

Comparative Evaluation of Passey Δ LogR, Statistical Regression, and Machine-Learning Models for Total Organic Carbon Estimation in parts of Gongola Basin, Nigeria

Edeh Ifeanyichukwu Samuel¹, Soronnadi-Ononiwu Geoffrey Chijioke², and Chiazor Fortune Ijeoma².

¹Center for Petroleum Geosciences, University of Port Harcourt.

²Department of Geology, Faculty of Science, University of Port Harcourt

Corresponding Author's Email:

DOI: <https://doi.org/10.51584/IJRIAS.2026.11090077>

Received: 23 September 2026; Accepted: 28 September 2026; Published: 14 October 2026

ABSTRACT

Geochemical methods for TOC determination required for source rock evaluation during exploration is accurate, but costly, time-consuming, and labor intensive, leading to alternative approaches. For the present study, a total of 829 laboratory measured TOC are available from Well A and Well B. This research evaluated and compared the efficacy of petrophysical, statistical, and advanced machine learning models for predicting Total Organic Carbon (TOC) in well A and well B in Gongola Basin using R^2 and RMSE. The data was divided into 70% for training and cross-validation and tested on the remaining 30%. The results show that Tree-based machine learning models outperformed both Passey overlay and multilinear regression. The accuracy of Passey overlay, multilinear regression and regularized linear regression models of Lasso (L1) Regression, Ridge (L2) Regression, and LASSO LARS were low possibly because their prediction were based of linear relationship between features and TOC. The study indicates that the linear relationship between TOC and the predictor variables is generally weak to very weak in Well A, and ranges from very weak to moderate in Well B. Among the five high-performing tree-based models, Gradient Boosted Tree and Decision Tree demonstrated the highest predictive accuracy score of 0.98 in feature group 3 in Well B. This was followed in descending order of performance by, Extra Tree, XGBoost and Random Forest. These models overcame poor linear relationship limitations by dynamically mapping non-linear relationship between TOC and log signatures / rock physics parameters. Global feature importance was evaluated by aggregating the magnitude of the SHAP values. Neutron porosity (NPHI), by resistivity (RES), and Lambda-Rho are the top three features considered by most of the models which geologically implies high concentration of hydrogen rich kerogen, thermal maturity of the source rock and variability in TOC being related to lithofacies changes. All the models predicted TOC better in well B than well A probably because well B exhibited better log-to-TOC relationship and has more homogenous lithology.

Keywords: Machine learning, Total Organic Carbon, Gongola Basin and Passey Δ LogR

INTRODUCTION

The global energy landscape continues to depend significantly on hydrocarbon resources despite the rapid growth of renewable and other alternative energy sources. Oil and natural gas remain important components of the global energy mix and any disrupt in the production, or availability of hydrocarbons can have significant implications for energy security, commodity prices and global economic activity. To sustain continuous oil and gas production, there is a need to discover new hydrocarbon reserves while minimizing exploration costs. To achieve this, data driven approaches is required to extract greater value from existing geological and geophysical data. Characterising petroleum system in frontier basin is a herculean task. This is because of high uncertainty and poor geological understanding around the petroleum system which lowers the chances of success. The lack of geological understanding is rooted in scarcity of data and limited

scientific research on source rock evaluation. TOC is a fundamental parameter in evaluation of source rock potential of a petroleum system. Total Organic Carbon (TOC) is a key geochemical parameter used to assess source rock quality, hydrocarbon generation potential, and petroleum system characteristics. Although laboratory-based methods such as Rock-Eval pyrolysis and LECO analysis provide accurate TOC measurements, they are costly, time-consuming, and limited by sample availability and depth coverage. As a result, continuous TOC evaluation in frontier basins remains a challenge. In Gongola Basin, most studies rely on laboratory geochemical analyses. Very few studies apply machine learning for TOC prediction on exploratory and limited scope. No comprehensive comparison exists between Passey ΔLogR , Multilinear Regression, and multiple tree-based ensemble algorithms for TOC prediction in the Gongola Basin. The contribution of geological, petrophysical, and rock-physics features to prediction performance has not been systematically evaluated. The objective of this study is to evaluate the performance of Passey, multilinear and machine learning model in Gongola and determine the best model for the basin. Artificial intelligence and machine-learning techniques provide potential mechanisms for identifying complex relationships within these datasets that may not be readily captured using conventional empirical or deterministic approaches.

Geological Setting and Stratigraphy

Benue Trough, and particularly the Upper Benue Trough, have been presented by several researchers like Benkhelil, 1982; Petters, 1982; Dike, 1993, 2002; Obaje, 1994; and Zaborski et al., 1997. The Upper Benue Trough is divided into the Gongola and Yola arms, with the Gongola Basin constituting a major depocentre within its northern segment. The Gongola Basin, located in the northern part of the Upper Benue Trough in northeastern Nigeria as shown in figure 1, forms part of the extensive Cretaceous intracontinental rift system that developed across West and Central Africa. The Benue Trough extends approximately 800 km in a north-northeast to south-southwest direction and contains up to 6,000m of Cretaceous to Tertiary sedimentary deposits. Its origin is closely associated with Early Cretaceous extensional tectonic activity linked to the opening of the South Atlantic Ocean. This extensional phase was later interrupted by the Santonian compressional event, which caused intense folding, faulting, uplift, magmatic activity, and significant reconfiguration of the basin architecture (Benkhelil, 1989; Zaborski et al., 1997). The geological evolution of the Gongola Basin reflects a complex history of rifting, subsidence, marine incursions, compression, and regression. These tectonic and sedimentary processes-controlled basin development, accommodation space, sediment influx, and depositional settings, resulting in the accumulation of continental, transitional, and marine sediments. Stratigraphically, the basin succession comprises, in ascending order, the Precambrian Basement Complex, Bima Sandstone, Yolde Formation, Kanawa Formation, Gulani Formation, Fika Formation, Gombe Formation, and Kerri-Kerri Formation as shown in figure 2 (Zaborski et al., 1997). The oldest sedimentary unit, the Bima Sandstone, unconformably overlies the Precambrian Basement Complex and represents the initial infill of the basin. It is composed predominantly of sandstones deposited in fluvial and alluvial environments during the early stages of rift development. Above the Bima Sandstone lies the Yolde Formation, which records the transition from continental to marginal-marine settings during the Cenomanian marine transgression. This formation consists of interbedded sandstone, siltstone, and shale deposited under variable nearshore and shallow-marine conditions

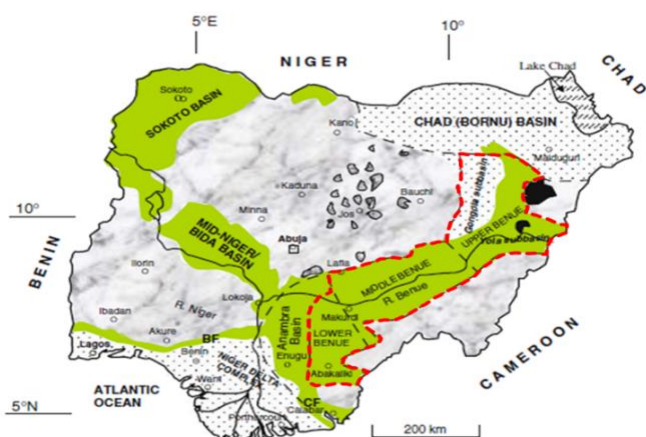


Figure 1: Generalized Geologic Map of Nigeria showing Gongola Sub-Basin, Benue Trough and other

Inland the sedimentary Basins. (Modified after Obaje, N.G. 2009).

The Pindiga Formation constitutes the principal marine succession within the basin and is characterized by thick intervals of shale, limestone, and calcareous mudstone deposited during peak marine flooding. Among its subdivisions, the Kanawa and Fika members are particularly significant because of their organic-rich shale intervals, which have attracted considerable interest for petroleum-system evaluation and source-rock studies. In several intervals, restricted circulation and oxygen-deficient depositional conditions enhanced the preservation of organic matter, thereby contributing to the development of potential hydrocarbon source rocks (Obaje, 2004; Abubakar et al., 2011; Hamis et al., 2024). Nwojiji et al. (2013) reported interbedding of shale, siltstone and sandstone in Pindiga formation.

The overlying Gombe Formation represents the regressive phase of basin evolution and marks the gradual transition from marine to deltaic, estuarine, and continental depositional environments. It is composed mainly of sandstone, siltstone, shale, and localized coal-bearing horizons, reflecting increased terrigenous sediment supply during the retreat of the Cretaceous sea. The sedimentary succession is capped by the Paleocene Kerri-Kerri Formation, consisting predominantly of fluvial and floodplain deposits that signify post-marine continental sedimentation following regional regression.

The diverse lithological units and depositional environments preserved within the Gongola Basin provide an excellent framework for Total Organic Carbon (TOC) prediction from well-log data. Variations in lithology, organic-matter abundance, depositional setting, burial history, and thermal maturity generate complex relationships between TOC and conventional well-log measurements such as gamma ray, resistivity, density, neutron porosity, and sonic transit time. These complexities often challenge traditional empirical TOC estimation methods and create opportunities for applying more advanced statistical and machine-learning techniques. Consequently, the Gongola Basin serves as an ideal natural laboratory for the development, validation, and comparison of empirical, statistical, and machine-learning models aimed at improving quantitative TOC estimation and source-rock characterization.

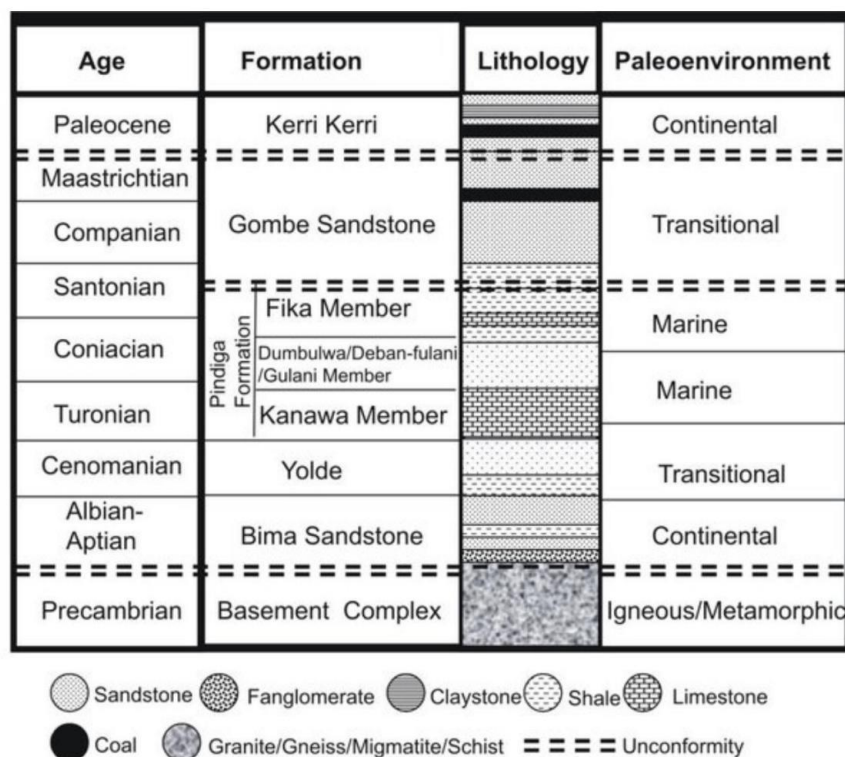


Figure 2: Stratigraphy of Gongola Basin after Zaborsky et al., 1997.

Previous investigations of the Gongola Basin have largely relied on conventional geochemical techniques, including TOC measurement, Rock-Eval pyrolysis, Hydrogen Index (HI), Tmax, vitrinite reflectance,

biomarker analysis, and stable isotope studies for source-rock evaluation. However, the application of Passey's ΔLogR method and machine learning-based TOC prediction remains limited.

Early studies by Akande et al. (1998) and Obaje et al. (1999, 2004) identified the Yolde, Pindiga, Fika, and Gombe formations as potential source rocks with highly variable organic richness. TOC values ranged from as low as 0.07 wt.% in the Bima and Fika formations to as high as 20.20 wt.% in coal-rich intervals of the Gombe Formation, while the Yolde Formation recorded the highest shale TOC values of up to 12.5 wt.%, suggesting significant source-rock potential.

A detailed geochemical assessment by Adedosu et al. (2014) using samples from the Kolmani River-1 well showed that the Gombe, Pindiga, Yolde, and Bima formations contain TOC values generally above the minimum threshold required for hydrocarbon generation. The study identified predominantly Type III/IV kerogen, indicating a gas-prone petroleum system, with maturity levels ranging from immature to early oil-window conditions, particularly within the Yolde and parts of the Pindiga formations. The authors further suggested that deeper intervals of the Bima Formation could represent effective mature source rocks.

The discovery of approximately 33 BCF of gas in the Kolmani River-1 well and the work of Abubakar et al. (2008) confirmed the existence of an active petroleum system within the basin and demonstrated that several Cretaceous formations contain hydrocarbon-generating organic matter. More recently, Didi et al. (2024) identified the Yolde, Pindiga, Fika, and Gombe formations as the principal source rocks of the basin and concluded that the Yolde Formation remains the most prospective source-rock unit due to its relatively higher TOC content and favorable maturity characteristics.

In contrast to the extensive conventional geochemical studies, predictive TOC estimation studies remain scarce. One of the few documented attempts is the work of Longman et al. (2024), who applied machine learning techniques to well-log data from Kolmani River-2 and Kolmani River-3 wells. The study integrated several machine learning algorithms and incorporated Passey's ΔLogR parameter, which demonstrated strong correlation with measured TOC and emerged as one of the most important predictive variables. Using Random Forest regression, the authors achieved R^2 values of 0.62 and 0.76, demonstrating the potential of combining ΔLogR and machine learning for reliable TOC prediction from well logs.

MATERIALS AND METHODS

The study integrated conventional petrophysical interpretation, statistical analysis, and advanced machine learning workflows using well logs, measured TOC, and mud logs information to develop and validate TOC prediction models for the Well A and Well B. Sampling intervals in Well A were generally between 10 and 20 ft, whereas Well B was sampled at wider intervals ranging from 30 to 90 ft. A total of 829 shale samples were used with 70% for training and 30% for testing the model. The TOC measurement served as reference data to calibrate well log derived TOC. Wireline logs from Well A and Well B from Gongola Basin were utilized in the study. The log consists of five open hole logs which are Compressional sonic, Gamma ray, Neutron Porosity, Resistivity and Density. Shear sonic was derived from compressional sonic using empirical relationship. Key rock physics parameters were derived using equations below.

$$\text{Acoustic impedance (Zp)} = V_p \cdot \rho \dots\dots\dots \text{Equation 1}$$

$$\text{Shear Wave Impedance (Zs)} = V_s \cdot \rho \dots\dots\dots \text{Equation 2}$$

$$\text{Velocity Ration (VpVs)} = V_p / V_s \dots\dots\dots \text{Equation 3}$$

$$\text{Lambda-Rho } (\lambda\rho) = (\rho V_p)^2 - 2(\rho V_s)^2 \dots\dots\dots \text{Equation 2}$$

$$\text{Mu-Rho } (\mu\rho) = (\rho V_s)^2 \dots\dots\dots \text{Equation 3}$$

This study employed three complementary TOC prediction approaches:

The Passey Overlay Technique (Passey et al., 1990) is one of the most widely used conventional petrophysical methods for estimating Total Organic Carbon (TOC) from well logs. The method involves overlaying resistivity and sonic logs after quality control and establishing an organic-lean shale baseline where the two curves coincide. The separation between the resistivity and sonic logs (ΔLogR) is then calculated and used, together with the Level of Organic Maturity (LOM), to estimate TOC. Positive ΔLogR values indicate organic-rich intervals, and the resulting TOC estimates are calibrated against laboratory-measured TOC values. The method provides continuous TOC prediction along the wellbore and has been widely adopted due to its simplicity and reliance on commonly available well logs.

Despite its widespread use, the Passey method has several limitations. First, the establishment of the baseline relies on subjective and iterative log scaling, which can introduce interpreter bias. Second, the method assumes a fixed baseline and a constant background TOC value, assumptions that may not hold in heterogeneous shale formations with varying lithologies. Wang, et al (2026) noted that assumption of a linear relationship between measured TOC and logarithmic resistivity and porosity logs is a major limitation. Furthermore, the approach requires thermal maturity information, typically expressed as Level of Organic Maturity (LOM), which may not always be available. In addition, the original model primarily relies on resistivity and sonic responses and may not adequately capture the complex geological and geochemical variability present in many unconventional reservoirs. These limitations can lead to inaccuracies in TOC estimation, particularly in formations characterized by significant lithological heterogeneity and varying maturity conditions.

To address these shortcomings, Wang et al. (2016) proposed revised ΔLogR models that extended the original Passey approach by incorporating Gamma Ray (GR) and thermal maturity (Tmax) in addition to ΔLogR . Using the Devonian Duvernay Shale as a case study, they demonstrated that integrating additional geological and geochemical parameters improves the characterization of organic richness and maturity variations, thereby enhancing TOC prediction accuracy in heterogeneous shale reservoirs. Further improvement was proposed by Zhao et al. (2017), who addressed one of the major weaknesses of the Passey method, namely its dependence on a fixed baseline. Zhao et al. introduced a mathematically derived variable baseline that accounts for lithological variability within shale formations. Unlike the original Passey model and the revised model of Wang et al. (2016), Zhao's approach does not require maturity-related parameters such as LOM, vitrinite reflectance (Ro), or Tmax. As a result, it offers a more flexible and robust framework for TOC estimation in formations where maturity data are unavailable or where lithological variations significantly influence log responses. In this study, the TOC were estimated for well A and well B from the conventional Passey method but the improvements by Wang et al. (2016) and Zhao et al. (2017) were not used. The performance was evaluated against laboratory-measured TOC data and compared with multilinear regression and machine learning approaches to determine the most accurate method for TOC prediction.

Advances in rock physics have enabled the derivation of additional attributes from conventional well logs. Parameters such as compressional velocity (Vp), shear velocity (Vs), bulk density, acoustic impedance (AI), shear impedance (SI), Vp/Vs ratio, Poisson's ratio, lambda-rho ($\lambda\rho$), and mu-rho ($\mu\rho$) have been widely investigated for TOC prediction. Organic matter generally reduces rock stiffness and density, resulting in lower acoustic velocities and impedances. Consequently, several studies have observed inverse relationships between TOC and attributes such as density, acoustic impedance, and elastic moduli. Sondergeld et al. (2010) demonstrated that organic-rich shale formations exhibit distinctive elastic properties compared to organic-lean intervals due to the presence of kerogen and variations in mineral composition. Their work highlighted the importance of integrating petrophysical and geomechanical properties for shale characterization. Similarly, Vernik and Nur (1992) showed that increasing organic content significantly affects seismic velocities and elastic moduli, providing a theoretical foundation for using rock physics parameters as indicators of TOC. Both Wang et al. (2016) and Zhao et al. (2017) incorporated rock physics-based variables alongside conventional log measurements in their improved TOC estimation models.

Multiple Linear Regression to establish statistical relationships between TOC and well-log/rock-physics parameters. Multiple Linear Regression was performed. TOC was used as the dependent variable, while well logs and rock-physics parameters including Gamma Ray, Compressional Slowness, Resistivity, Neutron

Porosity, Density, Shear Slowness, Vp/Vs, P-Impedance, S-Impedance, Lambda-Rho, and Mu-Rho served as predictor variables. The resulting regression equation was then used to predict TOC in new datasets.

Machine Learning Models to capture complex nonlinear relationships and improve prediction accuracy through advanced algorithms. It began with rigorous data Preparation and Cleaning that include removing irrelevant columns such as well names and dataset names. Followed by converting incorrectly formatted numeric fields to numerical data types. Removing unit rows and eliminating records containing null or invalid values (e.g., -9999). To improve model performance and computational efficiency, all variables were standardized using Z-score normalization, transforming each feature to a mean of zero and a standard deviation of one. TOC was defined as the target variable, while eleven well-log and rock-physics parameters were used as input features. To reduce multicollinearity and prevent misleading statistical relationships, highly correlated variables (correlation coefficient > 0.8 – 0.9) were grouped or excluded. The dataset was split into 70% for Training and 30% for Testing. K-Fold Cross-validation was applied. Several machine learning algorithms were evaluated, including: Random Forest, Extra Trees, Decision Tree, Gradient Boosted Tree, XGBoost Regression, Support Vector Machine (SVM), Ridge Regression (L2), Lasso Regression (L1), and LASSO-LARS. The global feature importance was evaluated by aggregating the magnitude of the SHAP values. The results from all approaches were validated against laboratory-measured TOC data and compared using performance metrics such as R^2 , and RMSE to identify the most reliable TOC prediction method for Well A and Well B.

RESULTS AND DISCUSSION

Geochemical Evaluation

Obaje et al. (2004) had observed that formations from the upper Benue trough generally have low TOC and HI contents. According the research work, samples from Bima, Yolda and Pindiga formations yielded poor source rock quality. Also, poor TOC and HI values were observed in Gombe formation apart from two samples from Doho locality with TOC of 20.2 and 6.84 and HI values of 178 and 176. However earlier studies of Akande et al. (1998) and Obaje et al. (1999) had separately reported TOC values of 12.5 wt.% from the Yolde Formation and 2.4 wt.% from the lower Pindiga Formation, respectively. For the present study, a total of 829 laboratory measured TOC are available with the highest TOC value being 3.382 and the lowest as 0.31. The TOC values have been aggregated into sources rock classification based on the scheme of Peters and Cassa (1994) as shown in table 1 below. However, this does not follow any stratigraphic pattern and was not tied to formation.

Table 1 - TOC ranges for well A and well B

TOC Ranges	Source Rock Ranking	Well A	Well B
0 – 0.5	Poor	119	38
0.5 – 1	Fair	325	148
1 – 2	Good	119	69
2 – 4	Very Good	6	3
>4	Excellent	0	0

Passey Overlay

Resistivity and sonic log were used in implementing Passey overlay techniques. The plot of measure TOC versus Compressional sonic as well as the plot of measured TOC versus resistivity shows no correlation between TOC and these two well log parameters for both wells as seen in figures 3 to 6. The R^2 values are

0.011 and 0.027 for resistivity and density respectively in well According to figure 3 and 4. For the second well, R^2 values is 0.265 for sonic and 0.190 for resistivity (figures 5 and 6). These plots show weak and very weak linear relationship between measured TOC and the two logs.

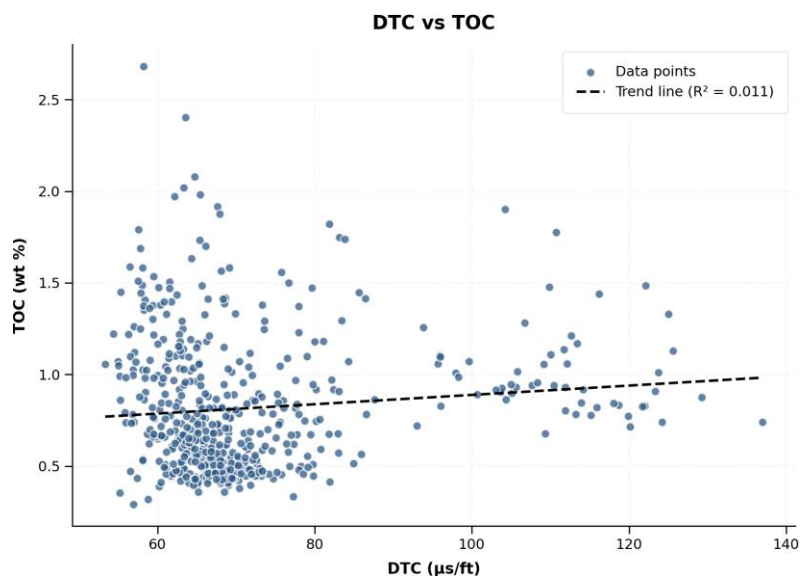


Figure 3 Plot of measured TOC versus Compressional sonic in Well A

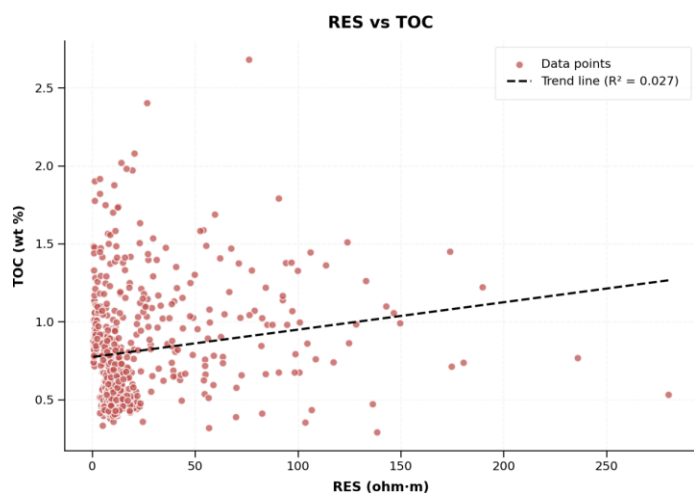


Figure 4 Plot of measured TOC versus Resistivity in Well A

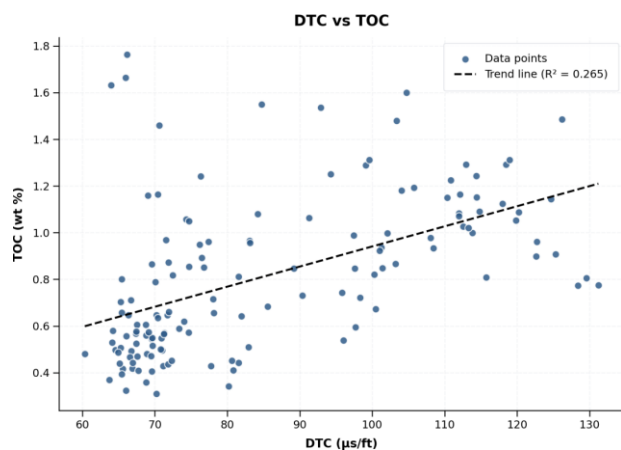


Figure 5 Plot of measured TOC versus Compressional sonic in Well B

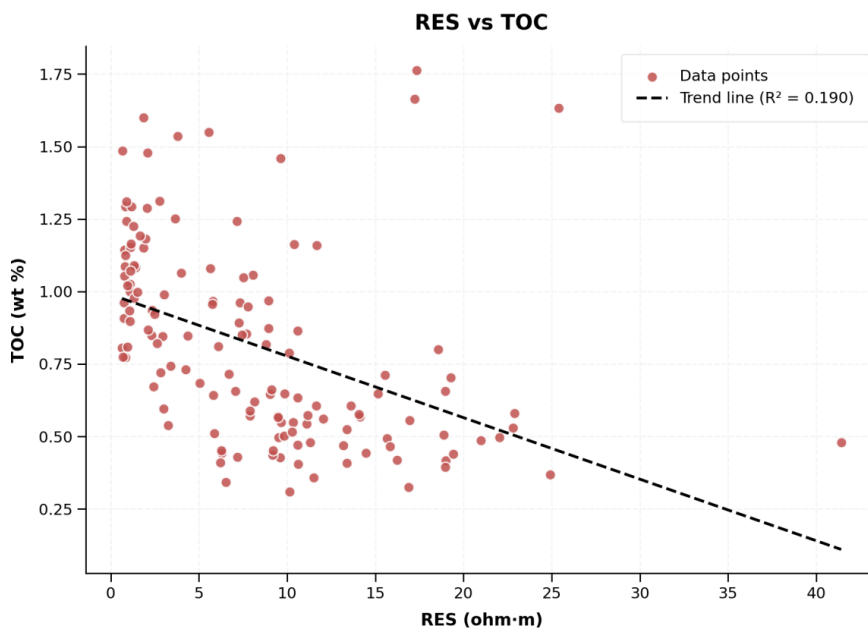


Figure 6 Plot of measured TOC versus Resistivity in Well B

TOC values were predicted for various formations using Passey Overlay method after obtaining level of organic maturity (LOM) of 12 from the cross plot of measured TOC against $\Delta \log R$. The result indicates very poor agreement with laboratory measured TOC and predicted TOC using Passey method according to figures 7 to 11 and tables 2 to 3. The highest R^2 value obtained using Passey Overlay method is 0.393 for Fika formation as seen in table 3. Poor performance of Passey overlay in predicting TOC as found in this study is consistent with observation of some previous workers in other basins. Alshakhs and Rezaee (2017) found Passey $\Delta \log R$ evaluation of TOC not suitable for Goldwyer Shale Formation in the Canning Basin. Similarly, Abiodun et al. (2018) determined Total Organic Carbon (TOC wt%), in Paleogene Well 002 in Western Niger Delta and the results of TOC derived from geochemical analysis correlated poorly with that estimated using Passey Overlay model. The low R^2 value may be due to the poor linear relationship between measured TOC and compressional sonic and resistivity used in Passey overlay method as observed in figures 3 to 6. Meanwhile, previous workers had already identified that assumption of linear relationship is a major limitation of Passey model (Wang, et al., 2016).

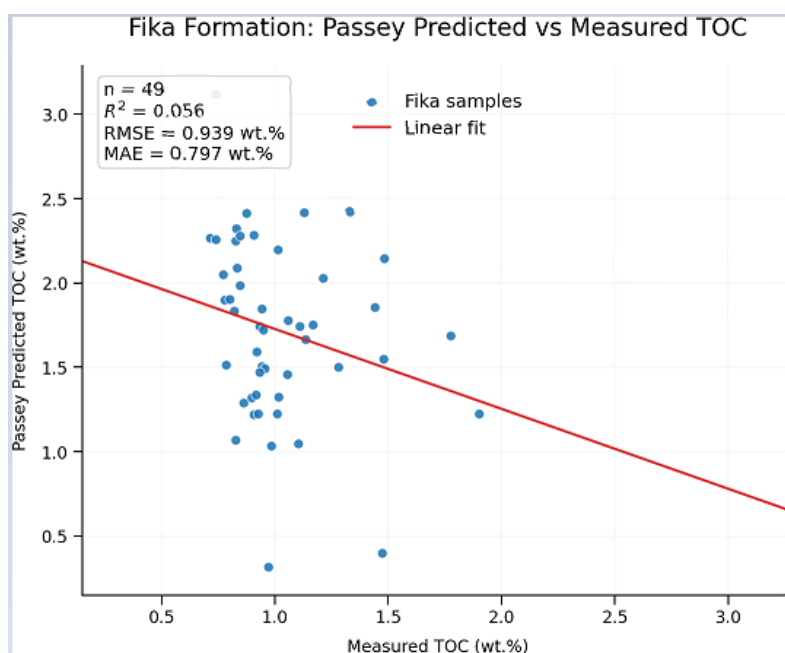


Figure 7 Passey Predicted TOC for Fika Formation

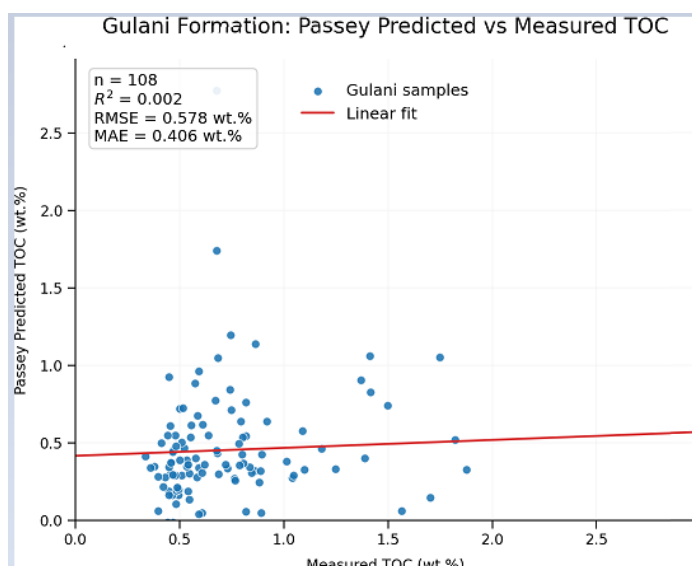


Figure 8 Passey Predicted TOC for Gulani Formation

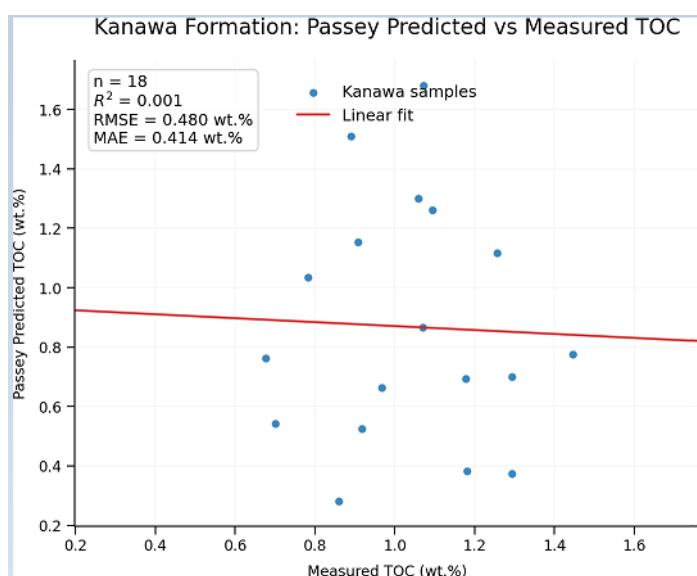


Figure 9 Passey Predicted TOC for Kanawa

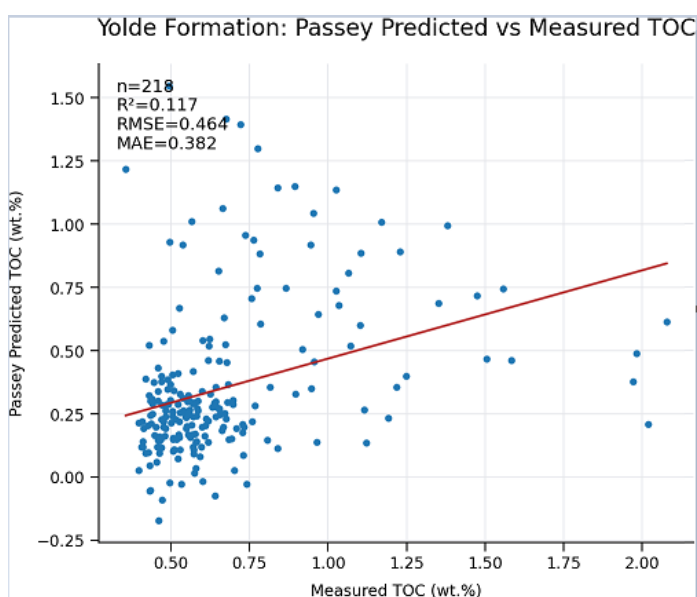


Figure 10 Passey Predicted TOC for Yolde Formation

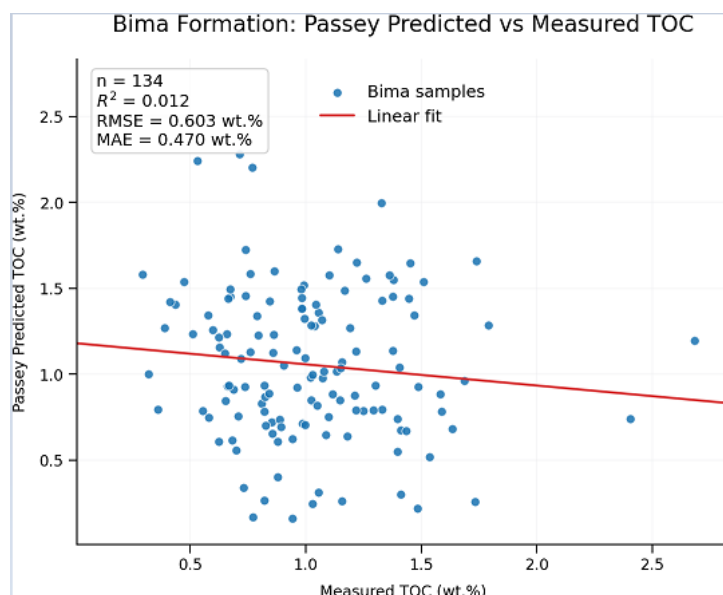


Figure 11 Passey Predicted TOC for Bima Formation

Table 2 Coefficient of Determination (R^2) for Passey Determined TOC in Well A

Formations	R^2
Fika	0.0557
Gulani	0.0023
Kanawa	0.0013
Yolde	0.1174
Bima	0.0118
All Formation	0.1137

Table 3 Coefficient of Determination (R^2) for Passey Determined TOC in Well B

Formations	R^2
Fika	0.3930
Pindiga	0.0647
Yolde	0.0836
Bima	0.0023
All Formation	0.2605

Multilinear Regression

To examine the relationship between TOC and well logs / rock physics, Pearson correlations was applied. Pearson correlation coefficient in Well A revealed a very weak association between measured TOC and well logs and rock physics parameters except Neutron porosity log which indicted a weak relationship to

measured TOC according to Evans (1996) framework as seen figure 12. Since Pearson's coefficient only detects straight-line relationships, this implies the relationship could be nonlinear or there is no relationship at all. There is a need to predict the TOC looking for nonlinear patterns. However, in Well B, the Pearson correlation coefficient revealed a very weak association between measured TOC and Gamma ray, Density and Vp/Vs whereas there is moderate relationship between TOC and the rest of the parameters which are compressional sonic, shear sonic, Neutron Porosity, P-Impedance, resistivity, S-Impedance, Lambda-Rho, Mu-Rho following Evans (1996) criteria according to figure 13. Based on the Pearson correlation analysis, different feature groups were selected to reduce high inter-feature correlations and mitigate the risk of multicollinearity, as presented in Tables 4 and 5. Highly correlated features often contain redundant information and may adversely affect model performance. Therefore, the feature groups were developed using a redundancy-grouping approach, whereby features exhibiting strong pairwise correlations were separated. This approach was adopted because correlation coefficients greater than 0.8 or 0.9 between predictor variables are generally considered indicative of potential multicollinearity (Jong, 2019).

Table 4 Feature Groups for well A

Groups	Features
Group 1	Compressional sonic, Shear sonic, Gamma Ray, Neutron Porosity, Acoustic Impedance, Resistivity, Density, Shear Impedance, Vp/Vs, Lambda-Rho and Mu-Rho
Group 2	Gamma Ray, Neutron Porosity, Density, Shear Impedance, Vp/Vs, Lambda-Rho and Mu-Rho
Group 3	Gamma Ray, Compressional sonic, Density, Shear Impedance, Vp/Vs, Lambda-Rho and Mu-Rho
Group 4	Gamma Ray, Compressional sonic, Neutron Porosity, P-Impedance and RHOB

Table 5 Feature Groups for well B

Groups	Features
Group 1	Compressional sonic, Shear sonic, Gamma Ray, Neutron Porosity, Acoustic Impedance, Resistivity, Density, Shear Impedance, Vp/Vs, Lambda-Rho and Mu-Rho
Group 2	Compressional sonic, Gamma Ray, Resistivity, Density, Vp/Vs and Lambda-Rho
Group 3	Gamma Ray, Resistivity, Density, Vp/Vs, Lambda-Rho and Mu-Rho
Group 4	Neutron Porosity, Gamma Ray, Density, Shear Impedance, Vp/Vs, Lambda-Rho and Mu-Rho

Just like Passey Overlay computed TOC, Multilinear regression computed TOC is generally low in Well A and well B as seen in table 4 and table 5 except for Fika formation in Well B as shown in table 5. Also, it was observed that the multilinear regression has higher prediction accuracy in Well B compared to Well A from tables 4 and 5 although the overall R^2 values are generally low in all the feature groups. This low R^2 values finding is not consistent with the work of Haris et al., (2017) who reported a good correlation between multilinear regression estimated TOC with measured TOC from the laboratory. Similarly, the finding of Alshakhs and Rezaee (2017) and Manaf's (2017) in Canning Basin and South Sumatra basin respectively also diverge from the result obtained in Gongola Basin. However, the multilinear regression

outperformed the Passey empirical model in both Well A and Well B as seen in table 4 and 5. From the Pearson correlation results in figures 12 and 13, there is generally very weak to weak linear relationships between TOC and most of the variables according to the framework of Evans (1996). This weak to very linear relationship implies means there is non-linear relationship between TOC and some of the features or no relationship at all. This is a pointer to the reason behind the low performance of multilinear regression in this research.

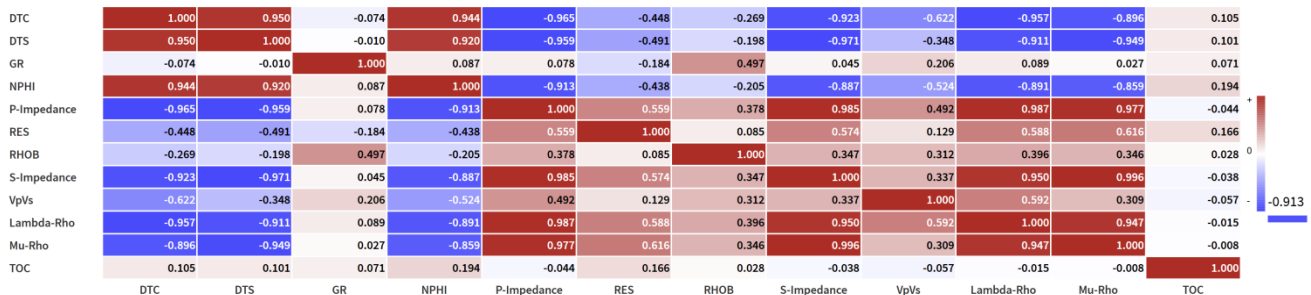


Figure 12 Pearson correlation coefficient for well A

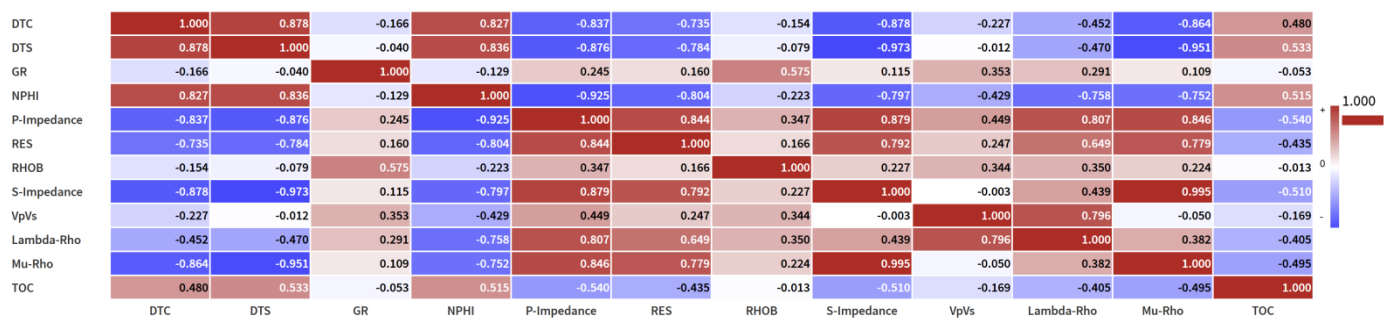


Figure 13 Pearson correlation coefficient for well B

Table 4 Coefficient of Determination (R²) for Multilinear Regression Estimated TOC in well A

Formations	Group 1	Group 2	Group 3	Group 4
All Formations	0.1983	0.1802	0.1301	0.1568
Fika	0.2331	0.1577	0.1740	0.1354
Gulani	0.2051	0.1904	0.1104	0.1858
Kanawa	0.6696	0.4320	0.4635	0.3395
Yolde	0.2232	0.1916	0.1659	0.1796
Bima	0.0834	0.0572	0.0598	0.0487

Table 5 Coefficient of Determination (R²) for Multilinear Regression Estimated TOC in well B

Formations	Group 1	Group 2	Group 3	Group 4
All Formations	0.3861	0.3337	0.3492	0.3323
Fika	0.8000	0.6576	0.6615	0.4901

Pindiga	0.4985	0.3210	0.3310	0.3841
Yolde	0.4036	0.2521	0.2379	0.2499
Bima	0.5971	0.4595	0.4457	0.4929

Machine Learning Model

A total of nine machine learning models were tested of which three are regularized linear regression, one is kernel-based algorithm and the other five are tree-based algorithms. The best performing machine learning models are the tree-based algorithms according to figure 14 to 17. This is because they were able to capture the non-linear relationship between the log responses and rock physics parameters with the measured TOC. The machine learning result are examined under three categories.

Regularized Linear Regression Models

Three regularized linear regression models were tested. Low performance was recorded by these regularized linear regression models namely Lasso (L1) Regression, Ridge (L2) Regression, and LASSO LARS according to figures 16 and 17. This confirms that the assumption of a linear relationship between the log responses and TOC by Passey et al (1990) is not correct as already noted by Wang et al (2016). This observation prompted the need for more complex algorithm since machine learning models based on linear relationship are performing low in predicting TOC in this study.

Support Vector Machine (SVM)

The performance of support vector machine (SVM), a kernel-based learning algorithm, was also consistently low in both wells with a maximum of 0.33 and 0.64 R^2 in well A and well B respectively according to figure 16 and 17. In this study, support vector machine (SVM) was optimized with Radial Basis Function (RBF) kernel only. This result outperformed the Passey overlay method, the multilinear regression and the regularized linear regression models according to figure 16 and 17. However, the support vector machine (SVM) did not perform well in this study compared to the work of Bolandi, et al., (2017) who analyzed the organic richness of source rocks in the Kazhdumi Formation, in the Persian Gulf basin and reported a good coefficient of determination for support vector machine (SVM). Bolandi, et al, (2017) reported that the SVM optimized with Radial Basis Function (RBF) kernel outperformed ANN, achieving higher accuracy of 93.69% compared to 90.77%. Whereas the same optimization yields low result in Gongola basin as seen in figure 16 and 17. Similarly, Wang, et al., (2018) tested three models for predicting TOC from shale gas well in Tonghua Basin, China and found particle swarm optimization-least square support vector machine (PSO-LSSVM) outperforming artificial neural network based backpropagated algorithm (ANN-BP), and least square support vector machine (LSSVM). Rui, et al., (2019) also found Support-Vector Regression Machine (SVR) optimized with Particle Swarm Optimization (PSO) having higher correlation coefficient compared with the results of an improved $\Delta\log R$ method and multilayer perception method. May be testing Support Vector Machine (SVR) optimized with Particle Swarm Optimization (PSO) will yield better accuracy in Gongola basin data, however, the current research did not explore this. The research of Tan, et al, (2015) who explored the use of support vector regression (SVR), which is an extension of support vector machine (SVM), for predicting TOC in organic shales in Jiumenchong formation and found it outperforming other predictive models such as Passey's $\Delta\log R$ and conventional regression models.

Tree-based Machine Learning Models

Out of the five high performing tree-based algorithms tested, Gradient Boosted Tree and Decision Tree had the highest accuracy score of 0.98, followed by Extra Tree, XGBoost and Random Forest respectively according to figures 16 and 17 of Well A and Well B. Hyperparameter tuning was carried out to further optimize the tree-based algorithms of Gradient Boosted Trees, Decision Tree, XGBoost, Random Forest and Extra Trees for well A and well B. However, Gradient Boosted Trees and Decision Tree were overfitting in

some of the feature groups in Well B according figure 17. The result of these tree-based machine learning models agrees with some previous works. Sun, et al (2023) experimented three machine learning models of which two are tree-based models of random forest (RF), and XGBoost to predict the TOC content using well logs. The findings shows that the two tree-based machine learning models performed better than the other model. The performance of Random Forest model provides the best prediction for the TOC content, with $R^2 = 0.915$, $MSE = 0.108$, and $MAE = 0.252$. The study is a good pointer of how robustness and scalability ensemble tree-based models are in predicting TOC as observed in this research work in Gongola basin also. The finding of this research also agrees with result of the research by Wu et al. (2025) which showed good prediction of total organic carbon using an optimized XGBoost machine learning model from well logs obtained in the Qiongzhusi Formation of the central Sichuan Basin. The finding demonstrates that the XGBoost framework achieves better predictive precision evident in higher coefficient of determination (R^2) of 0.81, a reasonable lower mean squared error (MSE), and smaller error fluctuations. The coefficient of determination (R^2) in this study for XGBoost model was better than the finding of Wu et al (2025) as shown in figures 16 and 17. Also, Zhao et al., (2026) optimize XGBoost framework with comprehensive feature engineering leading to strong TOC prediction performance of 0.9395 for R^2 value. The author did the research with well log data from the Shahejie Formation in the Dongying Depression, Bohai Bay Basin, China. However, from Gongola Basin, in this present study we obtained high coefficient of determination of 0.94 without feature engineering but rather a combination of well logs and rock physics parameters as seen in figure 16 to 17 and table 4 to 5. The excellent finding in using Gradient Boosting Tree in this study conforms with the report of Zhang et al. (2022) in unconventional shale reservoirs in the Linfen area of the Qinshui Basin, China. The finding of Zhang et al. (2022) shows that Gradient Boosting Decision Tree (GBDT) yielded exceptional determination coefficient R^2 of 0.9254, a minimized RMSE of 0.032, and the highest overall evaluation weight of 0.42. The finding of this present study yielded a determination coefficient that exceeded the result of Zhang et al. (2022) as seen in figure 17 with a record of 0.98 for well B. The major finding of this work is the high performance of five three based models as seen in figure 13 and 14.

Similar to this result is the work of Siddique et al. (2025) that examined the use of tree-based machine learning models such as Extra Trees (ET), Gradient Boosting (GB) and XGBoost, to predict total organic carbon (TOC) and compared result with rock-eval pyrolysis measured total organic data in the Shahejie Formation of the Dongying Depression, Bohai Bay, China. Siddique et al. (2025) findings show that Extra Trees and Gradient Boosting Trees significantly outperform the conventional Passey method in predicting total organic carbon (TOC) from well logs data. The Extra Trees model achieves superior predictive accuracy with an R^2 of 0.9825 and an RMSE of 0.2483. The result of Extra Trees in this study is lower than the determination of Siddique et al. (2025) as shown in figure 16 and 17. The gross features considered important by most of the algorithm are neutron porosity, resistivity and Lambda-Rho are the most important features by most of the models. Considering neutron porosity as a very important feature means the potential source rock may have high concentration of hydrogen rich kerogen (solid organic matter) rather than actual fluid-filled micro-porosity. High feature importance for resistivity means the models are successfully identifying thermal maturity and hydrocarbon generation potential of the source rock.

This also agrees with the level of organic maturity of 12 obtained in this research. High Lambda-Rho could mean that variability in TOC is related to lithofacies changes. This may be associated with interbedding of shale, sand, and siltstone observed in some of the formations (Nwojiji et. Al., 2013). Interbedding of shale, sand and siltstone occurred in Gulani, Yolde and Bima formation in well A lithostratigraphy. Also, the lithostratigraphy of well B have similar interbedding in Pindiga, Yolde and Bima formation. However, well B also have more homogeneous lithology compared to well A. Therefore, the better predictive performance observed in Well B suggests that machine learning algorithms are more effective in geologically homogeneous formations where the relationships between TOC and predictor variables remain relatively consistent. The more uniform lithology in Well B likely reduced variability in log responses and rock physics properties, enabling the models to identify clearer patterns between the input features and TOC. In contrast, the greater lithological heterogeneity in Well A may have introduced complex and nonlinear variations in petrophysical and elastic properties, making it more challenging for the models to generalize the relationship between the predictor variables and TOC.

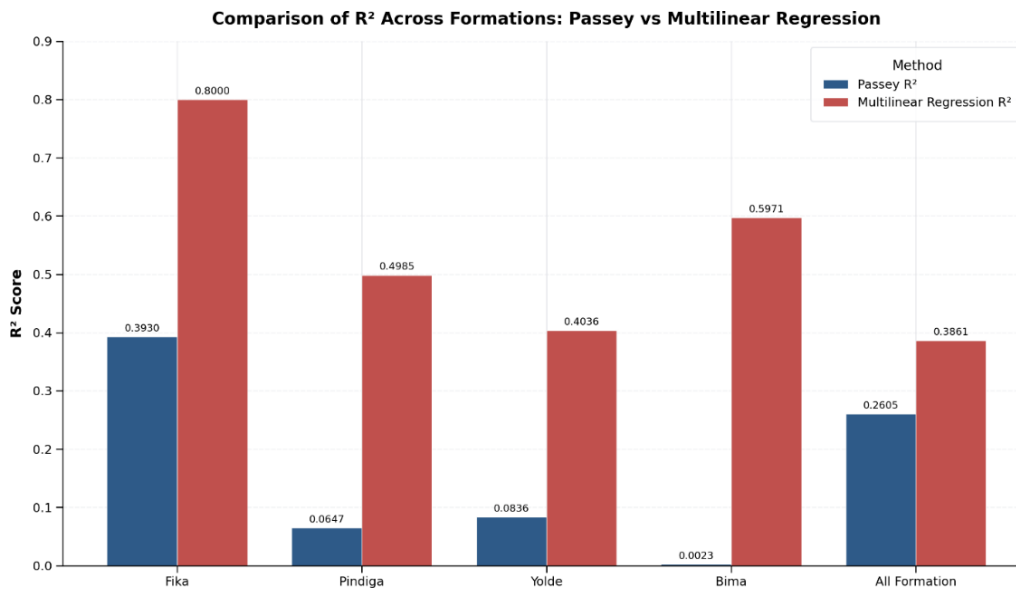


Figure 14: Comparison of R² Across Formations: Passey vs Multilinear Regression in Well A

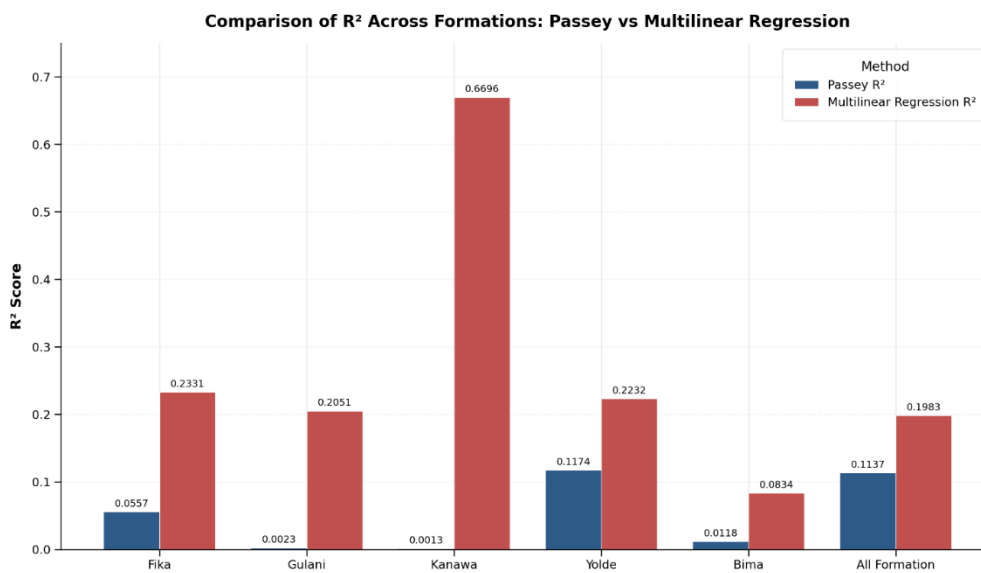


Figure 15: Comparison of R² Across Formations: Passey vs Multilinear Regression in Well B

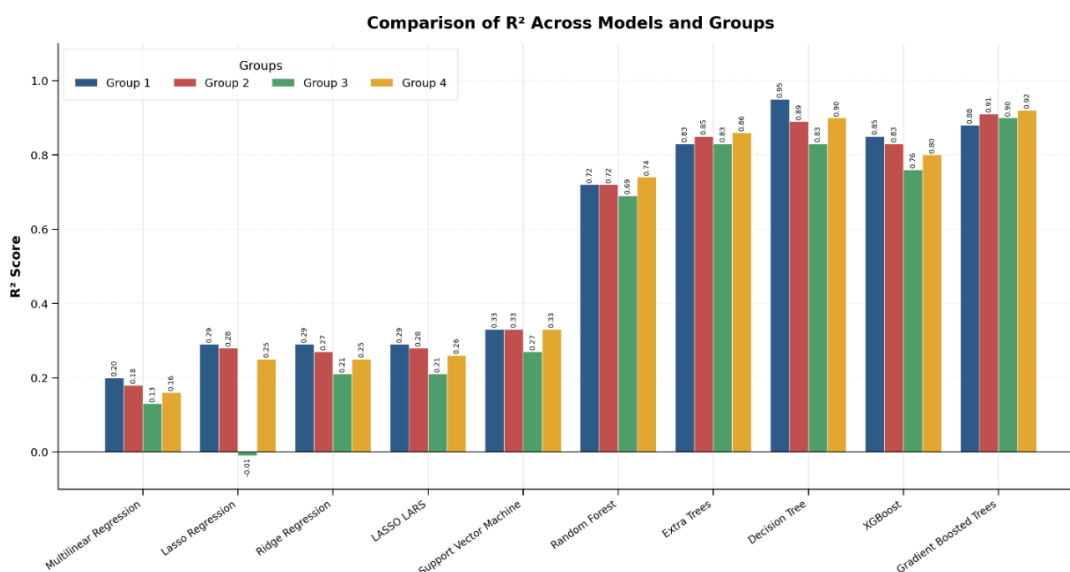


Figure 16 Comparison of R² Across Different Models and Feature Groups in Well A

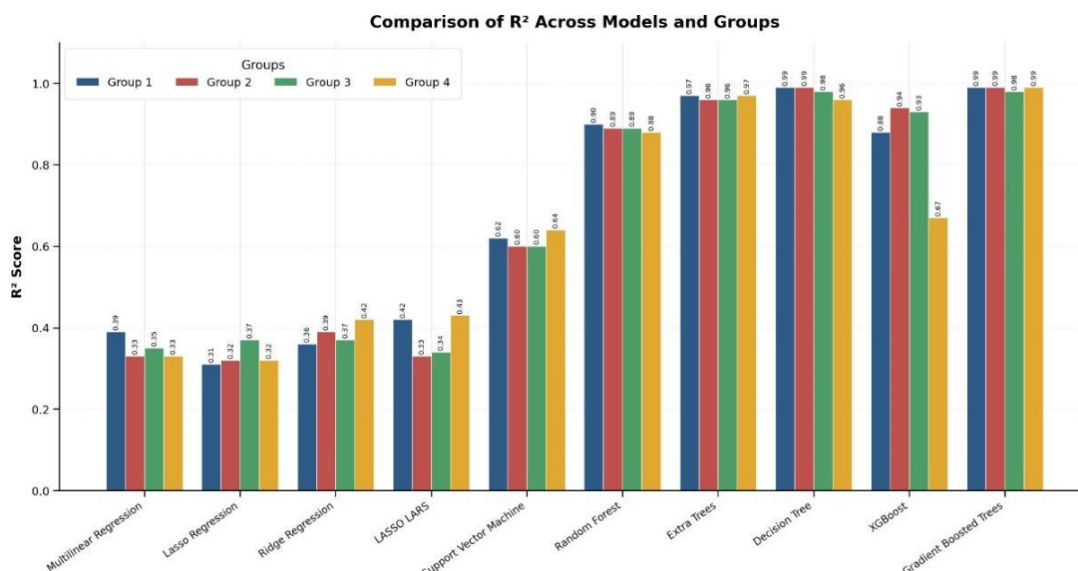


Figure 17 Comparison of R^2 Across Different Models and Feature Groups in Well B

CONCLUSION

- Prediction of TOC using Passey Overlay model, Multilinear regression models and regularized linear regression machine learning models are generally low.
- TOC predicted by Passey Overlay model, Multilinear regression model and regularized linear regression models are consistently low because of non-linear relationship between measured TOC and the well logs/ rock physics parameters.
- Machine learning models outperformed Passey Overlay and Multilinear models in prediction of TOC.
- The tree-based machine learning models had the best prediction of TOC.
- Among the five high-performing tree-based models, Gradient Boosted Tree demonstrated the highest predictive accuracy. It was followed in descending order of performance by Decision tree, extra tree, XGBoost and Random Forest
- The most prominent features considered by tree-based algorithms are neutron porosity, resistivity, and Lamda-rho.
- All the models predicted TOC better in Well B than Well A because of stronger relationship between TOC and well logs in Well B.

REFERENCE

1. Akande, S. O., Ojo, O. J., Erdtmann, B. D., & Hetenyi, M. (1998). Paleoenvironments, source rock potential and thermal maturity of the Upper Benue rift basins, Nigeria: implications for hydrocarbon exploration. *Organic Geochemistry*, 29, 531–542.
2. Alshakhs, M., & Rezaee, R. (2017). A New Method to Estimate Total Organic Carbon (TOC) Content, an Example from Goldwyer Shale Formation, the Canning Basin. *The Open Petroleum Engineering Journal*, 10, 118-133. doi:10.2174/1874834101710010118.
3. Abiodun, K., Saleh, A. S., Eze, S., & William, S. (2018). “Source Rock Analysis Using Well Logs In Western Niger Delta,” *Journal of Applied Geology and Geophysics*, Volume 6, Issue 2 Ver. I (Mar. – Apr. 2018), PP 70-85.
4. Bolandi, V., Kadkhodaie, A., & Farzi, R. (2017). “Analyzing organic richness of source rocks from well log data by using SVM and ANN classifiers: a case study from the Kazhdumi formation, the Persian Gulf basin, offshore Iran,” *Journal of Petroleum Science and Engineering*, vol. 151, pp. 224–234.
5. Evans, J. D. (1996). *Straight forward statistics for the behavioral sciences*. Brooks/Cole Publishing.
6. Benkhelil, J. (1989). The origin and evolution of the Cretaceous Benue Trough (Nigeria), *Journal of African Earth Sciences*. 8, 251-282.

7. Dike, E.F.C., 1993. Stratigraphy and structure of the Kerri-Kerri Basin, Northeastern Nigeria. *Journal of Mining and Geology*, 29(2), 77–93.
8. Hamis, M. B., Yandoka, B. M. S., and Usman, M. B. (2024). Facies distribution, paleoenvironment and hydrocarbon potential of Kanawa Member of Pindiga Formation, Gongola Basin, Northern Benue Trough, Nigeria. *Discover Geoscience*, 2(1), Article 46. <https://doi.org/10.1007/s44288-024-00052-y>
9. Manaf, P. E., Suparno, S., Haris, A., Usman, A., & Riyanto, A. (2017). Organic Shale Analysis using Geochemical Data and Seismic Attributes: Case Study of Talang Akar formation, South Sumatera Basin. *AIP Conference Proceedings*, 1862(030178), 1-5. doi:10.1063/1.4991282
10. Nwojiji C.N., Osterloff P., Okoro A. U., Ukeri P. O. (2013). Palynostratigraphy and Age of the Sequence Penetrated by the Kolmani River 1 Well in the Gongola Basin, Northern Benue Trough, Nigeria. *Journal of Geosciences and Geomatics*. Vol. 1(1), pp. 15-21; DOI:10.12691/jgg-1-1-3
11. Obaje, N. G., Ulu, O. K. and Petters, S. W. (1999). Biostratigraphic and geochemical controls of hydrocarbon prospects in the Benue trough and Anambra basin, Nigeria: *Nigerian Association of Petroleum Explorationists Bulletin*, v. 14, p. 18–54.
12. Obaje, N. G., Wehner, H., Scheeder, G., Abubakar, M. B. and Jauro, A. (2004). Hydrocarbon Prospectivity of Nigeria's Inland Basins: From the Viewpoint of Organic Geochemistry and Organic Petrology. *AAPG Bulletin*, Vol. 88, No. 3, pp. 325-353.
13. Passey, Q. R., Creany, S., Kulla, J. B., Moretti, F. J., & Stroud, J. D. (1990). A Practical Model for Organic Richness from Porosity and Resistivity Log. *American Association of Petroleum Geologists Bulletin*, 74(12), 1777-1794.
14. Petters, S. W. (1982): Central West Africa Cretaceous – Tertiary benthic foraminifera and Stratigraphy, *Paleontographica Abtelung, A*, **179**, 1–104.
15. Siddique, A. B., Alvee, A. M. M., Ana, P. D., Jahan, L. N., & Hashan, M. (2025). Tree-based, boosting, and stacked models for accurate prediction of total organic carbon from conventional well logs. *Energy Reports*, 13, 4491–4513. <https://doi.org/10.1016/j.egyr.2025.04.021>
16. Sun, J., Dang, W., Wang, F., Nie, H., Wei, X., Li, P., Zhang, S., Feng, Y., and Li, F. (2023), Prediction of TOC Content in Organic-Rich Shale Using Machine Learning Algorithms: Comparative Study of Random Forest, Support Vector Machine, and XGBoost, May 2023, *Energies* 16(10):4159, DOI: 10.3390/en16104159
17. Tan, M., Song, X., Yang, X., and Wu, Q. (2015). “Support-vector-regression machine technology for total organic carbon content prediction from wireline logs in organic shale: a comparative study,” *Journal of Natural Gas Science and Engineering*, vol. 26, pp. 792–802.
18. Rui, J., Zhang, H., Zhang, D., Han, F., & Guo, Q. (2019). “Total organic carbon content prediction based on support-vector regression machine with particle swarm optimization,” *Journal of Petroleum Science and Engineering*, vol. 180, pp. 699–706.
19. Wang, P., Chen, Z., Pang, X., Hu, K., Sun, M., & Chen, X. (2016). “Revised models for determining TOC in shale play: example from devonian Duvernay shale, western Canada sedimentary basin,” *Marine and Petroleum Geology*, vol. 70, pp. 304–319.
20. Wu, Q., Pang, H., Zhang, B., Jiang, F., Wu, L., Chen, J., Ma, M., & Huo, X. (2025). Application of shale TOC prediction model using the XGBoost machine learning algorithm: A case study of the Qiongzhusi Formation in central Sichuan Basin. *Carbonates and Evaporites*, 40(1), Article 8. doi.org
21. Zarboski, P., Ugodunlunwa, F. Idornigie, A., Nnabo, P and Ibe, K. (1997). Stratigraphy and structure of the Cretaceous Gongola basin, Northeastern Nigeria. *Bulletin des Centres de Recherches Exploration-Production. EIF-Aquitaine* 21, 153-185.
22. Zhao, P., Ma, H., Rasouli, V., Liu, W., Cai, J., & Huang, Z. (2017). “An improved model for estimating the TOC in shale formations,” *Marine and Petroleum Geology*, vol. 83, pp. 174–183.
23. Zhao, Z., Zhong, G., Diao, F., Ding, P., & He, J. (2026). A feature engineering and XGBoost framework for prediction of TOC from conventional logs in the Dongying Depression, Bohai Bay Basin. *Geosciences*, 16(1), 44. doi.org