# Comparative Analysis of Machine Learning Algorithms for Predicting Petrophysical Parameters from Well Logs

**Patricia Famoofo*[1], Faith Gboyinde[1], Gabriel Omolaiye[1], Salami Rasaki[2] and Sunday Amoyedo[2]**
[1]Department of Geology and Mineral Science, Kwara State University, Malete, Nigeria
[2]TotalEnergies Nigeria Limited, Victoria Island Lagos, Nigeria

## ABSTRACT

Predicting effective water saturation (SWE) of hydrocarbon reservoir sections from well-log data is crucial in petrophysical data analysis. This study focuses on enhancing the prediction accuracy of SWE through the application an optimization of various machine learning ad deep learning models. We compared the performance of three machine learning algorithms: Random Forest, XGBoost, and Support Vector Machine (SVM), alongside a deep-learning model employing Long Short-Term Memory (LSTM) networks. Hyperparameter optimization, critical for model parameters fine-tuning, was done using Optuna, a robust optimization framework. This adaptive approach enabled us to systematically explore hyperparameter spaces, thereby improving model accuracy and performance. The models were evaluated using two key regression metrics: Mean Squared Error (MSE) and the coefficient of determination (R2). In addition, we employed SHapley Additive exPlanations (SHAP) to interpret the model predictions. SHAP values provided insights into the impact of each feature on the target prediction, enhancing our understanding of the model's decision-making process and identifying the most influential variables in predicting SWE. Our results indicated that the XGBoost model outperformed the other models, achieving the lowest MSE and the highest $R^2$ value. This superior performance of the XGBoost model suggests that it is highly effective in predicting SWE, the target variable in our dataset. The findings underscore the importance of selecting appropriate algorithms and optimizing hyperparameters to enhance predictive accuracy in petrophysical data analysis.

**Keywords:** SWE · Machine Learning · Deep Learning · Hyperparameter Optimization · SHAP

## INTRODUCTION

In today's rapidly evolving technological landscape, the oil and gas industry faces increasing pressure to optimize reservoir characterization and management processes. Accurately predicting petrophysical properties from well logs is paramount for making informed decisions in exploration, production, and reservoir engineering (Aminzadeh & Dasgupta, 2013). Traditional methods, reliant on manual interpretation and empirical equations, often fall short in providing precise and scalable solutions to this challenge. Amidst these complexities, machine learning (ML) algorithms have emerged as transformative tools offering the potential to revolutionize well-log prediction by leveraging the wealth of information contained in well-log data. ML techniques have demonstrated remarkable capabilities in uncovering intricate patterns and relationships within well-log data which leads to enhanced accuracy and efficiency in the prediction of petrophysical properties from well logs, compared to traditional methods (Bhattacharya *et. al.,* 2020).

Well logging is an essential process in the oil and gas industry that involves the systematic collection of data from a wellbore. This data is then analyzed to provide valuable insights into the subsurface formations and reservoir properties. The data analysis helps determine various petrophysical properties of the reservoir, such as porosity, permeability, fluid saturation, and the presence of hydrocarbons (Ellis & Singer, 2007). Petrophysical properties are key characteristics of rocks that influence their behaviour in the context of oil and gas exploration and production. Gamma-ray logs measure natural radioactivity to help identify lithology, Resistivity logs assess how easily electricity can pass through the formation, indicating the presence of hydrocarbons or water, and Porosity logs (neutron and density) help determine the volume of water and hydrocarbon by providing the basis for saturation calculations (Rider & Kennedy, 2011). Water saturation (Sw) is the ratio of the

© Copyright 2025. Nigerian Association of Petroleum Explorationists.
All rights reserved.

The authors wish to thank NNPC Limited, NNPC Upstream Investment Management Services (NUIMS), Department of Geology and Mineral Science, Kwara State University, Malete, Nigeria, TotalEnergies Nigeria Limited, Victoria Island Lagos, Nigeria and NAPE for providing the platform to present the paper during the Annual Conference.

volume of water in the pore space to the total pore volume. Effective water saturation (SWE) is a key parameter in hydrocarbon reservoirs that represents the percentage of pore space filled with mobile water (water that hydrocarbons can displace during production). It excludes the water bound to the rock grains (irreducible water) and focuses on the moveable water that impacts hydrocarbon production (Asquith & Krygowski, 2004).

Several well logs can be combined to determine the effective water saturation of a formation with the primary combination being the resistivity and porosity logs. Archie's equation which is the primary mathematical expression for calculating effective water saturation allows for the integration of resistivity and porosity data to estimate fluid saturation in the reservoir (Archie, 1942). However, this traditional method of estimating effective water saturation has some limitations (Crain, 2010). The employment of machine learning algorithms can address the key limitations of conventional methods in predicting petrophysical properties from well logs. Since conventional methods are often time-consuming, labour intensive, and struggle to capture the complex, nonlinear relationships between well-log measurements and petrophysical properties, machine learning models such as random forest, extreme gradient boosting, and neural networks can automate the interpretation process, learn to intricate patterns in the data, and handle large-dimensional datasets more effectively. This will result in improved predictive accuracy, reduced human effort, and enhanced flexibility to adapt to diverse geological settings and changing reservoir conditions. In additional, the method will quantify uncertatinty and gain interpretable insights into the underlying relationships, which are valuable for informed decision-making in exploration and production operations (Zhang *et. al.,* 2018). This study aimed to evaluate different ML algorithms such as Support Vector Machine (SVM), Random Forest, and Gradient Boosting (XGBoost), as well as Long Short-Term Memory (LSTM), their efficacy in forecasting petrophysical properties which in this project is effective water saturation from well data, and to identify the most efficient methodology.

**Geology of the Offshore Niger Delta, Nigeria**

The Niger Delta is situated in the Gulf of Guinea and it extends throughout the Niger Delta province (Klett, 1997). From the Eocene to the present day, the delta prograde southwestwardly, forming depobelts that represent the most active portion of the delta at each stage of its development (Doust & Omatsola, 1990). The onshore portion of the Niger Delta province is delineated by the geology of the southern Nigeria and southwestern Cameroon. The Offshore boundary of the province is defined by the Cameroon volcanic line to the east and the eastern boundary of the Dahomey basin to the west. The province covers 300,000 km2 and includes the geologic extent of the Tertiary Niger Delta (Akata-Agbada) petroleum system. Offshore Nigeria is a significant area of interest for hydrocarbon exploration and production, situated within the Niger Delta Basin. This region is known for extensive oil and gas reserves which have been a major contributor to Nigeria's economy. For any given depobelt, gravity tectonics were completed before deposition of the Benin Formation and are expressed in complex structures, including shale diapirs, roll-over anticlines, collapsed growth faults crests, and steeply dipping, closely spaced flank faults (Evamy *et al.* 1978; Xiao & Suppe, 1992). The geology of the Offshore Niger Delta is complex and influenced by various geological processes including sedimentation, tectonics, and sea-level changes (Eustacy). The basin is made up of three major lithostratigraphic units: (1) Akata Formation dates back to the Palaeocene, is the oldest unit and consists primarily of marine shales, claystone, and silts. It is believed to have been deposited in deep marine environments under high-pressure conditions and serves as the primary source rock for hydrocarbons (Doust & Omatsola, 1990; Tuttle *et. al.,* 1999; Kulke, 1995). (2) Agbada Formation, a paralic sequence dates to Eocene. It is the middle unit comprised of an alternation of sandstones, shales, and siltstones deposited in shallow marine and deltaic environments. It is the main reservoir rock for oil and gas in the basin (Short & Stauble, 1967; Evamy et. al., 1978; Weber & Daukoru, 1975). And (3) Benin Formation, which is the youngest unit made up of predominantly fluvial and continental sands, deposited by river systems. This formation is typically non-hydrocarbon-bearing but represents the uppermost portion of the Niger Delta sequence. The formation date to the Oligocene (Avbovbo, 1978; Weber & Daukoru, 1975; Reijers *et.al.,*1997).

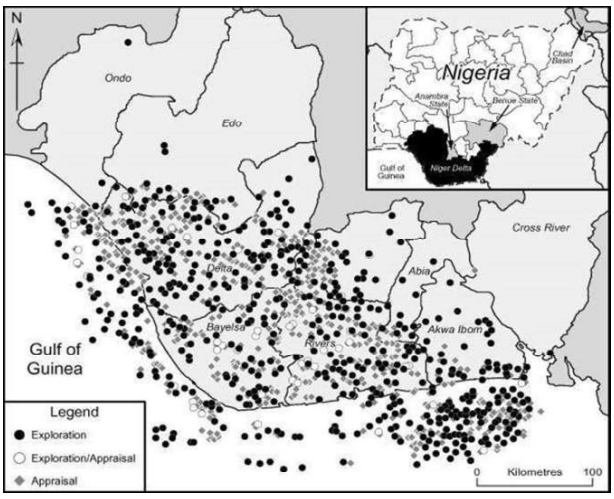


**Figure 1:** Map of the Niger Delta showing the categories of oil well. Adapted from Anifowose *et.al.* 2014.

**METHODOLOGY**

The machine learning (ML) concept using Python was

adopted in this research and the algorithms are Support Vector Machine (SVM), eXtreme Gradient Boosting Machines (XGBoost), Random Forest, and LSTM which are all popular machine learning algorithms used for classification and regression tasks. Some basic concepts used are briefly explained below:

**Artificial Intelligence (AI)**

In 1956, John McCarthy coined the phrase amid a time of considerable enthusiasm and hope for the technological industry. Computers and electronics, the size of whole rooms, were still in their infancy at the time. Early ideas of electronic intelligence in the future were inspired by these machines, despite their limited capacity for fundamental computations. These machines outperformed humans in terms of speed and efficiency (Nilsson, 2010). These days, without getting too science fiction-y, the best way to define artificial intelligence, or AI, is as "the automated processing by a computer that can carry out tasks typically associated with human intelligence."

**Machine Learning**

I. This is a significant branch of artificial intelligence (AI), focuses on developing algorithms and statistical models that allow computers to perform tasks without being explicitly programmed. It is a subset of AI that comes into play when systems are designed to learn and solve problems previously thought to be exclusive to human intelligence. In machine learning, computers are programmed to learn and adapt from experience, improving their performance on tasks by identifying patterns and making inferences, rather than being directly told the correct outcome (Mitchell, 1997; Mohri *et al.,* 2018). Machine learning methods can be used in predicting petrophysical logs without cost or time and needing additional data. Petrophysical issues in a part of the formation where it is not possible to do well logging operations or the obtained data are not reliable, are among the topics of interest in petroleum engineering (Kheirollahi *et al.,* 2023; Rajabi *et al.,* 2023). Machine learning is divided into (i) Supervised Learning, (ii) Unsupervised Learning and (iii) Reinforcement Learning (RL).

**Deep Learning**

Deep learning is a subset of machine learning that involves training neural networks with many layers (hence "deep") that can learn to represent data with increasing levels of abstraction to model complex patterns in data. Deep learning models, particularly deep neural networks, can automatically discover patterns and features from large amounts of data, making them highly effective for complex tasks. These deep neural networks are capable of learning hierarchical representations, making them highly effective for tasks such as image and speech recognition, natural language processing, and more.

I. **Support Vector Machine (SVM):** SVM is a supervised learning algorithm that is primarily used for the classification of tasks, but it can also be applied to regression. It works by finding the hyperplane that best separates the data points into different classes or groups. It uses a technique called the kernel trick to transform data and then based on the transformations, it finds the optimal boundary between the possible outputs. SVM can handle both linear and non-linear classification tasks through the use of different kernel functions, such as linear, polynomial, or radial basis function (RBF) kernels.

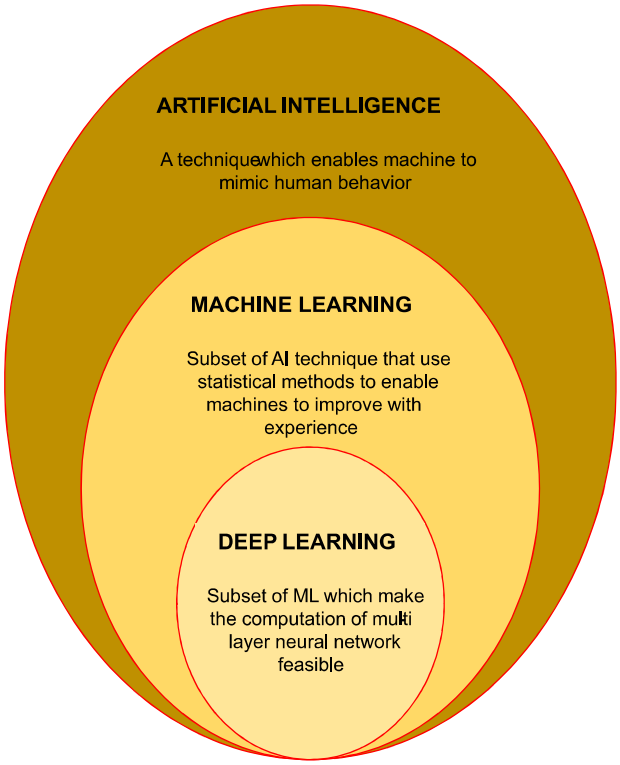


**Figure 2:** Schematization of the relationship between artificial intelligence, machine learning, and deep learning

ii. **eXtreme Gradient Boosting Machines (XGB):** is an ensemble learning technique that builds a strong predictive model by combining multiple weak learners sequentially. It works by iteratively fitting new models to the residual errors of the previous models, gradually reducing the overall prediction error. It is partially effective for regression and classification tasks and is known for its high predictive accuracy and flexibility.

iii. **Random Forest (RF):** It is an ensemble learning algorithm that builds a collection of decision trees and combines their predictions through averaging or voting. Each decision tree in a random forest is trained in a random subset of the data and a random subset of the features, which helps reduce overfitting and increase the diversity of the trees. Random Forest is known for its scalability, ease of use and ability to

handle high-dimensional data and large datasets.

**Long Short-Term Memory (LSTM):** It is a type of recurrent neural network (RNN) architecture designed to learn from sequences of data. Developed to overcome the limitations of traditional RNNs, LSTMs are particularly effective in capturing long-range dependencies and temporal patterns in sequential data. It contains memory cells that can store information for long periods, allowing the network to learn contextual information from previous time steps. LSTMs use three types of gates to control the flow of information: Forget Gate (decides which information to discard from the memory), Input Gate (determines what new information to add to the memory), and Output Gate (controls what information from the memory should be outputted to the next layer). It has a cell state that runs through the entire LSTM unit and carries relevant information across time steps, enabling the model to maintain context. LSTMs are trained using BPTT, which allows them to learn from the entire sequence of data, making them suitable for tasks where the order of inputs matters. The model can effectively model the relationships between different log measurements taken at successive depths, helping to reveal trends and correlations that may not be apparent using traditional methods. Well log datasets often contain missing values due to issues like equipment failure or data acquisition problems. LSTMs can learn to infer missing values based on the surrounding data, improving the robustness of predictions.

**Software Used in this Research**

This Study leveraged several key software tools: Anaconda, a Python distribution streamlining package management and environment creation for data science and machine learning; OpendTect (v7.0.4), an open-source seismic interpretation system used for data import, preprocessing, and integration with Python via its API (odbind); and Jupyter Notebook, a web application launched from Anaconda, facilitating interactive coding, visualization, and documentation for data analysis and machine learning workflows.

**Libraries and Modules Used -** Several Python libraries: NumPy for efficient numerical operations; odbind to interface with OpendTect software; Pandas for data manipulation and analysis; PyOD for outlier detection; Matplotlib and Seaborn for creating static and statistical visualizations; and ydata-profiling for generating comprehensive dataset reports.

**Data Analysis -** basic information about the collected data, the summary statistics and the structure of the dataset are inspected to understand their structure, contents, and relationships. Y-data profiling was used to generate profile reports to create comprehensive reports that summarize the well-log data. These reports include the following:

a. Descriptive Statistics: It provides a summary of key metrics such as mean, median, minimum, maximum values, number of variables, observation, and data distributions.
b. Visualization: Using histograms, box plots, and correlation metrics to visualize data distributions and relationships between variables in well logs data.
c. Missing Values Identification: Highlighted areas with missing or NaN values, which were used for subsequent cleaning steps.
d. Comparative analysis between different segments or datasets, aiding in identifying variations and anomalies.

**Data Cleaning**

This was done to correct or remove inaccurate records from a dataset. It ensures that the dataset is free from errors, inconsistencies, and missing values, which could negatively impact the performance of machine learning models. With the profile report from y-data profiling, data quality issues can be efficiently identified and addressed.

**Data Preparation**

Data preparation transforms raw data into format suitable for machine learning models. This step benefits from the insights gained during the data analysis and cleaning. This step is important for ensuring the quality and integrity of the data, which directly impacts the performance of machine learning models.

i. Data Transformation: Applying techniques like scaling, normalization, and encoding to convert data into a suitable format for analysis and modelling. Applying scaling such as standard scaling, min-max scaling, or robust scaling will bring all features to a similar scale, which is important for algorithms that rely on distance metrics.

ii. Feature Engineering: Creating new features or modifying existing ones to improve or enhance the predictive power of the model. Logarithmic transformation of resistivity log data (RT) was done because resistivity data, like many geophysical measurements, can be highly skewed with a long tail of high values as it ranges between 0 to 2000 ohms. This is because skewed data can affect the performance of machine learning algorithms, which often assume normally distributed data. Log transformation helps in reducing skewness and makes the resistivity data more normally distributed, leading to better model performance. Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.

iii. Data Splitting: This involves dividing the dataset into training and testing sets to evaluate the model's performance on unseen data. Machine learning enables learning some properties by a model from a dataset and applying them to new data. This is because a common

practice in machine learning is to evaluate an algorithm. The evaluation consists of splitting the dataset into two parts, one called the training set, with which the properties of the data are learnt, and the other called the testing set, on which to test these properties.

a. Training Set: The training set is a subset of the dataset used to train a machine learning model. The model learns the patterns, relationships, and features in the data from this set. The performance of the model is iteratively improved by adjusting the parameters based on the error it makes on the training set. In the context of a petrophysical dataset, the training set consists of well-log data where the petrophysical properties, such as porosity, permeability, and water saturation, are known. This data is used to train a machine learning model to predict these properties from well logs.

b. Testing Set: The testing set, also known as the validation set, is a separate subset or portion of the dataset used to evaluate the performance of the trained model. This set is not used during the training phase. This set is used to evaluate the model's accuracy and generalisation capability in predicting petrophysical properties on new, unseen data.

iv. Cross-Validation: Implementing cross-validation techniques to ensure the model generalizes well to different subsets of the data.

v. Dimensionality Reduction: Techniques like PCA to reduce the number of features while retaining important information, if necessary.

**Model Training**

Model training is the process by which a model learns to make accurate predictions or decisions from data. During training, the model iteratively adjusts its internal parameters to minimize the error between the predictions and the actual outcomes, using a defined loss function. This is achieved through a series of steps which include:

i **Model Initialization:** The model is initialized after data have been collected, pre-processed and split It involves algorithm selection which is done based on the problem or task (e.g., regression, classification). The selection of the feature scaling technique is done to get the best scaler using: trial.suggest_categorical. After this is done, the parameters or hyperparameters are set using: trial.suggest_integer, and trial.suggest_float where it applies to suggest the best values for each parameter which is eventually used to deploy the model.

ii. **Training process:** The training process in machine learning is the iterative procedure through which a model learns from data. It involves feeding the model with input data, adjusting its parameters to minimize the error between its prediction and actual values, and refining the model till it achieves satisfactory performance.

iii. **Model Evaluation:**

a. Validation: It involves evaluating the model's performance on the validation set to tune hyperparameters and prevent overfitting.

b. Testing: The final model's performance is assessed on a separate test set using the appropriate evaluation metrics for the task.

The model was deployed once it was trained and validated, to make predictions on new unseen data.

**Evaluation and Error Metrics**

Evaluation and error metrics are critical for assessing the performance of machine learning models. They provide insights into how well a model is performing and help identify areas for improvement. There are different types of evaluation metrics depending on the task that is performed; in this case, regression. Regression is a type of supervised learning task where the goal is to predict a continuous numerical value. The output variable is a real number, which means it can take on any value within a range.

I. **Mean Squared Error (MSE):** It measures the average squared difference between the predicted and actual values. Lower MSE values indicate better model performance. However, because it squares the errors, larger errors have a disproportionate impact on the metric. It is given by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where: $y_i$ is the actual value and $\hat{y}_i$ is the predicted value.

ii. **Root Mean Squared Error (RMSE):** It is the square root of the Mean Squared Error. It provides the interpretation of the average error magnitude in the same units as the target variable. It is more interpretable than MSE but still sensitive to large errors. It is given by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

iii. **R-Squared (R2):** It measures the proportion of variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, where the higher values indicate a better fit. An R2 value of 1 means that the model explains all the variability in the target variable. It is given by:

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

iv. Loss (or Training Loss): It is the error calculated on the training dataset for each epoch of training. It measures how well the model fits the training data.

v. Validation Loss: It is the error calculated on the validation dataset; a separate subset of data not used for

training. It is used to check how well the model generalizes to unseen data

## RESULTS AND DISCUSSION

After training and evaluation, each model is ready to make predictions on new, unseen data. The various prediction results from each model, the comparison of the models' evaluation metrics to determine the best model prediction, as well as the results of the residual analysis calculated for each model are presented in Figures 3-6, training loss and validation loss plot for LSTM (Figure 7).

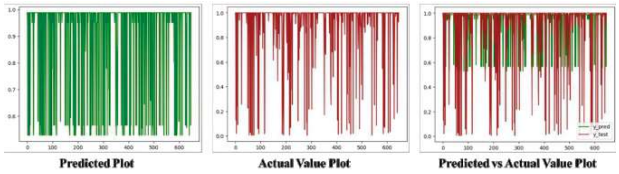


**Figure 3:** Random Forest Result.

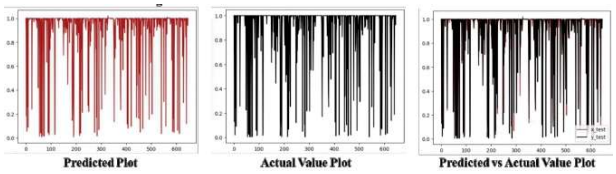


**Figure 4:** Extreme Gradient Boosting Result.

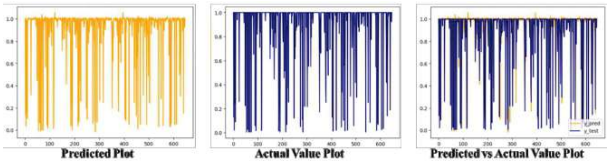


**Figure 5:** Support Vector Result.

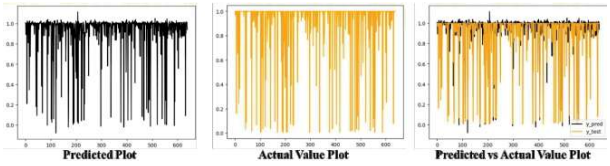


**Figure 6:** Long Short-Term Memory Result.

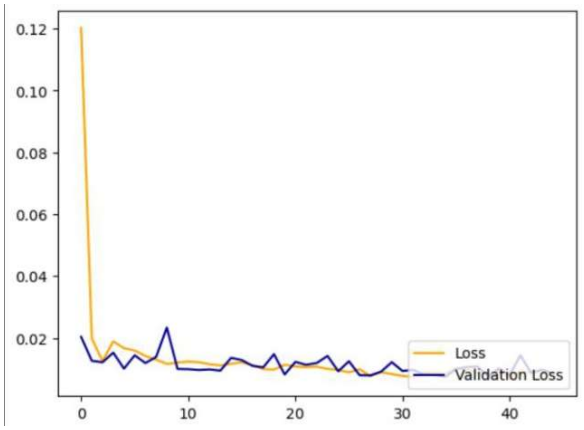


**Figure 7:** Training Loss and Validation Loss plot for LSTM model.

### Residual Analysis

A residual analysis was performed to assess the performance of the models. It involves examining the differences between the actual values and the predicted values which are known as the residuals.

Mathematically:

*Residual=Actual Value-Predicted Value*

Residual analysis helps determine how well the model has captured the data's pattern, to detect any unusual observations that may unduly influence the model. When the Residual values are plotted against the predicted values, any pattern or systematic structures suggest that the model might be missing some information (Figures 8 – 11).

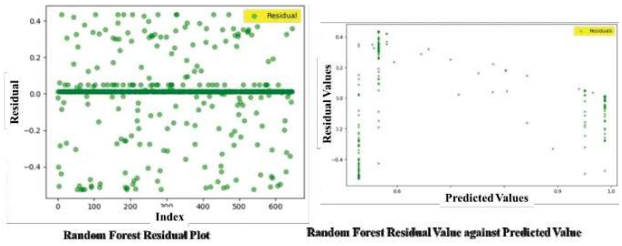


**Figure 8:** Random Forex.

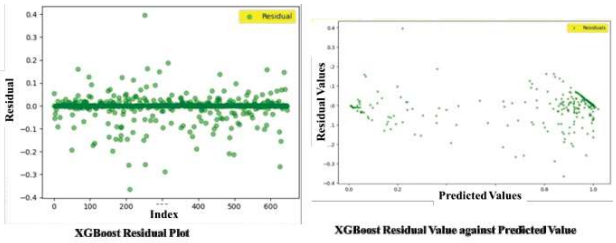


**Figure 9:** XGBoost.

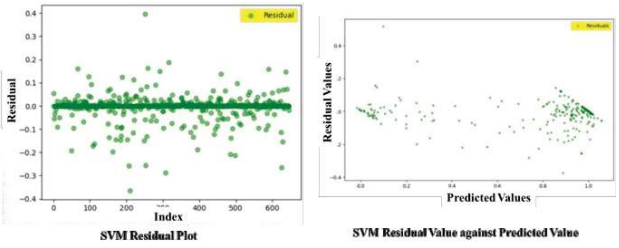


**Figure 10:** SVM.

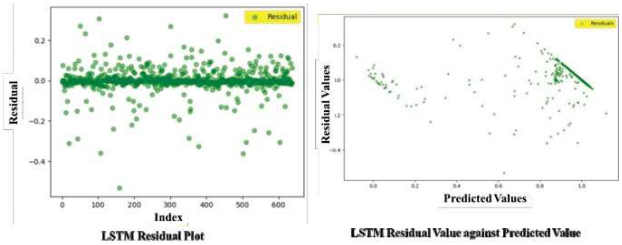


**Figure 11:** LSTM.

**Models' Metric Values Comparison**

The best MSE and R2 metric results from each model were compared to validate which model performed better. The XGBoost model demonstrated superior performance achieving the lowest MSE and the highest R2 (Figures 12-13). These values can be seen in the bar charts below:

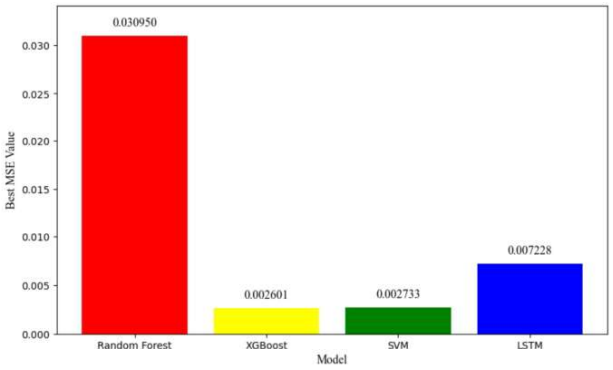


**Figure 12:** MSE Best Value Comparison Among all Four Models.

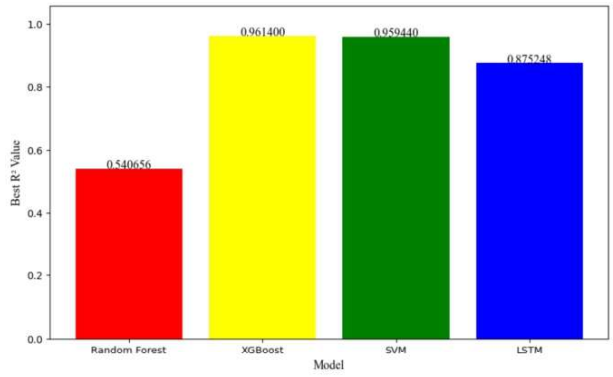


**Figure 13:** R2 Best Value Comparison Among all Four Models.

**Models' Prediction Interpretation**

SHapley Additive explanations (SHAP) were employed to interpret the models' prediction. It helps provide a unified measure of feature importance and how features influence individual predictions.

A SHAP summary plot shows the distribution of SHAP values for each feature. Each point represents a SHAP value for a feature in a specific prediction.

i. Y-axis: Features arranged by importance (average absolute SHAP value).

ii.X-axis: SHAP value showing the impact on the model output.

iii. Colour: Represents the feature value (red for high values, blue for low values).

Interpretation:

i. Feature Importance: The higher up a feature is, the more important it is. The average SHAP value's magnitude indicates the overall impact.

ii.Feature Effect: The spread of the SHAP values along the x-axis shows how much the feature impacts predictions. Wider spreads indicate more influence.

iii.Feature Value Impact: Colour indicates if higher or lower feature values increase or decrease the prediction.

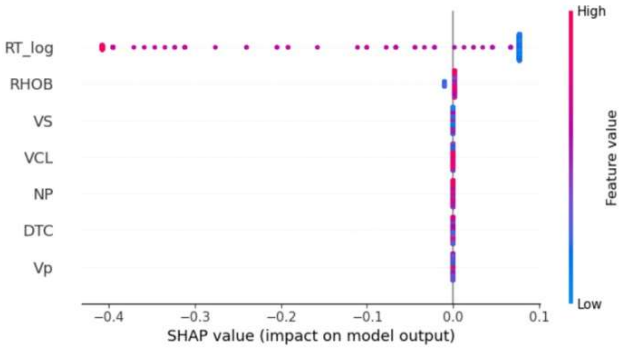


**Figure 14:** Random Forest Summary Plot.

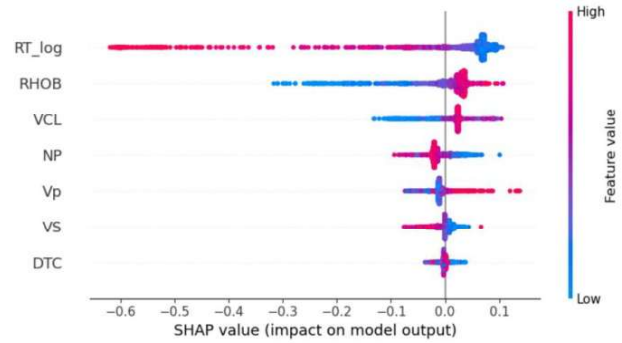


**Figure 15:** XGBoost Summary Plot.

**Discussion of Results**

**Random Forest**

The Random Forest (RF) prediction plot has a relatively uniform distribution of green vertical lines, indicating that the model was able to capture the overall variation in the data and make consistent predictions. The actual value plot shows a more varied and scattered pattern of red lines.

This suggests that the true values in the dataset have a higher degree of variability and complexity compared to the model's prediction (Figure 3). By overlaying the predicted values and the actual values, how well the model's predictions align with the true value can be visually assessed on the plot. From the plot, it can be seen that there is generally good alignment between the predicted value and the actual value, especially for the majority of the data points. This suggests that the RF model captured the overall trends and patterns in the data reasonably well (Figure 3). However, there are some areas where the predicted and actual values diverge more significantly. These discrepancies represent instances where the model's predictions do not match the true values in the dataset. This could be due to the inherent complexity of the data, limitations in the model's ability to capture certain relationships, or the presence of outliers in the data. between the model's outputs and the actual values in the data.

The RF residual plot shows the residuals of the predictions made by the Random Forest model across the entire dataset, plotted against the index of the test samples. The residual range is between -0.4 and 0.4. The residuals are scattered around the zero line, which suggests that the errors are evenly distributed and centred around zero. This indicates that RF predictions are not systematically biased. The spread of the residuals is relatively consistent across different levels of actual values which implies homoscedasticity (constant variance of errors) which is a sign of a well-behaved model (Figure 8). However, there are some data points with larger residuals, as shown by the green dots further away from the zero line. These points represent instances where the model's predictions deviate more significantly from the actual values. The presence of these larger residuals suggests that the model may have difficulty in capturing the complexity or variability of the data in certain regions for specific data points.

Homoscedasticity refers to the situation in which the variance of the residuals (errors) from a regression model is constant across all levels of the independent variable(s) which means that the spread of residuals is the same for all predicted values, (Gujarati & Porter, 2009). Heteroscedasticity, on the other hand, occurs when the variance of the residuals is not constant across all levels of the independent variable(s), which means that the spread of residuals varies at different levels of the predicted values, (Wooldridge, 2015). The plot of residuals against the actual values shows that the residuals are distributed evenly across the range of predicted values, with no clear patterns or trends emerging (Figure 8). The scattered distribution of the residuals around the zero line indicates that the model is generally capturing the underlying relationships in the data well.

Overall, the residual plot highlights the areas where the model has some difficulty, while the residuals vs predicted values plot confirms the overall consistency and quality of the model's predictions. Random Forest had the highest MSE and the lowest R2 value, indicating the poorest performance among the four models as it has more error in predictions and explains less variance in the data respectively, meaning it has the weakest explanatory power and ability to fit the data Figures 12 & 13). The SHAP summary plot for RF in Figure 14 shows the SHAP values for different features, including RT_LOG, RHOB, NP, Vp, VLC, VS, and DTC. The colour of the data points indicates the direction of impact on the model's output. RHOB and VCL have the highest positive SHAP values, indicating that they have the strongest positive impact on the model's output. This means that as the features increase, the model's output tends to increase as well. VP has the highest negative SHAP values, indicating that they have the strongest negative impact on the model's output. NP, DTC, and RT_log have relatively smaller SHAP values, both positive and negative, indicating that they have a moderate impact on the model's output compared to other features.

**XGBoost**

The XGBoost prediction plot displays the predicted values of the XGBoost model for each data point (represented by vertical red lines) for the entire dataset with the model's output ranging from 0.6 to 1.0 (Figure 4). The combined plot overlays the plot of the predicted values (red lines) and the plot of the actual values (black lines). The alignment between the predicted and actual plot suggests that the XGBoost model is generally able to capture the overall trends and patterns in the data, as the predicted values closely match the actual values for most data points (Figure 4). However, there are some instances where the predicted and actual values diverge, indicating areas where the model's predictions do not perfectly align with the true values in the dataset. The close alignment between the predicted and actual values suggests the model is performing well, but the discrepancies highlight opportunities for further model refinement and optimization to improve the overall accuracy.

The XGBoost residual plot shows the residuals of the predictions made by the eXtreme Gradient Boosting model. The residuals are scattered randomly around the zero line, indicating that the model captures the trend in the data well. The residuals range between -0.3 and 0.3. There is no visible pattern, and the variance appears to be consistent across the different index values (Figure 9). The plot of the residuals against the actual values shows that the residuals are randomly distributed around the zero. There is a slight funnel shape indicating the possibility of heteroscedasticity, but it is not very pronounced. Some residuals are closer to zero, showing that the model predictions are quite accurate for most data points (Figure 9). XGBoost had significantly lower MSE values compared to RF and also had the highest R2 value slightly better than SVM which also indicates that it is the best-performing model because it explains a large portion of the variance in the data, demonstrating the strongest model fit, ability to capture the underlying relationships in the data, and best performance in minimizing prediction error (Figures 12 & 13).

The SHAP summary plot for XGBoost shows that RHOB and Vp have the highest positive SHAP values, indicating that they have the strongest positive impact on the model's output. As these features increase, the model's output tends to increase as well. RT_log has the highest negative SHAP value, suggesting it has the strongest negative impact on the model's output. As it increases, the model's output tends to decrease. NP, VCL, and DTC have relatively smaller SHAP values, both positive and negative, indicating that they have a moderate impact on the model's output compared to other features. The RT_log and RHOB seem to have the most prominent SHAP values, indicating their importance in the model's performance. The XGBoost SHAP summary plot appears to capture more nuanced relationships between the

features and their impact on the model's output compared to the Random Forest SHAP summary plot (Figure 15).

**Support Vector Machine**

The combined plot overlays the plot of the predicted values (orange lines) and the plot of the actual values (blue lines). The alignment between the predicted and actual plots suggests that the SVM model is generally able to capture the overall trends for most data, as the predicted values closely match the actual values for most data points. However, there are some instances where the predicted and actual values diverge, indicating areas where the model's predictions do not perfectly align with the true values in the dataset. The close alignment between the predicted and actual values suggests the model is performing well, but the occasional discrepancies highlight opportunities for further model refinement and optimization to improve the overall accuracy (Figure 5).

The SVM residual plot shows the residuals of the predictions made by the SVM model. The residuals are also scattered randomly around the zero line like XGBoost. The residuals also range from -0.3 to 0.3. There is no visible pattern suggesting that the errors are distributed evenly. The plot of the residuals against the actual values shows that the residuals are randomly scattered around the zero line, similar to the XGBoost model. There is a slight funnel shape, indicating a minor presence of heteroscedasticity. There are a few residuals that deviate significantly, but most are closer to zero, indicating good model performance (Figure 10). SVM had low MSE and the high R2 values slightly close to XGBoost which means the model explains a large portion of the variance in the data just like XGBoost (Figures 12 & 13).

**Long Short-Term Memory**

The LSTM predicted value against the actual value plot shows good alignment between the predicted and actual values, suggesting that the LSTM model was able to capture the overall trend and patterns in the data. However, there are some discrepancies, especially at certain index values, where the predicted and actual values diverge, indicating the model could not accurately predict the exact values at those points (Figure 6). The LSTM loss and validation loss plot indicates the model is learning rapidly and improving its predictions on both the training and validation sets as evident from the decreasing curves. The validation loss being close to the training loss is a positive sign. It indicates that the model is generalizing well and not overfitting to the training data as overfitting would be characterized by a low training loss and a significantly higher validation loss.

The LSTM residuals plot shows the residuals for the model across different index values which represent time steps. The plot exhibits a scattered distribution of the residuals, with some clusters of higher residuals, indicating that the model may not be perfectly fitting the data, and there are some errors or deviations in the predictions. The plot of LSTM residuals against the predicted values visualizes the relationship between the predicted values and the residuals of the model. The scattered points suggest that the residuals are not uniformly distributed across the range of predicted values, meaning the model has varying performance or accuracy. It also shows a general trend where the residuals decrease as the predicted values increase, indicating the model may be better at predicting higher values compared to lower values (Figure 11). The LSTM model is the third best in terms of prediction accuracy based on its MSE value and its R2 value is also a good fit but lower than XGBoost and SVM (Figures 12 & 13).

## CONCLUSION

This Research explored and compared the performance of various machine learning (ML) algorithms in predicting petrophysical properties from well-log data. Based on the investigations carried out to achieve the specific objectives of the project, the following conclusions were drawn:

i. A comprehensive review of existing literature on the application of machine learning algorithms for well-log analysis was conducted which suggests that ML algorithms have been successfully applied to well-log analysis, providing improved accuracy and efficiency compared to traditional methods. The review also provided a solid foundation for understanding the current state of research and identifying potential gaps and areas for improvement.

ii. The meticulously preprocessed and prepared dataset of well-logs and petrophysical properties provides a robust foundation for the ML model development and evaluation. This step was very important as it ensures the quality and reliability of the data used in subsequent analyses.

iii. The implementation and evaluation of RF, XGBoost, SVM, and LSTM algorithms have demonstrated their potential in predicting petrophysical properties from well-log data. These algorithms were chosen based on their popularity and reported success in similar applications.

iv. The comparative analysis of the algorithms' performance has revealed the strengths and weaknesses of each method, with SVM exhibiting the highest accuracy and precision while maintaining computational efficiency. The analysis revealed that SVM and XGBoost outperformed Random Forest in trms of both MSE and R2 values, indicating their superior predictive capabilities for this specific dataset. SVM demonstrated a slight edge over XGBoost, particularly in terms of MSE values.

In conclusion, this project successfully demonstrated the potential of ML algorithms in enhancing the accuracy and efficiency of petrophysical property predictions from well-log data.

## RECOMMENDATION

Based on the comparative evaluation of the four ML algorithms of this project, the XGBoost is recommended as the most suitable algorithm for predicting petrophysical properties from well logs. It provided the best trade-off between predictive accuracy, precision, and computational efficiency, making it the preferred choice for this application. While the SVM model also performed well, the XGBoost model's superior predictive accuracy, as indicated by the lower MSE values, suggests that it is the most appropriate algorithm for the given well-log analysis task. To further enhance the project's impact, the following recommendations should be considered:

i. Explore the possibility of incorporating additional well-log features (such as Gamma-ray log, etc.) and geological data (such as lithology, stratigraphy, etc.) to improve the model's predictive capability for petrophysical properties.

ii. Investigate the interpretability of the XGBoost model to understand the underlying relationships between well-log features and petrophysical properties.

iii. Develop a user-friendly interface or tool to seamlessly integrate the XGBoost model into the oil and gas industry's well-log analysis workflow.

## REFERENCES CITED


Aggour, T. M. (2019). Application of Machine Learning in Lithofacies Classification Using Well Logs. Journal of Petroleum Science and Engineering, 175, 538-547.

Aminzadeh, F., & Dasgupta, S. (2013). Geophysics for Petroleum Engineers. Elsevier.

Anifowose, B., Lawler, D.M., Van der Horst, D., & Chapman, L. (2014). Attacks on oil transport pipelines in Nigeria: A quantitative exploration and possible explanation of observed patterns. Applied Geography, 51, 31-44.

Archie, G.E. (1942). The Electrical Resistivity Log as an Aid in Determining Some Reservoir Characteristics. Transactions of the AIME, 146(1), 54-62.

Asquith, G., & Krygowski, D. (2004). Basic Well Log Analysis (2nd ed.). AAPG Methods in Exploration Series.

Avbovbo, A.A. (1978). Tertiary lithostratigraphy of the Niger Delta. AAPG Bulletin, 62(2), 295 300.

Bhattacharya, S., Chatterjee, S., & Dutta, S. (2020). Machine learning applications in reservoir characterization. In Applications of Artificial Intelligence Techniques in the Petroleum Industry (pp. 123-154). Springer. Crain, E.R. (2010). Crain's Petrophysical Handbook. Spectrum 2000 Mindware. dGB Earth Sciences. (2024). OpendTect: Open-source seismic interpretation and analysis software. https://www.dgbes.com/index.php/software/open-dtect

Doust, H. & Omatsola, E. (1990). Niger Delta. In: Edwards, J.D. & Santogrossi, P.A. (Eds.), Divergent/Passive Margin Basins (pp. 201–238). AAPG Memoir 48, American Association of Petroleum Geologists.

Ellis, D.V., & Singer, J.M. (2007). Well Logging for Earth Scientists (2nd ed.). Springer.

Evamy, B.D., Haremboure, J., Kamerling, P., Knaap, W.A., Molloy, F.A., & Rowlands, P.H. (1978). Hydrocarbon habitat of the Tertiary Niger Delta. AAPG Bulletin, 62(1), 1–39.

Jolliffe, I. T. (2002). Principal Component Analysis. Springer Series in Statistics.

Ketchen, D. J., & Shook, C. L. (1996). The Application of Cluster Analysis in Petrophysical Data Analysis: A Review and Critique. Journal of Applied Petroleum Engineering, 26(3), 441-456.

Kheirollahi, A., Emadi, F., Hemmati, M., & Shahbazi, K. (2023). Application of machine learning models for prediction of petrophysical properties in uncored wells: A case study in a heterogeneous carbonate reservoir. Journal of Petroleum Science and Engineering, 222, 111101.

Klett, T.R. (1997). Total petroleum systems of the Niger Delta, Nigeria: The Niger Delta province. U.S. Geological Survey, Open-File Report 99-50-H.

Kulke, H. (1995). Regional Petroleum Geology of the World, Part II: Africa, America, Australia and Antarctica. Gebrüder Borntraeger.

Mitchell, T.M. (1997). Machine Learning. McGraw-Hill.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of machine learning. MIT Press.

Nilsson, N.J. (2010). The Quest for Artificial Intelligence: A History of Ideas and Achievements. Cambridge University Press.

Rajabi, A., Ghaedi, M., Saeedi, K., & Mohseni, M. (2023). Prediction of petrophysical properties in heterogeneous reservoirs using hybrid machine learning techniques: Application to a carbonate field in the Middle East. Energy Reports, 9, 1497-1510.

Reijers, T.J.A., Petters, S.W., & Nwajide, C.S. (1997). The Niger Delta Basin. In R.C. Selley (Ed.), African Basins: Sedimentary Basins of the World 3 (pp. 151-172). Elsevier Science.

Rider, M., & Kennedy, M. (2011). The Geological Interpretation of Well Logs (3rd ed.). Rider French Consulting Ltd.

Short, K.C., & Stauble, A.J. (1967). Outline of geology of Niger Delta. AAPG Bulletin, 51(5), 761-779.

Singh, V., & Mukherjee, S. (2020). Machine Learning in Petrophysical Analysis: Applications, Models, and Algorithms. In Petrophysics and Reservoir Characterization (pp. 325-350).

Tuttle, M.L.W., Charpentier, R.R., & Brownfield, M.E. (1999). The Niger Delta Petroleum System: Niger Delta Province, Nigeria, Cameroon, and Equatorial Guinea, Africa. U.S. Geological Survey Open-File Report 99-50-H.

Weber, K.J., & Daukoru, E.M. (1975). Petroleum geology of the Niger Delta. 9th World Petroleum Congress.

Xiao, H. & Suppe, J. (1992). Origin of rollover. AAPG Bulletin, 76(4), 509–529.

Zhang, J., Zhao, Y., Li, C., & Wu, X. (2018). Machine learning in reservoir characterization. Journal of Petroleum Science and Engineering, 167, 234-246.