Proceedings of ISERME 2025
19-20 September 2025
https://doi.org/10.31705/ISERME.2025.15

Performance Evaluation of Machine Learning Pipelines for Pore Pressure Prediction

Naweed MNM, Suheerman S, Dilkushan SMDKR, Thiruchittampalam S, and *Wickrama MADMG

Department of Earth Resources Engineering, University of Moratuwa, Sri Lanka

Corresponding author – maheshwari@uom.lk

Abstract
Accurate pore pressure prediction is critical for safe drilling operations. Conventional prediction methods, which rely on simplified empirical assumptions, often fail to capture the multivariate and non-linear relationships present in complex geological settings. Machine learning (ML) provides a data-driven approach that can model these complexities directly from well log data without relying on predefined physical equations. However, the practical application of ML is often inconsistent due to a lack of systematic understanding of how data preprocessing choices impact final model performance. This study aims to resolve this uncertainty by identifying the optimal combination of preprocessing strategy and ML algorithm for this task. A comparative analysis was conducted across four scenarios: raw data, outlier-capped data, feature-selected data, and combined preprocessing (outlier capping and feature selection) using six ML algorithms to systematically evaluate the effects of outlier capping and the removal of multicollinear features. The findings identify a tuned XGBoost model as the top performer ($R^2 = 0.9789$), achieving this optimal result on the raw, unprocessed dataset. This key finding, when analyzed in the context of the other experimental scenarios, demonstrates that removing linearly correlated features can be detrimental to advanced models and that the necessity of outlier treatment is algorithm dependent. This study concludes that while the data preparation strategy is universal, it is closely tied to algorithm choice, offering a context-aware framework to enhance model reliability and support interpretability in future research.

Keywords: Ensemble methods, Feature selection, Geomechanics, Hyperparameter tuning, Outlier capping, XGBoost

1 Introduction

Pore pressure is the pressure of fluids contained within the pore spaces of subsurface formations. Accurate estimation of pore pressure is fundamental to defining the safe mud weight window in drilling operations [1]. This operational margin is critical for preventing hazardous and costly events such as wellbore kicks, lost circulation, and blowouts, which pose significant risks to personnel, the environment, and project economics [1]. For decades, the industry has relied on conventional methods, such as Eaton [2], which correlate individual petrophysical properties with expected compaction trends. However, the reliability of these models diminishes in complex geologies because their underlying empirical assumptions often fail to capture the multivariate and non-linear relationships between formation properties.

The theoretical foundation of pore pressure prediction is based on Terzaghi's principle of effective stress, which relates total overburden stress to the stress borne by the rock matrix and the fluid pressure within its pores [3]. Conventional methods attempt to apply this principle by using empirical correlations, such as assuming a normal compaction trend (NCT) from well log data. The primary limitation of this approach is that these simplified, single-variable correlations often fail to capture the true multivariate and non-linear nature of subsurface systems, particularly in complex geologies [4].

ML has emerged as a data-driven approach capable of addressing this specific challenge. By processing a full suite of well logs simultaneously, ML algorithms can learn these complex relationships and cross-feature interactions directly from the data without reliance on predefined physical equations. Numerous studies have demonstrated the successful application of ML for pore pressure prediction [5], [6]. Studies have demonstrated the capability of artificial neural networks (ANNs) to capture complex non-linear relationships [7]. Concurrently, ensemble methods like random forest and XGBoost are widely utilized for their robustness and high performance on tabular petrophysical data [5].

Despite these successful applications, a critical methodological gap persists concerning the application of two common preprocessing steps: outlier treatment and feature selection. The existing literature often applies these techniques as a fixed, preliminary routine without systematically investigating their necessity or differential impact. For instance, while many studies acknowledge the need for robust data [5], it remains unclear whether a robust algorithm like XGBoost derives the same benefit from outlier capping as a sensitive algorithm like support vector regression (SVR). Similarly, it is not well-documented whether removing features based on linear correlation aids or harms complex models that might be capable of extracting unique non-linear information from them.

However, existing literature commonly adopts a single, fixed data preprocessing pipeline, such as outlier removal followed by feature selection, without systematically investigating its necessity or differential impact across various algorithm types [5]. This creates uncertainty in the methodology, as the optimal preprocessing strategy for an ensemble model may differ significantly from that for a sensitive kernel-based model.

Therefore, the problem this study addresses is the need to move beyond basic intermodal comparisons and determine the optimal combination of data preparation techniques and ML algorithms for pore pressure prediction. This research aims to resolve the ambiguity through the following objectives: (1) to systematically evaluate six distinct ML algorithms across four well-defined preprocessing scenarios; (2) to quantify the impact of outlier capping and feature selection on the performance of each model; and (3) to identify the optimal end-to-end pipeline for this prediction task.

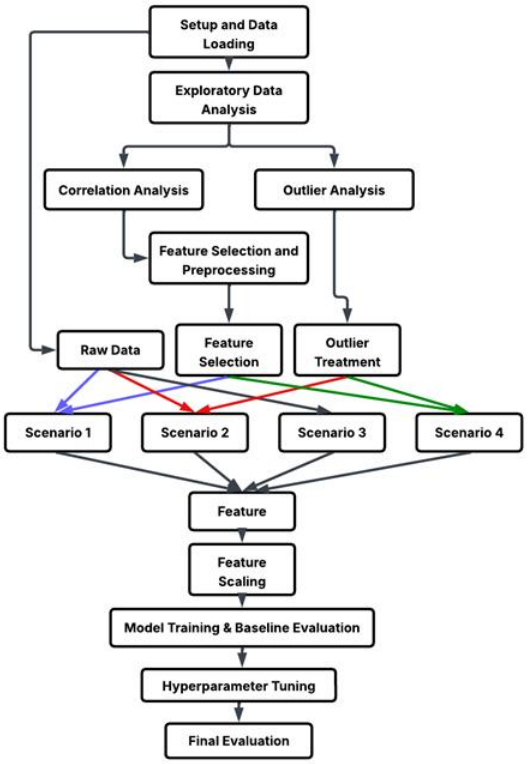

Figure 1: Methodological Workflow for the Four-Scenario Comparative Analysis

2 Methodology

The methodology employed in this study is illustrated in Figure 1. The workflow consists of two primary stages: first, an initial data processing phase to create four distinct experimental scenarios, and second, a standardized modeling pipeline through which each scenario is evaluated for a comprehensive comparative analysis.

2.1 Dataset Description

Table 1a: Descriptive statistics (Part 1)

Statistic	Depth (m)	Gr (API)	RHOB (g/cm³)	V_p (km/s)
Mean	139.71	92.09	1.81	-25.7
Std dev	74.66	8.95	0.14	162.89
Min	5.95	42.27	1.08	-999.25
Median	132.92	92.63	1.82	1.54
Max	335.88	114.99	2.12	1.72

Table 1b: Descriptive statistics (Part 2)

Statistic	V_{Sh}	Caliper (in)	Resistivity (ohm.m)
Mean	0.66	10.09	0.99
Std dev	0.45	0.63	0.27
Min	-0.16	9.42	0.36
Median	0.67	9.97	0.95
Max	46.21	16.38	2.87

Table 1c: Descriptive statistics (Part 3)

Statistic	Porosity (%)	Stress (MPa)	Pp (MPa)
Mean	59.44	2.54×10^6	1840.08
Std dev	6.69	1.45×10^6	219.43
Min	41.17	6.65×10^6	1416
Median	58.79	2.35×10^6	1823
Max	98.85	6.93×10^6	2314

The research utilizes an open-source dataset comprising petrophysical well logs from eight wells, sourced from a public GitHub repository [8]. While the dataset is reported to be from a U.S. oil field, specific geographical provenance is unavailable. The aggregated dataset contains 11,494 data points. The data includes nine predictor variables (Depth, Gamma Ray (GR), Bulk Density (RHOB), P-wave Velocity (V_p), Volume of Shale (V_{Sh}), Caliper, Porosity, Resistivity, and Stress) and one target variable, Pore Pressure (PP).

A statistical summary of the raw data is presented in Table 1a, Table 1b and Table 1c. The summary reveals two critical characteristics that informed the experimental design. First, the variables exist on vastly different scales (e.g., Stress has a mean of 2.54×10^6 MPa while Bulk Density has a mean of 1.81 MPa), underscoring the need for feature scaling. Second, the presence of physically unrealistic values, such as the minimum P-wave Velocity (V_p) of -999.25 km/s, provides clear statistical evidence of extreme outliers, justifying the study's investigation into outlier treatment methods.

2.2 Setup and Data Loading

The source data, consisting of eight separate Microsoft excel files corresponding to eight distinct wells, was loaded and aggregated, resulting in a unified dataset of 11,494 data points. A critical step in this initial data handling process was the creation of a “WELL identifier” column within each file's data before concatenation. This ensures that the provenance of every data row is preserved in the final master dataset, which is essential for traceability and potential future well-specific analyses. Following this, all eight datasets were combined into a single, unified data frame to serve as the basis for the subsequent experimental scenarios.

2.3 Exploratory Data Analysis (EDA)

The analysis began with the generation of a Pearson correlation matrix to quantify the linear relationships between all variables and to detect initial signs of multicollinearity [9]. This initial analysis revealed an unexpected statistical anomaly: the correlation between the V_{Sh} and GR logs was found to be weak. This result contradicts established petrophysical principles, where GR is a primary indicator for V_{Sh} and a strong positive correlation is expected [5].

It was hypothesized that this statistical distortion was caused by the presence of extreme outliers within these specific features. To test this hypothesis, a targeted outlier analysis was conducted, focusing on the V_{Sh} and GR variables. The interquartile range (IQR) method was chosen for this task due to its robustness to the non-normal, skewed distributions characteristic of petrophysical data.

The identified outliers in V_{Sh} and GR were then treated using capping. This method was chosen over simple deletion to mitigate the statistical influence of extreme values without discarding the entire data record, which may contain valid geological information. This targeted intervention resulted in the creation of a second, "capped" dataset for comparison against the original "raw" dataset. Both datasets were then used in the subsequent feature selection analysis and the main experimental design to fully evaluate the impact of this targeted data treatment.

2.4 Feature Selection

To address the strong multicollinearity detected between several features, a comparative analysis was conducted using univariate F-tests, recursive feature elimination (RFE), and embedded random forest feature importance. This analysis identified a subset of potentially redundant features, leading to the creation of a "full feature set" and a "reduced feature set" to be tested.

2.5 Data Preprocessing

2.5.1 Data Splitting

The dataset was partitioned into a training set (70%) and a testing set (30%). This ratio was chosen to provide a large, statistically stable test set for reliable performance evaluation, while leaving sufficient data for model training. The training set was consistently used for generating validation sets during baseline evaluation and hyperparameter tuning.

2.5.2 Feature Scaling

Following data splitting, feature standardization was applied to ensure all predictor variables shared a common scale with a mean of 0 and a standard deviation of 1. This preprocessing step was necessary for algorithms sensitive to feature magnitudes, such as K-nearest neighbors, support vector regression, and artificial neural networks [10]. To prevent data leakage, the scaler was fit solely on the training data and subsequently applied to both training and testing sets.

2.6 Model Training & Baseline Evaluation

Six machine learning algorithms were selected to cover a representative spectrum of learning paradigms: decision tree, K-nearest neighbors (KNN), SVR, random forest, XGBoost, and ANN to represent a diverse range of methodological paradigms, including foundational, kernel-based, ensemble, and deep learning approaches. This diversity ensures a comprehensive evaluation of their suitability for the petrophysical prediction task. An initial baseline evaluation was performed by training each model with its default hyperparameters on the prepared data. This step established an unbiased initial performance benchmark for each algorithm. To ensure a stable and reliable estimate of baseline performance, a 5-fold cross validation procedure was applied to the training data for each algorithm within each of the four experimental scenarios.

2.7 Hyperparameter Tuning

To unlock the full potential of each algorithm, a systematic hyperparameter tuning process was conducted. Randomized search was selected as it allows for efficient exploration of the hyperparameter space by focusing computational resources on a diverse set of parameter combinations. This approach increases the likelihood of identifying near-optimal configurations while reducing unnecessary computational cost [11]. A 5-fold cross-validation procedure was employed to ensure evaluation of each sampled combination.

2.8 Final Evaluation

Model performance was quantitatively assessed using three standard regression metrics (R-squared, mean absolute error, and root mean squared error). This combination provides a comprehensive view of accuracy, average error magnitude, and sensitivity to large errors.

2.8.1 R-Squared (R^2)

The proportion of variance in the target variable is predictable from the features. A value closer to 1 is better.

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

n: Total number of data points
y_i: Actual (observed) value for the i^{th} instance
$\hat{y}$: Predicted value for the i^{th} instance
$\bar{y}$: Mean of the actual observed values

2.8.2 Mean Absolute Error (MAE)

The average absolute difference between predicted and actual values, expressed in the units of the target variable. Lower is better.

$$MAE = \frac{1}{n}\sum_{n=1}^{n}(|y_i - \hat{y}_i|) \qquad (2)$$

All variables are the same as in Equation (1).

2.8.3 Root Mean Squared Error (RMSE)

The square root of the average of squared errors. It penalizes large errors more heavily than MAE. Lower is better.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (3)$$

All variables are the same as in Equation (1).

2.9 Experimental Design

To directly address the research gap concerning the impact of preprocessing, a four-scenario experimental design was implemented. This framework systematically tests the effect of outlier capping and feature selection, both individually and in combination. The four scenarios are:

- Full preprocessing (outlier capping + feature selection)

- Feature selection only (raw data with outliers)
- Outlier capping only (for V_{Sh} & GR)
- Raw data (no capping or feature selection)

Each of these four data configurations was subjected to the identical modeling pipeline (splitting, scaling, training, tuning, and evaluation), ensuring that any observed differences in performance could be confidently attributed to the specific preprocessing strategy employed.

3 Results and Discussion

3.1 Exploratory Data Analysis and Preprocessing Outcomes

The initial analysis of the raw, aggregated dataset revealed several key characteristics that guided the experimental design. The dataset was found to be complete with no missing values. The calculated Pearson correlation coefficient between the V_{Sh} and GR logs was only 0.28. This weak correlation contradicts established petrophysical principles, which dictate that GR is a primary indicator for shale content and a strong positive correlation is expected [5] and a subsequent capping procedure was applied exclusively to these two variables. The effect of this targeted intervention is clearly visualized in the correlation matrix shown in Figure 2.

The analysis of Figure 2 confirms the hypothesis. As shown in the heatmap, the coefficient of correlation between V_{Sh} and GR increased to 0.94 after the targeted capping. This demonstrates that the outliers were indeed masking the true underlying linear relationship between these two geologically linked parameters. Furthermore, the heatmap confirms the persistent and extremely high correlation between depth and Stress ($r \approx 1$) and the strong negative correlation between RHOB and Porosity ($r \approx -0.9$). This result validated the targeted capping as a necessary step for reliable statistical analysis and informed the creation of the "capped" datasets for the experimental scenarios.

3.2 Feature Selection Results

To address the identified multicollinearity, a comparative analysis of three feature selection methods was performed on the cleaned (capped) data. The results are summarized in Table 2.

While all methods identified stress as top-tier predictor, the more sophisticated methods (RFE and random forest importance) provided a clearer verdict on the redundant features. Both methods significantly down-ranked depth compared to stress, porosity compared to RHOB and prioritized V_{Sh} over GR. Based on the robustness of the random forest importance method, the features identified for removal in the "with feature selection" scenarios were depth, GR, and porosity.

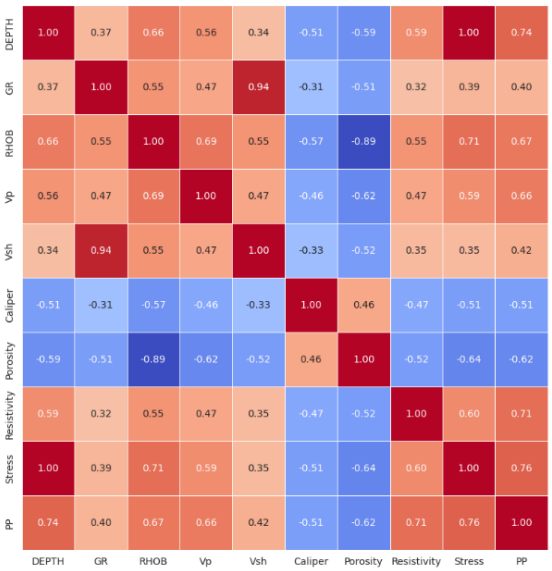

Figure 2: Correlation heatmap of features after capping

Table 2: Feature selection analysis on cleaned data (after capping)

Feature	F-Score (Univariate)	RFE Rank	RF Importance
Stress	15601.81	1	0.473722
Depth	14304.98	8	0.081773
Resistivity	8636.99	4	0.232932
RHOB	8720.29	5	0.018527
V_p	0.76	6	0.139250
Porosity	6769.38	7	0.006142
Caliper	2429.48	9	0.012542
V_{Sh}	2492.24	2	0.023211
GR	2125.14	3	0.011902

3.3 Model Performance Evaluation

The core findings of this research are presented in the performance evaluation figures (Figures 3 and 4), which compare R^2, MAE, and RMSE across all scenarios. Figure 3 summarizes the

results of the baseline models, while Figure 4 presents the final, optimized performance after hyperparameter tuning.

The baseline results immediately highlight the inherent robustness of the ensemble models (random forest and XGBoost), which consistently outperformed the other algorithms. They also provide initial evidence of SVR's sensitivity to outliers, as its performance is notably poor in the scenarios without capping.

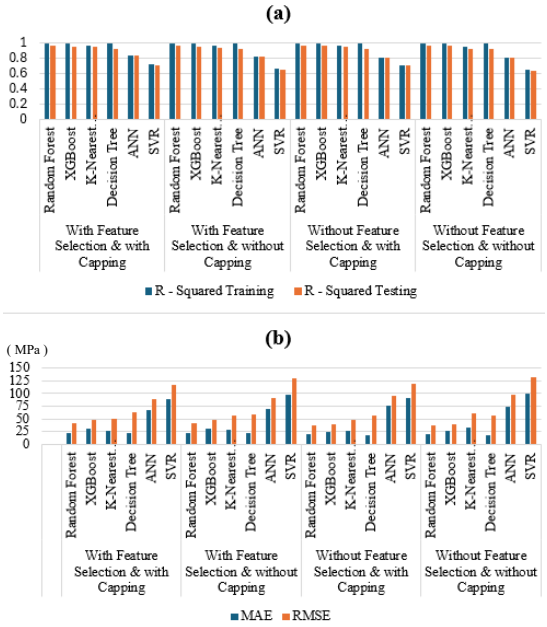

Figure 3: Evaluation Criteria before Tuning Across All Scenarios: (a) R^2 for training and testing; (b) MAE and RMSE on the testing set.

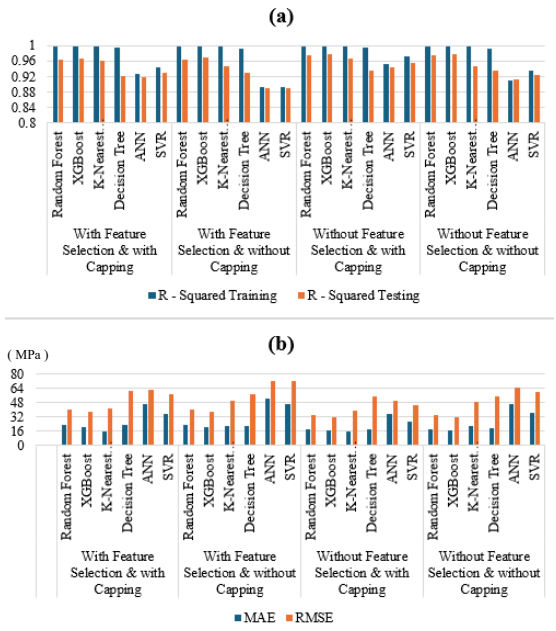

Figure 4: Evaluation criteria after hyperparameter tuning across all scenarios: (a) R^2 for training and testing; (b) MAE and RMSE on the testing set.

The results after tuning provide the basis for the main conclusions of this research. A deep analysis reveals several critical insights.

The single best performance was achieved by the tuned XGBoost model in scenario 4 (without feature selection & without capping). This configuration yielded the highest testing R^2 of 0.9789 and the lowest RMSE of 31.98 MPa. The optimal hyperparameters that produced this result were: n_estimators=700, max_depth=10, learning_rate=0.05, subsample=0.8, colsample_bytree=1.0, and gamma=0.

For the top-tier models (XGBoost and Random Forest), performance was consistently better when no features were removed. This suggests that the removed features, while linearly correlated, contained unique non-linear information that these advanced models were able to exploit.

The effect of outlier capping was highly model-dependent. For the robust XGBoost and random forest models, the impact was minimal. However, for the sensitive SVR model, outlier capping was critical, improving its R^2 score and transforming it into a competitive model.

Comparing Figure 3 and Figure 4 shows that tuning was universally beneficial, with the improvement seen in the SVR model, which was transformed from the worst-performing baseline model into a strong contender.

3.4 Diagnostic Plot Validation

To visually supplement the quantitative metrics, diagnostic plots were generated. Figures 5, 6 & 7 presents a curated selection of "Actual vs. Predicted" plots that tell a compelling story of model performance.

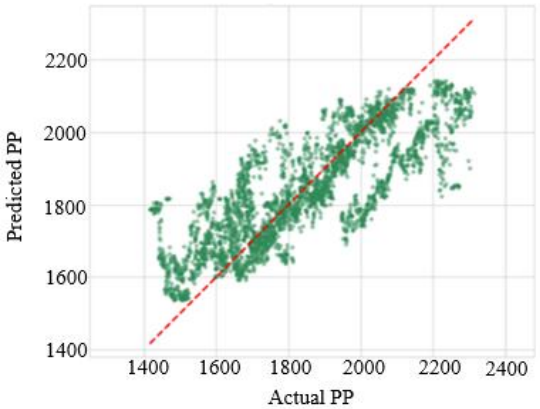

Figure 5: Baseline SVR (Scenario 1 - Testing)

Figure 5 illustrates a poor fit for the baseline SVR model, with significant scatters and a low R^2 of 0.6447.

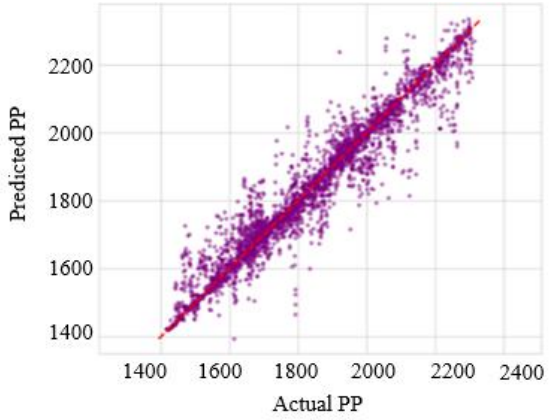

Figure 6: Tuned SVR (Scenario 1 - Testing)

Figure 6 shows the improvement after tuning, with the points clustering much more tightly around the reference line and the R^2 jumping to 0.8990. Finally, Figure 7 displays the optimized model's performance. The exceptionally tight clustering of the testing data around the 45-degree line visually confirms the high R^2 of 0.9789 and demonstrates the model's excellent generalization capability on unseen data.

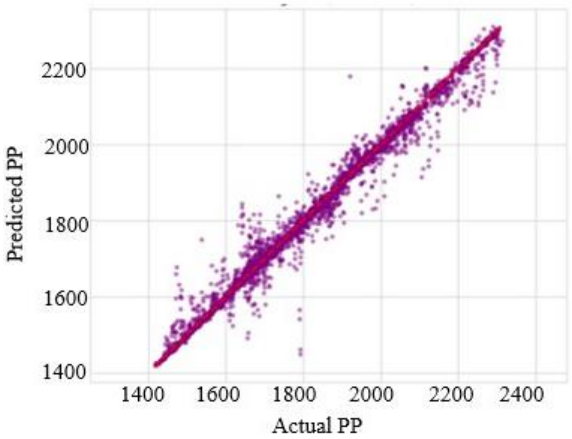

Figure 7: Tuned XGBoost (Scenario 4 - Optimized Model)

4 Conclusion

The results of this study provide an understanding of the relationship between data preprocessing and model performance in the context of pore pressure prediction. The most significant finding is the emergence of the Tuned XGBoost model, trained on raw, unprocessed data, as the optimized model. This outcome challenges the conventional wisdom that extensive data cleaning is a prerequisite for optimal performance and highlights the inherent robustness of state-of-the-art ensemble algorithms.

Two key insights explain this result. First, the analysis revealed that removing features based on linear correlation was detrimental to the top-performing models. The higher accuracy of XGBoost and random forest when using the full feature set suggests that these algorithms successfully extracted unique, non-linear information from features like depth that appeared redundant in a linear context. This demonstrates that for complex models, linear multicollinearity does not necessarily equate to informational redundancy.

Second, the study confirmed that the necessity of outlier treatment is highly model-dependent. While robust tree-based models were largely unaffected, the performance of the sensitive SVR model was critically dependent on outlier capping. This empirically validates that outlier treatment should not be a universal, default step but rather a targeted intervention required to enable sensitive models to perform competitively. The four-scenario design was therefore essential, as it created a level playing field to compare the optimal potential of every algorithm, rather than evaluating sensitive models on data to which they are fundamentally ill-suited.

The findings demonstrate that for a sufficiently powerful and robust algorithm like XGBoost, extensive data manipulation such as outlier capping and the removal of linearly correlated features may be unnecessary and even counterproductive. The optimal end-to-end pipeline was the one with the least preprocessing, prioritizing the algorithm's ability to learn complex patterns from unaltered data. This research provides strong empirical evidence that the best practices of data preparation are not universal but are instead intrinsically linked to the chosen algorithm's complexity and natural resilience. Future work should focus on applying model interpretability techniques to the best model and validating these findings on diverse geological datasets.

References

[1] S. Ramatullayev et al., “Formation pressure while drilling technology: Game changer in drilling overpressured reservoirs,” in *Abu Dhabi International Petroleum Exhibition & Conference*, Abu Dhabi, UAE, Nov. 2019. doi: 10.2118/198367-MS

[2] B. A. Eaton, "The theory of abnormal pore pressure prediction," *Journal of Petroleum Technology*, vol. 27, no. 1, pp. 21–26, 1975. doi: 10.2118/5173-PA.

[3] M. Azadpour et al., "Pore pressure prediction and modeling using well-logging data in one of the gas fields in south of Iran," *Journal of Petroleum Science and Engineering*, vol. 128, pp. 15–23, 2015. doi: 10.1016/j.petrol.2015.02.022.

[4] A. D. Ogbu et al., "Advances in machine learning-driven pore pressure prediction in complex geological settings," *Computer Science & IT Research Journal*, vol. 5, no. 7, pp. 1648–1665, 2024. doi: 10.51594/csitrj.v5i7.1350.

[5] J. Feng et al., "Pore pressure prediction for high-pressure tight sandstone in the Huizhou Sag, Pearl River Mouth Basin, China: A machine learning-based approach," *Journal of Marine Science and Engineering*, vol. 12, no. 5, p. 703, 2024. doi: 10.3390/jmse12050703.

[6] M. Sanei, A. Ramezanzadeh, and A. Asgari, "Applied machine learning-based models for determining the magnitude of pore pressure and minimum horizontal stress," Arabian Journal of Geosciences, vol. 17, no. 210, 2024. doi: 10.1007/s12517-024-11997-2.

[7] A. Abdelaal, S. Elkatatny, and A. Abdulraheem, "Data-driven modeling approach for pore pressure gradient prediction while drilling from drilling parameters," *ACS Omega*, vol. 6, no. 21, pp. 13807–13816, 2021. doi: 10.1021/acsomega.1c01340.

[8] Tammy Reservoir, "Pore-Pressure-Prediction-for-Oil-and-Gas [Source code]," GitHub, 2022. [Online]. Available: https://github.com/tammyreservoir/Pore-Pressure-Prediction-for-Oil-and-Gas

[9] P. Schober and C. Boer, "Correlation Coefficients: Appropriate Use and Interpretation," *ResearchGate*, 2018. https://www.researchgate.net/publication/323388613_Correlation_Coefficients_Appropriate_Use_and_Interpretation

[10] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*, 8th ed. Cengage, 2019.

[11] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 281–305, 2012.