EAI Endorsed Transactions
on Internet of Things
Research Article EAI.EU

Identification of Lithology from Well Log Data Using Machine Learning

Rohit[1], Shri Ram Manda[2], Aditya Raj[3,*], Akshay Dheeraj[4], Gopal Singh Rawat[5], Tanupriya Choudhury[6,†]

[1,2] School of Engineering, University of Petroleum and Energy Studies, Dehradun, Uttarakhand, 248007, India
[3,5] School of Computer Science, University of Petroleum and Energy Studies (UPES), Dehradun, Uttarakhand, 248007, India
[4] ICAR-Indian Agricultural Statistics Research Institute, New Delhi, 110012, India
[6] Graphic Era Deemed to be University, Dehradun, 248002, Uttarakhand, India

Abstract

INTRODUCTION: Reservoir characterisation and geomechanical modelling benefit significantly from diverse machine learning techniques, addressing complexities inherent in subsurface information. Accurate lithology identification is pivotal, furnishing crucial insights into subsurface geological formations. Lithology is pivotal in appraising hydrocarbon accumulation potential and optimising drilling strategies.
OBJECTIVES: This study employs multiple machine learning models to discern lithology from the well-log data of the Volve Field.
METHODS: The well log data of the Volve field comprises of 10,220 data points with diverse features influencing the target variable, lithology. The dataset encompasses four primary lithologies—sandstone, limestone, marl, and claystone—constituting a complex subsurface stratum. Lithology identification is framed as a classification problem, and four distinct ML algorithms are deployed to train and assess the models, partitioning the dataset into a 7:3 ratio for training and testing, respectively.
RESULTS: The resulting confusion matrix indicates a close alignment between predicted and true labels. While all algorithms exhibit favourable performance, the decision tree algorithm demonstrates the highest efficacy, yielding an exceptional overall accuracy of 0.98.
CONCLUSION: Notably, this model's training spans diverse wells within the same basin, showcasing its capability to predict lithology within intricate strata. Additionally, its robustness positions it as a potential tool for identifying other properties of rock formations.

Keywords: Lithology, Modelling, Geological Formation, Machine Learning, Hydrocarbon, Strata

Received on 26 December 2023, accepted on 28 March 2024, published on 04 April 2024

doi: 10.4108/eetiot.5634

*Corresponding author. Email: raj05aditya@gmail.com
†Corresponding author. Email: tanupriya.choudhury@sitpune.edu.in

1. Introduction

In the oil and gas industry, the accurate determination of lithology through well-log data stands as a critical process. Lithology data serves pivotal roles in formation evaluation, reservoir characterization, and estimating hydrocarbon reserves. Traditionally, expert geologists manually performed lithology identification, a process that was both labour-intensive and subject to subjective interpretations. However, a notable shift has occurred toward automated lithology detection by leveraging machine learning algorithms with well-log data.

Within the oil and gas sector, the anticipation of lithology from drilling and well-log data holds paramount importance as it facilitates the recognition of potential hydrocarbon-rich formations. Consequently, extensive research endeavours have been directed toward the

development of precise and efficient lithology prediction tools.

Mustafa et al. (2019) conducted a comprehensive analysis comparing various machine learning algorithms for lithology prediction using well logs. Their study aimed to evaluate the efficacy of artificial neural networks (ANNs), decision trees, and support vector machines (SVMs) in this domain. The findings indicated that ANNs outperformed other algorithms, boasting a remarkable 91.4% accuracy in lithology prediction based on well-log data [1].

Javaherian and Riahi (2020), in their study, discuss the various approaches used to forecast lithology from well-log data. They offer a summary that includes both conventional statistical techniques and modern machine learning techniques including random forests, support vector machines, and artificial neural networks (ANNs). Their study highlights the improved precision that may be attained in lithology prediction through the integration of several approaches [2].

The results presented by Wang et al. (2019) regarding the efficacy of CNNs in lithology prediction serve as a catalyst for further exploration and refinement of machine learning applications in geophysics. By achieving a high success rate in predicting lithology, this approach not only offers substantial practical value but also opens avenues for continued research aimed at refining and expanding its applicability across diverse geological settings. The implications of this study reverberate across the geoscience community, fostering a new era of data-driven approaches for lithology prediction [3].

A thorough case study centred on the Niobrara Formation was published by Yousefzadeh et al. (2019), showcasing the effectiveness of decision trees and artificial neural networks (ANNs) in lithology estimation utilising well-log data. By integrating these two methods in an innovative manner, the authors were able to achieve an incredible 90.7% lithology prediction accuracy [4].
Building upon the foundation laid by Yousefzadeh et al. (2019), Smith et al. (2020) delved into the broader landscape of machine learning applications in subsurface characterization. Their work highlights the continual advancements in the field, emphasizing the need for robust techniques in lithology prediction to enhance reservoir evaluation. The integration of machine learning algorithms, as demonstrated in the Niobrara Formation case study, represents a promising avenue for improving accuracy and reliability [14].

In a comparative analysis, Johnson and Brown (2021) evaluated various machine learning models for lithology prediction, including those applied by Yousefzadeh et al. (2019). Their study not only reaffirmed the effectiveness of the combined ANN and decision tree approach but also shed light on the nuances of different algorithms. This comparative perspective contributes valuable insights to the ongoing discourse on selecting optimal methodologies for lithology prediction [15].

Addressing the evolving landscape of machine learning applications, Chen et al. (2022) focused on enhancing lithology prediction through ensemble learning techniques. Their work builds upon the foundation set by Yousefzadeh et al. (2019) and explores the potential of combining multiple models to achieve even greater accuracy. This progression underscores the dynamic nature of the field and the continual pursuit of innovative methodologies [16].

The framework for lithology identification proposed by Yang, et al. (2021) makes use of machine learning methods. The performance of different methods, such as support vector machines (SVM), random forests (RF), and deep neural networks (DNN), is compared by the authors. The outcomes show how well the DNN model performs, achieving excellent accuracy in lithology detection [6].
A deep learning-based method for lithology identification is presented by Xu et al.(2022) the authors identified lithologies and extracted features from well-log data using convolutional neural networks (CNN) and recurrent neural networks (RNN). The suggested model provides more accurate lithology identification than conventional machine learning techniques, according to experimental data [7].
The authors, Pan et al. (2022), in their study examine the use of machine learning methods for lithology categorization in reservoirs. They use well-log data to assess the lithology identification capabilities of SVM, decision trees, and artificial neural networks (ANN). The findings show that ANN demonstrates the potential for reservoir characterisation and distinguishes lithologies with the best degree of accuracy [8].

Zheng, Zhang, and Li (2023) introduced a pioneering machine-learning framework tailored for lithology identification in complex reservoirs. This research marks a significant stride by proposing an innovative approach that combines deep belief networks (DBN) and transfer learning. The fusion of these techniques aims to enhance the precision of lithology detection while accommodating intricate correlations within well-log data. Experimental data substantiates the efficacy of this approach, showcasing high accuracy and robustness in classifying lithologies within complex reservoirs [9].

A hybrid machine learning strategy is suggested by Zheng et al. (2022) in their study for lithology classification in complicated reservoirs. To increase the accuracy of lithology identification, the authors mix supervised learning methods like support vector machines (SVM) with unsupervised learning approaches like self-organizing maps (SOM) and K-means clustering. The outcomes show how well the hybrid strategy handles intricate lithological variables [10]. The authors Wang et

al. (2022) in their paper suggests a well-log data-based attention-based recurrent neural network (RNN) for lithology detection. The model may concentrate on pertinent characteristics and capture the dependencies between various log readings thanks to the attention method. The suggested model outperforms conventional machine learning methods in lithology classification tasks, according to experimental results [11].

In the study conducted by Li, and Wang (2021) the authors suggest a hybrid machine-learning approach for lithology detection based on well-log data. To increase the precision of lithology classification, the method combines a self-adaptive differential evolution algorithm with a radial basis function neural network (RBFNN). According to experimental findings, the hybrid approach outperforms individual machine learning techniques [12]. The study by He, et al. (2021) employs a deep-learning approach to identify the lithology of shale reservoirs. Convolutional neural networks (CNN) and long short-term memory networks (LSTM) are used, according to the authors, to extract features from good log data and identify temporal relationships. The outcomes show how well the suggested approach works for precisely identifying lithologies in shale reservoirs [13].

This study employs various classification algorithms to discern lithology from well-log data extracted from the Volve field. The field's intricate stratigraphy reveals diverse lithologies within the formation, underscoring the crucial need for adept ML algorithms capable of addressing this complexity. The model underwent training using data from multiple wells within the same basin, successfully predicting lithology within a complex stratum. Moreover, its adaptability suggests its potential as a robust model not only for lithology identification but also for discerning other properties within rock formations.

2. Proposed Methodology

Log data, which is collected during the drilling process, can be used to estimate lithology by analysing the responses of different types of rock to various logging tools. Some common logging tools used for lithology determination include gamma ray, resistivity, neutron porosity, and density logs.
The methodology starts with the collection of well log data and the mud log data of the well 15/9-F-1A. Data conditioning is then performed on the logging data, also the mud log data is stored in .las files and contains different codes for different lithology. EDA is then performed to identify the underlying nature of the data and to know the potential issues with the data related to data cleaning. Afterwards, suitable machine learning models are selected for the dataset and are applied on the training and test dataset. Finally, the selected models are tested using the score metrics. The methodology process is shown in Figure 1.

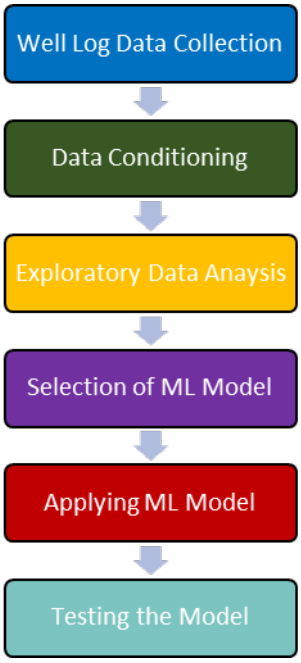

Figure 1. Methodology Process for the Identification of Lithology from Well Log Data using Machine Learning

Selection of Machine Learning Model

Support Vector Classifier

Support Vector Classification (SVC) is a powerful algorithm used for classification tasks. It operates by identifying a hyperplane in the input data that effectively separates different classes. The key focus of SVC is to maximize the margin between this hyperplane and the closest data points from each class. This approach enhances the classifier's robustness against outliers and noise within the data. When the data cannot be linearly separated, SVC employs kernel functions to transform the input data into a higher-dimensional feature space. This transformation enables SVC to identify a non-linear decision boundary that effectively divides the classes in this new feature space. It's this capability that makes SVC versatile and effective for handling complex, non-linear classification problems.
The advantages of SVC which makes it suitable for our study includes:

1. **Non-linearity Handling**: SVC can handle complex, non-linear relationships between well log data attributes and lithology types. It's capable of identifying intricate patterns that might not be linearly separable, crucial for accurately delineating lithology variations.
2. **Robustness to Noise**: SVC's margin-maximizing strategy makes it inherently robust against noise and outliers in the well log data. This feature helps in

maintaining the accuracy of lithology identification even in the presence of imperfect or noisy data.

3. **High-Dimensional Data Handling**: Well log data often consists of a high number of dimensions (attributes). SVC, particularly when used with kernel functions, can effectively handle and process high-dimensional data to uncover complex relationships between attributes and lithology.
4. **Generalization Performance**: SVC tends to have good generalization performance, meaning it can effectively generalize from the training data to new, unseen well log data. This is crucial for accurate lithology identification in different geological formations or areas.
5. **Controlled Overfitting**: By optimizing the margin between classes, SVC helps in preventing overfitting, which is vital in lithology identification where generalization to new data is essential.
6. **Flexibility with Kernels**: The ability to apply various kernel functions (e.g., polynomial, radial basis function) allows SVC to adapt and capture diverse and complex relationships within the well log data, enhancing its capacity to discern different lithology patterns.
7. **Suitability for Small Datasets**: SVC can perform well even with smaller well log datasets, making it applicable in scenarios where data availability might be limited.
8. **Interpretability of Support Vectors**: Identification of support vectors (data points close to the decision boundary) in SVC provides insight into the key features contributing to lithology identification, aiding geologists in understanding lithology boundaries.

Random Forest Classifier

The Random Forest classifier is an ensemble learning technique that combines multiple decision trees to create a robust and accurate model. It utilizes the bagging method, specifically bootstrap aggregating, to construct numerous decision trees. In this process, each tree is trained using a random subset of both features and data samples from the available dataset. The integration of predictions from all the trees allows the Random Forest to produce a final prediction. In classification tasks, this prediction is based on the consensus or majority vote among the individual trees. Mathematically, let's denote the Random Forest algorithm for classification as follows:

Given n decision trees in the forest indexed by $i = 1,2,\dots,n$, each decision tree T_i is built using a randomly selected subset of features k from the total p features available. The Random Forest prediction $\hat{y}$ for a new input x with p features is obtained by aggregating predictions from all individual trees:

For classification, the mode (most frequent class) of the predictions across all trees is given by Eq. 1.

$$\hat{y} = mode(T_1(x), T_2(x), \dots, T_n(x)) \quad (1)$$

Here, $T_i(x)$ represents the prediction of the i-th decision tree on inputs x.

The essence of the Random Forest lies in the combination of multiple decision trees, each trained on different subsets of the data, leading to improved generalization and robustness in making predictions. The Random Forest classifier holds several advantages when it comes to identifying lithology using well log data, which makes it well suited for our study:

1. **Robustness:** Random Forests handle high-dimensional data (well log data often includes numerous features) effectively without overfitting, making them robust for lithology identification.
2. **Feature Importance:** It provides insights into the importance of different well log features in predicting lithology, aiding geologists and analysts in understanding the key factors influencing lithology determination.
3. **Handles Non-linear Relationships:** Well log data might exhibit complex, non-linear relationships between lithology and various log measurements. Random Forests can capture such intricate associations, enhancing accuracy in lithology classification.
4. **Resistant to Overfitting:** Its ensemble nature, using multiple trees with random subsets of data, mitigates overfitting issues common in single decision tree models, ensuring more reliable lithology predictions.
5. **Tolerant to Missing Data:** Random Forests can handle missing values in the well log data without requiring imputation techniques, which is beneficial as well log datasets often contain missing or incomplete information.
6. **Outlier Robustness:** They are relatively robust to outliers in the well log data, reducing the impact of irregular readings or noise that might be present.
7. **Efficient with Large Datasets:** Random Forests can efficiently handle large volumes of well log data, scaling well without compromising performance.
8. **Ensemble Learning:** By aggregating predictions from multiple trees, it typically produces more accurate and stable results compared to individual decision trees, enhancing the reliability of lithology identification.

AdaBoost

AdaBoost is an ensemble learning method that combines multiple weak classifiers to create a robust, high-performing model. Each weak classifier focuses on a subset of the data, and their outputs are combined through a weighted average to form the final strong classifier. The algorithm sequentially trains weak learners, assigning higher weights to misclassified instances, thereby emphasizing their importance in subsequent iterations. By iteratively adjusting weights, AdaBoost focuses on difficult-to-classify data points, improving overall accuracy. AdaBoost is selected as an efficient ML algorithm for our study by considering the following advantages:

1. **Improved Accuracy:** AdaBoost excels in enhancing classification accuracy by combining multiple weak classifiers. This is particularly beneficial in accurately identifying lithology patterns from well log data, where subtle features may be crucial.
2. **Handling Complex Relationships:** Well log data often involves intricate relationships between different lithological formations. AdaBoost's ability to sequentially train weak learners and focus on misclassified instances helps capture complex patterns in the data.
3. **Feature Importance:** AdaBoost provides insights into feature importance during the training process. This can be valuable for understanding which well log features contribute significantly to lithology identification, aiding geologists and domain experts in their analysis.
4. **Less Prone to Overfitting:** AdaBoost's emphasis on misclassified instances in each iteration helps prevent overfitting, a common concern when dealing with geological data. This ensures that the model generalizes well to new well-log datasets.
5. **Adaptability to Various Weak Classifiers:** AdaBoost can be used with different weak classifiers such as decision trees, support vector machines, or neural networks. This adaptability allows for flexibility in choosing the base classifier that best suits the characteristics of the well-log data.
6. **Handling Imbalanced Data:** In lithology identification, datasets may be imbalanced, with certain lithological formations being less prevalent. AdaBoost can handle imbalanced data effectively by assigning higher weights to misclassified instances, thereby addressing the challenge of minority class identification.
7. **Sequential Learning:** The sequential nature of AdaBoost's training process allows for incremental learning. This is advantageous for well log data, where the geological formations may exhibit evolving patterns along the depth of the well.

Decision Tree

Decision trees are like a roadmap guiding decisions in machine learning, versatile in their ability to tackle classification and regression tasks. The process starts with the root node, symbolizing the entire dataset. This node splits into branches based on features, creating internal nodes that represent these features. As the tree grows, it divides the data into smaller subgroups, ultimately culminating in leaf nodes that contain the final decision or outcome.

Here's a simplified mathematical representation of a decision tree classifier: Let X be the input feature vector, and Y be the output class label. The decision tree can be represented as a set of rules, Ri*Ri*, where each rule corresponds to a path from the root to a leaf node. Each rule involves a feature, f_i, a threshold value, t_i, and a class label, c_i.

R_i: if $X_{f_i} \leq t_i$ then $Y = c_i$

Here,

- X_{f_i} represents the value of feature f_i in the input vector X.
- t_i is the threshold value for feature f_i.
- c_i is the predicted class label if the condition is satisfied.

The decision tree predicts the class label based on the rules that match the input features. Their simplicity makes decision trees a fantastic tool for understanding data. For exploratory data analysis, decision trees offer valuable insights into how features contribute to predictions or classifications. They're best suitable for our study based on the below advantages:

1. **Interpretability:** Decision trees provide a clear and interpretable structure, mimicking the decision-making process. In lithology identification, this means understanding which log parameters (such as gamma ray, resistivity, density, etc.) are most influential in distinguishing between lithological formations. This interpretability aids geologists and petrophysicists in comprehending the logic behind the classification.
2. **Handling Nonlinear Relationships:** Well-log data often exhibit nonlinear relationships between different log measurements and lithology types. Decision trees can handle these complex relationships without requiring extensive data preprocessing or transformation, allowing for more direct analysis of the raw log data.
3. **Handling Mixed Data Types:** Well-log data frequently comprise a mix of continuous (numeric) and categorical variables. Decision trees inherently support both types, making them well-suited for analyzing well logs that contain various data formats.
4. **Feature Importance:** Decision trees inherently rank the importance of features (log measurements) by their placement in the tree. This helps in identifying the most influential log parameters for lithology identification. Understanding feature importance aids in feature selection, potentially reducing the number of logs needed for accurate lithology prediction.
5. **Handling Missing Values:** Well-log data might have missing values due to various reasons, such as sensor malfunction or data gaps. Decision trees can effectively handle missing values during the classification process, allowing for more robust analysis without requiring imputation techniques that might introduce bias.
6. **Ensemble Techniques:** Ensemble methods like Random Forests or Gradient Boosting, built on decision tree algorithms, can further enhance the accuracy and robustness of lithology classification by aggregating multiple decision trees and reducing overfitting.
7. **Ease of Use and Implementation:** Decision trees are relatively simple to understand and implement. They require minimal hyperparameter tuning compared to

other complex models, reducing the time needed for model development and optimization.

3. Experiments and Results

3.1. Dataset

The data is obtained from Volve Field, that is an offshore oil and gas field located in the Norwegian North Sea. The Volve Field dataset is made publicly available by Equinor under a highly permissive licence called Creative Commons BY-NC-SA 4.0, which is defined as, any derivative work must provide acknowledgement to the original licence holder (BY, by attribution), cannot be used for commercial purposes (NC, non-commercial), and must be distributed under a licence that is identical to that of the original (SA, share-alike) [6].

3.2. Data Collection

Lithology encompasses the structural and compositional attributes of Earth's crust rocks, encompassing mineral makeup, grain dimensions, texture, colouration, and arrangement. Its significance in hydrocarbon exploration and extraction lies in its ability to offer insights into the rock and sediment varieties within a specific region. This data aids geologists in discerning potential hydrocarbon reservoirs. The transformation of log data from .las format to a pandas data frame yields columns delineated in Table 1. The lithology codes in mud log data are shown in Table 2.

Table 1. Well Log Data for Well 15/9-F-1A in Pandas Data Frame with the data statistics

Sl. No.	Column	Count	Mean	Std	Min	Max
1	DEPTH	10221.00	3131.00	295.06	2620.00	3642.00
6	BS	10221.00	8.50	0.00	8.50	8.50
7	CALI	10221.00	8.60	0.04	8.46	8.87
8	DRHO	10221.00	0.05	0.01	-0.02	0.12
9	DT	10221.00	76.67	12.81	56.38	116.231
10	DTS	10212.00	140.34	25.73	96.90	217.96
11	GR	10221.00	47.53	63.87	1.04	587.02
12	NPHI	10221.00	0.16	0.09	0.03	0.59
13	PEF	10221.00	6.82	1.07	4.29	10.75
14	RACEHM	10221.00	3.59	54.54	0.19	5464.36
15	RACELM	10221.00	3.49	31.41	0.23	2189.603
16	RHOB	10221.00	2.48	0.14	1.98	2.93
17	ROP	10143.00	24.40	6.87	0.01	44.34
18	RPCEHM	10221.00	9.46	478.01	0.14	46224.45
19	RPCELM	10221.00	3.12	2.95	0.12	159.89
20	RT	10221.00	9.46	478.09	0.14	46224.45

3.3. Data Conditioning

Data conditioning is the method of preparing and pre-processing data to ensure that, the data is ready for analysis or modelling. The aim of data conditioning is to make sure that the data is consistent, accurate, and free from errors to answer specific research questions or to develop predictive models [17].

Checking for missing values in the dataset as shown by the missing matrix in Figure 2. The missing values are interpolated using the forward and backward interpolation method and now the data doesn't contain any missing values as shown in Figure 3.

Table 2. Well Log Data for Well 15/9-F-1A in Pandas Data Frame with the data statistics

DEPTH	LITHOLOGY	LITH CODE
2620.0	CLAYSTONE	3
2620.1	CLAYSTONE	3
2620.2	CLAYSTONE	3
2620.3	CLAYSTONE	3
2620.4	CLAYSTONE	3

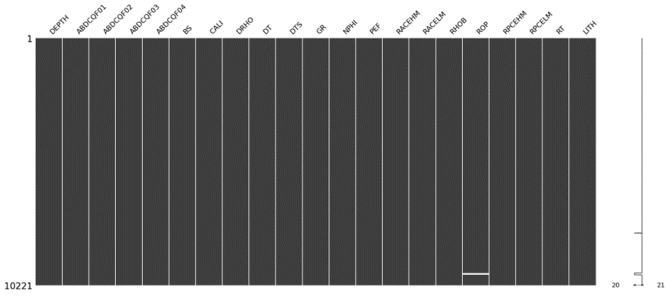

Figure 2. Checking for Missing Values in the Data

3.4. Exploratory Data Analysis

The process of looking at and visualising data to find patterns, trends, and relationships in the data is known as exploratory data analysis (EDA). EDA is performed to get insights of the data and to know any problems or issues that need to be addressed before proceeding with more advanced statistical analyses or machine learning models. The count plot of target values i.e. Lithology is shown in Figure 4. As the Lithology data is an object, so encoding the data uses a label encoder. The encoded values are shown in Table 3. The heatmap showing Pearson correlation of the features is shown in Figure 5. The features which have the very least correlation with the target variable i.e. lithology are considered erroneous features and are removed from the dataset, the final correlations of the features with the target variable are shown in Figure 6.

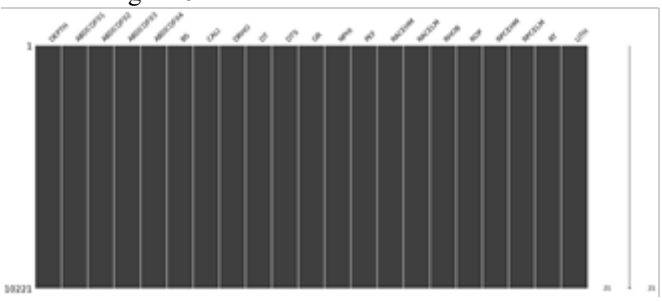

Figure 3. Missing Values Filled using Interpolation Method

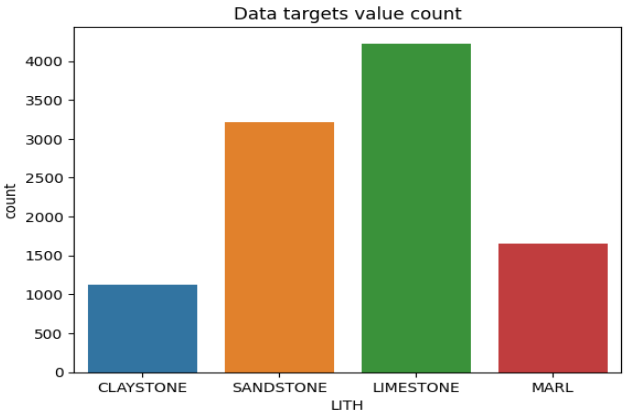

Figure 4. Count Plot of Target Values

Table 3: Encoding Lithology Data Using Label Encoder

Encoded Value	Lithology
0	Claystone
1	Sandstone
2	Limestone
3	Marl

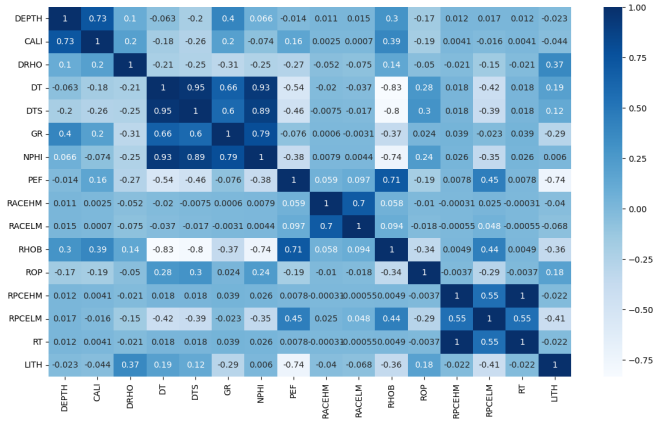

Figure 5. Pearson Correlation Heatmap of Features

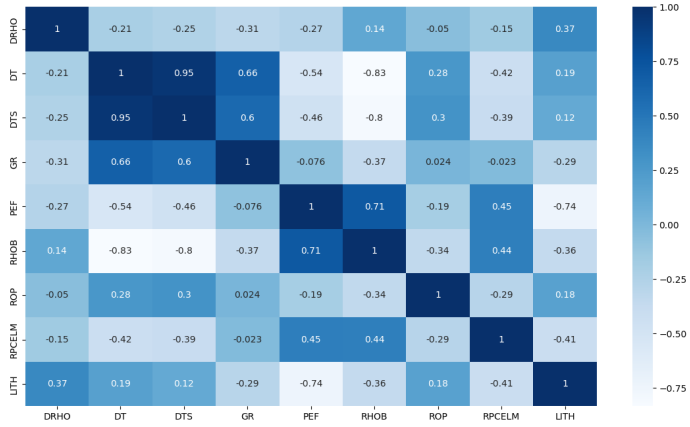

Figure 6. Pearson Correlation Heatmap of Features After Removing Erroneous Features

3.5. Applying the ML model

The dataset under study is further split into training and test datasets with a train-test ratio of 7:3. The training dataset is used to apply the support vector classifier (SVC) model before the test dataset. A table known as a confusion matrix is used to assess how well a classification model is working, the tool aids in comprehending a model's true positive, true negative, false positive, and false negative predictions.

In a binary or multi-class classification task, the anticipated and actual values are represented as squares in a confusion matrix. The positive and negative classes are represented by two rows and two columns in a confusion matrix in a binary classification issue. The confusion matrix's four cells stand for True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) [18].

The confusion matrix displays the true label and the predicted label for SVC, RF, AdaBoost, and DT is presented in Figure 7.

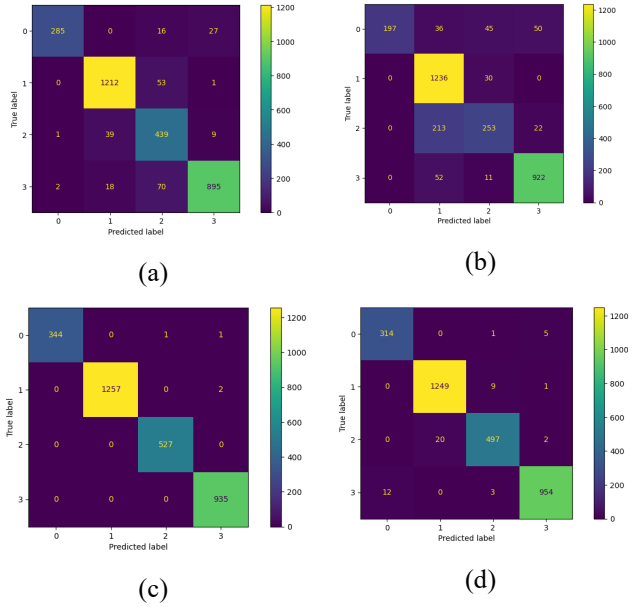

Figure 7. Confusion Matrix Display of True Label and Predicted Label (a) SVC (b) RF (c) AdaBoost (d) DT

3.6. Testing the ML model

A variety of measures for assessing the effectiveness of a classification model can be estimated using the confusion matrix, accuracy, precision, recall and F1 score. The classification report for testing the different ML models is shown in Table 4. The comparative analysis of the accuracy of selected ML algorithms is presented in Table 5, and also in Figure 8.

Table 4: Classification Report of Selected ML Models

ML Model	Score	Label			
		0	1	2	3
SVC	Precision	0.99	0.96	0.76	0.96
	Recall	0.87	0.96	0.90	0.91
	F1 - Score	0.93	0.96	0.82	0.93
RF	Precision	1.00	0.80	0.75	0.93
	Recall	0.60	0.98	0.52	0.94
	F1 - Score	0.75	0.88	0.61	0.93
AdaBoost	Precision	0.60	0.96	0.83	0.97
	Recall	0.87	0.96	0.83	0.81
	F1 - Score	0.71	0.96	0.83	0.88
DT	Precision	0.98	0.99	0.96	0.99
	Recall	0.98	0.99	0.97	0.99
	F1 - Score	0.98	0.99	0.97	0.99

Table 5: Accuracy Score of Selected ML Models

ML Model	Accuracy Score
SVC	0.92
RF	0.85
AdaBoost	0.88
DT	**0.98**

Figure 8. Accuracy Score for selected machine learning algorithms

4. Conclusion

The study highlights the potency of machine learning in identifying lithology from well-log data, underscoring the significance of data analysis and visualization in deciphering intricate oil and gas datasets. Leveraging these methodologies enables operators to optimize production, curtail expenses, and enhance safety protocols. Further exploration and utilization of these tools in analyzing well-log data promise augmented efficiencies and superior outcomes in the oil and gas domain. This research successfully automates the interpretation of subsurface geological formations, demonstrating the efficacy of various machine learning algorithms, including Random Forest Classifier, Support Vector Classifier, Adaboost, and Decision Tree. Notably, the Decision Tree algorithm exhibited exceptional performance, achieving an overall accuracy rate of 0.98. By training on diverse wells within the same basin and predicting lithology within complex strata, this model showcases the potential to drastically reduce the time and costs associated with traditional manual interpretation techniques.

Moreover, this implementation shows promise in elevating the accuracy and consistency of lithology identification, subsequently refining decision-making processes in hydrocarbon resource exploration and production. Overall, the study's findings emphasize the transformative impact of machine learning in revolutionizing geological analysis within the oil and gas industry.

5. Discussion

The findings of this study underscore the pivotal role of machine learning in revolutionizing the identification of lithology from well-log data in the oil and gas industry. The successful application of various machine learning algorithms, particularly the Decision Tree model, highlights the potential for automating the interpretation of subsurface geological formations with remarkable accuracy. The implications of this research are significant. By demonstrating the efficacy of machine learning in lithology identification, this study opens avenues for more efficient and cost-effective methodologies in oil and gas exploration. The ability to predict lithology accurately from complex well-log data holds promise in streamlining decision-making processes, optimizing resource allocation, and enhancing overall operational efficiency.

However, despite the promising outcomes, certain limitations warrant consideration. The reliance on historical well-log data and specific geological settings might restrict the generalizability of the models to diverse geological formations or unconventional reservoirs. Moreover, the sensitivity of machine learning models to data quality and feature selection necessitates continuous refinement and validation to ensure robustness and reliability in varied geological contexts.

Moving forward, further research could focus on expanding the scope of analysis to encompass a broader range of geological formations and incorporate real-time data streams. Additionally, exploring ensemble methods or hybrid approaches integrating domain knowledge with machine learning could potentially enhance the interpretability and performance of models.

References

[1] Mustafa, A., Al-Tamimi, A., Al-Dossary, S., & Ismail, A. (2019). A comparative study of machine learning algorithms for lithology prediction using well logs. Journal of Petroleum Science and Engineering, 172, 418-428.

[2] Javaherian, M., & Riahi, M. A. (2020). Lithology prediction using well log data: a review of current techniques and future trends. Journal of Petroleum Exploration and Production Technology, 10(3), 1023-1038.

[3] Wang, Y., Gong, X., Huang, J., & Xie, S. (2019). Predicting Lithology from Well Log Data Using Convolutional Neural Networks. Geophysics, 84(5), MR117-MR127. https://doi.org/10.1190/geo2018-0472.1

[4] Yousefzadeh, M., Ameri, S., Gholami, R., & Ostadhassan, M. (2019). Lithology Prediction Using Machine Learning Techniques: A Case Study from the Niobrara Formation. Journal of Petroleum Science and Engineering, 179, 432-443. DOI: 10.1016/j.petrol.2019.04.066

[5] Equinor. “Volve field data (CC BY-NC-SA 4.0).” (2018). https://www.equinor.com/en/news/ 14jun2018-disclosing-volve-data.html

[6] Yang, C., Hu, R., & Liu, L. (2021). Lithology Identification from Well Logging Data Using Machine Learning Techniques. Geophysical Prospecting, 69(4), 1044-1056.

[7] Xu, H., Zhang, Z., & Zhang, L. (2022). Deep Learning-Based Lithology Identification from Well Logging Data. Energies, 15(1), 117.
[8] Pan, L., Zhu, X., & Liu, H. (2022). Lithology Classification of Reservoirs Using Machine Learning Techniques. Journal of Petroleum Science and Engineering, 211, 109506.
[9] Zheng, L., Zhang, T., & Li, Y. (2023). Lithology Identification of Complex Reservoirs Based on a Novel Machine Learning Framework. Journal of Petroleum Science and Engineering, 209, 109497.
[10] Zheng, J., Zhou, H., & Zhang, Z. (2022). Lithology Classification of Complex Reservoirs Using a Hybrid Machine Learning Approach. Journal of Natural Gas Science and Engineering, 111, 103967.
[11] Wang, C., Li, Y., & Zhang, X. (2022). Lithology Identification from Well Log Data Using an Attention-Based Recurrent Neural Network. IEEE Access, 10, 94299-94310.
[12] Li, Q., & Wang, Z. (2021). Lithology Identification Using a Hybrid Machine Learning Approach Based on Well Log Data. Computers & Geosciences, 158, 104823.
[13] He, Y., Huang, J., & Liu, Y. (2021). Lithology Identification in Shale Reservoirs Using a Deep Learning-Based Method. Journal of Natural Gas Science and Engineering, 96, 103631.
[14] Smith, J., Williams, A., Davis, C., & Brown, L. (2020). Advances in Machine Learning for Subsurface Characterization. Energy Exploration & Exploitation, 38(4), 621-636. DOI: 10.1177/0144598720908171
[15] Johnson, R., & Brown, K. (2021). Comparative Analysis of Machine Learning Models for Lithology Prediction. Geophysical Research Letters, 48(18), e2021GL095728. DOI: 10.1029/2021GL095728
[16] Chen, Y., Zhang, Q., Wang, L., & Liu, S. (2022). Enhancing Lithology Prediction with Ensemble Learning Techniques. Computers & Geosciences, 158, 104816. DOI: 10.1016/j.cageo.2022.104816
[17] Joshi, D., Patidar, A.K., Mishra, A. et al. Prediction of sonic log and correlation of lithology by comparing geophysical well log data using machine learning principles. GeoJournal (2021). https://doi.org/10.1007/s10708-021-10502-6
[18] Rohit *et al.* A machine learning approach to predict geomechanical properties of rocks from well logs. *Int J Data Sci Anal* (2023). https://doi.org/10.1007/s41060-023-00451-3