Natural Resources Research (© 2021)
https://doi.org/10.1007/s11053-021-09988-1

Original Paper

Evaluation of Source Rock Potentiality and Prediction of Total Organic Carbon Using Well Log Data and Integrated Methods of Multivariate Analysis, Machine Learning, and Geochemical Analysis

Edwin E. Nyakilla,[1] Selemani N. Silingi,[1,3] Chuanbo Shen,[1,2,4] Gu Jun,[1,4] Alvin K. Mulashani,[1] and Patrick E. Chibura[1]

Received 9 July 2021; accepted 19 November 2021

In this study, integrated approaches based on multivariate analysis (MVA), machine learning (ML), and geochemical analysis are proposed to investigate the potential of hydrocarbon reserves and total organic carbon (TOC) prediction. These approaches employed the MVA technique as a future selection method in source rock evaluation. We used geochemical data from 30 core samples taken equally from wells SS-5 and SS-7. Geochemical parameters, namely TOC, free hydrocarbon, thermal pyrolysis hydrocarbon, hydrogen index, production index, and oxygen index, were determined for statistical evaluation. IBM SPSS statistical software and MATLAB (R2020a) were used for MVA and ML, respectively. The performance of the models built using MVA and ML were evaluated by, among others, coefficient of determination (R^2) and mean square error (MSE). Findings revealed that fair through good to excellent source rock with TOC ranging from 0.85 to 2.95 wt% are hosted in the Triassic beds of Tanga. A high 1.61% Ro at a mature peak of 463 °C predominates with the existence of type III/II kerogen that can produce both oil and gas. Considering TOC prediction from conventional well log data, optimized Gaussian process regression showed the best performance followed by MVA and support vector machine, giving the MSEs of 0.5629, 0.6172, and 0.7023, respectively. In terms of prediction accuracy, their R^2 values of 0.952, 0.9346, and 0.835, respectively, were in good agreement with the geochemical results. The concurrence of geochemical analysis, ML, and MVA revealed that the Tanga basin has great hydrocarbon potential of great economic importance. The study revealed that combining MVA and other methods can be applied to assess the hydrocarbon resource potential of other prospects around the globe.

KEY WORDS: Source rock, Geochemical analysis, Cluster analysis, Factor analysis, Pearson's correlation coefficient(r), Machine learning.

[1]Department of Petroleum Engineering, School of Earth Resources, China University of Geosciences, Wuhan 430074, China.
[2]Department of Petroleum Geology School of Earth Resources, China University of Geosciences, Wuhan 430074, China.
[3]Department of Geology, Earth Sciences Institute of Shinyanga, (ESIS, P.O.Box 1016 Shinyanga, Tanzania.
[4]To whom correspondence should be addressed; e-mail: cbshen@cug.edu.cn, gujun@cug.edu.cn

INTRODUCTION

The machine learning (ML) and Rock–Eval pyrolysis techniques have been widely used to evaluate source rocks. Despite the fact that previous approaches face vast challenges in terms of computing efficiency and Rock–Eval fallacy (Dembicki

Published online: 11 January 2022

Jr, 2009; Xie et al., 2018), an alternative method of multivariate analysis (MVA) as an adopted method in investigating source rock for a better result is required. The assessment of source rock is often predicted based on the amount of total organic carbon (TOC), the quality, and the capability of thermal maturation (Omran & Alareeq, 2018; Aziz et al., 2020). Rock–Eval pyrolysis is the most used technique in geochemical screening (Chalk et al., 1997; Lafargue et al., 1998; Mashhadi & Rabbani, 2015; Hakimi et al., 2017; Gentzis, 2018). It is also used for petroleum, soil, and sediments analysis at the industrial level (Carvajal-Ortiz & Gentzis, 2015; Mashhadi & Rabbani, 2015; Romero-sarmiento et al., 2016, 2017). Artificial neural networks (ANN), support vector machine (SVM), and Gaussian process regression (GPR) are the most employed methods of ML for predicting TOC and evaluation of source rock (Bolandi et al., 2017; Asante-Okyere et al., 2020; Mahmoud et al., 2020; Rui et al., 2020). However, ANN, SVM, and GPR classifiers may lead to the challenge of over-fit, iterative tuning of parameters, and selection of best kernel function (Xie et al., 2018; Golden et al., 2019; Rui et al., 2020; Mulashani, et al., 2021a, 2021b).

To provide the distinctive to these techniques, it's vitally important to integrate them with multivariate analysis (MVA), which is significantly connected with the principle of statistics, which means measuring and evaluating more than one independent statistical variable times-wise (Johnson & Wichern, 2002; Izenman, 2008). MVA has several sub-methods for variables analysis that lead to a much deeper improved analysis of source rock evaluation. MVA techniques of clustering by K-means, factor analysis, principal component analysis, and the person correlation coefficient were applied. The present study examines the performance of integrating geochemical, MVA, and machine learning techniques to enhance results accuracy.

The research firstly investigated the benefit of combining MVA, geochemical analysis, and ML in the evaluation of source rock and in the prediction of TOC. Secondly, we intended to ascertain the oil and/or gas hydrocarbon potential of the case study area. Thirdly, the study revealed the best method for source rock assessment. For further investigation, several wells have been drilled in this basin. This study revealed that a combination of multiple techniques can lead to improved source rock evaluation.

GEOLOGICAL SETTING

Tanga basin is one of the new potential pools for hydrocarbon reserves in Tanzania's coastal basins situated near the Kenya border in the northern part of Tanzania (Fig. 1). Coastal basins (Tanga, Ruvu, and Mandawa) were speculatively formed by the initiation process of the Permo-Triassic Continental rift. The Tanga basin is mainly influenced by the Tanga fault, which trends in NNE-SSW (Fig. 2). From Permo-Carboniferous to lower Jurassic, the column of sedimentary bedrocks in these basins starts with the continental Karoo sequence. The Karoo rocks are largely represented by conglomerate and shale (Ngerengere Formation) fluviatile arkosic sandstones that characterize the Selous-Ruvu-Tanga rift reservoir (Mbede & Dualeh, 1997; Kapilima, 2003). The Karoo sediments in the Tanga Basin are referred to as Tanga beds(Wopfner, 2002; Said et al., 2015). Tanga beds are formally classified as lower, middle, and upper sequences (Fig. 3).

MATERIALS AND METHODS

Sample Selection

In total, 30 core samples, 15 each from wells SS-5 and SS-7, were selected for analysis. Each sample was washed with dichloromethane to remove residual drilling fluids. Sixty (60) g of each sample was crushed into powder by a motor and pestle. Each powdered sample was put into a crucible for geochemical analysis (Chalk et al., 1997; Behar et al., 2001; Carvajal-Ortiz & Gentzis, 2015; Wu et al., 2017).

Geochemical Analysis of Source Rock

The geochemical investigation was performed by Rock–Eval pyrolysis 6, where the TOC, thermal maturity (T_{max}), free hydrocarbon (S_1), production index (PI), CO_2 released during thermal breakdown of kerogen (S_3), hydrogen index (HI), thermal pyrolysis hydrocarbon (S_2), and oxygen index (OI) were identified. The determination and quantification of parameters were used to assess the T_{max} and quality of the source rock (Li et al., 2018; El Hajj et al., 2019). A portion of about 69.98 mg from each sample was measured and examined by pyrolysis.

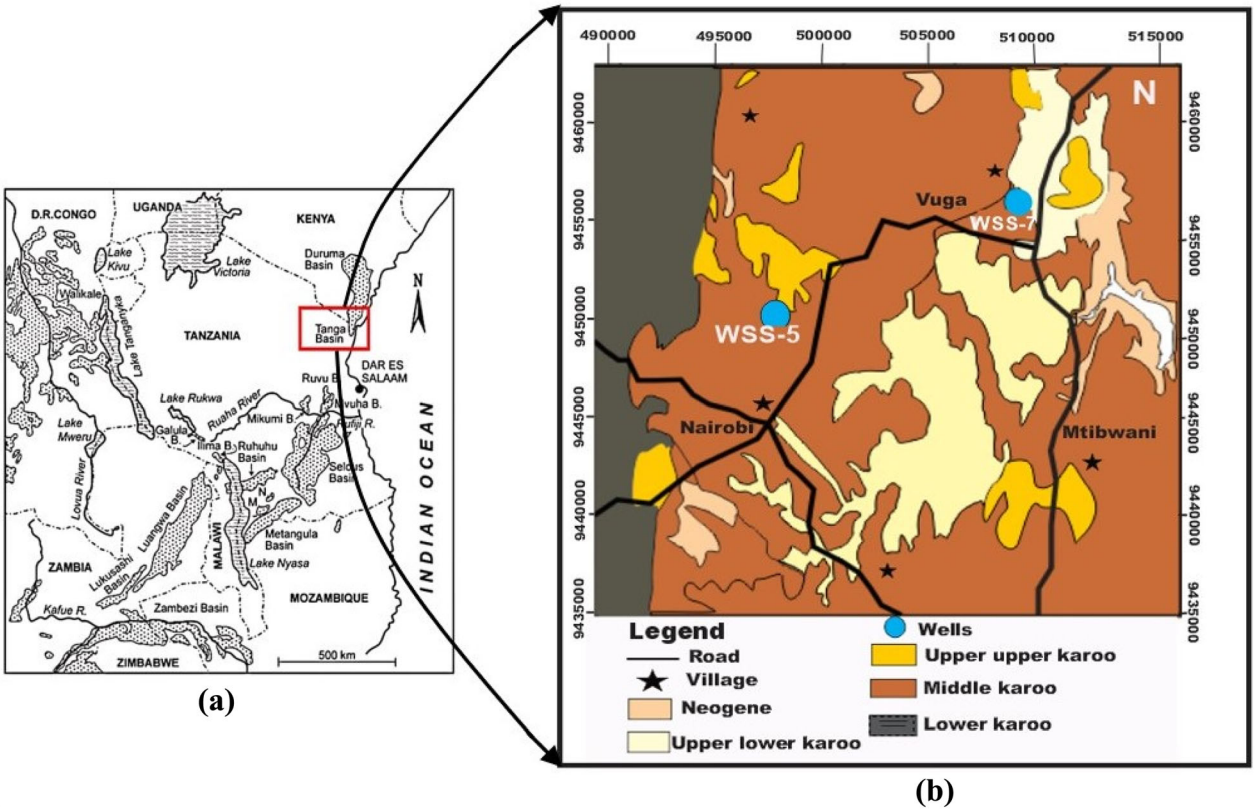

Figure 1. Location map of Tanzania country **a** and study area of Tanga basin **b** from which Well SS-5 and Well SS-7 were extracted.

Samples were kept in a helium-inert atmosphere at a constant temperature of 700 °C during parameters quantification. TOC Eq. 1, HI Eq. 2, OI Eq. 3, and PI Eq. 4 were calculated from pyrolysis data (Peters, 1986; Mashhadi & Rabbani, 2015; Hazra et al., 2017; Godfray & Seetharamaiah, 2019).

$$\%\mathrm{TOC} = 0.082(S_1 + S_2) + S_3]/10 \quad (1)$$

$$\mathrm{HI} = \frac{100 \times S_2}{\mathrm{TOC}} \quad (2)$$

$$\mathrm{OI} = \frac{100 \times S_3}{\mathrm{TOC}} \quad (3)$$

$$\mathrm{PI} = \frac{S_1}{[S_1 + S_2]} \quad (4)$$

Source Rock Evaluation Based on Multivariate Analysis

Different multivariate statistical techniques using well log data of GR (gamma-ray), DT (sonic), MSFL (resistivity), CALI (caliper), PEF (photo-electric effect), RHOB (bulk density), and NPHI (neutron porosity) were performed through IBM SPSS statistical tool (version 26). These techniques included *K*-means clustering, principal components analysis (PCA), Pearson's correlation coefficient (*r*), and factor analysis (FA) to enhance the accuracy of source evaluation and TOC prediction (Shen et al., 2019; Asante-Okyere et al., 2020).

K-Means Clustering

K-means clustering algorithm is defined as an intensive way of unsupervised machine learning techniques for which the dataset is categorized into *k* number that keeps the inner point of clusters as closer as possible while maintaining their area for predetermined non-overlapping clusters (Al-Mohair et al., 2015; El Nady et al., 2015a, 2015b). K-means clustering method was determined stepwise as indicated in model summary (Fig. 4) where Knee and Silhouette methods were applied to find the number of the cluster from well log data.

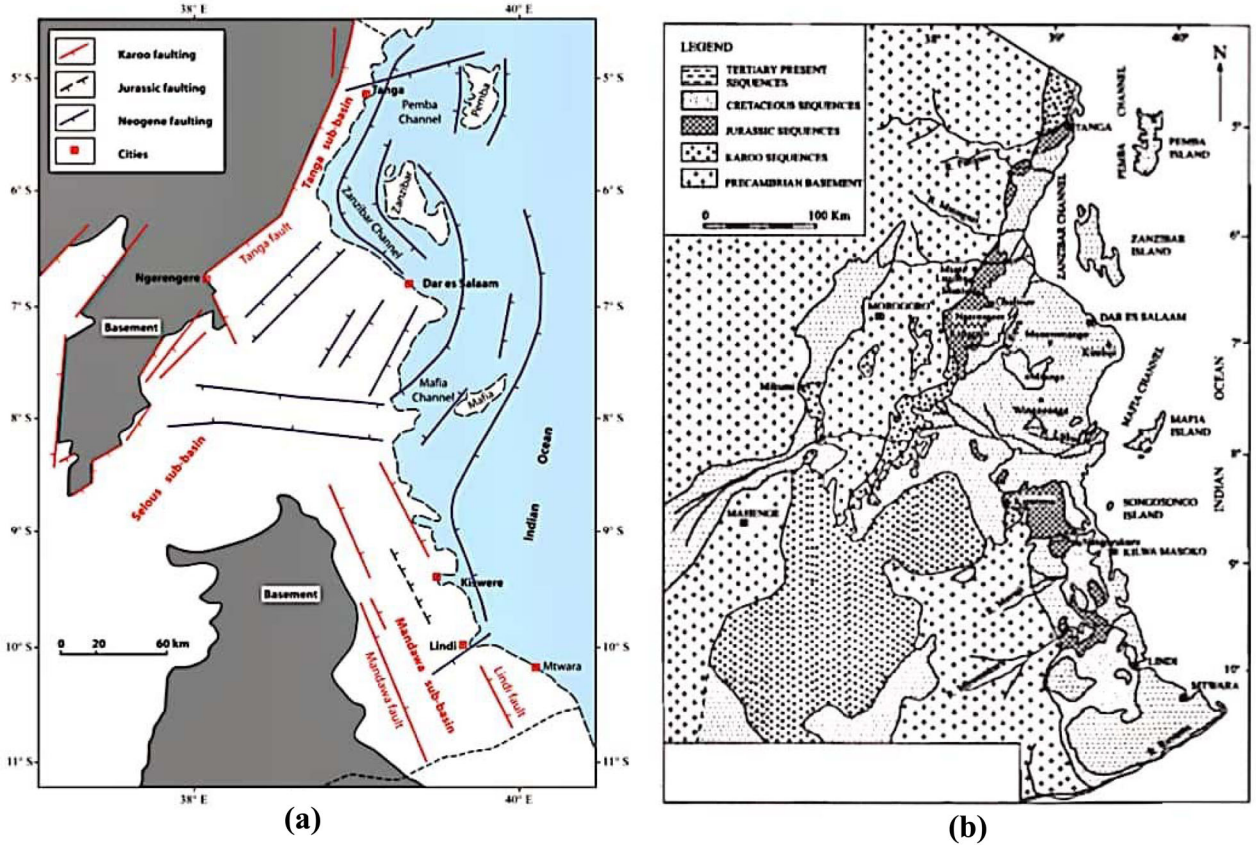

Figure 2. **a** Major structural faulting zones of Tanzania coastal basin. **b** Geological map of coastal basins (modified from Kapilima (2003) and Said et al. (2015).

The centroid for clusters was considered as:

$$D = \sqrt{(x-a)^2 + (x-b)^2 + (x-n)} \qquad (5)$$

ruwhere D represents the Euclidian distance of each selected well log data, *x, b, and n* are variables. Centroids for per cluster were identified as the sum of squared error by reducing the objective function (Edwards et al., 1999; Bramer, 2016; Zaremotlagh et al., 2016), thus:

$$E = \sum_{i=1}^{k} \sum\nolimits_{p \epsilon c_i} dist\left(G_p, C(i)\right)^2 \qquad (6)$$

where E represents the sum of square error of all selected data, G_p, represents well log data in space, $C(i)$ is the centroid. Then, for all selected well log data, the weighted summation and standard deviation, which represent the distinct centroids of all cases, were expressed respectively as:

$$A = \frac{1}{n} \sum_{i=n}^{n} ai \qquad (7)$$

where A is mean of specific variables collection, ai represents the sub-variables, and n represents the total count of variable numbers, and

$$S = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}} \qquad (8)$$

where S stands for standard deviation, x is variable value in the data, $\overline{x}$ is average of specific variables, n is number of variables in the data set.

Factor Analysis

Factor analyses in this work were executed through the principal axis factoring method to assist

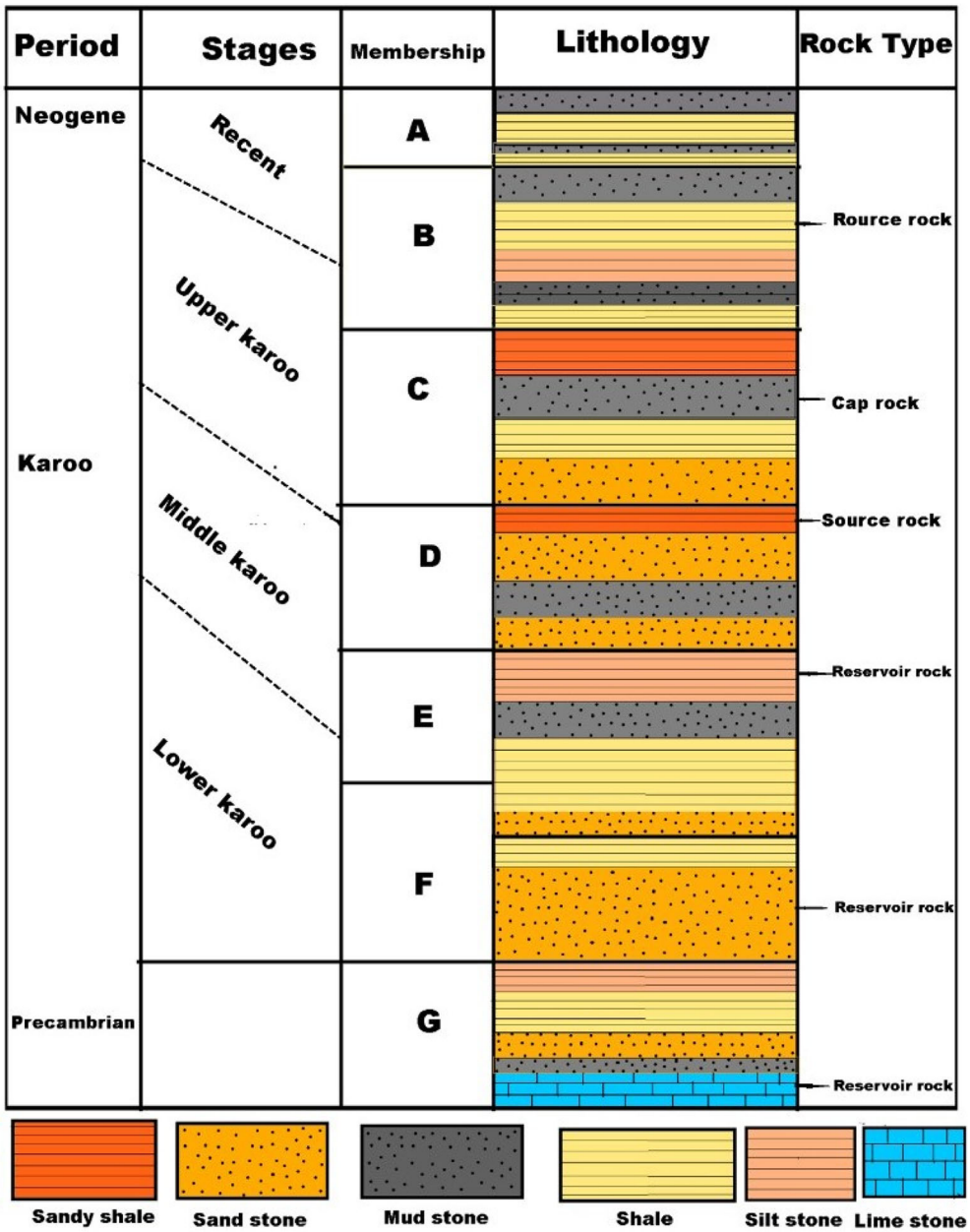

Figure 3. Illustration of the stratigraphy of the Tanga basin from the oldest lithology (Basement) to youngest lithology.

in well log data interpretation (Zhou et al., 1983; Zumberge, 1987; El Nady et al., 2015a, 2015b). Standardization of the data was made by subtracting the mean from the values of the corresponding log and then dividing the difference by the standard deviation to ensure that the well log data have the same content and format for analysis (Walden et al., 1992). The rotation was prohibited as criticized by Temple (1978), Pan et al. (2017), and Giannakopoulou et al. (2018).

Principal Components Analysis

Principal components analysis of the selected well log data was carried using SPSS statistical software. To determine the best statistical factor approach, rotation through varimax with Kaiser normalization was considered.

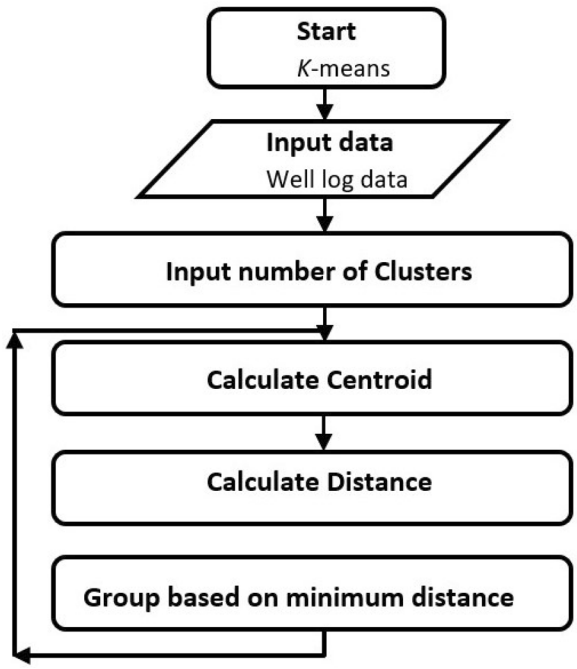

Figure 4. Schematic diagram of the model summary.

Pearson's Correlation Analysis

Analysis by Pearson correlation (r) was used to anticipate the linear relationship between well log data and TOC (Bolandi et al., 2017) to determine the importance of each factor in TOC prediction and reservoir assessment (Handhal et al., 2020). It was calculated as:

$$r_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} \tag{9}$$

where r is the linear correlation between two variables x, and y, σ_x represents the standard deviation of x, cov represents the covariance and σ_y represents the standard deviation of y. Well log data including GR, NPHI, DT, RHOB, LLD, PEF have great influence on rock formation as they measure organic-rich of rock, concentration hydrogen atom,

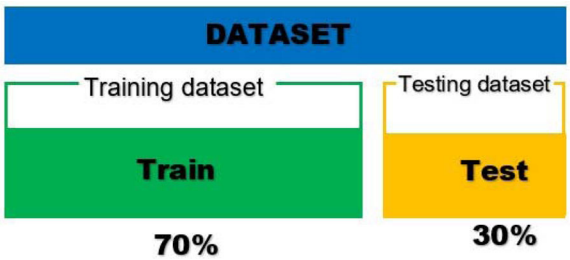

Figure 6. Division of the dataset into two sets (one for training and the other for testing).

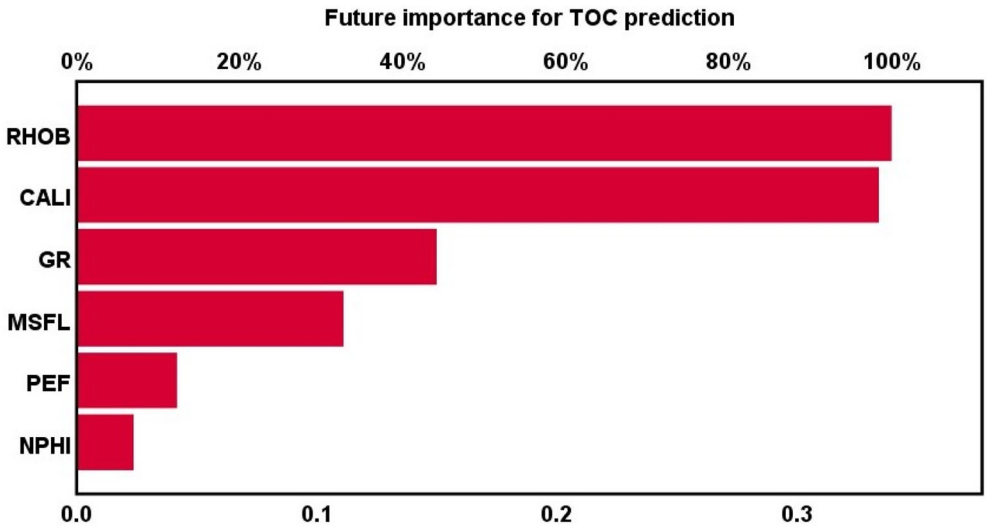

Figure 5. Variable significance values.

Table 1. Geochemical results for wells SS-5 and SS-7 of Triassic source Tanga basin

Sample number	Depth (m)	TOC (wt%)	S_1 (mg Hc/g)	S_2 (mg HC/g)	$S_1 + S_2$ (mg HC/g)	HI (mg Hc/g)	OI (mg Hc/g)	T_{max} (°C)	Ro (%)	PI (S_1/($S_1 + S_2$))	S_1/TOC (mg Hc/g/wt%)
WSS-5–10	138.67	0.98	1.13	2.72	3.85	277.55	33	460	1.12	0.29	1.15
WSS-5–12	143.82	1.07	1.25	3.12	4.37	291.59	19	451	0.96	0.29	1.17
WSS-5–15	157.41	1.15	1.18	3.34	4.52	290.43	16	461	1.14	0.26	1.03
WSS-5–16	160.82	1.01	0.98	3.25	4.23	321.78	22	456	1.05	0.23	0.97
WSS-5–17	174.23	1.51	1.73	3.64	5.37	241.06	16	456	1.05	0.32	1.15
WSS-5–27	191.89	0.85	0.98	3.12	4.1	367.06	21	454	1.01	0.24	1.15
WSS-5–29	1100.65	0.91	1.07	3.78	4.85	415.38	2	444	0.83	0.22	1.18
WSS-5–32	1110.30	1.24	1.27	3.45	4.72	278.23	17	456	1.05	0.27	1.02
WSS-5–34	1120.58	1.16	1.49	3.49	4.98	300.86	20	447	0.89	0.30	1.28
WSS-5–36	1130.52	2.59	2.69	4.51	7.2	174.13	23	453	0.99	0.37	1.04
WSS-5–42	1220.45	1.38	1.34	3.61	4.95	261.59	22	456	1.05	0.27	0.97
WSS-5–57	1240.81	2.75	2.59	4.74	7.33	172.36	14	455	1.03	0.35	0.94
WSS-5–59	1260.82	2.95	2.78	5.04	7.82	170.85	34	463	1.17	0.36	0.94
WSS-5–61	1280.82	2.51	2.53	4.55	7.08	181.27	30	463	1.17	0.36	1.01
WSS-aa5-66	1302.81	2.56	2.28	4.76	7.04	185.94	17	461	1.54	0.32	0.89
WSS-7–14	123.22	2.72	2.21	4.49	6.7	165.07	21	450	0.94	0.33	0.81
WSS-7–18	137.35	1.11	1.17	3.39	4.56	305.41	21	458	1.08	0.26	1.05
WSS-7–13	153.64	0.88	0.98	2.98	3.96	338.64	10	451	0.96	0.25	1.11
WSS-7–25	190.71	1.04	1.01	4.01	5.02	385.58	11	454	1.01	0.20	0.97
WSS-7–26	198.12	0.99	0.85	3.16	4.01	319.19	21	457	1.07	0.21	0.86
WSS-7–28	1109.22	1.12	1.25	3.09	4.34	275.89	22	456	1.05	0.29	1.12
WSS-7–32	1120.20	1.14	0.9	3.34	4.24	292.98	35	459	1.41	0.21	0.79
WSS-7–35	1131.73	1.51	1.11	3.35	4.46	221.85	28	458	1.08	0.25	0.74
WSS-7–37	1148.57	1.37	1.25	2.84	4.09	207.30	17	461	1.14	0.31	0.91
WSS-7–40	1210.16	2.42	2.57	4.49	7.06	185.54	29	458	1.58	0.36	1.06
WSS-7–48	1230.79	1.65	1.64	3.77	5.41	228.48	17	453	1.43	0.30	0.99
WSS-7–49	1250.23	2.52	0.79	4.64	5.43	184.13	28	460	1.52	0.15	0.31
WSS-7–50	1270.55	2.51	1.43	5.01	6.44	199.60	12	450	1.48	0.22	0.57
WSS-7–52	1290.70	2.46	2.78	4.38	7.16	178.05	24	461	1.4	0.39	1.13
WSS-7–53	1305.81	2.15	2.95	4.86	7.81	226.05	14	457	1.61	0.38	1.37

bulk density, matrix resistivity content, which affect the TOC level of rock formation.

MACHINE LEARNING ALGORITHM

ML is a computational technique that creates a model with the help of sample data. In current days some ML and deep learning techniques have been used for creating the model for TOC prediction (Amiri Bakhtiar et al., 2011; Mahmoud et al., 2019, 2020; Rui et al., 2020). Two ML algorithms have been used to predict TOC values from the Tanga basin. A complete selected well log data was used for ML analysis. SVM and GPR were the algorithms used for this purpose (Rui et al., 2020; Mulashani, et al., 2021a, 2021b). These two algorithms have been widely used in engineering, industrial, and medical aspect to perform various tasks like face detection, image classification, and making a reliable prediction (Priddy & Keller, 2005; Suzuki, 2011; Alquisom, 2016; Shalaby et al., 2019). Regression modeling through SVM and GPR algorithms was performed in MATLAB (R2020a) to assess the geochemical performance.

Support Vector Machine

The kernel function Eq. 10 was used in model building to predict TOC through support vector regression (Vapnik 2013). The key objective of SVM is to acquire the smallest loss function $f(x)$ and the curve as flat as possible (Kaloop et al., 2020), thus:

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) k(x_i, x) + b \qquad (10)$$

where kx_i, x stand for kernel function, $x_i and x$ are training and testing data respectively, $\alpha_i - \alpha_i^*$ are

Lagrangian multipliers. Optimization was done by the Lagrange multiplier (Azimi-Pour et al., 2020).

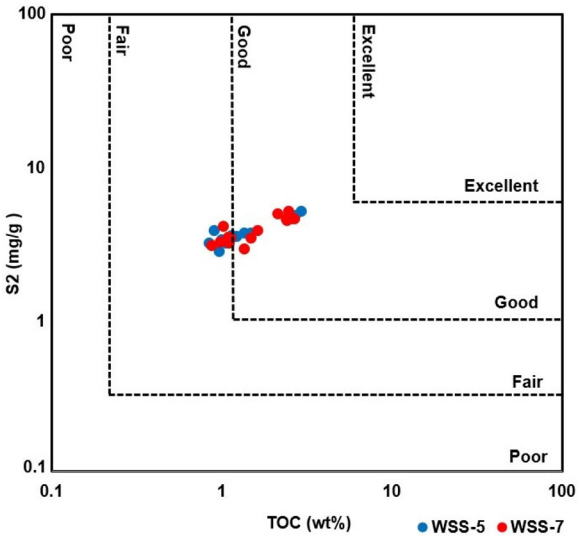

Figure 7. Plot of TOC vs S_2 to evaluate the generative ability of hydrocarbons showing fair to good source rock.

Gaussian Process Regression

GPR was performed through kernel function to measure similarities and predict the value of unknown data (Rui et al., 2020). The technique can compute a response for the model's input variables which makes it more beneficial than others (Wu et al., 2006; Kaloop et al., 2020). The output z of the GPR model is assumed through the function $(f(x))$ of Gaussian noise model, thus:

$$z_t = f(x_t) + e_t \tag{11}$$

where $f(x)$ stands for Gaussian process, e_t model noise which obeys normal distribution (Eq. 12), $e_t N(0, \sigma_n^2)$ with variance σ_e^2 and mean = 0 for the n observation (Kaloop et al., 2020).

$$f(x)\}^{\surd}(m(x), k(x, x')) \tag{12}$$

GPR involves covariance $k(x, x')$ and means functions $(m(x))$ of the data. The covariance identifies the dependence values existing at different input points $xandx'$. The k is kernel function, which is defined as:

$$k\left(x, x^{'}\right) = \sigma^2 \exp\left(-\frac{x - x^{'2}}{2l^2}\right) \tag{13}$$

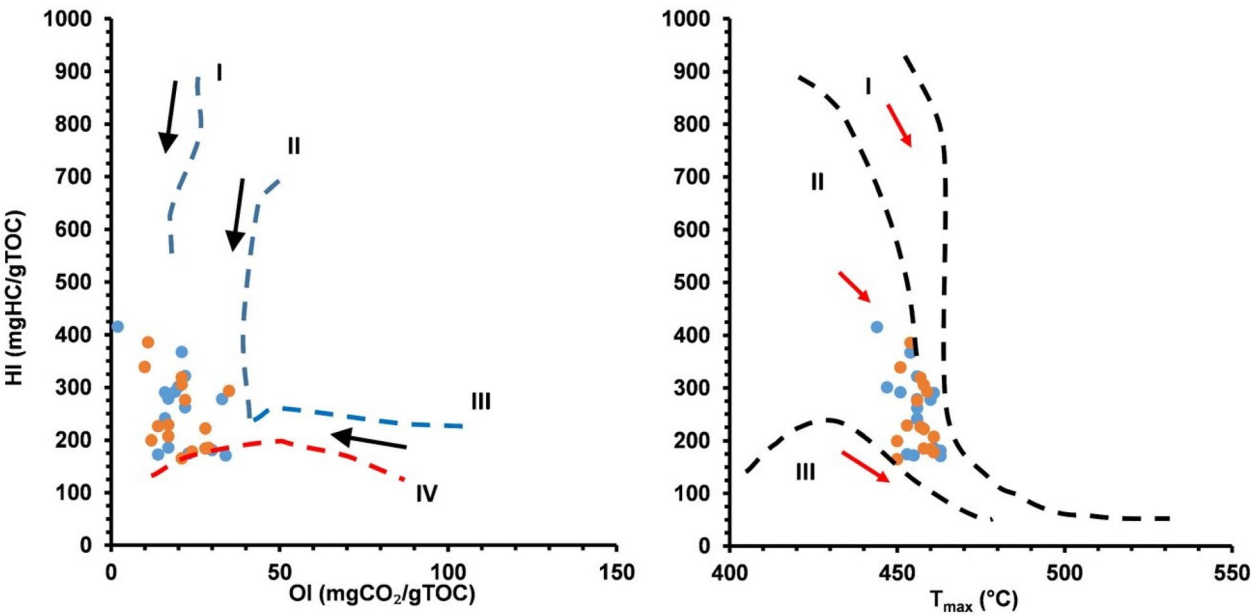

Figure 8. Cross-plot of Tanga shale formation indicating most samples are type II and type III kerogen.

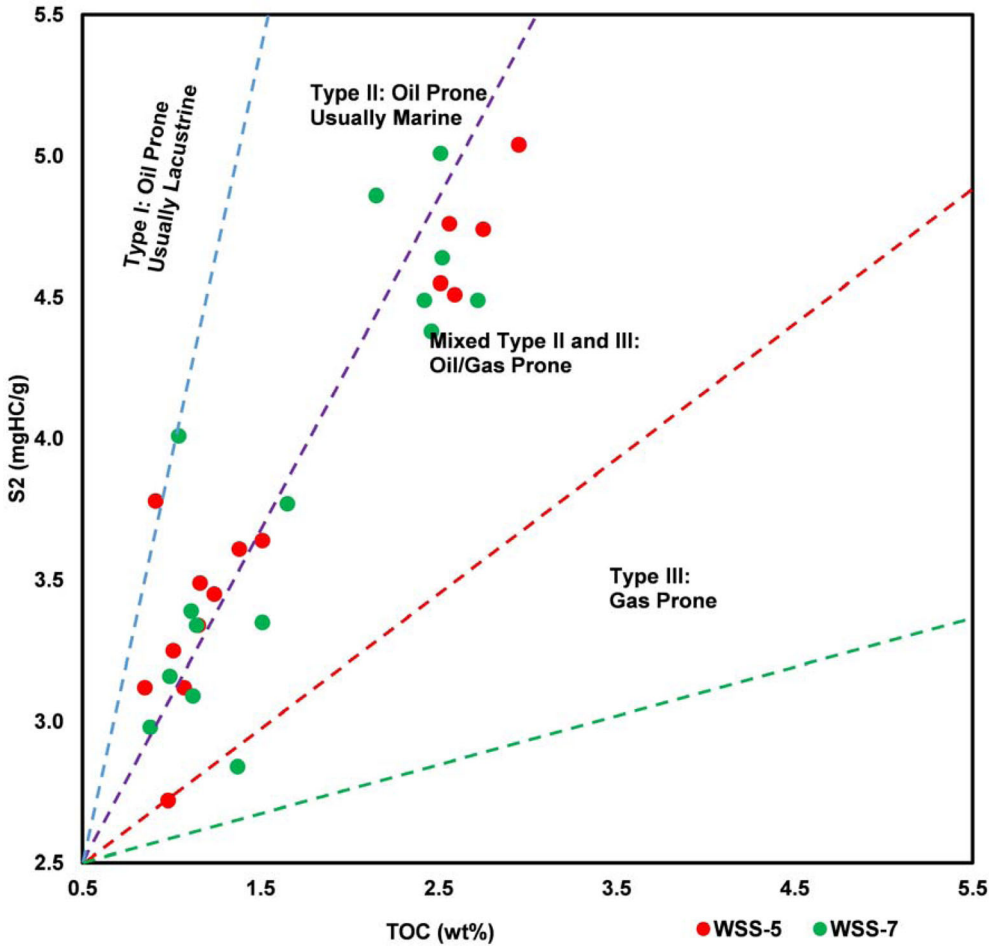

Figure 9. Plots of S_2 vs TOC showing that most of the studied samples are mixed type II and III (oil/gas prone).

where σ^2 stand for variance and l for scale length. The model data are calculated using Bayesian inference when kernel parameters are being detected.

TOC from geochemical study and well log data (GR, MSFL, CALI, PEF, RHOB, and NPHI) were used for TOC prediction. The importance of each log was ranked by considering its gain value. The importance of TOC prediction was increasing with the scores (Fig. 5), the order of importance was RHOB < CALI < GR < MSFL < PEF < NPHI. For cross-validation with two datasets, 70% as training data and 30% as testing data were considered (Fig. 6). The studied dataset involved 30 core samples taken equally from wells SS-5 and SS-7. The process was repeated 10 times to assess the stability of the algorithms. The best performing model was studied in terms of loss functions R^2 (coefficient of determination), RMSE (root mean square error), and MSE (mean square error) (Eqs.14, 15, and 16) because these functions are used for error estimation between real values (t_i) and predicted (p_i) (Asante-Okyere et al., 2020):

$$R = \frac{\sum_{i=1}^{n}(t-\bar{t})(p-\bar{p})}{\sqrt{\sum_{i=1}^{n}(t-\bar{t})+\sqrt{\sum_{i=1}^{n}(p-\bar{p})}}} \quad (14)$$

$$MSE = \frac{1}{n}\sum_{i-1}^{n}(t_i - p_i) \quad (15)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i-1}^{n}(t_i - p_i)} \quad (16)$$

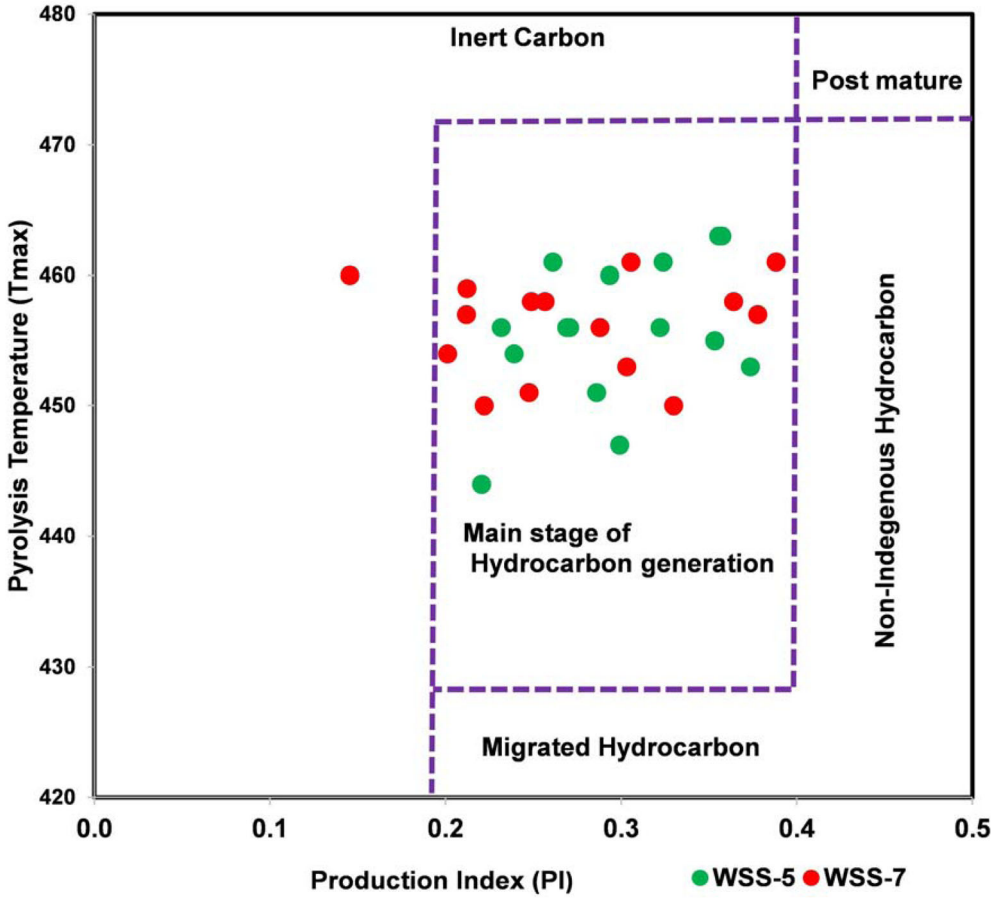

Figure 10. T_{max} versus PI cross-plot indicating that most of the samples are in the main stage of hydrocarbon generation.

where n is total number of data points, t is mean measured parameter value, and p is mean predicted parameter value.

RESULTS AND DISCUSSION

Geochemical Characterization of Source Rock

Table 1 displays geochemical data obtained from pyrolysis.

Quantity of Organic Matter

Characterization of the Triassic Tanga source rock samples was carried following the methodology by Peters (1986) and Omran and Alareeq (2018). The cross-plots of TOC (wt%) against S_2 were plotted to discriminate the indigenous hydrocarbon from a non-indigenous hydrocarbon. The TOC values of 0.85–2.95wt% indicated that the majority of samples were fair to good sources rock while some were excellent (Fig. 7) (El Kammar, 2015; Wang et al., 2020).

Evaluation of Source Rock Potentiality and Prediction

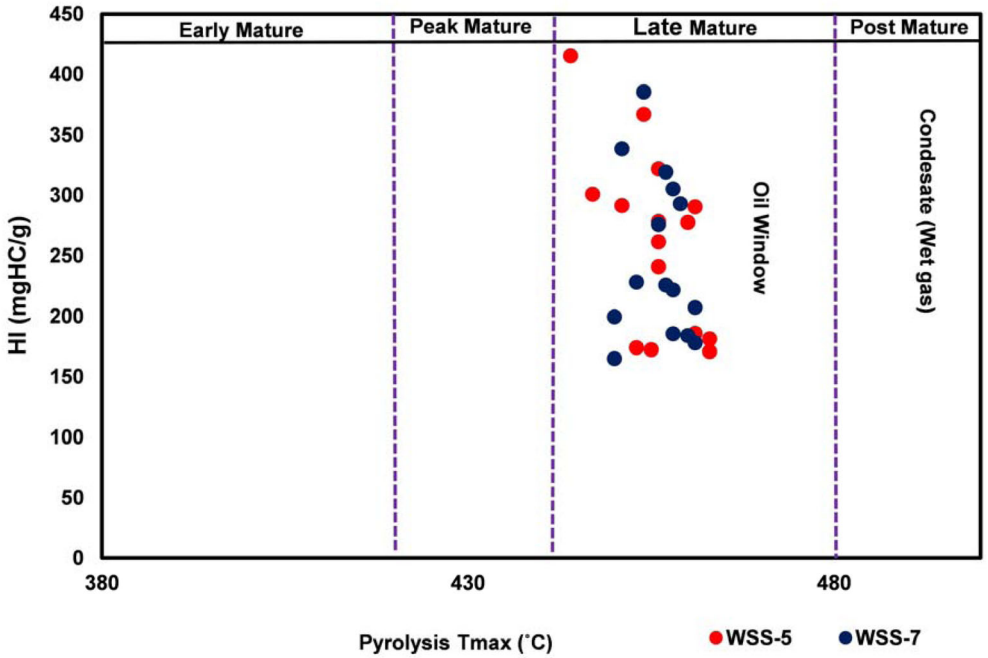

Figure 11. Cross-plot of HI against T_{max} indicating the thermal maturity stage of the basin where most of the data are in the late to post mature stage.

Type of Kerogen

HI was plotted against OI and T_{max} to anticipate kerogen categories of the Triassic Formation source rock. Figure 8 illustrates that the Tanga basin contained mostly Type II kerogen with the ability to generate oil with minor gas, and kerogen type III suggesting a source rock rich in gas prone (Langford & Blanc-Valleron, 1990; El Nady et al., 2015a, 2015b; Omran & Alareeq, 2018).

To infer kerogen classes in this formation, the S_2 against TOC (Fig. 9) was plotted additionally for complete verification of organic matter quality. We found that the wells contained both types II and II/III together with type III kerogen. The HI values of the analyzed samples ranged from 165.7 to 415.38 mg/g, which fall into the gas-oil generation window that reveals the kerogen type according to the classification by Peters (1986).

Thermal Maturity of Organic Matter

The Tanga shell formation yielded T_{max} ranging from 444 to 463 °C with mean of 456 °C (Table 1). The formation is interpreted as mature to post-mature class source rocks based on the T_{max} vs PI plot (Fig. 10) (Wu et al., 2017). The relationship depicted by this plot indicates that many of the assessed samples are in the stage of hydrocarbon generation (El Nady et al., 2015a, 2015b).

Source Rock Generation Potential

Figure 11 presents the HI vs T_{max} plot for source rock evaluation potential. It confirms that the Tanga reservoir is rich in oil with few samples showing gas-prone distribution within OM. In addition, the relationship between geochemical findings was assessed by correlation coefficient (R^2), (Eq. 10, Fig. 12). Regarding S_2 and TOC (Fig. 12a), there is

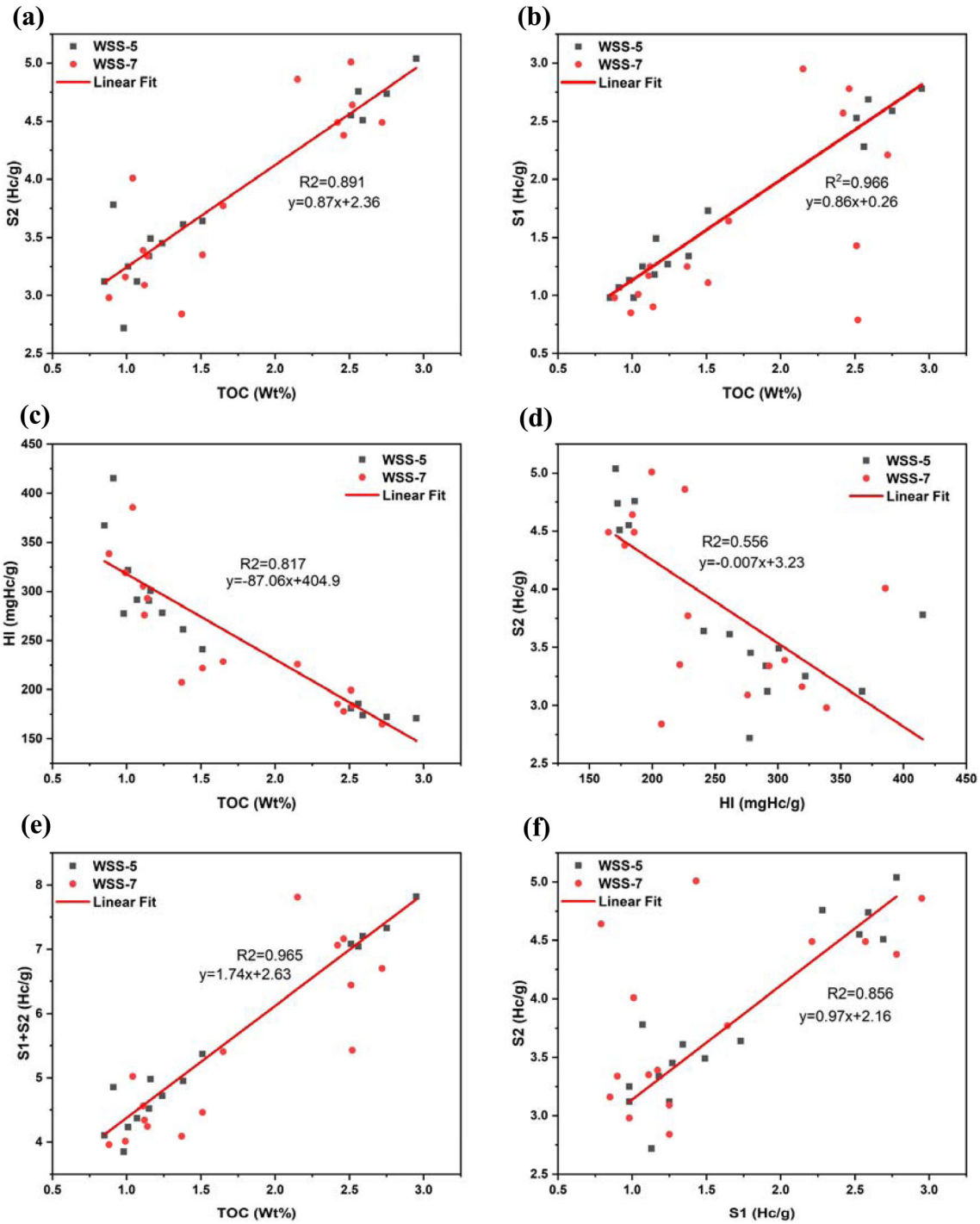

Figure 12. Cross-plots showing a linear relationship between **a** TOC and S_2, **b** TOC and S_1, **c** TOC and HI, **d** S_2 and HI, **e** $S_1 + S_2$ and TOC, and **f** S_2 and S_1.

Table 2. Results of ANOVA for wells SS-5 and SS-7 cluster

Well Logs	Mean Square	DF	MSE	RMSE	DF	F	SIG
NPHI	29.000	1	0.000	0.000	28	1,603,173,194.43	.000
RHOB	29.000	1	0.000	0.000	28	196,021,444.518	.000
PEF	29.000	1	0.000	0.000	28	4,798,945.391	.000
CALI	28.999	1	0.000	0.000	28	1,068,457.300	.000
MSFL	28.999	1	0.000	0.000	28	665,749.092	.000
GR	20.331	1	0.310	0.556	28	65.672	.000
TOC	4.264	1	0.381	0.6172	28	11.184	.002

Table 3. K-means distance and cluster membership

Case Number	Sample ID	Cluster	Distance	TOC
1	WSS-5–10	1	0.432	0.98
2	WSS-5–12	1	0.175	1.07
3	WSS-5–15	1	0.091	1.15
4	WSS-5–16	1	0.329	1.01
5	WSS-5–17	1	0.573	1.51
6	WSS-5–27	1	0.269	0.85
7	WSS-5–29	1	0.393	0.91
8	WSS-5–32	1	0.280	1.24
9	WSS-5–34	1	0.348	1.16
10	WSS-5–36	2	0.671	2.59
11	WSS-5–42	2	0.542	1.38
12	WSS-5–57	2	0.850	2.75
13	WSS-5–59	2	1.040	2.95
14	WSS-5–61	2	1.009	2.51
15	WSS-5–66	2	0.781	2.56
16	WSS-7–14	2	0.911	2.72
17	WSS-7–18	2	0.817	1.11
18	WSS-7–13	2	1.060	0.88
19	WSS-7–25	2	0.937	1.04
20	WSS-7–26	2	1.357	0.99
21	WSS-7–28	2	1.273	1.12
22	WSS-7–32	2	0.781	1.14
23	WSS-7–35	2	0.644	1.51
24	WSS-7–37	2	0.749	1.37
25	WSS-7–40	2	0.874	2.42
26	WSS-7–48	2	0.654	1.65
27	WSS-7–49	2	1.005	2.52
28	WSS-7–50	2	0.617	2.51
29	WSS-7–52	2	1.576	2.46
30	WSS-7–53	2	0.685	2.15

strong positive correlation ($R^2 = 0.891$); likewise, S_1 and TOC (Fig. 12b) depicted a highly positive correlation of ($R^2 = 0.966$), HI and TOC (Fig. 12c) showed strong negative correlation ($R^2 = 0.817$), and S_2 and HI (Fig. 12d) have moderate positive correlation ($R^2 = 0.552$). Moreover, there is high positive correlation of $R^2 = 0.856$ between $S_1 + S_2$ and TOC (Fig. 12e) and a strong positive correlation of $R^2 = 0.856$ between S_1 and S_2 (Fig. 12f). A good positive correlation ($R^2 = 0.981$) was anticipated by the model. According to the classification by Edwards et al. (1999) and El Nady et al. (2015a, 2015b), the relationship between TOC and HI as well as between $S_1 + S_2$ and TOC can be used to assess maturity level. The negative correlation between HI and TOC affirms variation in the occurrence of the former. High amounts of HI can often be found at a certain maturity level but it does not occur in less mature or over-mature stages. The highly strong positive correlation between $S_1 + S_2$ and TOC (Fig. 12e) indicates that wells SS-5 and SS-7 are highly related and that their hydrocarbon generating potential is high.

HYDROCARBON POTENTIAL BASED ON MULTIVARIATE STATISTICAL ANALYSIS

K-means Clustering

The optimal result of TOC prediction by the *K*-means model was achieved at the overall RMSE) of 0.6172 and MSE of 0.381 (Table 2). The results show that taking two clusters (*K*-means 1 and *K*-means 2 in Table 3) leads to the best bar graph clustering pattern (Fig. 13) because each log shows an almost equal range in their predictions. The minimum distance within points in a cluster (WCSS) gave a good average silhouette value of 0.7. The F-values of 11.184 for TOC, GR, MSFL, CALI, PEF, RHOB, and NPHI indicate the influence of each variable on clustering (Table 2). Moreover, RMSE and R (Eqs. 10, 12) were estimated to quantify the developed models (Table 3). The predicted result for TOC suggests a good performance agreement with

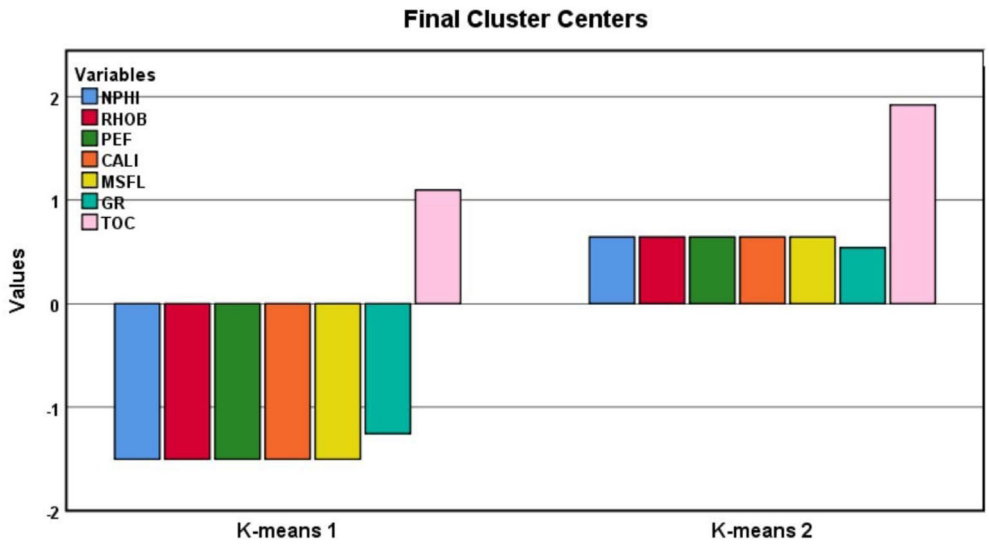

Figure 13. Cluster analysis of well log data from wells SS-5 and SS-7 indicating the effect of each variable in forming the clusters.

Table 4. Results of factor analysis by principal axis factoring

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	5.990	85.573	85.573

geochemical data. This implies that well log data have great influence on the prediction of TOC.

Factor Analysis

The findings of this analysis indicate that, based on principal axis factoring under eigenvalue extraction, one factor was the optimal solution for factor determination, as previously described by El Nady et al. (2015a, 2015b) and Pan et al. (2017). All conventional well log variables except one showed strong positive relationship with TOC (Table 4). In addition, the validity of the result was confirmed by checking the Kaiser–Meyer–Olkin (KMO) sphericity and Bartlett's test of sphericity (p-value) (Table 5). The KMO was 0.816, which indicates good variable identity.Bartlett's test was significant at a p-value of 0.00, which is reasonably strong for interpretation (Brian, 2004).

Principal Components Analysis

The relationship between well log data and TOC has been studied by Shalaby et al. (2019). The results in Table 6 show the best two components extracted by PCA. Component 1 includes GR, MSFL, CALI, PEF, RHOB, and NPHI, whose loadings are 0.847, 0.957, 0.956, 0.956, 0.957, and 0.957, respectively. These variables contributed effectively in determining TOC. Component 2 involves mainly TOC (with loading of 0.960), which contributed effectively in determining the percentage weight of TOC.

Figure 14 indicates the relative distribution of the variables within the data set, where the length dimension of the vector along the axis reflects a variable's contribution to the corresponding factor loading (Pan et al., 2017). The higher the axis receives relative factor loadings, the more the variable vector resembles the factor axis as indicated by component 1 (GR, RHOB, CALI, PEF, and MSFL) and component 2 (TOC). Table 7 presents the total variance explained by each component. Component 1 accounted for $> 76\%$ of the total variance while component 2 explained about 19.564% of the rotational variance. Positive correlations were observed among the variables (RHOB, GR, CALI, PEF, MSFL, and NPHI) within the dataset. This indicates that these variables were significantly related to each other in the factor determination.

Table 5. KMO and Bartlett's test results

Bartlett's and KMO		
Kaiser–Meyer–Olkin Measure of Sampling Adequacy		0.816
Bartlett's Test of Sphericity	Approx. Chi-Square	1371.390
	Df	21
	Sig. (*p*-value)	0.000

Table 6. PCA results: rotated component matrix of each factor loading

Well Logs	1	2
TOC	0.279	0.960
GR	0.847	0.242
MSFL	0.957	0.278
CALI	0.956	0.279
PEF	0.956	0.280
RHOB	0.957	0.278
NPHI	0.957	0.279
Extraction Method: Principal Component Analysis		
Rotation Method: Varimax with Kaiser Normalization		

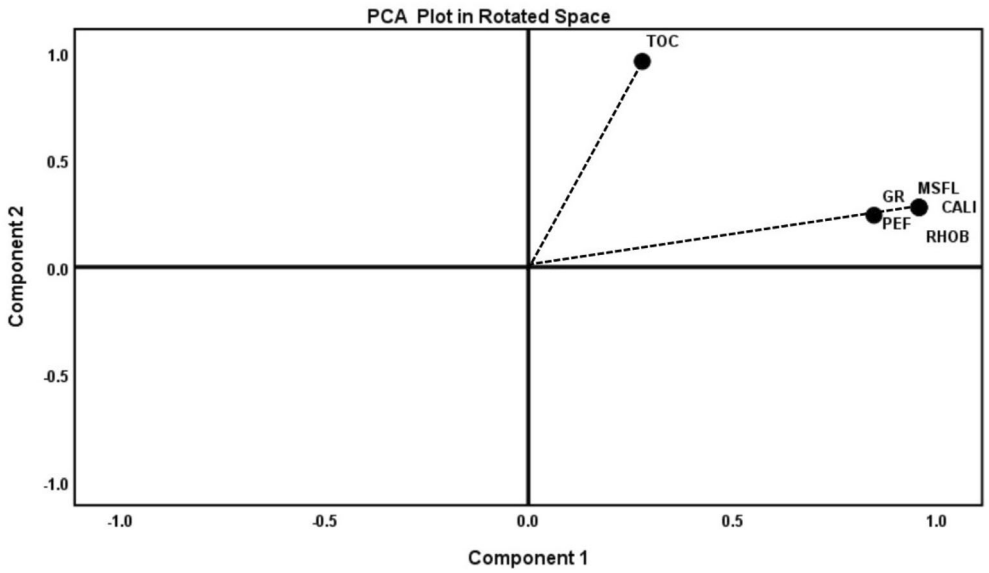

Figure 14. Results of PCA showing the interrelationship of variables in the spaces of components 1 and 2. Component 1 accounts for > 76% and component 2 accounts for 19.56% of the variance in the data.

Moreover, a reliability test was performed for each variable (Table 8) to check if they are reliable to each other or not. Figure 15 shows that all variables were statistically significant to each other, whereby GR, RHOB, CALI, PEF, NPHI, and MSFL showed higher inter-item correlation, and have cumulative factor loading contribution of over 96% at a confidence interval of 95% (Fig. 15, Table 8).

Table 7. PCA results: eigenvalues, and variance and cumulative variance explained

Component	Initial eigenvalues			Extraction sums of squared loadings		Rotation sums of squared loadings	
	Total	% of Variance	Cumulative %	% of Variance	Cumulative %	% of Variance	Cumulative %
1	6.065	86.647	86.647	86.647	86.647	76.718	76.718
2	0.674	9.635	96.283	9.635	96.283	19.564	96.283
3	0.260	3.717	99.999				
4	3.333E-5	0.000	100.000				
5	6.811E-6	9.731E-5	100.000				
6	2.131E-6	3.044E-5	100.000				
7	1.799E-8	2.570E-7	100.000				

Table 8. PCA reliability test results: inter-item correlation, mean, variance of the variables

Items	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha
TOC	0.0000000	34.374	0.536	0.991
GR	1.6736663	29.351	0.835	0.975
MSFL	1.6736663	27.955	0.988	0.963
CALI	1.6736667	27.960	0.987	0.963
PEF	1.6736677	27.958	0.987	0.963
RHOB	1.6736667	27.958	0.987	0.963

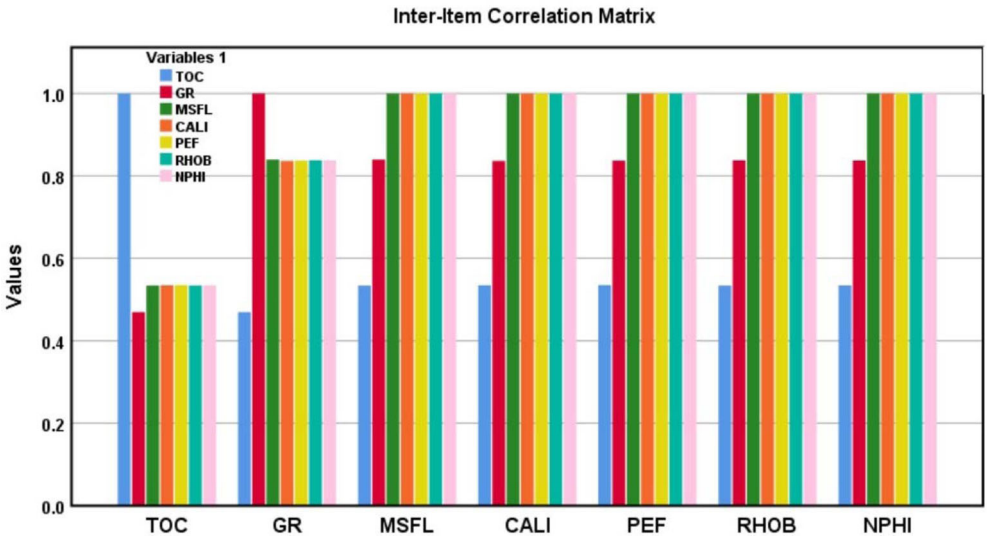

Figure 15. Reliability statistical test showing significance distribution values of each variable informing factor.

Pearson's Correlation Coefficient (r) Analysis

The results (Fig. 16) show that there were high positive correlations between DT and TOC (R^2 = 0.784) (Fig. 16a), NPHI and TOC (R^2 = 0.766) (Fig. 16b), GR and TOC (R^2 = 0.625) (Fig. 16c), MSFL and TOC (R^2 = 0.783) (Fig. 16d), RHOB and TOC (R^2 = 0.784) (Fig. 16e), PEF and TOC (R^2 = 0.7837) (Fig. 16f), respectively. Figure 17 presents the implication of the correlation

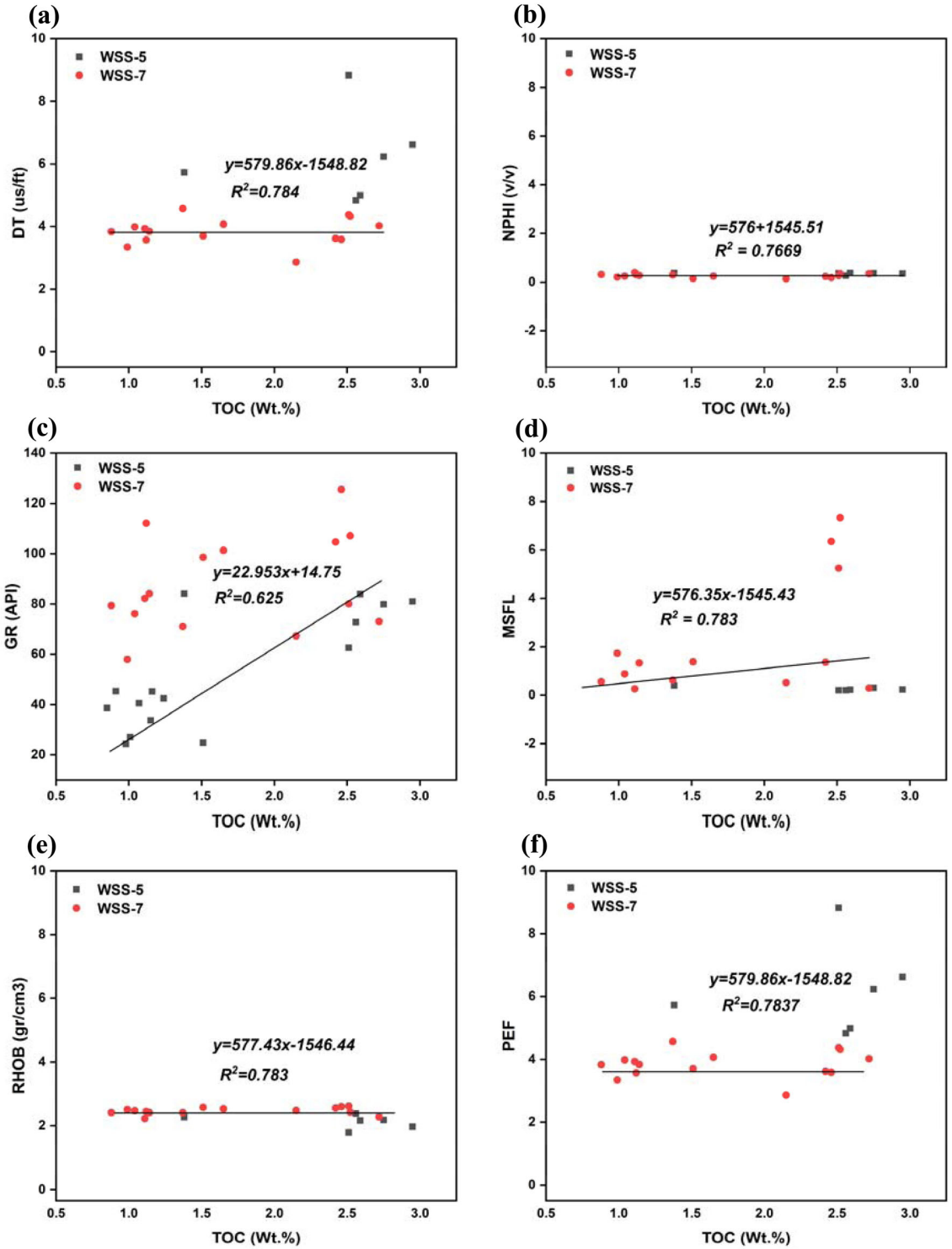

Figure 16. Relationships of TOC with **a** DT, **b** NPHI, **c** GR, **d** MSFL, **e** RHOB, and **f** PEF.

Figure 17. Heatmap of correlations among well log data (PEF, GR, RHOB, MSFL, CALI, and GR).

Table 9. Results of TOC prediction by MVA, SVM, and GPR, with well log data compared to predicted TOC

Sample ID	Depth (m)	TOC Core (wt%)	TOC (wt%) Predicted by MVA	TOC (wt%) Predicted by SVM	TOC (wt%) Predicted by GPR
WSS-5–10	138.67	0.98	1.34	1.47	1.34
WSS-5–12	143.82	1.07	1.34	1.49	1.34
WSS-5–15	157.41	1.15	1.35	1.49	1.37
WSS-5–16	160.82	1.01	1.33	1.48	1.33
WSS-5–17	174.23	1.51	1.48	1.48	1.48
WSS-5–27	191.89	0.85	1.33	1.43	1.33
WSS-5–29	1100.65	0.91	1.33	1.49	1.32
WSS-5–32	1110.3	1.24	1.33	1.68	1.33
WSS-5–34	1120.58	1.16	1.35	1.46	1.35
WSS-5–36	1130.52	2.59	1.87	1.73	1.87
WSS-5–42	1220.45	1.38	1.51	1.56	1.51
WSS-5–57	1240.81	2.75	1.87	1.71	1.87
WSS-5–59	1260.82	2.95	1.87	1.73	1.87
WSS-5–61	1280.82	2.51	1.83	1.73	1.76
WSS-5–66	1302.81	2.56	1.86	1.71	1.72
WSS-7–14	123.22	2.72	1.86	1.67	1.86
WSS-7–18	137.35	1.11	1.41	1.43	1.41
WSS-7–13	153.64	0.88	1.39	1.43	1.38
WSS-7–25	190.71	1.04	1.36	1.49	1.36
WSS-7–26	198.12	0.99	1.39	1.51	1.39
WSS-7–28	1109.22	1.12	1.47	1.44	1.47
WSS-7–32	1120.2	1.14	1.44	1.53	1.44
WSS-7–35	1131.73	1.51	1.48	1.55	1.48
WSS-7–37	1148.57	1.37	1.45	1.56	1.45
WSS-7–40	1210.16	2.42	1.78	1.65	1.79
WSS-7–48	1230.79	1.65	1.76	1.55	1.57
WSS-7–49	1250.23	2.52	1.78	1.76	1.79
WSS-7–50	1270.55	2.51	1.87	1.73	1.87
WSS-7–52	1290.7	2.46	1.76	1.78	1.76
WSS-7–53	1305.81	2.15	1.85	1.71	1.78

among the well log variables based on heat map by Pearson correlation. It can be observed that most of the well log variables have positive correlations with each other, such as NPHI, RHOB, PEF, CALI, and MSFL, although GR has a lower correlation with them possibly due to noise and low number instances within the dataset.

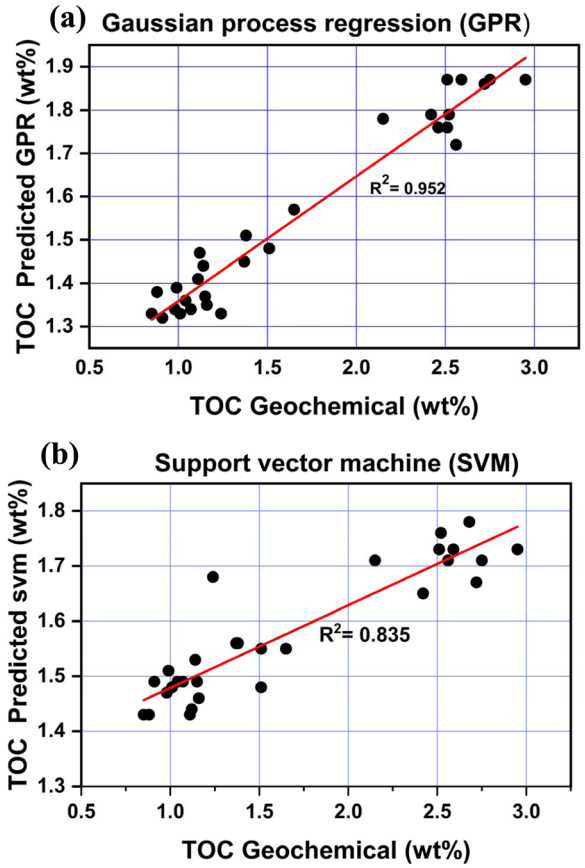

Figure 18. Relationship between TOC values from geochemical analysis and predicted TOC values from GPR and SVM modeling.

Table 10. Statistical prediction of optimized TOC model during training and testing

Model	MSE		RMSE	
	Train	Test	Train	Test
GPR	0.1586	0.3168	0.2805	0.5629
MVA	0.1905	0.3810	0.3087	0.6172
SVM	0.2467	0.4933	0.3512	0.7023

MACHINE LEARNING

As previously discussed by Shalaby et al. (2019), well log data of a reservoir are controlled by the availability level of organic matter. Hence, data for six-well log variables with good responses were selected as inputs and TOC from geochemical data as output. The data were trained in MATLAB R2020a to predict TOC. The method employed a K-fold cross-validation approach, which protects the data against over-fitting. Four algorithms from GRP and six algorithms from SVM were trained. All other algorithms performed worse than coarse Gaussian SVM and square exponential GPR (Table 9). Figure 18 shows the best results from SVM and GPR based on R^2 of predicted TOC values versus geochemical TOC data. GPR gave the best result with $R^2 = 0.952$ due to its capability in obtaining uncertainty of the predicted model and in yielding good output for the predicted values. SVM gave a lower R^2 of 0.835, indicating its complication in prediction.

The performance classification accuracy was assessed based on the optimum model of SVM and GPR through RMSE and MSE by observing the ones that produced the least error. The results (Table 10) show that the optimum GPR model gave RMSE of 0.5629 and MSE of 0.3168 at a prediction speed of 1200 obs/sec under Bayesian optimization) while the optimum SVM gave RMSE of 0.7023 and MSE of 0.4933 at a prediction speed of 1700obs/sec). Comparing the training and test errors (Figs. 19, 20), GPR gave a small MSE value that indicated its robustness more than other models. Both models tend to predict the TOC at a different level of lower error.

Figure 21 shows the predictability of MVA, SVM, and GPR models. The results show that all the models provided good results, which reveal that well log data are necessary important parameters for predicting TOC. However, the optimized GPR gave the best result ($R^2 = 0.952$) followed by MVA ($R^2 = 0.935$) whereas SVM provided the worst result ($R^2 = 0.835$).

CONCLUSIONS

The study proposed the advantage of integrating ML, geochemical, and MVA techniques that lead to improved prediction of TOC and source rock evaluation. Based on geochemical findings the Tanga shell was classified as fair to good source rocks. They have TOC contents ranging from 0.85 to 2.95 (wt%). They contain oil and gas prone characterized by type II together with III Kerogen laying in oil to condensate as mature source rocks. TOC prediction from MVA and ML models suggests a good agreement with geochemical analyzed values at R^2 above 0.8 for all the models applied. The prediction results from MVA and ML reveal that

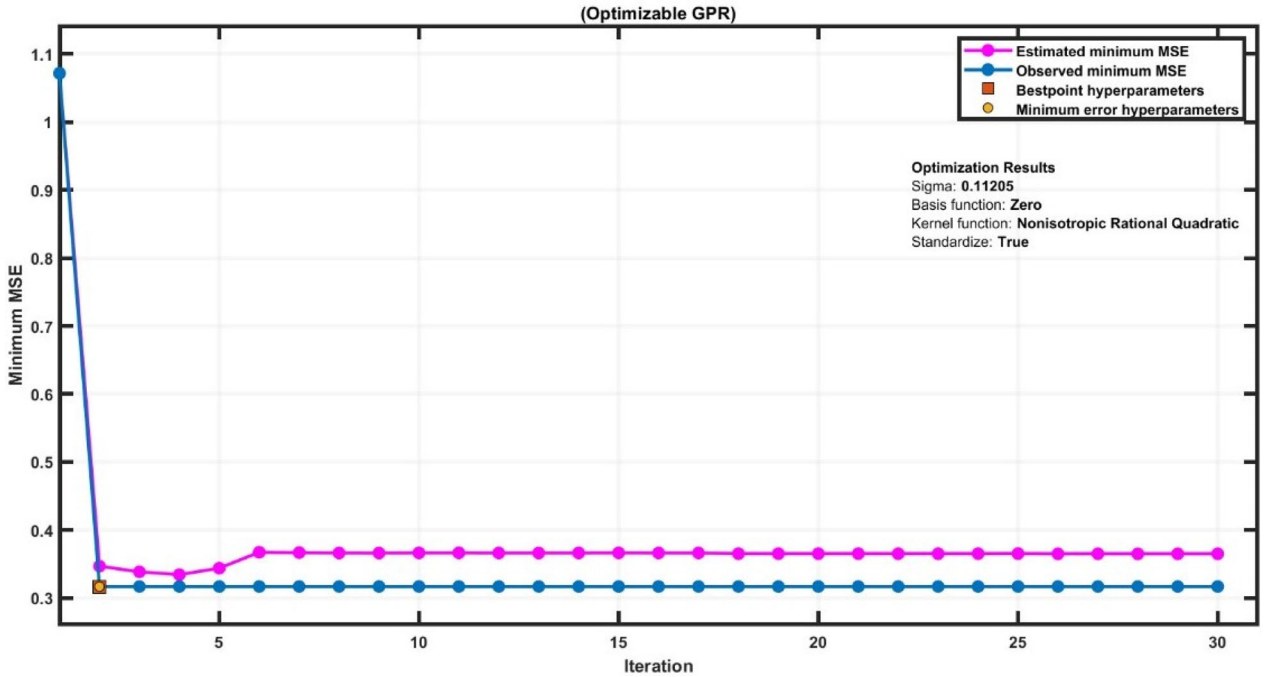

Figure 19. GRP optimized model training (pink curve) and testing (blue curve) for TOC prediction.

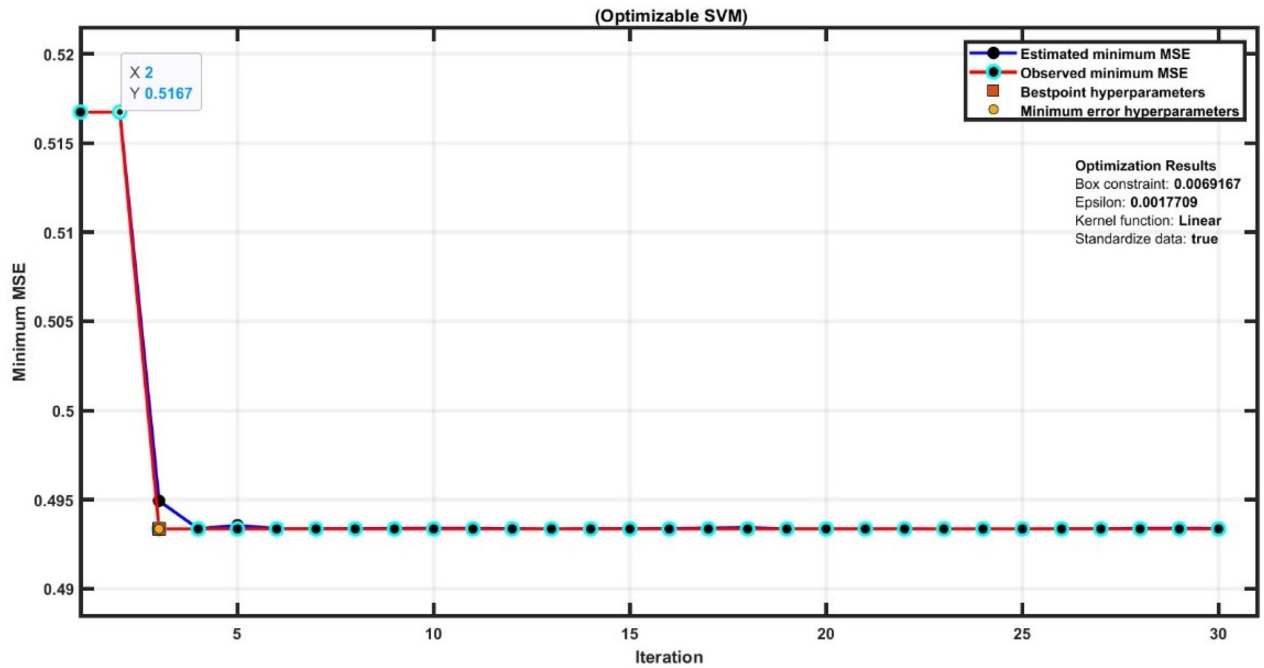

Figure 20. SVM optimized model training (pink curve) and testing (blue curve) for TOC prediction.

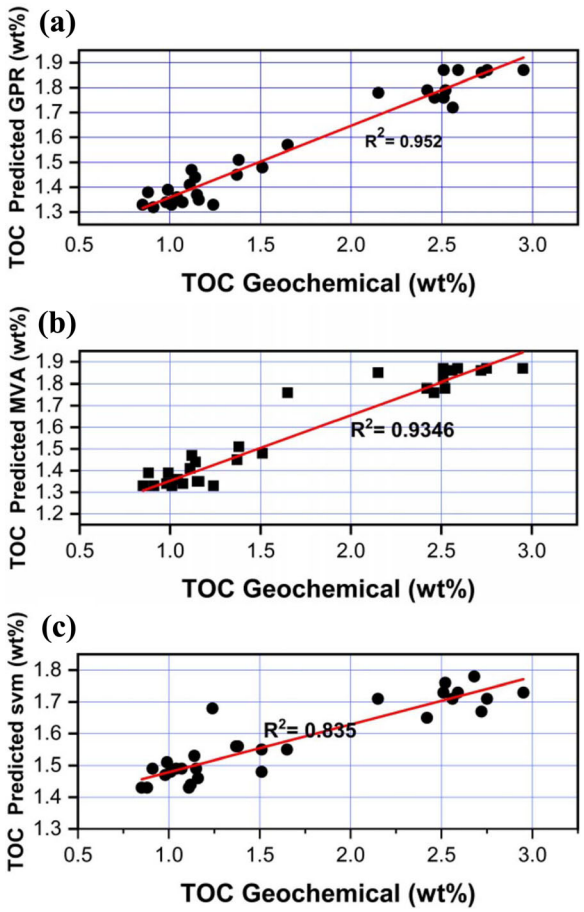

Figure 21. Plots of measured TOC values versus predicted TOC values by GPR, MVA, and SVM modeling.

optimized GPR was the best model having a lower value of RMSE = 0.5629, followed by MVA having an RMSE = 0.6172, and the least performance model was optimized SVM having *RMSE = 0.7023.* Generally, the consensus of geochemical, statistical, and machine learning methods suggests that the combination of this interactive method provides good results and can be applied in source rock evaluation worldwide.

ACKNOWLEDGMENTS

This work was supported by the Major National Science and Technology Programs in the ''Thirteenth Five-Year'' Plan period (No. 2017ZX05032-002-004), National Natural Science Foundation China (grant Nos.41972326 and 51774258), and the Fundamental Research Fund for the Central Universities, China University of Geosciences (Wuhan, No. CUGCJ1820). Department of Petroleum Engineering and International Education College at the China University of Geosciences provided helps for the success of this experimental study.

REFERENCES

Al-Mohair, H. K., Saleh, J. M., & Suandi, S. A. (2015). Hybrid human skin detection using neural network and K-means clustering technique. *Applied Soft Computing*, 33, 337–347.

Alquisom, M. M. (2016). Development of an artificial neural network based expert system to determine the location of horizontal well in a three-phase reservoir with a simultaneous gas cap and bottom water drive.

Amiri Bakhtiar, H., Telmadarreie, A., Shayesteh, M., Heidari Fard, M. H., Talebi, H., & Shirband, Z. (2011). Estimating total organic carbon content and source rock evaluation, applying ΔlogR and neural network methods: Ahwaz and Marun oilfields, SW of Iran. *Petroleum Science and Technology, 29*(16), 1691–1704.

Asante-Okyere, S., Shen, C., Ziggah, Y. Y., Rulegeya, M. M., & Zhu, X. (2020). A novel hybrid technique of integrating gradient-boosted machine and clustering algorithms for lithology classification. *Natural Resources Research, 29*(4), 2257–2273. https://doi.org/10.1007/s11053-019-09576-4.

Azimi-Pour, M., Eskandari-Naddaf, H., & Pakzad, A. (2020). Linear and non-linear SVM prediction for fresh properties and compressive strength of high volume fly ash self-compacting concrete. *Construction and Building Materials, 230*, 117021. https://doi.org/10.1016/j.conbuildmat.2019.117021.

Aziz, H., Ehsan, M., Ali, A., Khan, H. K., & Khan, A. (2020). Hydrocarbon source rock evaluation and quantification of organic richness from correlation of well logs and geochemical data: A case study from the sembar formation, Southern Indus Basin, Pakistan. *Journal of Natural Gas Science and Engineering, 81*, 103433. https://doi.org/10.1016/j.jngse.2020.103433.

Behar, F., Beaumont, V., & Penteado, H. L. D. B. (2001). Rock-Eval 6 technology: Performances and developments. *Oil & Gas Science and Technology, 56*(2), 111–134.

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, 151*, 224–234.

Bramer, M. (2016). Data for data mining. *Principles of data mining* pp. 9–19. Springer.

Carvajal-Ortiz, H., & Gentzis, T. (2015). Critical considerations when assessing hydrocarbon plays using Rock-Eval pyrolysis and organic petrology data: Data quality revisited. *International Journal of Coal Geology, 152*, 113–122.

Dembicki, H., Jr. (2009). Three common source rock evaluation errors made by geologists during prospect or play appraisals. *AAPG Bulletin, 93*(3), 341–356.

Edwards, D. S., Struckmeyer, H. I. M., Bradshaw, M. T., & Skinner, J. E. (1999). Geochemical characteristics of Australia's southern margin petroleum systems. *The APPEA Journal, 39*(1), 297–321.

El Hajj, L., Baudin, F., Littke, R., Nader, F. H., Geze, R., Maksoud, S., & Azar, D. (2019). Geochemical and petrographic analyses of new petroleum source rocks from the onshore upper jurassic and lower cretaceous of Lebanon. *International Journal of Coal Geology, 204*, 70–84.

El Kammar, M. M. (2015). Source-rock evaluation of the Dakhla Formation black shale in Gebel Duwi, Quseir area Egypt. *Journal of African Earth Sciences, 104*, 19–26.

El Nady, M. M., Lotfy, N. M., Ramadan, F. S., & Hammad, M. M. (2015a). Evaluation of organic matters, hydrocarbon potential and thermal maturity of source rocks based on geochemical and statistical methods: Case study of source rocks in Ras Gharib oilfield, central Gulf of Suez Egypt. *Egyptian Journal of Petroleum, 24*(2), 203–211. https://doi.org/10.1016/j.ejpe.2015.05.012.

El Nady, M. M., Ramadan, F. S., Hammad, M. M., & Lotfy, N. M. (2015b). Evaluation of organic matters, hydrocarbon potential and thermal maturity of source rocks based on geochemical and statistical methods: Case study of source rocks in Ras Gharib oilfield, central Gulf of Suez Egypt. *Egyptian Journal of Petroleum, 24*(2), 203–211.

Gentzis, T. (2018). International Journal of Coal Geology Geochemical screening of source rocks and reservoirs: The importance of using the proper analytical program. *International Journal of Coal Geology, 190*, 56–69. https://doi.org/10.1016/j.coal.2017.11.014.

Giannakopoulou, P., Petrounias, P., Tsikouras, B., Kalaitzidis, S., Rogkala, A., Hatzipanagiotou, K., & Tombros, S. (2018). Using factor analysis to determine the interrelationships between the engineering properties of aggregates from igneous rocks in Greece. *Minerals, 8*(12), 580.

Godfray, G., & Seetharamaiah, J. (2019). Geochemical and well logs evaluation of the Triassic source rocks of the Mandawa basin, SE Tanzania: Implication on richness and hydrocarbon generation potential. *Journal of African Earth Sciences, 153*, 9–16.

Golden, C. E., Rothrock, M. J., Jr., & Mishra, A. (2019). Comparison between random forest and gradient boosting machine methods for predicting Listeria spp. prevalence in the environment of pastured poultry farms. *Food Research International, 122*, 47–55.

Hakimi, M. H., Abdullah, W. H., & Ahmed, A. F. (2017). Organic geochemical characteristics of oils from the offshore Jiza-Qamar Basin, Eastern Yemen: New insight on coal/coaly shale source rocks. *Journal of Petroleum Science and Engineering, 153*, 23–35.

Handhal, A. M., Al-Abadi, A. M., Chafeet, H. E., & Ismail, M. J. (2020). Prediction of total organic carbon at Rumaila oil field, Southern Iraq using conventional well logs and machine learning algorithms. *Marine and Petroleum Geology, 116*, 104347. https://doi.org/10.1016/j.marpetgeo.2020.104347.

Hazra, B., Dutta, S., & Kumar, S. (2017). TOC calculation of organic matter rich sediments using Rock-Eval pyrolysis: Critical consideration and insights. *International Journal of Coal Geology, 169*, 106–115. https://doi.org/10.1016/j.coal.2016.11.012.

Izenman, A. J. (2008). Modern multivariate statistical techniques. *Regression, Classification and Manifold Learning, 10*, 970–978.

Johnson, R. A., & Wichern, D. W. (2002). Applied multivariate statistical analysis (Vol. 5, Issue 8). Prentice hall Upper Saddle River, NJ.

Kaloop, M. R., Kumar, D., Samui, P., Hu, J. W., & Kim, D. (2020). Compressive strength prediction of high-performance concrete using gradient tree boosting machine. *Construction and Building Materials, 264*, 120198.

Kapilima, S. (2003). Tectonic and sedimentary evolution of the coastal basin of Tanzania during the Mesozoic times. *Tanzania Journal of Science, 29*(1), 1–16.

Lafargue, E., Marquis, F., & Pillot, D. (1998). Rock-Eval 6 applications in hydrocarbon exploration, production, and soil contamination studies. *Revue De L'institut Français Du Pétrole, 53*(4), 421–437.

Landau S.E., & Brian, S. (2004). Handbook of statistical analyses using SPSS program..

Langford, F. F., & Blanc-Valleron, M.-M. (1990). Interpreting Rock-Eval pyrolysis data using graphs of pyrolizable hydrocarbons vs total organic carbon (1). *AAPG Bulletin, 74*(6), 799–804.

Li, M., Chen, Z., Cao, T., Ma, X., Liu, X., & Li, Z. (2018). International journal of coal geology expelled oils and their impacts on rock-eval data interpretation, Eocene Qianjiang formation in Jianghan Basin, China. *International Journal of Coal Geology, 191*, 37–48. https://doi.org/10.1016/j.coal.2018.03.001.

Mahmoud, A. A., Elkatatny, S., Ali, A., Abouelresh, M., & Abdulraheem, A. (2019). New robust model to evaluate the total organic carbon using fuzzy logic. *SPE Kuwait Oil & Gas Show and Conference*.

Mahmoud, A. A., Elkatatny, S., Ali, A., Abdulraheem, A., & Abouelresh, M. (2020). Estimation of the total organic carbon using functional neural networks and support vector machine. *International Petroleum Technology Conference 2020, IPTC 2020*. https://doi.org/10.2523/iptc-19659-ms.

Mashhadi, Z. S., & Rabbani, A. R. (2015). International journal of coal geology organic geochemistry of crude oils and cretaceous source rocks in the iranian sector of the Persian Gulf: An oil – oil and oil – source rock correlation study. *International Journal of Coal Geology, 146*, 118–144. https://doi.org/10.1016/j.coal.2015.05.003.

Mbede, E. I., & Dualeh, A. (1997). The coastal basins of Somalia, Kenya and Tanzania. In *Sedimentary Basins of the World* (Vol. 3, pp. 211–233). Elsevier.

Mulashani, A. K., Shen, C., Asante-okyere, S., Kerttu, P. N., & Abelly, E. N. (2021a). Group Method of Data Handling (GMDH) Neural Network for Estimating Total Organic Carbon (TOC) and hydrocar- bon potential distribution (S 1, S 2) using well logs. *Natural Resources Research*. https://doi.org/10.1007/s11053-021-09908-3.

Mulashani, A. K., Shen, C., Nkurlu, B. M., Mkono, C. N., & Kawamala, M. (2021b). Enhanced group method of data handling (GMDH) for permeability prediction based on the modified Levenberg Marquardt technique from well log data. *Energy*, 121915. https://doi.org/10.1016/j.energy.2021.121915.

Omran, A. A., & Alareeq, N. M. (2018). Joint geophysical and geochemical evaluation of source rocks–A case study in Sayun-Masila basin Yemen. *Egyptian Journal of Petroleum, 27*(4), 997–1012.

Pan, S., Horsfield, B., Zou, C., Yang, Z., & Gao, D. (2017). Statistical analysis as a tool for assisting geochemical interpretation of the upper triassic yanchang formation, Ordos Basin, Central China. *International Journal of Coal Geology, 173*, 51–64.

Peters, K. E. (1986). Guidelines for evaluating petroleum source rock using programmed pyrolysis. *AAPG Bulletin, 70*(3), 318–329.

Priddy, K. L., & Keller, P. E. (2005). *Artificial neural networks: an introduction* (Vol. 68). SPIE press.

Robison, C. R. (1997). Hydrocarbon source rock variability within the Austin chalk and Eagle Ford shale (upper cretaceous), East Texas, USA. *International Journal of Coal Geology, 34*(3–4), 287–305.

Romero-Sarmiento, M.-F., Euzen, T., Rohais, S., Jiang, C., & Littke, R. (2016). Artificial thermal maturation of source rocks at different thermal maturity levels: Application to the Triassic Montney and Doig formations in the Western Canada Sedimentary Basin. *Organic Geochemistry, 97*, 148–162.

Romero-sarmiento, M., Ramiro-ramirez, S., Berthe, G., Fleury, M., & Littke, R. (2017). International journal of coal geology geochemical and petrophysical source rock characterization of the Vaca Muerta formation, Argentina: Implications for unconventional petroleum resource estimations. *International Journal of Coal Geology, 184*, 27–41. https://doi.org/10.1016/j.coal.2017.11.004.

Rui, J., Zhang, H., Ren, Q., Yan, L., Guo, Q., & Zhang, D. (2020). TOC content prediction based on a combined Gaussian process regression model. *Marine and Petroleum Geology, 118*, 104429. https://doi.org/10.1016/j.marpetgeo.2020.104429.

Said, A., Moder, C., Clark, S., & Abdelmalak, M. M. (2015). Sedimentary budgets of the Tanzania coastal basin and implications for uplift history of the East African rift system. *Journal of African Earth Sciences, 111*, 288–295.

Shalaby, M. R., Jumat, N., Lai, D., & Malik, O. (2019). Integrated TOC prediction and source rock characterization using machine learning, well logs and geochemical analysis: Case study from the Jurassic source rocks in Shams Field, NW Desert Egypt. *Journal of Petroleum Science and Engineering, 176*, 369–380. https://doi.org/10.1016/j.petrol.2019.01.055.

Shen, C., Asante-Okyere, S., Yevenyo Ziggah, Y., Wang, L., & Zhu, X. (2019). Group method of data handling (GMDH) lithology identification based on wavelet analysis and dimensionality reduction as well log data pre-processing techniques. *Energies, 12*(8), 1509.

Suzuki, K. (2011). *Artificial neural networks: methodological advances and biomedical applications*. BoD–Books on Demand.

Temple, J. T. (1978). The use of factor analysis in geology. *Journal of International Association of Mathematics and Geology, 10*, 379–387.

Walden, J., Smith, J. P., & Dackombe, R. V. (1992). The use of simultaneous R-and Q-mode factor analysis as a tool for assisting interpretation of mineral magnetic data. *Mathematical Geology, 24*(3), 227–247.

Wang, J., Gao, Z., Kang, Z., Zhu, D., Liu, Q., Ding, Q., & Liu, Z. (2020). Geochemical characteristics, hydrocarbon potential and depositional environment of the Yangye Formation source rocks in Kashi sag, southwestern Tarim Basin, NW China. *Marine and Petroleum Geology, 112*, 104084.

Wopfner, H. (2002). Tectonic and climatic events controlling deposition in Tanzanian Karoo basins. *Journal of African Earth Sciences, 34*(3–4), 167–177.

Wu, G., Lü, Z. T., & Wu, Z. S. (2006). Strength and ductility of concrete cylinders confined with FRP composites. *Construction and Building Materials, 20*(3), 134–148.

Wu, X., Chen, Y., Zhao, G., Du, X., Zeng, H., Wang, P., Wang, Y., & Hu, Y. (2017). Evaluation of source rocks in the 5th member of the Upper Triassic Xujiahe formation in the Xinchang gas field, the Western Sichuan depression China. *Journal of Natural Gas Geoscience, 2*(4), 253–262.

Xie, Y., Zhu, C., Zhou, W., Li, Z., Liu, X., & Tu, M. (2018). Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. *Journal of Petroleum Science and Engineering, 160*, 182–193.

Zaremotlagh, S., Hezarkhani, A., & Sadeghi, M. (2016). Detecting homogenous clusters using whole-rock chemical compositions and REE patterns: A graph-based geochemical approach. *Journal of Geochemical Exploration, 170*, 94–106. https://doi.org/10.1016/j.gexplo.2016.08.017.

Zhou, D., Chang, T., & Davis, J. C. (1983). Dual extraction ofR-mode andQ-mode factor solutions. *Journal of the International Association for Mathematical Geology, 15*(5), 581–606.

Zumberge, J. E. (1987). Prediction of source rock characteristics based on terpane biomarkers in crude oils: A multivariate statistical approach. *Geochimica Et Cosmochimica Acta, 51*(6), 1625–1637.