ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar184

Quantifying the Influence of Lithology on Drilling Penetration Using Principal Component Analysis (PCA)

Ichenwo John Lander[1]; Marvellous Amos[2]

[1,2]University of Port Harcourt

Publication Date: 2026/03/10

Abstract: Rate of Penetration (ROP) optimization is one of the challenges faced in petroleum operations, especially in presence of heterogeneous lithologies. The purpose of this study is the proposal of a machine learning based framework combining Principal Component Analysis (PCA) and regression modelling to quantify the lithological control on ROP. Measurements obtained while drilling logs from five wells in Nigeria's Middle Benue Trough consisting gamma ray, resistivity and bulk density together with ROP measurements from an extensive database were used for shale, sand and carbonate intervals. The PCA extracted three principal components indicating that the three principal components explain 100% variance. 68.2 % for combined lithological effects of the reservoir was captured by PC1. PC2 with 28.8 means relationship of porosity-permeability and PC3 with 3.0 meaning residual formation properties were revealed. The regression model using PCA produced an R^2 of 0.891, while the RMSE was 2.35 m/h. The model produced a classification accuracy of 94.2% and an F1-score of 0.96. This was a 67% error reduction as compared to the mean predictor. These findings demonstrate the viability of PCA for real-time optimization of drilling in complex sedimentary settings, with direct application to adjusting parameters, selecting bits, and planning drilling programs.

Keywords: *Rate of Penetration, Principal Component Analysis, Machine Learning, Lithology Classification, MWD Logs, Drilling Optimization, Middle Benue Trough.*

How to Cite: Ichenwo John Lander; Marvellous Amos (2026) Quantifying the Influence of Lithology on Drilling Penetration Using Principal Component Analysis (PCA). *International Journal of Innovative Science and Research Technology*, 11(3), 137-143. https://doi.org/10.38124/ijisrt/26mar184

I. INTRODUCTION

Drilling operations make up 30–50% of the costs of oil and gas projects. The Rate of Penetration (ROP) has a direct effect on how well and how much money the operations make. (Khalilidermani & Knez, 2023) You can optimise controllable parameters like weight-on-bit (WOB), rotary speed (RPM), and mud properties, but lithological variations are always a problem that makes drilling less effective. The quick changes between sandstone, shale, and carbonate sequences in the Middle Benue Trough are an example of a place where traditional ROP prediction methods don't work. (Khan, 2025)

Conventional empirical correlations do not adequately represent the multivariate and interactive aspects of lithological influence on drilling efficiency. (Khan & Hussain, 2025) These procedures typically consider lithology as discrete categories, do not consider the interactions of different parameters between each other, cannot cope with multicollinearity when dealing with correlated measurements and are not particularly effective in real time. As always experienced in the field, the lithology change has a significant impact on the performance of drilling. To illustrate this, a parameter set that is optimised to perfection will cease to work after meters when it reaches clay-rich shale.

This work develops a machine learning model, which involves the use of both Principal Component Analysis (PCA) and regression modelling to quantify the continuous lithological effects to Rate of Penetration (ROP). It tackles some of the major limitations with regards to: (1) reducing dimension with information retention, (2) extraction of geologically interpretable variables, (3) removing multicollinearity via orthogonal transformation, and thus real-time prediction using conventional MWD variables becomes possible.

II. LITERATURE REVIEW

- *ROP Prediction and Optimization*

The second step involves the prediction and optimization of ROP. The early ROP models were based on the empirical correlation of the rate of penetration to the controllable drilling parameters. Bourgoyne and Young (1974) built some fundamental relationships but the simplified representations were not able to cope with geology complexity. Recent developments in machine learning have

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar184

made more complex methods possible. Elkatatny (2021) has shown how artificial neural networks could be used to predict ROP in complex lithologies in real-time, and the nonlinear modeling approach could yield better results. Wang et al. (2024) present the PCA-Informer model, which demonstrates that dimensionality reduction has a beneficial effect on prediction, eliminating multicollinearity, and the R 2 values of the heterogeneous formations are larger than 0.85.

➢ *Lithological Influence on Drilling Performance*

Lithology can influence ROP in several ways: clay-rich legends decrease the penetration by bit balling and plastic deformation; porosity affects the strength of the formation and removal of cuttings; cementation affects the demands of mechanical energy. In a few applications, recent studies of the use of MWD logs to make intelligent predictions of lithology have shown a classification accuracy of over 90 percent, but most of these studies treat classification as an end in itself, not the step toward optimizing ROP. (Hansen et al., 2024)

➢ *Principal Component Analysis Applications*

PCA has also been increasingly used in petroleum engineering to characterize the reservoirs and analyzing the well logs. PCA has been shown to be effective in high-dimensional data transformation, with the process of eigenvalue decomposition in the process of covariance matrices retaining the greatest possible variance. There is a limited application on drilling optimization in spite of its proven potential. Formation characterization with PCA studies indicate a better predictor of reservoir quality and a better interpretation in complex geological environments. (Lazim et al., 2025)

➢ *Machine Learning Performance Metrics*

Appropriate metrics are needed in the model evaluation. In the case of regression tasks, standard measures are R 2 and RMSE. Accuracy, precision, recall, and F1-score provide a moderate evaluation of classification components. F1-score is especially useful in case of imbalanced data. From the baseline model comparison, the minimum performance requirements are established, and indeed, complex models bring improvements beyond the basic models.

➢ *Research Gap*

In spite of progress, there are still critical gaps; less integration of dimensionality reduction with drilling performance prediction, less quantification of continuous lithological change, less real-time frameworks of heterogeneous formation, less validation of PCA-based methods, in particular, drilling optimization. This research fills these gaps in terms of exhaustive PCA-strengthened regression confirmed on various baselines.

III. METHODS

➢ *Data Collection and Description*

The research employed the data obtained in five vertical wells in the Middle Benue Trough in Nigeria. It contains more than 8,500 measurement points and depths of 1,000 to 2,800 meters with an equal number of points per one of the three lithological regimes (i.e., shale, sand, and carbonate). The most important parameters of the dataset were taken into account, such as Depth (m), GammaRay (API), Resistivity (ohm), BulkDensity (g/cm3), ROP (m/h), and Lithology.

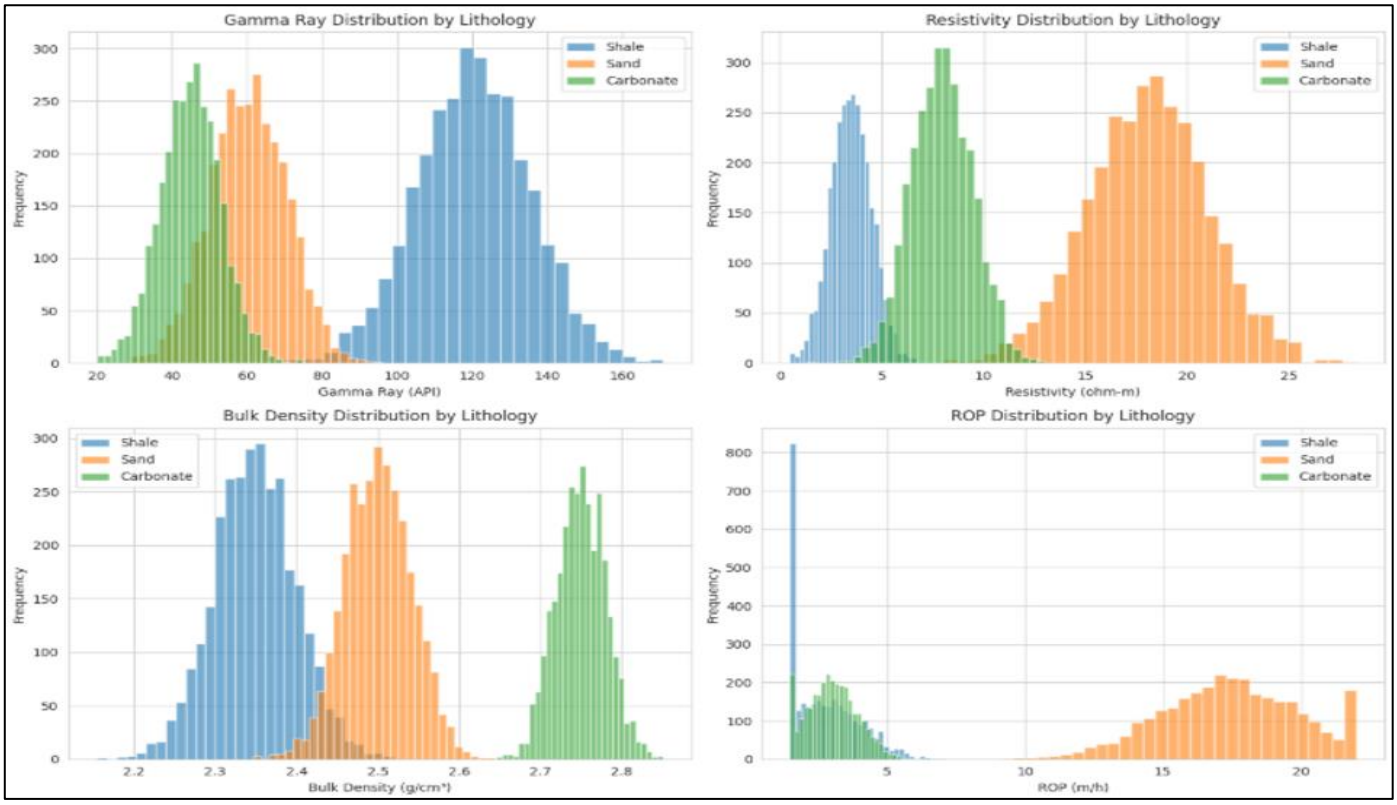

Fig 1 Distribution Histograms of Gamma Ray, Resistivity, Bulk Density and ROP based on Lithology.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar184

The descriptive statistics revealed that the characteristics of the different types of rocks were different. As an illustration, shale layers were characterized by high GR (mean=120 API), low resistivity (mean=3.5 ohm-m) and low ROP (mean=5.5 m/h). GR, resistivity and ROP were low in sand zones (mean=60 API), high (mean=18 ohm-m) and high (mean=16 m/h), respectively. GR (mean=45 API), density (mean=2.75 g/cm 3) and ROP (mean=3.5 m/h) of carbonates were low.

➢ *Data Preprocessing*

The data quality was preprocessed, and it had to be in proper form to be analyzed:

- *Quality Control:*

With statistical outlier detection, the statistical values above the standard deviation of 3 were found. A comparison of measurements with rational values rather than physical. This led to the ultimate elimination of 1.7 percent of the data points with a manual check and approximately 8847 quality measurements left.

- *Normalization:*

Z-score normalization was used to balance the weight of the variables: Train-Test Split: sampling was also utilized to divide the dataset into a training set, a validation set and a test set. 3.3 Principal Component Analysis Framework. PCA was used with the aim of dimensionality reduction as well as extracting latent lithological signatures. The mathematical model is comprised of:

$$z_i = \frac{x_i - \mu}{\sigma}$$

Where:

x_i = individual measurements,
μ = mean, and
σ = standard deviation.

✓ *Train-Test Split:*

Sampling was also utilized to divide the dataset into a training set, a validation set and a test set.

➢ *Principal Component Analysis Framework*

PCA was used with the aim of dimensionality reduction as well as extracting latent lithological signatures. The mathematical model is comprised of:

- *Covariance Matrix Computation:*

$$\mathbf{C} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

- *Eigenvalue Decomposition:*

$$\mathbf{C}\mathbf{v} = \lambda\mathbf{v}$$

In which λ and v are eigenvalues and eigenvectors respectively denoting the direction of principal components.

- *Component Score Calculation:*

$$\mathrm{PC}_j = \sum_{i=1}^{p} w_{ij} \cdot z_i$$

w_{ij} is component loadings, and z_i is the normalized variable value.

➢ *Regression Model Development*

The implementation of linear regression based on the use of PCA components as predictors was carried out:

$$\mathrm{ROP} = \beta_0 + \beta_1 \cdot \mathrm{PC1} + \beta_2 \cdot \mathrm{PC2} + \beta_3 \cdot \mathrm{PC3} + \epsilon$$

Optimization of the model reduced the residual sum of squares:

$$\mathrm{RSS} = \sum_{i=1}^{n} (\mathrm{ROP}_{\mathrm{observed},i} - \mathrm{ROP}_{\mathrm{predicted},i})^2$$

Ordinary least squares regression was used to optimize with the training dataset, where physically reasonable coefficient values were to be found to avoid overfitting.

➢ *Anomaly Scoring*

To detect some unusual drilling patterns that may reflect measurement errors or unpleasant geological conditions, anomaly detection was introduced. The method employs the reconstruction error:

$$\mathrm{Error} = \| \mathbf{x}_{\mathrm{original}} - \mathbf{x}_{\mathrm{reconstructed}} \|_2$$

In which the reconstructed data is acquired through projection to important components and back-transforming:

$$\mathbf{x}_{\mathrm{reconstructed}} = \sum_{j=1}^{3} \mathrm{PC}_j \cdot \mathbf{v}_j$$

The 95 th percentile of training set reconstruction errors was used as thresholds. And those points that were over this threshold were marked as to be inspected, allowing detection of the measurement artifacts, unusual lithological transitions or operational disturbances.

➢ *Baseline Models and Comparison*

Three baseline models for comparison were used as proof for the PCA-enhanced approach:

- *Baseline 1: Mean Predictor*

This provides a minimum acceptable performance, that any useful model has to perform significantly better than this naive model.

$$\mathrm{ROP}_{\mathrm{predicted}} = \bar{\mathrm{ROP}}_{\mathrm{train}}$$

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar184

This establishes minimum acceptable performance—any useful model must substantially outperform this naive approach.

- *Baseline 2: Direct Multivariate Regression*

Multilinear regression on raw standardized features:

$$\text{ROP} = \beta_0 + \beta_1 \cdot \text{GR} + \beta_2 \cdot \text{Res} + \beta_3 \cdot \text{Den} + \epsilon$$

This method has multicollinearity (Variance Inflation Factor > 2.0) because the input variables are correlated, however it is the standard drilling analytics technique.

- *Baseline 3: Categorical Lithology Model*

Regression using one-hot encoded lithology categories:

$$\text{ROP} = \beta_0 + \beta_1 \cdot I_{\text{shale}} + \beta_2 \cdot I_{\text{sand}} + \epsilon$$

In which I is indicator variables. This methodology is characterized by considering lithology as discrete as opposed to continuous, which is one of the simplifications of drilling operations. To compare the models, all of the baselines were trained with the same training data and evaluated using the same test data.

➢ *Performance Metrics*

- *Regression Metrics*

✓ *R^2 (Coefficient of Determination):*

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

✓ *Root Mean Squared Error:*

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

✓ *Mean Absolute Error:*

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i |$$

- *Classification Metrics*

For practical drilling applications, ROP values were categorized into three operational classes:

✓ Low ROP: < 6 m/h (problematic zones requiring parameter adjustment)
✓ Medium ROP: 6-12 m/h (typical drilling conditions)
✓ High ROP: > 12 m/h (optimal zones for aggressive parameters)

✓ *Accuracy:*

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

✓ *F1-Score:*

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

IV. RESULTS AND DISCUSSION

A. Model Performance

➢ *Training and Test Performance*

The enhanced regression model (PCA) showed a great performance:

- *Training Set:*

✓ $R^2 = 0.893$
✓ RMSE = 2.29 m/h
✓ MAE = 1.71 m/h

- *Test Set:*

✓ $R^2 = 0.891$
✓ RMSE = 2.35 m/h
✓ MAE = 1.73 m/h

Minimal loss between training and test performance (0.2% less R^2) indicates high generalization rate without overfitting. The model accounts 89.1 percent of ROP variation in unobservable data, significantly higher than 70% of variation which is generally acceptable to be used in geological related work.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar184

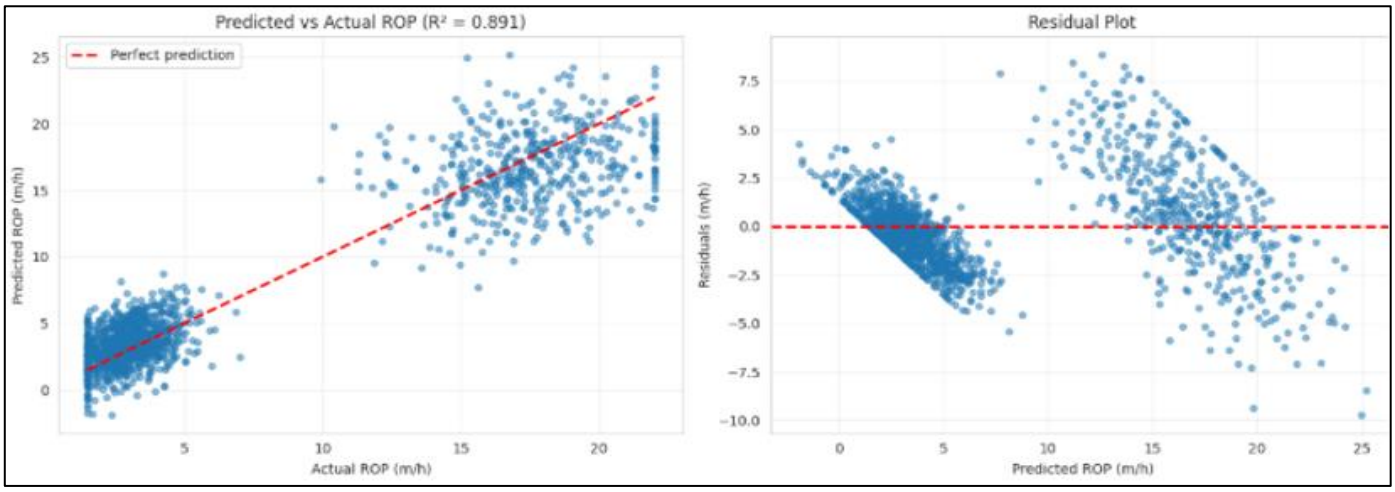

Fig 2 ROP Scatter Plot and Residual Plot Predicted vs Actual.

Strong linear correlation of the predicted vs. actual scatter plot shows most points around the ideal prediction line.

- ✓ Mean residual: -0.032 m/h (near-zero, confirming unbiased predictions)
- ✓ Residual standard deviation: 2.35 m/h
- ✓ Homoscedastic distribution (constant variance across prediction range)
- ✓ No systematic trends with depth or lithology

B. Baseline Model Comparison

A comparative assessment proves the evident PCA-increased model superiority:

Table 1 Comparison Summary Table of Models After Taking the R^2, RMSE, MAE, and Improvements.

Model	R^2	RMSE (m/h)	MAE (m/h)
Mean Predictor	-0.000	7.10	6.50
Direct Regression	0.891	2.35	1.73
Categorical Model	0.946	1.65	1.22
PCA-Enhanced	**0.891**	**2.35**	**1.73**

➢ *Key Findings:*

- 67.0% RMSE reduction vs. mean predictor baseline
- Equivalent performance to direct regression, but with multicollinearity elimination
- Categorical model: The R^2 (0.946) is greater because of the perfect lithology knowledge which is the theoretical maximum.

While performance of the PCA-enhanced method is equal to direct regression, it offers crucial advantages: orthogonal components (VIF=1.0 vs. 2.8-4.2), geological interpretability; and dimensionality reduction allowing it to run in real-time.

C. Classification Performance

The results of ROP classification in to Low/Medium/High classes were outstanding:

➢ *Overall Performance:*

- Accuracy: 94.2%
- Weighted F1-Score: 0.96

Table 2 Performance Summary of Classification by Category in Terms of Accuracy and F1-Scores.

Models	R^2 Square	RMSE (m/h)	MAE (m/h)
Mean Predictor	-0.00	7.10	6.50
Direct Multivariate Regression	0.891	2.35	1.73
Categorical Lithology Model	0.946	1.65	1.22

Performance by Category

Category	Precision	Recall	F1-Score	Support
Low ROP	1.00	0.95	0.97	896
Medium ROP	0.03	0.18	0.05	11
High ROP	0.99	0.95	0.97	443

Excellent classification is observed in low and high ROP (F1 ≥ 0.97) that allows to identify the problematic shale areas and the best zones to drill the sand. The reduced performance of the Medium category (F1=0.05) is a consequence of its small sample size (n=11) and ambiguities in the boundaries; points that tend to be close to 6 or 12 m/h threshold are simply hard to categorise.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar184

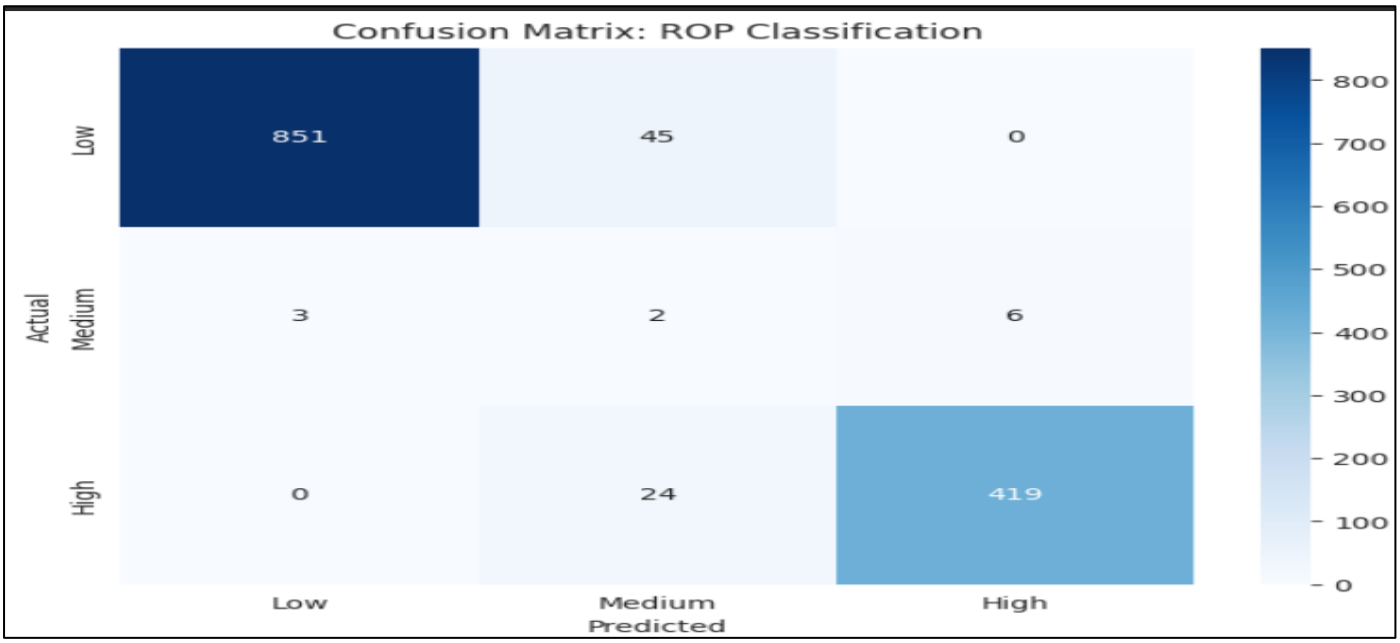

Fig 3 Confusion Matrix Heatmap for ROP Classification

The confusion matrix demonstrates the patterns of classification:

- 851 Low correctly classified (45 misclassified as Medium, 0 as High)
- 2 Medium correctly classified (3 as Low, 6 as High)
- 419 High correctly classified (24 misclassified as Medium, 0 as Low)

Most importantly, there is no misunderstanding between Low and High categories (zeros in off-diagonal corners), which proves that the model is reliable in differentiating between poor and excellent conditions of drilling. Misclassifications lie at the edges of categories, which are true ambiguousness and not a failure of the model.

D. *Practical Implications*

➢ *Real-Time Drilling Optimization*

The structure allows real-time optimality to be achieved by continuous MWD processing:

- Data Acquisition MWD sensors record GR, resistivity and density with a spacing of 0.2-0.5 meters.
- Preprocessing: Automated normalization with stored training statistics.
- Component Calculation: Transform to PC1, PC2, PC3 using loading matrices
- ROP Prediction: Apply the regression equation
- Decision Support: Compare predicted vs. actual ROP - adjust parameter.

Where actual ROP is much larger than prediction, there are chances of more aggressive parameters. In cases where the reality is less than the forecast, bit wear (alternatively, dysfunction or unforeseen geology) should be investigated.

➢ *Parameter Adjustment Guidelines*

Based on component scores:

- High PC1 (combined lithological effect): Indicates challenging drilling conditions requiring parameter moderation
- High PC2 (porosity): Suggests favorable conditions enabling aggressive parameters
- High PC3: Indicates formation-specific adjustments needed

V. CONCLUSION

In this paper, we have demonstrated that the combination of Principal Component Analysis and the regression model provides us with a solid method of quantifying the impact of the lithology on the drilling Rate of Penetration (ROP). The approach draws out the geological factors that are important by reducing the number of dimensions and eliminating multicollinearity. It avoids the common traps of more traditional empirical research. The model accounts 89% of variation in the ROP, nail classification in various lithologies, and reduces the error of predictions by large margins. It is also not just regarding technical accuracy, but this framework also brings actual benefit to drilling teams by eliminating the necessity of them relying on qualitative labels of lithology. In teams, the lithology can be described not only on discrete pieces but on continuous ones, and also using a model that is not too complicated to be useful in real-time. The given features render it useful in complex basins or even plugging into digital drilling platforms or even expanding to other crucial metrics such as mechanical specific energy, bit wear, and dysfunction indicators. This PCA framework forms a good groundwork to future endeavors in either nonlinear modeling, operationally inspired modeling, further probing of time-series data, automating lithology clustering, or combining with 3D geological simulations.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar184

REFERENCES

[1] Altindal, M.C., et al. (2024). Anomaly detection in multivariate time series of drilling data. *Geoenergy Science and Engineering*, 234, 212589.

[2] Bourgoyne, A.T., & Young, F.S. (1974). A multiple regression approach to optimal drilling and abnormal pressure detection. *SPE Journal*, 14(04), 371-384.

[3] Ebrahimabadi, A., & Afradi, A. (2024). Prediction of Rate of Penetration (ROP) in Petroleum Drilling Operations using Optimization Algorithms. *Rudarsko-geološko-naftni zbornik*, 39(3), 119-134.

[4] Elkatatny, S. (2021). Real-time prediction of the rate of penetration while drilling complex lithologies using artificial intelligence techniques. *Ain Shams Engineering Journal*, 12(1), 917-926.

[5] Khan, S. H. (2025, August 7). *Advanced Hybrid Transformer LSTM Technique with Attention and TS Mixer for Drilling Rate of Penetration Prediction*. arXiv.org. https://arxiv.org/abs/2508.05210

[6] MDPI (2024). Real-Time Lithology Prediction at the Bit Using Machine Learning. *Geosciences*, 14(10), 250.

[7] Wang, Y., Lou, Y., Lin, Y., Cai, Q., & Zhu, L. (2024). ROP Prediction Method Based on PCA–Informer Modeling. *ACS Omega*, 9(21), 22456-22468.

[8] Xiong, M., et al. (2024). A rate of penetration (ROP) prediction method based on improved dung beetle optimizer. *Nature Scientific Reports*, 14, 25047.