Multidiszciplináris Tudományok, 16. kötet, 1. sz. (2026) pp. 26–35. ***https://doi.org/10.35925/j.multi.2026.1.3***

REAL TIME FORMATION PREDICTION USING MACHINE LEARNING MODELS BASED ON DRILLING PARAMETERS

József Pap
PhD student, Research Institute of Applied Earth Sciences, University of Miskolc
3515 Miskolc-Egyetemváros, e-mail: zed.tylor@gmail.com

Norbert Péter Szabó
head of department, Department of Geophysics, University of Miskolc
3515 Miskolc-Egyetemváros, e-mail: norbert.szabo@uni-miskolc.hu

Krisztián Mátyás Baracza
director of institute, Research Institute of Applied Earth Sciences, University of Miskolc
3515 Miskolc-Egyetemváros, e-mail: krisztian.baracza@uni-miskolc.hu

Zoltán Tibor Turzó
head of department, Institute of Mining and Energy, University of Miskolc
3515 Miskolc-Egyetemváros, e-mail: zoltan.turzo@uni-miskolc.hu

Abstract

Accurate knowledge of formation characteristics is critical during well development, particularly in the drilling phase, as it guides trajectory planning, casing depth selection, and the design of bit, fluid, and cementing programs. Additionally, a precise lithological sequence is essential for determining optimal perforation depths and managing production. Real-time data interpretation is increasingly valuable for enabling faster decisions and reducing operational costs. Traditionally, formation interpretation relies on manual analysis of drilling logs and shaker samples by petrophysicists or geologists. This paper introduces a machine learning-based approach to predict formation types using real-time drilling and MWD parameters, demonstrating the potential to enhance accuracy and decision-making efficiency compared to conventional methods.

Keywords: *drilling data, python, machine learning, classifier model, confusion matrix*

1. Introduction

Accurate identification of subsurface lithology during drilling is essential for effective well planning and execution. Understanding the quality and thickness of penetrated formations allows for improved trajectory control, optimized casing depth selection, and informed choices regarding drilling fluids, bit types, and cementing strategies. These decisions, when made with confidence in real-time, contribute significantly to reducing non-productive time and overall operational costs. Conventionally, formation evaluation has relied on the manual interpretation of drilling logs and analysis of cuttings at the surface. While these approaches can be accurate, it is time-consuming and subject to human limitations, particularly when fast decisions are required for dynamic drilling environments. The growing availability of real-time drilling, and measurement-while-drilling (MWD) data presents an opportunity to enhance formation interpretation through automation (Mahmoud et al., 2021; Cranganu et al., 2015).

By leveraging machine learning techniques, it is possible to analyze complex patterns within these datasets and predict formation types with improved speed and accuracy (Hassaan et al., 2024). This paper presents a machine learning-based methodology for formation classification using real-time drilling and MWD parameters. The goal is to demonstrate how data-driven models performs on field data. In this study, three ensemble machine learning algorithms – AdaBoost, Gradient Boosting, and Random Forest – are employed to classify formation types based on real-time drilling and MWD data. These models are trained and evaluated to compare their predictive performance in terms of classification accuracy. By analyzing the results through confusion matrices and other evaluation metrics, the study aims to identify the most effective algorithm for real-time lithological prediction, thereby offering a practical framework for enhancing formation evaluation workflows in the field.

2. Data processing

The base input consisted of time- and depth-indexed drilling and mud logging records acquired from a selected well (referred to as *Well-1#*), which was drilled to a total vertical depth (TVD) of 1320 meters and a measured depth (MD) of 2500 meters. TVD represents the vertical distance from the surface to a specific point in the borehole, while MD corresponds to the actual length along the borehole trajectory. The lithological composition of the sampled formation segments is reflected in the drilling mud analyzed at the surface with some time delay and potential dilution. For the purposes of this study, this delay is assumed negligible, and the real-time mud composition is considered representative of the corresponding formation interval. Input variables for the machine learning model were selected based on their relation to drilling mechanics and rock response (Komadja et al., 2025). Chemical analysis of the host rock was not included, while hydrocarbon components were considered as accompanying markers, though their contribution to lithology identification was found to be insignificant—a useful negative finding.

Table 1. *Recorded drilling parameters*

Total depth	*m*	recorded total MD of the well
TVD depth	*m*	calculated from deviation survey data
Bit depth	*m*	the actual MD position of the drilling bit
Rate of penetration (ROP)	*m/h*	ability of drilling through rocks
Weight on Bit (WOB)	*t*	weight exerted to the formation below bit
Weight on Hook (WOH)	*t*	weight of the drilling string
Rev. per Minute (RPM)	*rpm*	revolution of the drilling string
Torque	*kNm*	the force required for drill string rotation
Standpipe Pressure (SPP)	*bar*	pressure at the standpipe
Pump flow rate	*l/min*	mud pumping rate
Active mud volume	m^3	cumulative mud volume in the active tanks
Vol +/–	m^3	change of mud volume in the active tanks
Equivalent circulation density (ECD)	g/cm^3	density needed for circulation
Mud Weight IN	g/cm^3	inlet mud weight
Mud Weight OUT	g/cm^3	outlet mud weight
Temperature IN	$°C$	inlet mud temperature

Temperature OUT	*°C*	outlet mud temperature
Hook Height	*m*	recoded drawworks hook position
Total res. gas	*ppm*	total background gas concentration
C1	*ppm*	methane concentration
C2	*ppm*	ethane concentration
C3	*ppm*	propane concentration
iC4	*ppm*	iso-butane concentration
nC4	*ppm*	normal butane concentration
iC5	*ppm*	iso-pentane concentration
nC5	*ppm*	normal pentane concentration
Bit time	*h*	cumulative time with the same drilling bit
Pump time	*h*	cumulative pumping time on the section

From the top of the intermediate section at 455.5 m MD, a comprehensive set of parameters was available. *Table 1* above lists the parameters recorded by the mud logging service, including drilling mechanics, hydraulic properties, and gas composition indicators.

The mud logging dataset was converted into a numerical format. Columns exhibiting a high incidence of invalid or erroneous values were systematically removed. For the time-dependent dataset, a secondary filtering process was applied to retain only those records corresponding to active drilling conditions. Specifically, records were preserved only if the bit depth matched the recorded total depth and if ROP, WOB and RPM were greater than a minimum value. Summary statistics describing the processed datasets are shown in *Table 2*, which includes key metrics such as sample count, mean, standard deviation, and selected percentiles for both depth- and time-dependent data. The original time-indexed dataset consisted of 293,760 entries, from which approximately 74% were excluded during preprocessing, resulting in a high-quality subset suitable for machine learning applications.

Table 2. *Description of depth dependent [top] and time dependent [bottom] dataset of Well-1*

	Tot Depth	ROP m/hr	...	BitTim	Pump Time
count	3680.000000	3680.000000	...	3680.000000	3680.000000
mean	1497.117120	20.354418	...	32.524671	47.451484
std	578.823027	8.040184	...	17.334859	25.791242
min	456.500000	10.000000	...	1.550000	2.980000
25%	1010.375000	14.150000	...	16.662500	23.895000
50%	1499.750000	15.710000	...	34.080000	46.770000
75%	1997.625000	26.642500	...	46.882500	70.215000
max	2499.500000	49.770000	...	64.630000	94.920000

	Total Depth	Bit Depth	...	Temperature In	Pump Time
count	76342.000000	76342.000000	...	76342.00000	76342.000000
mean	1619.511396	1619.344502	...	38.65872	47.585711
std	567.435078	567.430140	...	10.30165	26.680168
min	455.000000	455.000000	...	14.00000	0.210000
25%	1126.000000	1126.000000	...	29.00000	24.050000
50%	1706.000000	1706.000000	...	43.00000	45.710000
75%	2091.000000	2091.000000	...	46.00000	71.690000
max	2500.000000	2500.000000	...	52.00000	94.990000

To enrich the dataset with formation-specific information, natural gamma ray (GR) values recorded by the MWD tool were merged into the depth-based dataset. A total of 13,098 GR records were available, sampled at an average vertical interval of 0.156 meters. However, due to rounding limitations in the recorded depth values, many intervals did not align precisely with GR sample points. Therefore, gamma values at missing depth points were estimated via distance-weighted interpolation, using the five nearest neighbors above and below each target depth:

$$Z_0 = \sum_{i=1}^{n} w_i Z_i, \text{ and } w_i = \frac{d_i^{-1}}{\sum_{i-1}^{n} d_i^{-1}} \quad (1)$$

where Z_0 is the interpolated value, Z_i are the known neighboring values, d_i is the distance from the target point, and w_i represents the normalized weight for each neighbor (Szabó, 2024). A custom loop was implemented to dynamically generate distance lists, compute weights, and insert interpolated GR values into the main data frame. To check the relationships between input variables, the Pearson Correlation Coefficient was computed for each pair of features:

$$r_n = \frac{\sum_{k=1}^{n}(x_k - \bar{x}) * (y_k - \bar{y})}{\sqrt{\sum_{k-1}^{n}(x_k - \bar{x})^2 * \sum_{k-1}^{n}(y_k - \bar{y})^2}} \quad (2)$$

The resulting values were compiled into a correlation matrix. Several variable pairs exhibited strong positive or negative correlations, notably among the hydrocarbon gas channels and between Bit Time and Pump Time, indicating redundant behavior. Stratigraphic and lithological data from traditional drilling logs were digitized and appended to the master dataset.

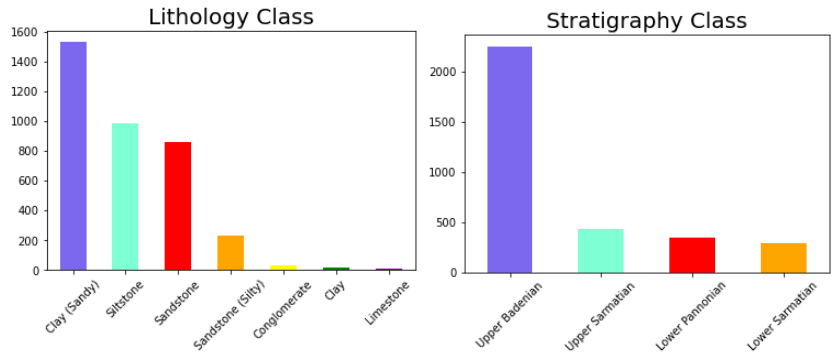

Figure 1. *Lithologic and stratigraphic distribution of samples*

Lithology class was selected as the target feature. To avoid model skewing due to imbalanced class representation, lithologies with insufficient sample counts were excluded. Prior to model training, exploratory data analysis was conducted to examine the distributions and relationships among remaining features. The objective was to prioritize features predominantly influenced by formation characteristics, and to discard those heavily affected by mechanical or operational factors. Among the input parameters, ROP, GR and Total Background Gas emerged as the most reliable lithology indicators. Therefore, initial models were trained using only these three features. Additional models incorporated *RPM*, *Torque*, and *WOB*, based on their moderate influence on lithology.

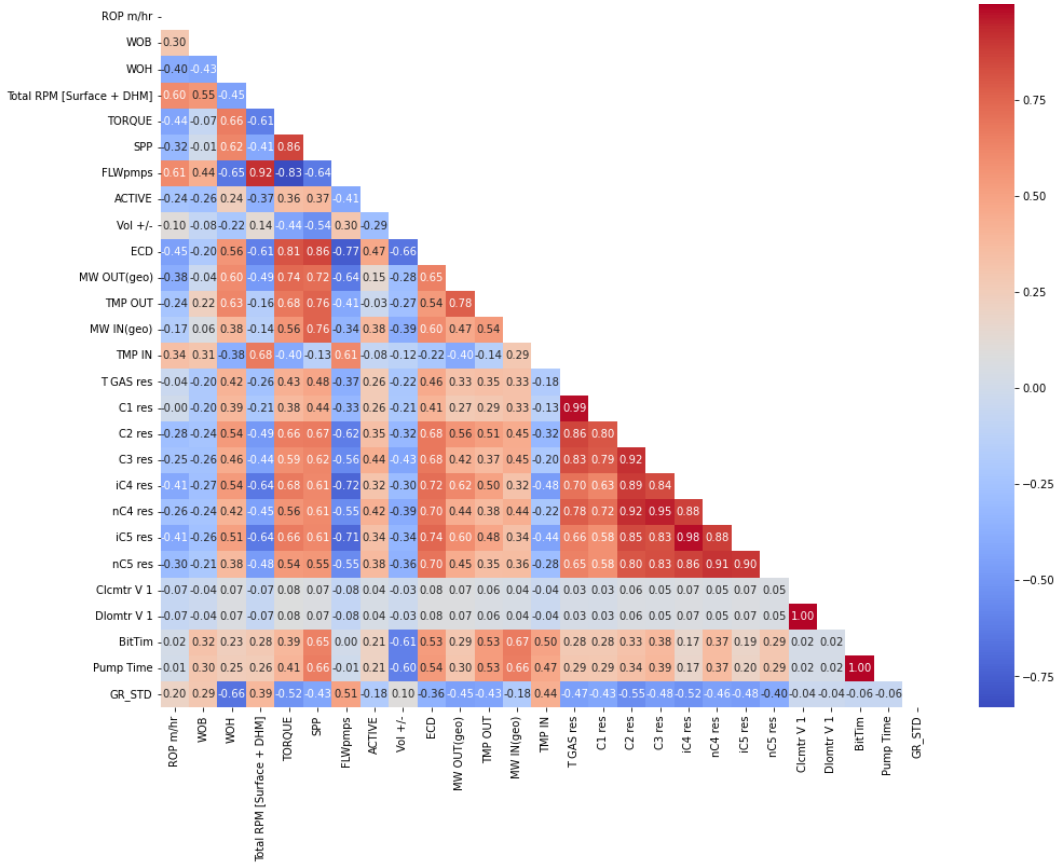

Figure 2. *Pearson correlation matrix of the input variables*

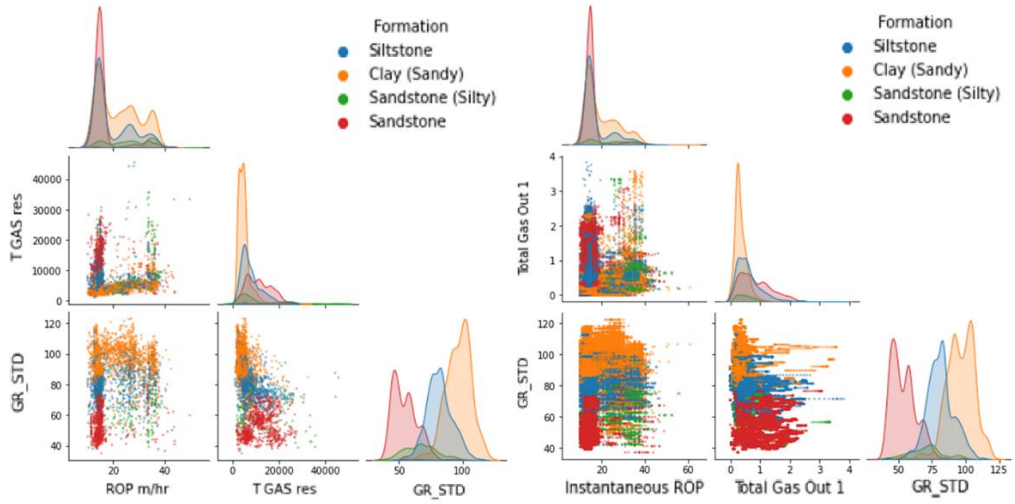

Figure 3. *Pair plots with KD overlays, for depth dependent [left] and time dependent [right] datasets*

Parameters such as Weight on Hook (WOH), Temperature IN/OUT, and cumulative Bit/Pump Time were excluded. WOH is strongly dependent on wellbore geometry and buoyancy, while temperature and cumulative operational times lack predictive value for real-time lithology classification.

3. Training the prediction models

Machine learning algorithms, including Random Forests, Decision Trees, and Neural Networks, have already been successfully applied to predict geophysical and lithological properties from MWD data, demonstrating high predictive accuracy and the importance of feature selection in drilling applications (Goldstein et al., 2025; Zhong et al., 2022). Building on these insights, this study focuses on ensemble-based classifiers—AdaBoost, Gradient Boosting, and Random Forest—to evaluate their performance in real-time lithology prediction. These models were integrated into a structured pipeline comprising preprocessing, resampling, training, and evaluation stages. The modeling workflow proceeded as follows:

1. **Feature and label definition:** The dataset was divided into feature variables (x) and target labels (y), where x represented the selected drilling and logging parameters, and y denoted the lithology class.
2. **Train-Test Split:** The dataset was split into training and testing subsets using a 70 : 30 ratio, ensuring representative class distribution in both sets.
3. **Feature Normalization:** *StandardScaler* was applied to normalize input features, transforming them to have zero mean and unit variance. This step improves model convergence and performance consistency.
4. **Class balancing via SMOTE:** To address class imbalance, the *Synthetic Minority Oversampling Technique* (SMOTE) was employed. Using 3 nearest neighbors, synthetic samples were generated for underrepresented classes within the training data.
5. **Model Initialization:** Three classifier models – AdaBoost, Gradient Boosting, and Random Forest - were instantiated with default hyperparameters for initial evaluation.
6. **Pipeline Construction:** A pipeline was created to sequentially execute data scaling, oversampling, and model fitting. This ensures reproducibility and integration of all preprocessing steps within the training loop.
7. **Evaluation Metrics:** Model performance was assessed using accuracy, precision, and recall metrics. These metrics were selected to provide a balanced evaluation of classifier effectiveness, particularly in multiclass scenarios.
8. **Cross-validation strategy:** A k-fold cross-validation scheme (k = 5) was applied to assess model generalizability and avoid overfitting. Cross-validation was conducted on the training data pipeline.
9. **Pipeline fitting:** Each pipeline was trained on the full training set, incorporating SMOTE and scaling prior to model learning.
10. **Prediction on test set:** After training, the SMOTE step was reapplied on the test set, followed by prediction using the trained pipeline.
11. **Performance visualization:** Results were visualized using confusion matrices and feature importance plots, enabling detailed inspection of classification behavior and variable influence.

4. Results: depth dependent models

Models trained on the depth-dependent dataset were evaluated on withheld test data. True lithology labels were masked during inference to simulate real-world deployment. AdaBoost classifier performed with 71.73% accuracy, while Gradient Boost model resulted in 81.84%. The most accurate model was Random Forest, with 93.20% overall precision, as seen on the table below. It is also observed that each model produced the lowest accuracy on Sandstone (Silty), followed by Siltstone samples. However Random Forest provided tolerable overall accuracy, the fact that it has 81.05% precision on predicting Sandstone (Silty) is not neglectable. Out of feature importance, gamma GR found to be proportionally the most influencing factor on the prediction.

Table. 3. *Model precision summary, and feature importance proportions of the depth-dependent prediction models*

MODEL PRECISION SUMMARY			
MODEL / LITHOLOGY	ADAPTIVE BOOSTING	GRADIENT BOOSTING	RANDOM FOREST
Clay (Sandy)	80,30%	88,93%	94,77%
Siltstone	59,98%	70,66%	90,59%
Sandstone	83,77%	93,23%	96,85%
Sandstone (Silty)	35,48%	57,10%	81,05%
SUM	**71,73%**	**81,84%**	**93,20%**

FEATURE IMPORTANCE PROPORTIONS			
MODEL / LITHOLOGY	ADAPTIVE BOOSTING	GRADIENT BOOSTING	RANDOM FOREST
GR_STD	38,00%	62,61%	47,19%
ROP m/hr	34,00%	22,18%	28,01%
T GAS res	28,00%	15,21%	24,80%
SUM	**100,00%**	**100,00%**	**100,00%**

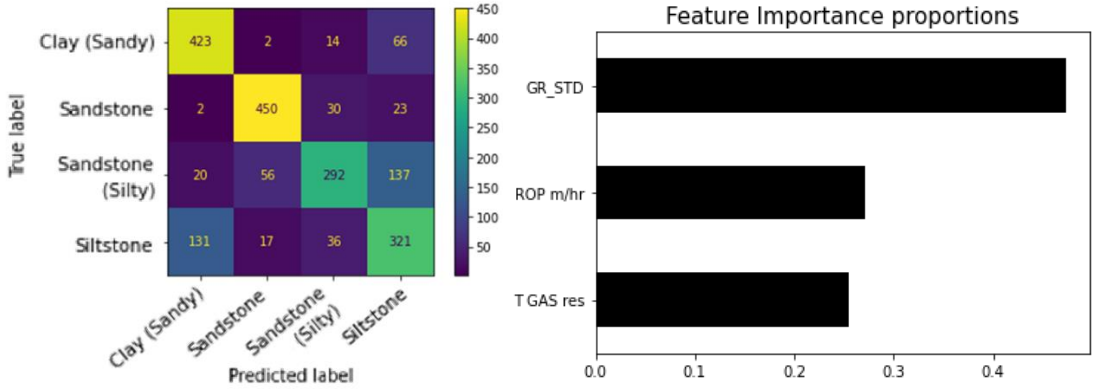

Figure 4. *Confusion matrix, and feature importance proportion plot, gathered from the depth dependent Random Forest prediction model*

5. Results: time dependent models

However, time dependent input had much more samples (76342 vs. 3680), significant improvement of precision was only observed at models with Random Forest classifier. While prediction accuracy slightly improved at Sandstone samples, from 96.85% to 99.52%, a great result was achieved on other samples as well.

Table. 4. *Model precision summary, and feature importance proportions of the time-dependent prediction models*

MODEL PRECISION SUMMARY			
MODEL / LITHOLOGY	ADAPTIVE BOOSTING	GRADIENT BOOSTING	RANDOM FOREST
Clay (Sandy)	80,69%	86,47%	98,79%
Siltstone	62,88%	78,96%	98,08%
Sandstone	89,96%	95,83%	99,52%
Sandstone (Silty)	14,19%	39,81%	96,08%
SUM	**71,72%**	**82,80%**	**98,68%**

FEATURE IMPORTANCE PROPORTIONS			
MODEL / LITHOLOGY	ADAPTIVE BOOSTING	GRADIENT BOOSTING	RANDOM FOREST
GR_STD	80,00%	78,00%	61,00%
ROP m/hr	6,00%	18,00%	16,00%
T GAS res	14,00%	5,00%	23,00%
SUM	**100,00%**	**101,00%**	**100,00%**

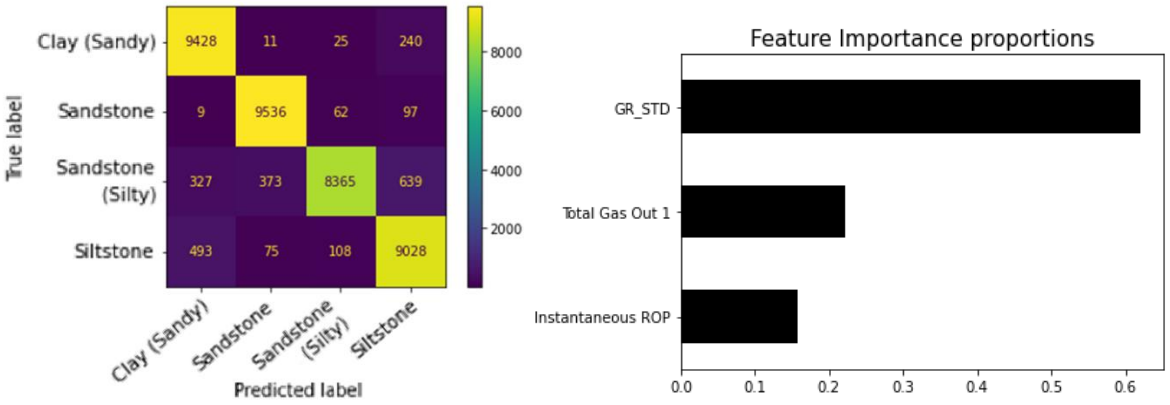

Figure 5. *Confusion matrix, and feature importance proportion plot, gathered from the time dependent Random Forest prediction model*

As shown on the figure, natural gamma ray remained the dominant predictive feature across all classifiers, with its importance increasing as model accuracy improved.

Given the consistently superior performance of the Random Forest algorithm, a final model was constructed by enriching the feature set. The time-dependent dataset was extended to include three additional parameters: Weight on Bit (WOB), Torque, and Revolutions Per Minute (RPM). These variables were chosen based on their moderate correlation with formation type and operational relevance (Ibrahim et al., 2023; Khalifa et al., 2023).

The updated Random Forest model achieved a remarkable 99.71% prediction accuracy, correctly classifying 74,387 out of 74,604 test samples. natural gamma ray remained the most impactful variable, although the contribution of mechanical drilling parameters was non-trivial. In the final model, Rate of Penetration (ROP) contributed the least (7%) to prediction, while WOB and Torque showed increased relevance.

6. Conclusion

This work investigated the efficiency of machine learning models for real-time lithology classification during drilling, using commonly recorded drilling and MWD parameters. Three models—Random Forest, Gradient Boosting, and AdaBoost—were trained and evaluated on both depth- and time-

dependent datasets. Random Forest consistently performed best, particularly on the larger time-dependent dataset, achieving an overall prediction accuracy of 99.71% when additional parameters such as WOB, Torque, and RPM were included. The results confirm that natural gamma ray remains the most influential feature for lithology prediction, while other drilling parameters contribute meaningfully when incorporated. Although certain lithology classes, such as silty sandstone, proved more challenging to classify, the overall performance demonstrates that the methodology has practical value for field applications.

This approach is not intended to replace traditional geological interpretation but to support it by providing rapid, consistent predictions that may reduce uncertainty during drilling. Its applicability is particularly relevant in operational environments requiring real-time decision-making, and further testing across different wells and formation types could extend and refine the methodology.

Furthermore, the observed correlations between mechanical drilling parameters—WOB, Torque, RPM, and ROP—suggest that a combined drillability metric, defined as:

$$D = \frac{Torque * RPM}{ROP} \tag{3}$$

which represents the drilling power per unit volume of the rock segment, may offer a physically meaningful proxy for formation-dependent drilling response. Although this study does not explicitly implement the D-parameter in the predictive models, it appears to be a promising concept for future research. Incorporating such a parameter could provide additional insight into rock-drilling interactions and potentially improve machine learning-based lithology classification in real-time applications.

7. Acknowledgements

The research was carried out in the framework of the GINOP-2.3.2-15-2016- 00010 “Development of enhanced engineering methods with the aim at utilization of subterranean energy resources” project of the Research Institute of Applied Earth Sciences of the University of Miskolc in the framework of the Széchenyi 2020 Plan, funded by the European Union, co-financed by the European Structural and Investment Funds.

References

[1] Mahmoud, A. A., Elkatatny, S., Al-Abdul Jabbar, A. (2021). Application of machine learning models for real time prediction of the formation lithology and tops from the drilling parameters. *Journal of Petroleum Science and Engineering*, 203, 108574. **https://doi.org/10.1016/j.petrol.2021.108574**

[2] Khalifa, H., Tomomewo, O. S., Berrehal, B. E. (2023). Machine Learning based real-time prediction of formation lithology and tops using drilling parameters with a web app integration. *Advances in Engineering*, 13, 2443–2467. **https://doi.org/10.3390/eng4030139**

[3] Ibrahim, A. F., Ahmed, A., Elkatatny, S. (2023). Applications of different classification machine learning techniques to predict formation tops and lithology while drilling. *ACS Omega*, 8, 42152–42163. **https://doi.org/10.1021/acsomega.3c03725**

[4] Hassaan, S., Mohamed, A., Ibrahim, A. F., Elkatatny, S. (2024). Real-time prediction of petrophysical properties using machine learning based on drilling parameters. *ACS Omega*, 9, 17066–17075. **https://doi.org/10.1021/acsomega.3c08795**

[5] Szabó, N. P. (2024). *Geostatistics*. Course Notes. University of Miskolc, Department of Geophysics.
[6] Cranganu, C., Breaban, M., Luchian, H. (2015). *Artificial Intelligent Approaches in Petroleum Geosciences*. Springer Cham. **https://doi.org/10.1007/978-3-319-16531-8**
[7] Komadja, G. C., Westman, E., Rana, A. et al. (2025). A machine learning approach to lithology classification in mining using measurement while drilling and exploration data. *Mining, Metallurgy & Exploration*, 42, 1955–1973. **https://doi.org/10.1007/s42461-025-01286-1**
[8] Goldstein, D., Aldrich, C., Shao, Q., O'Connor, L. (2025). Unlocking subsurface geology: A case study with measure-while-drilling data and machine learning. *Minerals*, 15, 241. **https://doi.org/10.3390/min15030241**
[9] Zhong, R., Salehi, C., Johnson, R. (2022). Machine learning for drilling applications: A review. *Journal of Natural Gas Science and Engineering*, 108, 104807. **https://doi.org/10.1016/j.jngse.2022.104807**