Integrated Hybrid Approach for Litho-facies Log Prediction Using Machine Learning Algorithms in Hydrocarbon Fields in the Niger Delta

[1] Jerome Asedegbega, [1] Alexander Nwakanma and [2] Richard Alakuko
[1] Gcube Integrated Services Limited, Lagos, Nigeria
[2] Federal University of Technology Akure

ABSTRACT

Well log interpretation is a vital procedure in the oilfield field development, especially because of its input throughout the exploration and production stages. This process becomes even more important as there exist few available alternatives to compensate for the lack of core samples in the study of facies (lithology) variation in a well. This is more complicated as the usual conventional expert oriented interpretational technique, are often associated with interpretational risk of subjective or technical errors. Thus, to overcome these challenges a Machine Learning (ML) approach was applied in this study to several hydrocarbon (HC) fields with the goal of predicting HC bearing units, in form of facies logs where the values of these logs were labels of classes separated based on their lithologic characteristics. Several ML algorithms were trained on the available log parameters including the litho-facies log to predict for target wells due to their superior performance, high evaluation rate, and low false discovery rate. This proposed hybrid ML algorithms approach was based on trained and tuned data relying on the expert's knowledge of the reservoirs. The resulting ML algorithms here, can be adapted to scan for overlooked pay and significantly mitigate the risk of false HC discoveries, which will further enhance future wells development in HC fields of similar geology in the Niger Delta.

Keywords: ML Algorithm, Hybrid, Facies Log, Hydrocarbon

INTRODUCTION

Well log interpretation is a vital procedure in the oil and gas upstream industry, especially because of its input throughout the exploration and production stages. This process becomes even more important as there exist few available alternatives to compensate for the lack of core samples in the study of facies variation (lithology) in a well. This is more complicated as the usual conventional expert oriented interpretational technique, are often associated with interpretational risk of subjective or technical errors.

Facies are uniform sedimentary bodies of rock which are distinguishable enough from each other in terms of physical characteristics (e.g., sedimentary structure, grain sizes), deposited under the action of a relatively uniform hydrodynamic regime in each depositional setting, where different types of facies are based on the following properties which include sedimentary facies, lithofacies, and seismic facies (Ibinabo, 2020).

The physical and organic characteristics found in these rock units usually provide some insight into the different process and systems (e.g., depositional environments) which may have occurred within the region of deposition, therefore, different combinations of facies with physical models and other geological data can help provide informative low-dimensional models of the geologic region, leading to better insights regarding the geology of the area (Ibinabo, 2020).

Reservoir rock characteristics are generally classified into geological facies to quantify the uncertainty in the classification of predictive modelling, accurate identification of reservoir facies is important to help understand geological variation in a proven target (Asedegbega *et al.,* 2021). For this study, seven (7) predictive machine learning models were applied to predict reservoir facies using thirty-eight (38) wells spread across twelve (12) onshore and offshore fields Niger Delta. Three (3) random blind test wells were selected to evaluate the measure of certainty for the predictive models built from training thirty-five (35) wells.

The authors wish to acknowledge the efforts of the team for the work and Gcube Integrated Services Limited for authorising the publications.

The hybrid machine learning model approach adopted was based on training and parameter tuning, of the predictive models such as Neural Network (NN), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), Extra Tree Model (ETM), K-Nearest Neighbour (KNN) and Decision Tree (DT) alongside the knowledge of the reservoirs. The resulting machine learning models can be adapted to scan for bypassed oil and significantly mitigate the risk of false hydrocarbon discoveries, which will further enhance future wells development in hydrocarbon fields of similar geology within the Niger Delta.

Geology of the Study Area

The Niger Delta is a prograding depositional complex within the Cenozoic Formation of Southern Nigeria. The basin is an extensional rift basin located in the Niger Delta and the Gulf of Guinea on the passive continental margin near the Western coast of Nigeria with the proven access to Cameroon, Equatorial Guinea and Sao Tome and Principle. It extends from the Calabar Flank and the Abakaliki Trough in Eastern Nigeria to the Benin Flank in the west and it opens to the Atlantic Ocean in the southern territory. The basin is very complex, and it carries high economic value as it contains a very productive hydrocarbon system. The Niger Delta basin is one of the largest subaerial basins in Africa. It has a subaerial area of about 75,000 km2, a total area of 300,000 km2, and sediment fill of 500,000 km3. The delta protrudes into the Gulf of Guinea as an extension from the Benue Trough and Anambra Basin Provinces (Evamy *et al.,* 1978). The composite tertiary sequence of the Niger Delta consists, in ascending order of the Akata, Agbada and Benin formations (Ejedawe *et al.,* 2002). They are composed of estimated 28,000 ft (8,535 m) of section of the approximate depocenter in the central part of the delta (Akpabio *et al.,* 2014). There is decrease in age basin ward, reflecting the overall regression of depositional environments within the Niger Delta clastic wedge, stratigraphic equivalent units to these three formations in eastern Nigeria. The formations reflect a gross coarsening upward progradational clastic wedge (Akpabio and Ojo, 2018), deposited in marine deltaic, and fluvial environments (Inyang *et al.,* 2015). The stratigraphic distribution of these rocks is poorly understood because of the lack of drilling information and outcrops (Akpabio *et al.,* 2014). Evamy *et al.,* (1978) described the mode formation, distribution and importance of growth faulting in the Niger Delta development. The growth faults are contemporaneous and more or less continuously active with deposition such that their throws increase with depth. The growth fault may be listric, typically cuspate normal faults, which flatten with depth into the thick clastic shaly sequence of the Akata Formation. The continuous growth of the faults after their inception, allows for greater sedimentation on the down-thrown blocks relative to the up-thrown blocks (Evamy *et al.,* 1978).

Location of Study Area

Figure 1: Map showing the location of wells within the Niger Delta (Source: Folium Library in Python).

MATERIALS AND METHODS

Materials

This study was carried out utilizing thirty-eight (38) wells acquired across twelve (12) fields, both within onshore and offshore areas of the Niger Delta. Available log parameters comprise of gamma ray log (GR), neutron log (NPHI), density log (RHOB), resistivity (RES), and shale volume (Vsh). Python programming language using Jupyter Notebook integrated development environment (IDE) was used as the interpretational tool to carry out the study.

Methods

The workflow utilized in carrying out this study include data preparation, data preprocessing, train model, evaluate model, test model and visualize model predictions. This workflow can be shown in Figure 2.

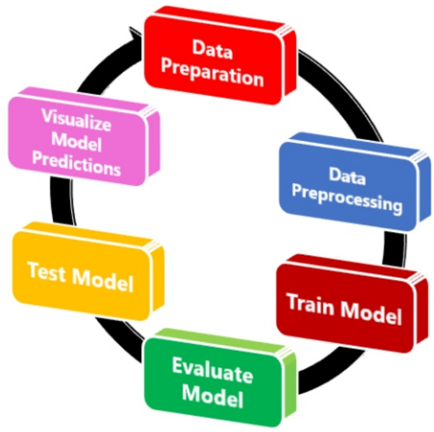

Figure 2: Integrated workflow utilized to carry out this study.

Data Preparation

This process involved a tedious procedure, where the well log data containing the five (5) log parameters gamma ray log (GR), neutron log (NPHI), density log (RHOB), resistivity (RES), and shale volume (Vsh) are prepared. Thirty-eight (38) wells containing this log parameters were concatenated during the process and exported to be used as a CSV file for the analysis.

Data Preprocessing

The CSV file is imported by reading the file. Python packages like, NumPy, Matplotlib, Seaborn and Pandas libraries were adopted for data wrangling and visualization task.

Exploratory Data Analysis

This procedure is carried out during the preprocessing workflow in order to carry out a critical process (initial performing investigations) on the dataset in view of discovering patterns, spot anomalies, test hypothesis and check assumptions with the help of summary statistics and graphical representations. Within this process the shape of the data is investigated and here, 548828 rows and 8 columns were recorded. The data info enabled the visualization of the data types. The log parameters were float64 for 6 columns, int64 for 1 column and 1 object columns while the memory usage indicated the data contain 33.5+ mb space of data.

Missing values were investigated where values of -999.25 was replaced as NAN (Not a Number) values. These missing values were filled for using the mean fillna function, that helps to replace the missing values with the estimated mean.

The descriptive statistical analysis which reveals the statistical summary of the data showing the count, mean, standard deviation, minimum, maximum, 25th, 50th, and 75th percentile of the data. Unique values obtained for the facies revealed that the count of sand, shale and shaly sand samples were estimated. A regression pairplot was visualized to reveal the skewness of the log parameters, the correlation pattern of the log plots whether negative or positive, which relates to the heatmap visualization of logs. A boxplot was also estimated and visualized to indicate outliers for each of the log parameters which revealed the degree of anomaly for each of the log parameters.

Model Training

The log parameter data was used to train the various machine learning classifiers algorithms to predict discrete samples of the different facies labels. Seven supervised learning algorithms: Neural Network (NN), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), Extra Tree Model (ETM), K-Nearest Neighbour (KNN) and Decision Tree (DT) were selected at random from the scikit learn package and utilized to build models to classify facies label as either shale, sand or shaly sand. Out of thirty-eight (38) wells, three (3) blind test wells (OB_2, E4, and P_3 wells) were randomly selected while the remaining thirty-five (35) wells were used to train the model. The train data was divided into 67% - 33%, 67 percent train and 33 percent test during model building. The stratified function was initiated to cover for the imbalance samples of the facies labels. The 67% train data was further divided into X_train and y_train while the 33% test data into X_test and y_test using the train_test_split method. A fit function is passed to the X_train and y_train while a predict function is passed to the X_test and y_test to evaluate the performance.

Model Evaluation

For the model evaluation, the measure of performance of the various model predictions were analyzed using confusion matrix, classification report, jaccard index and f1score.

RESULTS AND DISCUSSION

Well log visualization of the data revealing the log parameters for some of the training wells are shown on Figures 3, 4, 5, 6 and 7 for quick-look analysis.

Exploratory Data Analysis

From the data preprocessing workflow, the result of the exploratory data analysis carried out revealed that from the shape of the data, the dataset had 548828 rows and 8 columns. The data investigated showed the following missing or NAN values: Gamma ray, 255 samples, neuron 241475 samples, density 149513 samples, and resistivity 61665 samples (Figure 8). Unique values for the facies revealed that the count of sand, shale and shaly sand samples were estimated at 253687, 234028, and 10746 samples respectively (Figure 9). Results of the regression pairplot revealed the skewness of the distribution of facies labels across each of the log plots as shown in Figure 10, while the boxplot visualization revealed outlier values for gamma ray, neuron, density, resistivity, and shale volume (Figure 11). The result of the correlation of features (log parameters) showed that neutron and resistivity were negatively correlated while positive correlation was established with gamma ray, density and shale volume which showed how interrelated the features were to the facies labels (Figure 12).

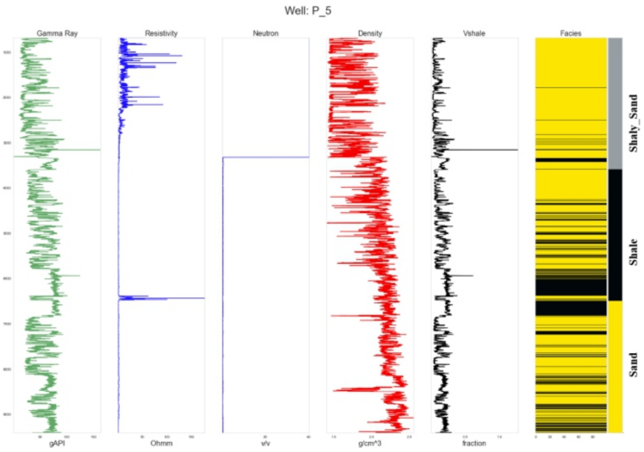

Figure 3: Well Log display for Well P_5.

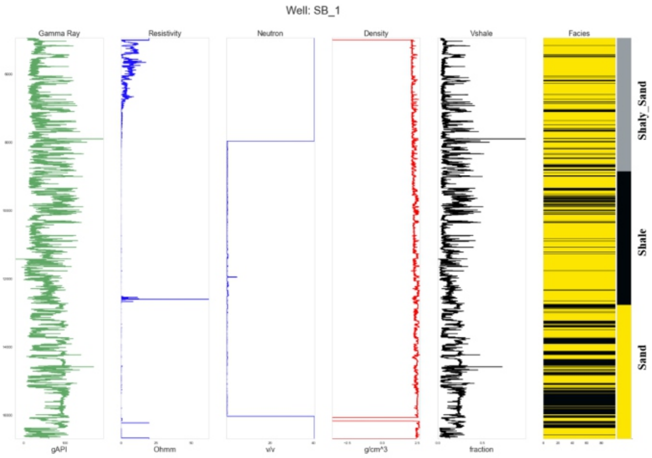

Figure 6: Well Log display for Well SB_1.

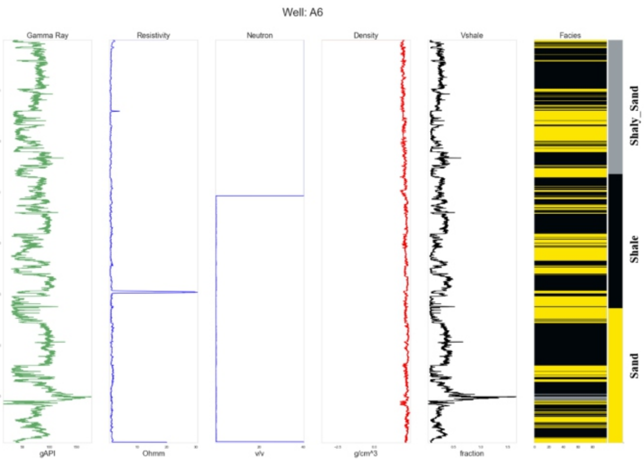

Figure 4: Well Log display for Well A6.

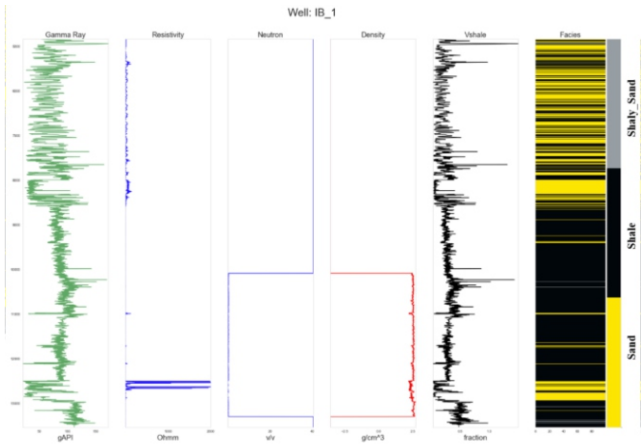

Figure 7: Well Log display for Well IB_1.

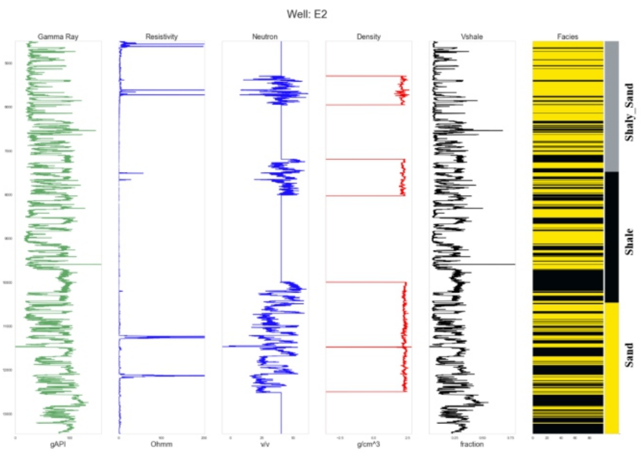

Figure 5: Well Log display for Well E2.

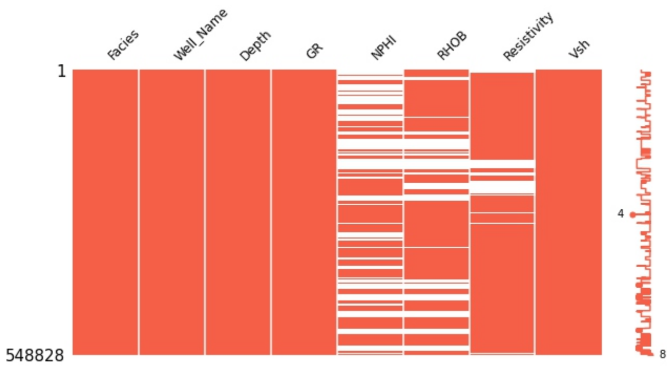

Figure 8: Missing values for gamma ray, neuron, density, and resistivity.

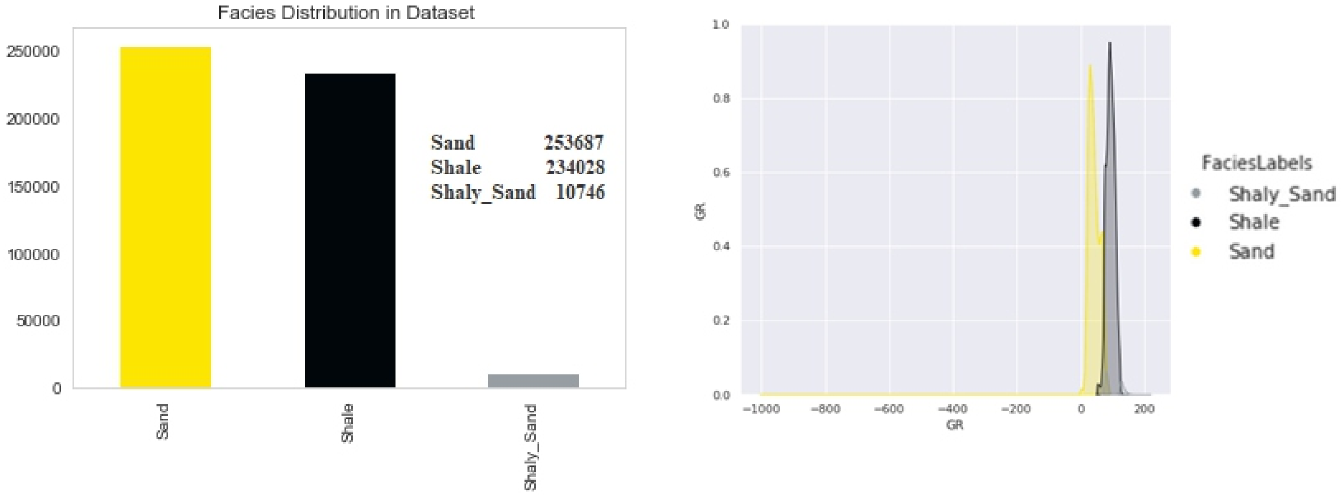

Figure 9: Facies label distribution of the data and the kernel density estimate showing the skewness of the facies.

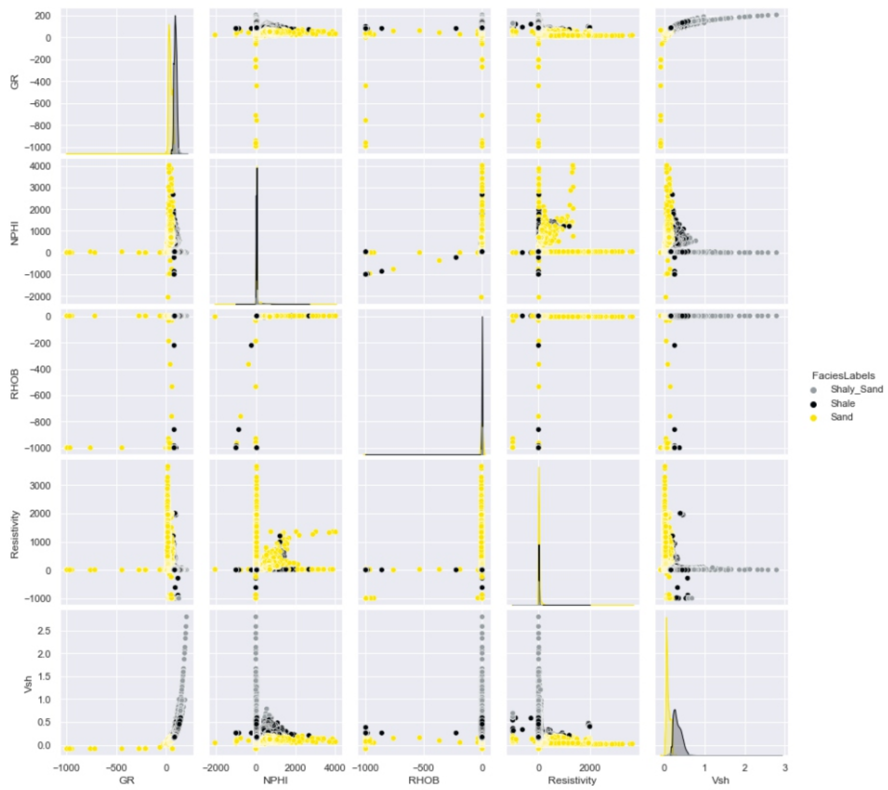

Figure 10: A Regression pairplot showing nature and distribution of log plot for log parameters.

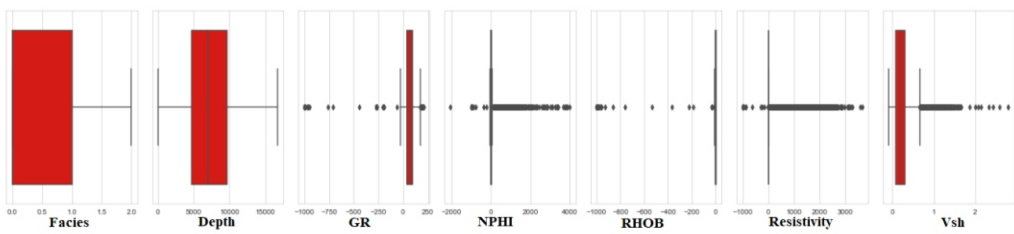

Figure 11: Boxplot distribution of log parameters showing outlier values.

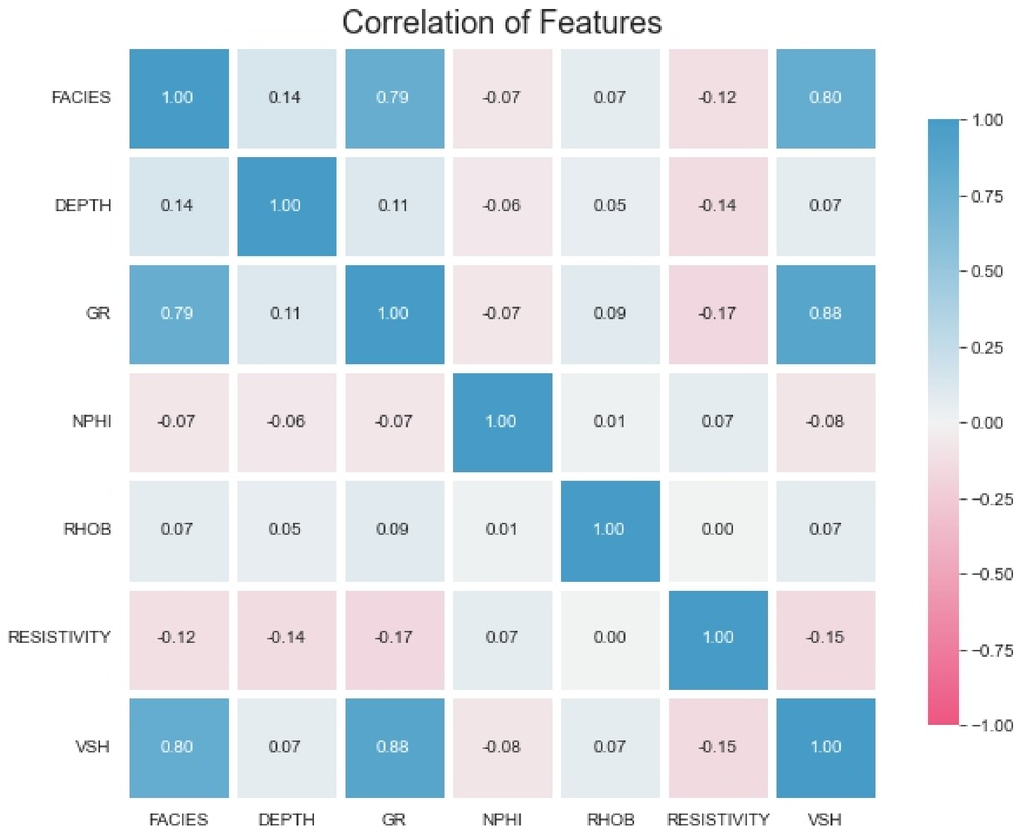

Figure 12: Correlation of features showing interrelated features and facies labels.

Model Building

Support Vector Machine

For the support vector machine model, regularization parameter which gives the inversely proportional strength of the regularization to C as 2, sigmoid kernel type, a degree of 3 and a scale gamma type were used to fit the X_train and y_train data. The evaluation of the model was 0.825 and 0.895 for Jaccard index and F1 score, respectively. The classification report revealed a precision score of 0.92 for sand, 0.86 for shale, 0 for shaly sand, a recall score of 0.89 for sand, 0.96 for shale, and 0 for shaly sand. This revealed that the model built could not effectively predict for shaly_sand facies label but did well in classifying sand and shale facies labels. This model built was further evaluated using the confusion matrix (Figure 13a) where 74656 samples of sand were predicted correctly, 74120 samples of shales were also predicted correctly but not sample of shaly_sand was predicted correctly.

Random Forest Model

For the random forest model, n_estimator of 20, gini criterion, maximum depth of 1 and bootstrap as False were used to fit X_train and y_train data. The evaluation of the model was 0.788 and 0.873 for Jaccard index and F1 score, respectively. The classification report revealed a precision score of 0.88 for sand, 0.89 for shale, 0 for shaly sand, a recall score of 0.92 for sand, 0.89 for shale, and 0 for shaly sand. This revealed that the model built could not effectively predict for shaly_sand facies label but did well in classifying sand and shale facies labels. This model built was further evaluated using the confusion matrix (Figure 13b) where 76645 samples of sand were predicted corrected, 78579 samples of shales were also predicted correctly but not sample of shaly_sand was predicted correctly.

Extra Tree Model

For extra tree model, n_estimator of 20, gini criterion, and a maximum depth of 3 were used to fit X_train and y_train data. The evaluation of the model was 0.673 and 0.798 for Jaccard index and F1 score, respectively. The classification report revealed a precision score of 0.90 for sand, 0.74 for shale, 0 for shaly sand, a recall score of 0.73 for sand, 0.93 for shale, and 0 for shaly sand. This revealed that the model built could not effectively predict for shaly_sand facies label but did well in classifying sand and shale facies labels. This model built was further evaluated using the confusion matrix (Figure 13c) where 61280 samples of sand were predicted corrected, 71521 samples of shales were also predicted correctly but not sample of shaly_sand was predicted correctly at all.

K Nearest Neighbour Model

For K Nearest Neighbour, n_neighbor of 2, leaf size of 10, and a power parameter for the minkowski metric P of 5 were fitted to the X_train and y_train data. The evaluation of the model was 0.908 and 0.95 results for Jaccard index and F1 score, respectively. The classification report revealed a precision score of 0.94 for sand, 0.97 for shale, 0.87 for shaly sand, a recall score of 0.98 for sand, 0.94 for shale, and 0.54 for shaly sand. This revealed that the model built could effectively predict for sand, shale, and shaly sand facies labels. This model built was further investigated using the confusion matrix (Fig. 13d) which revealed that 82082 samples of sand, 72549 samples of shale and 1905 samples of shaly sand were all predicted correctly. This particular model gave the best result yield in classifying for the facies labels.

Logistic Regression Model

For the logistic regression model building, dual was set to true, an inverse of regularization C of 1, solver of liblinear type, and a maximum iteration of 0.00000001 were used to fit the X_train and y_train data. The evaluation of the model was 0.449 and 0.613 results for Jaccard index and F1 score, respectively. The classification report revealed a precision score of 0.62 for sand, 0.64 for shale, 0 for shaly sand, a recall score of 0.75 for sand, 0.52 for shale, and 0 for shaly sand. This revealed that the model built could not effectively predict for shaly_sand facies label but did well in classifying sand and shale facies labels. This model built was further evaluated using the confusion matrix (Figure 13e) where 62924 samples of sand were predicted corrected, over 40000 samples of shales were also predicted correctly but not sample of shaly_sand was predicted correctly at all.

Decision Tree Model

For decision tree model building, gini criterion, a random spitter, a minimum sample split of 50 and a maximum depth of 1 were used to fit X_train and y_train data. The evaluation of the model was 0.565 and 0.715 results for Jaccard index and F1 score, respectively. The classification report revealed a precision score of 0.95 for sand, 0.64 for shale, 0 for shaly sand, a recall score of 0.54 for sand, 0.97 for shale, and 0 for shaly sand. This revealed that the model built could not effectively predict for shaly_sand facies label but did well in classifying sand and shale facies labels. This model built was further evaluated using the confusion matrix (Figure 13f) where 45281 samples of sand were predicted corrected, 75213 samples of shales were also predicted correctly but not sample of shaly_sand was predicted correctly at all.

Neural Network Model

For the neural network model, the multi-layer perceptron classifier was built, an alpha of 0.01, maximum iteration of 1, an initial learning rate of 0.0001, an Adam solver, hidden layer size of 100 and shuffle as False were used to fit X_train and y_train data. The evaluation of the model was 0.647 and 0.779 results for Jaccard index and F1 score, respectively. The classification report revealed a precision score of 0.89 for sand, 0.72 for shale, 0 for shaly sand, a recall score of 0.71 for sand, 0.91 for shale, and 0 for shaly sand. This revealed that the model built could not effectively predict for shaly_sand facies label but did well in classifying sand and shale facies labels. This model built was further evaluated using the confusion matrix (Figure 13g) where 59191 samples of sand were predicted correctly, over 70000 samples of shales were also predicted correctly but not sample of shaly_sand was predicted correctly at all.

Blind Test Well

The results of the blind test wells (OB_2, E4, and P_3 wells) revealed after applying the seven models Neural Network (NN), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), Extra Tree Model (ETM), K-Nearest Neighbour (KNN) and Decision Tree (DTM) built from training thirty-five (35) wells to test on three (3) well which were not included in training the models were robust enough to accurately classify for the facies labels for each of the blind test data. For OB_2 well (Figure 14), the evaluation result estimated was 0.91 & 0.95, 0.72 & 0.82, 0.89 & 0.93, 0.92 & 0.96, 0.58 & 0.71, 0.64 & 0.76, and 0.70 & 0.81 for Jaccard index and F1 score respectively, for E4 well (Figure 15), the evaluation result estimated was 0.64 & 0.78, 0.65 & 0.79, 0.65 & 0.72, 0.70 & 0.82, 0.69 & 0.81, 0.69 & 0.79, and 0.65 & 0.78 for Jaccard index and F1 score respectively, and for P_3 well (Figure 16), the evaluation result estimated was 0.90 & 0.95, 0.56 & 0.72, 0.90 & 0.95, 0.94 & 0.97, 0.40 & 0.57, 0.50 & 0.67, and 0.56 & 0.72 for Jaccard index and F1 score respectively, for support vector machine, random forest model, extra tree model, k nearest neighbour, decision tree model, linear regression model and neural network, respectively. Tables 1, and 2 reveals the description of the summary table for the evaluation of jaccard index and F1 scores for the various models built, for the training data and for each of the blind test wells. The K-Nearest Neighbour model did far better amongst the other six model with high prediction scores recorded for both Jaccard index and F1 score in well OB_2 (Figure 17), well E4 (Figure 18, and well P_3 (Figure 19). This model has therefore proved beyond every doubt to be efficient and accurate to predict for other wells within the same geologic terrain.

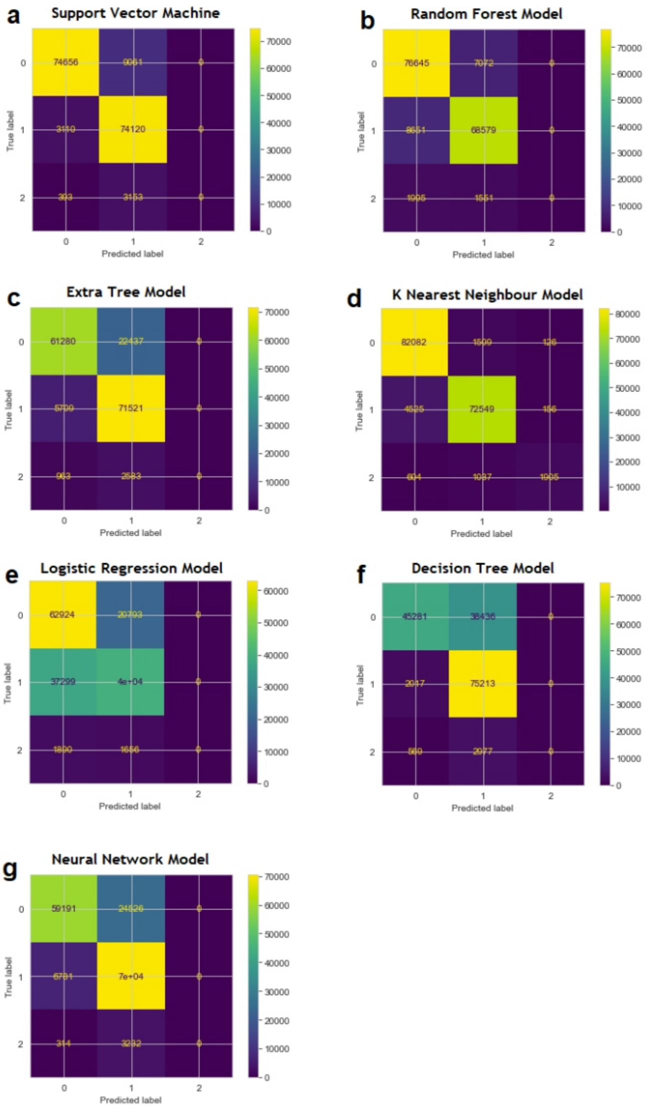

Figure 13: Plot of Confusion Matrix for Support Vector Machine, Random Forest, Extra Tree, K Nearest Neighbour, Logistic Regression, Decision Tree and Neural Network Models.

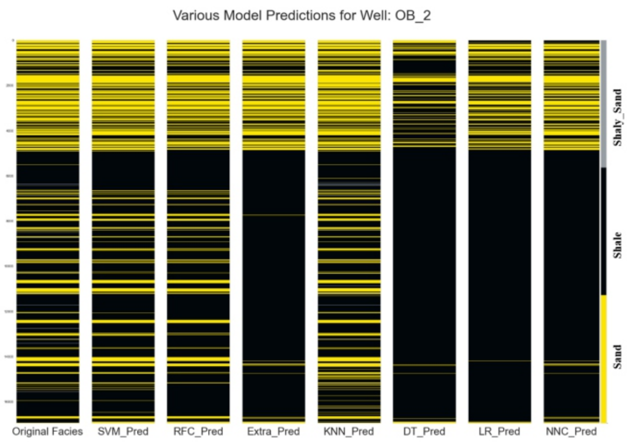

Figure 14: Results showing various model predictions for the OB_2 well.

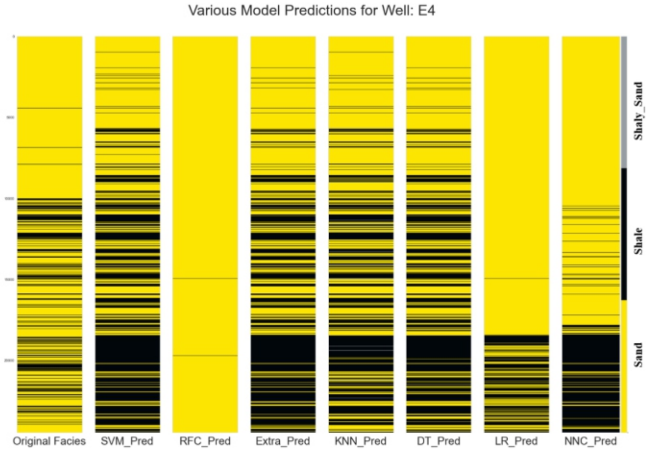

Figure 15: Results showing various model predictions for the E4 well.

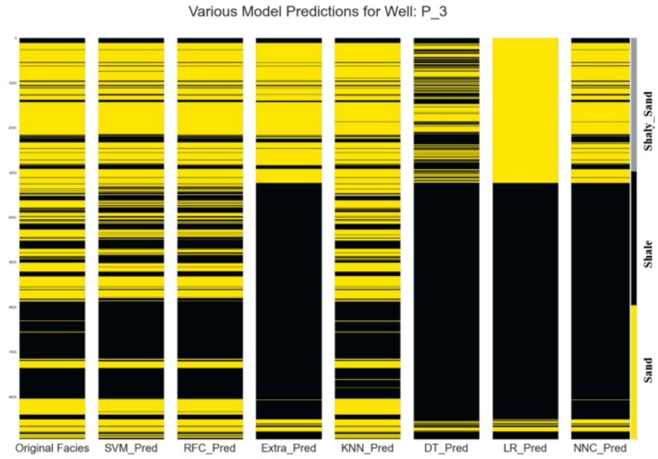

Figure 16: Results showing various model predictions for the P_3 well.

Table 1: Summary of Evaluation Metrics showing the Jaccard Index and F1 score for the various model results for the Training data and OB_2 Well.

	Training data		OB_2 Well	
Model type	**Jaccard index**	**F1-Score**	**Jaccard index**	**F1-Score**
SVM	0.82	0.89	0.91	0.95
RFC	0.79	0.87	0.72	0.82
ETM	0.67	0.80	0.89	0.93
KNN	0.91	0.95	0.92	0.96
DT	0.65	0.78	0.58	0.71
LR	0.45	0.61	0.64	0.76
NNC	0.65	0.78	0.70	0.81

Table 2: Summary of Evaluation Metrics showing the Jaccard Index and F1 score for the various model results for E4 Well and P_3 Well.

	E4 Well		P_3 Well	
Model type	Jaccard index	F1-Score	Jaccard index	F1-Score
SVM	0.64	0.78	0.90	0.95
RFC	0.65	0.79	0.56	0.72
ETM	0.65	0.72	0.90	0.95
KNN	0.70	0.82	0.94	0.97
DT	0.69	0.81	0.40	0.57
LR	0.69	0.79	0.50	0.67
NNC	0.65	0.78	0.56	0.72

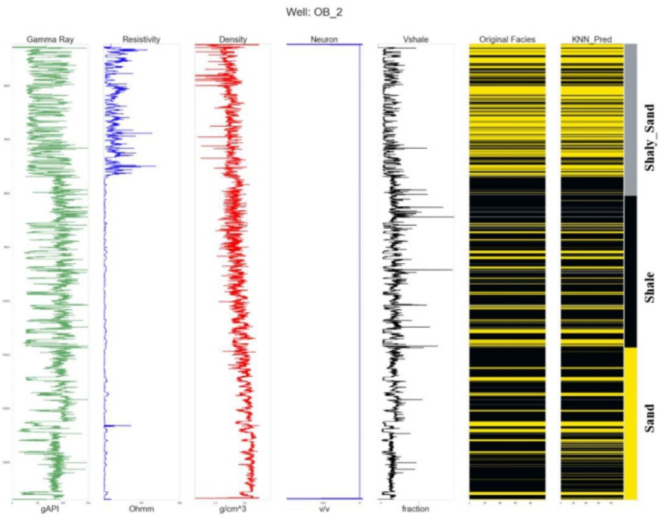

Figure 17: Results showing K-Nearest Neighbour Model Prediction for OB_2 Well.

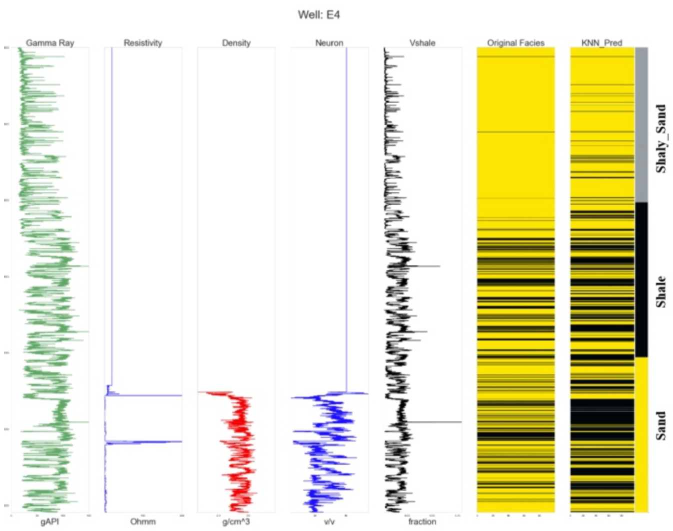

Figure 18: Results showing K-Nearest Neighbour Model Prediction for E4 Well.

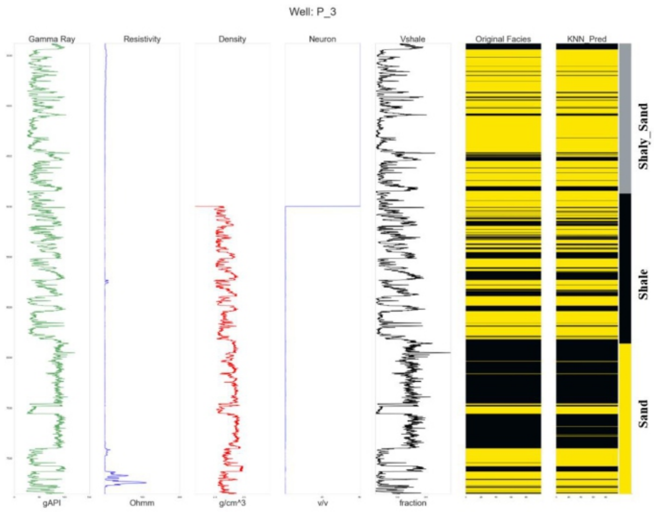

Figure 19: Results showing K-Nearest Neighbour Model Prediction for P_3 Well.

CONCLUSIONS

This study has demonstrated the efficiency of building machine learning models that can invariably help resolve one of the many challenges in exploration and production. The challenge of identifying lithology types in places with little or no well penetration is overcome using the constructed model for the different geologic settings. This machine learning technique is unique and precisely important for both geologists, geophysicists and petroleum engineers to easily apply and also identify facies (lithology), especially during exploration and drilling phases as it can aid in predicting formation types or depth of penetration ahead of drill bit.

The K Nearest Neighbour (KNN) model was found to be robust to accurately classify the facies labels amongst all the machine learning model built and utilized in blinding test in several wells. The measure of performance proved the fidelity of the KNN model in these fields and thus, can be carefully adapted to predict for missing well logs(s) for future exploration endeavours in the studied areas. The failure of other models to efficiently predict lithofacies in the blind tested wells, give credence to the fact that, it is not one model fit-for-all and to improve models level of performance in a particular well, parameter tuning with in-situ data is needed.

REFERENCES CITED

Akpabio, I., Ibuot, J. C., Agbasi O. E. & Ojo, O. T. (2014). Petrophysical Characterization of eight wells from Wire-line Logs, Niger Delta Nigeria. Asian Journal of Applied Science , 2(2) 105-109.

Akpabio, I. O. & Ojo, O. T. (2018). Characterization of hydrocarbon reservoir by pore fluid and lithology using elastic parameters in an X field, Niger Delta, Nigeria. International Journal of Advance Geosciences, 6 (2) 173-177.

Asedegbega, J., Ayinde, O. & Nwakanma, A. (2021). Application of Machine Learniing For Reservoir Facies Classification in Port Field, Offshore Niger Delta. One Petro (SPE Nigeria Annual International Conference and Exhibition, Lagos, Nigeria 2021), https://doi.org/10.2118/207163-MS.

Ejedawe, J., Fatumbi, A., Ladipo, K. & Stone, K.(2002). Pan - Nigeria exploration well look - back (Post-Drill Well Analysis). Shell Petroleum Development Company of Nigeria Exploration Report.

Evamy, B. D., Haremboure, J., Kamerling, P., Knaap, W. A., Molloy, F. A. & Rowlands, P. H. (1978). Hydrocarbon habitat of Tertiary Niger Delta. American Association of Petroleum Geologists Bulletin, 62, 277-298.

Ibinabo, B. (2020). Facies classification using unsupervised machine learning in geoscience. Towards Data Science, https://towardsdatascience.com/facies-classification-using-unsupervised-machine-learning-in-geoscience-8b33f882a4bf.

Inyang, N. J., Okwueze, E. E., & Agbasi, O. E. (2015). Detection of Gas Sands in the Niger Delta by Estimation of Poisson's Dampening-Factor (PDF) Using Wireline Log Data. Geosciences, 5, 46-51. doi:10.5923/j.geo.20150501.06.