JGE (Jurnal Geofisika Eksplorasi)
Vol. 12 No. 01, Maret 2026 (58-72) https://doi.org/10.23960/jge.v12i1.511

COMPARATIVE STUDY OF K-NEAREST NEIGHBORS (KNN) AND ARTIFICIAL NEURAL NETWORK (ANN) FOR LITHOLOGY CLASSIFICATION

STUDI KOMPARATIF K-NEAREST NEIGHBORS (KNN) DAN ARTIFICIAL NEURAL NETWORK (ANN) UNTUK KLASIFIKASI LITOLOGI

Ruth Agnesia Sasono S[1*], Rahma Ramadhani Herliana[2], M. Fadhil Hawari[3], Rizky Yustisia Sari[4], Stevy Canny Louhenapessy[5]

[1,2,5]Oil and Gas Engineering/Institut Teknologi Sumatera; Terusan Ryacudu St., Desa Way Hui, Kecamatan Jati Agung, Lampung Selatan 35365; Fax: (0721) 8030189
[3]Informatics Engineering/ Institut Teknologi Sumatera; Terusan Ryacudu St., Desa Way Hui, Kecamatan Jati Agung, Lampung Selatan 35365; Fax: (0721) 8030189
[4]Instrumentation and Automation Engineering/ Institut Teknologi Sumatera; Terusan Ryacudu St., Desa Way Hui, Kecamatan Jati Agung, Lampung Selatan 35365; Fax: (0721) 8030189

Received: 2026, February 23rd
Accepted: 2026, May 21st

Keywords:
Artificial Neural Network;
K-Nearest Neighbors;
Lithology Classification;
Machine Learning;
Well Log.

Correspondent Email:
ruthagnesia19@gmail.com

How to cite this article:
Sasono S, R.A., Herliana, R.R., Hawari, M.F., Sari, R.Y., & Louhenapessy, S.C. (2026). Comparative Study of K-Nearest Neighbors (KNN) and Artificial Neural Network (ANN) For Lithology Classification. *JGE* (*Jurnal*

Abstract. This study applies a machine learning approach to classify lithology using well log data from 14 wells in Ford County, Kansas, United States, to address the limitations of conventional interpretation, which is time-consuming and subjective due to overlapping log responses. Reference lithology labels were generated using predefined well-log interpretation criteria and grouped into four classes: sandstone, limestone, shale/clay, and coal. Two supervised learning algorithms, K-Nearest Neighbors (KNN) and Artificial Neural Networks (ANN), were evaluated and compared. The preprocessing stages included data cleaning by removing null values and inconsistencies, Z-score normalization, class balancing using SMOTE on the training data to prevent data leakage, and feature selection based on Pearson correlation. Model performance was evaluated using Classification Accuracy (CA), Area Under the Curve (AUC), Logarithmic Loss (Log Loss), and 5-fold cross-validation. The results indicate that ANN consistently outperformed KNN in lithology classification. ANN achieved classification accuracies above 95%, AUC values approaching 1.00, and low Log Loss, whereas KNN achieved testing accuracies of approximately 75-80% but exhibited lower cross-validation performance, indicating reduced robustness in intervals characterized by overlapping lithological responses. The optimal ANN architecture consisted of three hidden layers with 100-100-100 neurons and 100 training iterations. Visual evaluation of four test wells showed good agreement between the predicted and reference lithology distributions. These findings suggest that machine learning, combined with appropriate preprocessing techniques, can support lithology classification

Geofisika Eksplorasi), 12(01), 58-72.

from well log data. Among the evaluated models, ANN demonstrated superior capability in capturing nonlinear relationships between well log responses and lithological variations within the study area.

Abstrak. *Studi ini menerapkan pendekatan pembelajaran mesin untuk mengklasifikasikan litologi menggunakan data log sumur dari 14 sumur di Ford County, Kansas, Amerika Serikat, untuk mengatasi keterbatasan interpretasi konvensional, yang memakan waktu dan subjektif karena tumpang tindihnya respons log. Label litologi referensi dihasilkan menggunakan kriteria interpretasi log sumur yang telah ditentukan sebelumnya dan dikelompokkan menjadi empat kelas: batupasir, batugamping, serpihan/lempung, dan batubara. Dua algoritma pembelajaran terawasi, K-Nearest Neighbors (KNN) dan Jaringan Saraf Buatan (ANN), dievaluasi dan dibandingkan. Tahap pra-pemrosesan meliputi pembersihan data dengan menghilangkan nilai nol dan inkonsistensi, normalisasi skor Z, penyeimbangan kelas menggunakan SMOTE pada data pelatihan untuk mencegah kebocoran data, dan pemilihan fitur berdasarkan korelasi Pearson. Kinerja model dievaluasi menggunakan Akurasi Klasifikasi (CA), Area di Bawah Kurva (AUC), Kerugian Logaritmik (Log Loss), dan validasi silang 5-fold. Hasil menunjukkan bahwa ANN secara konsisten mengungguli KNN dalam klasifikasi litologi. ANN mencapai akurasi klasifikasi di atas 95%, nilai AUC mendekati 1,00, dan Log Loss yang rendah, sedangkan KNN mencapai akurasi pengujian sekitar 75-80% tetapi menunjukkan kinerja validasi silang yang lebih rendah, yang mengindikasikan berkurangnya ketahanan pada interval yang ditandai dengan tumpang tindih respons litologi. Arsitektur ANN yang optimal terdiri dari tiga lapisan tersembunyi dengan 100-100-100 neuron dan 100 iterasi pelatihan. Evaluasi visual dari empat sumur uji menunjukkan kesesuaian yang baik antara distribusi litologi yang diprediksi dan referensi. Temuan ini menunjukkan bahwa pembelajaran mesin, dikombinasikan dengan teknik pra-pemrosesan yang tepat, dapat mendukung klasifikasi litologi dari data log sumur. Di antara model yang dievaluasi, ANN menunjukkan kemampuan yang unggul dalam menangkap hubungan nonlinier antara respons log sumur dan variasi litologi di dalam area penelitian.*

1. INTRODUCTION

Reservoir characterization integrates geological and petrophysical analyses to quantify key rock properties such as porosity, permeability, fluid saturation, and lithology, which ultimately control reservoir performance (Esiri et al., 2024). Lithology identification is commonly performed through the interpretation of well log data, including gamma ray (GR), resistivity, density (RHOB), neutron (NPHI), sonic (DT), and spontaneous potential logs (SP) (Mohamed et al., 2019). However, conventional manual interpretation remains subjective and time-consuming (Sastra & Rohmana, 2024). The rapid development of machine learning (ML) offers an opportunity to automate subsurface interpretation by extracting numerical patterns from well log data. ML techniques have been successfully applied to predict various geophysical and petrophysical properties, including shear-wave velocity, porosity, and total organic carbon (TOC) (Rohmana et al., 2024; Wardhana et al., 2022; Wardhana & Pakpahan, 2021). For lithology classification, K-Nearest Neighbors (KNN) and Artificial Neural Network (ANN) are among the most widely used methods. KNN classifies samples based on distance similarity in feature space, whereas ANN is capable of capturing more

complex nonlinear relationships between log responses and lithological variation (Komara et al., 2024; Maulana et al., 2024). However, previous studies have generally focused on the application of individual algorithms or employed different datasets and evaluation procedures, making direct performance comparisons difficult. Therefore, this study conducts a systematic comparison of KNN and ANN using the same well log dataset, preprocessing workflow, and evaluation metrics to determine the most effective approach for lithology classification.

Previous comparative studies indicate that ML performance depends strongly on dataset characteristics, lithology distribution, and hyperparameter selection (Galupino & Dungca, 2022; Khatti & Grover, 2022; Marembayev et al., 2021). However, many studies remain limited in well coverage and do not adequately address class imbalance, which may reduce the detection of minor lithologies such as coal beneath dominant lithologies such as limestone or shale. Although KNN and ANN have been validated in various geoscience applications, including facies estimation, coal classification, shear-wave velocity prediction, shale gas modeling, and reservoir analysis (Adiwiguna et al., 2022; Prabowo et al., 2023; Putra et al., 2023; Septyandy & Subroto, 2023; Sudrazat et al., 2020), fewer studies have explicitly compared their performance under heterogeneous and imbalanced well log datasets.

Accordingly, this study serves as a benchmarking analysis of K-Nearest Neighbors (KNN) and Artificial Neural Networks (ANN) for lithology classification in Ford County, Kansas. KNN is used to assess the limitations of distance-based classification, whereas ANN is employed to evaluate the capability of a nonlinear model in capturing complex log–lithology relationships. The main contribution of this study lies in the integration of data balancing on a dataset from 14 wells to reduce bias toward minority lithologies and to provide a more systematic comparison between both algorithms. In addition, this study presents a direct comparison of KNN and ANN using data from 14 wells, a consistent preprocessing workflow, SMOTE-based class balancing, and clearly defined training, testing, and cross-validation procedures.

2. LITERATURE REVIEW

2.1. Lithological Interpretation Based on Well Logs

Lithological interpretation in subsurface petrophysics relies on the integration of well log parameters such as gamma ray, resistivity, density, neutron, sonic, and spontaneous potential. These logs provide diagnostic responses to lithological variations, although overlapping signatures between lithologies may complicate manual interpretation, particularly in heterogeneous formations (Mohamed et al., 2019; Nabilah et al., 2023; Sephiana et al., 2023). As a result, conventional interpretation remains vulnerable to subjectivity and reduced consistency when log responses are ambiguous (Sastra & Rohmana, 2024).

Table 1 summarizes the characteristic well-log response ranges of the main lithology classes considered in this study. These ranges were compiled from established petrophysical interpretation references (Asquith et al., 2004) and adapted to the characteristics of the study dataset, and were subsequently used as interpretation criteria to generate the reference lithology labels for model training and evaluation. Because log responses may overlap among lithologies, particularly in transitional intervals and heterogeneous formations, the labeling process relied on the integrated interpretation of multiple log parameters rather than strict thresholds from a single log.

Table 1. Lithological Classification of Rocks Based on Well Logging.

Lithology	Gamma Ray (API)	Density (g/cm^3)	Neutron (%)	Sonic (μs/ft)	Resistivity (Ω-m)	SP (mV)
Coal	< 20	< 2.0	40 – 60%	–	10 – 100	(–5) – 5

Shale/Clay	60 – 150	2.6	30 – 60%	62.5 – 167	1 – 20	0 – 5
Sandstone	15 – 50	2.65	5 – 25%	51.3 – 55.5	1 – 50	–50 – (–5)
Limestone	10 – 40	2.71	5 – 20%	47.5	80 – >100	0 – (–5)

2.2. Machine Learning for Lithology Classification

Machine learning has been widely applied in geoscience to predict subsurface properties such as shear-wave velocity, porosity, and total organic carbon, demonstrating its ability to model nonlinear relationships in petrophysical data (Rohmana et al., 2024; Wardhana & Pakpahan, 2021). In lithology classification, previous studies have shown that model performance depends strongly on dataset characteristics, class distribution, and hyperparameter selection. KNN has been reported to improve performance when weighted distance metrics are used in heterogeneous formations, but its distance-based nature may limit robustness in noisy or high-dimensional datasets (Wang et al., 2018). In contrast, ANN has been shown to perform well in more complex geological classification tasks because of its stronger nonlinear generalization capability.

Nevertheless, many previous studies remain limited by small datasets and insufficient treatment of class imbalance, which may reduce the detection of minority lithologies. Moreover, existing studies often present ML methods separately rather than through direct benchmarking under the same geological and preprocessing conditions. Therefore, this study compares KNN and ANN within a unified framework and incorporates class balancing to obtain a more rigorous evaluation of both models in heterogeneous lithological settings.

2.2.1. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a supervised learning algorithm that performs classification based on the distance proximity among data points in feature space. Its basic principle is that data with similar well log characteristics will be located near each other, so the lithology class of a new sample can be determined based on the majority class of its nearest neighbors. The distance between data points is calculated using the Euclidean Distance (Hamid & Mousavi, 2024).

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

where,

$d(x,y)$ = Euclidean distance between test data x and training data y

x_i = i-th feature/attribute of the test data

y_i = i-th feature/attribute of the training data

n = Number of features/attributes

Because well log parameters have different scales, Z-score standardization is applied to prevent feature dominance. The main weakness of KNN is its inability to handle overlapping log responses, which can cause high Log Loss and low confidence at complex lithological boundaries. Therefore, in addition to Classification Accuracy (CA), Log Loss analysis is needed to measure probabilistic accuracy in heterogeneous geological intervals.

The KNN process is illustrated in **Figure 1.** The blue points (shale) and orange points (sandstone) represent the training data, while the question mark (?) represents a new log sample. The K circle indicates the K nearest neighbors (K = 3). Classification is based on the majority vote, where 2 blue points and 1 orange point can classify the new sample as shale. The more similar the log responses are, the smaller the Euclidean distance, and the higher the likelihood that the lithology is the same.

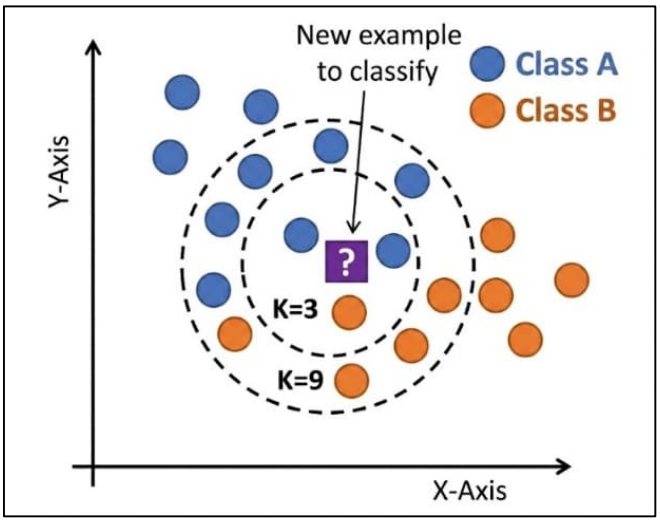

Figure 1. Illustration of the KNN classification process, where the blue and orange represent two different lithology classes, and the purple square represents a new data point classified based on the majority of its nearest neighbors.

2.2.2. Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) is a machine learning algorithm that is effective for modeling nonlinear relationships between input parameters and target classes. In this study, an ANN was used to identify complex relationships between well log parameters and rock lithology, particularly in heterogeneous and multidimensional data (Qamar & Zardari, 2023). The ANN architecture used in this study, consisting of an input layer, multiple hidden layers, and an output layer, is illustrated in **Figure 2.**

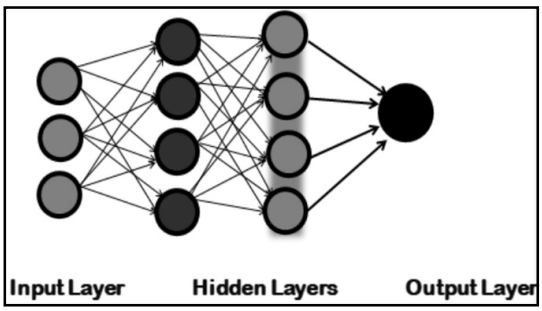

Figure 2. Artificial Neural Network (ANN) architecture consisting of an input layer, hidden layers, and an output layer. The input layer receives well log parameters, the hidden layers process the information through interconnected neurons, and the output layer produces the lithology classification result (Qamar & Zardari, 2023).

The model was trained using a supervised learning approach, with well log parameters as inputs and lithology as the target output. Training was performed using the backpropagation algorithm to adjust network weights and minimize classification error. Before training, the data were normalized using a Z-score to make the feature distribution more uniform. Model performance was evaluated using AUC and Log Loss. For the multiclass case, AUC was calculated using the One-vs-Rest (OvR) method to assess the model's ability to distinguish each lithology class. In contrast, Log Loss was used to measure the confidence of the predictions. Overall, the combination of these two metrics provides a comprehensive picture of the ANN classification performance.

3. RESEARCH METHODS

3.1. Data and Geological Conditions

This study used a well log dataset obtained from the Kansas Geological Survey (KGS) database (KGS, 2004), consisting of 14 wells located in Ford County, Kansas, United States (**Figure 3**). The selected wells are distributed across several sections within Township 29 South, Range 22 West (T29S-R22W), covering

Sections 2, 5, 6, 7, 8, and 33. The wells were selected based on the availability and completeness of the required logging parameters, including gamma ray, resistivity, neutron, density, sonic, and spontaneous potential. Each well covers different depth intervals depending on data availability and logging coverage. After data cleaning and preprocessing, the final dataset comprised 60,942 valid depth samples. These well logs were used to classify four lithology classes: sandstone, limestone, shale/clay, and coal.

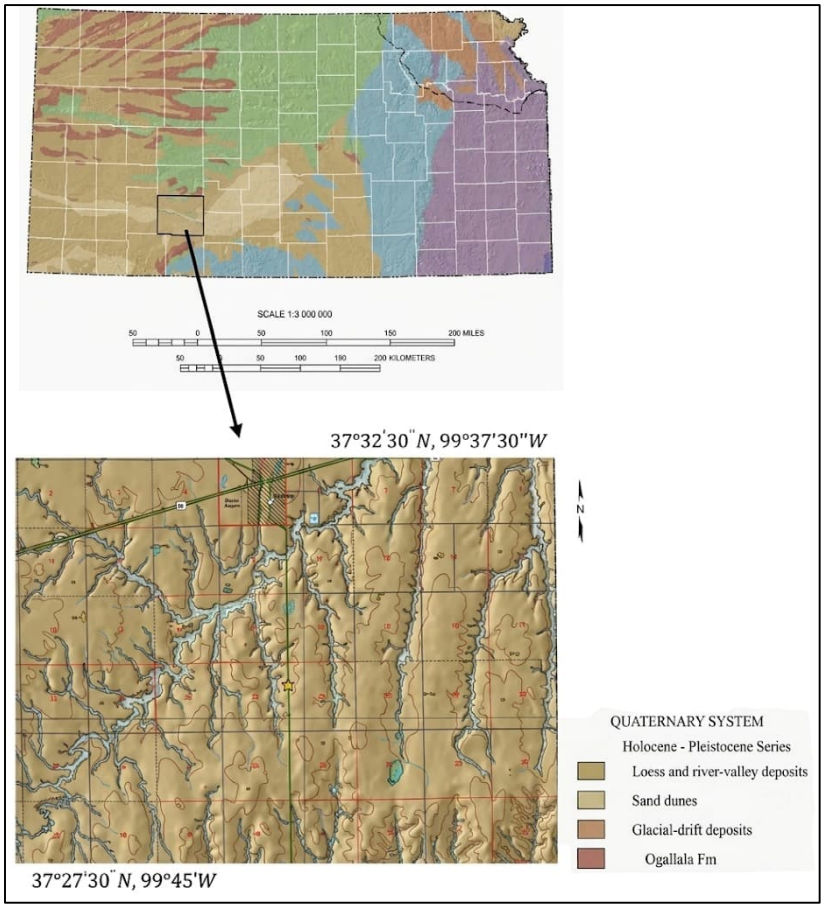

Figure 3. Map of Ford County, Kansas (Johnson & Woodburn, 2009).

3.2. Data Pre-processing

The first step in data processing involved data cleaning to remove null values and error values, such as -999.25, which are commonly found in LAS files. The cleaned data were then converted into CSV format to facilitate processing and analysis using Python. The second step was data labeling, in which a reference lithology class was assigned to each depth interval based on predefined well-log interpretation criteria summarized in **Table 1**. These labels are therefore considered log-derived labels rather than direct observations from core or cutting descriptions. Because the same well log parameters were subsequently used as input features for machine learning, there is a potential risk of circularity or pseudo-labeling. However, the objective of this study is to evaluate and compare the capability of KNN and ANN in reproducing a consistent log-based lithology interpretation. Future studies incorporating core, cutting, or independently interpreted lithology data would provide a more rigorous validation framework.

The labelled dataset was then divided at the well level to avoid information leakage between

the training and testing stages. A total of 10 wells were assigned to the training set and 4 wells were reserved for independent testing. The class distribution analysis indicated an imbalance, as shown in **Figure 4**, shale/clay constituted 72.2% of the original training data, followed by limestone (15.2%), sandstone (8.5%), and coal (4.1%). To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied only to the training data. SMOTE was not applied to the testing data to prevent data leakage and biased performance evaluation. In this study, the number of nearest neighbors for SMOTE generation was set to 5. After oversampling, the training data became more balanced, allowing the model to learn lithology patterns more representatively, while the testing data retained the original class distribution to ensure an unbiased evaluation of the model's generalization capability under realistic geological conditions.

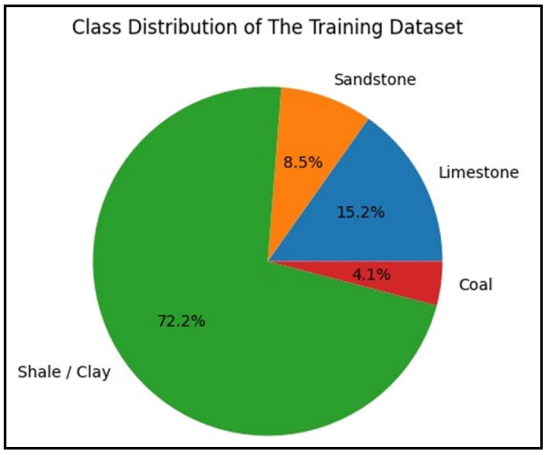

Figure 4. Original data distribution in the training data before SMOTE oversampling.

3.3. Feature Selection

Feature selection was conducted using Pearson correlation analysis to examine linear relationships among the well log parameters before applying ANN and KNN models. The correlation results were visualized in the form of a heatmap, as shown in **Figure 5**, where red indicates positive correlation, blue indicates negative correlation, and the color intensity reflects the strength of the correlation coefficient. A correlation threshold of $|r| > 0.9$ was used to identify highly redundant features. From the visualization in **Figure 5,** it can be observed that the SPOR variable exhibits a strong positive correlation with the DT variable, suggesting that both variables carry similar information, as an increase or decrease in one variable tends to be followed by a corresponding change in the other. Based on this threshold, SPOR was excluded from the model inputs to avoid redundancy, while DT was retained as a representative feature. It should be noted that SPOR was included in the heatmap visualization for completeness, but was not used as an input variable in either the ANN or KNN models.

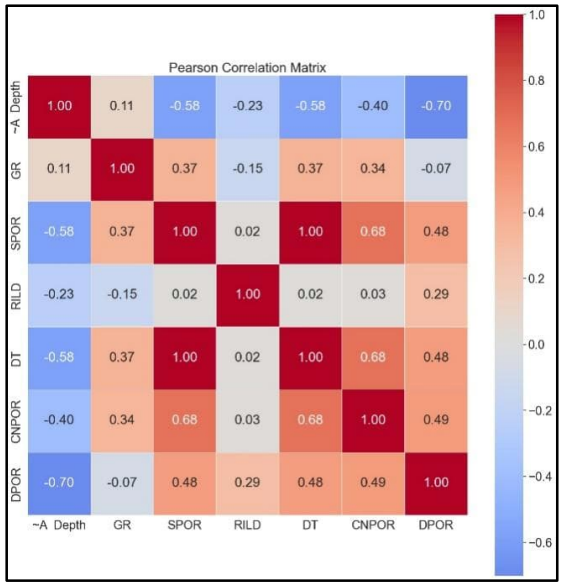

Figure 5. Pearson correlation heatmap of the input parameters. Red indicates positive correlation and blue indicates negative correlation, with color intensity representing the strength of the correlation coefficient (r). Depth was excluded from the modelling as it represents only vertical position, and SPOR was removed due to its strong positive correlation with DT and CNPOR, indicating potential data redundancy.

3.4. Cross-Validation Model

5-fold stratified cross-validation was used to evaluate the generalization ability of the model and reduce dependence on a specific data split. In this approach, the dataset was divided into five subsets while maintaining the proportional class distribution in each fold, with each subset randomly assigned, and each

subset used once as validation data while the remaining subsets were used for training. This method was applied to both KNN and ANN to assess performance stability, support more objective parameter selection, and minimize the risk of overfitting. It is acknowledged that since the data were obtained from 14 wells, a well-based splitting strategy would ideally be more appropriate to prevent potential data leakage between samples from the same well. However, due to software limitations, random stratified sampling was applied.

3.5. Model Design and Experimental Setup

The training scheme in this study was implemented using Orange Data Mining, an open-source data mining software built on the scikit-learn library that supports both KNN and ANN methods. The ANN model was constructed with a three hidden-layer architecture, where the number of neurons per layer varied between 50 and 100. Each architectural combination was trained under three different maximum iteration settings: 10, 100, and 1000. The ANN hyperparameters used during model training are summarized in **Table 2**. For the KNN model, training was conducted using K values ranging from 1 to 13, with Euclidean distance as the distance metric and uniform weighting applied to all neighboring points.

Table 2. Fixed hyperparameter settings used for ANN model training.

Hyperparameter	Value
Solver	Adam
Activation Function	ReLu
Regularization	0.0001
Learning Rate	Default Software
Batch Size	Default Software
Random State	Default Software (Replicable Training)

4. RESULTS AND DISCUSSION

4.1. K-Nearest Neighbors (KNN) Results

The K-Nearest Neighbors (KNN) model was evaluated using different K values ranging from 1 to 13. Model performance was assessed using Classification Accuracy (CA), Log Loss, and 5-fold cross-validation. The results indicate that increasing the value of K generally reduced Log Loss and improved model stability by producing smoother probabilistic predictions and reducing sensitivity to local noise. The independent testing results achieved classification accuracies of approximately 75–80%, with the highest value of 80.1% obtained at K = 1. However, the comparatively lower cross-validation performance suggests that the KNN model is sensitive to variations in the training data and may have limited robustness when applied to heterogeneous lithological datasets.

Table 2. KNN performance evaluation based on Accuracy, Log Loss, and 5-fold cross-validation.

K	Accuracy (CA)		Log Loss		Time		Cross Validation			
	Train	Test	Train	Test	Train	Test	Accuracy (CA)	Log Loss	Time (s)	
									Train	Test
1	1	0.801	0	7.183	0.151	1,118	0.979	0.758	0.552	4.495
2	0.986	0.782	0.012	6.273	0.150	1.187	0.967	0.603	0.517	4.175
3	0.987	0.789	0.024	5.762	0.148	1.210	0.967	0.510	0.738	6.333
4	0.977	0.775	0.035	5.345	0.152	1.284	0.958	0.456	0.561	4.728
5	0.976	0.779	0.045	4.981	0.169	1.339	0.957	0.426	0.668	7.097
6	0.968	0.769	0.054	4.699	0.150	1.239	0.950	0.401	0.572	4.524
7	0.968	0.772	0.064	4.465	0.146	1.234	0.949	0.381	0.687	7.074
8	0.961	0.762	0.072	4.254	0.149	1.256	0.943	0.357	0.537	4.659

9	0.960	0.766	0.081	4.061	0.154	1.267	0.942	0.348	0.533	6.028
10	0.955	0.759	0.088	3.897	0.150	1.279	0.937	0.330	0.559	4.942
11	0.954	0.761	0.096	3.789	0.147	1.314	0.935	0.326	0.564	6.684
12	0.949	0.754	0.103	3.651	0.153	1.296	0.931	0.320	0.580	5.151
13	0.948	0.757	0.109	3.521	0.147	1.303	0.930	0.316	0.645	6.666

This performance drop is mainly related to the overlap of well log responses between lithological classes, especially sandstone and shale/clay. In such intervals, the Euclidean-distance-based nature of KNN becomes less effective because adjacent samples may not represent the same lithology even if they are numerically close in feature space. As a result, KNN tends to misclassify transition zones and cannot fully capture the nonlinear structure of subsurface data. Among the tested configurations, K = 13 produced the lowest test Log Loss and the most stable cross-validation performance, However, despite achieving moderate to high classification accuracy on the independent testing dataset, the relatively lower cross-validation results indicate that the KNN model is less robust and has limited generalization capability compared with ANN for lithology prediction.

4.2. Artificial Neural Network (ANN) Results

The ANN model was tested using several network architectures with varying hidden-layer sizes and maximum iteration values (**Table 3**). Model performance was evaluated using AUC, Log Loss, and 5-fold cross-validation. The results indicate that ANN consistently achieved AUC values close to 1.00 on both training and testing data, demonstrating excellent class discrimination. Compared with KNN, ANN also produced lower and more stable Log Loss values, particularly when the number of training iterations increased from 10 to 100.

The selected ANN architecture (scheme 11) was the three hidden-layer configuration (100, 100, 100) with 100 iterations. This configuration was chosen because it provided a balanced trade-off among classification performance, model stability, and computational efficiency. Although several alternative configurations achieved comparable performance in terms of AUC, accuracy, and Log Loss, increasing the number of iterations to 1000 did not produce substantial improvements while significantly increasing computational cost. These results indicate that the model had reached a stable learning state at approximately 100 iterations. The strong performance of the ANN model can be attributed to its ability to capture nonlinear relationships between well log parameters and lithology classes, making it well suited for geologically complex and heterogeneous datasets compared with KNN.

Table 3. ANN application schemes with various parameter selections.

Scheme	Parameter		AUC		Log Loss		Time (s)		Cross Validation			
											Time (s)	
	Hidden Layer Size	Max Iteration	Train	Test	Train	Test	Train	Test	AUC	Log Loss	Train	Test
1	(50,50,50)	10	0.996	0.999	0.092	0.055	4.89	4.736	0.996	0.097	17.22	0.25
2	(50,50,50)	100	0.998	0.998	0.063	0.066	49.10	0.059	0.997	0.071	219.38	0.31
3	(50,50,50)	1000	0.998	0.998	0.062	0.064	51.45	0.063	0.997	0.068	246.09	0.29
4	(50,100,50)	10	0.997	0.998	0.086	0.111	8.36	0.079	0.996	0.094	34.91	0.34
5	(50,100,50)	100	0.998	0.997	0.063	0.077	65.88	0.087	0.997	0.068	359.64	0.34
6	(50,100,50)	1000	0.998	0.997	0.053	0.075	157.76	0.073	0.997	0.066	562.44	0.36
7	(100,100,50)	10	0.997	0.999	0.085	0.060	9.91	0.092	0.997	0.091	45.86	0.43

8	(100,100,50)	100	0.999	0.998	0.062	0.082	116.71	0.101	0.993	0.067	452.62	0.47
9	(100,100,50)	1000	0.999	0.998	0.060	0.074	109.39	0.109	0.993	0.066	601.64	0.43
10	(100,100,100)	10	0.999	0.998	0.125	0.067	11.40	0.105	0.995	0.091	44.69	0.465
11	(100,100,100)	100	0.999	0.997	0.057	0.080	126.5	0.105	0.993	0.067	537.93	0.522
12	(100,100,100)	1000	0.998	0.998	0.057	0.072	10.35	0.107	0.997	0.063	1110.2	10.29

4.3. Comparison of Lithology Prediction Results

Figure 6 until **Figure 9** show a visual comparison between the predicted lithology and the reference lithology for the four independent testing wells. These four wells were excluded from the training stage and served as hold-out wells to evaluate the model's generalization capability on unseen data. The results indicate that ANN closely reproduces the overall lithological trends with depth across all testing wells. The predicted intervals generated by ANN generally align well with the reference lithology, demonstrating the model's ability to capture both the main lithological succession and several thin-bed variations. In contrast, KNN produces more fragmented predictions and less consistent layer boundaries, particularly within transitional intervals where well log responses overlap.

This difference is especially evident in sandstone–shale/clay transitions, where similar log signatures make classification more difficult. ANN is able to represent these nonlinear and heterogeneous patterns more effectively, whereas KNN tends to misclassify intervals with overlapping responses because its decision rule is based primarily on local distance in feature space. Coal intervals are comparatively easier for both models to identify because they exhibit more distinctive log responses, especially low density and characteristic gamma ray behavior. The confusion matrix in Figure 10 supports the visual interpretation from Figures 6–9. ANN achieved overall accuracy above 95% with stable performance across classes, while KNN reached only about 60% accuracy and showed a higher rate of misclassification in ambiguous intervals. These results indicate that ANN provides a more reliable lithology prediction across the test wells and is better suited for heterogeneous subsurface datasets.

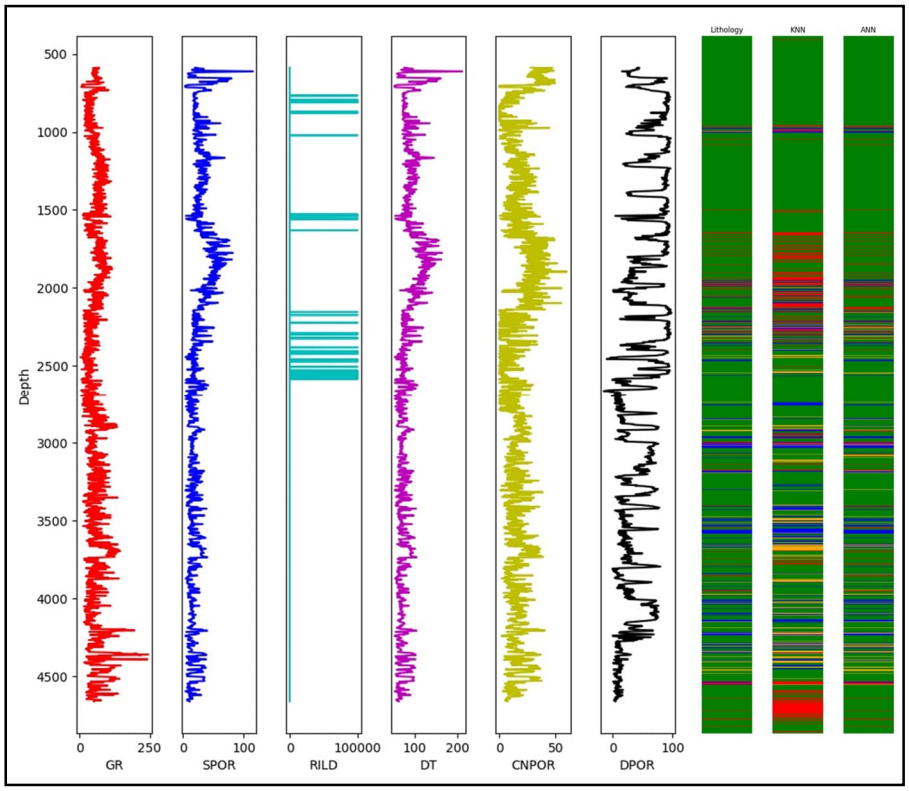

Figure 6. Prediction results of the KNN and ANN methods at the Close 3-2 well.

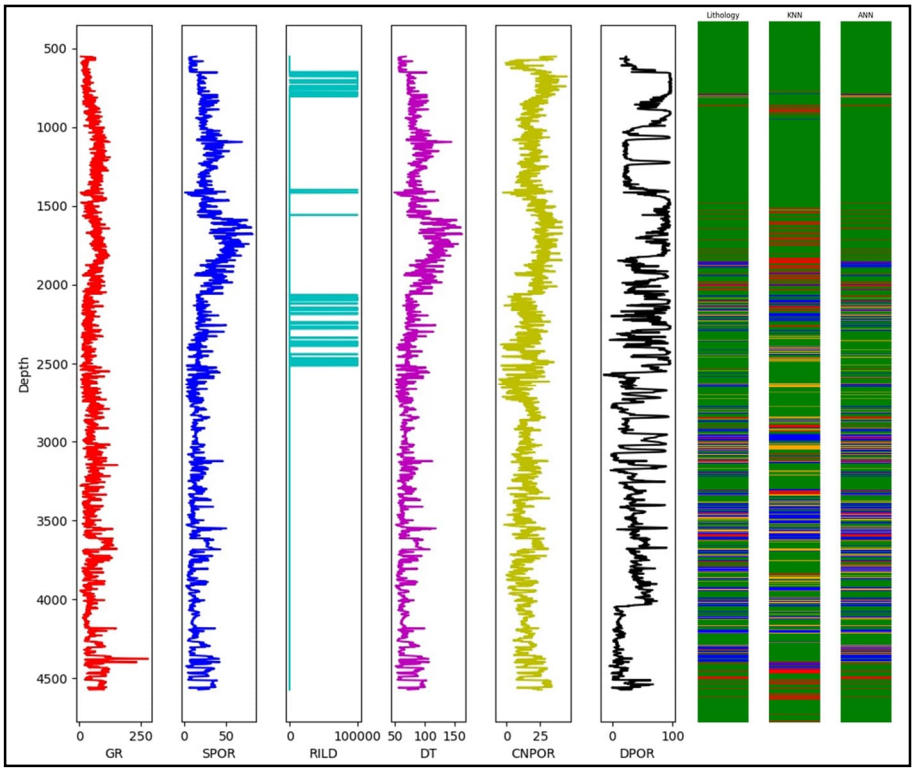

Figure 7. Prediction results of the KNN and ANN methods at the Mc-Fadden 'A' 1-33 well.

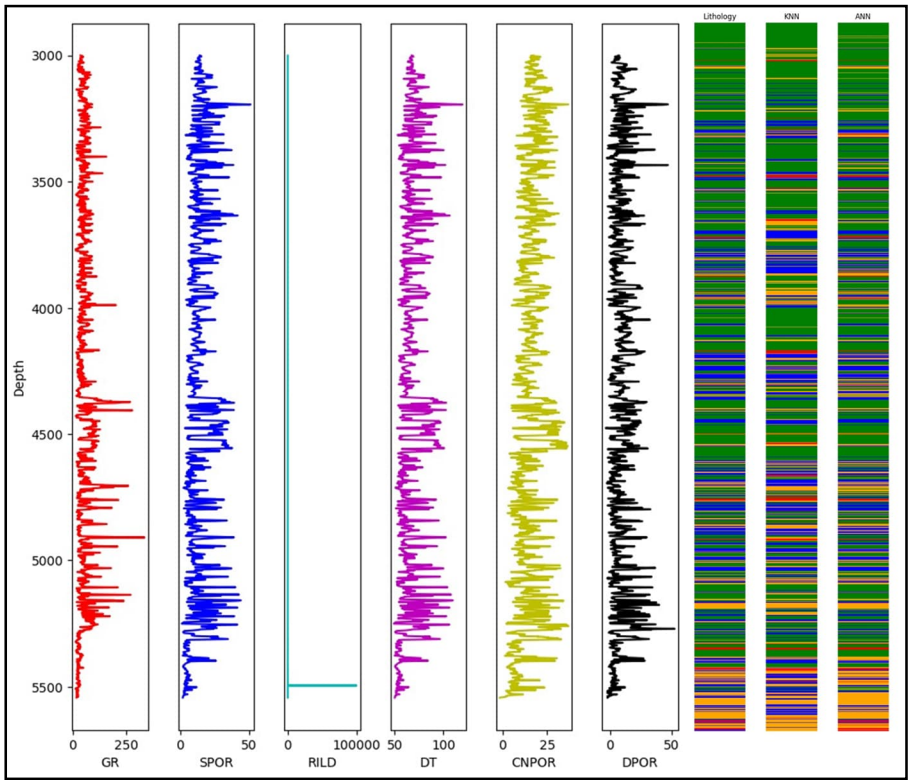

Figure 8. Prediction results of the KNN and ANN methods at the Ferkert Farms 1-8 well.

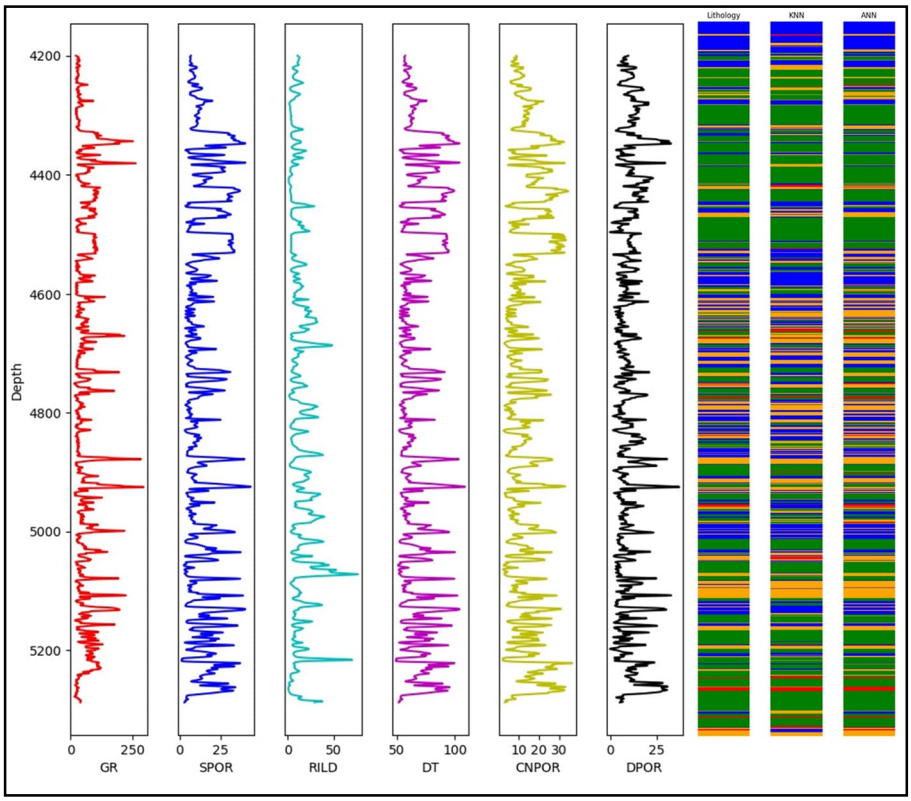

Figure 9. Prediction results of the KNN and ANN methods at the Ferkert Farms 5-8 well.

4.4. Limitations and Implications

The results show that ANN consistently outperformed KNN in lithology classification based on well log data. The optimal ANN configuration achieved accuracy above 95%, AUC values close to 1.00, and low, stable Log Loss across all cross-validation folds. In contrast, KNN achieved testing accuracies of approximately 75-80% but exhibited lower cross-validation performance and less stable predictions in intervals characterized by overlapping lithological responses. These findings indicate that nonlinear learning architectures are more suitable for heterogeneous subsurface datasets than distance-based classifiers

Several limitations should be noted. First, lithology labels were derived from well log interpretation rather than direct core data, which may introduce interpretive bias. Second, although SMOTE was applied only to the training data prior to cross-validation to prevent data leakage, the synthetic samples may not fully represent natural geological variability, as SMOTE generates interpolated samples based on feature-space proximity rather than actual geological processes. It is further acknowledged that applying SMOTE before cross-validation rather than within each fold as a pipeline may introduce a degree of optimistic bias in the reported performance metrics. This was adopted due to software limitations in Orange. Third, the study was limited to four main lithology classes and a single study area in Ford County, Kansas, which may restrict the generalizability of the results to other basins with different geological settings. Future studies should incorporate core-calibrated labels, multi-basin datasets, and more advanced deep learning approaches to improve robustness and generalization.

From a practical perspective, ANN-based lithology classification offers clear benefits for subsurface characterization. It can reduce subjectivity in manual interpretation, accelerate multi-well analysis, and improve consistency in complex geological settings. Therefore, machine learning can serve as an effective decision-support tool for geophysicists and reservoir engineers in data-driven subsurface interpretation.

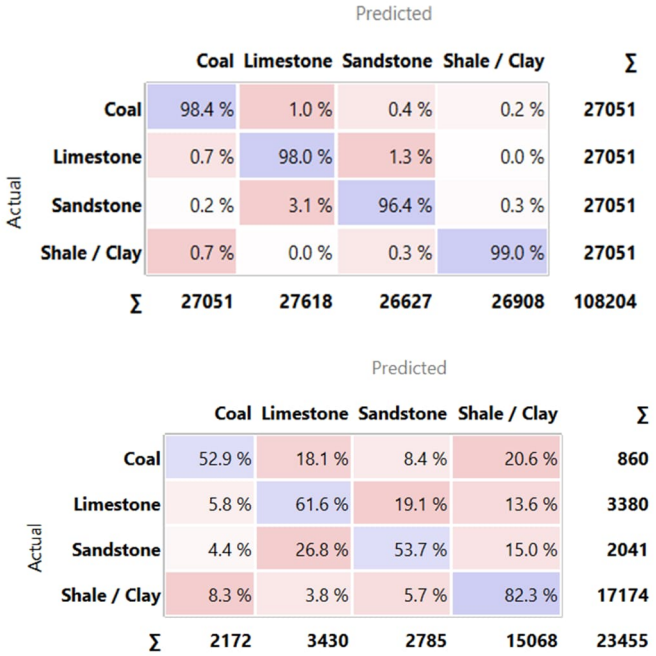

Figure 10. Confusion matrices of ANN (above) and KNN (below) lithology prediction results. ANN demonstrates superior

performance (>95% accuracy) with stable classification across classes, whereas KNN exhibits lower accuracy (±60%) and higher misclassification in overlapping lithological intervals.

5. CONCLUSION

This study demonstrates the application of machine learning for lithology classification using well log data within the scope of six log paramters, four lithology classes, and 14 wells from Ford County, Kansas. The results indicate that ANN and KNN are capable of performing lithology prediction under the given dataset, with ANN showing more stable behavior across cross-validation compared to KNN. The use of SMOTE improved class balance and model stability, while 5-fold cross-validation confirmed good generalization performance within the dataset used in this study.

From a practical perspective, ANN can support faster multi-well lithology interpretation, reduce subjectivity in manual analysis, and assist subsurface characterization. Although increasing the number of training iterations reduced Log Loss to a certain extent, excessive iterations did not yield significant performance gains and reduced computational efficiency. Overall, ANN with balanced data and optimized hyperparameters proved to be the most robust model in this study. However, its applicability remains dependent on the six log parameters and four lithology classes used here, and further adaptation is required for different geological settings.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to the Kansas Geological Survey at the University of Kansas for providing the dataset used in this study.

REFERENCES

Adiwiguna, Y. P., Purbaseno, H., Syuhara, H., Pribadi, F. J., Sarbani, A., & Kusuma, A. D. (2022). Membangun Tool Prediksi yang Komprehensif Guna Meningkatkan Akurasi Produk Akhir Pencampuran Kualitas Batubara Menerapkan Perbandingan Kolaboratif pada Algoritma Regresi Hingga Neural Network. *Indonesian Mining Professionals Journal*, *4*(2 November 2022). https://doi.org/10.36986/impj.v4i2.79

Asquith, G. B., Krygowski, D., Henderson, S., & Hurley, N. F. (2004). *Basic Well Log Analysis (2nd ed.*). American Association of Petroleum Geologist.

Esiri, A. E., Jambol, D. D., & Ozowe, C. (2024). Enhancing Reservoir Characterization With Integrated Petrophysical Analysis and Geostatistical Methods. *Open Access Research Journal of Multidisciplinary Studies*, *7*(2), 168–179. https://doi.org/10.53022/oarjms.2024.7.2.0038

Galupino, J. & Dungca, J. (2022). Machine Learning Models to Generate a Subsurface Soil Profile : A Case Of Makati City , Philippines. *International Journal of GEOMATE*, *23*(95), 57–64. https://doi.org/https://doi.org/10.21660/2022.95.3029

Hamid, S. & Mousavi, R. (2024). A Novel Approach To Classify Lithology of Reservoir Formations Using GrowNet and Deep - Insight With Physic - Based Feature Augmentation. *Energy Science & Engineering*, *July*, 4453–4477. https://doi.org/10.1002/ese3.1895

Johnson, W.C. & Woodburn, T. L. (2009). *Surficial Geology of Ford County, Kansas* (Map Series 108). Kansas Geological Survey.

Kansas Geological Survey (KGS) (2004). *Oil and Gas Well Log Database And Geological Data*. University of Kansas. https://kgs.ku.edu/data-and-maps

Khatti, J. & Grover, K. S. (2022). Determination of Suitable Hyperparameters of Artificial Neural Network for the Best Prediction of Geotechnical Properties of Soil. *International Journal For Research in Applied Science & Engineering Technology (IJRASET)*, *10*(V May 2022).

Komara, E., Nugraha, M. F., & Rafi, M. (2024). Lithology Prediction Using K-Nearest Neighbors (KNN) Algorithm Study Case In Upper Cibulakan Formation. *Geomatics International Conference 2024 (GEOICON 2024), IOP Conference Series Earth and Enviromental Science.* https://doi.org/10.1088/1755-1315/1418/1/012062

Marembayev, T., Kurmangaliyev, D., Bekbaouv, B., & Amanbaek, Y. (2021). A Comparison of Machine Learning Algorithms in Predicting Lithofacies: Case Studies from Norway and Kazakhstan. *Energies*, *14*(1986), 1–16.

https://doi.org/https://doi.org/10.3390/en14071896

Maulana, A., Lepong, P., Rayzy, D., Sutaji, P., Munir, R., Fisika, S., Mulawarman, U., Geofisika, S., & Mulawarman, U. (2024). Identifikasi Keberadaan Hidrokarbon Menggunakan Inversi Impedansi Akustik dengan Algoritma Artificial Neural Network. *Jurnal Geosains Kuta Basin*, *7*(1).

Mohamed, I. M., Mohamed, S., Petroleum, K., Mazher, I., & Chester, P. (2019). Formation Lithology Classification : Insights into Machine Learning Methods. *Society of Petroleum Engineers*, *SPE-196096*.

Nabilah, P., Masrurah, Z., Ikhlas, & Putra, R. R. (2023). Identifikasi Lapisan Akuifer Wilayah Aceh Besar Berdasarkan Korelasi Data Electrical Logging dan Cutting. *Jurnal Geofisika Eksplorasi*, *09*(02), 131–141. https://doi.org/https://doi.org/10.23960/jge.v9i2.279

Prabowo, U. N., Ferdiyan, A., & Raharjo, S. A. (2023). Comparison of Facies Estimation Using Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) Algorithm Based on Well Log Data. *Aceh International Journal of Science and Technology*, *12*(August), 246–253. https://doi.org/10.13170/ aijst.12.2.28428

Putra, Army, M. I. R., Adiwarman, M., & Rassarandi, F. D. (2023). Pengolahan Citra Digital Untuk Penentuan Jenis Batubara di Pt Bukit Asam, Tbk. Dengan Menggunakan Metode Algoritma K-Nearest Neighbor. *Jurnal Ilmiah Teknik Dan Sains (JITS)*, *1*(1), 6–10. https://doi.org/https://doi.org/10.62278/jits.v1i1.2

Qamar, R. & Zardari, B. A. (2023). Artificial Neural Networks : An Overview. *Mesopotamian Journal of Computer Science*, 2023, 124–133. https://doi.org/10.58496/ MJCSC/2023/015

Rohmana, R. C., Triwanti, D., Setiyaningrum, P. R., Perminyakan, T., Perminyakan, T., & Perminyakan, T. (2024). Penerapan Machine Learning dalam Penentuan Porositas Batuan : Studi Kasus Menggunakan Regresi Linier Berganda dan Regresi KNN pada Data Log Sumur Application of Machine Learning in Rock Porosity Determination : Case Study Using Multiple Linear Regression. *Jurnal Teknik (JT) FT UMT*, *13*(02), 42–50.

Sastra, M. M. & Rohmana, R. C. (2024). Perbandingan Metode Klasifikasi Machine Learning : Studi Kasus Prediksi Jenis Litologi Berdasarkan Data Well Log Pada Formasi Sleipner , North Sea Comparison of Machine Learning Classification Methods : Case Study of Lithology Type Prediction Based on Wel. *Jurnal Teknik (JT) FT UMT*, *13*(02), 51–63.

Sephiana, S. E., Karyanto, & Sinambela, R. Z. (2023). Analisis Petrofisika dalam Mengidentifikasi Zona Hidrokarbon pada Formasi Tualang dan Lakat. *Jurnal Geofisika Eksplorasi*, *9*(3), 165–183. https://doi.org/https://doi.org/10.23960/jge.v9i3.273

Septyandy, M. R., & Subroto, E. A. (2023). Penentuan Model Total Organic Carbon dengan Menggunakan Metode Artificial Neural Network dan Adaptive Neuro Fuzzy Inference System untuk Estimasi Potensi Gas Serpih di Cekungan Jawa Barat Utara. *Jurnal Ilmiah Teknologi Informasi Asia*, *17*(1), 83–96.

Sudrazat, S. D., Purba, B., Wijaksono, E., Pranowo, W., & Hibatullah, M. I. (2020). Prediksi Kecepatan Gelombang S dengan Machine Learning pada Sumur "S-1", Cekungan Sumatera Tengah Indonesia. *Lembaran Publikasi Minyak Dan Gas Bumi*, *54*(1), 29–35.

Wang, X., Yang, S., Zhao, Y., & Wang, Y. (2018). Lithology Identification Using An Optimized KNN Clustering Method Based on Entropy-Weighed Cosine Distance in Mesozoic Strata of Gaoqing Field, Jiyang Depression. *Journal of Petroleum Science and Engineering*, *166*, 157–174. https://doi.org/10.1016/j.petrol.2018.03.034

Wardhana, S. G., Aldi, M., & Siregar, R. I. (2022). Prediksi Kecepatan Gelombang Geser (Vs) Menggunakan Machine Learning di Sumur X. *Jurnal Geofisika Eksplorasi*, *08*(01), 67–77.

Wardhana, S. G. & Pakpahan, H. J. (2021). Algoritma Komputasi Machine Learning untuk Aplikasi Prediksi Nilai Total Organic Carbon (TOC). *Lembaran Publikasi Minyak Dan Gas Bumi*, *55*(2), 75–87.