



Original Paper

Group Method of Data Handling (GMDH) Neural Network for Estimating Total Organic Carbon (TOC) and Hydrocarbon Potential Distribution (S_1 , S_2) Using Well Logs

Alvin K. Mulashani,^{1,2} Chuanbo Shen,^{1,5} Solomon Asante-Okyere,^{1,3} Paulus N. Kerttu,⁴ and Elieneza N. Abelly¹

Received 10 February 2021; accepted 14 June 2021
Published online: 26 June 2021

Recent advancement in computing capabilities has brought to light the application of machine learning methods in estimating geochemical data from well logs. The widely employed artificial neural network (ANN) has intrinsic problems in its application. Therefore, the objective of this study was to present a group method of data handling (GMDH) neural network as an improved alternative in predicting total organic carbon (TOC), S_1 , and S_2 from well logs. The study used bulk density, sonic travel time, deep lateral resistivity log, gamma-ray, spontaneous potential, neutron porosity well logs as input variables to predict TOC, S_1 , and S_2 of the Nondwa, Mbuo, and Mihambia Formations in the Triassic to mid-Jurassic of the Mandawa Basin in southeast Tanzania. The TOC prediction results indicated that the GMDH model trained well while generalizing better across the testing data than both ANN and $\Delta\log R$. Specifically, the GMDH provided TOC testing predictions having the least errors of 0.40 and 0.45 for mean square error (MSE) and mean absolute error (MAE), respectively, as compared to 1.27 and 0.81, 0.68 and 0.7, 1.4 and 0.89 obtained by back-propagation neural network (BPNN), radial basis function neural network (RBFNN), and $\Delta\log R$, respectively. For S_1 and S_2 , the ANN models performed excellently during training but were unable to produce similar results when tested on the completely unseen well data. This represents a clear case of over-fitting by ANN. During testing, the GMDH avoided over-fitting and outperformed ANN by obtaining the least MSE of 0.04 and 1.16 and MAE of 0.07 for S_1 and S_2 , respectively, while BPNN achieved MSE and MAE of 0.08 and 0.17 for S_1 , 1.96, and 0.9 for S_2 , and RBFNN obtained MSE and MAE of 0.15 and 0.25 for S_1 and 1.4

¹Key Laboratory of Tectonics and Petroleum Resources, Ministry of Education, China University of Geosciences, Wuhan 430074, China.

²Department of Geoscience and Mining Technology, College of Engineering and Technology, Mbeya University of Science and Technology, Box 131 Mbeya, Tanzania.

³Department of Petroleum and Natural Gas Engineering, School of Petroleum Studies, University of Mines and Technology, Tarkwa, Ghana.

⁴Faculty of Geophysics and Geomatics, China University of Geosciences, Wuhan 430074, China.

⁵To whom correspondence should be addressed; e-mail: cb-shen@cug.edu.cn

and 0.87 for S_2 . Hence, the improved generalization performance of the GMDH makes it an improved form of a neural network for TOC , S_1 , and S_2 prediction. The proposed model was further adopted to predict the geochemical data and determine the source rock quality for the East Lika-1 well, which has no core data.

KEY WORDS: Group method of data handling, Total organic carbon, Hydrocarbon potential distribution, Artificial neural network.

INTRODUCTION

Source rock reservoirs are rapidly growing into a major energy resource player as conventional hydrocarbon fields across the world continue to diminish. To determine accurately the potential of an unconventional shale reservoir requires in-depth knowledge of the distribution of the organic geochemical properties. Generally, the geochemical parameters provide an understanding of the organic matter content, type, and thermal distribution of the source rock (Cappuccio et al., 2020; Curiale & Curtis, 2016; Liu et al., 2020). The fundamental organic geochemical parameters include total organic carbon (TOC), which describes the amount of organic matter. S_1 represents the amount of free hydrocarbon present in the rock sample before performing the pyrolysis process while S_2 is the hydrocarbon formed during the pyrolysis of the sample. The S_2 values give a general indication of the prevailing hydrocarbon generating potential. Rock-Eval pyrolysis is the most commonly used pyrolysis method for measuring the quantity of emitted carbon dioxide and hydrocarbons (Carvajal-Ortiz & Gentzis, 2015; Hakimi et al., 2020; Mani et al., 2017; Shalaby et al., 2019).

The most reliable method of quantifying the source rock is by performing organic geochemical analysis on core samples in the laboratory. However, coring is an expensive exercise to be conducted on all wells (Evenick, 2020; Mahmoud et al., 2017). In cases where there may not be enough core data, readily available drill cuttings are commonly used to compensate. The challenges of using drill cuttings are their difficulty for reconciling with depth and they can be contaminated (Bai & Tan, 2020). Concerning this, attempts have been made to generate TOC values from geophysical well logs based on the knowledge that well log parameters can detect the presence of organic matter. The two frequently employed conventional techniques are the Schmoker and $\Delta\log R$ models. The Schmoker model esti-

mates TOC using the reciprocal of bulk density. However, the Schmoker approach is influenced heavily by the reservoir or geological characteristics (Schmoker, 1979; Mahmoud et al., 2019; Xiong et al., 2019). The widely adopted conventional method is the $\Delta\log R$ method proposed by Passey et al. (1990) using a porosity and resistivity log (Tenaglia et al., 2020). It is well documented that the shortcoming of the $\Delta\log R$ is the varying log-baseline in different wells, formations, and depositional environments (Charsky & Herron, 2013; Mahmoud et al., 2020). Zhu et al. (2019b) improved the classical $\Delta\log R$ method by incorporating shale mineralogy to develop the dual difference $\Delta\log R$ ($DD\Delta\log R$). The challenge with the implementation of the $DD\Delta\log R$ is that it requires the mineral composition of shale rock samples.

The successful application of artificial intelligence in hydrocarbon exploration and production in recent years has seen the adoption of machine learning models in predicting TOC from well log data. The advantage of machine learning models is their ability to learn and adapt to the dynamics of reservoir conditions such as formation and depositional environment while making use of the entire suite of well logs for a better TOC prediction (Mahmoud et al., 2020; Tariq et al., 2020; Wang et al., 2019a, 2019b; Zhu et al., 2019a). Artificial neural network (ANN) has been the predominantly used machine learning method to predict geochemical data in studies such as (Amiri Bakhtiar et al., 2011; Mahmoud et al., 2020; Shalaby et al., 2020). From these researches, ANN outperformed conventional methods like $\Delta\log R$ due to its ability to map out patterns within the suite of input well logs and geochemical data. However, ANN suffers from intrinsic shortcomings such as overfitting and low convergence speed due to the constant manual tuning of model parameters like the number of hidden nodes, weights, and biases (Asante-Okyere et al., 2020; Bai & Tan, 2020; Zhu et al., 2018, 2019a). Several researches have proposed new con-

cepts and improved machine learning algorithms as an alternative to the standard ANN. Gaussian process regression (GPR) has been implemented for the prediction of *TOC* by Yu et al. (2017) and Rui et al. (2020). However, GPR requires the user to specify the best kernel function to achieve the optimal prediction results. Support vector machine (SVM) and extreme learning machine (ELM) as an improved machine learning techniques have been proposed in determining *TOC* values in shale reservoirs (Shi et al., 2016; Tan et al., 2015; Wang et al., 2018). Similar to GPR, these machine learning models require an iterative tuning of training parameters to obtain the best performance. In addition to the aforementioned machine learning methods, deep learning techniques have been proposed to improve the evaluation of *TOC* (Wang et al., 2019a, 2019b; Zhu et al., 2019a, 2020).

Based on this, we examined for the first time the applicability of the group method of data handling (GMDH) as an improved neural network model in estimating *TOC*, S_1 , and S_2 data from geophysical well logs. The GMDH neural network possesses a self-organizing nature to automatically tune model parameters and generate the optimal model structure during training. Unlike other machine learning models trained to predict *TOC*, the GMDH neural network does not require a manual adjustment of learning parameters to generate the best outcome as summarized in Table 1. A polynomial function of relevant input variables is also generated by GMDH neural network. To ascertain the performance of the proposed GMDH neural network in predicting *TOC*, S_1 , and S_2 , its results were fairly compared with ANN algorithms of backpropagation, radial basis function, and Passey's conventional method of $\Delta \log R$.

GEOLOGICAL SETTING AND DATA DESCRIPTIONS

Geological Setting

The Mandawa Basin is a rift basin situated within the coastal belt of Tanzania. It extends from the border of Kenya in the north to the border of Mozambique in the south. The basin occupies an area of 16,000 km² bordered by the Indian ocean with offshore basins in the east and bounded by the basement of metamorphic terrain on the west side (Caracciolo et al., 2020; Hudson & Nicholas, 2014;

Kagya, 1996). It circumscribes the Ruvuma Saddle to the south, which separates the Ruvuma basin from the Mandawa basin and by the Rufiji trough to the north (Fig. 1a). Karoo rifting (Permian–Triassic) was the major tectonic event that dominated the development of the Mandawa Basin. The depositional origin and the Mandawa Basin history are strongly influenced by the Gondwana break-up (Delvaux, 2001; Hou, 2015; Nerbråten, 2014).

The Mandawa Basin is divided into the Kilwa, Mandawa, Pindirolu, Mavuji and Songo Songo Groups (Fig. 2). During Early Cretaceous to Late Jurassic, the basin was subsiding at a higher rate in which clastic sediments were major deposits. The prograding of the Mavuji groups and Mandawa groups were then deposited in alluvial and river deposits (Emanuel et al., 2020; Fossum et al., 2020). The coastal Mandawa Basin, under the outer-mid shelf environment, subsided at a constant stage rate from Bajocian to Priabonian–Paleogene, which resulted in the deposition of the Kilwa Group. The Kilwa Group comprises the Kivinje, Nangurukuru, Masoko, and Pande Formations. The Mavuji Group consists of the Makonde, Kitiruka, and Kihuluhulu Formations, which are three time-equivalent formations (Emanuel et al., 2020; Hou, 2015; Hudson, 2011).

The Nondwa Formation in an unconformable manner overlays the deltaic-fluvial from Upper Triassic to Lower Jurassic Mihambia and Mbuo Formations, which underlie the middle Jurassic Mtumbei Formation separated by the unconformity stage of the Aelenian-Bajocian. The basin was occupied with shoreface to offshore fine-grained, siltstones, mudstones and, sandstones of the Mbuo Formation (Einvik-Heitmann et al., 2015; Smelror et al., 2018). The study area consists of four exploration wells, which are Mbate-1, Mbuo-1, Mita Gamma-1, and East Lika-1 (Fig. 1b). It is important to remember that Mita Gamma-1 intersected both the Nondwa Formation and silicic limestones of the Mihambia Formation; Mbate-1 well is found in the Kivinje Formation while Mbuo-1 well intersected both the Nondwa and Mbuo Formations and, lastly, East Lika-1 intersected the Mbuo and Nondwa sequences.

Data Descriptions

Conventional well log suite of bulk density log (RHOB), sonic travel time (DT), deep lateral

Table 1. Advantages and limitations of conventional and machine learning models previously used to estimate *TOC*

Model	Merits	Limitations
Schmoker's model	It is easy to implement because it only requires the density log	The results are heavily influenced by the type of geological formation or reservoir
$\Delta\log R$	It represents the widely used computational model for estimating <i>TOC</i> from well logs	Baselines used for estimation can vary in different wells, formation, and depositional environment
$DD\Delta\log R$	An improved form of the $\Delta\log R$ proposed by Zhu et al. (2019b) takes into consideration large lateral changes in mineral composition and wellbore size	This method is dependent on the mineral content of the shale rock sample
ANN	ANN is a better <i>TOC</i> predictive model than the $\Delta\log R$ approach (Amiri Bakhtiar et al., 2011; Mahmoud et al., 2020; Shalaby et al., 2020)	It requires a constant model parameter adjustment in order to achieve the optimal results
SVM	SVM proved to be a better predictive model than the ANN (Tan et al., 2015; Wang et al., 2018)	It requires the user to manually adjust the model training parameters to obtain the best results
GPR	GPR outperformed ANN when estimating <i>TOC</i> (Rui et al., 2020; Yu et al., 2017)	It involves user-defined parameters in the form of kernel type before the best predictions can be known
ELM	ELM is an improved form of neural network that generates better results than ANN (Shi et al., 2016)	However, it needs a manual tuning of training parameters to obtain the best model structure

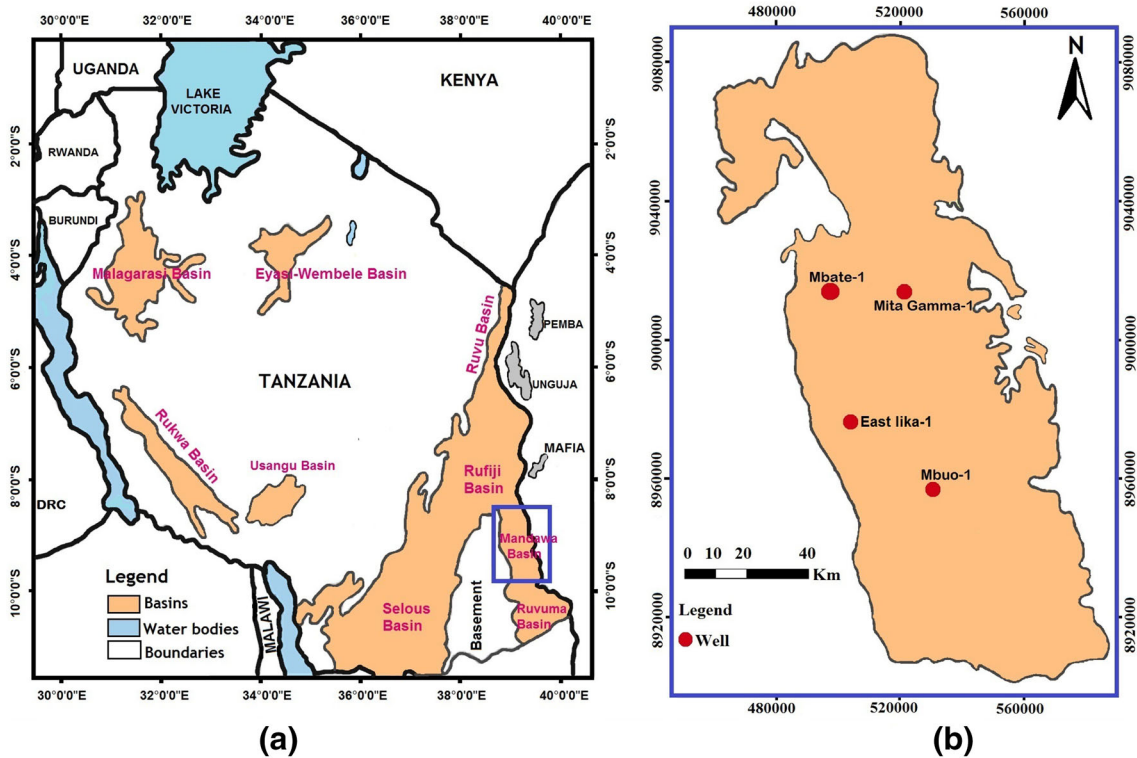


Figure 1. a Location of the Mandawa Basin. b Locations of exploration wells used in this study (Hudson, 2011).

resistivity log (LLD), gamma-ray (GR), spontaneous potential log (SP), neutron porosity (NPHI), and measured geochemical results of *TOC*, *S*₁ and *S*₂ values obtained from Mandawa Basin were used in

this study. Three wells namely Mbate-1, Mbuo-1, and Mita Gamma-1, which have a complete set of well log suites and core *TOC*, *S*₁, and *S*₂ data, were employed to develop the machine learning models.

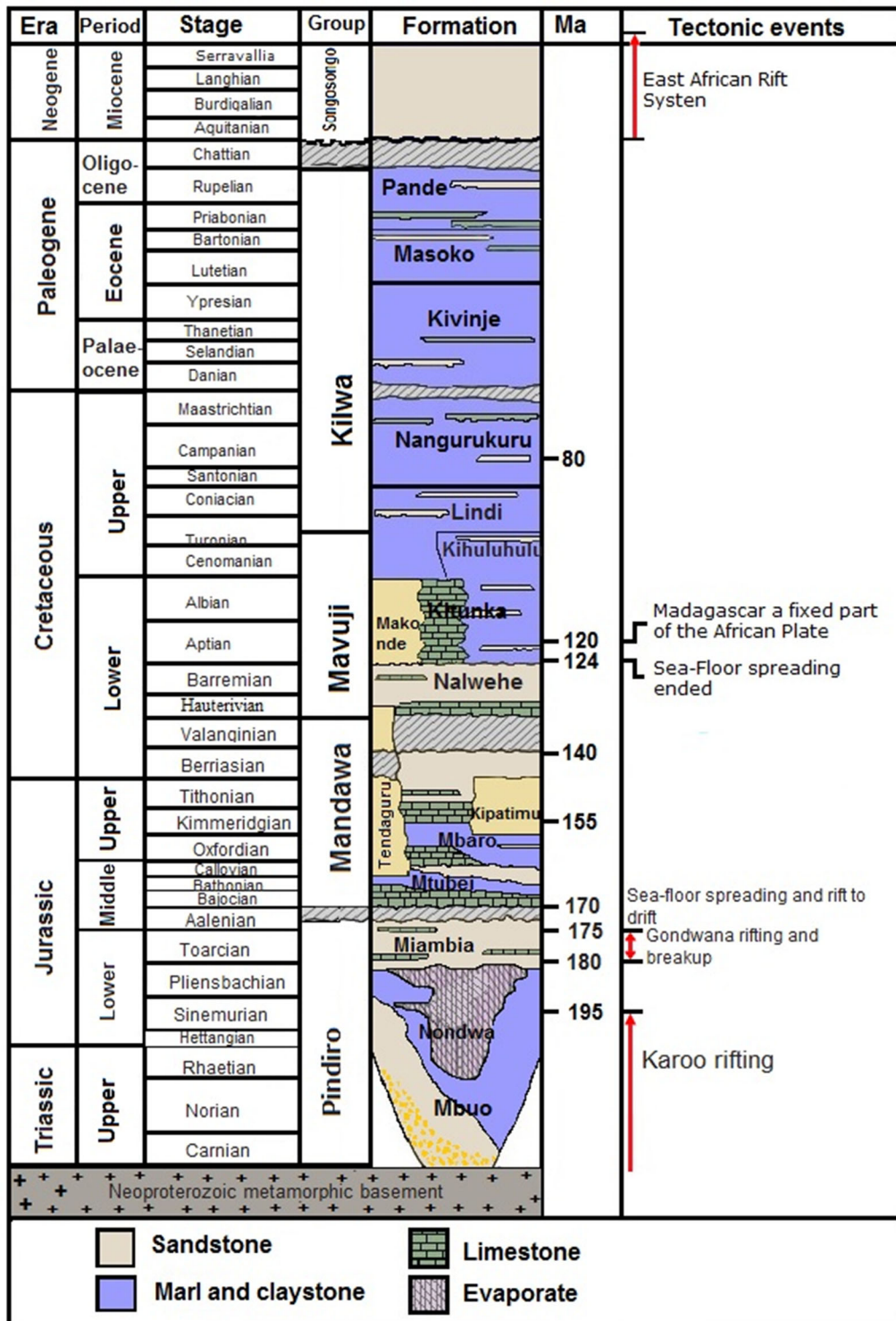


Figure 2. The Mandawa Basin stratigraphical chart (from Fossum et al., 2019).

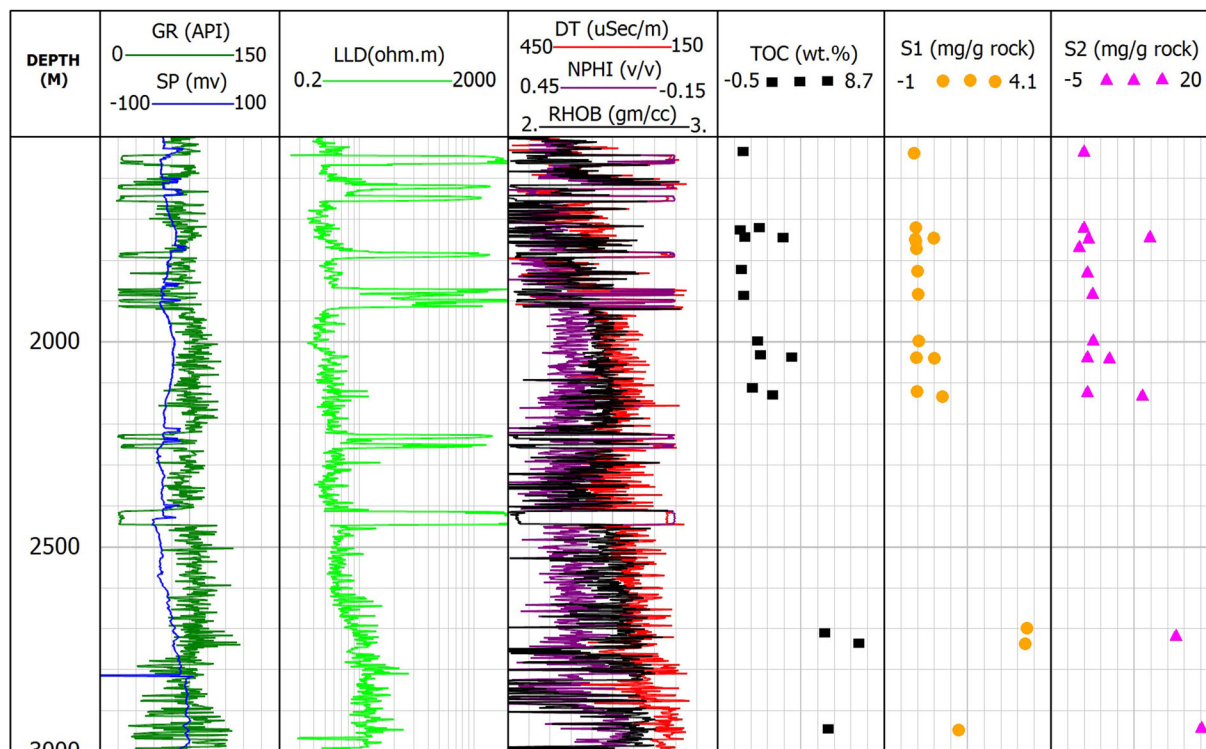


Figure 3. Well log suite and measured TOC , S_1 and, S_2 for Mbuo-1 well used for training.

Mbate-1 and Mbuo-1 wells consisted of 33 sample data for TOC , and 25 sample data for both S_1 and S_2 , which were used to train the models. The developed models were tested on the Mita Gamma-1 well, which consisted of 20 sample data of TOC , S_1 , and S_2 . The well log suite and geochemical results of Mbate-1, Mbuo-1, and Mita Gamma-1 wells are illustrated in Figures 3, 4 and 5. The input data of well logs and output data of TOC , S_1 , and S_2 were normalized within the range of [0, 1] to avoid a case of bias treatment during the model development.

METHODS

$\Delta \log R$ Technique

Passey et al. (1990) introduced the method known as $\Delta \log R$ for obtaining TOC from wireline data. The proper scaling of both the sonic transit time log and resistivity curve is a crucial initial step in the application of this technique, utilizing gamma-ray logs, fine-grained, organic-poor, non-source rock intervals to define baselines for the sonic and resis-

tivity logs. Except in the case of an organic-rich source or a hydrocarbon reservoir interval, these two curves should be parallel and should be overlap (Passey et al., 1990, 2010). The curve of resistivity will react and respond to fluid formation and the sonic curve will react to the presence of kerogens of low velocity/density. At each depth, the corresponding separation of the two curves, i.e., $\Delta \log R$ [Eq. (1)], can be estimated and used to evaluate organic-rich intervals. As long as the quantity of organic maturity (LOM) of an interval is defined or can be measured, the value of $\Delta \log R$ at a specified depth can now be explicitly used in the TOC estimation [Eq. (2)]. The LOM is a typical numerical scale that reflects a whole petroleum generation thermal spectrum (Hood et al., 1975).

$$\Delta \log R = \log_{10} \left(\frac{R}{R_{\text{baseline}}} \right) + n * (\Delta t + \Delta t_{\text{baseline}}) \quad (1)$$

$$TOC = (\Delta \log R) * 10^{(2.2970 - 0.1688 * LOM)} \quad (2)$$

where the curve separation expressed in logarithmic cycles of resistivity is $\Delta \log R$, R is the resistivity

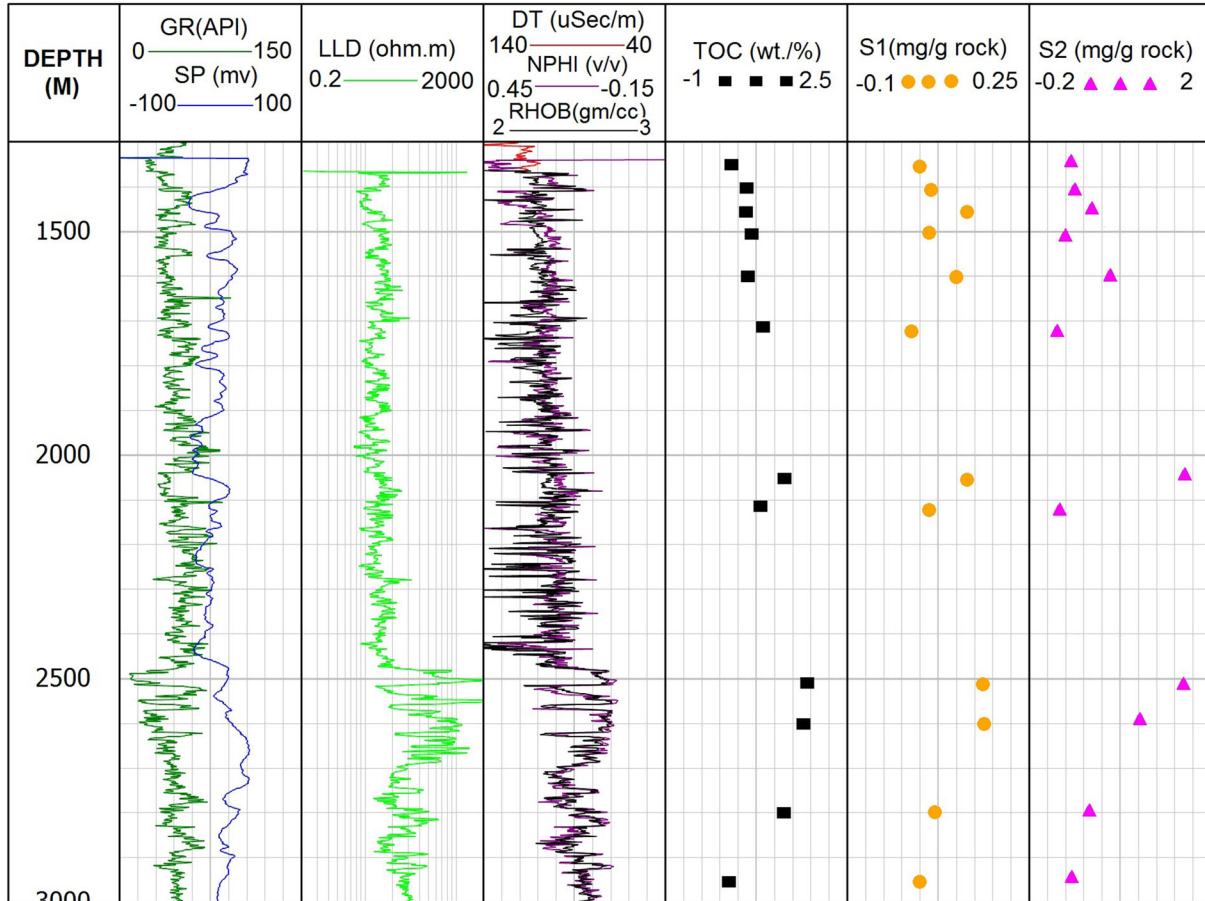


Figure 4. Well log suite and measured TOC , S_1 and, S_2 for Mbate-1 well used for training.

measured in V_m by the logging tool, Δt is the time taken during the transition, $R_{baseline}$ is resistivity equivalent to the $\Delta t_{baseline}$ value during non-source rocks reached the curved baseline, and n is created on the ratio of quantity of transition time cycle to one resistivity cycle usually given as 0.02. The baseline values are derived when 1 logarithmic deep resistivity value equivalent to 50 sonic overlaid with the DT log scaling on the deep resistivity log. An overlap between the high deep resistivity log and high sonic DT log in the lithology indicates the presence of a potential area in the zone in the crossover.

Artificial Neural Network (ANN)

An ANN is a computational model that mimics the human brain’s role to learn from instances and

discover solutions to complex problems in decision-making and classification challenges (Shanmuganathan, 2016). An ANN uses nonlinear and complex types of hypotheses $h_{w,b}(x)$ to describe and interpret the predictions of new instances, provided the training dataset x_i with a vector of dimension m function feature, thus:

$$h_{w,b}(x) = f(w^T x + b) = \frac{1}{1 + \exp(-(w^T x + b))} \quad (3)$$

where x_i is the input, w is weight allocated to each input x_i , b is the feature bias, and the output is $h_{w,b}(x)$. The inputs used in this study were the well logs of DT, RHOB, LLD, GR, SP, and NPHI. To determine the output network values of TOC , S_1 , and S_2 , forward propagation is carried out based on:

$$z^{l+1} = w^l \alpha^l + b^l \quad (4)$$

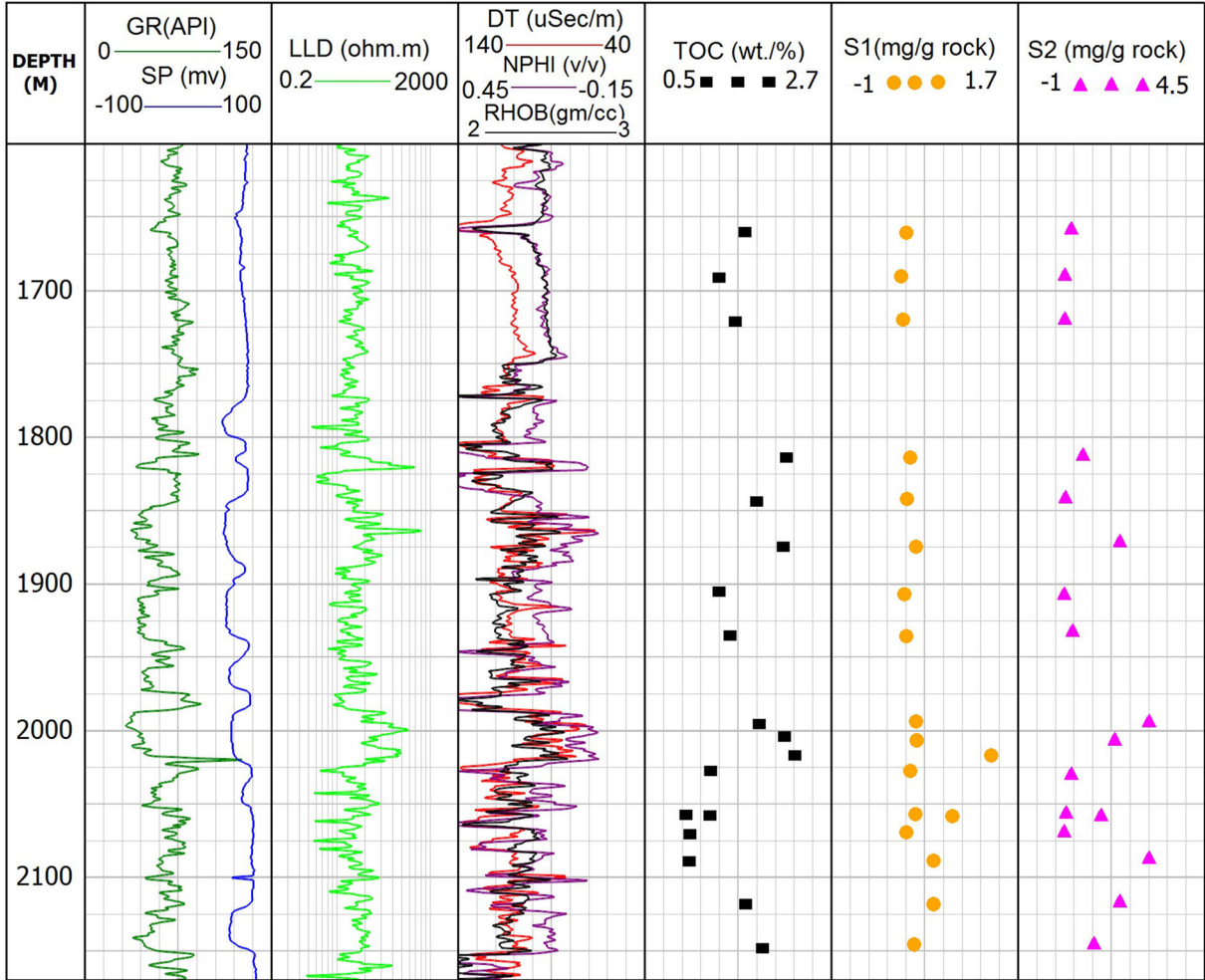


Figure 5. Well log suite and measured TOC, S₁, and S₂ for Mita gama-1 well used for testing.

$$\alpha^{l+1} = f(z^{l+1}) \tag{5}$$

where l denotes the neural network layer, z_i^l denotes the cumulative weighted number of inputs in layer l for neuron i , and α_i^l signifies the output after the activation function of z_i^l . The sigmoid activation function for every neuron is represented by $f(z)$. Depending on the activation functions and, model weights a hypothesis $h_{w,b}(x)$ is determined after the forward propagation step. The error function $J(w, b) - \frac{1}{2} \|h_{w,b}(x) - y\|^2$ is then determined with the purpose of decreasing the loss (error) function for the network training. The loss function is a technique of assessing how well the specific algorithm models the dataset given. Since the error

function $J(w, b)$ is given as a non-convex function, Adaptive Moment Estimation (Adam) or Stochastic Gradient Descent (SGD) as the optimization algorithms may be utilized to modify the weights from each neuron to minimize the local error (Zhang et al., 2019). The SGD updates weight (w) and bias (b), where α is the learning rate, as seen in Eqs. (6) and (7).

$$w_{ij}^l = w_{ij}^l - \alpha \frac{\partial}{\partial w_{ij}^l} J(w, b) \tag{6}$$

$$b_{ij}^l = w_{ij}^l - \alpha \frac{\partial}{\partial w_{ij}^l} J(w, b) \tag{7}$$

The efficient method to calculate the partial derivatives to adjust the weights (w) can be given by

the backpropagation algorithm [Eq. (8)]. The partial derivatives of the hidden layer can also be calculated and used to update weights $j = l - 1; j > 2; j - -$ thus:

$$\delta_l = \frac{\partial}{\partial z_l^l} \frac{1}{2} \|h_{w,b}(x) - y\|^2 = -(y_i - \alpha_i^l) f'(z_i^l) \quad (8)$$

$$\delta^j = ((w^l)^T \delta^{j+1}) f'(z^j) \quad (9)$$

The neurons throughout the previous layer are assigned a penalty δ^j , for the weights, and the local error of each layer j is modified based on the penalty δ^j . According to Qian (1999), the momentum coefficient is used to enhance the SGD mostly in the proper path to avoid becoming local minimum stuck during optimization; this is achieved by adding a fraction number of the updated layer of the past step time to the updated current layer. Proper weight adjustment results in lower error rates and hence, increasing its generalization, makes the model more efficient.

The present study employed the radial basis function neural network (RBFNN) and back-propagation neural network (BPNN). The optimal model structure for BPNN and RBFNN models was achieved based on the trial-and-error method by tuning the number of hidden neurons and the spread parameter, respectively, in order to achieve the least error margin. The gradient descent approach was adopted for RBFNN. The model learning process for BPNN was achieved for 1000 epochs with a learning rate of 0.03 and a momentum coefficient of 0.7. The ANN models were developed, coded, and implemented in MATLAB software R2019b.

Group Method of Data Handling (GMDH) Network

The objective of the GMDH model is to find a function, \hat{f} , which is used as an approximation rather than a real function, f , to estimate the output, t , for a defined input vector $U = (u_1, u_2, u_3, \dots, u_n)$, close to its real output, p (TOC, S_1, S_2). Therefore, given single output with n multi-input data pair of observations, so that:

$$t_i = f(u_{i1}, u_{i2}, u_{i3}, \dots, u_{in}) \quad (i = 1, 2, 3, \dots, M) \quad (10)$$

With any given input vector $U = u_{i1}, u_{i2}, u_{i3}, \dots, u_{in}$, the GMDH network can then be trained to estimate the output values t , meaning:

$$t_i = \hat{f}(u_{i1}, u_{i2}, u_{i3}, \dots, u_{in}) \quad (i = 1, 2, 3, \dots, M) \quad (11)$$

The GMDH builds the general relationship in the context of a mathematical description between output and input parameters called the reference to solve this problem. The problem here is to determine the GMDH network so that the square difference between the predicted and the actual output is minimized as:

$$\sum_{i=1}^M [\hat{f}(u_{i1}, u_{i2}, u_{i3}, \dots, u_{in}) - p_i]^2 \rightarrow \min \quad (12)$$

The polynomial series known as the Kolmogorov–Gabor polynomial, which is the complex discrete form of the Volterra function (Anastasakis & Mort, 2001; Najafzadeh & Azamathulla, 2013), can express the general link between input parameters and output parameter in the mode of:

$$p = a_o + \sum_{i=1}^N a_i u_i + \sum_{i=1}^N \sum_{j=1}^N a_{ij} u_i u_j + \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^N a_{ijk} u_i u_j u_k + \dots \quad (13)$$

Equation (12) can be simplified using the simplified GMDH network’s partial quadratic polynomial equation as (Shen et al., 2019):

$$t = W(u_i, u_j) = a_o + a_1 u_i + a_2 u_j + a_3 u_i^2 + a_4 u_j^2 + a_5 u_i u_j \quad (14)$$

The mathematical relationship between the variables of input–output specified in Eq. (13) is generated by this network of connected neurons. The coefficients of weighting (Eq. (14)) are computed using techniques of regression to minimize the variation between an actual (p) and predicted (t) output, as for each variable pair of u_i and u_j input is minimized (Armaghani et al., 2020). The schematic type of the GMDH network architecture is shown in Figure 6.

Using the quadratic equation [Eq. (14)], a tree of polynomials is created in which the coefficients of weighting can be found by employing the least square approach technique. The quadratic function of weighting coefficients W_i is obtained to match the output optimally in the complete set of output–input data pairs as:

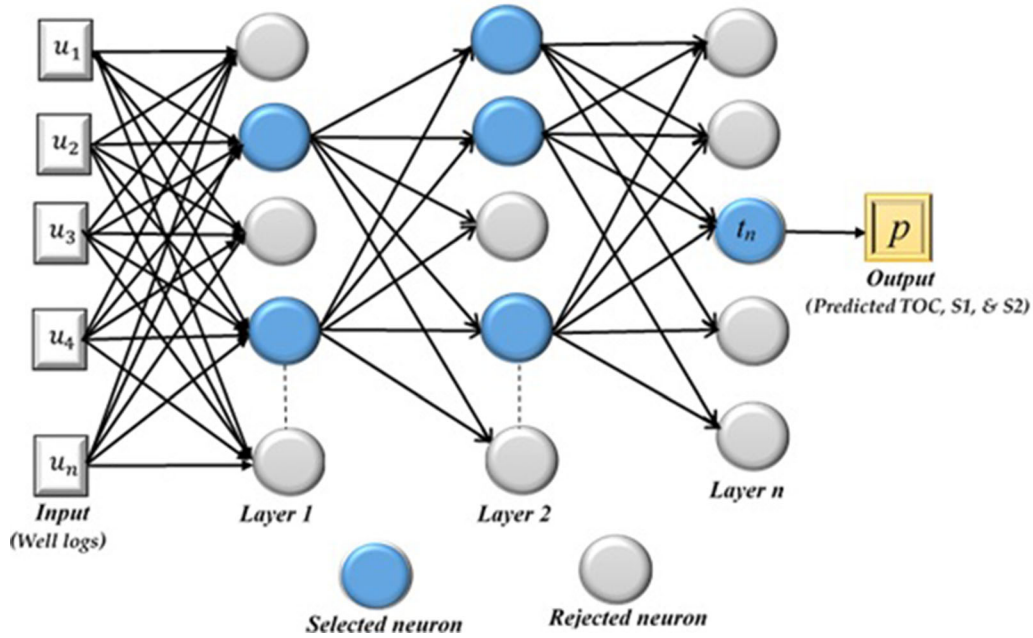


Figure 6. Type of the GMDH network architecture.

$$E = \frac{\sum_{i=1}^M (p_i - W_i())^2}{M} \rightarrow \min \quad (15)$$

The possibilities for dual independent variables among the overall n input parameters are drawn to fit the general form of GMDH algorithms to build the regression polynomial defined in the form of Eq. (13), which suits better in the least-square sense of the dependent observations ($p_i, i = 1, 2, 3 \dots M$). Consequently $C_n^2 = n(n - 1)/2$, quadratic polynomial neurons can be constructed from observations $\{(p_i; u_{xi}, u_{yi}); (i = 1, 2, 3 \dots M)\}$ for various $x, y \in \{1, 2, 3, \dots n\}$ in the feed-forward network's first layer. Triples of M data can now be built $\{(p_i; u_{xi}, u_{yi}); (i = 1, 2, 3 \dots M)\}$, using those $x, y \in \{1, 2, 3, \dots n\}$ in the form of:

$$\begin{bmatrix} u_{1x} & u_{1y} & : & p_1 \\ u_{2x} & u_{2y} & : & p_2 \\ \dots & \dots & : & \dots \\ u_{mx} & u_{my} & : & p_m \end{bmatrix} \quad (16)$$

For every row of the m data triples, using the quadratic sub-expression shown in the form of Eq. (14), the following matrix expression can be found directly as:

$$Aa = P \quad (17)$$

where a represents an undefined vector of quadratic polynomial weighting coefficients in Eq. (14):

$$a = [a_0, a_1, a_2, a_3, a_4, a_5]^T \quad (18)$$

where T indicates matrix transposition:

$$P = [p_1, p_2, p_3, \dots, p_M]^T \quad (19)$$

Equation (16) represents a vector of outputs' observation written as:

$$A = \begin{bmatrix} 1 & u_{1x} & u_{1y} & u_{1x}u_{1y} & u_{1x}^2 & u_{1y}^2 \\ 1 & u_{2x} & u_{2y} & u_{2x}u_{2y} & u_{2x}^2 & u_{2y}^2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & u_{mx} & u_{my} & u_{mx}u_{my} & u_{mx}^2 & u_{my}^2 \end{bmatrix} \quad (20)$$

The least-square approach derived from the technique of multiple-regression analysis results in normal equations being solved, which is in the form of:

$$a = (A^{TA})^{-1}A^TP \quad (21)$$

For the whole set of triples of m data, Eq. (21) specifies the vector for the best quadratic weighting coefficients given in Eq. (10). The GMDH model was also coded and implemented in MATLAB software R2019b.

RESULTS AND DISCUSSIONS

The GMDH and ANN models were developed using MATLAB R2019b. The results from the machine learning predictive models were compared fairly using mean square error (*MSE*) and mean absolute error (*MAE*) as statistical indices. The differences between the predictions and the measured *TOC*, S_1 , and S_2 values were highlighted by *MSE* and *MAE*. For comparison of the models, *MSE* and *MAE* approaching zero proves that a model is an accepted predictor. The mathematical expressions for *MSE* and *MAE* are given, respectively, as:

$$MSE = \frac{1}{n} \sum_{i=1}^n (t_i - p_i)^2 \quad (22)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |t_i - p_i| \quad (23)$$

where p is the predicted outcome, t is the measured value, and n is the total number of observations. During training, *MSE* and *MAE* approaching zero mean that a model trained well. However, more emphasis is placed on the ability of the trained model to perform well when tried on the withheld testing data. Therefore, testing results (*MSE* and *MAE*) approaching zero signify the best performing model with an improved generalization capacity.

TOC Prediction

The baseline for the resistivity and sonic well logs when estimating *TOC* using $\Delta \log R$ were observed at 335.83 $\mu\text{s/m}$ and 0.866 Ωm for Mbuo-1, 52.30 $\mu\text{s/m}$ and 42.47 Ωm for Mbate-1, 68.77 $\mu\text{s/m}$ and 38.84 Ωm for Mita Gamma-1. The estimates from $\Delta \log R$ generated *MSE* and *MAE* of 8.795 and 1.66, respectively, for Mbate-1 and Mbuo-1; for Mita Gamma-1, the *MSE* and *MAE* were 1.418 and 0.890, respectively (Table 1). However, the ANN models of BPNN and RBFNN performed better than the conventional $\Delta \log R$ approach. The optimal BPNN model structure observed after training the *TOC* data was 6 inputs, a hidden layer with 12 neurons, while 6 inputs and a hidden layer with a spread parameter of 0.07 were identified for RBFNN. From Table 1, BPNN provided predictions having *MSE* and *MAE* of 0.606 and 0.56, respectively, for train-

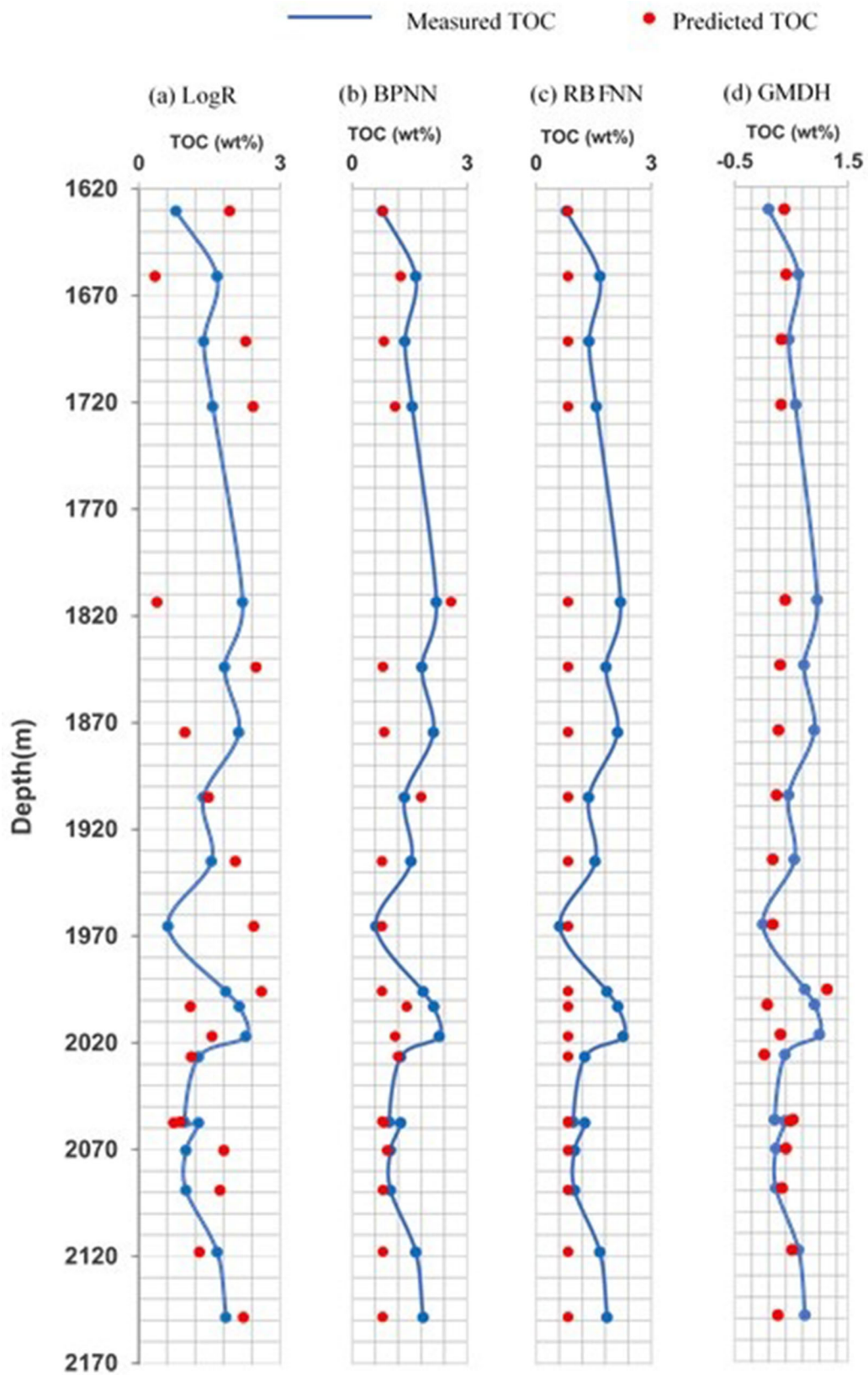
Table 2. Statistical parameters of *TOC* predictive models during training and testing

Model	<i>MSE</i>		<i>MAE</i>	
	Train	Test	Train	Test
GMDH	0.0184	0.4042	0.0988	0.455
BPNN	0.606	1.2689	0.5634	0.8195
RBFNN	0.0064	0.6813	0.0197	0.704
Log R	8.795	1.418	1.6643	0.8904

ing; for testing, *MSE* and *MAE* were 1.269 and 0.819, respectively. It can be seen from Table 2 that RBFNN trained well with *MSE* and *MAE* of 0.0064 and 0.019 but failed to generate similar results when the model was tested on the Mita Gamma-1 well data as its predictions had *MSE* and *MAE* of 0.681 and 0.70, respectively. The best performing model was the GMDH, which trained well with its outcome having *MSE* and *MAE* of 0.018 and 0.098, respectively, while generalizing better than $\Delta \log R$ and ANN on Mita Gamma-1. The GMDH provided *MSE* and *MAE* of 0.4 and 0.455, respectively, for *TOC* during testing as expressed in Table 2. The performance of the *TOC* models is compared in Figure 7.

S_1 and S_2 Prediction

For the hydrocarbon potential distribution parameters, S_1 and S_2 , the results from RBFNN when training with Mbuo-1 and Mbate-1 data were excellent. RBFNN had *MSE* and *MAE* of 8×10^{-6} and 0.0008, respectively, for S_1 while generating 0.0012 and 0.299, respectively, for S_2 (Table 3). The optimal model structure of the S_1 RBFNN model was 6 inputs, a hidden layer with a parameter spread of 0.03. The S_2 RBFNN model that generated the best outcome had 6 inputs and a hidden layer with spread parameter of 0.01. Unsurprisingly, RBFNN was unable to generate similar results for the Mita Gamma 1 data during testing. During testing, RBFNN achieved relatively high *MSE* and *MAE* of 0.15 and 0.245, respectively, for S_1 , 1.14, and 0.87, respectively, for S_2 . On the other hand, the best BPNN model structure that produced the least error rate for S_1 was 6 input, a hidden layer with 6 neurons. The S_2 BPNN that had the optimal architecture was 6 inputs and a hidden layer with 2 neurons. From Table 3, the S_1 BPNN model achieved *MSE*



◀ **Figure 7.** Plots of GMDH, ANN, and $\Delta\log R$ predictions and measured TOC .

Table 3. Statistical parameters of S_1 predictive models during training and testing

Model	<i>MSE</i>		<i>MAE</i>	
	Train	Test	Train	Test
GMDH	0.0011	0.04327	0.0243	0.0702
BPNN	0.0102	0.0846	0.0555	0.1669
RBFNN	0.000006	0.1511	0.0008	0.245

Table 4. Statistical parameters of S_2 predictive models during training and testing

Model	<i>MSE</i>		<i>MAE</i>	
	Train	Test	Train	Test
GMDH	0.0069	1.1597	0.0489	0.0702
BPNN	5.52	1.9615	1.3281	0.8988
RBFNN	0.0012	1.4144	0.2994	0.87

and *MAE* of 0.0102 and 0.056, respectively, for training. During testing, it generated predictions scoring 0.085 and 0.167 for *MSE* and *MAE*, respectively (Table 3). The BPNN for predicting S_2 data produced comparatively the worst estimates with *MSE* and *MAE* of 5.52 and 1.33, respectively, and 1.96 and 0.9, respectively, for testing (Table 4). The BPNN outcome for the testing data of the Mita Gamma-1 well is illustrated in Figure 8.

The prediction from the GMDH gave an error values of 0.0011 and 0.024 for S_1 , and 0.007 and 0.049 during training in the case of S_2 (Tables 3 and 4). Looking at the generalization performance, the GMDH tested better than BPNN and RBFNN when it produced testing results with *MSE* and *MAE* of 0.04, 0.0702, respectively, for S_1 , and 1.16 and 0.07 for S_2 , respectively. This makes the GMDH the best performing model compared to ANN. Figure 9 describes the predictions of the GMDH, BPNN, and RBFNN.

It was recognized that the performance of the GMDH in estimating S_2 was generally low compared to the TOC and S_1 results. This is due to the

fact that the Mita Gamma-1 well, which was used as the testing data, had a high variation of S_2 values. The high variation in S_2 values of the testing data generated an imbalance distribution of the training and testing observations. However, the GMDH generalized better than ANN when handling the high variation of S_2 values. Consequently, ANN models failed to account for the changes in S_2 values of Mita-Gamma 1 well as observed in Figure 9.

Quality of Organic Matter and Hydrocarbon Generation Potential for East Lika 1

The constructed GMDH prediction model in this paper was further used to estimate the TOC , S_2 , and S_1 for the East Lika-1 well, which had no core geochemical data. Figure 10 shows the well logs from East Lika-1 and the GMDH predicted values for TOC , S_1 , and S_2 . The TOC prediction showed a generally constant value across the entire depth of the well. According to the generated results from GMDH, we can confirm that East Lika-1 well has a low-quality source rock due to the near-constant TOC values across the entire depth of the well. Based on the relatively high S_1 values compared to S_2 from depths of 2060 m to 2140 m, it can be assumed that there exists evidence of hydrocarbon migration.

CONCLUSIONS

The application of machine learning methods in evaluating geochemical data such as TOC from well log parameters has revealed the shortcomings of artificial neural networks (ANN). Therefore, this study proposed a group method of data handling (GMDH) neural network as a means of offering an improved performance when predicting TOC , S_1 , and S_2 . In line with this, Mbuo-1 and Mbate-1 wells were considered as training data, while the predictive capability of the models was judged on the Mita-Gamma-1 well.

When estimating TOC , the ANN models of radial basis function neural network (RBFNN) and backpropagation neural network (BPNN) trained well but failed to generate similar results when the developed models were tested on Mita-Gamma-1 well. The conventional $\Delta\log R$ model obtained the

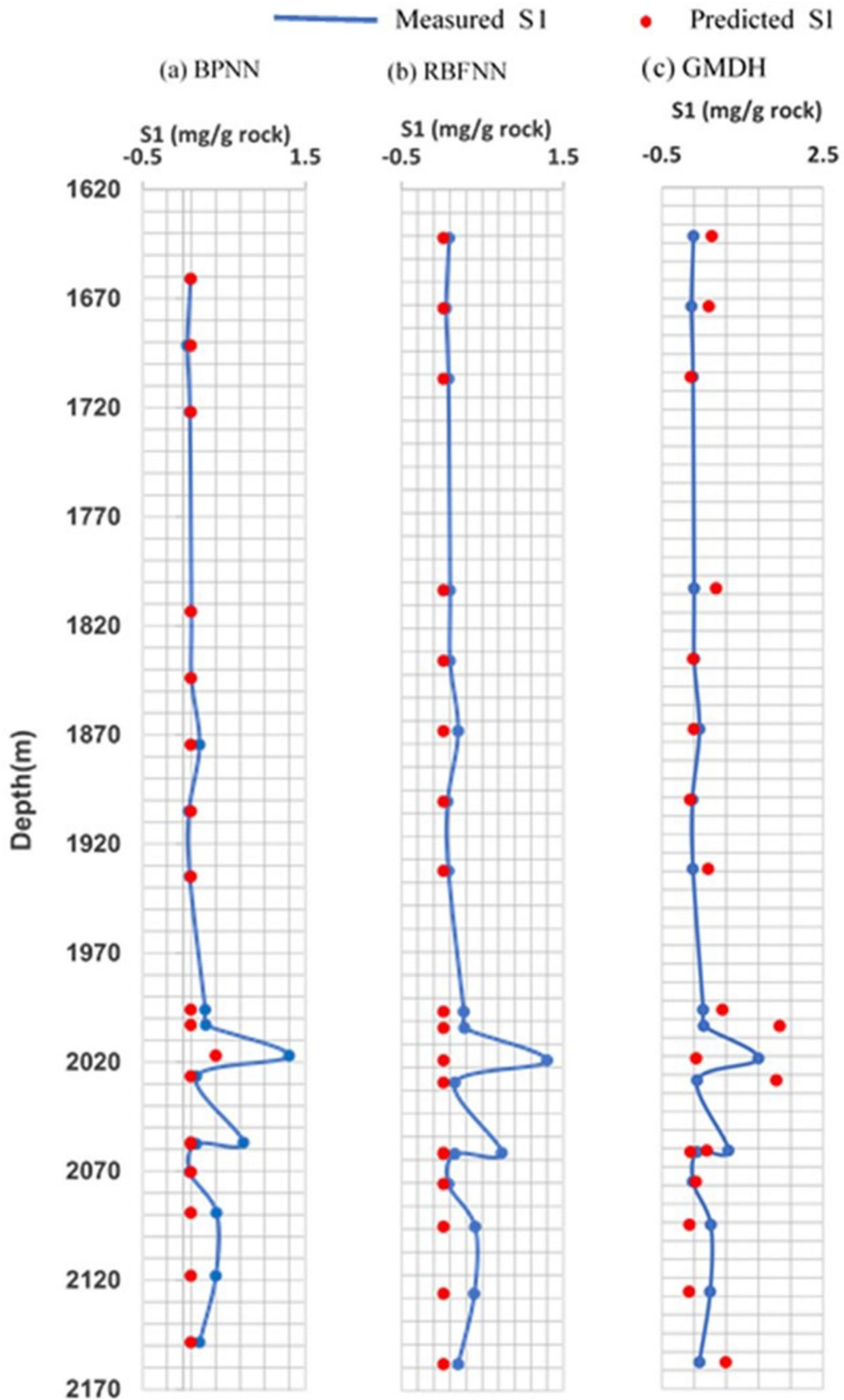


Figure 8. Plots of GMDH, BPNN, and RBFNN predictions and measured S_1 .

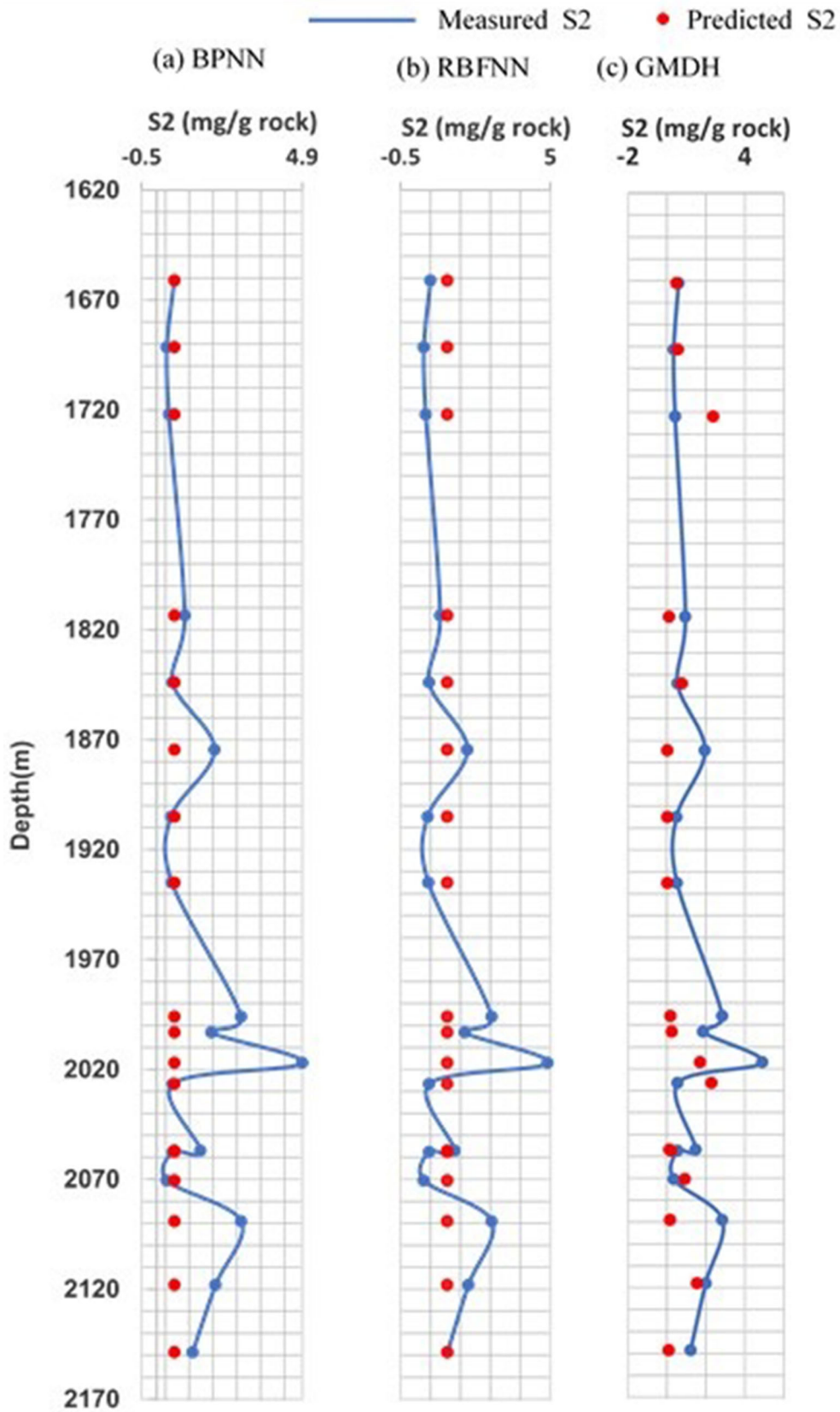


Figure 9. Plots of GMDH, BPNN, and RBFNN predictions and measured S₂.

highest error rate from its estimates. However, the GMDH performed well during training and further produced the least error margin compared to ANN and $\Delta \log R$ during testing.

A similar scenario was observed when estimating the hydrocarbon potential parameters of S_1 and S_2 . RBFNN and BPNN generated close to perfect training results for both S_1 and S_2 but failed to generate similar results when tested on the Mita-Gamma well data. The GMDH provided the least

error margins despite its excellent training capabilities. Based on the outcomes, we suggest GMDH as an improved machine learning method as an alternative to the ANN algorithms for estimating geochemical results like TOC , S_1 , and S_2 . The proposed GMDH was further used to determine the TOC , S_1 and S_2 results for East Lika-1 well, which has no core geochemical data. The predictions revealed a low-quality source rock for East Lika-1 well and a pos-

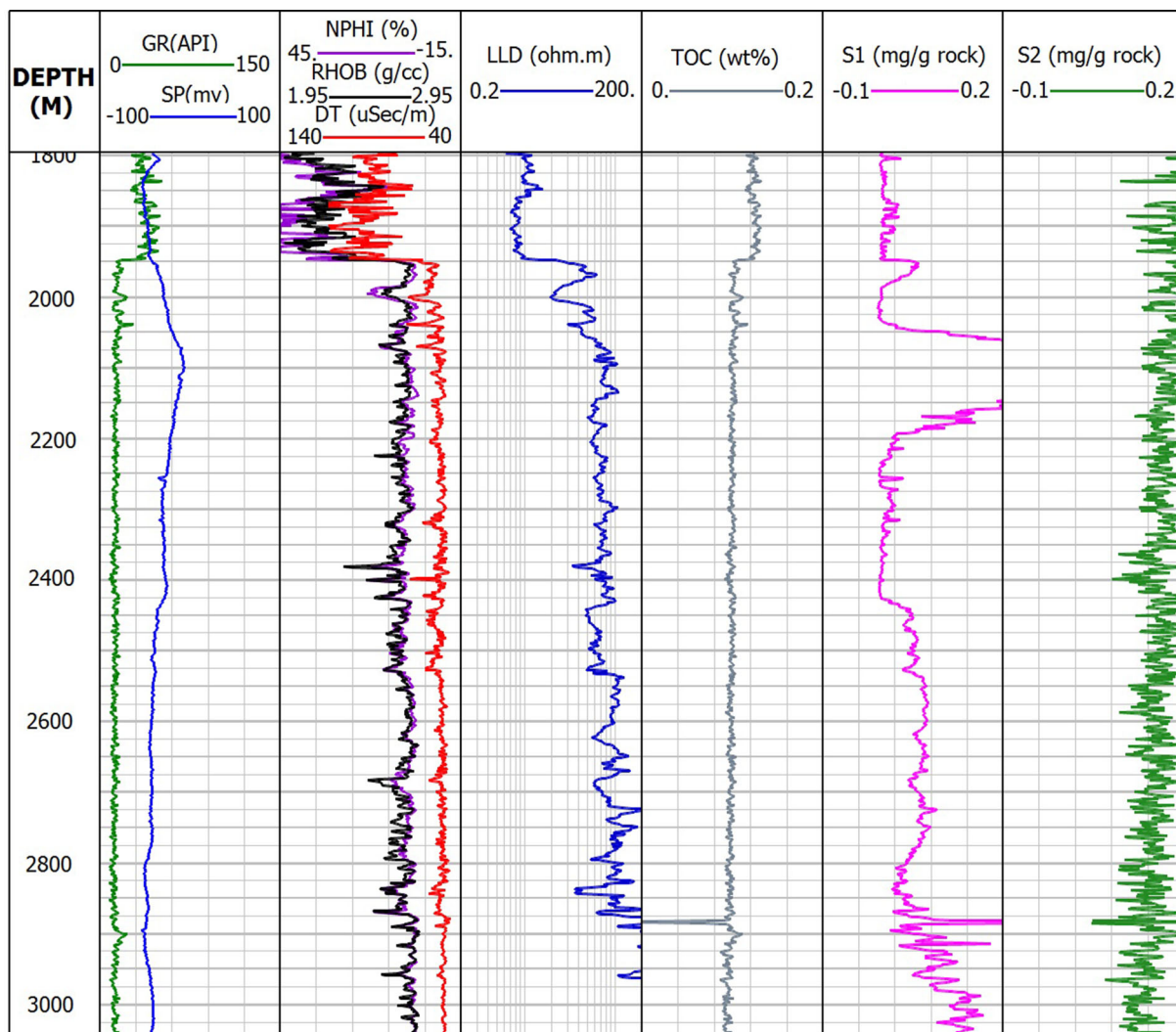


Figure 10. Well logs from East Lika-1 and GMDH predicted values for TOC , S_1 , and S_2 .

sible hydrocarbon migration occurring between the depths of 2060 m and 2140 m.

ACKNOWLEDGEMENTS

This work was supported by the Major National Science and Technology Programs in the “Thirteenth Five-Year” Plan period (No. 2017ZX05032-002-004), the Outstanding Youth Funding of Natural Science Foundation of Hubei Province (No. 2016CFA055).

REFERENCES

- Amiri Bakhtiar, H., Telmadarreie, A., Shayesteh, M., Heidari Fard, M., Talebi, H., & Shirband, Z. (2011). Estimating total organic carbon content and source rock evaluation, applying AlogR and neural network methods: Ahwaz and Marun oilfields, SW of Iran. *Petroleum Science and Technology*, 29(16), 1691–1704.
- Anastasakis, L., & Mort, N. (2001). The development of self-organization techniques in modelling: a review of the group method of data handling (GMDH). *Research Report-University of Sheffield Department of Automatic Control and Systems Engineering*, 47, 191–204.
- Armaghani, D. J., Momeni, E., & Asteris, P. (2020). Application of group method of data handling technique in assessing deformation of rock mass. *Metaheuristic Computing Applied*, 1, 1–18.
- 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.
- Bai, Y., & Tan, M. (2020). Dynamic committee machine with fuzzy-c-means clustering for total organic carbon content prediction from wireline logs. *Computers & Geosciences*, 146, 104626.
- Cappuccio, F., Porreca, M., Omosanya, K. O., Minelli, G., & Harishidayat, D. (2020). Total organic carbon (TOC) enrichment and source rock evaluation of the Upper Jurassic-Lower Cretaceous rocks (Barents Sea) by means of geochemical and log data. *International Journal of Earth Sciences*, 7, 1–12.
- Caracciolo, L., Andò, S., Vermeesch, P., Garzanti, E., McCabe, R., Barbarano, M., et al. (2020). A multidisciplinary approach for the quantitative provenance analysis of siltstone: Mesozoic Mandawa Basin, southeastern Tanzania. *Geological Society, London, Special Publications*, 484(1), 275–293.
- 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.
- Charsky, A., & Herron, S. (2013). Accurate, direct total organic carbon (TOC) log from a new advanced geochemical spectroscopy tool: Comparison with conventional approaches for TOC estimation. In *AAPG annual convention and exhibition, Pittsburgh, PA, May, 2013* (pp. 19–22).
- Curiale, J. A., & Curtis, J. B. (2016). Organic geochemical applications to the exploration for source-rock reservoirs—A review. *Journal of Unconventional Oil and Gas Resources*, 13, 1–31.
- Delvaux, D. (2001). Karoo rifting in western Tanzania: Precursor of Gondwana break-up. *Contributions to geology and paleontology of Gondwana in honor of Helmut Wopfner: Cologne, Geological Institute, University of Cologne*, 111–125.
- Einvik-Heitmann, V., Dypvik, H., Hou, G., Fossum, K., Nerbraten, K., Karega, A., et al. (2015). The early cretaceous Kihuluhulu formation of the Mandawa Basin. In *First EAGE Eastern Africa petroleum geoscience forum, 2015* (Vol. 2015, pp. 1–1, Vol. 1). European Association of Geoscientists & Engineers.
- Emanuel, A., Kasanzu, C., & Kagya, M. (2020). Geochemical characterization of hydrocarbon source rocks of the Triassic-Jurassic time interval in the Mandawa basin, southern Tanzania: Implications for petroleum generation potential. *South African Journal of Geology*, 17, 121–134.
- Evenick, J. C. (2020). Late Cretaceous (Cenomanian and Turonian) organofacies and TOC maps: Example of leveraging the global rise in public-domain geochemical source rock data. *Marine and Petroleum Geology*, 111, 301–308.
- Fossum, K., Dypvik, H., Haid, M. H., Hudson, W. E., & van den Brink, M. (2020). Late Jurassic and Early Cretaceous sedimentation in the Mandawa Basin, coastal Tanzania. *Journal of African Earth Sciences*, 104013.
- Fossum, K., Morton, A. C., Dypvik, H., & Hudson, W. E. (2019). Integrated heavy mineral study of Jurassic to Paleogene sandstones in the Mandawa Basin, Tanzania: Sediment provenance and source-to-sink relations. *Journal of African Earth Sciences*, 150, 546–565.
- Hakimi, M. H., Abdullah, W. H., Lashin, A. A., Ibrahim, E.-K.H., & Makeen, Y. M. (2020). Hydrocarbon generation potential of the organic-rich Naifa Formation, Say'un–Masila Rift Basin, Yemen: Insights from geochemical and palynofacies analyses. *Natural Resources Research*, 29(4), 2687–2715.
- Hood, A., Gutjahr, C., & Heacock, R. (1975). Organic metamorphism and the generation of petroleum. *AAPG Bulletin*, 59(6), 986–996.
- Hou, G. (2015). *Late Cretaceous Sedimentation (Mavuji Group) in Mandawa Basin, Tanzania*.
- Hudson, W. (2011). *The geological evolution of the petroleum prospective Mandawa Basin southern coastal Tanzania*. Trinity College Dublin.
- Hudson, W., & Nicholas, C. (2014). The Pindiuro Group (Triassic to Early Jurassic Mandawa Basin, southern coastal Tanzania): Definition, palaeoenvironment, and stratigraphy. *Journal of African Earth Sciences*, 92, 55–67.
- Kagya, M. L. (1996). Geochemical characterization of Triassic petroleum source rock in the Mandawa basin, Tanzania. *Journal of African Earth Sciences*, 23(1), 73–88.
- Liu, C., Zhao, W., Sun, L., Wang, X., Sun, Y., Zhang, Y., et al. (2020). Geochemical assessment of the newly discovered oil-type Shale in the Shuangcheng area of the northern Songliao Basin, China. *Journal of Petroleum Science and Engineering*, 196, 107755.
- Mahmoud, A. A., Elkatatny, S., Ali, A., Abdurraheem, A., & Abouelresh, M. (2020). Estimation of the total organic carbon using functional neural networks and support vector machine. In *International petroleum technology conference, 2020*. International Petroleum Technology Conference.
- Mahmoud, A. A., Elkatatny, S., Ali, A. Z., Abouelresh, M., & Abdurraheem, A. (2019). Evaluation of the total organic carbon (TOC) using different artificial intelligence techniques. *Sustainability*, 11(20), 5643.
- Mahmoud, A. A. A., Elkatatny, S., Mahmoud, M., Abouelresh, M., Abdurraheem, A., & Ali, A. (2017). Determination of the total organic carbon (TOC) based on conventional well logs

- using artificial neural network. *International Journal of Coal Geology*, 179, 72–80.
- Mani, D., Kalpana, M., Patil, D., & Dayal, A. (2017). Organic matter in gas shales: Origin, evolution, and characterization. In *Shale Gas* (pp. 25–54). Elsevier.
- Najafzadeh, M., & Azamathulla, H. M. (2013). Group method of data handling to predict scour depth around bridge piers. *Neural Computing and Applications*, 23(7–8), 2107–2112.
- Nerbråten, K. B. (2014). *Petrology and sedimentary provenance of Mesozoic and Cenozoic sequences in the Mandawa Basin*.
- Passey, Q., Creaney, S., Kulla, J., Moretti, F., & Stroud, J. (1990). A practical model for organic richness from porosity and resistivity logs. *AAPG Bulletin*, 74(12), 1777–1794.
- Passey, Q. R., Bohacs, K., Esch, W. L., Klimentidis, R., & Sinha, S. (2010). From oil-prone source rock to gas-producing shale reservoir-geologic and petrophysical characterization of unconventional shale gas reservoirs. In *International oil and gas conference and exhibition in China, 2010*. Society of Petroleum Engineers.
- Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1), 145–151.
- 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*, 104429.
- Schmoker, J. W. (1979). Determination of organic content of Appalachian Devonian shales from formation-density logs: Geologic notes. *AAPG Bulletin*, 63(9), 1504–1509.
- 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.
- Shalaby, M. R., Malik, O. A., Lai, D., Jumat, N., & Islam, M. A. (2020). Thermal maturity and TOC prediction using machine learning techniques: Case study from the Cretaceous-Paleocene source rock, Taranaki Basin, New Zealand. *Journal of Petroleum Exploration and Production Technology*, 10, 2175–2193.
- Shanmuganathan, S. (2016). Artificial neural network modelling: An introduction. In *Artificial neural network modelling* (pp. 1–14). Springer.
- 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.
- Shi, X., Wang, J., Liu, G., Yang, L., Ge, X., & Jiang, S. (2016). Application of extreme learning machine and neural networks in total organic carbon content prediction in organic shale with wireline logs. *Journal of Natural Gas Science and Engineering*, 33, 687–702.
- Smelror, M., Fossum, K., Dypvik, H., Hudson, W., & Mweneinda, A. (2018). Late Jurassic-Early Cretaceous palynostratigraphy of the onshore Mandawa Basin, southeastern Tanzania. *Review of Palaeobotany and Palynology*, 258, 248–255.
- Tan, M., Song, X., Yang, X., & Wu, Q. Z. (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*, 26, 792–802.
- Tariq, Z., Mahmoud, M., Abouelresh, M., & Abdurraheem, A. (2020). Data-driven approaches to predict thermal maturity indices of organic matter using artificial neural networks. *ACS Omega*.
- Tenaglia, M., Eberli, G. P., Weger, R. J., Blanco, L. R., Sánchez, L. E. R., & Swart, P. K. (2020). Total organic carbon quantification from wireline logging techniques: A case study in the Vaca Muerta Formation, Argentina. *Journal of Petroleum Science and Engineering*, 107489.
- Wang, P., Peng, S., & He, T. H. (2018). A novel approach to total organic carbon content prediction in shale gas reservoirs with well logs data, Tonghua Basin, China. *Journal Natural Gas Science and Engineering*, 55, 1–15.
- Wang, H., Wu, W., Chen, T., Dong, X., & Wang, G. (2019a). An improved neural network for TOC, S1 and S2 estimation based on conventional well logs. *Journal of Petroleum Science and Engineering*, 176, 664–678.
- Wang, J., Gu, D., Guo, W., Zhang, H., & Yang, D. (2019b). Determination of total organic carbon content in shale formations with regression analysis. *Journal of Energy Resources Technology*, 141(1).
- Xiong, H., Wu, X., & Fu, J. (2019). Determination of total organic carbon for organic rich shale reservoirs by means of cores and logs. In *SPE annual technical conference and exhibition, 2019*. Society of Petroleum Engineers.
- Yu, H., Rezaee, R., Wang, Z., Han, T., Zhang, Y., Arif, M., & Johnson, L. (2017). A new method for TOC estimation in tight shale gas reservoirs. *International Journal of Coal Geology*, 179, 269–277.
- Zhang, B., Chen, C., Ye, Q., Liu, J., & Doermann, D. (2019). Calibrated stochastic gradient descent for convolutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence, 2019* (Vol. 33, pp. 9348–9355).
- Zhu, L., Zhang, C., Zhang, C., Wei, Y., Zhou, X., Cheng, Y., et al. (2018). Prediction of total organic carbon content in shale reservoir based on a new integrated hybrid neural network and conventional well logging curves. *Journal of Geophysics and Engineering*, 15(3), 1050–1061.
- Zhu, L., Zhang, C., Zhang, C., Zhang, Z., Nie, X., Zhou, X., et al. (2019a). Forming a new small sample deep learning model to predict total organic carbon content by combining unsupervised learning with semisupervised learning. *Applied Soft Computing*, 83, 105596.
- Zhu, L., Zhang, C., Zhang, Z., Zhou, X., & Liu, W. (2019b). An improved method for evaluating the TOC content of a shale formation using the dual-difference $\Delta\log R$ method. *Marine and Petroleum Geology*, 102, 800–816.
- Zhu, L., Zhang, C., Zhang, C., Zhang, Z., Zhou, X., Liu, W., et al. (2020). A new and reliable dual model-and data-driven TOC prediction concept: A TOC logging evaluation method using multiple overlapping methods integrated with semi-supervised deep learning. *Journal of Petroleum Science and Engineering*, 188, 10694.