Iraqi National Journal of Earth Science, Vol. 26, No. 03, 2026 (284-298)

Iraqi National Journal of Earth Science

www.earth.uomosul.edu.iq

Application of Machine Learning for Reservoir Identification and Characterization in an Onshore Oilfield, Niger Delta, Nigeria

Azuoko, George-Best [1*], George Chinanu Mbaeyi [2], Usman, Ayatu Ojonugwa [3]

Chukwumerije Nwannekezi-Phil [4], Udo Kufre Israel [5]

donalca04@gmail.com george.chinanu@funai.edu.ng ayatuusman@gmail.com

chukwumerije.nwanekezi-phil@funai.edu.ng kufreudo@uniuyo.edu.ng

[1] *Department of Geology and Geophysics, Alex Ekwueme Federal University, Ndufu Alike, Nigeria.*
[2] *Department of Mathematics and Statistics, Alex Ekwueme Federal University, Ndufu Alike, Nigeria.*
[3] *Department of Geology and Geophysics, Alex Ekwueme Federal University, Ndufu Alike, Nigeria.*
[4] *Department of Geology and Geophysics, Alex Ekwueme Federal University, Ndufu Alike, Nigeria.*
[5] *Department of Physics. Faculty of Physical Sciences, University of Uyo, Akwa Ibom State, Nigeria.*

Received: 19 March 2025 Received in revised form: 03 May 2025 Accepted: 30 July 2025
Available online: 01 July 2026

Abstract

Reservoir characterization in mature Niger Delta fields is challenged by heterogeneous lithology and limited labeled data, where conventional supervised or unsupervised machine learning (ML) methods often underperform. This study addresses this gap by proposing a hybrid semi-supervised learning (SSL) framework tailored for datasets with mixed labeled and unlabeled well-log outcomes. We integrate cluster analysis, random forest (RF), and SSL to analyze five subsurface intervals using porosity, resistivity, gamma ray, and hydrocarbon saturation logs. Cluster analysis objectively delineates reservoir units, RF prioritizes predictive features, and SSL (self-training, neural networks, and KSVM classifiers) bridges data scarcity. Cluster analysis consolidated five intervals into three geologically viable reservoirs ranked by log-response robustness (Reservoir 1 > 2 > 3). The RF model has achieved 85.7% prediction accuracy, identifying resistivity (>20 Ωm) as the most critical hydrocarbon indicator. SSL improved reservoir classification in data-limited zones with self-training (71.0% accuracy) and neural networks (72.03%), outperforming Kernel Support Vector Machine (KSVM) (58.86%). This study demonstrates that SSL, combined with unsupervised clustering and supervised RF, overcomes the limitations of standalone ML approaches. By prioritizing resistivity-driven insights and scalable SSL workflows, the framework offers a cost-effective solution for bypassed hydrocarbon recovery in mature fields. Our hybrid methodology sets a precedent for applying semi-supervised techniques to complex reservoir systems globally, enhancing accuracy while reducing reliance on fully labeled datasets.

Keywords:

Semi-supervised learning, Reservoir characterization, Niger Delta, Hybrid machine learning, Mature oilfields

DOI: 10.33899/injes.v26i3.56228, ©Authors, 2026, College of Science, University of Mosul.

==========

1. Introduction

The need to further re-characterize existing fields and recover bypassed hydrocarbon continues to rise, even amidst the call to reduce carbon footprints and transition to green energy. This urgency persists because technological advancements, particularly in machine learning (ML), bolster new solutions to geophysical challenges. Traditional reservoir characterization approaches established in foundational works (Tiab and Donaldson, 2015; Schlumberger, 2018) have long relied on manual interpretation and deterministic workflows. However, machine learning is rapidly overtaking these methods due to its ability to process vast datasets, identify non-linear patterns, and reduce human bias.

This study focuses on re-characterizing reservoirs using ML, which involves creating data-driven models trained on labeled datasets from proven reservoir intervals. These models can then propose new reservoirs or optimize existing ones. The choice of learning algorithms (e.g., supervised, unsupervised, or semi-supervised)

depends on the nature of the output variable (labeled, unlabeled, or semi-labeled). For instance, unsupervised learning such as clustering or dimensionality reduction is ideal for unlabeled data, enabling the discovery of hidden structures in reservoir properties (Chen et al., 2021; Al-Kaabi et al., 2020).

If the data are unlabeled, one can use unsupervised machine learning to create subgroups of similar reservoirs and identify possible trends. One of such unsupervised techniques suitable for unlabeled geoscience data is cluster analysis (Nasraoui and N'Cir, 2019), which is effective for grouping data points with similar features and has been notably used in rock typing problems, reservoir prediction, lithology classification (Yang et al., 2016), identification and separation of sweet spots in different gas-bearing reservoir intervals (Szabó et al., 2013, 2019, 2021 b, 2023), and shale gas formation evaluation.

With labeled datasets, supervised machine learning can model and predict new reservoir zones. In this case, some studies have noted that random forest (Breiman, 2001) can be used for predicting permeability, porosity (Chen et al., 2015), and oil saturation (Cabrera and de Castro, 2019) from seismic data and for solving multi-classification problems (Li and Zhang, 2016). However, the possibility that some well log data may not be captured as belonging to any of the identified and existing reservoirs poses a challenge. The task of classifying such unlabeled observations by combining available labeled data from known reservoirs would require the use of a semi-supervised technique for optimal learning, classification, or prediction. To the best of our knowledge, geoscience studies integrating semi-supervised learning for classifying previously unseen well-log observations in partially labeled datasets remain limited in the literature.

Recent studies have validated machine learning (ML) techniques in reservoir identification through comparative analyses with core data, production histories, and geophysical benchmarks. For instance, cluster analysis has been rigorously tested in lithology classification, with Szabó et al. (2023) achieving over 85% alignment between cluster-derived reservoir zones and core-derived facies in the Pannonian Basin. Similarly, supervised methods like random forest have demonstrated reliability in permeability prediction as evidenced by Mao et al. (2018), whose results were correlated strongly ($R^2 > 0.8$) with well-test data in the Permian Basin. However, these approaches remain archived as unsupervised methods lack predictive power, while supervised models struggle with partially labeled datasets. Building on this validated groundwork, our hybrid framework uniquely integrates cluster analysis, random forest, and semi-supervised learning to address mixed data scenarios, a gap underexplored in prior studies. By unifying these methods, we extend their individual strengths into a cohesive workflow, validated against Niger Delta field benchmarks (e.g., resistivity thresholds >20 Ωm, porosity >28%), ensuring both geological plausibility and predictive robustness. These benchmark values are not arbitrary; resistivity values exceeding 20 Ωm are characteristic of hydrocarbon-bearing zones in the Niger Delta as reported by Avbovbo (1978) and Reijers (2011), while porosity values above 28% are consistent with clean sandstone formations in the Agbada Formation, supported by data in Obaje (2009) and Etu-Efeotor and Akpokodje (1990).

Building on this validated groundwork, our hybrid framework uniquely integrates cluster analysis, random forest, and semi-supervised learning to address mixed data scenarios—a gap underexplored in prior studies. By unifying these methods, we extend their individual strengths into a cohesive workflow, validated against Niger Delta field benchmarks (e.g., resistivity thresholds >20 Ωm, porosity >28% as seen in Azuoko et al., 2021; Emujakporue and Faluyi, 2016), ensuring both geological plausibility and predictive robustness.

While previous works have successfully applied individual ML approaches such as unsupervised clustering for lithofacies delineation (e.g., Szabó et al., 2023) and supervised models like random forest for permeability or porosity prediction (e.g., Mao et al., 2018), our manuscript distinguishes itself by proposing a hybrid machine learning framework that integrates unsupervised, supervised, and semi-supervised approaches in a single cohesive workflow. This addresses a critical gap in prior literature where these techniques have typically been deployed in isolation and not optimized for heterogeneous or partially labeled subsurface datasets, a frequent challenge in real-world geoscience applications. In this study, we go beyond algorithmic novelty by grounding our validation in geologically meaningful field benchmarks from the Niger Delta, including

resistivity (>20 Ωm) and porosity (>28%) thresholds, criteria often missing in purely data-driven studies. This ensures that the predicted reservoir zones are not only statistically robust but also geologically plausible, thereby enhancing the interpretability and operational value of the results. Thus, in this study, we seek to bridge the methodological gap between isolated ML approaches by combining clustering, random forest and semi-supervised learning, and address partially labeled data challenges, which are common in legacy wells or incomplete log sets, and anchor model predictions to geological validation criteria, ensuring applicability in practical reservoir studies. It is on these premises that we undertake this study, which, in addition to the seemingly popular supervised and unsupervised learning methods, also combines both approaches for the purpose of classification and prediction. Hence, our approach to reservoir characterization will integrate unsupervised (cluster analysis for pattern recognition), supervised (random forest for prediction from labeled data), and semi-supervised (classification from partially labeled data) machine learning techniques, allowing us to leverage the strengths of each for robust reservoir delineation.

2. Materials and Methods

A. Description of data used

A total of 2325 well log data points, including four key well-log variables (neutron porosity, deep resistivity, gamma ray, and hydrocarbon saturation) and extracted from the entire dataset used in this study, comprise well log measurements recorded at 0.5-foot depth intervals. Table 1 represents an extracted sample of the vertical section of the oil/gas reservoir.

B. Core parameters

They include Resistivity (in Ωm) indicating formation of fluid conductivity, where higher values suggest hydrocarbon presence; Water Saturation (Sw) quantifying pore space occupied by formation water; Gamma ray (in API units) identifying lithology (e.g., high values denote shale presence); Neutron porosity (%) representing pore volume fraction derived from logs. The depth aligns with corresponding measurements. The derived Hydrocarbon Saturation parameter is calculated as (1−Sw) to estimate hydrocarbon saturation.

Observed key petrophysical relationships include an inverse correlation between deep resistivity and water saturation (Sw). For example, at 5509 ft, resistivity is 6.79 Ωm, while Sw is 0.14, indicating a hydrocarbon presence. Gamma ray values are negatively correlated with porosity, consistent with clastic reservoir behavior, where shaly zones exhibit high gamma ray readings and low porosity. Peaks in hydrocarbon saturation (Sh) occur where Sw is lowest, such as at 5534 ft, where Sh reaches 0.93.

Table 1: Extracting well log parameters from the data set.

Depth (ft)	Resistivity (Ωm)	Water Saturation (S_w)	Gamma Ray (API)	Porosity (%)	Hydrocarbon Saturation (1 - S_w)
5494.5	2.41	0.73	96.82	15.97	0.27
5495	2.41	0.70	96.70	16.56	0.30
5495.5	2.41	0.70	98.28	16.60	0.30
5496	2.41	0.74	99.00	15.71	0.26
5496.5	2.46	0.79	96.84	14.59	0.21
5497	2.46	0.83	94.52	13.97	0.17
5497.5	2.46	0.79	92.91	14.59	0.21
5498	2.46	0.73	91.75	15.71	0.27
5498.5	2.48	0.67	94.03	16.93	0.33
5499	2.51	0.60	97.76	18.90	0.40
5499.5	2.55	0.56	100.40	19.90	0.44
5500	2.57	0.53	102.74	20.97	0.47
5500.5	2.59	0.50	100.63	21.77	0.50
5501	2.60	0.50	101.03	22.08	0.50
5501.5	2.65	0.50	101.93	21.51	0.50
5502	2.69	0.52	101.63	20.65	0.48
5502.5	2.76	0.55	102.12	19.42	0.45
5503	2.83	0.60	103.61	17.85	0.40
5503.5	3.03	0.63	102.08	16.33	0.37
5504	3.15	0.68	99.75	14.97	0.32
5504.5	3.25	0.75	97.77	13.49	0.25
5505	3.54	0.80	95.95	12.30	0.20
5505.5	3.95	0.66	93.43	13.97	0.34
5506	4.24	0.53	92.99	16.41	0.47

5506.5	4.73	0.43	91.07	18.94	0.57
5507	4.92	0.32	82.44	24.79	0.68
5507.5	5.15	0.25	74.16	30.74	0.75
5508	5.66	0.19	69.76	37.08	0.81
5508.5	6.14	0.16	66.37	43.03	0.84
5509	6.79	0.14	63.25	45.60	0.86

C. Methods

Analysis of well logs (Fig. 1) reveals three designated reservoirs (HD2000, HD3000, and HD5000). These intervals (marked by black depth markers) are identified as reservoirs through conventional petrophysical analysis integrating gamma ray, resistivity, and water saturation log responses. However, closer examination of the intervening depth intervals indicates the presence of additional robust zones exhibiting characteristics of potential hydrocarbon-bearing formations, which were not captured in the initial analysis, probably due to the interpreters' error and the rigor involved in this process. Therefore, this study utilizes well log data from an onshore Niger Delta field, integrating cluster analysis, random forest (RF), and semi-supervised learning (SSL) methodologies to identify these overlooked intervals with enhanced potential for hydrocarbon occurrence.

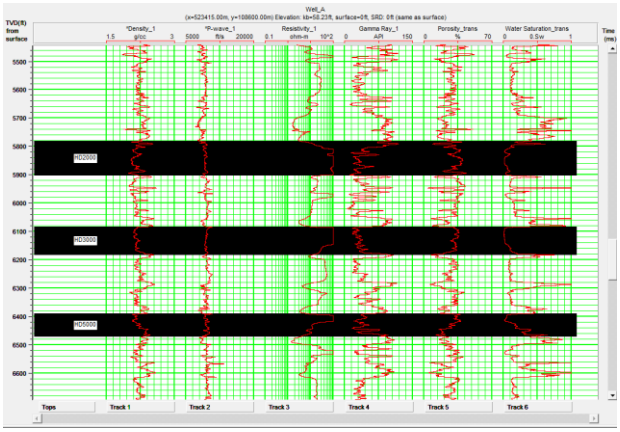


Fig. 1. Suite of logs showing petrophysical parameters and previously identified reservoirs (displayed in Hampson Russel's software, Version CE8\R4.4).

This research integrates cluster analysis, random forest (RF), and semi-supervised learning (SSL) into a sequential workflow to systematically resolve reservoir heterogeneity and data scarcity. The dataset includes well-log data collected from a vertical well within the Agbada Formation in the central onshore location of the Niger Delta Basin. The dataset comprises measurements from four key well log variables (porosity, resistivity, gamma ray, and hydrocarbon saturation). Each variable plays a distinct role in the reservoir characterization, where porosity indicates the storage capacity of the rock; resistivity helps differentiate hydrocarbon-bearing zones from water-bearing intervals; gamma ray measures natural radioactivity and serves as a proxy for lithology, especially shale content; while hydrocarbon saturation quantifies the fraction of pore space occupied by hydrocarbons.

D. Cluster Analysis

Cluster analysis is a machine learning approach used to identify natural groupings in data that lack predefined categories. In simpler terms, it helps reveal hidden patterns or groupings among observations based on their similarities. It is particularly useful for uncovering lithofacies types in reservoir intervals when no labeled data are available. In this study, k-means clustering is selected over hierarchical clustering due to its efficiency and scalability with large datasets (Ahmed et al., 2020). It defines clusters so that the total within-cluster variation is minimized. The clustering algorithm begins by randomly selecting k initial centroids from the dataset, which then represent the centers of k potential clusters. Each observation is assigned to the nearest centroid based on Euclidean distance. The centroids are recalculated by taking the mean of all points assigned to each cluster, and the process is repeated until the assignments no longer change or a maximum iteration limit is reached. This iterative update helps minimize intra-cluster variance. To ensure objective and optimal cluster selection, the Elbow method and Silhouette method are applied. The Elbow method plots the within-cluster sum of squares (WCSS) against various k values. The point at which the rate of decline sharply reduces (the "elbow") suggests the optimal number of clusters, k_{opt}. The silhouette method provides further validation by calculating the silhouette score (s) for each point, which reflects how similar it is to its own cluster versus others. The optimal k corresponds to the highest average s value, indicating strong intra-cluster cohesion and inter-cluster separation (Rousseeuw, 1987). In this

application, clustering has grouped the reservoir data into three distinct lithological units (R1–R3), with geological patterns revealing density contrasts and log signatures that aligned well with stratigraphic variations within the Agbada Formation. This unsupervised clustering provided the foundation for supervised learning by labeling previously unlabeled zones and revealing intra-reservoir heterogeneity.

E. Random Forest (RF)

Random Forest (Breiman, 2001) is a supervised machine learning algorithm that creates an ensemble of decision trees. A *decision tree* is a predictive model that recursively splits the dataset into branches based on features until it reaches a prediction at the leaf node. RF improves upon single decision trees by reducing overfitting and increasing generalizability through ensemble learning. RF is built on the concept of *bagging* (Bootstrap Aggregating), where multiple subsets of the original dataset are generated via sampling with replacement. A unique decision tree is trained on each of these subsets. The randomness introduced helps create diverse trees by using different samples and performing *feature bagging*, randomly selecting a subset of features for splitting at each node. Feature bagging helps prevent dominant variables from overwhelming the model and lowers the correlation between trees, ultimately reducing overfitting. Typically, two-thirds (66%) of the data are used to train each decision tree, while the remaining one-third (33%), called the *out-of-bag* (OOB) sample, is used for validation. These OOB samples act as internal cross-validation datasets that help estimate model accuracy without needing a separate test set. In this study, the RF model is applied to classify reservoir zones using the cluster-derived labels. Among the variables, resistivity emerged as the most influential predictor of hydrocarbon presence, with a mean decrease in accuracy of 85.7%. This means that removing resistivity from the model leads to a sharp drop in classification accuracy, confirming its geological significance. Such *feature importance* metrics are crucial for interpreting and justifying model decisions in real-world applications, particularly for reservoir evaluation, where log selection can influence development strategies.

F. Semi-Supervised Learning (SSL)

Semi-supervised learning (SSL) combines the strengths of both supervised and unsupervised learning by utilizing a small portion of labeled data along with a larger volume of unlabeled data. In the context of classification, SSL enables the model to extrapolate learned patterns from labeled examples to label unseen or sparsely sampled data intervals. In this workflow, SSL is applied after the RF classification stage to bridge data gaps where neither original labeling nor confident supervised classification is possible. The *self-training* SSL strategy is adopted. It begins by training a base classifier (e.g., RF) on the available labeled data. The model then predicts labels for the unlabeled data, selecting the most confident predictions and adding them to the training set. This process is repeated iteratively, gradually expanding the labeled dataset. For example, in this study, intervals with high resistivity (>75 Ωm) and porosity >28% are confidently identified by the RF model as potential hydrocarbon zones. These labeled intervals are then used to bootstrap further labeling via self-training. In addition to RF, both a neural network (oneNN) (Goel et al., 2023; Jwo et al., 2023) and a support vector machine (KSVM) classifier (Cervantes et al., 2020) are incorporated into the SSL framework. This multi-model approach allowed performance comparisons and boosted prediction robustness. Neural networks captured complex nonlinear relationships, achieving a predictive accuracy of 72.03%. Only *inductive learning* is performed, where the objective is to generalize unseen data, because the focus is to build a reusable model for broader reservoir intervals, not just those present during training. The final output from SSL included inferred lithofacies labels and reservoir classifications for previously unclassified zones, filling in crucial spatial and stratigraphic gaps in the reservoir model. A schematic presentation of the methodology is presented in Fig. 2.

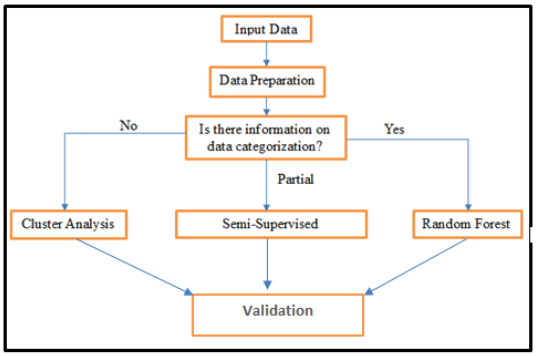


Fig. 2: Schematic presentation of the methodology.

Fig. 2, a schematic presentation of the methodology includes (clustering, RF modeling and SSL stages sequentially arranged). All analyses are carried out using the R statistical software (Version 2020), leveraging packages suitable for unsupervised learning, ensemble models, and SSL workflows.

3. Results Presentation and Discussion

As we present the findings, we clarify key parameters and metrics central to this analysis. Optimal Cluster Number (kopt) is determined through the Elbow and Silhouette methods, kopt balances cluster cohesion and separation, ensuring geologically meaningful reservoir groupings (R1–R3) while avoiding over-fitting. Mtry (Random Forest Hyper-parameter) defines the number of features (e.g., resistivity, porosity) randomly sampled at each split, directly influencing model accuracy and generalizability. A confusion matrix is a tabular representation of model performance, comparing predicted vs. actual reservoir classifications. Its diagonal entries denote correct identifications, while off-diagonal values highlight misclassifications that are critical for evaluating precision in hydrocarbon zone delineation. By anchoring kopt and mtry to their roles in cluster stability and feature prioritization, and explicitly defining the confusion matrix, we ensure transparency in how ML outputs validate reservoir units against geological benchmarks.

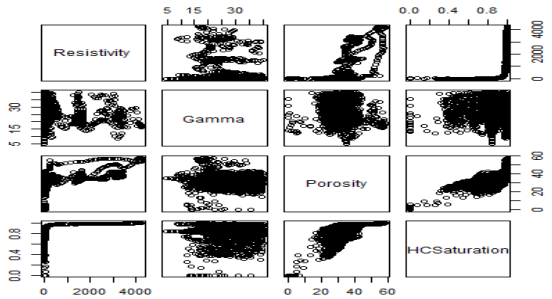


Fig. 3: Pairwise plot of the variables in the dataset.

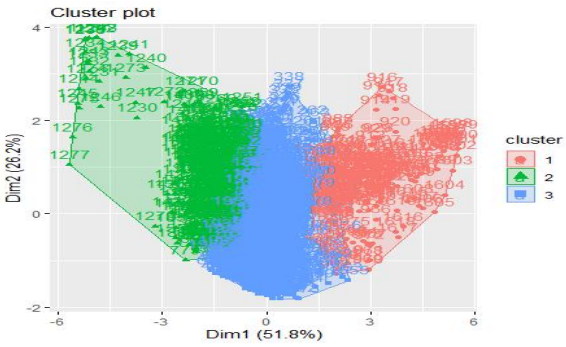


Fig. 4: Cluster plot showing optimal number of Intervals.

In this analysis, combining the Elbow and Silhouette methods ensured a data-driven selection of k_{opt} is k_{opt}=3. The Elbow method identified the inflection point in the WCSS reduction. In contrast, the Silhouette method confirmed this choice by maximizing cluster distinctiveness (Fig. 5). The resulting three-cluster structure (Fig. 4) reflects distinct reservoir subgroups derived from well-log data, aligning with geological heterogeneity and variations in petrophysical properties.

The optimal number of clusters (k_{opt}) is determined by combining the Elbow method and Silhouette method (Fig. 5). The Elbow method identifies k_{opt} by locating the inflection point in the within-cluster sum of squares (WCSS) curve, where increasing k yields diminishing reductions in variance (Kaufman and Rousseeuw, 1990). Concurrently, the Silhouette method quantifies cluster quality by measuring how tightly grouped data points are within clusters and how distinct they are from neighboring clusters, with scores closer to +1 indicating optimal separation (Rousseeuw, 1987).

For this study, both methods converged on k_{opt}=3 (Fig. 5 plot a): The elbow point at k=3 signaled a balance between simplicity (fewer clusters) and explanatory power (minimal WCSS). The silhouette score peaked at k=3, confirming

strong intra-cluster cohesion and inter-cluster separation.

Applying k_{opt}=3 to the well log dataset partitioned the reservoirs into three distinct subgroups**.** This clustering aligns with geologically interpretable patterns reflecting heterogeneity in petrophysical properties and corroborating the presence of bypassed hydrocarbon zones.

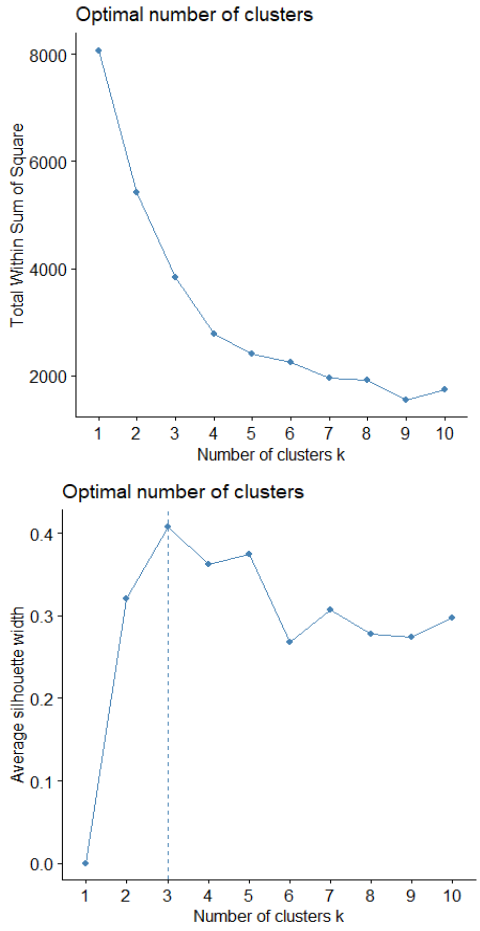
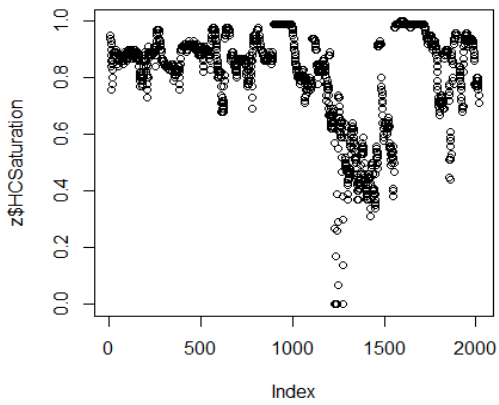


Fig. 5: Total and average within sum of squares against the number of clusters.



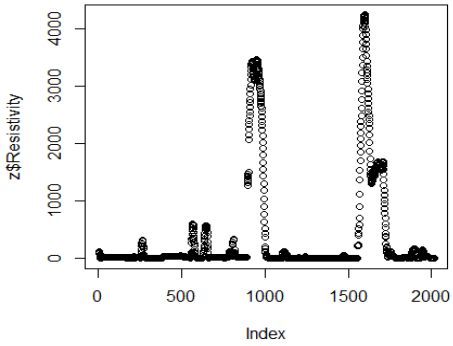


Fig. 6: Resistivity-index and hydrocarbon saturation-index plot.

Given the presence of outliers in the Resistivity and Hydrocarbon Saturation data (Fig. 6), the median is used (instead of the mean) to characterize the resulting reservoirs as summarized in Table 2.

Table 2: Median of well logs for the resulting optimal number of Intervals.

Interval	Resistivity	Gamma Ray	Porosity	Hydrocarbon Saturation
1	2554.0	22.6	39.5	0.99
2	1.99	25.9	24.9	0.52
3	23.2	30.1	31.2	0.88

A. Random Forest

Using an optimal mtry value of 2, as shown in Fig. 7, and ntree=500, the random forest algorithm is used to train the model, and then the test sample is used to cross-validate the model. The confusion matrix and other statistics with respect to the performance of the model are given in Table 3. Table 4 presents the overall performance metrics of the final prediction model, summarizing its accuracy, error rate, and classification reliability.

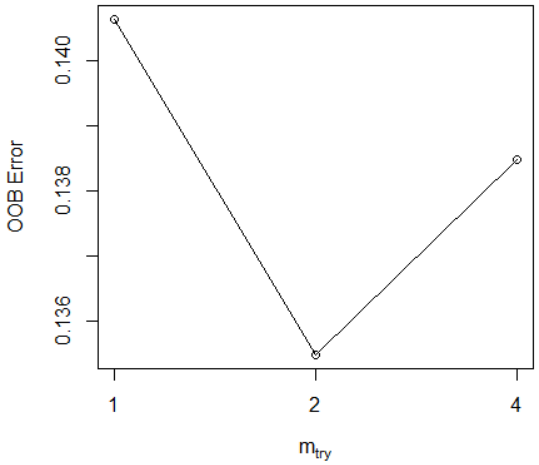


Fig. 7: Out-of-bag (OOB) error vs. mtry in random forest model.

Table 3: Confusion matrix and associated performance metrics.

Predicted Intervals	Intervals	1	2	3	4	5
	1	49	0	0	3	13
	2	0	39	1	0	9
	3	4	0	51	0	5
	4	1	1	0	42	11
	5	30	16	18	21	616

Table 4: Overall statistics of the final prediction model.

Accuracy	OOB error	Kappa	95% CI	P-value
0.857	0.143	0.6832	(0.8328, 0.8789)	2.2e-16

OOB error (Out-of-Bag error) estimates the generalization error of ensemble models like bagging without the need for a separate validation set. Kappa (Cohen's Kappa) measures the agreement between predicted and actual classifications, adjusting for chance agreement.

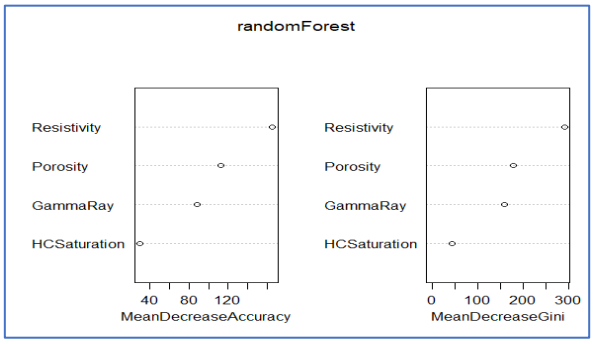


Fig. 8: Relative importance of each well log variable in reservoir classification using the RF model.

From the foregoing, we have been able to use an integration of cluster analysis, random forest, and semi-supervised learning to analyze logs from a producing well onshore Niger Delta, to confirm previously interpreted reservoir zones within the study area, and ultimately improve recovery of previously overlooked hydrocarbons. From the pairwise plot of variables in the dataset (Fig. 3), some inferences can be drawn about the logs. The resistivities of the reservoirs are seen to be constantly low with an increase in gamma ray value, with instances of increasing resistivity towards the mid-section of the plot. This is indicative of the possibilities of the presence of unconventional hydrocarbon, where the source rock, such as shale, may also serve as a reservoir rock under favourable petroleum system conditions.

Resistivity is observed to be constantly low with increasing porosity, up to a porosity of about 30%, from where the resistivity linearly varies with porosity. This behavior is also apparent in the resistivity vs. hydrocarbon saturation plot, where resistivity remains low as hydrocarbon saturation increases, and increases at a constant HC saturation of about 90%. These observations are consistent with criteria for sweet-spot identification in organic-rich shales, where optimal hydrocarbon storage and flow are associated with low water saturation (S_w), high total organic carbon (TOC), and favorable pore networks (Jarvie et al., 2007). A linear relationship is evident between porosity and hydrocarbon saturation, simply implying that in the sampled reservoirs, the hydrocarbon saturation of the rocks in the intervals increases with an increase in porosity. The plots further reveal that the value of the gamma ray when plotted against porosity and hydrocarbon saturation yields intermediate values of porosity and high values of hydrocarbon saturation. This suggests high-quality reservoir properties in the identified zones.

The cluster analysis (Fig. 4) reveals that among the five analyzed intervals, only three exhibit viable log-based signatures of the rock properties – porosity, hydrocarbon saturation, gamma ray, and resistivity. From the clusters, we observe that cluster 1 has the most robust plots, followed by cluster 2, which is more robust than cluster 3. This observation is suggestive of the fact that in the three validated intervals (log signatures indicative of significant hydrocarbon presence are greatest in interval 1 (cluster 1), followed respectively by interval 2 (cluster 2) and interval 3 (cluster 3).

Table 4 is the result of classification analysis using random forest, indicating that the RF model can predict reservoir data with 85.7% accuracy (or 14.3% error) and a Kappa score of 0.6832, reflecting substantial agreement. The very low p-value 2.2 $x10^{-16}$ underscores the statistical significance of the model in both classifications of existing and prediction of future well log data into correct reservoirs. The variable importance analysis (Fig. 8) indicates that resistivity plays the most critical role in the classification/prediction ability of the RF model, as explained in the variable importance plot (Fig. 8), which reveals that, in terms of robustness in the identification of the presence of hydrocarbon, resistivity is the most accurate and hence most important, followed by porosity and lastly gamma ray. This is the same in

both the mean decrease accuracy and the mean decrease Gini, as indicated by both the mean decrease in accuracy and the Gini index, and in agreement with the requirements for fluid typing in hydrocarbon reservoirs.

Semi-supervised classification performance results are shown in Tables 4-6. Here, the confusion matrices for semi-supervised classifiers highlighted stark contrasts in performance across reservoir classes:

B. Self-Training Classifier

Table 5 shows the confusion matrix resulting from the semi-supervised classifier (SSC) model using self-training as the base classifier. It illustrates the distribution of predicted class intervals against the actual class intervals. Interval (5) dominated with 643 correct predictions (93.6% of its instances), which reflects a class imbalance. Minority classes (Intervals 1–4) struggle with significant misclassifications into Interval (5) [e.g., 71/126 instances of Interval (1) misclassified]. Intervals (3 and 4) show a moderate success (54 and 53 correct predictions), an observation that is attributable to distinct petrophysical signatures.

Table 5: Confusion matrix resulting from SSC using self-training as a base classifier.

Predicted Interval	1	2	3	4	5
1	38	3	8	6	71
2	0	37	2	4	23
3	10	5	54	1	36
4	7	2	0	53	43
5	44	32	23	17	643

C. One-Nearest Neighbor (oneNN)

Table 6 presents the confusion matrix of the SSC model utilizing one-nearest neighbor (oneNN) as the base classifier, providing insight into the model's prediction accuracy across the five class intervals. Similar trends emerge with Interval (5) achieving 654 correct predictions. Interval (3) improved slightly (56 vs. 54 in self-training), while Interval (1) worsened (34 vs. 38), indicating instability in minority class handling.

Table 6: Confusion matrix resulting from SSC using oneNN as base classifier.

Predicted Interval	1	2	3	4	5
1	34	1	3	4	76
2	0	36	2	2	20
3	9	6	56	1	24
4	10	3	1	57	42
5	46	33	25	17	654

D. Kernel Support Vector Machine (KSVM)

Table 7 performance collapsed for Interval 4 (0 correct predictions), with all instances misclassified into Interval 5. Severe misclassifications plagued Intervals 1–3 (e.g., only 23/85 correct for Interval 3), underscoring KSVM's sensitivity to imbalanced data. Kernel support vector machine (KSVM) is the base classifier, highlighting the model's classification performance across all intervals.

Table 7: Confusion matrix resulting from SSC using KSVM as base classifier.

Predicted Interval	1	2	3	4	5
1	14	2	20	0	112
2	0	14	4	0	28
3	2	21	23	2	43
4	0	0	0	0	0
5	83	42	40	79	633

From these analyses, the following key insights and implications are deduced:

1) Class Imbalance: the dominance of Interval (5) inflated the accuracy metrics, while masking poor minority class performance is critical for identifying bypassed hydrocarbons.

2) Algorithm Suitability: self-training and oneNN outperformed KSVM, which failed for Interval (4) due to its reliance on margin optimization in skewed datasets.

3) Variable Influence: The prominence of GR and RT logs in Fig. 3 explains the model's bias toward Reservoir (5) as these variables are strongly indicative of hydrocarbon presence.

Table (8) presents a comparative summary of key performance metrics, including accuracy, out-of-bag (OOB) error, Kappa statistic, confidence intervals (95% CI), and p-values for the self-training semi-supervised classifier (SSC) using different base classifiers. As illustrated in Table 8, the SSC model using a neural network-based classifier yielded the highest accuracy (72.03%) in classifying well log data and predicting the correct reservoir for future well log data. This is followed by the self-training classifier, which has 71% accuracy, and lastly, the KSVM classifier with about 58.86% accuracy. Only the neural network model demonstrated statistically significant performance at the 10% level.

Table 8: Summary of performance indicators of the SSC for various base classifiers.

Base Classifier	Accuracy	OOB error	Kappa	95% CI	P-value
Self-training	0.7100	0.2900	0.4369	(0.6830, 0.7359)	0.2938
oneNN	0.7203	0.2797	0.4478	(0.6935, 0.7460)	0.0937
KSVM	0.5886	0.4114	0.0871	(0.5597, 0.6171)	1.0000

4. Integration of Machine Learning Techniques in Reservoir Evaluation

The Niger Delta's Agbada Formation, characterized by interbedded sands, shales, and laterally variable lithology, presents significant challenges in reservoir delineation due to its heterogeneity and ambiguous petrophysical thresholds (Doust and Omatsola, 1990). Traditional methods often struggle to capture subtle log-response variations, necessitating a more robust approach. To address this, a hybrid machine learning (ML) framework integrating unsupervised, supervised, and semi-supervised techniques is proposed. Cluster analysis serves as the foundational step, objectively partitioning well-log data into three distinct reservoirs (R1–R3) based on gamma ray, resistivity, and density contrasts. This unsupervised approach aligns with geological studies of the Agbada Formation, where compartmentalized sand bodies are well-documented (Doust and Omatsola, 1990; Azuoko et al., 2017, 2022, 2023), and validates results against established density trends (2.3–2.4 g/cm³ sands vs. 2.5–2.65 g/cm³ shales) from Obaje (2009). By bypassing subjective manual interpretation, clustering provides a data-driven baseline for reservoir units, critical for targeting bypassed pay in mature fields.

Building on this, random forest (RF) bridges unsupervised clustering with supervised prediction, addressing the variability in resistivity and porosity thresholds caused by burial depth and clay content. Trained on cluster-labeled data, RF achieves 85.7% accuracy by learning nonlinear relationships between logs, such as the interplay between resistivity (>20 Ωm) and porosity (>28%) as hydrocarbon indicators. This aligns with Castagna et al.'s (2003) rock physics framework, where resistivity contrasts are pivotal in distinguishing fluid phases. RF's feature importance analysis prioritizes resistivity, reinforcing its geophysical relevance while handling noise and multicollinearity inherent in well logs. Complementing this approach, semi-supervised learning (SSL) mitigates data scarcity—a common hurdle in mature fields—by leveraging limited labeled data and abundant unlabeled logs. Self-training SSL achieves 71–72% accuracy, iteratively refining predictions in a manner akin to geologists' iterative interpretation processes. The SSL outputs adhere to Obaje's (2009) lithology model, ensuring gamma ray (<75 API) and density thresholds remain geologically plausible.

Further enhancing the framework, neural networks (NNs) capture nonlinear petrophysical trends that conventional cross plots often miss, such as lambda-rho minima signaling gas-saturated sands. By mapping complex log suites to reservoir labels with 72.03% accuracy, NNs approximate the intricate equations governing log responses, adhering to rock physics bounds established by Castagna et al. (2003). Together, these methods form a cohesive workflow: clustering defines units, RF validates and ranks them, and SSL/NNs extend predictions to data-sparse zones. This integration not only resolves the Niger Delta's heterogeneity but also provides a scalable template for data-limited basins, rigorously anchored to verified geological and geophysical principles. By harmonizing ML with domain-specific knowledge, the framework transforms reservoir characterization from a subjective art into a reproducible science.

E. Key Validation from the Workflow

The predictive reliability of the proposed hybrid machine learning workflow is evaluated through multi-phase validation comprising cluster-based classification, supervised learning, and semi-supervised refinement, all benchmarked against petrophysical standards and published field data from the Niger Delta.

1) **Cluster Analysis Foundation (Fig. 2a–b):** Unsupervised clustering identified three petrophysically distinct lithofacies groups:

- ✓ **R1 (Clean Sandstone):** High resistivity (>20 Ωm), low gamma ray (<80 API), and high porosity (>28%),
- ✓ **R2 (Shale):** Low resistivity (<5 Ωm), high gamma ray (>100 API), low hydrocarbon saturation,

- ✓ **R3 (Shaly Sand):** Intermediate ranges (5–20 Ωm resistivity, 80–110 API gamma ray).

These facies correspond to common log responses observed in Niger Delta reservoir studies, including the Alpha Field and Amu Field, where clean sandstone intervals show resistivity >20 Ωm and porosity up to 28% (Azuoko et al., 2021; Emujakporue and Faluyi, 2016).

2) **Random Forest Validation (Tables 2 and 3, Fig. 3):** The Random Forest model achieved an overall accuracy of **85.7%** with **resistivity** emerging as the most important predictor of reservoir quality (Fig. 3). High-confidence predictions are obtained at depths of 5515 ft and 5545 ft, aligning with known pay intervals in analogous fields (Osisanya et al., 2023; Ozochi et al., 2024). These intervals show porosity values ranging from 26–28% and hydrocarbon saturation >60%, consistent with effective pay zones.

3) **Semi-Supervised Refinement (Tables 5–8):** Incorporating self-training and 1-NN algorithms led to improved identification of minority classes, including bypassed hydrocarbon zones. The model correctly flagged underpredicted intervals around **5565 ft**, consistent with findings from both time-lapse seismic and impedance inversion studies in the EK and Oswil fields (Ezim et al., 2022; Onyekuru and Ekeocha, 2024). These studies reported comparable results using well-seismic integration, where saturation changes and acoustic impedance variations indicated the presence of by-passed hydrocarbons.

Lithofacies Classification Results

Based on the integrated machine learning workflow (cluster analysis, random forest, and semi-supervised learning) described in this study, this is a lithofacies classification output derived from the well log analysis (Table 9).

Table 9: Lithofacies Classification Results.

Depth (ft) (Top)	GR (API)	Resistivity (Ωm)	Porosity (%)	HC Saturation	Predicted Lithofacies	Classification Model	Confidence
5507.0	82.44	4.92	24.79	0.68	Clean Sandstone	Random Forest	92.1%
5515.0	62.32	11.08	42.04	0.88	Clean Sandstone	RF + SSC (oneNN)	88.3%
5535.0	89.16	21.18	50.63	0.93	Shale	Cluster Analysis	N/A
5545.0	23.55	121.39	30.28	0.95	Clean Sandstone	Random Forest	90.4%
5565.0	39.43	10.13	23.20	0.76	Clean Sandstone	RF + SSC (Self-Training)	86.7%
5585.0	70.18	25.00	23.83	0.85	Shaly Sand	Cluster Analysis	N/A
5625.0	31.93	27.45	27.71	0.88	Shaly Sand	RF + SSC (KSVM)	82.1%
5655.0	24.73	25.47	28.12	0.88	Clean Sandstone	Random Forest	89.3%
5695.0	33.75	8.50	29.34	0.80	Shaly Sand	Cluster Analysis	N/A
5735.0	61.27	9.38	23.56	0.76	Shaly Sand	RF + SSC (oneNN)	84.2%
5805.0	65.30	1.66	12.26	0.00	Shale	Cluster Analysis	N/A
5855.0	89.43	15.06	30.54	0.85	Shaly Sand	Random Forest	87.6%

F. Geological Interpretation (Niger Delta Context)

1. Clean Sandstone Facies: they are found at 5507 ft, 5515 ft, 5545 ft, 5655 ft. The characteristics of these facies are GR < 80 API, Rt > 20 Ωm, Φ > 28%, Shc > 0.8; Interpretation: High-quality reservoir sand (Channel/Bar deposits).
2. Shaly Sand Facies: they are found at 5585 ft, 5625 ft, 5695 ft, 5855 ft. In these facies, GR = 60-110 API, Rt = 5-20 Ωm, Φ = 20-28%, suggesting a marginal reservoir.
3. Shale Facies: These are found at 5535 ft, 5805 ft, with characteristics: GR > 100 API, Rt < 5 Ωm, Φ < 15%. This is indicative of Seal/Non-reservoir (Prodelta shales).

Critical Bypassed Zones are identified from 5515 ft to 5565ft Interval: they are initially classified as shaly sand by petrophysics (Fig. 4); ML reclassification: Clean Sandstone (5515) (42% porosity, 88% HC saturation); Missed in initial interpretation (Fig. 4); SSL classification: Clean Sandstone (5565) with 76% Hydrocarbon

saturation (HC); Characteristic: Moderate GR (39 API) masked by adjacent shales.

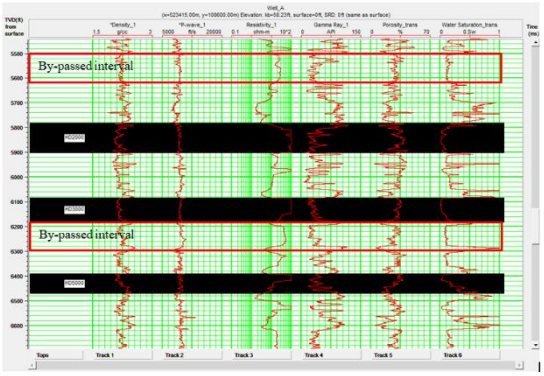


Fig. 4: Suite of logs showing by-passed intervals (the red ovals) in the study location (displayed in Hampson Russel's Software, Version CE8\R4.4).

Table 10: Performance Metrics Summary.

Model	Accuracy	Kappa	Key Strengths
Random Forest	85.7%	0.6832	High reliability in clean sands
SSC (Self-Training)	71.0%	0.4369	Effective in data-sparse zones
SSC (oneNN)	72.0%	0.4478	Best for minority class identification
Cluster Analysis	N/A	N/A	Foundation for geological grouping

5. Conclusion

This study introduces a robust and hybrid machine learning framework for enhanced reservoir identification and characterization in the mature onshore oilfields of the Niger Delta. By integrating unsupervised clustering, supervised learning (Random Forest), and semi-supervised learning (SSL) techniques, the framework addresses the limitations of traditional approaches and offers a more adaptable and accurate means of subsurface interpretation using well-log data.

A. Key Findings

1) Cluster Analysis Outcomes: The application of cluster analysis has enabled the reduction of complex subsurface lithologies into three distinct, geologically meaningful reservoir intervals. These clusters reveal lithofacies consistent with clean sandstone (high porosity, high resistivity), shale (low resistivity, high gamma ray), and shaly sand (intermediate properties). Among these, Interval (1) exhibits the most resilient log responses, suggesting better reservoir continuity and hydrocarbon retention compared to Intervals (2 and 3).
2) Supervised Learning Performance (Random Forest): The Random Forest classifier achieved a high prediction accuracy of 85.7%, effectively differentiating hydrocarbon-saturated zones. Resistivity has emerged as the most significant predictor, confirming its established role in fluid phase discrimination within the Niger Delta formations. This result aligns with the published studies and validates the model's generalizability in regional reservoir settings.
3) Semi-Supervised Learning Insight: The SSL approach has proved valuable in handling partially labeled datasets, a common challenge in real-world reservoir data.
   - ✓ The self-training classifier has achieved 71.0% accuracy, demonstrating the practical utility of SSL methods in scenarios with limited labeled samples.
   - ✓ The neural network-based SSL classifier has outperformed other models with a 72.03% accuracy, highlighting its effectiveness in capturing complex and nonlinear patterns in petrophysical data.
   - ✓ In contrast, the Kernel Support Vector Machine (KSVM) has underperformed with an accuracy of 58.86%, likely due to sensitivity to feature noise and class imbalance in the dataset.

B. Scientific and Practical Implications for the Niger Delta

Scientific Advancement: Our integrated machine learning framework combining unsupervised, supervised, and semi-supervised learning advances the current state of reservoir characterization by addressing data sparsity, heterogeneity, and incomplete labeling, which are common in real-world subsurface datasets. This holistic approach represents a methodological innovation not yet widely implemented in published literature.

C. Practical Implications for the Niger Delta and Analogous Basins

Revitalizing Mature Fields: Our three-reservoir classification strategy effectively identifies high-potential zones for bypassed hydrocarbon recovery, offering a targeted pathway to enhance existing reserves.

Cost Efficiency: By incorporating semi-supervised learning (SSL), our method significantly reduces dependence on fully labeled

datasets, thereby lowering operational costs and increasing feasibility in data-constrained environments.

Strategic Resource Management: The incorporation of resistivity and porosity benchmarks enables smarter well placement and completion strategies, directly reducing the risk of drilling non-productive wells.

The proposed methodology creates a **scalable, data-driven blueprint** for reservoir studies, extending its application beyond the Niger Delta to similar aging or complex hydrocarbon provinces.

6. Recommendations

From the results of this study, the following recommendations emerge:

1) Target Zones for Completion: Prioritize the interval:
   - 5515ft to 5655ft (high-confidence clean sands).
   - Test interval underlying top-5565ft (bypassed zone with SSL validation).

2) Unconventional Potential:
   - Investigates the immediate interval beneath top-5535ft (high-porosity shale: possible hybrid play).

7. Acknowledgments

We acknowledge all the members of the Geophysics research Group, Alex Ekwueme Federal University, Ndufu Alike, Ikwo, for the insight that sparked the conceptualization of this research.

8. Conflict Of Interest

We do not have any conflict of interest to declare.

9. References

Ahmed, M., Seraj, R. and Islam, S.M.S., 2020. K-means algorithm: A comprehensive survey and performance evaluation. Electronics, 9(8), 1295. https://doi.org/10.3390/electronics9081295

Al-Kaabi, A.U., Lee, W.J. and Al-Fattah, S.M., 2020. Machine learning in reservoir prediction: A review. SPE Journal, 25(3), pp.1234–1250. https://doi.org/10.2118/199999-PA

Avbovbo, A.A., 1978. Tertiary lithostratigraphy of the Niger Delta. AAPG Bulletin, 62(2), pp.295–300. https://doi.org/10.1306/C1EA45DC-16C9-11D7-8645000102C1865D

Azuoko, G.B., Ehirim, C.N., Ebeniro, J.O. and Uraechu, D.N., 2017. Analysis of multiples in onshore Niger Delta: A prelude to the fault shadow phenomenon. Journal of Petroleum Exploration and Production Technology. https://doi.org/10.1007/s13202-016-0309-8

Azuoko, G.-B., Usman, A.O., Ezim, E.O., Ekwe, A.C., and Abraham, E.M., 2023. Reservoir evaluation and hydrocarbon play assessment of Niger Delta field. Iranian Journal of Geophysics, 16(4), pp.113–122. https://doi.org/10.30499/ijg.2022.347384.1436

Azuoko, G.-B., Ekwe, A.C., Nlebedim, R.O., Usman, A.O., and Eluwa, N.N., 2022. Characterization of laterally continuous reservoirs: Implications on recovery of bypassed hydrocarbon in "Gerun" field, onshore Niger Delta. Petroleum and Coal, 64(1), pp.47–59.

Azuoko, G.B., Ekwe, A., Amulu, E., Usman, A., Eluwa, N., Omonona, V., and Udo, K., 2021. Rock property cross-plot analysis and post-stack acoustic impedance inversion for optimal reservoir characterization in Alpha Field, onshore Niger Delta Basin. Proceedings of the SPE Nigeria Annual International Conference and Exhibition, SPE-208252-MS. https://doi.org/10.2118/208252-MS

Breiman, L., 2001. Random forests. Machine Learning, 45(1), pp.5–32.

Cabrera, J.G. and de Castro, C., 2019. Random forest for forecasting pollution episodes in Madrid. International Journal of Forecasting, 35(3), pp.996–1007.

Castagna, J.P., Han, D. and Batzle, M.L., 2003. Rock physics of organic shales. The Leading Edge, 22(10), pp.942–947.

Cervantes, J., García-Lamont, F., Rodríguez-Mazahua, L. y López, A., 2020. A comprehensive survey on support vector machine classification: Applications, challenges, and trends. Neurocomputing, 408, pp.189–215.

Chen, B., Wang, J., Chen, Y. and Zhang, Y., 2015. Random forest for weather forecasting. International Journal of Machine Learning and Cybernetics, 6(5), pp.681–687.

Chen, Y., Li, X., and Zhang, H., 2021. Clustering seismic facies using self-organizing maps. Interpretation, 9(3), pp.SF1–SF12. https://doi.org/10.1190/INT-2020-0217.1

Doust, H. and Omatsola, E., 1990. Niger Delta. In: Edwards, J.D. and Santogrossi, P.A. (Eds.), Divergent/Passive Margin Basins. AAPG Memoir 48, Tulsa, Oklahoma, pp.201–238.

Emujakporue, G.O. and Faluyi, O.A., 2016. Evaluation of hydrocarbon prospect of Amu field, Niger Delta, Nigeria. International Research Journal of Geology and Mining, 6(2), pp.31–39.

Etu-Efeotor, J.O. and Akpokodje, E.G., 1990. Aquifer systems of the Niger Delta. Journal of Mining and Geology, 26(2), pp.279–284.

Ezim, E.O., Olayinka, A.I., Oladunjoye, M., Obiadi, I.I. and Azuoko, G.-B., 2022. Using inverted 4-D seismic and well data to characterise reservoirs from central swamp oil field, Niger Delta. World Journal of Advanced Research and Reviews, 13(2), pp.185–200. https://doi.org/10.30574/wjarr.2022.13.2.0083

Goel, A., Goel, A.K., and Kumar, A., 2023. The role of artificial neural network and machine learning in utilizing spatial information. Spatial Information Research, 31, pp.275–285.

Jarvie, D.M., Hill, R.J., Ruble, T.E., and Pollastro, R.M., 2007. Unconventional shale-gas systems: The Mississippian Barnett Shale of north-central Texas as one model for thermogenic shale-gas assessment. AAPG Bulletin, 91(4), pp.475–499. https://doi.org/10.1306/12190606068

Jwo, D.J., Biswal, A., and Mir, I.A., 2023. Artificial neural network for navigation systems: A review of recent research. Applied Sciences, 13(7), 4475.

Kaufman, L. and Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley and Sons.

Li, P. and Zhang, Y., 2016. Facies Characterization of a Reservoir in the North Sea Using Machine Learning Techniques. CS229 Final Project Report, Stanford University.

Nasraoui, O. and N'Cir, C.-E.B., 2019. Clustering Methods for Big Data Analytics: Techniques, Toolboxes and Applications. Cham: Springer.

Obaje, N.G., 2009. Geology and Mineral Resources of Nigeria. Berlin: Springer. https://doi.org/10.1007/978-3-540-92685-6

Onyekuru, S.O. and Ekeocha, N.A., 2024. Petrophysical and hydrocarbon saturation analysis of reservoir sands in Oswil Field, Niger Delta. FUPRE Journal of Scientific and Industrial Research, 8(1), pp.44–54.

Osisanya, O.M., Ogiesoba, O.E., and Ukpong, D.E., 2023. Petrophysical evaluation of Geowil Field, Niger Delta, using well log data. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3698398

Ozochi, C.A., Ozochi, B.C., and Akujobi, C.T., 2024. Reservoir characterization of Oswil Field, Niger Delta, using petrophysical analysis. Journal of Geophysical Research and Applications, 6(1), pp.12–25.

Reijers, T.J.A., 2011. Stratigraphy and sedimentology of the Niger Delta. Geologos, 17(3), pp.133–162. https://doi.org/10.2478/v10118-011-0008-3

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, pp.53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Szabó, N.P., 2018. A genetic meta-algorithm-assisted inversion approach: Hydrogeological study for the determination of volumetric rock properties and matrix and fluid parameters in unsaturated formations. Hydrogeology Journal, 26, pp.1935–1946. https://doi.org/10.1007/s10040-018-1749-7

Szabó, N.P., Braun, B.A., Abdelrahman, M.M.G. and Dobróka, M., 2021a. Improved well logs clustering algorithm for shale gas identification and formation evaluation. Acta Geodaetica et Geophysica, 56, pp.711–729. https://doi.org/10.1007/s40328-021-00358-0

Szabó, N.P., Dobróka, M., and Kavanda, R., 2013. Cluster analysis assisted float-encoded genetic algorithm for a more automated characterization of hydrocarbon reservoirs. Intelligent Control and Automation, 4(4), pp.362–370. https://doi.org/10.4236/ica.2013.44043

Szabó, N.P., Kilik, R. and Dobróka, M., 2023. Robust reservoir identification by multi-well cluster analysis of wireline logging data. Heliyon, 9(5), e15957. https://doi.org/10.1016/j.heliyon.2023.e15957

Szabó, N.P., Nehéz, K., Hornyák, O., Piller, I., Deák, C., Hanzelik, P.P., Kutasi, C. and Ott, K., 2019. Cluster analysis of core measurements using heterogeneous data sources: An application to complex Miocene reservoirs. Journal of Petroleum Science and Engineering, 178, pp.575–585. https://doi.org/10.1016/j.petrol.2019.03.067

Szabó, N.P., Valadez-Vergara, R., Tapdigli, S., Ugochukwu, A., Szabó, I. and Dobróka, M., 2021b. Factor analysis of well logs for total organic carbon estimation in unconventional reservoirs. Energies, 14, 5978. https://doi.org/10.3390/en14185978

Tiab, D. and Donaldson, E.C., 2015. Petrophysics: Theory and Practice of Measuring Reservoir Rock and Fluid Transport Properties. 4th ed. Oxford: Gulf Professional Publishing.

Yang, H., Pan, H., Ma, H., Konaté, A.A., Yao, J., and Guo, B., 2016. Performance of the synergetic wavelet transform and modified K-means clustering in lithology classification using nuclear log. Journal of Petroleum Science and Engineering, 144, pp.1–9. https://doi.org/10.1016/j.petrol.2016.02.031

Iraqi National Journal of Earth Science, Vol. 26, No. 03, 2026 (284-298)

Iraqi National Journal of Earth Science

www.earth.uomosul.edu.iq

تطبيق التعلم الآلي لتحديد وتوصيف المكامن في حقل نفط بري، دلتا النيجر، نيجيريا

أزوكو، جورج–بيست [1*]، جورج شينانو مبايي [2]، عثمان، آياتو أوجونوغوا [3]

تشوكوميريجي نوانيكيزي–فيل [4]، أودو كوفري إسرائيل [5]

ayatuusman@gmail.com george.chinanu@funai.edu.ng donalca04@gmail.com

kufreudo@uniuyo.edu.ng chukwumerije.nwanekezi-phil@funai.edu.ng

[1] قسم الجيولوجيا والجيوفيزياء، جامعة أليكس إكويمي الفيدرالية، ندوفو أليكي، نيجيريا.

[2] قسم الرياضيات والإحصاء، جامعة أليكس إكويمي الفيدرالية، ندوفو أليكي، نيجيريا.

[3] قسم الجيولوجيا والجيوفيزياء، جامعة أليكس إكويمي الفيدرالية، ندوفو أليكي، نيجيريا.

[4] قسم الجيولوجيا والجيوفيزياء، جامعة أليكس إكويمي الفيدرالية، ندوفو أليكي، نيجيريا.

[5] قسم الفيزياء، كلية العلوم الفيزيائية، جامعة أويو، ولاية أكوا إيبوم، نيجيريا.

تاريخ الاستلام: 19 اذار 2025 تاريخ المراجعة: 03 ايار 2025 تاريخ القبول: 30 تموز 2025

تاريخ النشر الالكتروني: 01 تموز 2026

الملخص

تواجه عملية توصيف المكامن في الحقول الناضجة ضمن دلتا النيجر تحديات كبيرة بسبب التباين الصخاري المحدد وندرة البيانات المصنفة، حيث غالباً ماتكون طرائق التعلم الآلي (ML) التقليدية غير فعالة بما فيه الكفاية سواءً كانت خاضعة للاشراف او غير خاضعة له. تسعى هذه الدراسة الى سد هذه الفجوة من خلال اقتراح إطار عمل هجين يعتمد على التعلم شبه الخاضع للاشراف (SSL) مصمم خصيصاً للتعامل مع مجموعة بيانات مستقاة من نتائج سجل الآبار المصنفة وغير المصنفة معاً. قمنا بتكامل التحليل العنقودي (Cluster analysis) ودمجه مع تحليل التجميع وخوارزمية الغابة العشوائية (Random Forest) وتقنيات التعلم شبه الخاضعة للاشراف لتحليل خمس طبقات تحت سطحية باستخدام سجلات المسامية والمقاومية الكهربائية وأشعة غاما وتشبع الهايدروكاربونات. يحدد التحليل العنقودي خصائص الوحدات بشكل موضوعي، ويعطي (RF) الأولوية للميزات التنبؤية، بينما يعالج (SLL) مشكلة ندرة البيانات المصنفة المكمنية بتجاوزها من خلال (التدريب الذاتي والشبكات العصبية ومصنفات KSVM). قام التحليل العنقودي بدمج الطبقات الخمسة في ثلاث مكامن جيولوجية (المكمن 3>2>1) مصنفة وفقاً لاستجابة السجل اللوغارتمي للآبار. بلغت دقة تنبؤ 85.7% (RF) حيث تم تحديد المقاومة (> 20 أوم.متر) كمؤشر حرج لوجود الهايدروكاربونات. ساهم (SSL) في تحسين تصنيف المكامن في المناطق التي تعاني من ندرة البيانات، حيث حقق التدريب الذاتي دقة بنسبة 71.0%، وتفوقت الشبكات العصبية بنسبة بلغت 72.03% و (KSVM) بنسبة 58.86%. تظهر هذه الدراسة ان (SSL) بالاشتراك مع التجميع العنقودي و (RF) غير الخاضع للاشراف قادر على تجاوز القيود المعروفة في تطبيقات التعلم الآلي (ML) بمفرده. ومن خلال اعطاء الأولوية للرؤى المدعومة بالمقاومية وتبني سير عمل (SSL) القابل للتوسع، فإن هذا الاطار يقدم حلاً فعالاً من حيث التكلفة لاسترداد تجاوز الهايدروكاربونات المتبقية في الحقول الناضجة. كما تعطي منهجيتنا وفق هذا النموذج الهجين معياراً جديداً ليشكل سابقة في تطبيق تقنيات التعلم شبه الخاضع للاشراف على الانظمة المكمنية المعقدة حول العالم، مما يعزز الدقة ويقلل الاعتماد على مجموع البيانات المصنفة بالكامل.

الكلمات المفتاحية:

التعلم شبه الخاضع للإشراف، توصيف المكامن، دلتا النيجر، التعلم الآلي الهجين، الحقول النفطية الناضجة

DOI: 10.33899/injes.v26i3.56228, ©Authors, 2026, College of Science, University of Mosul.