European Journal of Computer Science and Information Technology 11 (5), 30-49, 2023
Print ISSN: 2054-0957 (Print),
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

A Comparison of Two Machine Learning Techniques for the Prediction of Initial Oil in Place in the Niger Delta Region

[1]Ekemini Johnson, [2]Okure Obot, [2]Udoinyang Inyang and [2]Julius Akpabio
[1]Ritman University, Ikot Ekpene, Nigeria. [2]University of Uyo, Uyo, Nigeria

doi:https://doi.org/10.37745/ejcsit.2013/vol11n53049 Published October 21 2023

Citation: Johnson E., Obot O., Inyang U.,and Akpabio J.. (2023) Comparison of Two Machine Learning Techniques for the Prediction of Initial Oil in Place in the Niger Delta Region, European Journal of Computer Science and Information Technology 11 (5), 30-49

ABSTRACT: *Conventionally, the knowledge of experts on the drilling features of a potential oil well is practically used to predict the volume of initial oil in place. Experts used different knowledge-based models such as volumetric, material balancing, analogy to predict the initial oil in place. In this study, 816 datasets were collected from Shell petroleum development company (SPDC) where the volumetric method is used for their prediction. These datasets were preprocessed and applied on two machine learning techniques of random forest and supervised vector regressor to predict the initial oil in place and the results obtained were compared with that obtained from SPDC.The results of computation using 4 principal features from the 9 features were closer to that obtained from SPDC than the computations using all the 9 features. The results of computations with random forest were also compared with that of supervised vector regressor. The results of random forest covary strongly (0.970) with the field results more than that of the support vector regressor (0.832). The uniqueness of this study is shown in the use of 4 predicting features (independent variables) to obtain prediction values that are very close to that obtained in the field with 9 features. This is obtained with random forest, so it can be recommended as a reliable machine technique for the prediction of initial oil in place in the Niger delta region.*

KEYWORD: Machine learning, Randomforest, Support vector regressor, Volumetric, Material balance, Analogy, Initial-oil-in-place. Niger Delta.

INTRODUCTION

The knowledge of the volume of initial oil in place in a reservoir is needed before drilling of an oil well begins. This is necessary to determine the cost effectiveness of the process of exploration. Sometimes the estimates undertaken to determine the Initial Oil in-Place (IOIP) mislead experts to huge financial loss and the attendant environmental degradation. This makes such experts to

devise various methods of prediction with a view to finding an optimal method. Such methods include; the volumetric method which uses the size of the reservoir and the features of rocks and fluids in the surrounding areas,the material balancing method uses a mathematical equation that establishes a relationship between the volume of oil and some other parameters such as water, gas, pressure on the reservoir etc, andthe analogy method that predicts based on the similarity features of some reservoirs.

These methods present one common weakness of predicting an over-estimate or under-estimate volume because the parameters are based on the knowledge of individual experts so are very subjective and prone to error. The knowledge of the experts is driven by rules and like all rules t are brittle more so when presented with imprecise and noisy data (Obot et al, 2022). Observations of the behaviour of the operations of the reservoir over a period of time could offer a better alternative if such observations are reliable and are subjected to a reliable method of prediction or forecast model. Classical data-driven forecast models like moving average, weighted average and exponential smoothing have been applied successfully in business forecast. Unfortunately, these models lack the requisite intelligence to drive a critical and sensitive venture like oil exploration and drilling.

Artificial intelligence (AI) mimics a human intelligence and has demonstrated tremendous successes in prediction, classification, scheduling among other tasks. Recently, machine learning has exhibited great reliability and promises as a vehicle driving AI. Machine learning algorithms learn from data to make prediction in a supervisory or non-supervisory mode. The techniques used to realize this include; Random Forest, Supervised Vector Machine, Artificial Neural Networks, Naïve Bayes among several others.

Random forest (RF) combines the results of random multiple decision trees to give a single result. It can handle both classification and regression estimation using random datasets and features. Support vector machine (SVM) is used for both linear regression estimation and binary classification by taking advantage of its ability to separate the thin hyperplane of the parameters. Artificial Neural Networks (ANN) mimic the biological neurons, it is inspired by the biological knowledge of the brain and comprises a large number of simple interconnected networks that help in classification and regression tasks. Naïve Bayes algorithm is based on Bayes theorem and is used for classification task to model the distribution of inputs to an assigned class.

Aside from having a reliable model for prediction, one also needs a reliable source of data in order to get an accurate and reliable result. To this end, 816 datasets of past drillings and their corresponding outputs were gathered from Shell Petroleum Development Company (SPDC) in Nigeria. The datasets are made up of 9 features which are independent variables and 1 dependent variable, the predicted value. The nine features are reservoir permeability, reservoir porosity, water

cut, oil rate, gas oil ratio, column thickness, liquid rate, gas prediction rate, and water production rate.

This study is therefore set to compare two machine learning algorithms for predicting the volume of initial oil in place in the Niger Delta region. The objectives include to; (i) collect datasets of reservoir parameters from SPDC that had been used to predict the volume of oil for some years. (ii) implement a random forest machine language algorithm (iii) implement a support vector regression algorithm (iv) Evaluate the results of each of the algorithms (v) Compare the results of the evaluation with respect to the results obtained from the field with a view to adopting the closet of the two algorithms to the prediction done by the experts in the field.

The rest of the paper is organized as follows; in Section 2, related literatures are reviewed while in Section 3, the collected datasets and the methodology of the research are described. In Section 4, the results of the experiment carried out are shown and discussed. Section 5 proffers recommendations and draws the conclusion of the study.

REVIEW OF RELATED LITERATURE

Based on the fracture density obtained by core analysis in a carbonate reservoir located in the Ordos Basin, in northwest China, three types of fracture density (low fracture density, medium fracture density, and high fracture density) of the target formation were identified inLi et al (2018). The study investigated the effect of fractures on acoustic logging signals in the time and frequency domains by the Hilbert-Huang Transform (HHT) and extracted 11 features in the time domain and nine features in the frequency domain. The results indicate that the fracture density has a greater effect on the attenuation of intrinsic mode function 2 (IMF2) and IMF3 components for three different types of formation by empirical-mode decomposition analysis. The energy of the Stoneley wave and S-wave has higher sensitivity than the P-wave. Compared with the time domain, the distribution in the high-frequency domain has a greater correlation with fracture density by the Hilbert spectrum and marginal spectrum.The low accuracy (52-59%) of features in time domain cannot effectively reflect the degree of fracture development.

Li et al (2020) investigates the semi-supervised learning methodfor lithology identification, and proposes a semi-supervised lithology identification workflow. In the workflow,the Laplacian support vector machine is employed to achieve semi-supervised learning. The feature similarityand depth similarity are introduced to reveal the data distribution characteristic, thus enabling us to elevate theclassification performance of the Laplacian support vector machine dealing with the issue of lacking labels.K-means clustering wasused to select the k well loggingsamples labeled by experts, based on theLaplacian support vector machine algorithm

tobeconducted. The proposed method is compared with the supervised method and excludesthe unlabeled data and the semi-supervised method without considering the depth similarity. Thecomparison experiments are conducted on four datasets collected from the Jiyang Depression, Bohai BayBasin, China. The experimental results show that the utilization of unlabeled data can improve identificationperformance, especially that of minority lithology classes. It is also verified that the information provided byfeature similarity and depth similarity are helpful for lithology identification.It is found that although the global accuracies of the three methods are high, the accuracies of SVM-based lithology identification method on minority classes decrease sharply compared with those in the case of the equally labeled dataset.

Chen et al (2021) proposed a new prediction model base on support vector machine (SVM) for pure/impure Carbon dioxide (CO_2) and crude oil system. This studywas based on 147 sets of Minimum Miscibility Pressure (MMP) data from the literature with full information on reservoir temperature,oil composition and gas composition. The main control factors were screened by several statisticalmethods. Unlike the conventional prediction models that are verified by only prediction accuracy, learning curve and single factor control variable. Analysis was further validated by the proposed model to obtain the optimum results.It was concluded that the Radial basis function (RBF) kernel prediction model cannot reflect the effect and the relative degree of influence of two features on MMP.

Susantoro et al (2023) aimed at examining the utilization of random forest classification for oil and gas exploration in the Central Sumatra Basin based on subsurface and surface data. Subsurface data included gravity data, basement structure maps, seismic interpretation maps (6 variables), and surface data, including Landsat 8 OLI data, Shuttle Radar Topography Mission Digital Elevation Model (SRTM DEM), surface geological maps, drainage pattern maps, and topographical maps (32 variables). The research was conducted on random forest with sample training in oil-and-gas-proven and potential areas (6 classes) and non-oil-and-gas-potential areas (5 classes). Based on the Ntree 600 parameter, Mtry k, and node size 2, unexplored oil and gas potential areas were identified. This consists of a very high potential area in the south of the Bengkalis graben; three locations of high potential areas located on the banks of the Kiri trough and close to the Aman and Balam grabens; two locations of medium potential category II located east and northwest of the Bengkalis graben; and a medium potential area category located east of the Bengkalis graben.The study was limited because no feature relationship was established.

Yu et al (2022) analyzed factors of productivity of tight conglomerate reservoirs based on random forest algorithm. The study proposed an evaluation model of major productivity controlling factors of thetight conglomerate reservoir to provide a reference for oil recovery based on a randomforest

(RF) machine-learning algorithm. The productivity factors were investigated fromtwo aspects: petrophysical facies that are capable of indicating the genetic mechanism ofgeological dessert and engineering dessert parameters forming complex fracture networks.. The results indicated that the RF model produced excellent results with only 12misclassifications across the entire data set of 627 samples that represent <2% error. Brittleness and maximum horizontal stress are considered the leastimportant indexes, with values of less than 5%. Reservoir quality and oil saturation were confirmed as the major controlling factorsand material foundation for oil wells' high and stable production.The limitation of the study was lack of experimental validation of the results presented and the applicability of the current RF model has not been fully understood.

Vo Thanh et al (2020) introduced the use of artificial neural networks (ANN) in estimating the ability of Residual Oil Zones (ROZs) to store CO_2 and recover oil. The training database was created using the uncertainty parameters, which also included the geological aspects and well operations. The cumulative oil production, cumulative CO_2 storage, and cumulative CO_2 retained were then constructed using a total of 351 numerical samples that were simulated. Findings showed that the created ANN model had an exceptional prediction performance with a high correlation coefficient (R^2) of over 0.98 and a total root mean square error of less than 2%. Additionally, four real ROZs in the Permian Basin were used to test the accuracy and stability of ANN models. The developed ANN models reduced the computational time for the optimization process in ROZs.

Mohammadi et al (2021) proposed a Genetic algorithm based support vector machine regression for prediction of SARA analysis in crude oil samples using Attenuated total reflectance - Fourier transform infrared (ATR-FTIR) spectroscopy.saturates, aromatics, resins, and asphaltenes. A hybrid of genetic algorithm (GA) and support vector machine regression (SVM-R) model was applied to predict SARA analysis of crude oil samples from different Iranian oil field using ATR–FTIR spectroscopy. The result of GA-SVM-R model were compared with genetic algorithm-partial least square regression (GA-PLS-R) model. Correlation coefficient (R^2) and root mean square error (RMSE) for calibration and prediction of samples were also calculated, in order to evaluate the calibration models for each component of SARA analysis in crude oil samples. The performance of GA-SVM-R is found to be reliableso that it can be successfully applied as an alternative approach for the quantitative determination of the SARA analysis of crude oil samples.

Jung et al (2018) undertook a Geological model sampling using PCA-assisted support vector machine for reliable channel reservoir characterization. The research performed PCA for figuring out main geological characteristics of reservoir models. A SVM classifier was trained using 10% models which show the most similar or dissimilar well oil production rates (WOPR) with the true values (5% for each category). Then, the other 90% models are classified by the trained SVM. The

researchers selected models on the side of lower WOPR errors. By repeating the classification process, they can select reliable models which have similar geological trend with the true reservoir model.However, history matching with the sampled ensemble offers reliable characterization results by figuring out proper channel trend and gives dependable prediction of future performances. Increase in computational efficiency using only selected geological models in history matching was obtained

Zang et al (2019) developed a hybrid scoring system for Enhanced Oil Recovery (EOR) screening by combining conventional screening guidelines and random forest algorithm. First, the screening guidelines were established by compiling 977 EOR projects from various publications in different languages from existing literature. Boxplots were used to detect the special cases for each reservoir/fluid property and to present the graphical screening results. To avoid the experts' bias, the weighting factors for each EOR technique were determined through the application of the random forest algorithm. The scoring system was then established by the fuzzification of reservoir/fluid property scores and the computation of composite screening scores. A case study was used to demonstrate that with a simple input of reservoir/fluid information, the novel scoring system could effectively provide recommendations for EOR selection by ranking scores.The case study result showsthat the established novel hybrid scoring system could provide discriminative EOR screening results for the selected field.

Hadavimoghaddam et al (2021) compared the predictions of Dead Oil Viscosity using Machine Learning and Classical Correlations. This paper implements six machine learning models: random forest (RF), lightgbm, XGBoost, multilayer perceptron (MLP) neural network, stochastic real-valued (SRV) and SuperLearner to predict dead oil viscosity. More than 2000 pressure– volume–temperature (PVT) data were used for developing and testing these models. The results show that the functional form f(γ_{API}, T), has the best performance, and additional correlating parameters might be unnecessary. Furthermore, SuperLearner outperformedother machine learning (ML) algorithms as well as common correlations that are based on the metric analysis. The SuperLearner model can potentially replace the empirical models for viscosity predictions on a wide range of viscosities (any oil type). Ultimately, the proposed model is capable of simulating the true physical trend of the dead oil viscosity with variations of oil API gravity, temperature and shear rate.

Zou et al (2021) predicted Porosity with Uncertainty Quantification from Multiple Seismic Attributes using Random Forest. The study proposes a Random Forest (RF) based method using multiple seismic attributes to predict the underground porosity distribution with uncertainty quantification. The standard deviation of base models' predictions is used to quantify the regression uncertainty of RF.The application of the proposed method on seismic data shows its potential to

characterize spatially varying reservoir parameters, and the quantified uncertainty profile offers insights into risk evaluation for hydrocarbon exploration and development. The study failed to compare results from the study with baseline systems.

Wang et al(2022) proposeda novel hybrid model (NRF)combining neural network (NN) and random forest (RF) was proposed based on welllogging data to predict the porosity and saturation of shale gas reservoirs. The databaseincludes six horizontal wells, and the input logs include borehole diameter, neutron,density, gamma-ray, and acoustic and deep investigate double lateral resistivity log.The porosity and saturation were chosen as outputs. The NRF model withindependent and joint training was designed to extract key features from well log dataand physical parameters.The NRF model has a similar data distribution with measured porosity and saturation, which demonstrates the NRFmodel can achieve greater stability. It was proven that the proposed NRF model cancapture the complex relationship between the logging data and physical parameters moreaccurately, and can serve as an economical and reliable alternative tool to give a reliableprediction. The developed model is however limited to shale gas reservoir.

Sandunilet al (2023) investigated the effects ofn_estimators, max_features and min_samples_leafto predict porosity of Volve oil field in the North Sea. Depth, gamma ray logs, neutron porosity logs and resistivity logs parameterswere used as inputs, while calculated porosity was used as target outputs to develop the Random Forest Regression (RFR) models. The RFR models were developed through: (i) tuning each hyperparameter individually, (ii) tuning hyperparameters by coupling them into three groups and, (iii) tuning all three hyperparameters at once. Results showed that tuning maxfeatures had a higher impact on improving the performance of the RFR model when predicting porosity of Volve oil field in the North Sea.

Gamal et al (2021) developed an Intelligent system for Prediction of Rock Porosity while Drilling Complex Lithology in Real Time. The paper aims at predicting the rock porosity in real time while drilling complex lithology using machine learning. Two intelligent models were developed utilizing random forest (RF) and decision tree (DT) techniques. The drilling parameters include weight on bit, torque, standpipe pressure, drill string rotation speed, rate of penetration, and pump rate. Forbuilding the models 3767 data points and 1676 data pointsfor validation were used to develop the models. Thecollected datasets have complex lithology of carbonate, sandstone, and shale. Sensitivity and optimization on different parameters for each technique were conducted to ensure optimum predictionthe training and testing results showed that, for the best RF model parameters, R2 of 0.99 and 0.90 with AAPE of 1.5 and 7% was observed for the training and testing datasets, respectively. VAF recorded 99.44% and 95.76%, while the a20 index was 1 and

0.93 for training and testing phases, respectively.The results indicated the strong porosity prediction capability for the two models.

Al-AbdulJabbaret al (2020) implemented an artificial neural network (ANN) technique to predict the porosity in the reservoir section from the drilling parameters. The data used to build the ANN model are based on real field data (2,800 data points) that were obtained from two horizontal wells (Well A and Well B). The data from Well A were used to train and test the ANN model with a training/ testing ratio of 7:3. More than 30 sensitivity analyses were performed to select the optimum ANN model's design parameters. Well B data were used to validate the developed ANN model.The obtained results showed that ANNs can be used effectively to predict the porosity from the drilling parameters in the reservoir section with an average correlation coefficient of approximately 0.96 and a root mean square error (RMSE) of almost 0.018. The best ANN parameter combination was with two layers, 30 neurons per layer with Levenberg-Marquardt training function and tan-sigmoid as the transfer function. The validation process confirmed that the ANN porosity model was able to predict the porosity of Well B with a correlation coefficient of 0.907 and an RMSE of 0.035.

Rosid et al (2019) carried out a study that aimed at classifying Carbonate reservoir rock type using comparison of Naïve Bayes and Random Forest method. The comparison is done so that the best method can be chosen from these two methods. Several assessments are done in the comparing process. Random Forest method has a better result in estimating the error value of the training model, comparing the core permeability with predicted permeability, and the level of accuracy in predicting rock type classification. However, the Naive Bayes method has a separately well distribution of seismic parameters compared to the Random Forest method, which is the most important thing. This shows that in wells that have no core data, estimation of the permeability value was done using the Naive Bayes method.

Ahmed et al (2019) Predicted Pore and Fracture Pressures using Support Vector Machine (SVM) .Inthe study, a real fielddata that contain the log data and real time surface drilling parameters were utilized by SVM to predict the pore and fracture pressures. SVM predicted the pore and fracture pressures with a high accuracy where the coefficient of determination (R^2) is greater than 0.995. In addition, it can estimate the pore pressure without the need for pressure trends and predict the fracture pressure from only the real time surface drilling parameters which are easily available.

Otchere et al (2021) did a comparative analysis of ANN and SVM models in prediction of petroleumreservoir properties.This review focuses on ANN with different shallow modelsused in reservoir characterization. The SVM and Relevant VectorMachine (RVM) have over the years

emerged as competitive algorithms where in most casesbased on this review it outperformed the ANN. This makes it preferable than the ANN whenthere are limited data sets. Finally, hybridization of multiple algorithms methodologies alsoshowed improved performance over singularly applied algorithms offering a pathway inimproving reservoir characterization based on supervised machine learning as future scope ofwork.

METHODOLOGY

In consultation with a Petroleum Engineer, 816 data sets were collected from Shell Petroleum Development Company (SPDC). The engineer assisted in data cleaning, normalizing and some outliers were discovered and resolved, getting rid of 152 data points, leaving 664 data points used in this study. The attributes of the data are Reservoir Permeability (RPE), Reservoir Porosity (RO), Water Cut (WC), Oil Rate, Gas (OGR), Oil Ratio (OR), Oil Column Thickness (CT), Liquid Rate (LR), Gas Production Rate, Water Production Rate and Oil in Place. The data sample and the transformed data are shown on Table 1 and 2 respectively.Data transformation involves converting raw data into a well-suitable format for the machine learning models. Reservoir Permeability column was converted to numeric values by taking the average of the range of Permeability, Reservoir Porosity column was converted to numeric values by taking the average of the porosity and the Oil Column thickness was converted to numeric values by subtracting the lower oil boundary from the upper oil boundary and this created a dataset containing numeric values. To transform data to suitable format, Min-Max Scaling (Normalization) method was adopted.This method scales the features to a specified range, usually [0, 1]. The formula for Min-Max Scaling is:

$$X\_normalized = (X - X\_min) / (X\_max - X\_min) \quad (1)$$

where X is the original feature and X={ x_1,x_2,…x_n}, X_min is the minimum value of the feature in the dataset, and X_max is the maximum value of the feature in the dataset.

The tools used for the study are Random Forest (RF) and Support Vector Regression (SVR). During the training phase, each Regressor is trained independently by its own replicated training data set via a bootstrap method. The data was divided into two sets of training and testing sets. The training set is 80% of the data, while the remaining 20% is for the Testing.The Mean Squared Error (MSE), Mean Absolute ErrorMAE), R-squared Score ($R^{2)}$, Explained Variance Score (EVS) andMedian Absolute Error (MedAE) metrics were used to measure the performance of the models.

European Journal of Computer Science and Information Technology 11 (5), 30-49, 2023
Print ISSN: 2054-0957 (Print),
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Table 1: Data Sample

Liquid Rate mmbbl/d	Oil Rate mmbbl/d (Qo)	Gas Oil Ratio Mcf/mmbbl	Water Cut %	Gas produced Mcf/d (Qg)	water production Rate mmbbl/d (Qw)	Average Reservoir permeability (K) MD	Average Reservoir porosity (Ø)	Oil Column thickness ft		Oil in Place mmbbl (Np)
1592.55	1592.55	1.38	0	2197.719	0	1000-5000 MD	20-30%	253.92 ft @ 8331.21 ft	0.03	
1254.68	1254.68	1.35	0	1693.818	0	1000-5000 MD	20-30%	253.92 ft @ 8331.21 ft	0.09	
1348.97	1348.97	1.47	0	1982.9859	0	1000-5000 MD	20-30%	253.92 ft @ 8331.21 ft	0.18	
1334.97	1334	1.44	0.07	1920.96	0.97	1000-5000 MD	20-30%	253.92 ft @ 8331.21 ft	0.26	
1392.53	1389.77	1.33	0.2	1848.3941	2.76	1000-5000 MD	20-30%	253.92 ft @ 8331.21 ft	0.35	
1381.1	1378.35	1.39	0.2	1915.9065	2.75	1000-5000 MD	20-30%	253.92 ft @ 8331.21 ft	0.42	
1349.29	1346.58	1.38	0.2	1858.2804	2.71	1000-5000 MD	20-30%	253.92 ft @ 8331.21 ft	0.5	
1356.89	1354.18	1.36	0.2	1841.6848	2.71	1000-5000 MD	20-30%	253.92 ft @ 8331.21 ft	0.57	
1351.71	1349	1.4	0.2	1888.6	2.71	1000-5000 MD	20-30%	253.92 ft @ 8331.21 ft	0.63	

Table 2: Normalized data Sample

Liquid_ Rate	Oil_ Rate	Oil_Column _thickness	Gas_Oil_ Ratio	Water_ Cut	Gas_ produced	water_ production_ Rate	Average_Reservoir _permeability	Average_ Reservoir_ porosity	Oil_in _Place
0.2895	0.2863	1.0000	0.0288	0.0000	0.0620	0.7017	1.0000	1.0000	0
0.7242	0.7230	1.0000	0.0248	0.0000	0.1355	0.7017	1.0000	1.0000	0.0058
0.9242	0.9238	1.0000	0.0303	0.0000	0.2101	0.7017	1.0000	1.0000	0.0144
0.9107	0.9103	1.0000	0.0240	0.0000	0.1651	0.7017	1.0000	1.0000	0.0221
1.0000	1.0000	1.0000	0.0248	0.0000	0.1875	0.7017	1.0000	1.0000	0.0307
0.7356	0.7344	1.0000	0.0278	0.0000	0.1535	0.7017	1.0000	1.0000	0.0375
0.8945	0.8941	1.0000	0.0278	0.0000	0.1868	0.7017	1.0000	1.0000	0.0451
0.8037	0.8028	1.0000	0.0396	0.0000	0.2368	0.7017	1.0000	1.0000	0.0519

Architectural Design of the System

The Architectural design of the system is depicted in Figure 1

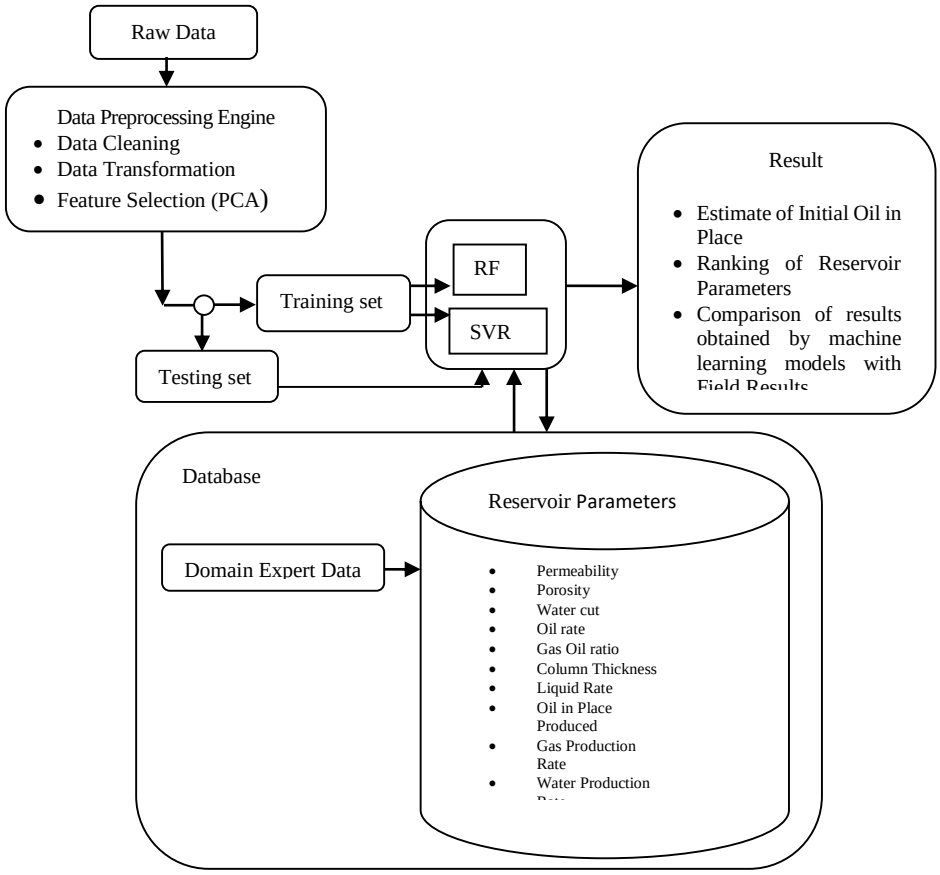

Figure 1: Architectural Design of the Prediction of Initial-Oil-in-Place in the Niger Delta Region.

The system architecture consist of some components which include; Raw data, Data Preprocessing Engine, Database, Machine Learning Models and Results. Raw data are data obtained from SPDC for the purpose of this research. Data Preprocessing Engine cleans and transforms data. The machine learning models (RF and SVR) are used to predict the Initial Oil in Place in relation to the database that stores the reservoir parameters. Result component of the system architecture shows the results that will be obtained after a successful implementation of the System

RESULTS AND DISCUSSION

A principal component Analysis (PCA) was conducted on the features and four out of the nine input features were selected based on their Eigen values and Explained Variance Percentage. The

Eigen values, Explained Variance Percentages and Cumulative Explained Variance Percentage of the features are shown on Table 1. The selected features are Liquid Rate, Oil Rate, Oil Column Thickness and Gas Oil Ratio.The decision of using four input features was arrived at using domain expert knowledge and literature. According to Araújo and Santos (2018), features with eigen values of 0.5 and above are stable; hence the decision of using four features.

The hyperparameter tuning results for RF models using 9 and 4 input features are shown on Table 2 and 3 respectively, while that of SVR models using 9 and 4 input features are shown on Table 4 and 5 respectively.

Table 1: Eigen Values and corresponding Percentage Explained Variance for OIP features

Features	Eigen Values (EV)	Explained Variance Percentage(EVP)	Cumulative Explained Variance Percentage (CEVP)
Liquid Rate	2.1229	0.3533	0.3533
Oil_Rate	1.8288	0.3043	0.6576
Oil Column Thickness	1.3552	0.2255	0.8831
Gas_Oil_Ratio	0.5561	0.0925	0.9756
Reservior Porosity	0.1461	0.0244	1.0000
Reservoir Permaebility	0.0000	0.0000	1.0000
Water cut	0.0000	0.0000	1.0000
Gas Production Rate	0.0000	0.0000	1.0000
Water Poduced	0.0000	0.0000	1.0000
Total	6.0091	1.0000	1.0000

Table 2: Hyperparameter Tuning Results for RF using 9 input features

Parameters	n_estimators	MSE	MAE	R^2	EVS	MedAE
{'max_features': 'auto'}	50	0.009	0.049	0.885	0.885	0.019
{'max_features' : 'auto}	150	0.009	0.050	0.885	0.885	0.020
{'max_features': 'auto'}	100	0.010	0.049	0.819	0.819	0.019
{'max_features': 'sqrt'}	100	0.0010	0.060	0.802	0.802	0.025

Table 3: Hyperparameter Tuning Results for RF using 4 input features

Parameters	n_estimators	MSE	MAE	R^2	EVS	MedAE
{'max_features': 'auto'}	50	0.008	0.048	0.886	0.886	0.018
{'max_features': ' auto')	150	0.0010	0.050	0.880	0.880	0.019
{'max_features': 'auto'}	200	0.0010	0.060	0.815	0.815	0.018
{'max_features': ' sqrt'}	100	0.010	0.054	0.881	0.881	0.026

Table 4: Hyperparameter Tuning Results for SVR using 9 input features

Parameters	MSE	MAE	R^2	EVS	MedAE
{'C':0.01, 'epsilon': 0.2}	0.037	0.158	0.446	0.455	0.145
{'C':0.01, 'epsilon': 0.01}	0.041	0.155	0.385	0.385	0.116
{'C': 0.01, 'epsilon': 0.1}	0.044	0.166	0.359	0.345	0.133
{'C': 0.01, 'epsilon': 0.2}	0.049	0.187	0.271	0.350	0.181

Table 5: Hyperparameter Tuning Results for SVR using 4 input features

Parameters	MSE	MAE	R^2	EVS	MedAE
{'C':1, 'epsilon': 0.01}	0.036	0.156	0.448	0.456	0.145
{'C': 0.01, 'epsilon': 0.01}	0.044	0.164	0.383	0.416	0.117
{'C': 0.01, 'epsilon': 0.1}	0.044	0.169	0.387	0.394	0.143
{'C': 0.01, 'epsilon': 0.2}	0.047	0.184	0.348	0.363	0.194

The total IOIP estimated by experts in the Field, RF and SVR using 4 features are shown on Table 6. The percentage errors of Models using 4 features for prediction are shown in Table 7. The Line Graph showing the comparison of IOIP predictions using 9 and 4 features is shown in Figure 2.

Table 6: IOIP predictions of experts in the Field, RF and SVRModels using 4 features

Field Results (mmbbl)	RF Results (mmbbl)	SVR Results (mmbbl)
328.11	330.73	339.92

Table 7: IOIP Prediction Percentage Errors (PE) for the TwoModels using 4 input Features

	RF	SVR
PE	0.798	3.599

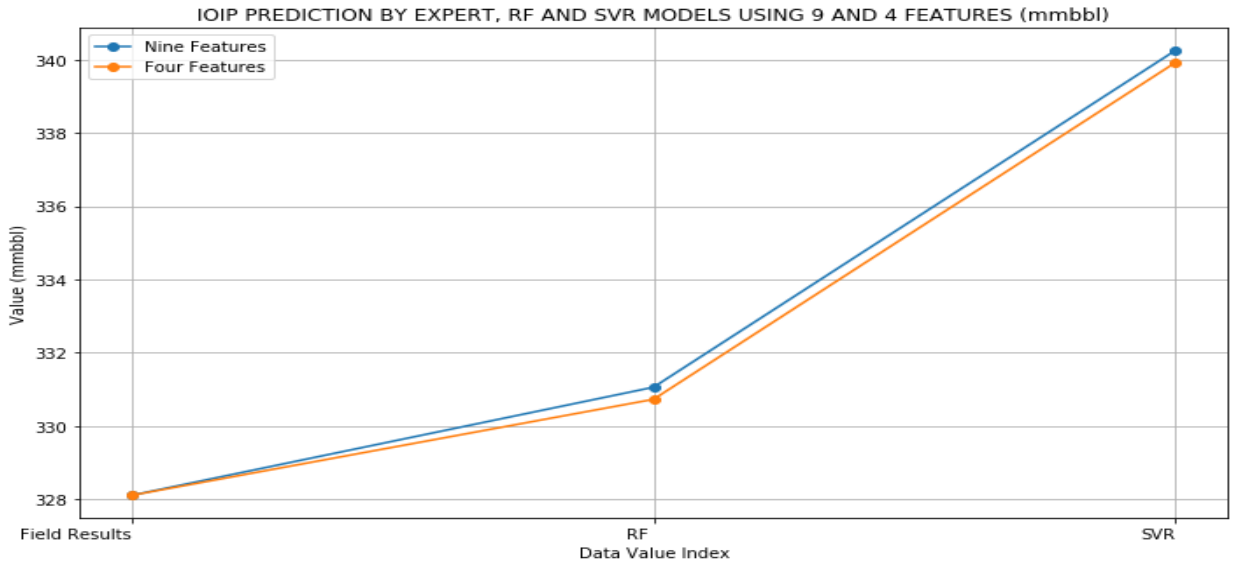

Figure 2: Line Graph Comparison of the Prediction of RF and SVR Models.

The Scatter Plot of the field results against the computedresults for RF and SVR are depicted in Figure 3 and 4 respectively, while the group bar chart for the comparison of the RF and SVR models is shown in Figure 5.

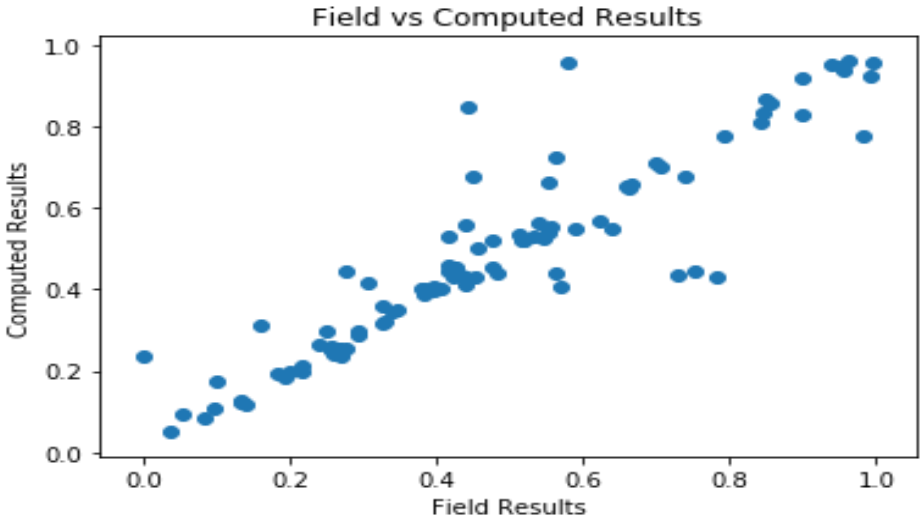

Figure 3: Scatter Plot of Field Results against the Computed Results in RF.

In Figure 4, the relationship between variables is high, positive and linear. There are seven (7) outliers with the farthest point apart being 0.5 units. The points form a tight cluster around the diagonal line (indicating a strong positive correlation between field and computed results).

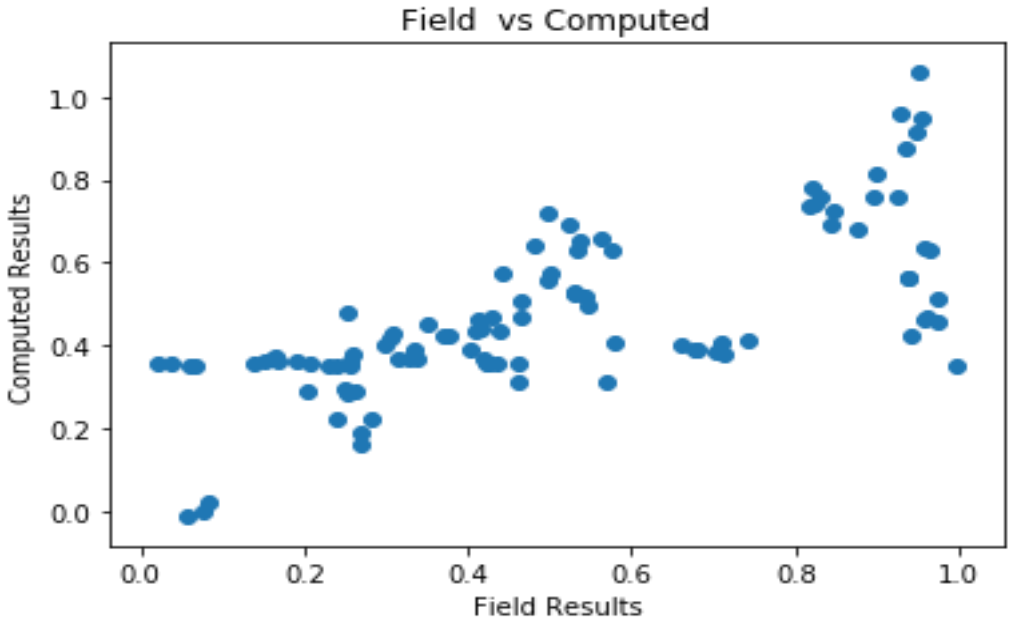

Figure 4: Scatter Plot of Field Results against Computed Results in SVR

The SVR model shows a null/ no relationship. The points do not form a tight cluster around the diagonal line (indicating that there is no strong positive correlation between field and computed results). There are several outliers with the furthest point apart being 0.7 units.

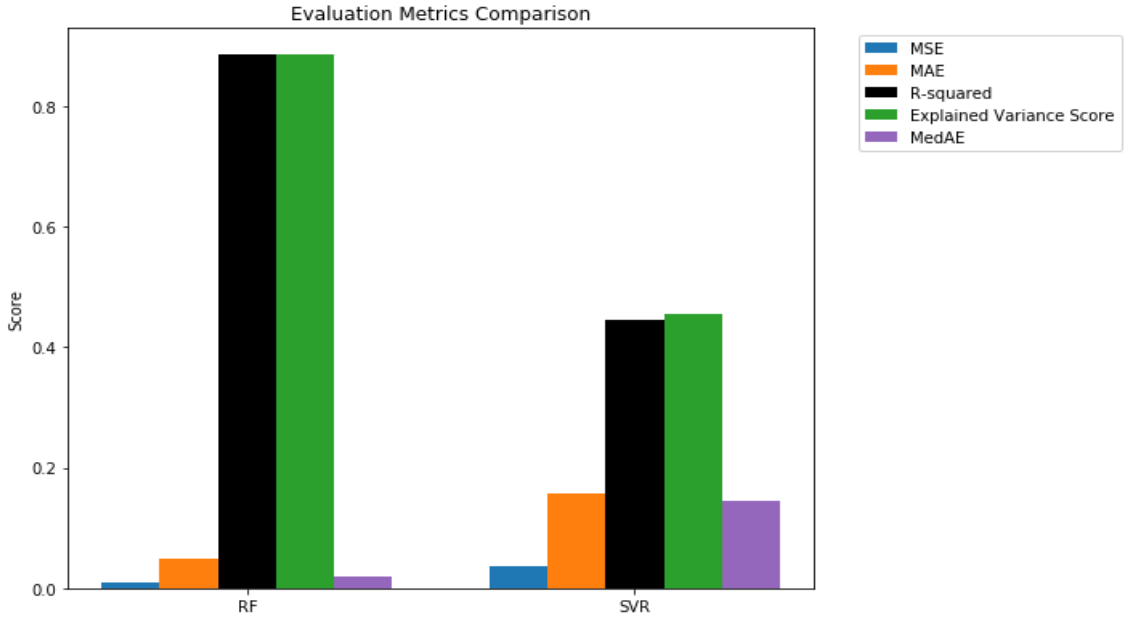

Figure 5: Evaluation Metrics Comparison of the RF and SVR models.

DISCUSSION

The error metrics used in the evaluation of the models include Mean Squared Error (MSE), Mean Absolute Error (MAE), R-Squared Score (R^2), Explained Variance Score (EVS) and Median Absolute Error (MedAE). Visualization tools used are scatter plot, line graph and grouped bar chart. PCA was conducted on the datasets to select4 features(Liquid Rate, Oil Rate, Oil Column Thickness and Gas Oil Ratio)based on their eigen values. Random Forest and Support Vector Regression were used to predict IOIP in the study and results obtained were compared with that obtained from the field as shown in Table 6 and Figure 2.

CONCLUSION AND RECOMMENDATIONS

The analysis of the two regression models for predicting the relationship between variables has provided valuable insights into their performance and suitability for the given dataset. The models were evaluated based on multiple assessment criteria, including Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared (R^2), Explained Variance Score (EVS), and Median Absolute Error (MedAE).The models that used 4 input features performed better than their counterparts that used 9 input variables.

Random Forest (RF) with 4 input variables exhibited a strong positive linear relationship betwee n the variables, as evidenced by its low MSE of 0.008, MAE of 0.048, high R^2 of 0.886, EVS of 0.886, MedAE of 0.18 and predicted IOIP of 330.73 mmbbl, while RF using 9 input variables ha d a MSE of 0.009, MAE of 0.049, R^2 of 0.885, EVS of 0.885, MedAE of 0.019 and predicted IOI P of 331.06 mmbbl. The scatter plot revealed a tight cluster of points around the diagonal line, in dicating a robust correlation.

Support Vector Regression (SVR) using 4 inputs features, in contrast, demonstrated weaker predictive performance, with higher error values compared to the RF model. The SVR model using 4 input variables had a MSE of 0.036, MAE of 0.156, $R^2$of 0.448, and EVS of 0.456 and MedAE of 0.145.. The scatter plot did not show a compact convergence of points around the diagonal line, indicating a lack of significant positive correlation. Additionally, the presence of several outliers, including one extreme outlier at 0.7 units apart, suggests limited robustness in handling the data. The uniqueness of this study is in the use of 4 out of the 9 predicting features (independent variables) to obtain a prediction that is closer to that obtained with 9 features in the field. With this finding, companies can save cost of exploring the other 5 features and concentrate on the 4 features.

It is recommended that drilling datasets from other companies be collected and combined with that used in this study on the same models to investigate an improvement on the results obtained from this study. An integration of knowledge-based reasoning used by the experts with the data-driven work done in this study is also recommended. An ensemble of regression tools could also be used to predict IOIP and results compared with the ones obtained in this study.

References

Ahmed S, A., Mahmoud, A. A., Elkatatny, S., Mahmoud, M., & Abdulraheem, A. (2019). Prediction of pore and fracture pressures using support vector machine. In international technology conference (p. D021S018R002). IPTC.

Al-AbdulJabbar, A., Al-Azani, K., &Elkatatny, S. (2020). Estimation of reservoir porosity from drilling parameters using artificial neural networks. Petrophysics, 61(03), 318-330.

Araújo, J. M., & Santos, T. L. M. (2018). Control of a class of second-order linear vibrating Systems with time-delay: Smith predictor approach. Mechanical Systems and Signal Processing, 108, 173-187.

Chen, H., Zhang, C., Jia, N., Duncan, I., Yang, S., & Yang, Y. (2021). A machine learning model for predicting the minimum miscibility pressure of CO2 and crude oil system based on a support vector machine algorithm approach. Fuel, 290, 120048.

Gamal, H., Elkatatny, S., Alsaihati, A., & Abdulraheem, A. (2021). Intelligent prediction

for rock porosity while drilling complex lithology in real time. Computational intelligence and neuroscience, 2021, 1-12.

Hadavimoghaddam, F., Ostadhassan, M., Heidaryan, E., Sadri, M. A., Chapanova, I., Popov, E., ... &Rafieepour, S. (2021). Prediction of dead oil viscosity: Machine learning vs. classical correlations. Energies, 14(4), 930.

Li, T., Wang, R., Wang, Z., Zhao, M., & Li, L. (2018). Prediction of fracture density using genetic algorithm support vector machine based on acoustic logging data. Geophysics, 83(2), D49-D60.

Li, Z., Kang, Y., Feng, D., Wang, X. M., Lv, W., Chang, J., & Zheng, W. X. (2020). Semi-supervised learning for lithology identification using Laplacian support vector machine. Journal of Petroleum Science and Engineering, 195, 107510.

Jung, H., Jo, H., Kim, S., Lee, K., & Choe, J. (2018). Geological model sampling using PCA-assisted support vector machine for reliable channel reservoir characterization. Journal of Petroleum Science and Engineering, 167, 396-405.

Mohammadi, M., Khorrami, M. K., Vatani, A., Ghasemzadeh, H., Vatanparast, H., Bahramian, A., & Fallah, A. (2021). Genetic algorithm based support vector machine regression for prediction of SARA analysis in crude oil samples using ATR-FTIR spectroscopy. Spectrochimica Acta part A: Molecular and biomolecular spectroscopy, 245, 118945.

Obot, O., Attai, K and Onwodi, G. (2022). Integrating Knowledge-Driven and Data-Driven Methodologies for Efficient Clinical Decision Support System In Diverse Perspectives and State-of-the-art Approaches to the Utilization of data-driven Clinical Decision Support Systems (pp.1-28). IGI-Global

Otchere, D. A., Ganat, T. O. A., Gholami, R., & Ridha, S. (2021). Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models. Journal of Petroleum Science and Engineering, 200, 108182.

Rosid, M. S., Haikel, S., & Haidar, M. W. (2019, November). Carbonate reservoir rock type classification using comparison of Naïve Bayes and Random Forest method in field “S” East Java. In AIP Conference Proceedings (Vol. 2168, No. 1). AIP Publishing.

Sandunil, K., Bennour, Z., Ben Mahmud, H., &Giwelli, A. (2023, June). Effects of Tuning Hyperparameters in Random Forest Regression on Reservoir's Porosity Prediction. Case Study: Volve Oil Field, North Sea. In 57th US Rock Mechanics/Geomechanics Symposium. OnePetro.

Susantoro, T. M., Wikantika, K., Suliantara, S., Setiawan, H. L., Harto, A. B., & Sakti, A. D. (2023). Applying random forest to oil and gas exploration in Central Sumatra basin Indonesia based on surface and subsurface data. Remote Sensing Applications: Society and Environment, 32, 101039.

Vo Thanh, H., Sugai, Y., and Sasaki, K. (2020). Application of artificial neural network for predicting the performance of CO2 enhanced oil recovery and storage in residual oil zones. Scientific reports, 10(1), 1-16.

Wang, M., Feng, D., Li, D., & Wang, J. (2022). Reservoir Parameter Prediction Based on the Neural Random Forest Model. Frontiers in Earth Science, 10, 888933.

Yu, Z., Wang, Z., Jiang, Q., Wang, J., Zheng, J., & Zhang, T. (2022). Analysis of factors of productivity of tight conglomerate reservoirs based on random forest algorithm. ACS omega, 7(23), 20390-20404.

Zhang, N., Wei, M., Fan, J., Aldhaheri, M., Zhang, Y., & Bai, B. (2019). Development of a hybrid scoring system for EOR screening by combining conventional screening guidelines and random forest algorithm. Fuel, 256, 115915.

Zou, C., Zhao, L., Xu, M., Chen, Y., & Geng, J. (2021). Porosity prediction with uncertainty quantification from multiple seismic attributes using random forest. Journal of Geophysical Research: Solid Earth, 126(7), e2021JB021826.