Scientific Contributions Oil & Gas, Vol. 49. No. 1, March: 489 - 510

SCIENTIFIC CONTRIBUTIONS OIL AND GAS

Testing Center for Oil and Gas
LEMIGAS

Journal Homepage:http://journal.lemigas.esdm.go.id
ISSN: 2089-3361, e-ISSN: 2541-0520

Machine Learning for Reservoir Characterization: Lithology Prediction Using Support Vector Machine (SVM) in The "VISA" Field, East Kalimantan

Muhammad Faiz Nugraha[1], Eki Komara[2], Abdul Halim Abdul Latiff [1], Wien Lestari[2], Edy Wijanarko[3]

[1]Teknologi Petronas University
Teknologi PETRONAS University Seri Iskandar, Perak, Malaysia

[2] Institut Teknologi Sepuluh Nopember
Teknik Kimia Street, Keputih, Surabaya, East Java 60111, Indonesia

[3]Testing Centre for Oil and Gas Technology LEMIGAS
Ciledug Raya Street, Cipulir, South Jakarta City 12230, Jakarta, Indonesia

Corresponding Author : Eki Komara (komara@its.ac.id)

Manuscript received: January 7th, 2026; Revised: January 26th, 2026
Approved: February 27th, 2026; Available online: March 2th, 2025; Published: March 2th, 2026.

ABSTRACT - This research is conducted in the "VISA" field of the Balikpapan Formation, located in the Kutai Basin, one of Indonesia's largest hydrocarbon basins. The lithology of this formation is primarily sandstone and shale, which are significant for hydrocarbon exploration and production. Determining the initial lithology is an essential step for understanding the characteristics of the well data during processing. Consequently, the Support Vector Machine (SVM) algorithm is implemented in this study to predict lithology using well data. This investigation employs data from four wells: VISA-9, VISA-13, VISA-36, and VISA-39. The prediction results are subsequently visualized as well logs and lithology distribution histograms to make the results easier to interpret based on three interpreted lithology categories: sandstone, shale, and coal. Performance evaluations indicate that limitations remain in the SVM classification. The error range obtained in Experiments 1 and 2 is 11–22% compared to the actual lithology. However, Experiment 3 demonstrates substantial improvement by utilizing three training datasets, which reduces the error rate to 5% (a 7% improvement from previous experiments). Overall, the SVM method can effectively classify rock lithology; however, the model still requires optimization to minimize residual errors during the prediction process. Ultimately, this investigation demonstrates that SVM can be successfully applied to predict lithology using well log parameters.

Keywords: lithology prediction, support vector machine, radial basis function, accuracy

How to cite this article:

Muhammad Faiz Nugraha, Eki Komara, Abdul Halim Abdul Latiff, Wien Lestari, Edy Wijanarko 2026, Machine Learning for Reservoir Characterization: Lithology Prediction Using Support Vector Machine (SVM) in The "VISA" Field, East Kalimantan, Scientific Contributions Oil and Gas, 49 (1) pp. 489 - 510. DOI org/10.29017/scog.v49i1.2032

INTRODUCTION

This research was conducted in the "VISA" field of the Balikpapan Formation, which is situated in the Kutai Basin, one of the largest hydrocarbon basins in Indonesia (Abhimantra 2021). The lithology of this formation is primarily composed of Sandstone and shale, which are significant for the exploration and production of hydrocarbons (Asquith et al., 2004). Determining the initial lithology is an essential step for comprehending the characteristics of the well data in the context of well data processing (Augusto & Martins 2009).

Consequently, the Support Vector Machine (SVM) algorithm was implemented in this investigation to predict lithology using well data (Burges 1998). This investigation employs data from four wells: VISA-9, VISA-13, VISA- 36, and VISA-39. The prediction results are subsequently visualized as well as logs and lithology distribution histograms to facilitate the interpretation of the results based on three lithology categories: Sandstone, shale, and Coal (Al Ghaithi & Prasad 2020).

Performance evaluations indicate that limitations remain in the SVM classification of well data parameters (Cortes & Vapnik 1995). The error range obtained in Experiments 1 and 2 was 11 –22% compared to the actual lithology. However, the fourth experiment demonstrated substantial improvement by utilizing three training datasets, which reduced the error rate to 5% (a 7% accuracy increase from the previous experiment). Overall, the SVM method can be employed to classify rock lithology based on the actual well data (Géron 2022). However, the model still requires optimization due to residual errors that occur during the prediction process.

Ultimately, this investigation demonstrates that SVM can be successfully applied to the lithology prediction process using well log parameters.

LITERATURE REVIEW

Regional Geology

The Kutai Basin in East Kalimantan is one of the largest sedimentary basins in Indonesia, renowned for its abundant hydrocarbon reserves. Geologically, the basin was formed by tectonic activity that began during the Tertiary Period, approximately 65 million years ago, and it continues to undergo sedimentation to the present day (Abhimantra 2021). The structural complexity of the Kutai Basin is the result of significant erosion and sedimentation processes, alongside the influence of past extensional tectonic activity (Husein 2015). The sediments accumulating in this basin are primarily composed of clastic deposits, such as Sandstone, mudstone, and lignite (Handoyo et al., 2018). As a foreland basin, the Kutai Basin was subjected to depositional processes driven by erosion from the Schwaner Mountains to the southwest and other mountain ranges encircling East Kalimantan (Hidayat et al., 2021).

In the Kutai Basin, hydrocarbons are present in both conventional reservoirs, such as deltaic sandstones, and as shale gas held within organic-rich shale formations (Khawarizmy et al., 2020). Additionally, the basin has the potential to become a significant site for shale gas development, which is currently a major focus in alternative energy exploration (Sunarjanto et al., 2014).

The depositional environments within the basin are categorized into three primary zones: the shelf, the slope, and the basin (Battu 2026). The shelf zone is dominated by deltaic sediments (Lunt 2019), which frequently function as hydrocarbon reservoirs. Conversely, the slope and basin zones accumulate deep-sea sediments (Figure 1).

Basic parameters for determining lithology based on well data

Serra (1983) stated that Sandstone typically exhibits a gamma ray value between 15 and 75 API due to a predominant quartz mineral content and a minor amount of clay minerals. The rock is generally compact and hard, which is reflected in sonic logs by a rapid travel time of 55 to 70 µs/ft (Doust 2008). The resistivity of Sandstone is contingent upon the fluid occupying its pores, typically ranging from 2 to 15 ohm-m (Ellis, 2007). Furthermore, the density log value of Sandstone is between 2.00 and 2.65 g/cc, and its porosity ranges from 10% to 25% (Horsfall et al., 2013).

Shale is characterized by gamma ray values exceeding 75 API due to an abundance of clay minerals, which emit natural radiation (Guo et al., 2005). Sonic logs indicate that shale is fine-grained and less dense than tighter formations, resulting in a slower travel time of 70 to 100 µs/ft (Munadi 2007). Because shale is frequently saturated with bound water, its resistivity is low, generally ranging from 0.5 to 2 ohm-m (Zaemi et al., 2022). Shale porosity is typically low and ineffective, ranging from 5% to 10%, while its density ranges from 2.40 to 2.60 g/cc (Killeen 1982). Lastly, Coal typically yields a gamma ray value below 10 to 15 API due to a nearly complete absence of clay content.

Sonic logs indicate that Coal contains organic material and has a complex structure, evidenced by slow travel times of 100 to 140 µs/ft (Moherek et al., 2015). The resistivity of Coal is extremely high, ranging from 100 to 2,000 ohm-m. While Coal has a high total porosity, ranging from 30% to 50%, it is ineffective for fluid storage (Passey 2010). The material's lightweight structure results in low-density readings on well logs, ranging from 1.20 to 1.80 g/cc (Qi et al., 2021).

Support vector machine (SVM)

A Support Vector Machine (SVM) is a machine learning algorithm that is used for regression and

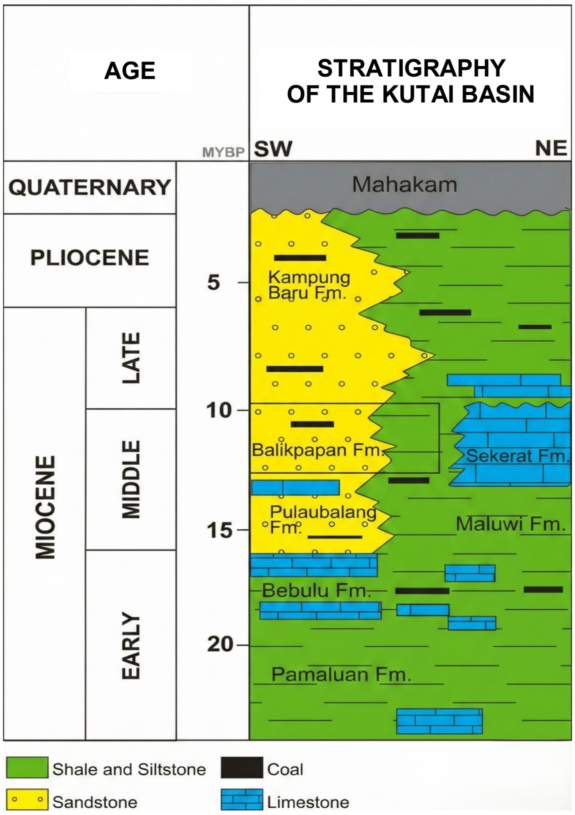

Figure 1. Regional Stratigraphy of the Kutai Basin (Abhimantra 2021)

classification (Sebtosheikh et al., 2015). It maximizes the margin, or distance, between the two classes in order to identify an optimal *hyperplane* (decision boundary) as a maximum margin classifier (Mahesh 2020). Support vectors are the specific data points that are located on the margin's edge. These "limit instances" function as the constraint conditions for the optimization problem (Sebtosheikh et al., 2015). To transform data that is not linearly separable into a higher-dimensional space where a linear boundary can be identified, SVM utilizes kernel functions (such as polynomial, RBF, or sigmoid) (Usama et al., 2019). Furthermore, a regularization parameter (*C*) is introduced by the soft margin notion to manage misclassified data. This parameter strikes a compromise between the need to minimize the classification error and maximize the margin (Wardhana et al., 2025).

Nonlinear classification support vector machine

In Support Vector Machines (SVM), the capacity to manage data that is not linearly separable is referred to as nonlinear classification. For nonlinear data, a simple linear hyperplane is insufficient to distinguish between distinct classes (Shalev-Shwartz., 2014). To address this, SVM employs a kernel function to map the data from its original feature space (a lower-dimensional space) into a higher-dimensional space. Data that was initially not linearly separable in the original space can become linearly separable in this newly transformed space, enabling the SVM to identify an optimal separating hyperplane.

Support vector machine parameter

To transform data that is not linearly separable into a higher-dimensional space where a linear boundary can be identified, SVM utilizes kernel functions (such as Polynomial, Radial Basis Function [RBF], or Sigmoid). Furthermore, the soft margin concept introduces a regularisation parameter (C) to manage misclassified data. This parameter strikes a balance between the need to minimize classification errors and maximize the margin.

$$min\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\delta_i \qquad (1)$$

min	:	goal to find the lowest possible value
w	:	weight vector that defines the position of the boundary
$\|\|w\|\|^2$	:	margin size
C	:	penalty parameter
n	:	total number of training data points
i	:	individual data point
δ_i	:	error margin for a misclassified data point
$\sum_{i=1}^{n}\delta_i$	:	Summary of all errors across training data

The gamma parameter (γ) is the subsequent parameter, and it is employed to ascertain the complexity of a margin in the kernel function. The training of the SVM algorithm is also influenced by this parameter. In this investigation, the gamma parameter functions as a hyperparameter for the prediction of lithology. The gamma value is determined on a scale of 0 to 1. The classification conducted is more narrow and complex as the gamma value increases. However, the likelihood of overfitting is also increased. A broader classification is achieved with small gamma values; however, underfitting may occur. The gamma parameter is represented by the following equation :

$$\gamma = \frac{1}{2\sigma^2} \qquad (2)$$

A unconstrained parameter is denoted by σ. The function value approaches 1 when $x \approx x'$, while it approaches 0 (in the limit) as x deviates from x'. This can be interpreted as a similarity measure for the conventional value of σ2=1 (Hastie, 2009). Additionally, the feature will adjust more consistently as σ2 increases, resulting in a lower variance and a higher bias. In contrast, the feature will exhibit a less uniform pattern when σ2 values are smaller, which leads to a higher variance and lower bias (Figure 2).

The Radial Basis Function (RBF) kernel is employed in this investigation. This kernel function depends on the distance between two data points, which is typically quantified using the Euclidean distance. The RBF kernel operates on the principle that the value of the function decreases as the distance from the centre point increases (Zamri et al., 2022).

Conversely, the kernel's value approaches its maximum when the points are in close proximity. The RBF equation is expressed as follows:

$$K(x,x') = esp(-\gamma||x - x'||^2) \quad (3)$$

In the input space, x and x' represent two feature vectors (data point), and the gamma (γ) parameter regulates the degree of influence that one data point has on another. The extent of influence is reduced as the gamma (γ) value increases.

METHODOLOGY

Research data

This study utilizes well data from the "VISA" field within the Balikpapan Formation, located in the Kutai Basin of East Kalimantan. The data spans varying depths from 1,800 to 3,000 MD. The dataset comprises four wells, with the details regarding parameter completeness outlined in Table 1.

Table 1. Dataset Detail

Well	LOG PARAMETER						
	GR	DT	LLD	ILD	NPHI	PHIN	RHOB
VISA-9	✓	✓	✓	-	✓	-	✓
VISA-19	✓	✓	✓	-	✓	-	✓
VISA-36	✓	✓	-	✓	-	✓	✓
VISA-39	✓	✓	-	✓	-	✓	✓

Determination Of lithology based on log parameters

The data has been imported, then the next stage is the interpretation stage qualitative. Qualitative interpretation is carried out with a quick look interpretation based on the value of each data logging parameter (Figure 3). The results of lithological interpretation in each well obtained three types of lithology, namely Sandstone, shale, and Coal. After completing qualitative interpretation, four wells data were exported back into LAS File.

Dataset preparation

Table 2. Experiment Dataset Ratio

Experiment	Ratio	Train Dataset	Test Dataset
1	25:75	VISA 9	VISA 13 VISA 36 VISA 39
2	50:50	VISA 9 VISA 36	VISA 13 VISA 39
3	75:25	VISA 13 VISA 36 VISA 39	VISA 9

The interpreted well data was imported into a Jupyter Notebook environment using the Python programming language. Prior to loading the data, the necessary Python libraries were imported to enable the required file-handling functions. Subsequently, the data from the four wells was loaded using the Pandas library. Once imported,

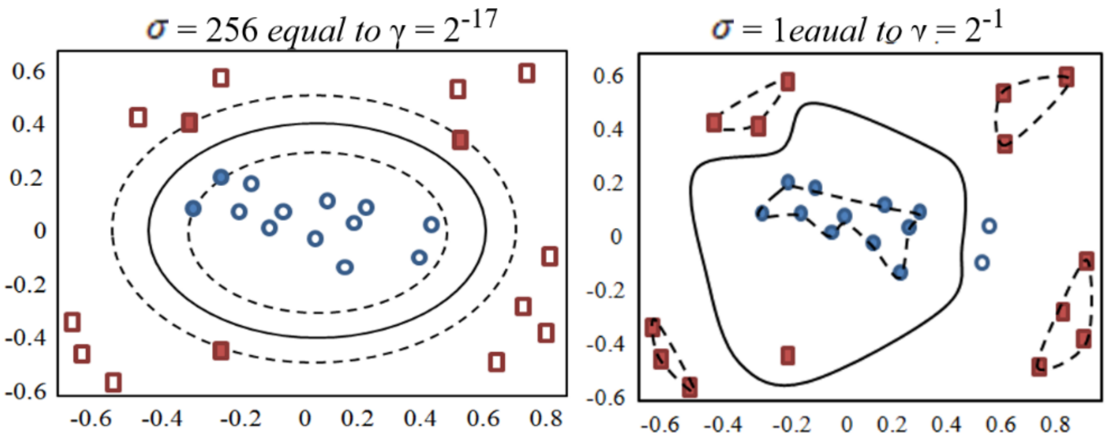

Figure 2. Variations on Gaussian RBF Kernel with Variations of Parameter γ

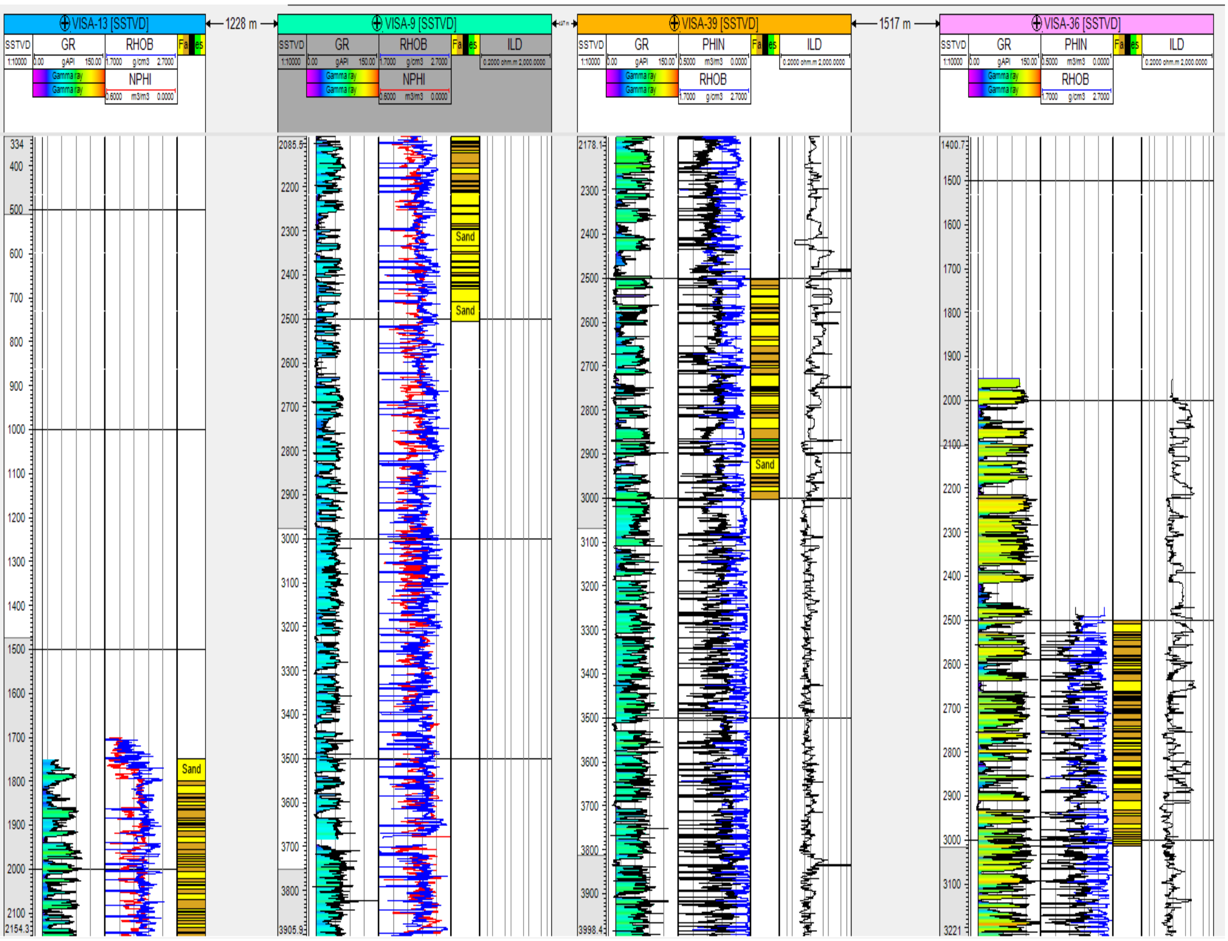

Figure 3. Determination of Lithology Based On Log Parameters

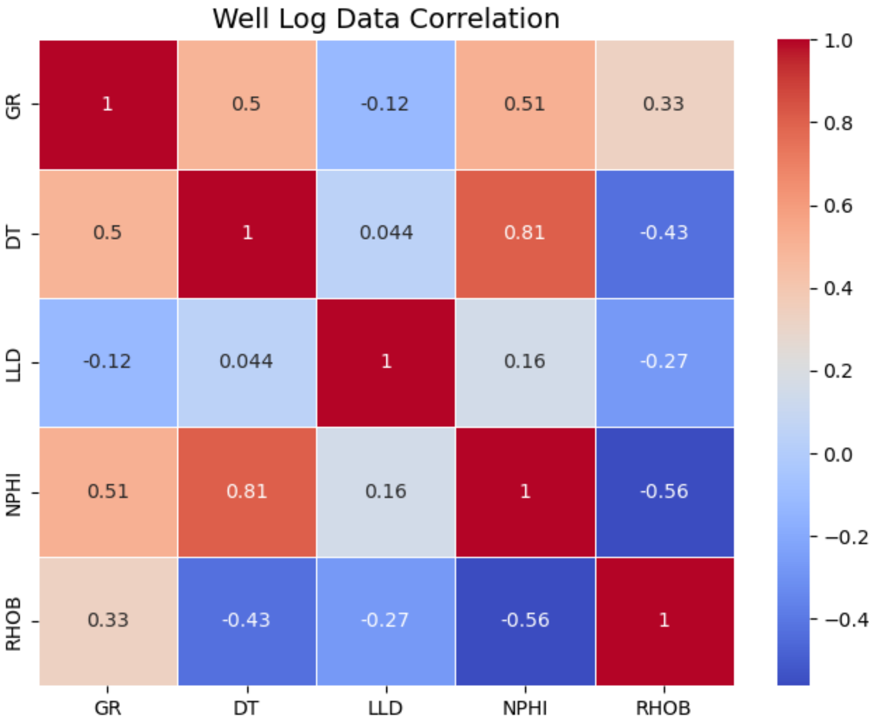

Figure 4. Correlation of Well Log Data Parameters

the well data was converted into DataFrames and divided into training data (comprising VISA -13, VISA-36, and VISA-39) and test data (specifically VISA-9). The experiment was conducted three times, with the specific data ratios detailed in Table 2.

Build support vector machine

The Support Vector Machine (SVM) model relies on three main parameter which is C, gamma (γ), and the kernel. In this study, GridSearch CV a module within the scikit-learn library is utilized to identify the most suitable parameter values for the well data. The regularisation parameter (C) is evaluated across the range [0.1, 1, 10, 100]. These values are selected to observe how effectively the SVM determines the optimal margin for the data during lithology prediction. Similarly, the gamma (γ) parameter is tested across the range [1, 0.1, 0.01, 0.001] to identify its optimal value. Finally, the Radial Basis Function (RBF) kernel is employed. This kernel is designed to map data into higher-dimensional spaces and is commonly utilized for conducting nonlinear classification in large dataset. Lithology predictions on both the training and testing datasets are conducted once the SVM model has been constructed using optimized parameters. This approach ensures that the model avoids overfitting or underfitting during the prediction phase. The analysis focuses on evaluating the performance of the prediction results to assess the effectiveness of the SVM model on the utilized dataset and its ability to analyze the data accurate.

Analysis of the performance results base on dataset train dataset test

Following the lithology predictions, the model's overall performance will be analyzed. The specific evaluation metrics used for this assessment include precision, F1-score, and overall prediction accuracy. Furthermore, the SVM prediction results will be directly compared against the actual well lithology (the ground truth) established prior to the prediction process.

RESULT AND DISCUSSION

Data well log correlation

The correlation data in Figure 4 indicate that the NPHI log parameter strongly correlates with other parameters, as demonstrated by a DT log correlation coefficient of 0.81. Conversely, the RHOB log parameter exhibits the least correlation with the other log parameters. Furthermore, the correlation results of the well log parameters suggest that a balanced data distribution yields a more reliable model response, and that an improper combination of datasets may lead to overfitting.

Correlation of log and lithology data

The resulting pairplots illustrate the relationships among various well log parameters including DEPTH, GR, DT, LLD, NPHI, and RHOB and demonstrate how each parameter is associated with distinct lithologies (Sandstone, shale, and Coal) (Figure 5). The diagonal elements of the pairplot display the individual distribution of each parameter across the three lithology types. Shale exhibits higher GR values than the other lithologies, while Coal displays distinctive distribution patterns for several parameters, notably NPHI and RHOB. Furthermore, the scatterplots indicate a potential positive correlation between RHOB and NPHI, suggesting that these two parameters increase in tandem. Conversely, the absence of a discernible correlation pattern between DEPTH and GR suggests that these parameters are not directly related.

Results of training and testing experiment 1 training dataset experiment 1

The results of the VISA-9 lithology evaluation, including the confusion matrix and performance metrics, are presented in Table 3. Sandstone exhibits the highest recall value of 0.90, indicating that nearly all actual sandstone instances are successfully identified. However, its precision is only 0.64, which implies a high number of false positives.

Table 3. Lithology Prediction Results in VISA-9 Train Dataset Experiment 1

Lithology	Confusion matrix			
	True positive (TP)	False positive (FP)	False negative (FN)	True negative (TN)
Sandstone	234	130	27	410
Shale	363	31	139	268
Coal	30	13	8	750

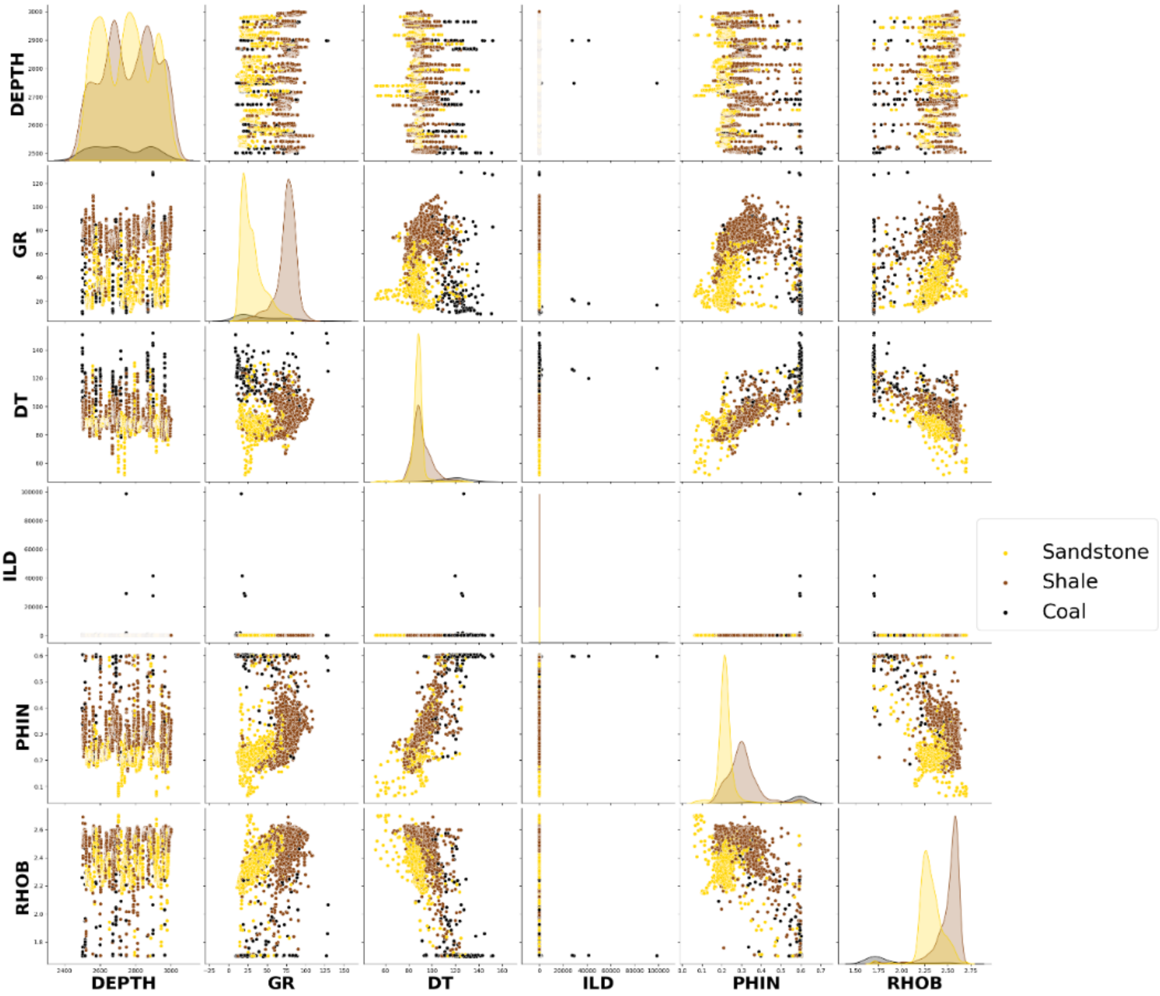

Figure 5. Correlation of Lithology Distribution to Well Data Parameter

Table 4. Performance Parameters Result in VISA-9 Train Dataset Experiment 1

Performance			
Accuracy	Precision	Recall	F1 Score
0.78	0.64	0.9	0.75
	0.92	0.72	0.81
	0.7	0.79	0.74

In Table 3, the precision and recall values for shale lithology are 0.72 and 0.81, respectively, suggesting that the shale predictions are more balanced and resistant to false positives. Conversely, coal lithology exhibits the highest precision value of 0.79 but the lowest Recall of 0.74. This indicates a high number of false negatives resulting from the model's inability to detect Coal during the prediction process (Table 4). Overall, the prediction model employed to analyze the VISA-9 test data was found to be more effective at identifying Sandstone than Coal. This outcome is likely attributed to the difficulty in recognizing coal patterns due to their limited representation within the dataset.

Figure 6 illustrates a comparison between the two datasets, where the blue bars represent the actual lithology distribution and the red bars represent the predicted distribution. Sandstone and Coal are over-predicted, as their predicted values exceed the actual values. Conversely, shale is under -predicted, because its predicted values are lower than the actual values. These misclassifications are likely caused by the relatively limited number of coal samples and the overlapping well log characteristics of Sandstone and shale, which are

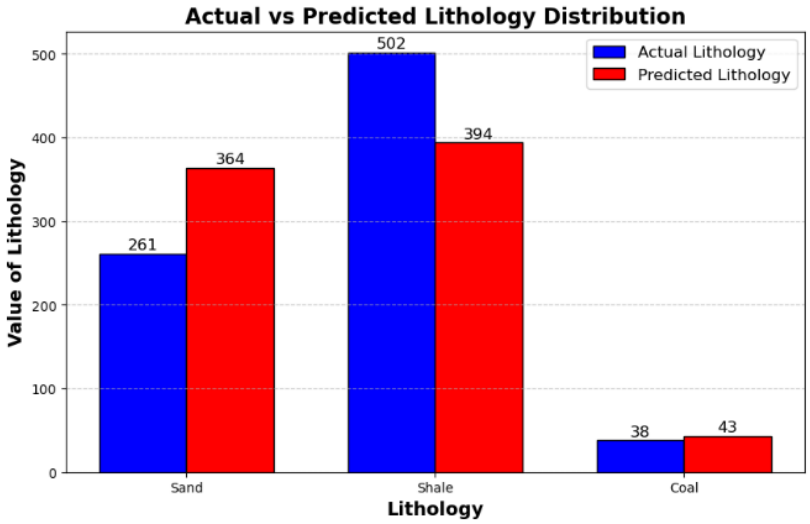

Figure 6. Lithology Distribution Histogram in VISA-9 Experiment 1

obscured by the larger proportion of other lithologies in the dataset.

Testing Dataset Experiment 1

The prediction results presented in Table 5 yield an overall accuracy of 0.80. The precision value of 0.79 for sandstone lithology indicates that 79% of all sandstone predictions are correct. However, there are still 78 false positives, resulting from the misidentification of shale or Coal as Sandstone.

The shale lithology in this well exhibits a precision of 0.87, suggesting fewer false positives compared to Sandstone. Finally, coal lithology has a precision of 0.50 and a recall value of 0.94; this indicates that although most actual coal instances are successfully detected, approximately 50% of the predicted coal samples are incorrectly classified and are actually Sandstone or shale.

Table 5. Lithology Prediction Results in VISA-13 Test Dataset Experiment 1

Lithology	Confusion matrix			
	True positive (TP)	False positive (FP)	False negative (FN)	True negative (TN)
Sandstone	290	78	53	380
Shale	348	51	92	310
Coal	17	17	1	766

Table 6. Performance Parameters Result in VISA-13 Test Dataset Experiment 1

Performance			
Accuracy	Precision	Recall	F1 Score
0.8	0.98	0.59	0.74
	0.78	0.95	0.86
	0.51	0.96	0.67

The overall accuracy is 0.80, as demonstrated in Tables 7 and 8. The precision for sandstone lithology is 0.98, indicating that nearly all sandstone predictions are correct. However, the recall value of 0.59 suggests that a significant number of actual sandstone instances are undetected, having been misclassified as shale or Coal, which is reflected by a relatively high false negative count of 523. Furthermore, the prediction for shale lithology yields a precision value of 0.78 and a recall of 0.95, indicating that the model successfully identifies the vast majority of actual shale. Nevertheless, there are still a substantial number of false positives - specifically 506 - which implies that the model frequently misclassifies Sandstone or Coal as shale. Finally, due to the limited quantity of coal data, the SVM model only learns limited characteristics from the available samples. This results in the lowest precision of 0.51 and a recall of 0.96 for coal lithology.

Table 7. Lithology Prediction Results in VISA-36 Test Dataset Experiment 1

Lithology	Confusion Matrix			
	True positive (TP)	False positive (FP)	False negative (FN)	True negative (TN)
Sandstone	1458	119	161	1596
Shale	1330	139	203	1662
Coal	162	126	20	3026

Table 8. Performance Parameters Result in VISA-36 Test Dataset Experiment 1

Performance			
Accuracy	Precision	Recall	F1 Score
0.86	0.92	0.9	0.91
	0.91	0.87	0.89
	0.56	0.89	0.69

In Table 9, the VISA-39 prediction results are evaluated, showing an overall accuracy of 0.86. The precision and recall values for sandstone lithology are 0.92 and 0.90, respectively, suggesting that the prediction results are highly accurate with a relatively low number of false negatives in comparison to the true positives. In contrast to the other lithologies, shale has a relatively high number of false negatives, leading to a lower recall value (Table 10). Concurrently, coal lithology has a precision value of 0.56, which is considered low due to the significant discrepancy between the predicted and actual results.

Table 9. Lithology Prediction Results in VISA-39 Test Dataset Experiment 1

Lithology	Confusion matrix			
	True positive (TP)	False positive (FP)	False negative (FN)	True negative (TN)
Sandstone	1458	119	161	1596
Shale	1330	139	203	1662
Coal	162	126	20	3026

Table 10. Performance Parameters Result in VISA-39 Test Dataset Experiment 1

Performance			
Accuracy	Precision	Recall	F1 Score
0.86	0.92	0.9	0.91
	0.91	0.87	0.89
	0.56	0.89	0.69

Figure 7 illustrates the comprehensive model evaluation. The accuracy results indicate minor overfitting, as the testing data yields a lower accuracy than the training data. Additionally, the F1-score for the testing dataset decreased, suggesting a reduction in the model's generalization capability. While Sandstone and shale exhibit stable precision values, Coal shows a decline in precision. The disparity in prediction distributions suggests that bias persists, likely as a result of the unbalanced data split (25% training data and 75% testing data).

Results of training and testing experiment 2 training dataset experiment 2

Table 11. Lithology Prediction Results in VISA-9 Train Dataset Experiment 2

Lithology	Confusion matrix			
	True positive (TP)	False positive (FP)	False negative (FN)	True negative (TN)
Sandstone	234	130	27	410
Shale	363	31	139	268
Coal	30	13	8	750

The evaluation results for VISA-9 indicate that Sandstone has the highest Recall (0.90), indicating that the SVM effectively identified nearly all sandstones. Nevertheless, the precision for Sandstone is only 0.64, suggesting that there are still numerous false positives (falsely identifying shale and anthracite as Sandstone). Coal has a precision of 0.7 and a recall of 0.74, suggesting that the prediction is still missing some coal.

Table 12. Performance Parameters Result in VISA-9 Train Dataset Experiment 2

Performance			
Accuracy	Precision	Recall	F1 Score
0.78	0.64	0.9	0.75
	0.92	0.72	0.81
	0.7	0.79	0.74

Sandstone has the highest Recall (0.90) in the VISA-9 evaluation results, indicating that the SVM effectively identified nearly all sandstones. On the other hand, the precision for Sandstone is only 0.64, suggesting that there are still numerous false positives (misclassification of shale and lignite as

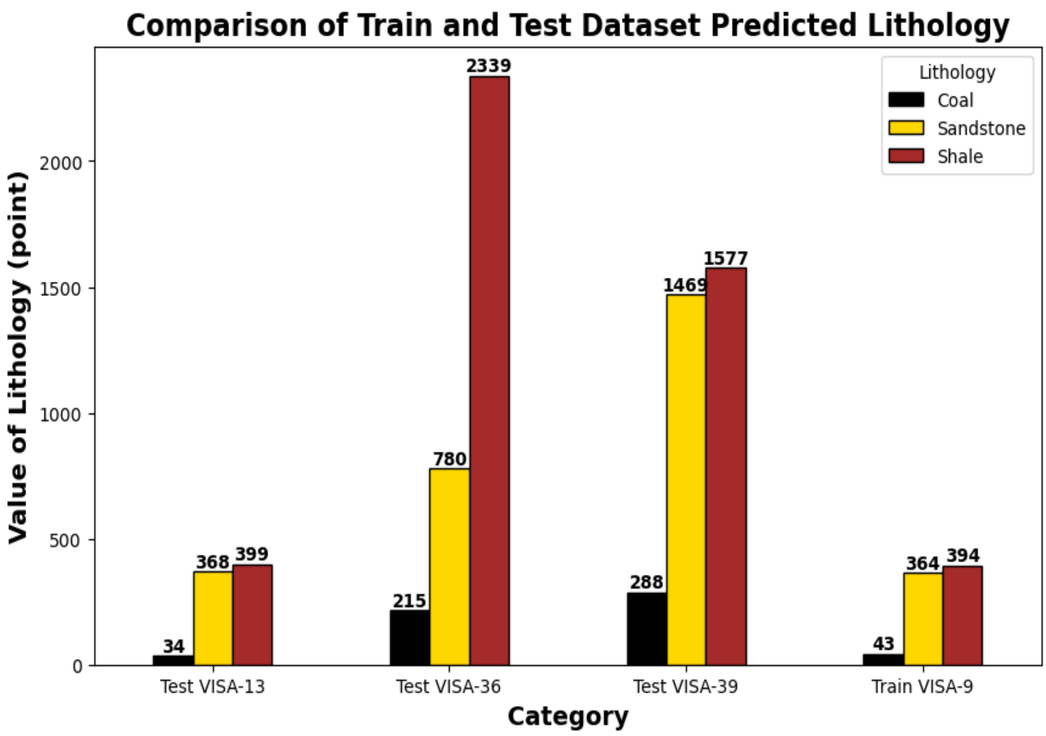

Figure 7. Comparison of Lithology Prediction Results of Train and Test Datasets in Experiment 1

Sandstone). Coal's precision is 0.7, and its Recall is 0.74, suggesting that the prediction is still missing some coal.

Table 13. Lithology Prediction Results in VISA-36 Train Dataset Experiment 2

Lithology	Confusion matrix			
	True positive (TP)	False positive (FP)	False negative (FN)	True negative (TN)
Sandstone	941	32	349	2012
Shale	1819	332	110	1073
Coal	110	100	5	3119

Table 14. Performance Parameters Result in VISA-36 Train Dataset Experiment 2

Performance			
Accuracy	Precision	Recall	F1 Score
0.86	0.97	0.73	0.83
	0.85	0.94	0.89
	0.52	0.96	0.68

The SVM is accurate in predicting Sandstone in VISA-36, but it still misses many sandstones, as evidenced by the very high precision (0.97) and lesser Recall (0.73). Shale's Recall is 0.94, suggesting that it is more consistent in its predictions. Coal's precision of 0.68 suggests that there is still a significant amount of undetected Coal. However, the SVM's high Recall of 0.96 suggests that it is more cautious in its predictions of Coal.

The distribution of VISA-9 and VISA-36 lithology is compared in Figure 8 to the results of SVM prediction and the actual data. In VISA-9, the sandstone lithology is actually 261 and expected to be 364 (+39.5% over-prediction). The shale lithology is actually 502 and expected to be 394 (-21.5% under-prediction).

The coal lithology is actually 38 and expected to be 43. In VISA-36, the sandstone lithology is actually 1929 and anticipated 2339, which is a 21.3% overestimate. The shale has an actual value of 1290 and a predicted value of 973, which is an underestimate. The Coal has an actual value of 115 and a predicted value of 210. It is evident that the model has a propensity to over-predict Coal and Sandstone, while shale has a propensity to under-predict, particularly in VISA-9. This suggests that the characteristics of shale and Sandstone are related and can lead to misclassification, while tiny quantities of Coal are frequently concealed by other lithologies.

Testing dataset experiment 2

Table 15 contains the confusion matrix metrics and the performance evaluation of the SVM in predicting Sandstone, shale, and Coal, all of which are included in the evaluation of the lithology prediction results for the VISA-13 well. The SVM demonstrated a high level of accuracy, with an aggregate score of 87%, suggesting that it was capable of accurately generalizing the lithology patterns in this test dataset. The SVM achieved a high recall of 0.90 for sandstone lithology, indicating that the model accurately identified 90% of the Sandstone in the actual data.

Table 15. Lithology Prediction Results in VISA-13 Test Dataset Experiment 2

Lithology	Confusion matrix			
	True positive (TP)	False positive (FP)	False negative (FN)	True negative (TN)
Sandstone	310	65	33	393
Shale	373	33	67	328
Coal	18	2	0	781

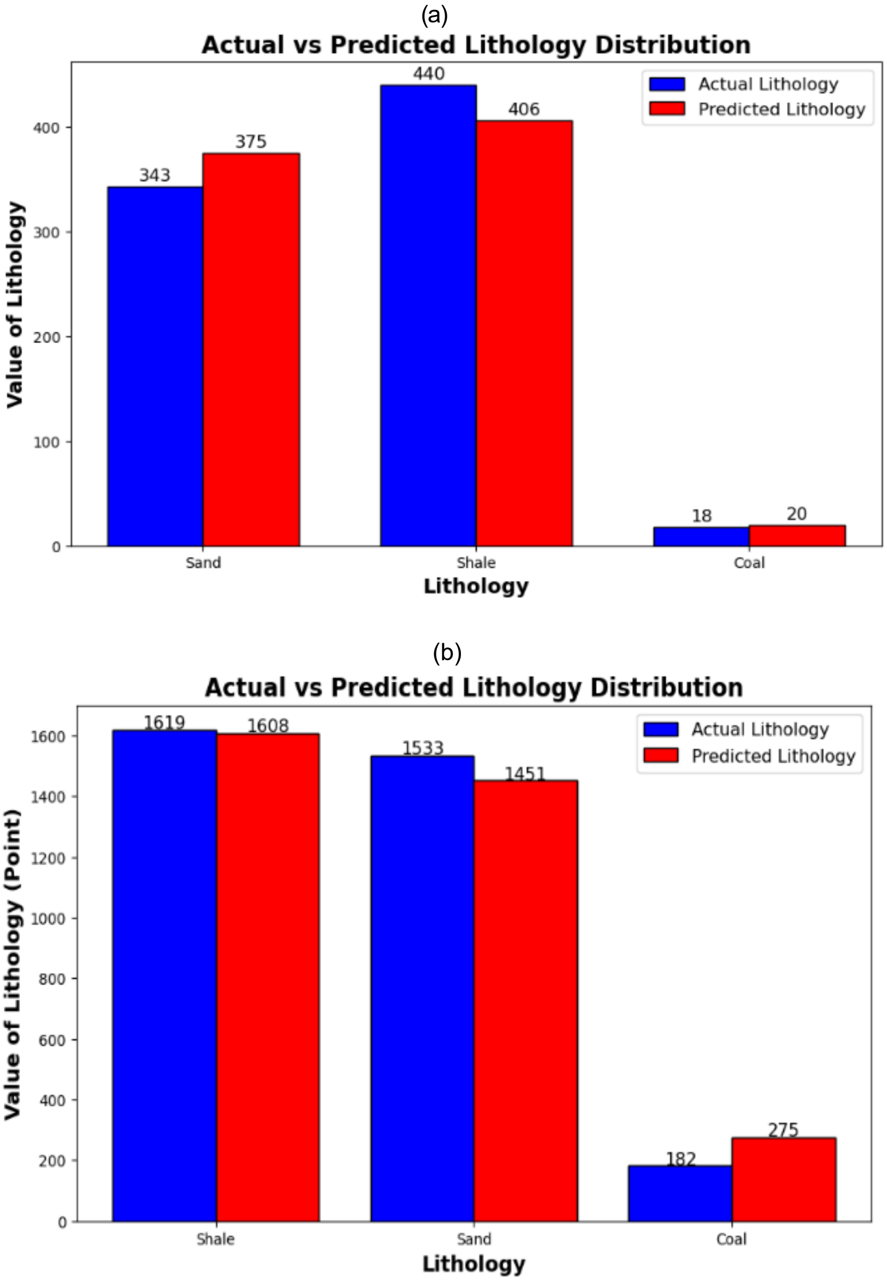

Figure 8. Histogram of Actual vs Predicted Lithology Distribution in Wells (a) VISA-9 (b) VISA-36

Table 16. Performance Parameters Result in VISA-13 Test Dataset Experiment 2

Performance			
Accuracy	Precision	Recall	F1 Score
0.87	0.83	0.9	0.86
	0.92	0.85	0.88
	0.9	1	0.95

Nevertheless, the precision for Sandstone was only 0.83, suggesting that 17% of all sandstone predictions were false positives, as shale or Coal was incorrectly classified as Sandstone. The SVM was capable of reliably identifying the majority of the shale in the dataset, as evidenced by its precision of 0.92 and Recall of 0.85 for shale lithology. Nevertheless, 67 false negatives were identified, indicating that some shale was misclassified as Sandstone or Coal. SVM demonstrated satisfactory performance in coal lithology, with a precision of 0.9 and a false positive value of only 2.

Table 17. Lithology Prediction Results in VISA-39 Test Dataset Experiment 2

Lithology	Confusion matrix			
	True positive (TP)	False positive (FP)	False negative (FN)	True negative (TN)
Sandstone	1488	120	131	1595
Shale	1336	115	197	1686
Coal	165	110	17	3042

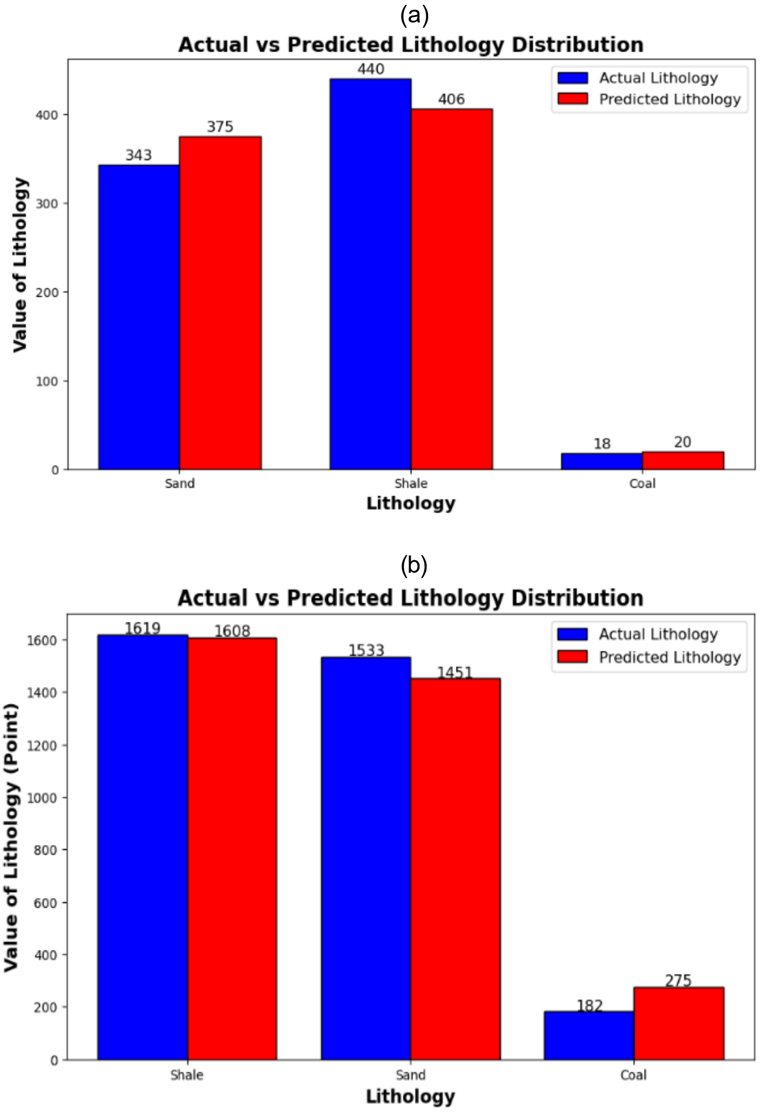

Figure 9. Histogram of Actual vs Predicted Lithology Distribution in Wells (a) VISA-13 (b) VISA-39

The results in Table 17 and Table 18 indicate an overall accuracy of 89%, which is higher than that of the VISA-13 well. This suggests that the SVM has superior generalization capabilities in this particular well, despite the fact that there are still some misclassifications, particularly in the case of Coal. The SVM demonstrated a precision of 0.93 and a recall of 0.92 for sandstone lithology, suggesting that it was capable of accurately identifying Sandstone with only a few false positives and false negatives.

Table 18. Performance Parameters Result in VISA-39 Test Dataset Experiment 2

Performance			
Accuracy	Precision	Recall	F1 Score
0.89	0.93	0.92	0.92
	0.92	0.87	0.9
	0.6	0.91	0.72

The model also demonstrated satisfactory performance in the field of shale lithology, with a precision of 0.92 and a recall of 0.87. This suggests that the SVM was highly accurate in identifying shale; however, there were still 197 false negatives, indicating that some shale was misclassified as Sandstone or Coal. The SVM was unable to predict the coal well in VISA-39 accurately. The SVM's precision was 0.6 and its Recall was 0.91, suggesting that it still overlooked some coal in the dataset. Seventeen false negatives were detected, suggesting that some coal was incorrectly identified as shale or Sandstone.

The lithology distribution histograms for the VISA-13 and VISA-39 wells illustrate the contrast between the actual data and the predictions generated by the Support Vector Machine (SVM) model. In these histograms, blue represents the actual lithology distribution, while red represents the distribution based on model projections. The distribution for the VISA-13 well indicates that the model has a propensity to over-predict Coal and Sandstone while under-predicting shale. In comparison, the VISA-39 well exhibits a broader spectrum of discrepancies between the predicted and actual data. However, the model still demonstrates over-prediction for Coal in VISA- 39, albeit to a lesser extent than in the VISA-13 well.

The efficacy of the SVM in predicting lithology was assessed through a comparative analysis of the training and testing datasets. This assessment involved comparing the lithology prediction results between the sets while examining accuracy, F1-score, recall, and precision, all of which were

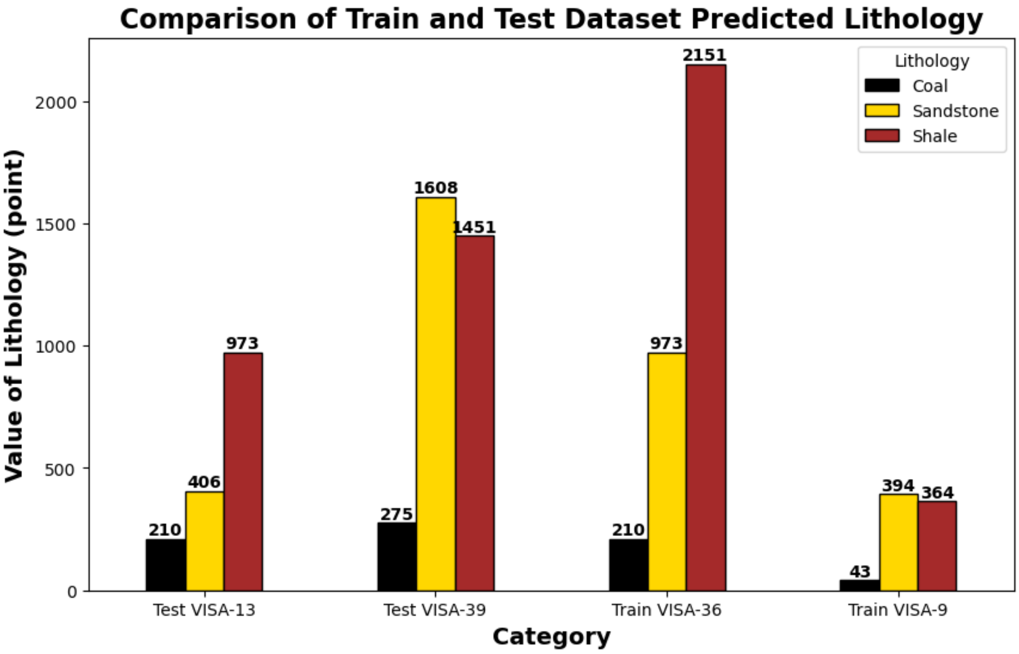

Figure 10. Comparison of Lithology Prediction Results of Train and Test Dataset in Experiment 2

represented in histograms (Figure 10). The SVM did not experience overfitting, as the accuracy, F1-score, and Recall were generally higher on the testing dataset than on the training dataset.

Furthermore, the accuracy and F1-score for Sandstone and shale remained high across both training and testing, suggesting that the SVM was stable in identifying these lithologies. However, the precision and Recall for Coal remained lower than those of other lithologies particularly in the VISA-36 dataset indicating that the model continues to struggle with differentiating Coal from other rock types.

Results of training and testing experiment 3 training dataset experiment 3

In Table 19, SVM predicts 304 sandstone points, with 68 false positives (FP) and 39 false negatives (FN). Consequently, the actual number of sandstone points is 343 points, while the predicted number is 372 points, resulting in a 29- point discrepancy, suggesting a modest overprediction. The actual number of points for shale was 440, while the SVM predicted 405 points, resulting in an underprediction of 35 points. This was due to the fact that there were 366 true positives (TP) and 74 false negatives (FN). In contrast, Coal had a total of 18 actual points, while the model predicted 24 points, with 6 false positives. This suggests that the SVM slightly overpredicted Coal (Table. 20).

Table 19. Lithology Prediction Results in VISA-13 Train Dataset Experiment 3

Lithology	Confusion matrix			
	True positive (TP)	False positive (FP)	False negative (FN)	True negative (TN)
Sandstone	304	68	39	390
Shale	366	39	74	322
Coal	18	6	0	777

Table 20. Performance Parameters Result in VISA-13 Train Dataset Experiment 3

Performance			
Accuracy	Precision	Recall	F1 Score
0.85	0.82	0.89	0.85
	0.90	0.83	0.87
	0.75	1	0.86

Table 21 illustrates a more consistent accuracy on Sandstone, with an actual value of 1,290 points and a model prediction of 1,270 points, a mere 20- point margin of error. Nevertheless, the model's prediction of 1,850 samples was significantly lower than the actual number of 1,929 samples, resulting in a 79-point underprediction of shale. In the case of Coal, an error occurred as the model predicted 214 samples, despite the actual number being 115 (Table. 22). This resulted in an overprediction.

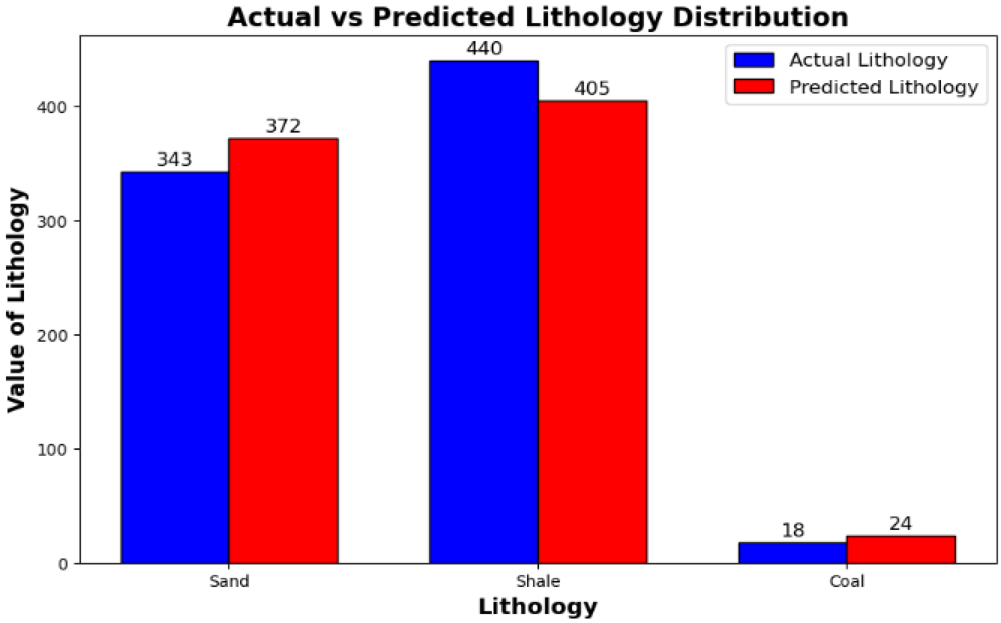

Figure 11. Histogram of Actual vs Predicted Lithology Distribution in VISA-13

Table 21. Lithology Prediction Results in VISA-36 Train Dataset Experiment 3

Lithology	Confusion matrix			
	True positive (TP)	False positive (FP)	False negative (FN)	True negative (TN)
Sandstone	1144	126	146	1918
Shale	1720	130	209	1275
Coal	110	104	5	3115

Table 22. Performance Parameters Result in VISA-36 Train Dataset Experiment 3

Performance			
Accuracy	Precision	Recall	F1 Score
0.89	0.90	0.89	0.89
	0.93	0.89	0.91
	0.51	0.96	0.67

The actual (1.619) and predicted (1.617) counts are only 2 points apart in Table 23, indicating that the model performs better on Sandstone. Nevertheless, shale continued to experience underprediction by 70 points, while Coal again experienced overprediction by 72 points (Table 24). This suggests a consistent pattern of model errors in coal prediction.

Table 23. Lithology Prediction Results in VISA-39 Train Dataset Experiment 3

Lithology	Confusion matrix			
	True positive (TP)	False positive (FP)	False negative (FN)	True negative (TN)
Sandstone	1514	103	105	1612
Shale	1373	90	160	1711
Coal	176	78	6	3074

Table 24. Performance Parameters Result in VISA-39 Train Dataset Experiment 3

Performance			
Accuracy	Precision	Recall	F1 Score
0.91	0.94	0.94	0.94
	0.94	0.90	0.92
	0.69	0.97	0.81

Testing dataset experiment 3

The lithology prediction model for the VISA- 9 well, as demonstrated by the analysis of Table 25, exhibits an exceptional level of performance, with an overall accuracy of 0.95 (95%). Despite the fact that the Coal class had the fewest samples (38

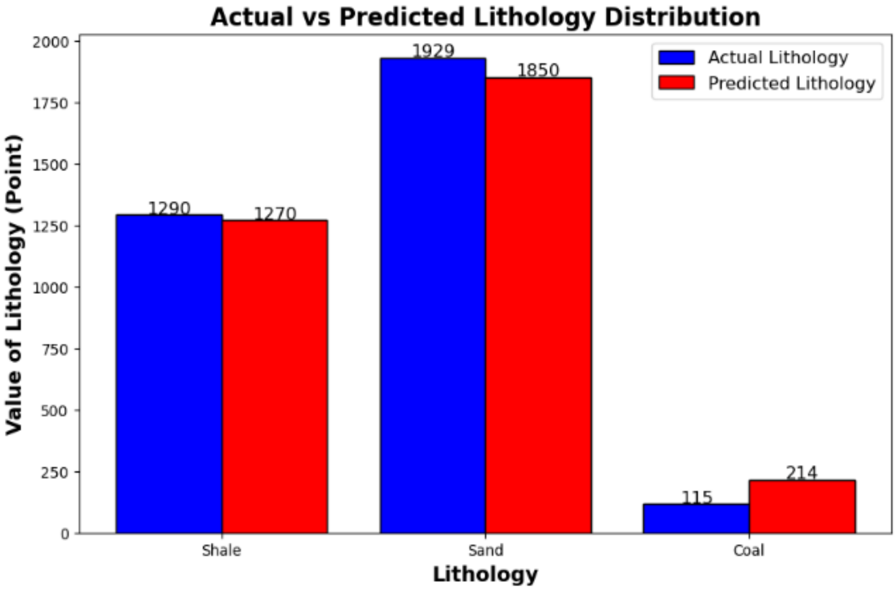

Figure 12. Histogram of Actual vs Predicted Lithology Distribution in VISA-36

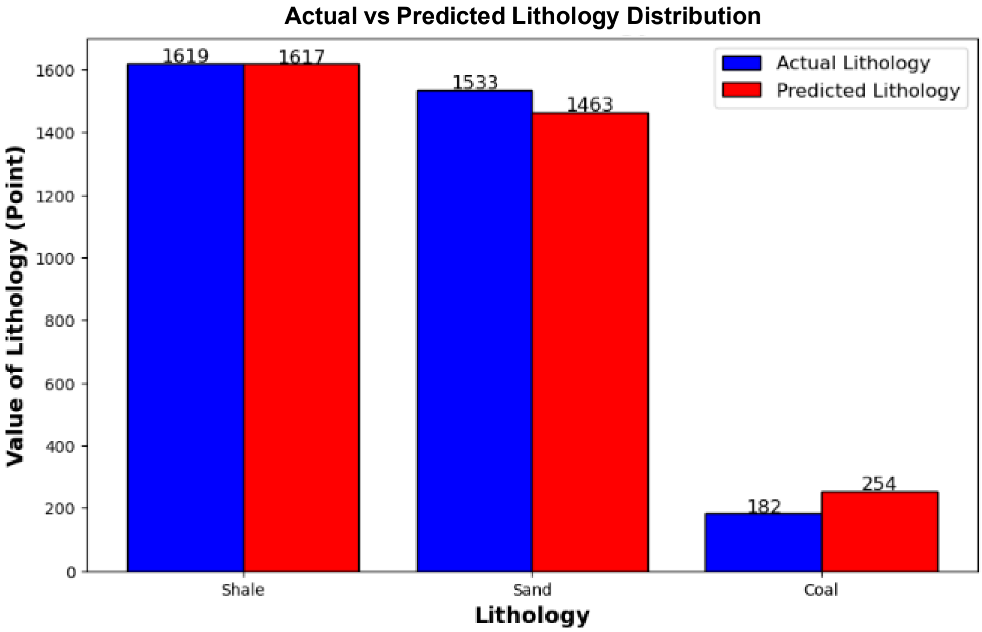

Figure 13. Histogram of Actual vs Predicted Lithology Distribution in VISA-39

actual samples), it was able to accomplish a Recall of 1 (or 100%). This was the most exceptional, nearly flawless performance. This indicates that the model accurately identified all coal samples without any false negatives (0 False Negative) and with a precision of 0.97 (only 1 False Positive), resulting in the highest F1 Score of 0.99 (Table 26). The model exhibits an exceptional precision of 0.99 for the Shale class, which is the predominant lithology, indicating that it is nearly certain to be accurate when it predicts "Shale" (502 actual samples).

Table 25. Lithology Prediction Results in VISA-9 Test Dataset Experiment 3

Lithology	Confusion matrix			
	True positive (TP)	False positive (FP)	False negative (FN)	True negative (TN)
Sandstone	253	23	8	517
Shale	479	7	23	292
Coal	38	1	0	762

Nevertheless, the Recall is 0.95, indicating that the model "missed" 23 Shale samples (23 False Negatives), which is consistent with the underprediction remark in the accompanying text. The Sandstone class, which consisted of 261 actual samples, also demonstrated exceptional performance, with a Recall of 0.97 (Table 26). This indicates that it was highly dependable in identifying nearly all Sandstone, with only 8 False Negatives. Nevertheless, its Precision of 0.92 is the lowest of the three, as a result of 23 False Positives (FP).

Table 26. Performance Parameters Result in VISA-9 Train Dataset Experiment 3

Performance			
Accuracy	Precision	Recall	F1 Score
0.95	0.92	0.97	0.94
	0.99	0.95	0.97
	0.97	1	0.99

The 23 FP figure for Sandstone is highly likely to be associated with the 23 FN for Shale. This suggests that the primary source of confusion for the model is the misclassification of certain Shale as Sandstone, resulting in a slight overprediction of Sandstone. VISA-9 is the sole testing well, while the other three wells (VISA-13, VISA-36, and VISA-39) are included in the training dataset. This information provides a profound and significant understanding of the SVM model's behavior. It is evident that Sandstone and Shale lithologies are the predominant classes in all wells, regardless of whether they are in training or testing. Nevertheless, the primary emphasis of this analysis is on the minority class,

specifically Coal. In Figure 15, the presence of "significant overprediction" in the training dataset is specifically emphasized in the accompanying data. This assertion is visually and unequivocally corroborated by the graph: the model predicts extremely high and relatively consistent quantities of Coal in all three training wells, including 210 points in VISA-13, 214 points in VISA-36, and even 283 points in VISA-39. This implies that the SVM model may have "learned" to be either too sensitive or too aggressive in its identification of signals that may be associated with Coal during the learning process, resulting in a significant overestimation of their number in the training data.

Nevertheless, in Figure 16, the most significant and advantageous aspect of this analysis is the stark contrast observed in the testing well. The model did not exhibit this severe overprediction behavior when it was presented with entirely novel, previously unseen data (the VISA-9 Test). The model, in contrast, only predicted 39 points for Coal.

By integrating this information with the data from Table 4.16 above, it is evident that the 39-point prediction is not only a small number, but also a highly accurate result (with 38 actual samples, 1.0 Recall, and 0.97 Precision). This is an ideal situation, as it suggests that the model has demonstrated excellent generalization capabilities and is robust, despite the fact that it may have developed some bias or indications of overfitting on the minority class during the training phase (as

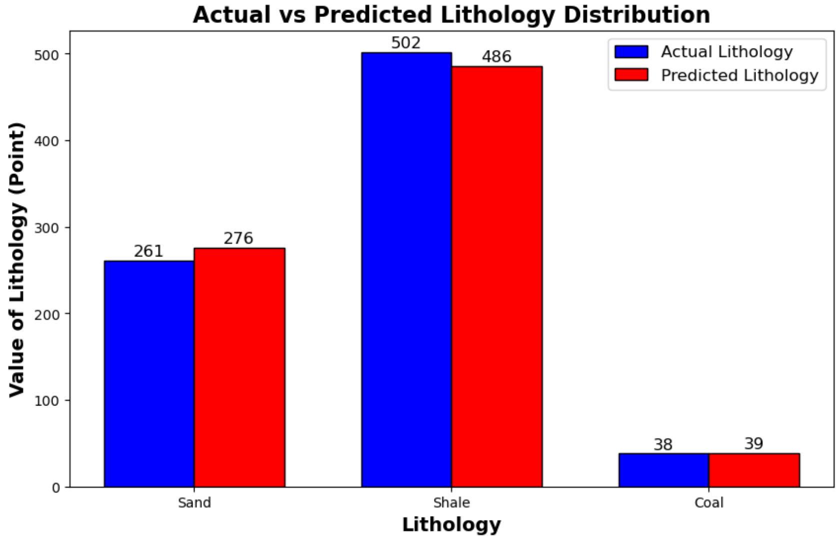

Figure 14. Histogram of Actual vs Predicted Lithology Distribution in VISA-9

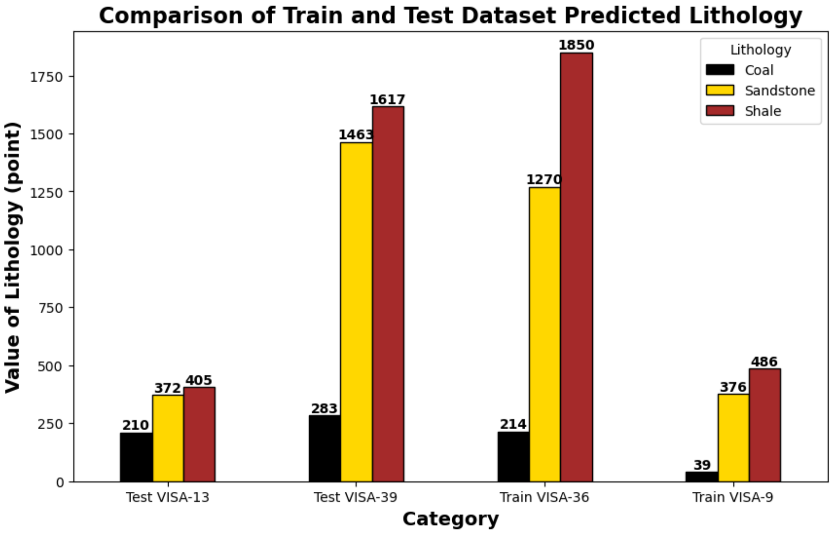

Figure 15. Comparison of Lithology Prediction Results of Train and Test Dataset in Experiment 3

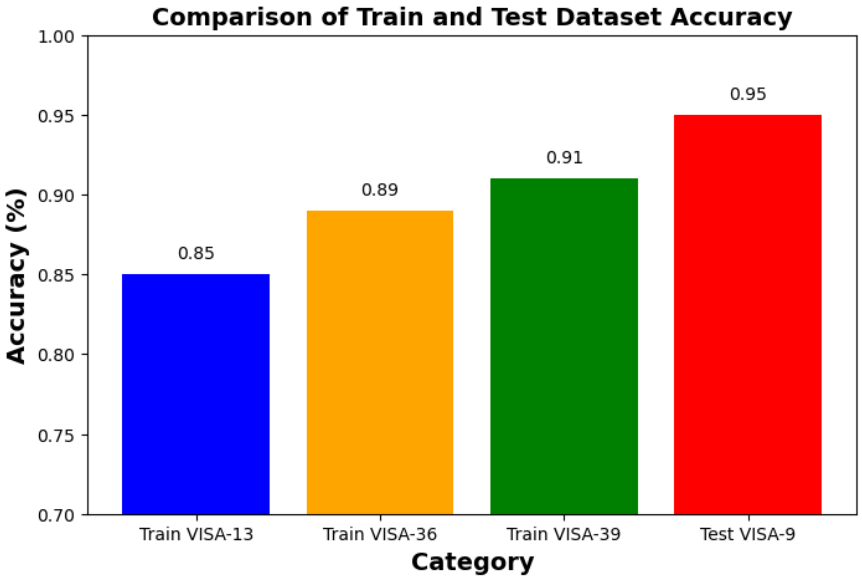

Figure 16. Histogram Comparing Training and Testing Accuracy in Experiment 3

indicated by the 200+ points figure). The model effectively "repeated its mistakes" and accurately identified much smaller and realistic amounts of Coal in new fields, thereby demonstrating its efficacy and reliability for practical applications (Figure 16).

CONCLUSION

Based on the research conducted, the Support Vector Machine (SVM) algorithm can be effectively deployed to predict lithology in the VISA-9, VISA-13, VISA-36, and VISA-39 wells. This prediction is predicated on well log parameters including GR, DT, NPHI, ILD, LLD, PHIN, and RHOB. The findings indicate that SVM performance is significantly influenced by the ratio between the training and testing datasets. In Experiments 1 and 2, the performance was suboptimal, with prediction error levels ranging from 11% to 22% relative to the actual lithology. However, the third experiment, which utilized a training data ratio of 75% and a test data ratio of 25%, yielded substantial improvements. In this configuration, the prediction error was reduced to only 5%, and the performance value increased by 7% compared to the previous experiments. Furthermore, Experiments 1, 2, and 3 demonstrated that the use of either adjacent or distant wells did not result in a significant difference in evaluation results, with only a 1–2% variance observed.

ACKNOWLEDGEMENT

The authors gratefully acknowledge financial support from the Institut Teknologi Sepuluh Nopember for this work, under project scheme of the Publication Writing and IPR Incentive Program (PPHKI) 2025.

GLOSSARY OF TERMS AND SYMBOLS

Terms & Symbol	Definition	Unit
C	Regularisation Parameter	
Dataset	Structured Collection Of Data Used To Train, Validate, And Evaluate An Algorithm	
F1-Score	Machine Learning Evaluation Metric That Measures A Model's Accuracy	
FN	The Number Of Instances Where The Model Incorrectly Predicts The Absence Of The Target Class When It Is Actually Present.	
FP	This Is Statistically Referred To As A Type I Error.	
Precision	Measures The Accuracy Of A Model Positive Predictions	

RBF	Radial Basis Function
Recall	Metric That Measures A Classification Model Ability To Identify All Relevant Instances Of The Positive Class Within A Dataset
SVM	Support Vector Machine
TN	The Number Of Instances Where The Model Correctly Identifies The Absence Of The Target Class
TP	The Predicted Outcome And The Actual Data Both Indicate The Positive Condition.

REFERENCES

Abhimantra, S. (2021). Geologi Dan Studi Sikuen Stratigrafi Formasi Balikpapan, Lapangan "Minggiran "Cekungan Kutai, Kalimantan Timur. Jurnal Ilmiah Geologi PANGEA, 2(2).

Asquith, G. B., Krygowski, D., & Gibson, C. R. (2004). Basic well log analysis (Vol. 16, pp. 305 -371). Tulsa: American Association of Petroleum Geologists.

Augusto, F. D. O. A., & Martins, J. L. (2009). A well- log regression analysis for P-wave velocity prediction in the namorado oil field, Campos basin. Revista Brasileira de Geofisica, 27, 595- 608.

Al Ghaithi, A., & Prasad, M. (2020). Machine learning with artificial neural networks for shear log predictions in the Volve field Norwegian North Sea. In SEG Technical Program Expanded Abstracts 2020 (pp. 450-454). Society of Exploration Geophysicists.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2), 121-167.

Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.

Battu, D. P. (2026). Characterization Of Deltaic Source Rocks And Hydrocarbon Potential In The Lower Kutai Basin. Scientific Contributions Oil and Gas, 49(1), 55-67.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297.

Doust, H., & Noble, R. A. (2008). Petroleum systems of Indonesia. Marine and Petroleum Geology, 25(2), 103-129.

Ellis, D. V., & Singer, J. M. (2007). Well logging for earth scientists (Vol. 692). Dordrecht: Springer.

Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.".

Guo, G., Diaz, M. A., Paz, F., Smalley, J., & Waninger, E. A. (2005). Rock typing as an effective tool for permeability and water-saturation modeling: a case study in a clastic reservoir in the Oriente Basin. In SPE Annual Technical Conference and Exhibition? (pp. SPE-97033). SPE.

Horsfall, O. I., Omubo-Pepple, V. B., & Tamunobereton-ari, I. (2013). Correlation analysis between sonic and density logs for porosity determination in the south-eastern part of the Niger Delta Basin of Nigeria. Asian Journal of Science and Technology, 4(1), 1-5.

Husein, S. (2015). Petroleum and Regional Geology of Northeast Java Basin, Indonesia-Excursion Guide Book for Universiti Teknologi Petronas Malaysia. Department of Geological Engineering Universitas Gadjah Mada.

Hidayat, H., Setiawan, J. J. H., Ibrahim, A., Marjiyono, M., & Junursyah, G. L. (2021). Studi Magnetotelurik (MT) untuk Mendelineasi Potensi Regional Gas Serpih Bawah Permukaan Berdasarkan Properti Tahanan Jenis di Cekungan Kutai, Kalimantan Timur. Jurnal Geologi dan Sumberdaya Mineral, 22(2), 107-114.

Handoyo, F., Fourier, D. E. L., Reza, R., & Harnanti, Y. P. (2018). Estimasi Parameter Fisis Batuan Berdasarkan Citra Batuan (Digital Rock Physics) Studi Kasus: Lapangan Minyak Bumi Blok Cepu, Jawa Tengah, Indonesia. Jurnal Geofisika, 16(01), 21-26.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

Khawarizmy, A., Haryanto, I., Ganjar, R. M., & Firmansyah, Y. (2020). Asies Pengendapan dan Porositas Batuan Karbonat Formasi Kujung I Cekungan Jawa Timur Utara Lapangan "Tengah". Geoscience Journal, 4(1), 61-67.

Killeen, P. G. (1982). Gamma-ray logging and interpretation. In Developments in Geophysical Exploration Methods—3 (pp. 95-150). Dordrecht: Springer Netherlands.

Lunt, P. (2019). The origin of the East Java Sea basins deduced from sequence stratigraphy. Marine and Petroleum Geology, 105, 17-31.

Moherek, A., Mukherjee*, S., Garrison, N., Caraway, A., Medina, R., & Sarmah, B. B. (2016, August). Application of High-Resolution Blended Resistivity Measurement in Characterization of Unconventional Plays. In Unconventional Resources Technology Conference, San Antonio, Texas, 1-3 August 2016 (pp. 2706-2717). Society of Exploration Geophysicists, American Association of Petroleum Geologists, Society of Petroleum Engineers.

Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9(1), 381-386.

Munadi, S. (2007). The Effect Of Oil Content On Sonic Wave Propagation (Analyses from Well Log Data). Scientific Contributions Oil and Gas, 30(2), 33-37. https://doi.org/10.29017/SCOG.30.2.983

Passey, Q. R., Bohacs, K. M., Esch, W. L., Klimentidis, R., & Sinha, S. (2010, June). From oil-prone source rock to gas-producing shale reservoir–geologic and petrophysical characterization of unconventional shale-gas reservoirs. In SPE International Oil and Gas Conference and Exhibition in China (pp. SPE- 131350). SPE.

Peng, J., Jury, E. C., Dönnes, P., & Ciurtin, C. (2021). Machine learning techniques for personalized medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges. Frontiers in pharmacology, 12, 720694.

Qi, Q., Fu, L. Y., Deng, J., & Cao, J. (2021). Attenuation methods for quantifying gas saturation in organic-rich shale and tight gas formations. Geophysics, 86(2), D65-D75.

Serra, O. E. (1983). Fundamentals of well-log interpretation.

Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.

Sunarjanto, D., Suliantara, S., Iskandar, U. P., & Nainggolan, M. T. (2014). Sistem Informasi Geogra untuk Optimasi Eksplorasi dan Pengembangan Wilayah Migas Geographic Information System for Optimization Exploration Oil and Gas Area Development. Lembaran publikasi minyak dan gas bumi, 48(1), 1-12. https://doi.org/10.29017/LPMGB.48.1.225

Sebtosheikh, M. A., & Salehi, A. (2015). Lithology prediction by support vector classifiers using inverted seismic attributes data and petrophysical logs as a new approach and investigation of training data set size effect on its performance in a heterogeneous carbonate reservoir. Journal of Petroleum Science and Engineering, 134, 143-149.

Sukma, R. F. N., Rosid, M. S., & Wijanarko, E. (2025). Telisa Formation Characterization Using Seismic Acoustic Impedance Inversion in the Akasia Area of the Central Sumatra Basin. Scientific Contributions Oil and Gas, 48(2), 29-40. doi.org/10.29017/scog.v48i2.1774

Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K. L. A., Elkhatib, Y., ... & Al-Fuqaha, A. (2019). Unsupervised machine learning for networking: Techniques, applications and research challenges. IEEE access, 7, 65579-65615.

Wardhana, S. G., Pakpahan, H. J., Simarmata, K., Pranowo, W., & Purba, H. (2021). Algoritma komputasi machine learning untuk aplikasi prediksi nilai total organic carbon (TOC). Lembaran publikasi minyak dan gas bumi, 55 (2), 75-87.

Zamri, N., Pairan, M. A., Azman, W. N. A. W., Abas, S. S., Abdullah, L., Naim, S., ... & Gao, M. (2022). A comparison of unsupervised and supervised machine learning algorithms to predict water pollutions. Procedia Computer Science, 204, 172-179.

Zaemi, F. F., Rohmana, R. C., & Atmoko, W. (2022). Uncovering The Potential of Low

Resistivity Reservoirs Through Integrated Analysis: A Case Study from The Talang Akar Formation in The South Sumatra Basin. Scientific Contributions Oil and Gas, 45(3), 169-181. https://doi.org/10.29017/SCOG.45.3.1258