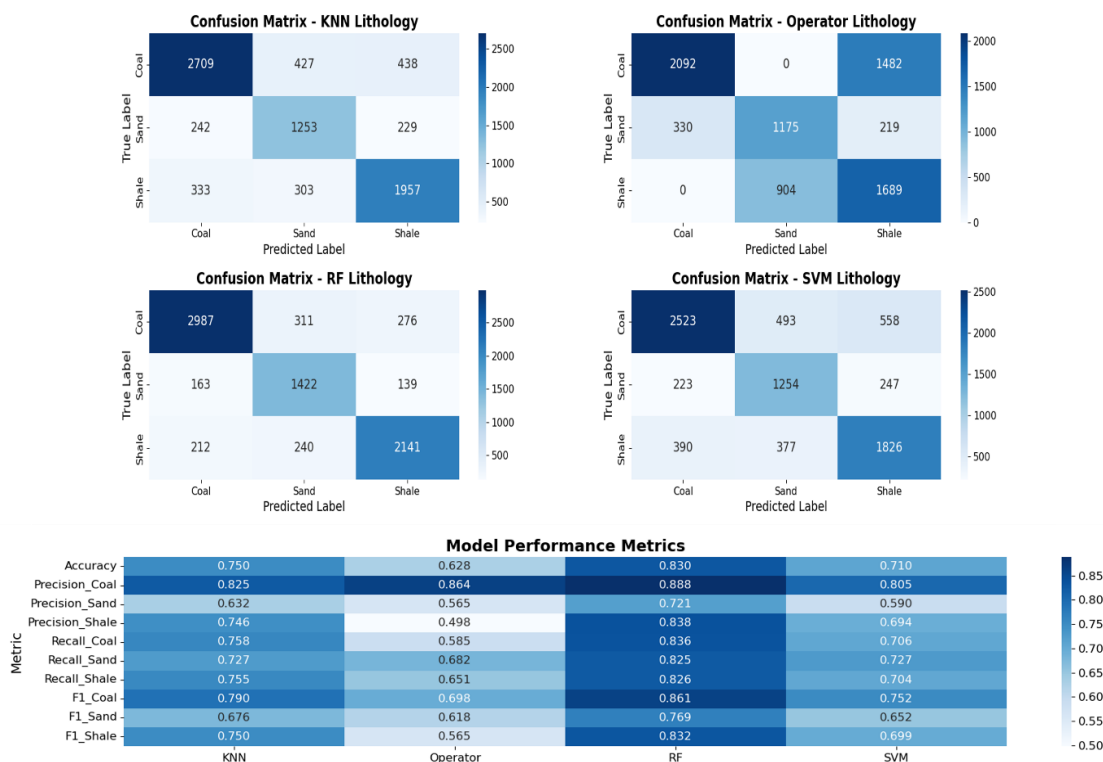


Fig. 5. Confusion matrices and model performance metrics for blind test of well X

Рис. 5. Матрицы ошибок и показатели эффективности модели для независимого тестирования (blind test) на скважине X

The confusion matrices likely reveal challenges in distinguishing between coal and shale, particularly for the Operator model. These findings highlight the potential of RF in coal geology research and suggest that improvements in data quality or algorithm design could enhance the performance of Operator and SVM models. Furthermore, blind test results for Well X reinforce the dominance of the RF model, especially in coal seam identification. RF ensemble learning approach enables it to handle complex, non-linear relationships in geological data, reducing overfitting and improving prediction accuracy. Compared to KNN distance-based method, SVM reliance on support vectors, and the manual adjustments in Operator Lithology, RF consistently delivers higher Precision and Recall – particularly for the coal class –demonstrating its robustness and generalization capability.

Conclusions and recommendations

This study underscores the advantages of ML-based lithology classification over traditional coal seam identification techniques. Conventional methods, such as manual well log interpretation or empirical threshold-based approaches, often suffer from inconsistency, reliance on subjective expert judgment, and difficulties in resolving complex lithological transitions. In contrast, the RF, KNN, and SVM models developed in this study provide a more accurate, automated, and scalable

framework for classifying sand, shale, and coal formations.

Through meticulous optimization via the Grid Search algorithm, the RF model emerged as the top-performing classifier, exhibiting markedly higher accuracy and reliability compared to other approaches. Unlike traditional methods, which falter under complex geological conditions, RF harnesses an ensemble of decision trees to discern intricate patterns within well log data, enhancing predictive accuracy and robustness. To evaluate its practical utility, a blind test was conducted using data from a new well (Well X), excluded from the training. The results affirm RF ability to maintain high classification accuracy, demonstrating robust generalization to unseen data.

Overall, this research highlights the maturity of ML techniques like RF, KNN, and SVM in lithology classification, moving beyond the need to revalidate their superiority over operator-based methods. Instead, the focus shifts toward addressing the model optimization problem, streamlining deployment, and developing ready-to-use software for industry applications. Future investigations will prioritize expanding the dataset, integrating additional well log parameters, and exploring automated deployment strategies to enhance classification accuracy and adaptability across diverse geological contexts, thereby providing practical tools for the coal mining industry.

REFERENCES/СПИСОК ЛИТЕРАТУРЫ

1. Gordon H.W. Jr., Kehn T.M., Carter M.D., Culbertson W.C. Coal resource classification system of the U.S. Geological Survey. *U.S. Geological Survey*, 1983, Circular 891, 65 p. DOI: <https://doi.org/10.3133/cir891>
2. Miller M.S., Moore M. Geophysical logging and exploration techniques in the Appalachian coal fields. *SPE Annual Technical Conference and Exhibition*. Dallas, Texas, September 1980. SPE-9466-MS.
3. Chatterjee R., Paul S. Application of cross-plotting techniques for delineation of coal and non-coal litho-units from well logs. *Geomaterials*, 2012, vol. 2 (4), pp. 94–104.
4. Hatherly P. Overview on the application of geophysics in coal mining. *International Journal of Coal Geology*, 2013, vol. 114, pp. 74–84.
5. Zhou B., Guo H. Applications of geophysical logs to coal mining – some illustrative examples. *Resources*, 2020, vol. 9 (2), Article 11.
6. Zhou B., O'Brien G. Improving coal quality estimation through multiple geophysical log analysis. *International Journal of Coal Geology*, 2016, vol. 167, pp. 75–92.
7. Zhang N., Liu C. Radiation characteristics of natural gamma-ray from coal and gangue for recognition in top coal caving. *Scientific Reports*, 2018, vol. 8, Article 190.
8. Jalil S., Rashid M. Analysis of natural radioactivity in coal and ashes from a coal fired power plant. *Chemical Engineering Transactions*, 2015, vol. 45, pp. 1549–1554.
9. 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, vol. 40 (5), pp. 599–605.
10. Sriniaiah 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, vol. 7 (1), Article 322.
11. 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, vol. 230, pp. 258–265.
12. Mclean C. *Pseudo proximate analysis: method using wireline logs to estimate components of coal bearing rock matrix without control data*. Master's thesis. Western Cape, 2015. 115 p.
13. Keskinsezer A. Determination of coal layers using geophysical well-logging methods for correlation of the Gelik-Zonguldak and Kazpınar-Amasra (Bartın) coalfields, Turkey. *Geomechanics and Geophysics for Geo-Energy and Geo-Resources*, 2019, vol. 5 (3), pp. 223–235.
14. 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 Journal*, 2020, vol. 25, pp. 1241–1258.
15. Yegireddi S., Bhaskar G. U. Identification of coal seam strata from geophysical logs of borehole using adaptive neuro-fuzzy inference system. *Journal of Applied Geophysics*, 2009, vol. 67 (1), pp. 9–13.
16. Maxwell K., Rajabi M., Esterle J. Spatial interpolation of coal properties using geographic quantile regression forest. *International Journal of Coal Geology*, 2021, vol. 248, Article 103861.
17. Guo J., Yang R., Han F. Coal seam roof and floor lithology prediction for underground coal gasification based on deep residual shrinkage network. *Energy Science & Engineering*, 2015, vol. 13 (4), pp. 1361–1374.
18. 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, vol. 268, Article 104208.
19. 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, vol. 217, Article 103314.
20. Kumar T., Seelam N. K., Rao G. S. Lithology prediction from well log data using machine learning techniques: a case study from Talcher coalfield, Eastern India. *Journal of Applied Geophysics*, 2022, vol. 199, Article 104605.
21. Maxwell K., Rajabi M., Esterle J. Automated classification of metamorphosed coal from geophysical log data using supervised machine learning techniques. *International Journal of Coal Geology*, 2019, vol. 214, Article 103284.
22. 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, vol. 14 (18), Article 8195.
23. 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, vol. 14 (22), Article 10381.
24. 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, vol. 59 (3), pp. 463–490.
25. Shimaoka A.M., Ferreira R.C., Goldman A. The evolution of crisp-dm for data science: methods, processes and frameworks. *SBC Computing Reviews*, 2024, vol. 4 (1), pp. 28–43.
26. Dastjerdy B., Saeidi A., Heidarzadeh S. Review of applicable outlier detection methods to treat geomechanical data. *Geotechnics*, 2023, vol. 3 (2), pp. 375–396.
27. Petrelli M. *Machine learning for earth sciences: using python to solve geological problems*. Springer Nature, 2023. 209 p. DOI:10.1007/978-3-031-35114-3
28. Lipovetsky S. Explanatory model analysis: explore, explain and examine predictive models. *Technometrics*, 2022, vol. 64 (3), pp. 423–424. DOI: <https://doi.org/10.1080/00401706.2022.2091871>
29. Wallace K., Gourgues C., Bernet C., Atkinson A.D., Yotter T. Why most models fail: the importance of blind validation. *Proceedings of the 7th Unconventional Resources Technology Conference*. Denver, Colorado, USA, 2019. pp. 489–502. DOI: <https://doi.org/10.15530/urtec-2019-396>

Information about the authors

Vu Hong Duong, Lecturer, Hanoi University of Mining and Geology, 18, Vien street, Duc Thang Ward, Bac Tu Liem District, Hanoi, 11900, Vietnam; vuhongduong@humg.edu.vn; <https://orcid.org/0009-0007-0923-0880>

Nguyen Tien Hung, Cand. Sci. (Eng.), Lecturer, Hanoi University of Mining and Geology, 18, Vien street, Duc Thang Ward, Bac Tu Liem District, Hanoi, 11900, Vietnam; nguyentienhung.dk@humg.edu.vn; <https://orcid.org/0009-0001-4849-7058>

Nguyen Minh Hoa, Cand. Sci. (Geol.-Mineral.), Principal Official, Office of the Government, 1, Hoang Hoa Tham street, Ngoc Ha Ward, Ba Dinh District, Hanoi, 100000, Vietnam; nguyenminhhoa@chinhphu.vn; <https://orcid.org/0000-0002-6555-662X>

Nguyen The Vinh, Cand. Sci. (Eng.), Associate Professor, Hanoi University of Mining and Geology, 18, Vien street, Duc Thang Ward, Bac Tu Liem District, Hanoi, 11900, Vietnam; nguyenthevinh@humg.edu.vn

Nguyen Xuan Duy, Trainee, Vietnam Petroleum Institute, 167, Trung Kinh street, Hanoi, 10000, Vietnam; duyxn@vpi.pvn.vn

Received: 27.06.2025

Revised: 15.09.2025

Accepted: 03.02.2026

Информация об авторах

Ву Хонг Зыонг, преподаватель, Ханойский университет горного дела и геологии, Вьетнам, 11900, г. Ханой, ул. Вьен, 18; vuhongduong@humg.edu.vn; <https://orcid.org/0009-0007-0923-0880>

Нгуен Тиен Хунг, кандидат технических наук, преподаватель, Ханойский университет горного дела и геологии, Вьетнам, 11900, г. Ханой, ул. Вьен, 18; nguyentienhung.dk@humg.edu.vn; <https://orcid.org/0009-0001-4849-7058>

Нгуен Минь Хоа, кандидат геолого-минералогических наук, главный специалист, Аппарат Правительства Вьетнама, Вьетнам, 100000, г. Ханой, ул. Хоанг Хоа Тхам, 1; nguyenminhhoa@chinhphu.vn; <https://orcid.org/0000-0002-6555-662X>

Нгуен Тхе Винь, кандидат технических наук, доцент, Ханойский университет горного дела и геологии, Вьетнам, 11900, г. Ханой, ул. Вьен, 18; nguyenthevinh@humg.edu.vn

Нгуен Суан Зуи, стажер, Вьетнамский нефтяной институт, Вьетнам, 10000, г. Ханой, ул. Чунг Кинь, 167; duyxn@vpi.pvn.vn

Поступила в редакцию: 27.06.2025

Поступила после рецензирования: 15.09.2025

Принята к публикации: 03.02.2026