NABEA AL-AJYAL FOUNDATION PRESS
Online ISSN: 3078-5413
CyberSystem Journal (CSJ)
csj.nabea.pub
CyberSystem Journal Research Article

Enhancing Lithology Prediction through Preprocessing Techniques and Machine Learning Models

Cherif Mesroua[1*], Ibrahim Lahouel[2], Basma Hamrouni[3], Khadra Bouanane[4], Faiza Zidouni[5] and Wafa Kafi[6]

[1,2,3,4]Department of Computer Science, Faculty of New Technologies of Information and Communication (FNTIC), University of Kasdi Merbah Ouargla, Ouargla, Algeria

[5,6]Department of Physics, University of Science and Technology Houari Boumediene (USTHB), Algiers, Algeria

* Corresponding Author: **Cherif Mesroua**, Email: mesroua.cherif@univ-ouargla.dz.

***Abstract*:** Lithology classification through well log interpretation is a fundamental task in reservoir characterization, enabling accurate delineation of subsurface formations and assessment of hydrocarbon potential. However, measurements are rarely full, and missing data intervals are prevalent due to operational difficulties or logging device failure. Thus, imputation of missing data from down-hole well logs is a prevalent issue in subsurface processes. Our work a strong emphasis on the preprocessing phase and data imputation, acknowledging that missing data in well logging is a prevalent problem that can have a major impact on classification results. Our work is part of the FORCE2020 Lithology Classification Competition. Our method underlines how important extensive data preprocessing is for improving model performance, including regression-based imputation, normalization, and class balancing by SMOTE. Traditional models like Random Forest and XGBoost were able to produce reliable results in the challenging FORCE2020 Lithology Classification. By leveraging multiple models, we aim to enhance the accuracy and robustness of our predictions, addressing the challenges posed by missing data and ensuring a more reliable classification process. We show that the Random Forest model obtains the greatest accuracy of 95% using the FORCE 2020 dataset from 118 wells in the Norwegian Sea. This study emphasizes how crucial thorough data imputation and preprocessing techniques are to raising the precision and dependability of lithology classification.

Access this article online

***Keywords*:** Lithology classification, FORCE 2020 dataset, Machine learning

1. Introduction

Increasing drilling efficiency is a primary goal of well drilling. Improving real-time

drilling efficiency and developing automation are essential to supplying the increasing demand for hydrocarbons [1-8]. Real-time lithology determination during drilling operations is essential for increasing drilling efficiency. Significant variations in the azimuth and inclination of the well axis might result from various formation layer types, increasing the risk of vibrations and downtime [9-12]. Drilling efficiency is increased by accurate lithology type prediction and the related rock strength, which aid in borehole stability and Rate of Penetration (ROP) analysis. However, it might be difficult to determine lithology using the conventional way of looking at the cuttings that are received at the shale shaker, especially in interbedded sections. Using petrophysical and drilling information, this task offers a chance to use machine learning (ML) techniques to forecast the lithology near the

Received February 8, 2025; Revised March 1, 2025; Accepted April 2, 2025; Published June 30, 2025
https://doi.org/10.57238/csj.2025.1007

drill bit. Overcoming the drawbacks and delays of conventional techniques and successfully resolving sensor off set difficulties, such a methodology allows for a more precise and timely determination of rock kinds [13].

Lithology classification is a cornerstone of subsurface geological analysis, particularly in hydrocarbon exploration and reservoir characterization. Accurate identification of lithofacies from well log data enables geoscientists to delineate subsurface formations, assess resource potential, and optimize extraction strategies [1,2]. Traditional methods, which rely on manual interpretation by experts, are not only time-consuming but also prone to subjective biases, especially when dealing with heterogeneous or complex reservoirs [3-4]. The integration of machine learning (ML) techniques has revolutionized this field by introducing automated, data-driven approaches that enhance both efficiency and consistency in lithology prediction [1-5].

The FORCE 2020 Machine Learning Competition emerged as a pivotal initiative in advancing ML applications for lithology classification, providing a standardized dataset of 118 wells from the Norwegian Sea with diverse well-log measurements including gamma-ray (GR), resistivity (RDEP), density (RHOB), and neutron porosity (NPHI) along side interpreted lithofacies and lithostratigraphy [6,7]. This dataset has served as a benchmark for evaluating the robustness of predictive models under realistic conditions, such as imbalanced class distributions and geologically informed error penalties. For instance, misclassifying shale as marlincurred a lower penalty than confusing shale with anhydrite, reflecting the practical nuances of geological interpretation [7]. The competition attracted 329 teams globally, with top-performing models demonstrating marginal differences in accuracy on blind test data, underscoring the challenges of overfitting and distribution shifts between training and real-world datasets [6 ,7].

Recent studies have leveraged this dataset to explore a wide array of ML algorithms. For example, [5] compared linear models, k-nearest neighbors, and gradient-boosted decision trees (GBDTs), identifying CatBoost as the top performer in a Siberian oil field study with 86 wells and six physical parameters[2]. Similarly, an analysis of the Baikouquan Formation in China's Junggar Basin revealed that ensemble methods like Random Forest and Extreme Gradient Boosting (XGBoost) consistently outperformed linear classifiers such as logistic regression, achieving accuracy margins of up to 15%[1]. In the Niuxintuo Block of China's Liaohe Oil field, support vector machines (SVMs) achieved 93% accuracy in distinguishing six lithology classes, surpassing Bayesian discriminate analysis (58.2%) and other ML models like convolutional neural networks (CNNs) [3]. These findings highlight the superiority of nonlinear and ensemble methods in capturing complex relationships between well-log parameters and lithology [1-3].

In the FORCE 2020 Lithofacies Prediction competition, the top-performing teams, Olawale, GIR, and Lab.ICA, used careful preprocessing techniques in conjunction with tree-based ensemble approaches to address the challenging task of classifying lithofacies using incomplete well log data. Using the XGBoost algorithm and a 10-fold stratified cross-validation strategy, Olawale, the winner, obtained the best accuracy on the blind test data. His method included feature engineering via windowing and gradient computations, but it was noteworthy that he dropped uncommon curves

like SGR and DTS rather than impute any missing values. Context-aware imputation was a major focus of the GIR team, which also employed XGBoost. They used the physical relationships between logs to rebuild missing values. They placed second thanks to their outstanding classification performance, which was aided by the addition of polynomial features and non-local gradients to the model input. The third-place Lab.ICA team used a Random Forest classifier with five-fold cross-validation and their feature set contained normalized log values, log gradients, and information on geological formations. They used a straightforward median imputation technique to fill in the missing values.

While the top teams used a variety of approaches, including XGBoost with feature engineering, context-aware imputation, and Random Forest with median imputation, there were significant discrepancies in their methodologies. Olawale's decision to eliminate imputation and drop unusual curves may have improved accuracy in some circumstances, but it also risks missing crucial data patterns and limiting model applicability. The GIR team's use of physical relationships for imputation, while clever, may not be as effective in capturing all potential data nuances, especially in complex datasets. The Lab.ICA team's use of basic median imputation may be considered insufficient for dealing with more complex missing data patterns. Neither of these methods completely embraces a comprehensive model selection strategy that includes numerous algorithms.

To address the above challenges, a key aspect of our methodology lies in the advanced preprocessing, this significantly contributed to the improved performance of even standard models. We conducted a detailed analysis of missing values and removed columns with more than 50% missing data to reduce noise. Remaining gaps were imputed using linear interpolation, a method that better captures data continuity compared to the mean imputation adopted by

other competitors. Continuous features were normalized using MinMaxScaler, and categorical variables such as GROUP and FORMATION were excluded due to their limited predictive value. Moreover, to address the severe class imbalance, we employed SMOTE to generate over 8.6 million synthetic samples for the minority class, balancing the dataset and enhancing model learning. These preprocessing strategies proved crucial in boosting model performance and underscore the idea that a well-designed preprocessing workflow can significantly elevate the effectiveness of even traditional machine learning models in complex tasks such as lithology classification. We use a broad set of models, including RNN, Random Forest, and XGBoost, with comprehensive adjustment of many hyperparameters to improve performance.

We place a strong emphasis on the preprocessing phase and data imputation, acknowledging that missing data in well logging is a prevalent problem that can have a major impact on classification results.

By implementing a more comprehensive method that handles missing values using advanced imputation techniques, we hope to reduce any biases caused by incomplete data and assure more accurate and resilient model performance. This method not only improves prediction reliability but also increases the model's ability to generalize across a wide range of real-world data circumstances.

2. Methodology

This study aims to classify lithologies from well log data using machine learning models. The methodology involves four major steps: data acquisition, pre-processing, feature selection, model training and evaluation (see **Error! R eference source not found.**).

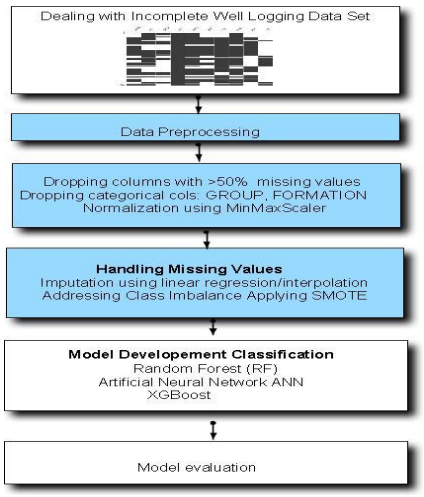

Figure 1. Workflow for our lithology prediction methods.

A key aspect of our methodology lies in the advanced preprocessing, which significantly contributed to the improved performance of even standard models. We conducted a detailed analysis of missing values and removed columns with more than 50% missing data to reduce noise. Remaining gaps were imputed using linear interpolation, a method that better captures data continuity compared to the mean imputation adopted by other competitors. Continuous features were normalized using MinMaxS-caler, and categorical variables such as GROUP and FORMATION were excluded due to their limited predictive value. Moreover, to address the severe class imbalance (103 vs. 107,000 samples), we employed SMOTE (Synthetic Minority Over-sampling Technique) to generate over 8.6 million synthetic samples for the minority class, balancing the dataset and enhancing model learning. These preprocessing strategies proved crucial in boosting model performance and underscore the idea that a well-designed preprocessing workflow can significantly elevate the effectiveness of even traditional machine learning models in complex tasks such as lithology classification.

2.1 Dataset

The dataset used in this work comes from the FORCE 2020 Machine Learning Competition on lithology prediction [5]. It comprises well log data from multiple wells, containing 29 petrophysical measurements such as gamma ray (GR), neutron porosity (NPHI), bulk density (RHOB), sonic logs, and others, along with lithology labels used for classification, the number of samples in this dataset is more than 1.7 million samples.

2.2 Data Preprocessing

To prepare the dataset for model training, the following preprocessing steps were applied . Handling Missing Values. Missing values were analyzed across all columns. Columns with more than 50% missing values were removed to eliminate sparsity and reduce noise. Imputation and Normalization. Remaining missing values were imputed using linear interpolation. Continuous features were normalized using MinMaxScaler to ensure consistent feature scaling. other competitors used mean interpolation.

Dropping Categorical columns. Categorical features such as GROUP and FORMATION were dropped because they were irrelevant to the prediction.

Dealing with unbalanced classes. We encountered a significant class imbalance, with the minority class having only 103 samples, while the majority class had 107,000 samples, We used SMOTE (Synthetic Minority Over-sampling Technique), a library for handling imbalanced

datasets, This technique generated synthetic data (the new samples number more than 8.6 million samples) points for the minority class, effectively increasing the dataset size and balancing the class distribution. Target Variable Preparation. Categorical features such as GROUP and FORMATION were dropped because they were irrelevant to the prediction.

2.3 Feature selection

After pre-processing, feature selection was conducted using domain knowledge (GR is mostly used to determine the lithology) and correlation analysis. Key features such as CALI, RDEP, RHOB GR, NPHI, PEF were retained based on their result of the correlation.

2.4 Model Development

Two machine learning models were developed and compared:

- **Artificial Neural Network (ANN).** A fully connected feed-forward neural network was implemented using Tensor Flow/Keras. The architecture consisted of input (6), 3 hidden layers (128, 64, 64 neurons), and output layers (12 neurons) with ReLU and softmax activations. Dropout layers were added to prevent overfitting.
- **Random Forest Classifier (RF).** A Random Forest classifier was implemented using scikit-learn. we used different n estimators starting from 10 to 50.
- **XGBoost Classifier (XGB).** An XGBoost classifier was implemented using the XGBoost library in Python. This gradient boosting method is known for its high performance and efficiency in handling structured data.

2.5 Model training

- **Artificial Neural Network (ANN).** Both models were trained on the preprocessed dataset. the ANN was trained using Adam optimizer with sparse categorical cross entropy loss, learning rate = 0.001, 10 epochs (model could be improved add more training loops) , batch size = 32, validation set was the 0.2 part from the train data, the
- model took approximately 2 hours to train. we used the ANN here because after used SMOTE the size of the data increased significantly and we assumed with larger size ANN could outperform the random forest and xgBoost aslo XGBoost is known to overfit small datasets if not finely tuned. In contrast Random Forest uses bootstrapped samples and averaging to reduce overfitting, ANN with dropout layers and regularization can generalize well if trained carefully, and benefits from GPU acceleration for faster training.
- **Random Forest Classifier (RF).** we tried different hyperparameters, we focused more on the n estimators and that makes the difference, we used 10 folds cross validation.we also tried another approach which is to train 12 binary ANN models with are relatively simpler and combine them for the final prediction but this gave good result on the validation data 89% and bad result on the testing set 20% (this was before the class balancing), we used the same approach using logiticre.
- **XGBoost Classifier (XGB).** The XGBoost model was trained on the same preprocessed dataset. As a gradient boosting algorithm, it is designed to optimize performance through an ensemble of decision trees.

2.6 Model Evaluation

Evaluation was performed using metrics such as accuracy, precision, recall, and F1-score. The ANN achieved an accuracy of 89% on validation but 61% on leader-board test and this might be related because the test data is to an extent different from the train, while the Random classifier and with a special pre-processing which is removing just the features with more than 50% and using 15 estimators achieved a superior performance of 95.8% in both accuracy and F1-score on validation but 69.4% on leader-board test.

3. Results and Discussion

The performance of the Artificial Neural Network (ANN) and Random Forest (RF) models was evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score. Table I summarizes the performance of both models on the test dataset.

Table 1. Performance Comparison Of ANN, Random Forest Models And XGBOOST

Metric	ANN	RF	XGBoost
Accuracy(valid/test)	0.89 / 0.61	**0.96/0.69**	0.88/0.55
Precision(valid/test)	0.89 / 0.54	**0.96/-**	0.87/0.39

Recall(valid/test)	0.89 / 0.61	**0.96/-**	0.96/0.56
F1-score(valid/test)	0.89 / 0.54	**0.96/0.62**	0.96/0.46

The results indicate that the Random Forest model outperformed the Artificial Neural Network across all evaluation metrics. The superior performance of the Random Forest classifier can be attributed to its ensemble nature, which combines multiple decision trees to reduce overfitting and increase model stability. Moreover, RF handles heterogeneous features and missing data more effectively compared to neural networks, which require more preprocessing and are sensitive to hyperparameter tuing.

In contrast, while the ANN achieved reasonably high accuracy, its performance was comparatively lower. The ANN is more flexible and powerful in modeling non-linear relationships, but it typically requires a larger amount of data and careful regularization techniques to avoid overfitting. In this study, the relatively small training dataset and class imbalance may have limited the ANN's generalization capability.

4. Conclusion

This paper investigated the application of advanced machine learning techniques for automated lithology classification using well log data from the FORCE 2020 competition dataset. We implemented and compared three widely used models: an Artificial Neural Network (ANN) ,a Random Forest (RF) and XGBoost classifier. Both models were trained on a preprocessed dataset featuring standard well logs after handling missing values, encoding categorical features, and scaling numerical data.

The experimental results demonstrated the effectiveness of machine learning for this geological interpretation task. The Random Forest model achieved a superior classification accuracy of 95% on the validation set, significantly out-performing the ANN model, which achieved 89% accuracy.

This highlights the robustness and suitability of ensemble methods like Random Forest for analyzing tabular well log data and capturing complex relationships relevant to lithology prediction.The study confirms that machine learning can provide accurate and efficient lithology predictions, offering valuable support for subsurface exploration and reservoir characterization workflows. The high accuracy achieved by the RF model underscores its potential as a practical tool for geoscientists seeking to automate or augment log interpretation.

Future research could explore several avenues. Investigating other advanced techniques like Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTMs) that could potentially yield further performance improvements. Additionally, testing the developed models on datasets from different geological settings would be valuable for assessing their generalization capabilities and practical applicability in diverse exploration scenarios.

This study's primary shortcoming is that, rather than creating original classification methods, we mainly focused on investigating a reliable data preprocessing procedure. However, this approach proved to be highly effective in the context of lithology classification, yielding promising results on the FORCE2020 dataset. We intend to build on this research in the future by applying increasingly complex artificial intelligence models to improve prediction accuracy and generalization capability.

References

[1] learning methods for high-quality reservoir identification: A case study of Baikouquan formation in Mahu Area of Junggar Basin, NW China," *Energies*, vol. 15, no. 10, p. 3675, 2022.doi:

[2] Y. Meshalkin *et al.*, "Well-logging-based lithology prediction using machine learning," *Data Science in Oil & Gas 2020*, Eur. Assoc. Geosci. & Eng., vol. 2020, no. 1, 2020.

[3] Z. Fan *et al.*, "Logging-data-driven lithology identification in complex reservoirs: an example from the Niuxintuo block of the Liaohe oilfield," *Front. Earth Sci.*, vol. 13, p. 1491334, 2025.doi: 10.3389/feart.2025.1491334 .

[4] S. R. Manda, R. Rohit *et al.*, "Identification of lithology from well log data using machine learning," 2024.doi: 10.4108/eetiot.5634 .

[5] T. Merembayev, R. Yunussov, and Y. Amirgaliyev, "Machine learning algorithms for classification geology data from well logging," in *Proc. 14th Int. Conf. Electron. Comput. Comput. (ICECCO)*, pp. 1–6, IEEE, 2018.doi: 10.1109/icecco.2018.8634775 .

[6] R. A. Mardani, M. Mardani, and D. Trad, "Rock facies imbalanced classification with over-sampling and under-sampling techniques," *arXiv preprint*, arXiv:2306.06509, 2023.

[7] S. A. Garini *et al.*, "Enhanced lithology classification in well log data using ensemble machine learning techniques," in *Proc. 2024 IEEE Int. Conf. Artif. Intell. Mechatronics Syst. (AIMS)*, pp. 1–5, IEEE, 2024.doi: 10.1109/aims61812.2024.10512485 .

[8] A. Hallam, D. Mukherjee, and R. Chassagne, "Multivariate imputation via chained equations for elastic well log imputation and prediction," *Appl. Comput. Geosci.*, vol. 14, p. 100083, 2022.doi: 10.31223/x57k6q .

[9] A. Elmgerbi, E. Chuykov, G. Thonhauser, and A.

Nascimento, "Machine learning techniques application for real-time drilling hydraulic optimization," in *Proc. Int. Petroleum Technol. Conf. (IPTC)*, Riyadh, Saudi Arabia, 21–23 Feb. 2022.doi: 10.2523/iptc-22662-ms

[10] T. Burak and S. Akin, "Estimation of downhole inclination in directionally drilled geothermal wells," in *Proc. World Geothermal Congr. 2020+1*, Reykjavik, Iceland, pp. 1–5, 2021.doi:

[11] A. Sharma, M. Al Dushaishi, and R. Nygaard, "Evaluating PDC bit-rock interaction models to investigate torsional vibrations in geothermal drilling," *Geothermics*, vol. 20, p. 103060, 2024.doi: 10.2139/ssrn.4724858

[12] A. Sharma, M. Al Dushaishi, and R. Nygaard, "Fixed bit rotary drilling failure criteria effect on drilling vibration," in *Proc. 55th U.S. Rock Mechanics/Geomechanics Symp.*, Online, pp. 1–6, 2021.

[13] T. Burak, "Application of artificial neural networks to predict the downhole inclination in directionally drilled geothermal wells," M.S. thesis, Middle East Tech. Univ., Ankara, Türkiye, 2018.

How to cite this article

Mesroua C., Lahouel I., Hamrouni B., Bouanane K., Zidouni F., & Kafi W., "Enhancing Lithology Prediction through Preprocessing Techniques and Machine Learning Models," *CyberSystem J.*, vol. 2, no. 1, pp. 65-70, 2025. doi: 10.57238/csj.2025.1007

Access this article online