Energy 284 (2023) 129232

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy

Deep learning integrated approach for hydrocarbon source rock evaluation and geochemical indicators prediction in the Jurassic - Paleogene of the Mandawa basin, SE Tanzania

Christopher N. Mkono [a,b,*], Shen Chuanbo [a,**], Alvin K. Mulashani [c,***], Grant Charles Mwakipunda [b]

[a] Key Laboratory of Tectonics and Petroleum Resources, Ministry of Education, China University of Geosciences, Wuhan, 430074, China
[b] Department of Petroleum Engineering, School of Earth Resources, China University of Geosciences, Wuhan, 430074, China
[c] Department of Geoscience and Mining Technology, College of Engineering and Technology, Mbeya University of Science and Technology, Box 131, Mbeya, Tanzania

ARTICLE INFO

Handling Editor: Wojciech Stanek

Keywords:
Total organic carbon
Group method of data handling
Hydrocarbon potentials
Thermal maturity
Source rocks
Energy

ABSTRACT

The world's energy demands are growing at an unprecedented rate, and the exploration of new hydrocarbon sources is more important than ever. Therefore, the objective of this study was first to quantitatively analyze hydrocarbon source rock potentiality of the Triassic-Jurassic of Mandawa Basin based on the generalized group method of data handling neural network (g-GMDH), Machine learning, and Geochemical using well logs data. Then a novel g-GMDH was presented to predict a continuous geochemical log profile of TOC, Tmax, S1, and S2. It was observed that the basin's hydrocarbon source rocks are classified as fair to very good source rocks with TOC contents ranging from 0.5 to 8.7 wt%. The source rocks contain mixed kerogen type II and III, which are oil and gas-prone, ranging from immature to mature source rocks. The results of the predictive models indicated that the g-GMDH model trained better whilst generalizing well throughout the testing data than both GPR and SVM models. Specifically, the g-GMDH when tested on unseen data had the least value of MSE = 0.18, 2.35, 0.08, and 61.74 for TOC, Tmax, S1, and S2 respectively, and MAE = 0.45, 1.37, 0.17 and 11.55 for TOC, Tmax, S1 and S2 respectively. The g-GMDH model was further applied to assess the source rock and predict the geochemical information in the East Lika well, which lacks core data. The proposed model can offer rapid and real-time values of geochemical indicators and are independent of laboratory-dependent parameters therefore, can be adopted as an improved technique for evaluating source rocks in frontier basins.

1. Introduction

The global demand for energy is increasing at an alarming rate, driven by a growing population, urbanization, and industrialization [1]. According to the International Energy Agency (IEA), global energy demand is set to rise by 4.6% in 2023, with fossil fuels remaining the dominant energy source, accounting for 75% of the total energy mix [2–4]. However, the outbreak of COVID-19 pandemic in 2020 has significantly impacted the energy sector, resulting in a decrease in demand by 4% in 2020 [5]. As the demand for energy continues to rise, the search for new sources of energy becomes increasingly critical. One of the most significant sources of energy is fossil fuels, particularly oil and gas, which are extracted from sedimentary rocks [6–9]. The quality and quantity of these fossil fuels are largely determined by the properties of the source rock, such as total organic carbon (TOC) content, thermal maturity (Tmax), and hydrocarbon generation potential (S1 and S2). On the other hand, the analysis of source rock potentiality in sedimentary basins has been limited due to the scarcity of appropriate samples for geochemical analysis [10,11]. Analysis has mostly depended on a few chosen samples, including cuttings samples, or sidewall cores cuttings samples, which mostly fail to offer a comprehensive account of the sequence of lithological changes.

The most reliable method of quantifying the source rock and determining TOC, Tmax, S1 and S2 parameters is by performing organic

[*] Corresponding author. Key Laboratory of Tectonics and Petroleum Resources, Ministry of Education, China University of Geosciences, Wuhan, 430074, China.
[**] Corresponding author.
[***] Corresponding author.
E-mail addresses: chrismkono@cug.edu.cn (C.N. Mkono), cbshen@cug.edu.cn (S. Chuanbo), alvinmulashani@yahoo.com (A.K. Mulashani).

https://doi.org/10.1016/j.energy.2023.129232
Received 25 June 2023; Received in revised form 28 August 2023; Accepted 29 September 2023
Available online 29 September 2023
0360-5442/© 2023 Elsevier Ltd. All rights reserved.

Table 1
Summary of the previous study used for prediction of geochemical parameters.

References	Method	Parameters	Inputs	Limitation
Ahangari, Daneshfar [41]	PSO-LSSVM	TOC, S1 and S2	DEN, CNL, RT, GR, and AC	It is prone to overfitting
Shalaby, Malik [42]	BRNN	Tmax and TOC	GR, RHOB, RT, DTC and NPHI	Requires a large amount of computational resources
Alizadeh, Maroufi [43]	ANN	TOC and S2	DT and RT	It is time consuming
Amosu, Imsalem [44]	SVM	TOC	GR, ILD and DT	Highly sensitive to the choice of parameters
Mandal, Rezaee [46]	ELM	TOC	GR, RHOB, LLD, DT and NPHI	It is a black box model, which means it's difficult to interpret
Handhal, Al-Abadi [45]	kNN	TOC	GR, RT, DN, AC and NCL	It is sensitive to the value of k
Barham, Ismail [47]	FFNN	Tmax and TOC	BD, GD, RT, DTC, DTSH, SGR, U, TH and K	It is computationally expensive

geochemical analysis on core samples in the laboratory [12]. However, coring is an expensive exercise and time-consuming to be conducted on all wells [13,14]. In cases where there may not be enough core data, readily available drill cuttings are usually used to compensate [15,16]. The challenges of using drill cuttings are that it is difficult to reconcile with depth and can be contaminated [17]. Due to that, attempts have been made to generate geochemical profile values from geophysical well logs based on the knowledge that well log parameters can detect the presence of organic matter of a source rock [18,19]. On the contrary, heterogenicity of sedimentary basins, complex mineral matrices and lithology present in reservoirs requires the user to understand the density of the mineral matrix for better interpretation of well logging data such as sonic and density well loggings [20]. Due to the low porosity of these reservoirs, the logging response to porosity might be readily misinterpreted by other mineral information. As a result, determining the continuous profile of geochemical parameter of sedimentary basins formation remain challenging.

In recent years, there has been a growing interest in using machine learning techniques to evaluate the source rock potentiality and predict geochemical parameters from well log data. This is because machine learning models have the advantage of being able to adapt and learn to the dynamic conditions of the reservoir such as depositional and formation environment while utilizing the complete set of well logs for better prediction [21–23]. Several studies such as [24–38] have demonstrated the effectiveness of machine learning techniques in predicting geochemical parameters from well logs. Mulashani, Shen [39] without considering source rock potentiality used standard GMDH model to predict TOC, S1, and S2 from well log data. Their results showed that the standard GMDH model outperformed Passey's conventional method of ΔlogR and standard ANN algorithms of BPNN and RBFNN. These results highlight the superior performance of the standard machine learning over conventional methods. Wang, Wu [40] used a convolutional neural network (CNN) model to predict TOC, S1 and S2 from well log data of the Shahejie Formation. According to their findings, the CNN model was able to accurately predict these parameters, with R^2 ranges from 0.74 to 0.84.

Ahangari, Daneshfar [41] used the hybrid model of PSO-LSSVM to predict the three geochemical parameters of TOC, S1 and S2 from well

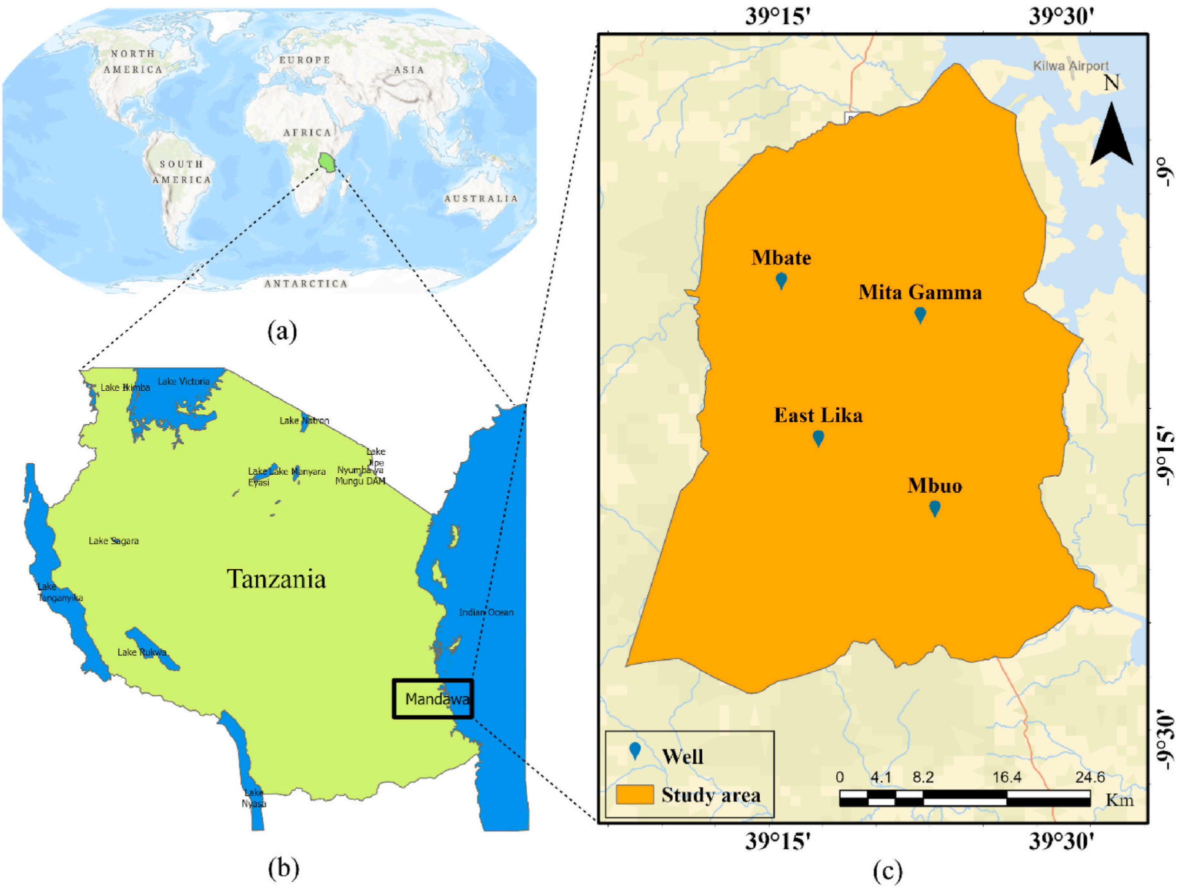

Fig. 1. Location of study area from (a) World view (b) Tanzania map (c) wells used in the study.

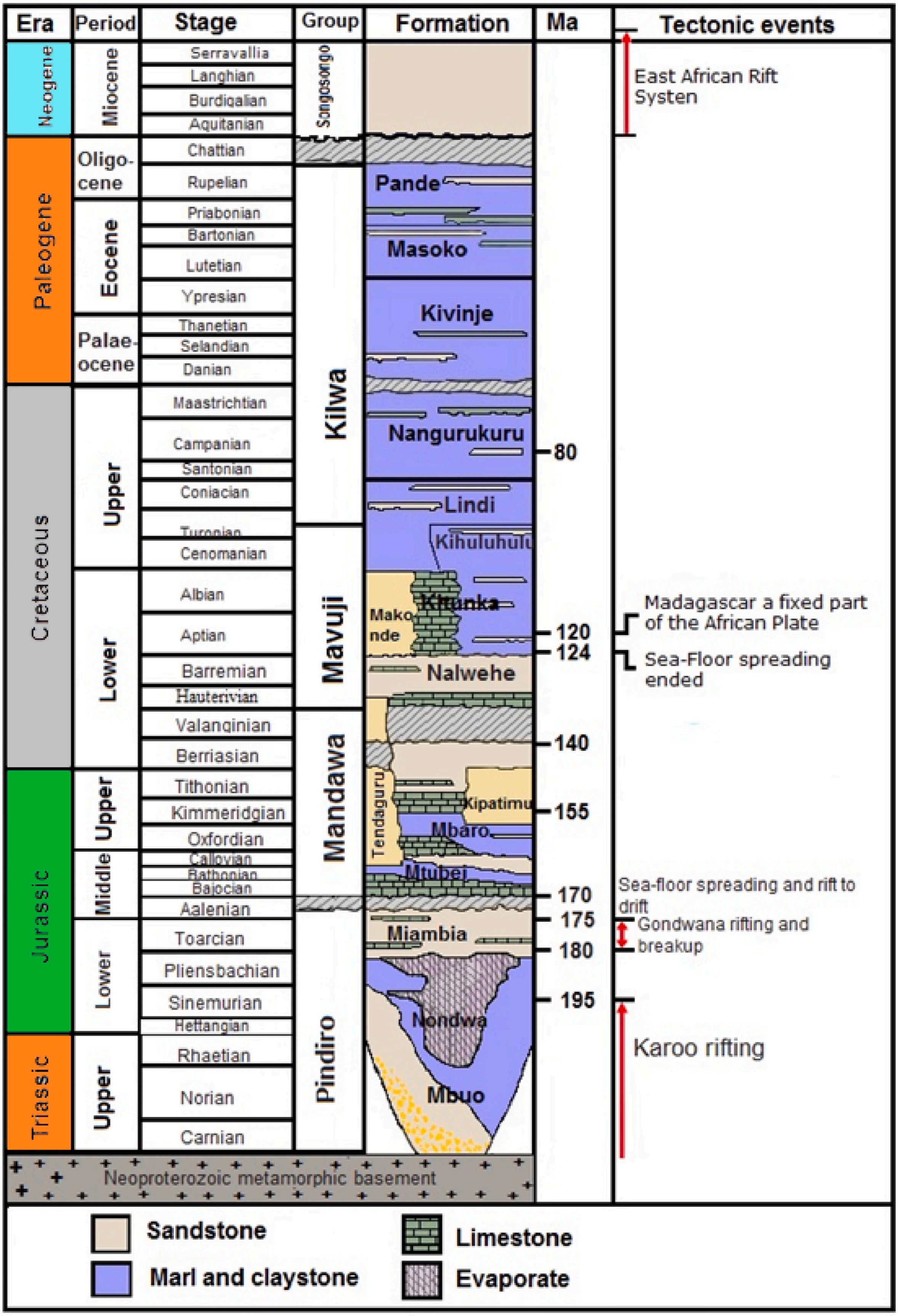

Fig. 2. Lithostratigraphy of the Mandawa basin adapted from Ref. [39].

Table 2
Statistical parameters of the input data used to build the model.

Well names	Statistical features	DT (μs/f)	GR (API)	LLD (ohm.m)	NPHI (%)	RHOB (g/cc)	SP (mV)
Mita Gamma	Minimum	59.31	33.18	1.52	3.12	2.23	52.48
	Maximum	127.73	88.65	22.15	37.79	2.47	80.33
	Average	95.63	62.61	6.29	19.24	2.37	66.71
	Standard Deviation	15.19	13.84	5.05	8.92	0.08	9.90
Mbate	Minimum	185.57	5.70	0.88	0.09	2.14	−31.06
	Maximum	439.10	84.08	6.33	0.42	2.58	−5.61
	Average	285.17	53.43	2.96	0.29	2.35	−23.16
	Standard Deviation	76.98	28.42	1.73	0.08	0.14	7.89
Mbuo	Minimum	217.33	57.95	0.68	0.14	1.78	−32.43
	Maximum	444.30	125.54	43.08	0.39	2.61	−2.56
	Average	307.75	85.38	8.35	0.28	2.39	−19.21
	Standard Deviation	54.46	16.79	10.94	0.07	0.21	7.39

logs data. Another study conducted by Ref. [42] revealed that BRNN outperformed the GPR, SVM, RF and LR in the prediction of Tmax and TOC from well logs. The results demonstrated that the machine learning techniques can capture the nonlinear relationship between the well-log data, TOC and Tmax, which may not be fully understood by existing linear models. Alizadeh, Maroufi [43] estimated the TOC and S2 by utilizing ANN model. Moreover, the TOC was successful predicted from well logs data by different researchers using SVM, ELM and kNN [44–46]. Barham, Ismail [47] developed FFNN model for estimation of Tmax and TOC from conventional well logs.

Although, most of these computational learning models attain the greatest performance but the user requires to specify regularization parameters and the optimum model parameters can be achieved through manual adjustment of training parameters. Similar, most of these models suffers some drawbacks including low computational speed, overfitting and converging at local minima. Thus, it is necessary to develop more advanced machine learning algorithms that can improve the accuracy of source rock potentiality and predictions while overcoming these limitations. Table 1 Summaries the limitations of previous machine learning techniques used to predict geochemical parameters. Based on the literature survey, it was observed that there have been limited studies that focus on the application of machine learning methods for predicting complete profile of geochemical results such as TOC, Tmax, S1, and S2. However, hydrocarbon source rock potentiality assurance of organic matter and geochemical analysis of sedimentary basin formations requires estimation of a complete profile of thermal maturity parameters of TOC, Tmax, S1, and S2. This makes the estimation of all geochemical parameters from geophysical well-log data and hydrocarbon source rock potentiality analysis a promising research area in the application of machine learning.

Therefore, the objective of this study was first to integrate a novel self-organizing neural network approach of generalized group method of data handling (g-GMDH), generated geochemical, wireline log and cuttings sample data to produce detailed hydrocarbon source rocks distributions in the sediment successions of the Mbuo, Mbate and Mita Gamma wells of the Mandawa Basin in South-East Tanzania. Then we utilized a novel g-GMDH neural network to predict continuous changes in geochemical indicators suite of TOC, Tmax, S1 and S2. The g-GMDH algorithm is a powerful data-driven approach for modeling complex nonlinear relationships between variables and dealing with multidimensional data. It is particularly useful because it can automatically select the most important features and interactions, which can reduce the risk of overfitting and enhance the model's performance for generalization. This algorithm works by using a divide-and-conquer strategy to build a series of nested models. At each level, the model selects the best input variables from the preceding layer, resulting in a highly flexible and adaptable approach that is well-suited for a wide range of data and modeling tasks. To the best of our knowledge, this is the first study to apply the g-GMDH method for the prediction of source rock potentiality, organic richness and maturity parameters based on well log data. The findings of this study will provide valuable insights into the application of machine learning techniques in the energy industry and will contribute to the development of more efficient and cost-effective exploration strategies.

2. Geological setting and stratigraphy of the study area

The Mandawa basin is located on the southern coast of Tanzania (Fig. 1b), bordered to the north by the Rufiji River and to the south by the Ruvuma Saddle. The basin's geological evolution has been studied by various scholars, including [48–50]. The Karoo rifting, Gondwana breakup, East African rift system, and opening of the Somali basin are the primary factors that influenced the Mandawa basin's evolution [51–53].

The Mandawa basin's depositional history was mainly influenced by the Gondwana breakup. Before the breakup of Gondwana, the depositional environment was continental with both deltaic and fluvial deposits dominating the area [54]. As the rifting and drifting developed, the Paleo-Tethys transgression led to the formation of restricted barrier reefs and marine embayment, isolating several saline lagoons during the early to middle Jurassic [55]. During the late Jurassic to early Cretaceous, the basin was exposed to rapid subsidence, leading to the deposition of clastic sediments, including the fluvial and alluvial deposits of the Mandawa and Mavuji groups. The constant decrease of the coastal Mandawa basin's mid-to-outer shelf zone during the Paleogene led to the development of the Kilwa group [56].

The Mandawa basin's source rock consists of Nondwa shales of the lower Jurassic Pindiro Group and Mbuo Claystone of the upper Triassic Pindiro Group [57]. The Mandawa Basin's sedimentary sequence ranges from Triassic to Neogene and it is controlled by coastal deposits to shallow marine shelf of carbonate, evaporitic and siliciclastic facies. The study area consists of Mbate, Mbuo, and Mita Gamma exploration wells (Fig. 1c).

The basin comprises five primary groups: Mandawa, Kilwa, Pindiro, Songosongo, and Mavuji (Fig. 2). Kilwa Group consists of four formations which are Masoko, Pande, Kivinje, and Nangurukuru. The group is made up of a uniform sedimentary package that consists series of clays, claystones, and marls, as well as abundant fossils such as benthic foraminifera and nummulites [58]. The Mbuo, Mihambia, and Nondwa Formations of Pindiro Group are the one that defines Karoo's sediments [59]. The Nondwa Formation, which has a peculiar border surface, covers the Mbuo Formation, which is the Pindiro Group's first sediments. Two sedimentary members make up the Mbuo Fm. The base of the Mbuo sandstone is where the transition from metamorphic rocks to clastic sediments took place, whilst the Nondwa formation's evaporates covered the upper part member of the Mbuo claystone [60]. Fluvial, alluvial and lacustrine are the environment in which the deposition of the Mbuo Formation took place [61].

The Mavuji group is made up of three formations which are Makonde, Kitiruka, and Kihuluhulu. The Kihuluhulu formation has the strongest exposed outcrops and is the easiest to reach compared to the other two formations. The Kihuluhulu Formation sits between the

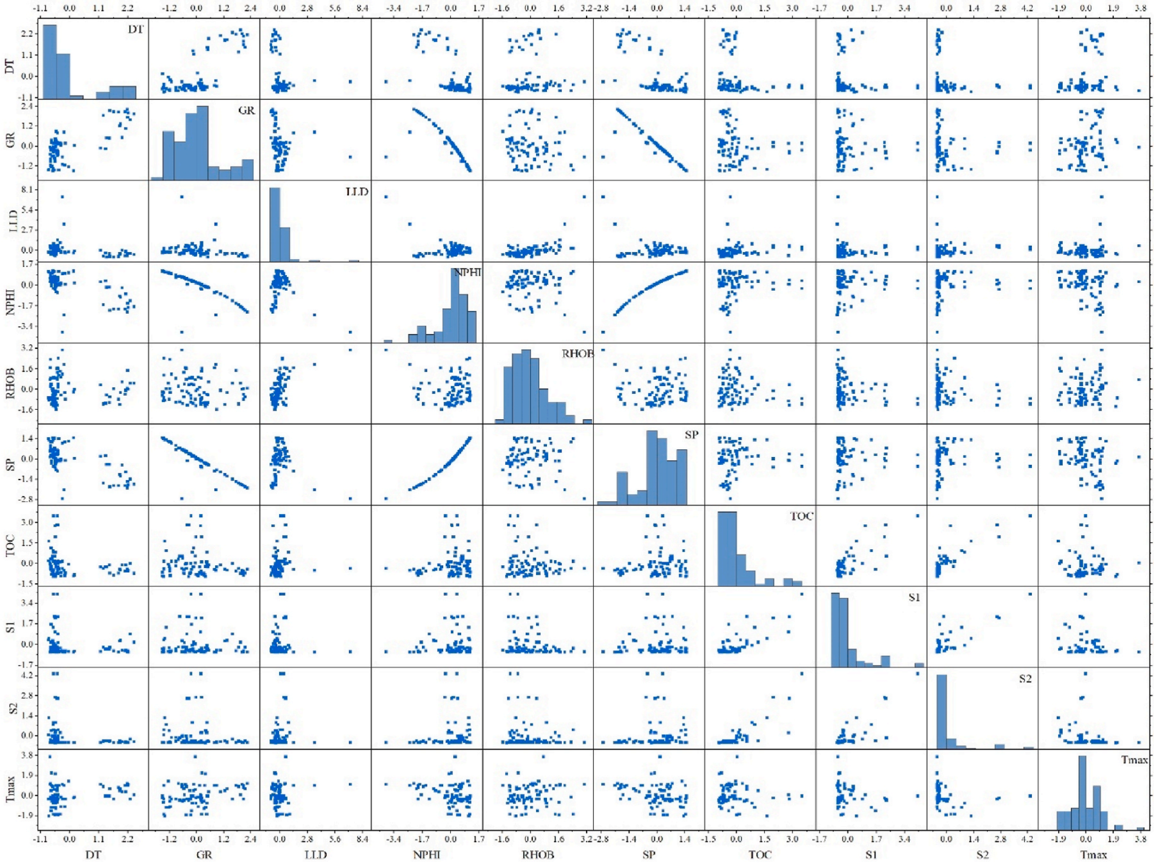

Fig. 3. Pair-plot of the data after power transformation with Box-Cox technique.

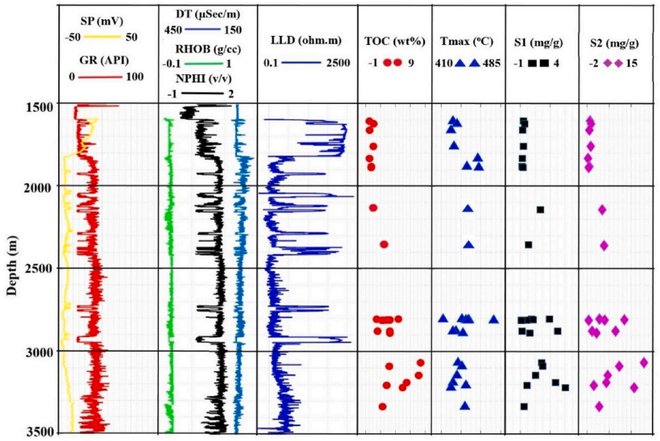

Fig. 4. Well logs and geochemical data for Mbuo well.

Kipatimu and Nangurukuru Formation, separated by the unconformity of Turonian-Santonian. The Kihuluhulu Formation fine-grained sandstones, siltstones, and mudstones from the shoreface to offshore filled the basin [62]. The Kilwa Group has been suggested to include the newly described Lindi Formation (upper Albian – Coniacian) [63].

3. Material and methods

3.1. Wireline logs acquisition and data handling

Samples and logs used were acquired from Mbate, Mita Gamma and Mbuo wells in an uncased wellbore with 0.15 m between each data point. According to TPDC standards, well log measurements suit of SP, DT, GR, NPHI, LLD and RHOB were processed and made available in Log ASCII Standard (LAS) format. The data in LAS format were then carefully handled to a spreadsheet (Excel®) with an interval corresponding to sample cuttings intervals. During data processing, feature selection (variable selection) was performed to identify and delete obsolete, unnecessary, and redundant data attributes that have a negative impact on a predictive model's accuracy or may minimize the model's accuracy. The statistical analysis of all the data from three wells is shown in Table 2 Source rock evaluation and interpretation of wire logs involved the utilization of both high resolution and low-resolution data.

To avoid overfitting and bias, the data were normalized by using Box-Cox transformations technique (Equation (1))

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{(\lambda)} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log\,(y) & \text{if } \lambda = 0 \end{cases} \tag{1}$$

Whereas y is a list of n strictly positive numbers. The selected technique enables the computational learning algorithm to execute faster, improves the model's accuracy, reduces the overfitting, and also it decreases the complexity of the model [64]. It's better to note that before doing normalization, the resistivity data were first log-transformed to a better distribution of data which improves prediction performance. After variable selection, and normalization the data was then distributed using pair-plot in Fig. 3.

Three wells namely Mbate, Mbuo, and Mita Gamma, which have a complete set of well log suites and measured core data of TOC, Tmax, S1, and S2 data, were employed to develop the machine learning models. 58 samples data from Mbuo and Mbate were used to train the models. The developed models were tested on the unseen Mita Gamma well which consisted of 25 sample data of TOC, Tmax, S1, and S2. Data were

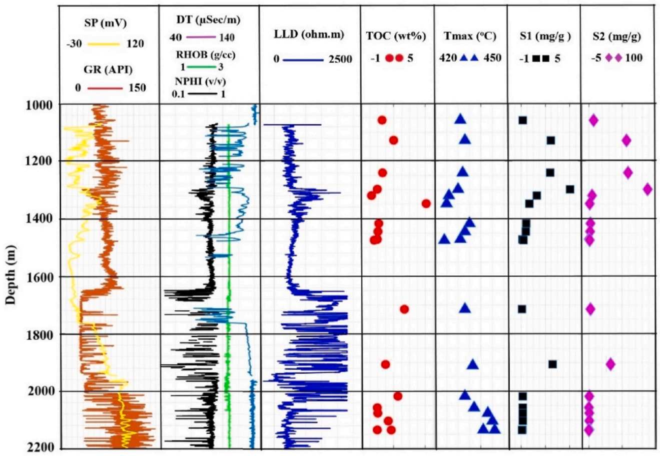

Fig. 5. Well logs and geochemical data for Mbate well.

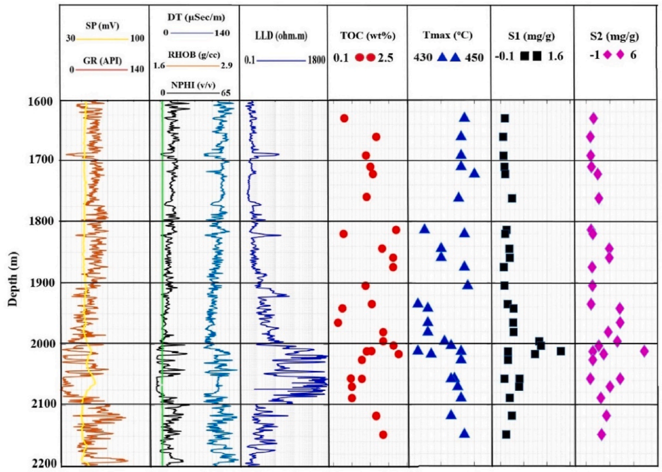

Fig. 6. Well logs and geochemical data for Mita Gamma well.

carefully split into training and testing to ensure a balanced dataset to avoid high variance whilst covering a range of lithological and maturity variations present in the basin. The core samples were taken from the depth intervals 1660–3335 m for Mbuo well, 1058–2135 m for Mbate well and for Mita Gamma well were from 1630 to 2150 m interval. The well log suite and geochemical results of Mbate, Mbuo, and Mita Gamma wells are illustrated in Fig. 4, Figs. 5 and 6.

3.2. *Rock-eval pyrolysis technique*

The Mandawa basin shows promise for hydrocarbon exploration. However, its development has been limited due to lack of comprehensive data and geological complexities, which contribute to uncertainties in geochemical parameters prediction and analysis. The study area consists of four exploration wells, which are Mbate, Mbuo, Mita Gamma, and East Lika. Mita Gamma, Mbate and Mbuo wells intersected both the Nondwa Formation and silicic limestones of the Mihambia Formation of the Mandawa basin whereas only East Lika well intersected the Nondwa sequences (Fig. 2). The rock-eval pyrolysis was performed to provide measured data as a reference point for thermal maturity parameters of TOC (wt. %), volatile hydrocarbon (S1 in mg HC (hydrocarbon)/g rock), hydrocarbon derived from kerogen pyrolysis (S2 in mg HC/g rock), the temperature at the highest yield of S2 (Tmax in °C). The rock-eval pyrolysis followed in this study were based on the methodology described by Ordoñez, Vogel [65]. The samples were then taken to a laboratory for analysis; 67 g of each sample was crushed, sieved into a powder form, and then extracted and analyzed. The Rock-Eval was run with a temperature schedule of 25 °C min^{-1}, where the final temperature in the pyrolysis oven exceeds 750 °C, and in the oxidation oven 800 °C.

3.3. *Gaussian process regression (GPR)*

Gaussian process regression (GPR) is a statistical machine learning technique that models an unknown function as a Gaussian process with a mean and covariance function. The mean and covariance function can be based on prior knowledge or not [66,67]. The covariance hyperparameters in a Gaussian process model are estimated from the data using a Type II maximum likelihood method. The data is first centered to assume a zero-mean function [68]. The output at new (test) inputs is then predicted by computing the predictive posterior distribution. More details about GPR are provided in supplementary file.

3.4. *Support Vector Machines (SVM)*

Support Vector Machine (SVM) is a powerful machine learning algorithm used for classification and regression analysis. It works by finding the hyperplane in a high-dimensional space that best separates the different classes [44]. The SVM algorithm finds the best hyperplane to separate data points of different classes with the least error. The hyperplane is chosen to maximize the distance between it and the nearest data points from each class [69]. The SVM algorithm can handle non-linearly separable data by utilizing a kernel function to map the input data into a higher-dimensional space. In this expanded space, the data points might become linearly separable, enabling the SVM algorithm to identify a decision boundary [70]. More explanation about SVM is found in supplementary file.

3.5. *Generalized structure of GMDH (g-GMDH) model*

The g-GMDH neural network is an advanced data analysis, prediction, and modeling tool, offering numerous advantages in a wide range of applications. Firstly, its robustness and adaptability allow it to effectively handle complex, nonlinear, and noisy data, making it suitable for diverse domains [71]. Secondly, the automatic model selection feature eliminates the need for predefined structures, iteratively generating and evaluating models to find the optimal fit while reducing overfitting risk. Additionally, g-GMDH's interpretability fosters a better understanding of relationships between input and output variables, aiding informed decision-making. Its scalability enables the efficient processing of large datasets through parallel processing capabilities, making it an excellent choice for big data applications and real-time analysis. Lastly, the network's reduced training time, attributed to its inductive nature and iterative model generation process, allows for faster model development and deployment, enhancing overall performance and efficiency.

The g-GMDH neural network is well-suited for predicting geochemical parameters because it can handle large and complex datasets, and it can automatically identify and extract relevant features from the data. By analyzing source rock potentiality, the g-GMDH neural network can learn patterns and relationships that can then be used to make accurate predictions. The goal for GMDH is to identify a function $\widehat{f}$ that is used as an estimation rather than an actual function, f, to evaluate the output (geochemical parameters), y, with a given input vector $U = (u_1, u_2, u_3, \ldots, u_n)$, as close to its actual output as possible, p (geochemical parameters). As a result, if you have a single output and n data pairs with numerous inputs, you'll get:

$$y_i = f(u_{i1}, u_{i2}, u_{i3}, \ldots., u_{in}) \quad (i = 1, 2, 3, \ldots., M) \tag{2}$$

The output t can be evaluated by GMDH using any given input vector $U = u_{i1}, u_{i2}, u_{i3}, \ldots., u_{in}$, implying:

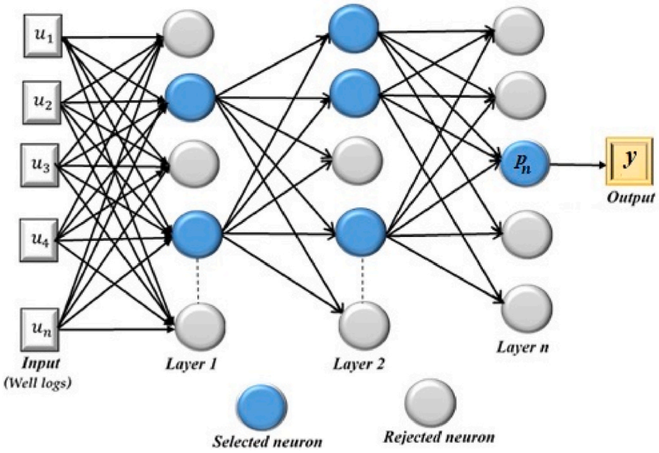

Fig. 7. Type of the GMDH network design.

$$y_i = \widehat{f}(u_{i1}, u_{i2}, u_{i3}, ..., u_{in}) \quad (i = 1, 2, 3, ..., M) \tag{3}$$

To tackle this difficulty, a general relationship is established by GMDH within the reference of the output and input parameters. The goal is to find the GMDH network that minimizes the square difference between the expected and actual output, as follows:

$$\sum_{i=1}^{M} [\widehat{f}(u_{i1}, u_{i2}, u_{i3}, ..., u_{in}) - p_i]^2 \rightarrow \min \tag{4}$$

The Kolmogorov-Gabor polynomial, also identified as the polynomial series, or the Volterra function in complex discrete form is known as the Volterra series [72,73], can serve to illustrate how input and output parameters generally relate to one another in the mode of:

$$p = a_o + \sum_{i=1}^{N} a_i u_i + \sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij} u_i u_j + \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} a_{ijk} u_i u_j u_k + ... \tag{5}$$

The GMDH network can simplify Equation (5) by using a partial quadratic polynomial equation [71].

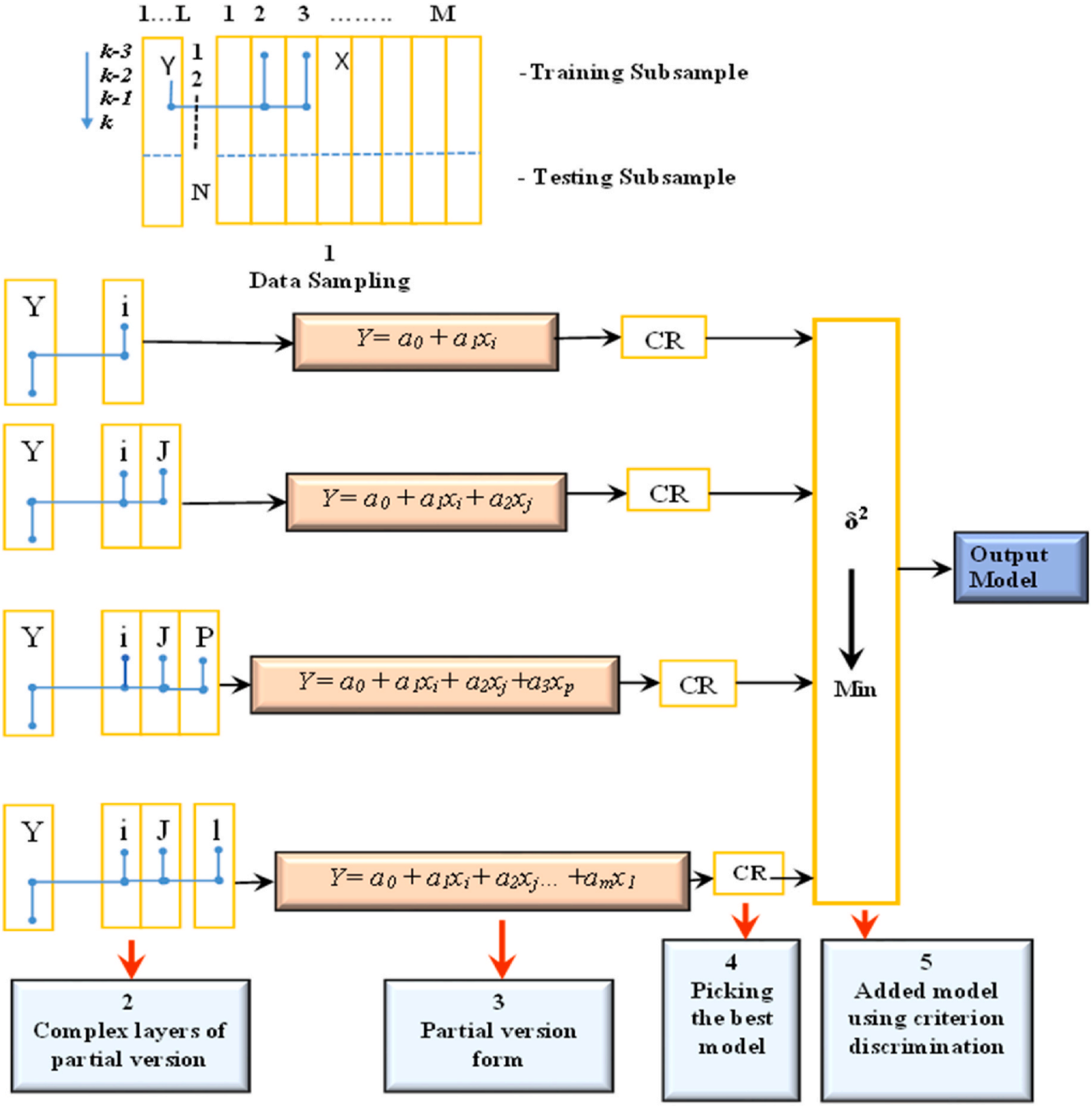

Fig. 8. Flowchart of the generalized structure of combinatorial GMDH algorithm.

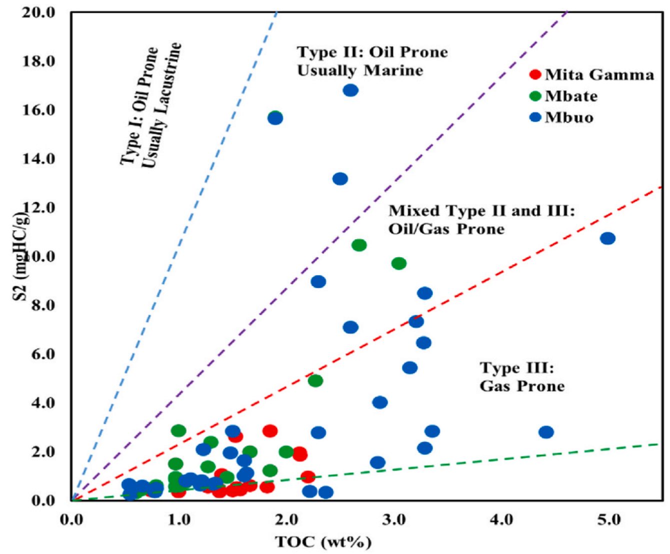

Fig. 9. Modified Van Krevelen plot indicating types of kerogens of the samples from Mandawa. (Hydrocarbon Potential (S2) versus pyrolysis TOC).

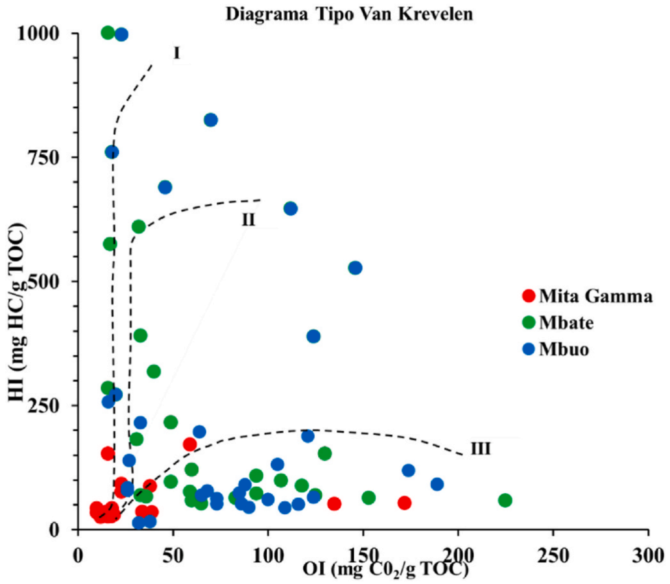

Fig. 10. Modified Van Krevelen diagram showing HI vs OI for kerogen classification for the Nondwa, Mihambia, and Mbuo Formations.

$$y = W(u_i, u_j) = a_o + a_1u_i + a_2u_j + a_3u_i^2 + a_4u_j^2 + a_5u_iu_j \tag{6}$$

This network of associated neurons generates the mathematical association between the input-output variables stated in Equation (4). The variance between the predicted *(y)* and actual *(p)* is reduced by computing the weighting in Equation (5) coefficients using regression techniques as each pair of the inputs (u_i and u_j) is minimised [74]. Fig. 7 depicts the GMDH network design in a schematic form.

A tree of polynomials is formed using the quadratic equation from the provided equation (5), where the least squares method can be used to compute the weighting coefficients. For each pair of output-input data, a weighting coefficients W_i of quadratic function is derived as follows:

$$E = \frac{\sum_{i=1}^{M} (p_i - W_i())^2}{M} \rightarrow \min \tag{7}$$

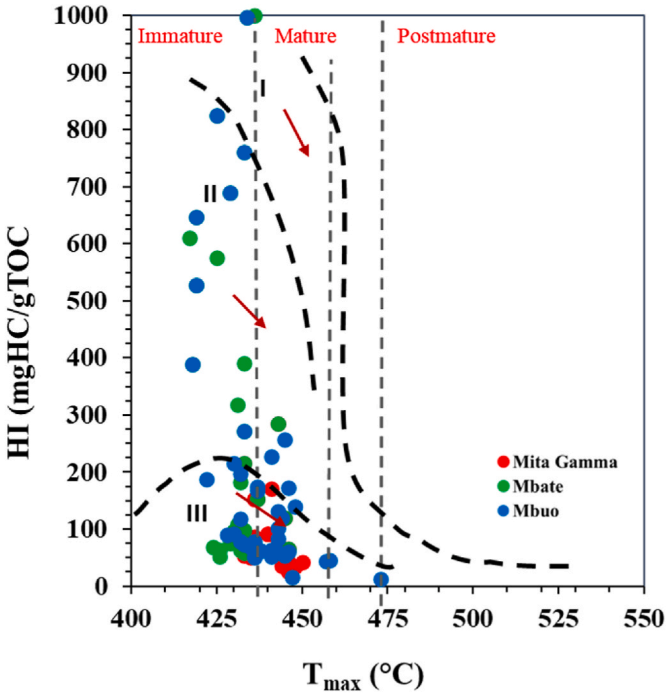

Fig. 11. Modified Van Krevelen plot showing thermal maturity and kerogen types of Hydrogen Index (mg HC/g TOC) vs pyrolysis Tmax (°C).

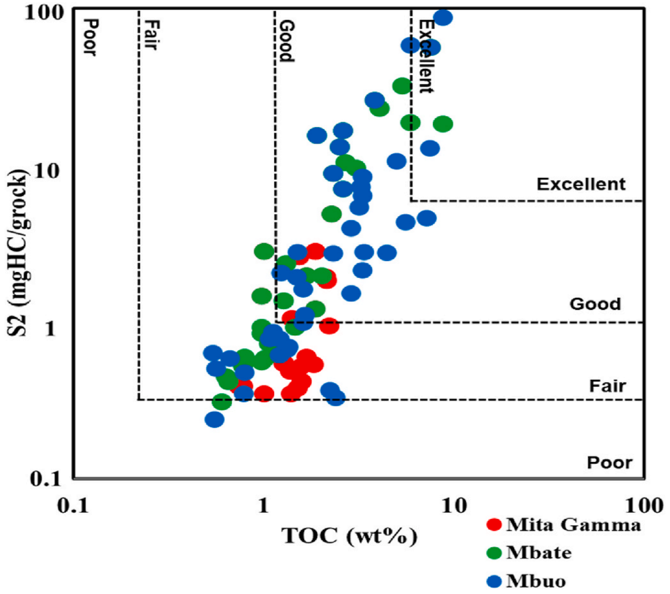

Fig. 12. Source rock characterization based on genetic potential and richness of the analyzed samples.

In order to match the general form of GMDH algorithms, the possibility for dual independent variables among the total n input parameters is drawn. Then construction of the regression polynomial is described in equation (7), which in the sense of least-square suits the dependent observations better. Consequently $C_n^2 = n(n-1)/2$, observations can create quadratic polynomial neurons $\{(p_i; u_{xi}, u_{yi}); (i = 1, 2, 3....M)\}$ for various $x, y \in \{1, 2, 3, ...n\}$ in the first layer of feed-forward network. The triples of M data can be created $\{(p_i; u_{xi}, u_{yi}); (i = 1, 2, 3....M)\}$, making use of $x, y \in \{1, 2, 3, ...n\}$ the form of:

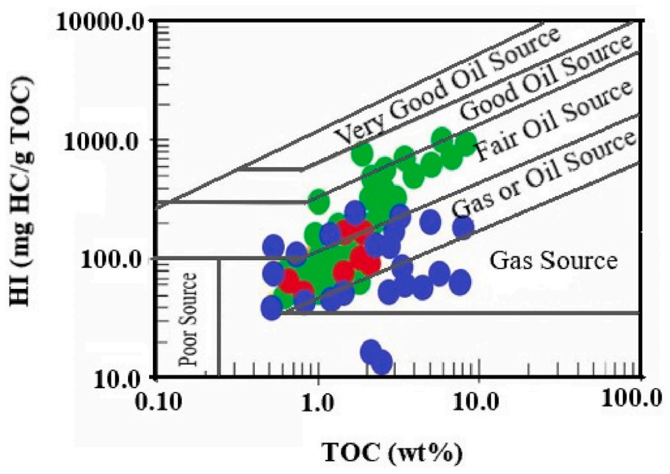

Fig. 13. Source rock characterization for the type of hydrocarbon to be generated for analyzed samples from Mandawa.

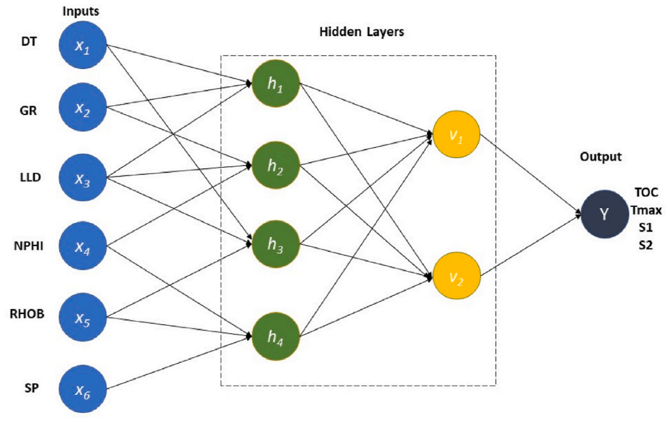

Fig. 14. g-GMDH model structure.

Table 3
Hyperparameter setting for g-GMDH model.

Hyperparameters	TOC	Tmax	S1	S2
Population size	100	50	80	50
Mutation rate	0.2	0.1	0.2	0.2
Crossover rate	0.7	0.6	0.8	0.7
Number of hidden layers	2	2	2	2
Number of neurons in each layer	20	12	16	10
Stopping criterion	200 iterations	100 iterations	150 iterations	200 iterations
Selection rate	0.6	0.5	0.6	0.6

Table 4
Error performance results in TOC prediction during training and testing.

Model	MSE		MAE	
	Train	Test	Train	Test
g-GMDH	0.1137	0.1841	0.2019	0.4503
GPR	0.1686	0.3399	0.3005	0.6010
SVM	0.2671	0.5033	0.3492	0.6990

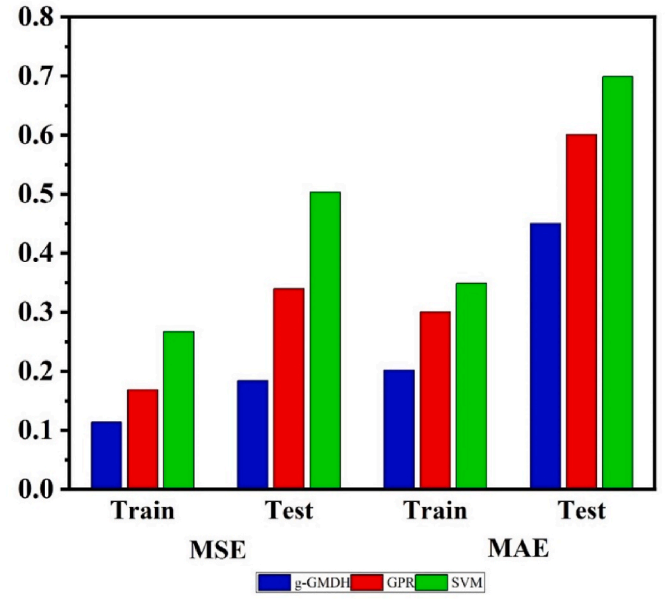

Fig. 15. Performance of the g-GMDH, GPR and SVM models during training and testing in TOC prediction.

$$\begin{bmatrix} u_{1x} & u_{1y} & : & p_1 \\ u_{2x} & x_{2y} & : & p_2 \\ \cdots & \cdots & : & \cdots \\ u_{mx} & u_{my} & : & p_m \end{bmatrix} \tag{8}$$

It is possible to express the following matrix expression directly using the quadratic sub-expression shown in Equation (8) as follows:

$$Aa = P \tag{9}$$

A vector of unknown weighting factors in quadratic polynomial is illustrated as *a* in Equation (10):

$$a = [a_0, a_{,1} a_2, a_3, a_4, a_5]^T \tag{10}$$

Superscript, T, indicating the transposition of matrix:

$$P = [p_1, p_2, p_3, \ldots p_M]^T \tag{11}$$

Standard equations are solved using an approach known as least-square, which was created using the multiple regression analysis idea, which is in the form of:

$$a = \left(A^{TA}\right)^{-1} A^T P \tag{12}$$

Equation (12) signifies the optimum quadratic weighting coefficients given the vector in Equation (2). To tackle the issue of linear dependency and equation complexity, the g-GMDH can build a higher-order polynomial [75]. By minimizing the fitness function in Equation (13), In order to build the best model structure, the sub-samples are utilized during training:

$$\mathrm{AR}(s) = \frac{1}{\mathrm{N_B}} \sum_{i=1}^{\mathrm{N_B}} (\mathrm{z_i} - \mathrm{z_i}(\mathrm{B}))^2 \tag{13}$$

The fitness function is based on the d-fold cross-validation criteria, which randomly selects training and testing subsamples considering all information in data samples. A comprehensive search is conducted on models classified of similar complexity, allowing the entire search termination rule to be planned. The models are compared to the measured samples, and the process is repeated till the criterion is reached. Because of the constraint imposed by the computation time, it is suggested to increase the number of parameters in the model after a specific number of iterations, as this will improve the model's

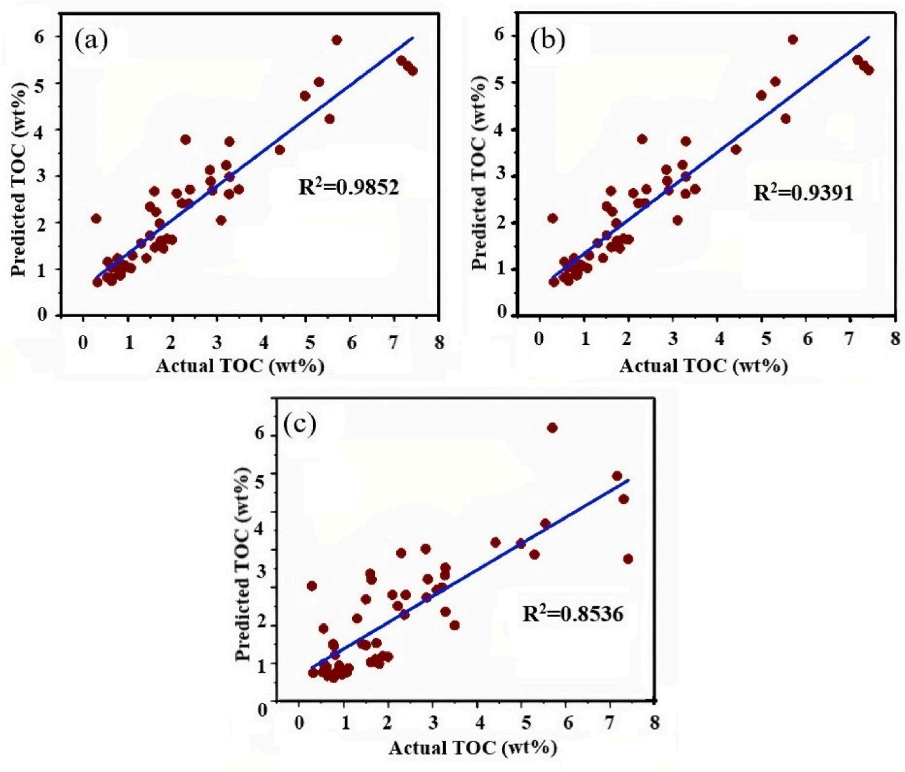

Fig. 16. Cross plots between actual and prediction TOC values using (a) g-GMDH, (b) GPR and (c) SVM models during training.

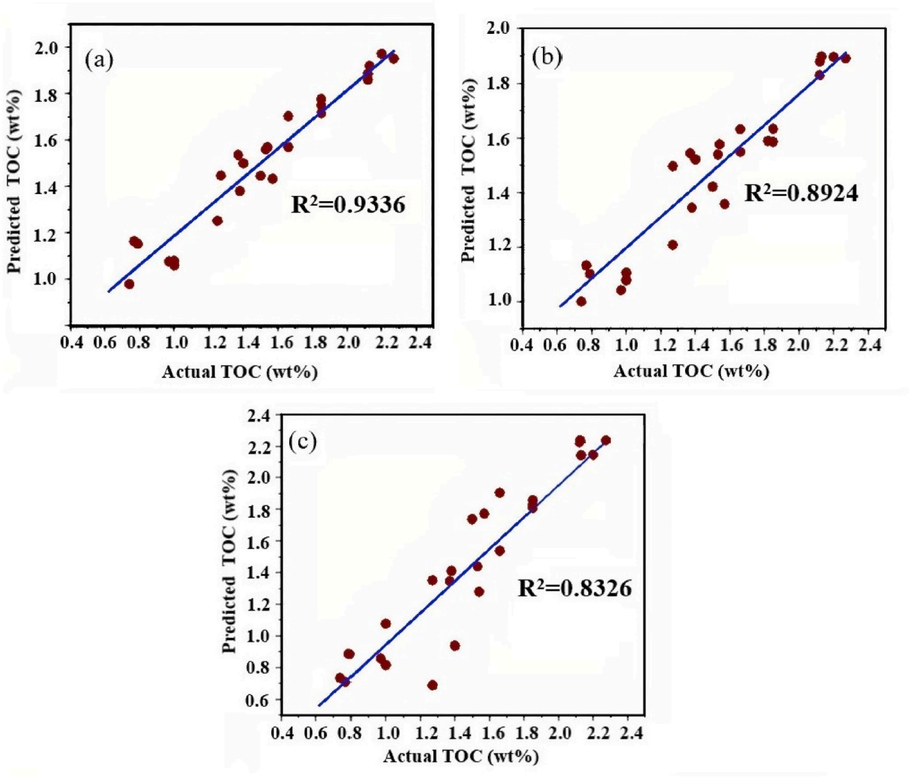

Fig. 17. Cross plots between actual and prediction TOC values using (a) g-GMDH, (b) GPR and (c) SVM models during testing.

performance. Then, in the set of best-chosen variables, complete arranging approaches are used until progress is minimal. This gives the option of including more input information and saving successful elements between layers to get the best model. In the first phase, the user defines the data sample for the model. Several layers are used to express model complexity during the second step. In the third step, the best

Table 5
Error performance results in Tmax prediction during training and testing.

Model	MSE		MAE	
	Train	Test	Train	Test
g-GMDH	3.38425	2.35512	1.02119	1.37521
GPR	11.23226	6.29555	2.23773	1.35261
SVM	17.02091	9.14027	3.23219	2.56066

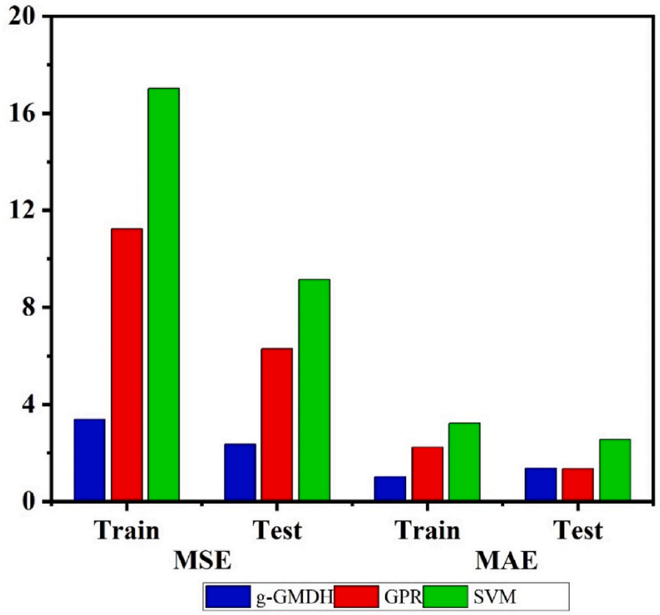

Fig. 18. Performance of the g-GMDH, GPR and SVM models during training and testing in Tmax prediction.

models are created, and the best model is chosen in the fourth step. During the fifth stage, discriminating criteria are used to complete the extra model definition, as shown in Fig. 8.

4. Results and discussion

4.1. Geochemical evaluation of Mandawa source rocks

This study examined the quality and quantity of organic matter in Triassic to Jurassic source rocks from selected Mandawa basin wells, along with their potential to generate oil and gas. The results revealed that the Nondwa Formation has relatively high TOC values, ranging between 0.6 and 8.7 wt%, whereas the TOC content in the Mbuo Formation vary from 0.5 to 7.4 wt%. Compared to all evaluated stratigraphic sections, the Mihambia rocks showed the lowest TOC content ranging from 0.7 to 2.1 wt%. S2 is widely used to identify types of organic matter during Rock-Eval pyrolysis by assessing the hydrocarbon-generating potential [76]. The Mbuo formation showed S2 values ranging from 0.2 to 13 mg HC/g rock, Mihambia and Nondwa Formations had S2 values ranging from 0.39 to 2.83 mg HC/g rock and 0.31–88.52 mg HC/g rock respectively.

The indicator or measure of the hydrocarbon's type (oil or gas) to be generated is termed as the hydrogen index (HI) and has values between 51 and 1000 mg HC/g TOC. The oxygen index (OI) varies from 16 to 225 mg/g. At a maximum of S2, the temperature is termed as Tmax which is a measure of source rock maturity. Tmax exhibited temperature values that varies from 417 to 473 °C, suggesting maturation differences from immature to mature source rocks. The geochemical results from this study provide valuable insights into the hydrocarbon-generating potential of Triassic to Jurassic source rocks in the Mandawa basin. The Nondwa Formation and Mbuo Formation show promising characteristics for oil and gas generation, while the Mihambia rocks exhibit lower potential. The findings from this research contribute to our understanding of the petroleum system in the Mandawa basin and can assist in future exploration and production efforts in the region.

4.1.1. Quality of organic matter

The determination of kerogen type for a given source rock is critical for estimating oil and gas potential. The Hydrocarbon Potential distribution namely TOC and S2 indicators were used to assess the quality of the formation units on kerogen types. Followed Clayton and Ryder [77], the TOC versus S2 cross plot was used with the support of kerogen type characterization using the Oxygen Index (OI) and Hydrogen Index (HI) technique, as described by Ref. [78].

In analyzing the source rock, the kind of organic matter (kerogen) is regarded the second most essential criterion. Physicochemical approach may also be used to distinguish the kerogen type. The differences in organic matter are due to its original composition. The organic matter in prospective source rocks must be of a kind that may produce petroleum. Waples [79] determined that organic matter is divided into three categories. Many researchers such as, Cruz Luque and Aguilera [80] modified Van Krevelen diagram to present three different types of organic matter, the graph of hydrogen index (HI) versus the oxygen index (OI). The authors explained that type II-III combined kerogen with HI values between 200 and 350 mg HC/g TOC and/or S2/S3 values within 5 and 10 is expected to create both oil and gas. Type II ($350 < HI < 700$ and/or $10 < S2/S3 < 15$) and I ($HI > 700$ and/or $S2/S3 > 15$) kerogen may produce liquid hydrocarbons and is typically generated from marine and lacustrine organic matter. On the other hand, samples containing HI values less than 50 mg HC/g TOC and/or S2/S3 ratios less than 1 consists of inert components with no capacity to generate hydrocarbons.

According to the results, Mbate, Mbuo, and Mita Gamma wells intersected most of the Lower Jurassic/Triassic Mbuo Formation source rocks are kerogen Type III, with a few samples confirming mixed kerogen Types II/III (Fig. 9). Variations in TOC generated by either oxidation or CO_2 addition to the system might create the observed scattered points. mixed terrestrial/marginal and Terrestrial marine depositional environments are characterized by kerogen Type III and mixed II/III, respectively. The existence of Type I, II, and III is also shown in the modified Van Krevelen diagram utilizing Oxygen Index (OI) and Hydrogen Index (HI), with most of the samples suggesting Type II and Type III. Fig. 10 illustrates that Types II and III account for the bulk of the samples, whereas Types I are insignificant.

The majority of the Lower Jurassic Nondwa Formation data plots in the gas-prone Type III, grading to mixed Types II/III, Type II and Type I, which usually are of terrestrial origin. A few samples scatter in the mixed Types II/III fields (Fig. 9). These samples indicate terrestrially derived organic matters. On the other hand, the organic matter contents in the Mihambia rocks are mainly composed of gas-prone Type III and inert gas (Figs. 9 and 10).

4.1.2. Thermal maturity

In this study, the pyrolysis technique was used to determine the thermal maturity degree of the source rocks; data for Tmax (^{0}C), and PI, were computed as described by Espitalié, Deroo [78]. With small outliers in the immature and condensate zones, the plot of HI vs Tmax demonstrates that the Mita gamma and Mbuo wells of the Mbuo and Mihambia Formations source rocks are inside the mature zone of the oil window (Fig. 11). This suggests that these formations have reached an advanced stage of thermal maturity, indicating their potential for hydrocarbon generation. Mbate well of the Nondwa Formation source rocks, on the other hand, are seen to be in the immature zone, grading into the mature oil window (Fig. 11). This implies that the Nondwa Formation is in an early stage of maturation and is gradually progressing towards the point of oil generation. The Tmax maturation range is mainly influenced by the type of kerogen. Because of the intricacy of

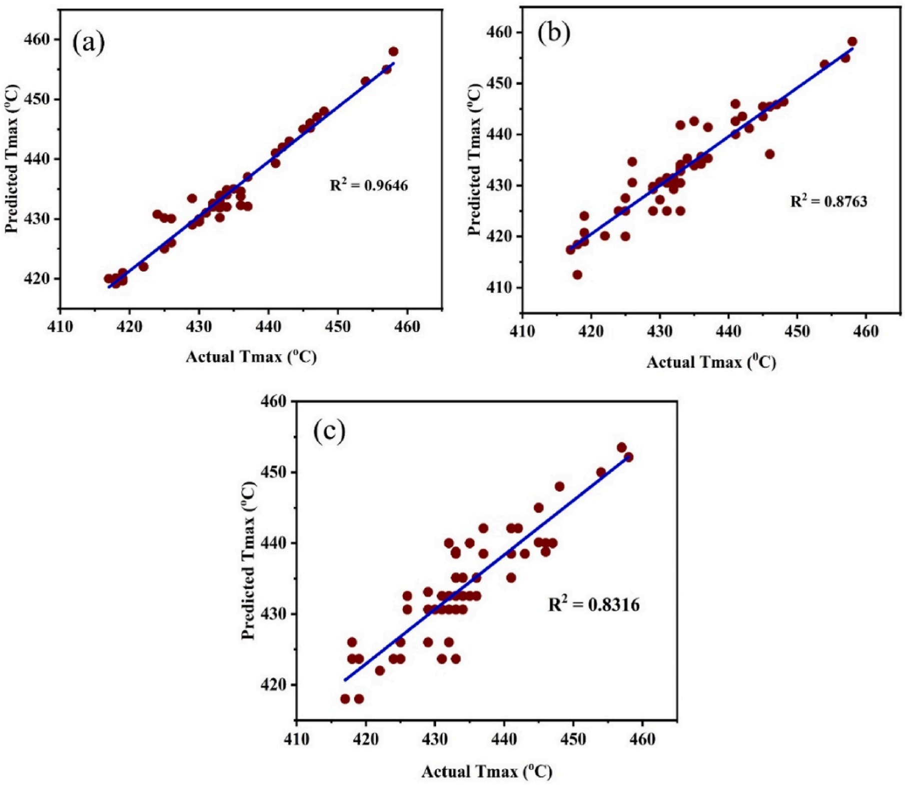

Fig. 19. Cross plots between actual and prediction Tmax values using (a) g-GMDH, (b) GPR and (c) SVM models during training.

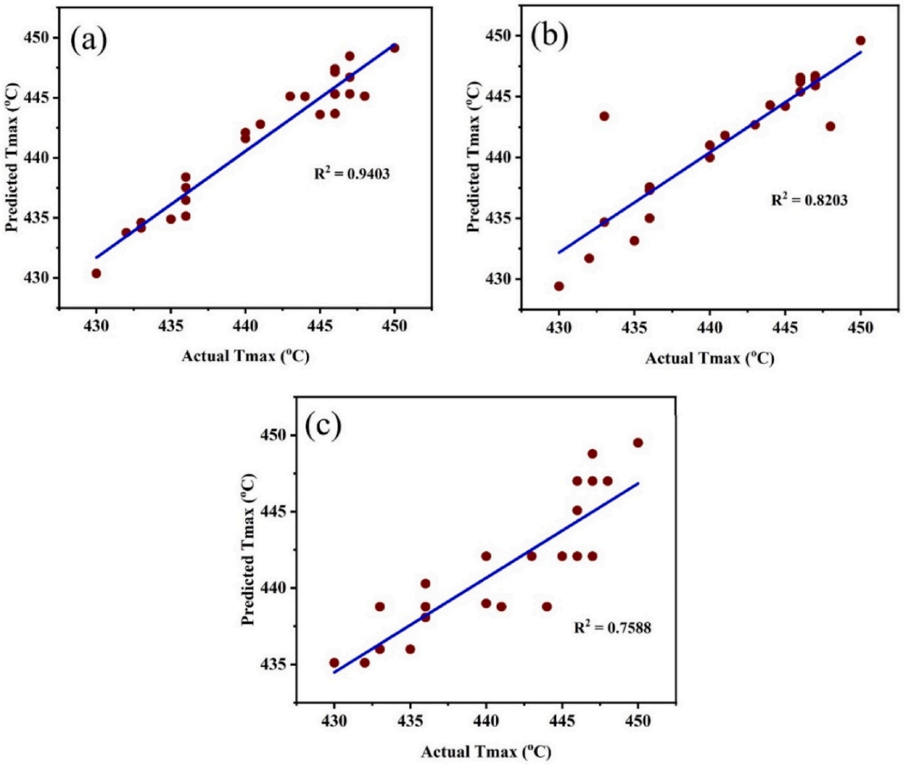

Fig. 20. Cross plots between actual and prediction Tmax values using (a) g-GMDH, (b) GPR and (c) SVM models during testing.

Type II kerogens' molecular structure, source rocks containing Type I kerogens have a small range, but Type II kerogens have a greater range [81]. As a result, the study findings highlight the significance of kerogen type in influencing the maturation process of source rocks. Tmax values for source rocks from the Lower Jurassic Mbuo Formation (Mbuo well) vary from 422 to 473 °C (mean = 440 °C) in this analysis, suggesting a

Table 6
Error performance results in S1 prediction during training and testing.

Model	MSE		MAE	
	Train	Test	Train	Test
g-GMDH	0.00511	0.08563	0.03725	0.17526
GPR	0.06057	0.15598	0.15955	0.20868
SVM	0.07486	0.38414	0.19644	0.39439

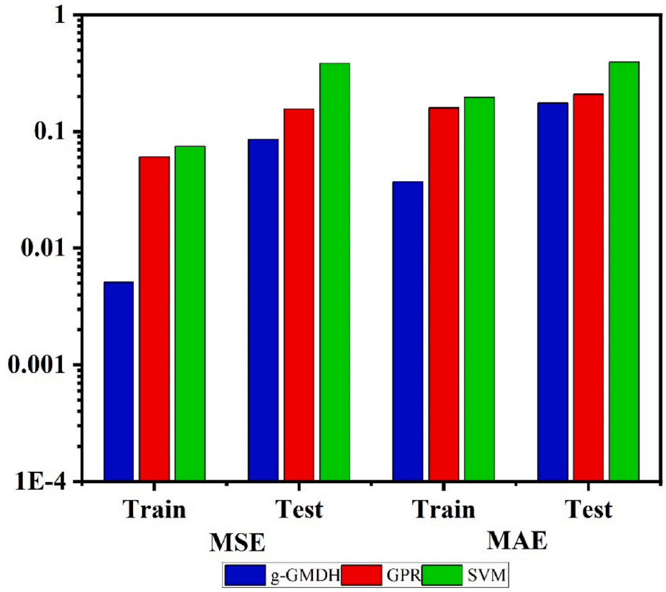

Fig. 21. Performance of the g-GMDH, GPR and SVM models during training and testing in S1 prediction.

mature source rock with significant hydrocarbon potential. Similarly, the Mid-Jurassic Mihambia Formation (Mita gamma well) exhibits an average Tmax value of 437 °C, suggesting a mature level of kerogen, further supporting its suitability as a potential source of hydrocarbons. On the other hand, the Lower Jurassic Nondwa Formation (Mbate well) displays the lowest Tmax values ranging from 417 to 446 °C (mean = 431 °C), which signifies an immature to early mature grade. Although it is not fully matured yet, it still shows potential for oil generation as it approaches the mature stage.

4.1.3. Source rock generation potential

According to Hunt [82], source rocks with generating potential (GP) (mg HC/g rock) < 2; 2 to 5; 5 to 10 and > 10 are considered to present poor, fair, good, and very good generation potential, respectively. The generation potential versus TOC cross plot (Fig. 12) is applied with the backing of the HI versus TOC cross plot of Jackson, Powell [83] to interpret generation potential with respect to source rock's kerogen type and maturity.

The important information required in the initial exploration stages is the presence or absence of effective source rock. By analyzing the results from cross plots, the hydrocarbon generation potential varies with different kerogen types and maturity levels. The GP versus TOC cross plot is a useful tool for assessing the hydrocarbon potential of source rocks of the Mandawa basin. the source rocks with high TOC content and relatively low GP, indicate a high quantity of organic matter, but with poor hydrocarbon-generating potential. On the other hand, source rocks with both high TOC and high GP are indicative of excellent hydrocarbon generation potential.

The Mbuo Formation with respect to current maturity is indicated to have poor to very good gas generation potential; averages: TOC = 2.41 wt%, S2 = 2.79 mg HC/g rock, HI = 102.44 mg HC/g TOC) and the source rock is more likely to produce the gas at peak maturity at these potential ratings as well. The TOC, S1, and S2 data further suggest gas generation potential for Mbuo. On the other hand, the data for the Nondwa Formation suggest the existence of fair, very good, and excellent source rocks trends for oil and gas generation. However, some isolated zones are indicated to be poor sources (Fig. 12; averages: TOC = 2.35 wt%, S2 = 12.28 mg HC/g rock, HI = 312.28 mg HC/g TOC).

Despite a limited number of data, the source rocks in Mihambia Formation are suggested to be fair to a good source of kerogen Type III, and as such, they have the potential for gas generation (Figs. 11 and 13). The mean TOC, S2 and HI values for this formation are 1.51 wt%, 1.58 mg HC/g rock, and 97.57 mg HC/g TOC, respectively. Their HI and TOC cross-plot in Fig. 13 provides more evidence for these findings.

Overall, the assessment suggests that the Mbuo Formation is more likely to produce gas at peak maturity, given its gas-prone nature with moderate to high HI values and relatively low S2 values. On the other hand, the Nondwa Formation exhibits a more complex hydrocarbon generation potential, with indications of both oil and gas-prone source rocks. The presence of fair, very good, and excellent source rock trends within the Nondwa Formation suggests a broader range of hydrocarbon possibilities.

4.2. Hydrocarbon potential based on machine learning

4.2.1. g-GMDH model development

The proposed g-GMDH model consisted of six input neurons and two hidden layers with four neurons, h_1, h_2, h_3, and h_4, for first layer and two neurons, v_1 and v_2 for second layer. The output of the model was represented as y. The model was coded in MATLAB R2021a. Fig. 14 presents a neural network structure for the proposed model while the hyperparameters setting which produce the best results for the model are presented in Table 3.

4.2.2. Performance indicators

In this study, we implemented g-GMDH, GPR, and SVM models in MATLAB R2021a. The models were run on window 11 operating system with 2.8 GHz Intel Core i7-1065G7 processor. The mean absolute error (MAE), coefficient of determination (R^2) and mean square error (MSE) were the statistical measures used to evaluate the selected models' performance. R^2 measures the strength and direction of the linear relationship of the selected model variables, MSE measures the relative average square of the errors and represents the stability or quality of the models while MAE describes the errors model in terms of expressing the same phenomenon between paired observations. The R^2, MSE, and MAE mathematical expressions are given in supplementary file.

4.2.3. TOC prediction

The results of the prediction of total organic carbon (TOC) from the training and testing data using three different models was presented in Table 4. For the training data, the g-GMDH model achieved the lowest MAE and MSE values of 0.2019 and 0.1137, respectively. The GPR model also performed relatively well, with MSE values of 0.1686. The SVM model had the highest MSE value of 0.2671. For the testing data, the g-GMDH model achieved the lowest MAE and MSE values of 0.4503 and 0.1841, respectively. The GPR model had MAE and MSE values of 0.6010 and 0.3399, respectively. The SVM model had the highest MSE value of 0.5033 and MAE value of 0.6990. As shown in Fig. 15, findings show that the g-GMDH model had the best overall performance for predicting TOC from the training and testing data. The GPR models also achieved relatively good performance, while the SVM model had the poorest performance. Furthermore, the results from the testing data suggest that the g-GMDH model was more robust and had better predictive accuracy compared to the other models.

Due to the ability to get uncertainty of the anticipated value, the models had a high level of accuracy, as the R^2 values of the training and testing datasets were above 0.8. The g-GMDH model had the highest

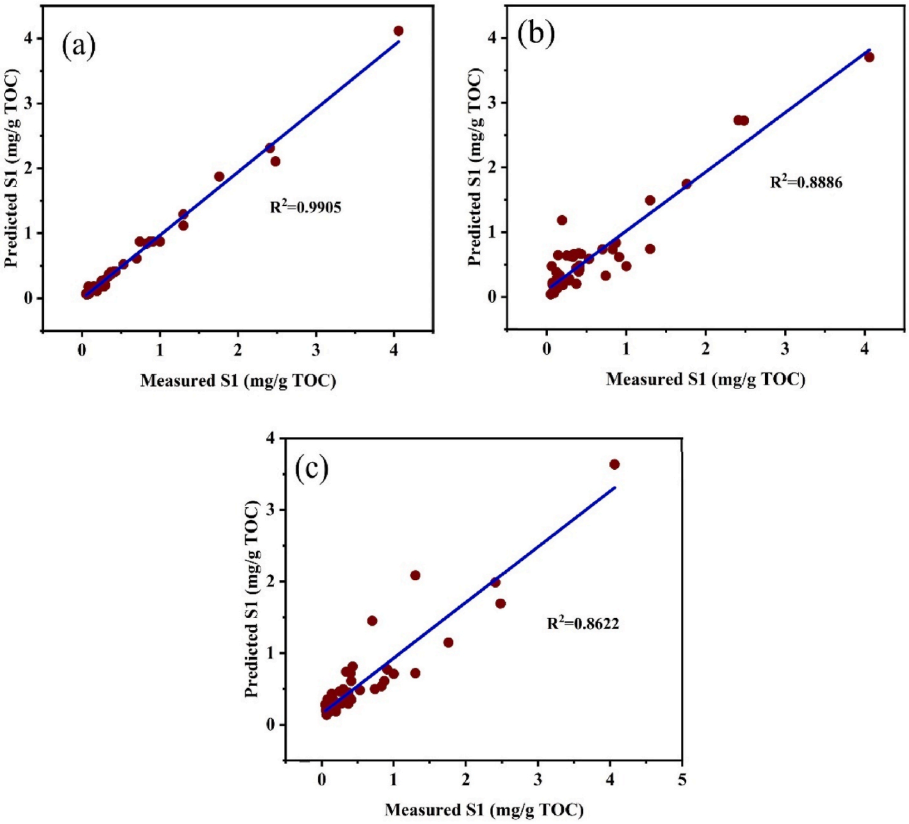

Fig. 22. Cross plots between actual and prediction S1 values using (a) g-GMDH, (b) GPR and (c) SVM models during training.

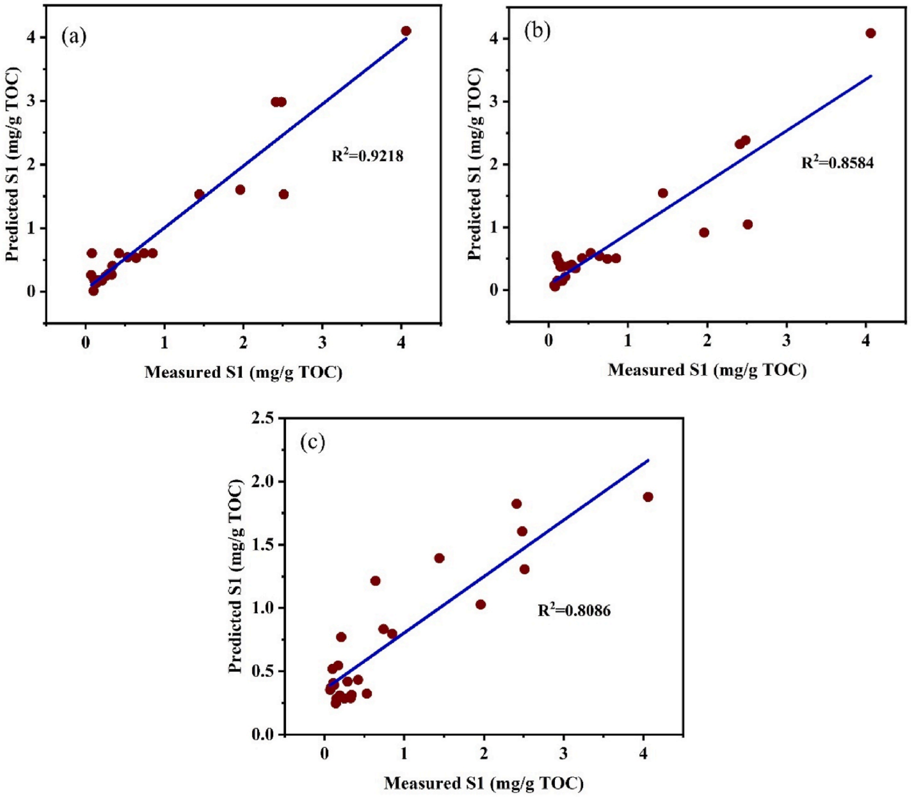

Fig. 23. Cross plots between actual and prediction S1 values using (a) g-GMDH, (b) GPR and (c) SVM models during testing.

Table 7
Error performance results in S2 prediction during training and testing.

Model	MSE		MAE	
	Train	Test	Train	Test
g-GMDH	11.55488	61.74456	1.27891	4.1445
GPR	45.47719	93.79075	3.76621	5.0137
SVM	51.76251	131.97055	4.90752	6.47348

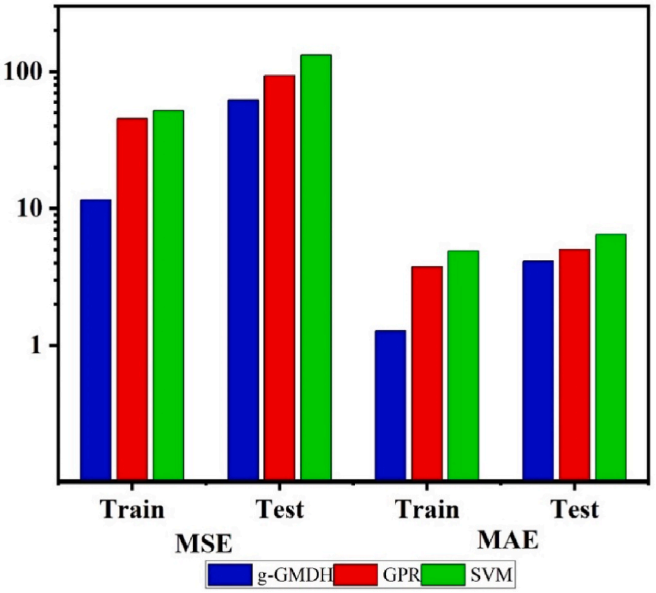

Fig. 24. Performance of the g-GMDH, GPR and SVM models during training and testing in S2 prediction.

performance, with an R^2 value of 0.9852 for the training dataset (Fig. 16) and 0.9336 for the testing dataset (Fig. 17). The GPR model had the second-highest performance, with an R^2 value of 0.9391 for the training dataset and 0.8924 for the testing dataset. The SVM model had the lowest performance, with an R^2 value of 0.8536 for the training dataset and 0.8326 for the testing dataset. Overall, the three models were able to accurately predict the TOC with high accuracy, which indicates that these models can be used to successfully predict the TOC.

The robustness and predictive accuracy of the g-GMDH model, particularly on the testing data, indicate that it is a suitable and reliable approach for predicting TOC in the studied dataset. The results also highlight the importance of choosing an appropriate modeling technique when predicting complex geological parameters like TOC, and the g-GMDH model seems to be a promising choice for this particular application. One of the key findings of the study is that the TOC content in the basin varies significantly, with some intervals having high TOC values that indicate good petroleum potential. This observation has important implications for petroleum exploration and production in Tanzania, as it suggests that identifying intervals with high TOC values can help target areas with good petroleum potential.

4.2.4. *Tmax prediction*

Table 5 presented the results of models obtained during the prediction of Tmax in training and testing phases. The g-GMDH was the most effective model among the other two models with the lowest MSE and MAE values both for training and testing data as shown in Fig. 18. The MSE and MAE values for the training data were 3.38425 and 1.02119, respectively for this model. Similarly, for the test data, the MSE and MAE values were 2.35512 and 1.37521 respectively. It can be seen that GPR had a mean squared error (MSE) of 11.23226 for training and 6.29555 for testing, and a mean absolute error (MAE) of 2.23773 for training and 1.35261 for testing. On the other hand, SVM had an MSE of 17.02091 for training and 9.14027 for testing, and an MAE of 3.23219 for training and 2.56066 for testing. Overall, g-GMDH had a lower MSE and MAE than other models, indicating that it is more accurate in predicting the Tmax.

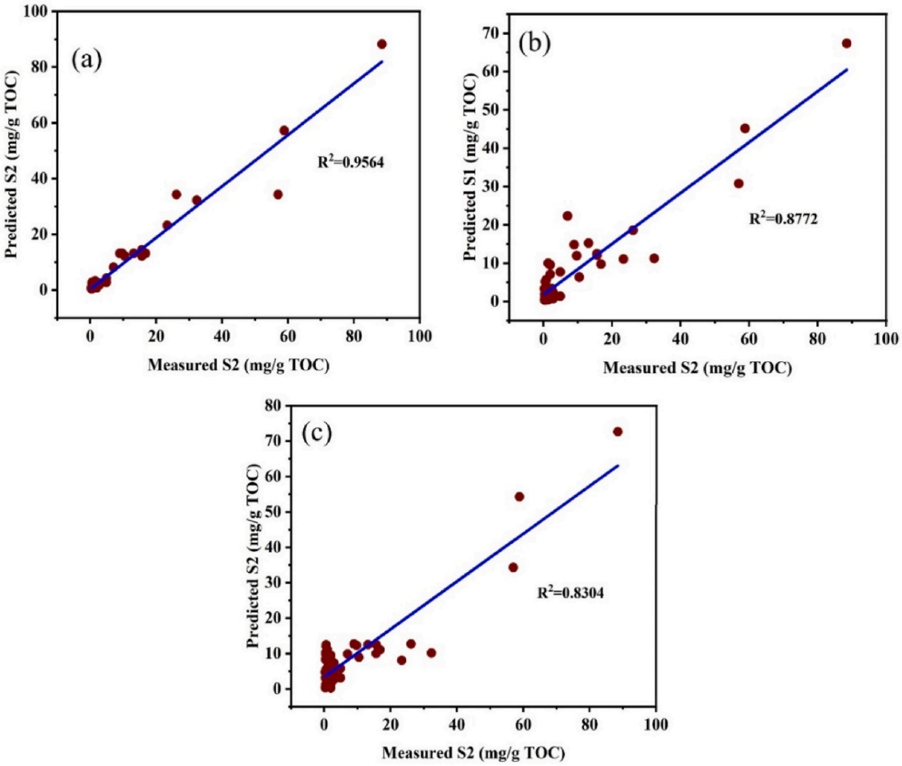

Fig. 25. Cross plots between actual and prediction S2 values using (a) g-GMDH, (b) GPR and (c) SVM models during training.

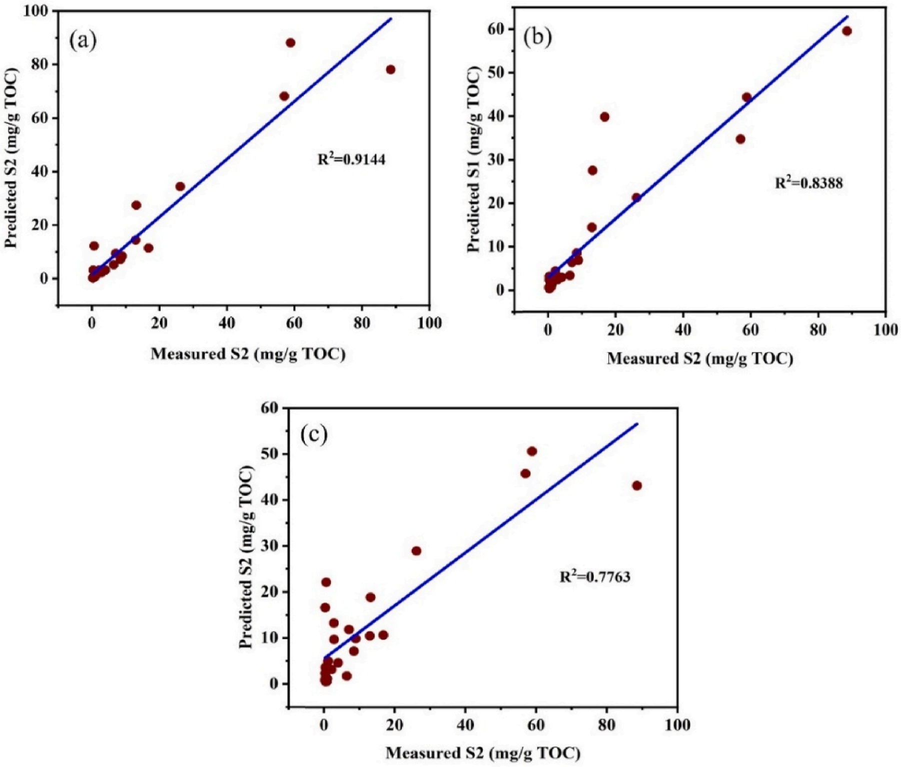

Fig. 26. Cross plots between actual and prediction S2 values using (a) g-GMDH, (b) GPR and (c) SVM models during testing.

g-GMDH also had a lower error in both training and testing sets, suggesting that it is better at generalizing its predictions. The lower error of g-GMDH can be attributed to its ability to capture nonlinear relationships between features in the dataset.

Furthermore, the results obtained from the prediction of Tmax using three different models are presented in Figs. 19 and 20, respectively. The R^2 scores of the g-GMDH, GPR and SVM models on the training dataset are 0.9646, 0.8763 and 0.8316, respectively while for the testing datasets are 0.9403, 0.8203 and 0.7588, respectively. From these results, it is clear that g-GMDH has the highest R^2 score both on the training and test datasets, indicating that it is the best model for predicting Tmax. This is followed by GPR and SVM, The R^2 scores for all the models are relatively high on the training dataset, but the scores drop slightly on the test dataset, suggesting that the models are performing well on the training data but not as well on the test data. This may be due to over-fitting of the data on the training set. To improve the results, it is necessary to use regularization techniques or cross-validation to minimize the effect of over-fitting. The higher accuracy of g-GMDH can be attributed to its inherent ability to capture nonlinear relationships between various features present in the dataset. Standard ML models of GPR and SVM might struggle to efficiently model complex, nonlinear interactions in the data, leading to less performance compared to g-GMDH. The flexibility of g-GMDH in discovering and incorporating such nonlinear relationships allowed it to provide more accurate predictions for Tmax.

4.2.5. S1 prediction

The present study is aimed to analyze the results of three machine learning models for the prediction of S1 during training and testing. The models used are g-GMDH, GPR and SVM. The results are presented in Table 6 and Fig. 21. The g-GMDH model was found to be the most accurate model with the lowest MSE (0.00511) and MAE (0.03725) values in the training dataset. The MAE and MSE values in the testing dataset are 0.17526 and 0.08563, respectively, indicating a slight increase in the prediction error. The GPR model showed slightly higher error in the training dataset compared to the g-GMDH model, with MAE and MSE values of 0.15955 and 0.06057, respectively. In the testing dataset, the MAE and MSE values are 0.20868 and 0.15598, respectively while that of SVM had an error margin of MAE and MSE values of 0.39439 and 0.38414, respectively. Regarding the GPR and SVM models, their relatively higher errors in both the training and testing datasets indicate limitations in capturing the underlying patterns of the S1 data. These models might struggle to handle nonlinearity or complex interactions, leading to inferior performance compared to g-GMDH.

The results of the prediction of S1 during both the training and testing phases of the modeling process are visualized in Figs. 22 and 23, respectively. The best overall results were obtained with the g-GMDH model, which had an R^2 of 0.9905 for training dataset and 0.9218 for testing dataset. This indicates that the model was able to accurately predict S1 values in both datasets. The GPR performed slightly worse, with an R^2 of 0.8886 for training dataset and 0.8584 for testing dataset. Finally, the SVM model had the lowest R^2 values, with 0.8622 for training dataset and 0.8086 for testing dataset. The slightly lower R^2 values observed for the GPR and SVM models indicate that they might not be as effective in capturing the complex patterns and variations present in the S1 data. The performance differences could be attributed to the models' respective capabilities in handling nonlinear relationships, which are often prevalent in geological datasets. Overall, the results suggest that the g-GMDH model is the most reliable for predicting S1 values in both the training and testing datasets.

4.2.6. S2 prediction

The results for the S2 prediction are presented in Table 7. The g-GMDH model had the lowest MSE and MAE errors for both the training and testing datasets. The MSE and MAE errors for the training dataset were 11.55488 and 1.27891 respectively. Similarly, the MSE and MAE errors for the testing dataset were 61.74456 and 4.14456 respectively. The GPR model gave slightly higher errors than the g-GMDH model. The MSE and MAE errors for the training dataset were 45.47719 and 3.76621 respectively. Similarly, the MSE and MAE errors for the testing dataset were 93.79075 and 5.0137 respectively. Finally, The SVM model

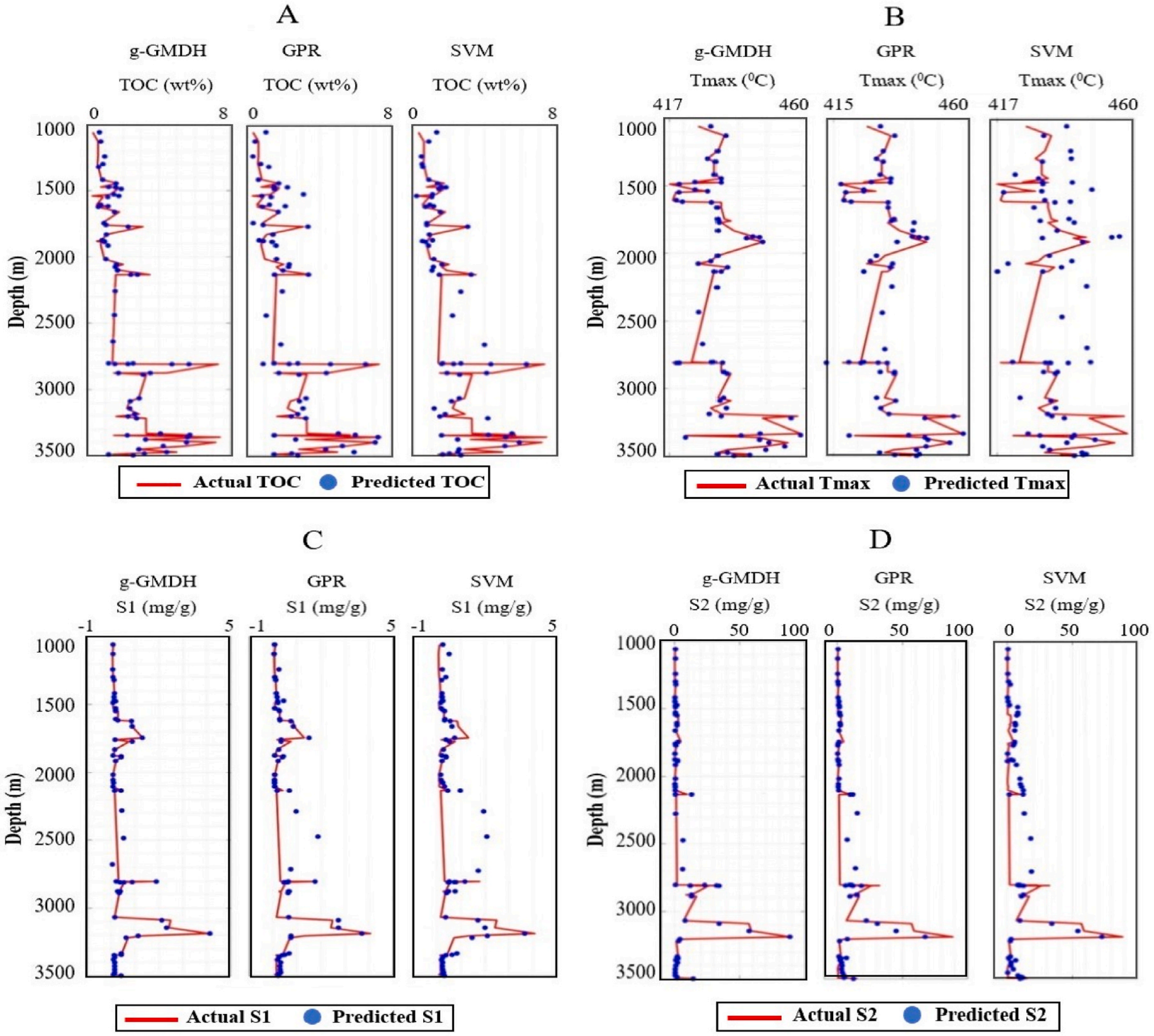

Fig. 27. Plots between predicted g-GMDH, GPR, SVM and actual values for (A) TOC (B) Tmax (C) S1 (D) S2.

had the highest errors among the other two models. The MSE and MAE errors for the training dataset were 51.76251 and 4.90752 respectively. Similarly, the MSE and MAE errors for the testing dataset were 131.97055 and 6.47348 respectively. Overall, the g-GMDH model had the lowest MSE and MAE errors for both the training and testing datasets (Fig. 24). The lower MSE and MAE errors of the g-GMDH model indicate its ability to provide more accurate predictions and better capture the underlying patterns in the S2 data. The g-GMDH model's strength lies in its ability to handle nonlinear relationships and complex interactions in the dataset, which are common in petroleum geology data. This indicates that the g-GMDH model is the most suitable model for the prediction of S2.

The performance of the three models used to predict S2 was further evaluated with respect to R^2. The results of the training and testing are shown in Figs. 25 and 26. The highest R^2 value of 0.9564 was obtained by the g-GMDH model, followed by the GPR model with 0.8772 and the SVM model with 0.8304 during the training phase. In the testing phase, the g-GMDH model achieved the highest R^2 value of 0.9144, followed by GPR with 0.8388 and SVM with 0.7763. It is evident from the results that the g-GMDH model outperformed the other two models. The g-GMDH model achieved a higher R^2 value of 0.9564 during the training phase, thus indicating its high efficacy in predicting S2. Similarly, the g-GMDH model also achieved the highest R^2 value of 0.9144 during the testing phase, which further validates its superior performance. Moreover, the high R^2 values obtained by the g-GMDH model indicate its strong ability to provide accurate predictions, making it a valuable tool for petroleum geologists in predicting S2 values. The g-GMDH model's ability to handle nonlinear relationships and complex interactions in the data allows it to better approximate the true S2 values. On the other hand, the slightly lower R^2 values obtained by the GPR and SVM models suggest that they may not be as effective in capturing the complexities present in the S2 data. These models might struggle to handle nonlinear relationships and could be limited in their predictive accuracy compared to the g-GMDH model. These results indicate that g-GMDH is the most suitable model for predicting S2 values. Its superior performance compared to GPR and SVM highlights its potential as an essential tool in petroleum geology for accurately predicting S2, which is crucial for various reservoir characterization and exploration tasks.

4.3. Model comparison

The study compared the performance of three different machine learning models of g-GMDH, GPR, and SVM, in predicting four different parameters, TOC, Tmax, S1, and S2. The results of the study showed that g-GMDH performed the best in predicting all four parameters, followed by GPR and SVM in that order. From Fig. 27 it revealed that g-GMDH model had the highest performance in predicting TOC, Tmax, S1, and S2, indicating its superiority in handling the complexity of the datasets. GPR model showed good performance in predicting the parameters, but not as good as g-GMDH. It is still considered to be a powerful model for

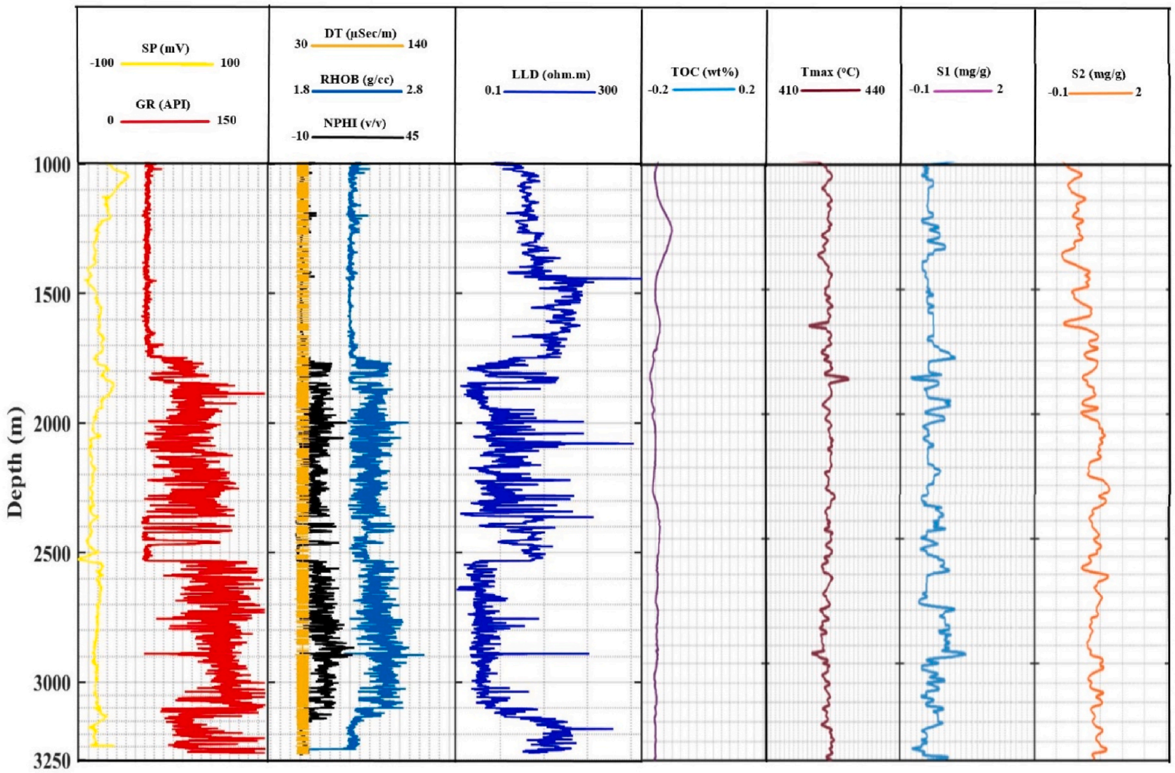

Fig. 28. Well logs and g-GMDH predicted values for TOC, Tmax, S1, and S2 from East Lika.

handling complex data. SVM showed reasonable performance in predicting the parameters, but not as good as g-GMDH or GPR. SVM is known for its ability to handle large datasets and non-linear relationships, but in this study, it did not perform as well as the GPR models. SVM showed the lowest performance in predicting the parameters, indicating that it may not be the best choice for handling the complexity of the datasets used in this study. In conclusion, the study showed that g-GMDH performed the best in predicting the four parameters of interest, followed by GPR and SVM. This suggests that g-GMDH and GPR are powerful models for handling complex datasets with non-linear relationships, while SVM may not be the best choices for such datasets. However, the performance of these models may vary depending on the dataset and task, and more research is needed to confirm these findings.

4.4. Model verification using East Lika well

After the g-GMDH model show the success in prediction of geochemical parameters of TOC, Tmax, S1 and S2 from three wells in Mandawa basin. The East Lika well was used to test the method, as it had no geochemical parameters. Fig. 28 shows the well logs of East Lika and predicted geochemical data by g-GMDH model. The predicted TOC showed a very small difference ranged from 0 to 0.2 wt% which according to the classification it falls into the category of poor source rock. Similarly, for the Tmax the results showed that the source rock is still immature with the range of 420–438 °C, for the case of S1 and S2 the results ranged from 0.1 to 4.67 mg/g. The evidence for hydrocarbon migration can also be observed from depth 2700–2950 m due to the relatively higher S1 value compared to the S2.

5. Conclusion and recommendation

This study proposed an approach of generalized of group method of data handling (g-GMDH) as a novel method in the source rock evaluation and prediction of TOC, Tmax, S1, and S2 from well logs data. The following conclusion can be made from the above results

(1) Based on geochemical findings the Mandawa basin can be classified as fair to very good source rocks. TOC contents range from 0.5 to 8.7 (wt.%). Mandawa basin contains oil and gas prone characterized by mixed kerogen type II and III laying in oil to condensate as immature to mature source rocks.
(2) Furthermore, this study revealed that the g-GMDH model was the most accurate model for the prediction of TOC, Tmax, S1, and S2 with the R^2 value greater than 0.9 and low errors margin in both training and testing phases. The GPR and SVM models showed inferior performance, indicating that they may not be the best choice for predicting these parameters. The findings of this study have practical implications for the energy industry, as accurate prediction of TOC, Tmax, S1, and S2 is crucial for successful exploration and production of hydrocarbons.
(3) The proposed g-GMDH method was applied to estimate the TOC, Tmax, S1 and S2 values for the East Lika-1 well, which lack core geochemical data. The results indicated a poor-quality source rock in the well and suggested the likelihood of hydrocarbon migration between the depths of 2700 m and 2950 m.

In the future, researchers may investigate how combining different types of data, such as seismic and mineralogical data, could improve the accuracy of predicting geochemical parameters. They should also work on improving the model parameters to increase prediction accuracy and efficiency. Furthermore, the model suggested in this study can also be applied to predicting other reservoir parameters, such as Porosity and water saturation.

CRediT authorship contribution statement

Christopher N. Mkono: Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing. **Shen Chuanbo:** Supervision, Funding acquisition, Resources, Conceptualization, Writing – review & editing. **Alvin K. Mulashani:** Methodology, Data curation, Software, Writing – review & editing. **Grant Charles**

Mwakipunda: Methodology, Software, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Data availability

The data that has been used is confidential.

Acknowledgments

This work was supported by the Major National Science and Technology Programs in the "Thirteenth Five-Year" Plan period (No. 2017ZX05032-002-004), the innovation Team Funding of Natural Science Foundation of Hubei Province (No. 2021CFA031). The first author gratefully acknowledges financial support from China Scholarship Council (CSC).

Nomenclature

DEN	Density
CNL	Compensated Neutron
RT	Resistivity
GR	Gamma Ray
AC	Acoustic
ILD	Induction
DT	Sonic Travel Time
RHOB	Bulk Density
BPNN	Backpropagation Neural Network
RBFNN	Radial Basis Function Neural Network
LLD	Deep Lateral Resistivity
NPHI	Neutron Porosity
DN	Density
NCL	Neutron
BD	Bulk Density
GD	Grain Density
DTSH	Shear Slowness
DTC	Compressional Slowness
SGR	Spectral Gamma Ray
U	Uranium
TH	Thorium
K	Potassium
TPDC	Tanzania Petroleum Development Corporation
RF	Random Forest
LR	Linear Regression
PSO	Particle Swarm Optimization
LSSVM	Least Square Support Vector Machine
ELM	Ensemble Learning Machine
FFNN	Feed Forward Neural Network
KNN	k Nearest Neighbour

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.energy.2023.129232.

References

[1] Su M, Wang Q, Li R, Wang L. Per capita renewable energy consumption in 116 countries: the effects of urbanization, industrialization, GDP, aging, and trade openness. Energy 2022;254:124289.

[2] IEA. World energy outlook. Paris: IEA; 2022.

[3] Karakurt I, Aydin G. Development of regression models to forecast the CO2 emissions from fossil fuels in the BRICS and MINT countries. Energy 2023;263:125650.

[4] Ozdemir AC. Decomposition and decoupling analysis of carbon dioxide emissions in electricity generation by primary fossil fuels in Turkey. Energy 2023;273:127264.

[5] Global IEA. Energy review. Paris: IEA; 2021.

[6] Novotnik B, Nandy A, Venkatesan SV, Radović JR, JDl Fuente, Nejadi S, et al. Can fossil fuel energy be recovered and used without any CO2 emissions to the atmosphere? Rev Environ Sci Biotechnol 2020;19(1):217–40.

[7] Rezaei A, Siddiqui F, Dindoruk B, Soliman MY. A review on factors influencing the rock mechanics of the gas bearing formations. J Nat Gas Sci Eng 2020;80:103348.

[8] Siddik M, Islam M, Zaman A, Hasan M. Current status and correlation of fossil fuels consumption and greenhouse gas emissions. Int J Energy Environ Econ 2021;28:103–19.

[9] Wang K, Ma L, Taylor KG. Microstructure changes as a response to CO2 storage in sedimentary rocks: recent developments and future challenges. Fuel 2023;333:126403.

[10] El Diasty WS, El Beialy S, Mostafa A, Abo Ghonaim A, Peters K. Chemometric differentiation of oil families and their potential source rocks in the Gulf of Suez. Natural Resources Research 2020;29:2063–102.

[11] Green H, Šegvić B, Zanoni G, Omodeo-Salé S, Adatte T. Evaluation of shale source rocks and clay mineral diagenesis in the permian basin, USA: inferences on basin thermal maturity and source rock potential. geosciences 2020;10(10):381.

[12] Aziz H, Ehsan M, Ali A, Khan HK, Khan A. 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. J Nat Gas Sci Eng 2020;81:103433.

[13] Mahmoud AA, ElKatatny S, Abdulraheem A, Mahmoud M, Omar Ibrahim M, Ali A. New technique to determine the total organic carbon based on well logs using artificial neural network (white box). Paper presented at the SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition, Dammam, Saudi Arabia, 24-27 April 2017. SPE-188016-MS.

[14] Evenick JC. Late Cretaceous (Cenomanian and Turonian) organofacies and TOC maps: example of leveraging the global rise in public-domain geochemical source rock data. Mar Petrol Geol 2020;111:301–8.

[15] Wu Z, Zhao X, Pu X, Wang E, Dong X, Li C. Petroleum resource potential evaluation using insights based on hydrocarbon generation, expulsion, and retention capabilities—a case study targeting the Paleogene Es1 formation, Qikou Sag. J Petrol Sci Eng 2022;208:109667.

[16] Wu Y, Liu C, Jiang F, Hu T, Lv J, Zhang C, et al. Geological characteristics and shale oil potential of alkaline lacustrine source rock in Fengcheng Formation of the Mahu Sag, Junggar Basin, Western China. J Petrol Sci Eng 2022;216:110823.

[17] Alves GM, Júnior IP. Microwave remediation of oil-contaminated drill cuttings–A review. J Petrol Sci Eng 2021;207:109137.

[18] Jia W, Zong Z, Qin D, Lan T. Integrated well-log data and seismic inversion results for prediction of hydrocarbon source rock distribution in W segment, Pearl River Mouth Basin, China. Geoenergy Science and Engineering 2023:212233.

[19] Khalil Khan H, Ehsan M, Ali A, Amer MA, Aziz H, Khan A, et al. Source rock geochemical assessment and estimation of TOC using well logs and geochemical data of Talhar Shale, Southern Indus Basin, Pakistan. Front Earth Sci 2022;10:969936.

[20] Chen J, Zhang G, Chen H, Yin X. The construction of shale rock physics effective model and prediction of rock brittleness. Conference the construction of shale rock physics effective model and prediction of rock brittleness. SEG, p. SEG-2014-0716.

[21] Mulashani AK, Shen C, Nkurlu BM, Mkono CN, Kawamala M. Enhanced group method of data handling (GMDH) for permeability prediction based on the modified Levenberg Marquardt technique from well log data. Energy 2022;239:121915.

[22] Abdel-Fattah MI, Mahdi AQ, Theyab MA, Pigott JD, Abd-Allah ZM, Radwan AE. Lithofacies classification and sequence stratigraphic description as a guide for the prediction and distribution of carbonate reservoir quality: a case study of the Upper Cretaceous Khasib Formation (East Baghdad oilfield, central Iraq). J Petrol Sci Eng 2022;209:109835.

[23] Ahmadi MA, Chen Z. Comparison of machine learning methods for estimating permeability and porosity of oil reservoirs via petro-physical logs. Petroleum 2019;5(3):271–84.

[24] Zhang H, Wu W, Wu H. TOC prediction using a gradient boosting decision tree method: a case study of shale reservoirs in Qinshui Basin. Geoenergy Science and Engineering 2023;221:111271.

[25] Zhang W, Shan X, Fu B, Zou X, Fu L-Y. A deep encoder-decoder neural network model for total organic carbon content prediction from well logs. J Asian Earth Sci 2022;240:105437.

[26] Wang H, Lu S, Qiao L, Chen F, He X, Gao Y, et al. Unsupervised contrastive learning for few-shot TOC prediction and application. Int J Coal Geol 2022:104046.

[27] Nyakilla EE, Silingi SN, Shen C, Jun G, Mulashani AK, Chibura PE. 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. Natural Resources Research 2022;31(1):619–41.

[28] Khalil Khan H, Ehsan M, Ali A, Amer MA, Aziz H, Khan A, et al. Source rock geochemical assessment and estimation of TOC using well logs and geochemical data of Talhar Shale, Southern Indus Basin, Pakistan. Front Earth Sci 2022:1593.

[29] Liu X, Lei Y, Luo X, Wang X, Chen K, Cheng M, et al. TOC determination of Zhangjiatan shale of Yanchang formation, Ordos Basin, China, using support vector regression and well logs. Earth Science Informatics 2021;14(2):1033–45.

[30] Tan M, Song X, Yang X, Wu Q. Support-vector-regression machine technology for total organic carbon content prediction from wireline logs in organic shale: a comparative study. J Nat Gas Sci Eng 2015;26:792–802.
[31] Siddig O, Ibrahim AF, Elkatatny S. Application of various machine learning techniques in predicting total organic carbon from well logs. Comput Intell Neurosci 2021:2021.
[32] Wood DA. Predicting total organic carbon from few well logs aided by well-log attributes. Petroleum 2023;9(2):166–82.
[33] Saporetti C, Fonseca D, Oliveira L, Pereira E, Goliatt L. Hybrid machine learning models for estimating total organic carbon from mineral constituents in core samples of shale gas fields. Mar Petrol Geol 2022;143:105783.
[34] Siddig O, Abdulhamid Mahmoud A, Elkatatny S, Soupios P. Utilization of artificial neural network in predicting the total organic carbon in devonian shale using the conventional well logs and the spectral gamma Ray. Comput Intell Neurosci 2021: 2021.
[35] Siddig O, Mahmoud AA, Elkatatny S. New robust model to evaluate the total organic carbon using artificial neural networks and spectral gamma-ray. Conference new robust model to evaluate the total organic carbon using artificial neural networks and spectral gamma-ray. OnePetro.
[36] Shi X, Wang J, Liu G, Yang L, Ge X, Jiang S. Application of extreme learning machine and neural networks in total organic carbon content prediction in organic shale with wire line logs. J Nat Gas Sci Eng 2016;33:687–702.
[37] Mahmoud AA, Gamal H, Elkatatny S, Alsaihati A. Estimating the total organic carbon for unconventional shale Resources during the drilling process: a machine learning approach. J Energy Resour Technol 2022;144(4).
[38] Mahmoud AA, Elkatatny S. Novel empirical correlation for estimation of the total organic carbon in devonian shale from the spectral gamma-ray and based on the artificial neural networks. J Energy Resour Technol 2021;143(9).
[39] Mulashani AK, Shen C, Asante-Okyere S, Kerttu PN, Abelly EN. Group method of data handling (GMDH) neural network for estimating total organic carbon (TOC) and hydrocarbon potential distribution (S1, S2) using well logs. Natural Resources Research 2021;30(5):3605–22.
[40] Wang H, Wu W, Chen T, Dong X, Wang G. An improved neural network for TOC, S1 and S2 estimation based on conventional well logs. J Petrol Sci Eng 2019;176: 664–78.
[41] Ahangari D, Daneshfar R, Zakeri M, Ashoori S, Soulgani BS. On the prediction of geochemical parameters (TOC, S1 and S2) by considering well log parameters using ANFIS and LSSVM strategies. Petroleum 2022;8(2):174–84.
[42] Shalaby MR, Malik OA, Lai D, Jumat N, Islam MA. Thermal maturity and TOC prediction using machine learning techniques: case study from the Cretaceous–Paleocene source rock, Taranaki Basin, New Zealand. J Pet Explor Prod Technol 2020;10(6):2175–93.
[43] Alizadeh B, Maroufi K, Heidarifard MH. Estimating source rock parameters using wireline data: an example from Dezful Embayment, South West of Iran. J Petrol Sci Eng 2018;167:857–68.
[44] Amosu A, Imsalem M, Sun Y. Effective machine learning identification of TOC-rich zones in the Eagle Ford Shale. J Appl Geophys 2021;188:104311.
[45] Handhal AM, Al-Abadi AM, Chafeet HE, Ismail MJ. Prediction of total organic carbon at Rumaila oil field, Southern Iraq using conventional well logs and machine learning algorithms. Mar Petrol Geol 2020;116:104347.
[46] Mandal PP, Rezaee R, Emelyanova I. Ensemble learning for predicting TOC from well-logs of the unconventional goldwyer shale. Energies 2021;15(1):216.
[47] Barham A, Ismail MS, Hermana M, Padmanabhan E, Baashar Y, Sabir O. Predicting the maturity and organic richness using artificial neural networks (ANNs): a case study of Montney Formation, NE British Columbia, Canada. Alex Eng J 2021;60(3): 3253–64.
[48] Geiger M, Clark D, Mette W. Reappraisal of the timing of the breakup of Gondwana based on sedimentological and seismic evidence from the Morondava Basin, Madagascar. J Afr Earth Sci 2004;38(4):363–81.
[49] Nicholas CJ, Pearson PN, McMillan IK, Ditchfield PW, Singano JM. Structural evolution of southern coastal Tanzania since the Jurassic. J Afr Earth Sci 2007;48 (4):273–97.
[50] Fossum K, Dypvik H, Haid MHM, Hudson WE, Mvd Brink. Late Jurassic and early cretaceous sedimentation in the Mandawa Basin, coastal Tanzania. J Afr Earth Sci 2021;174:104013.
[51] Hudson W. The geological evolution of the petroleum prospective Mandawa Basin southern coastal Tanzania. 2011. Trinity College Dublin.
[52] Kapilima S. Tectonic and sedimentary evolution of the coastal basin of Tanzania during the Mesozoic times. Tanzan J Sci 2003;29:1–16.
[53] Reeves C. The development of the East African margin during Jurassic and Lower Cretaceous times: a perspective from global tectonics. Petrol Geosci 2018;24(1): 41–56.
[54] Mtabazi EG, Boniface N, Andresen A. Geochronological characterization of a transition zone between the Mozambique belt and unango-marrupa complex in SE Tanzania. Precambrian Res 2019;321:134–53.
[55] Tuck-Martin A, Adam J. Eagles GJBR. New plate kinematic model and tectono-stratigraphic history of the East African and west Madagascan margins 2018;30(6): 1118–40.
[56] Bown PR, Jones TD, Lees J, Randell R, Mizzi J, Pearson PN, et al. A Paleogene calcareous microfossil Konservat-Lagerstatte from the Kilwa Group of coastal Tanzania 2008;120(1–2):3–12.
[57] Maganza NE. Petroleum system modelling of onshore Mandawa Basin-southern. Tanzania: Institutt for geologi og bergteknikk; 2014.
[58] Nicholas CJ, Pearson PN, Bown PR, Jones TD, Huber BT, Karega A, et al. Stratigraphy and sedimentology of the upper cretaceous to Paleogene Kilwa group, southern coastal Tanzania. J Afr Earth Sci 2006;45(4):431–66.
[59] Hudson W, Nicholas C. The Pindiro group (triassic to early Jurassic Mandawa Basin, southern coastal Tanzania): definition, palaeoenvironment, and stratigraphy. J Afr Earth Sci 2014;92:55–67.
[60] Fossum K, Morton AC, Dypvik H, Hudson WE. Integrated heavy mineral study of Jurassic to Paleogene sandstones in the Mandawa Basin, Tanzania: sediment provenance and source-to-sink relations. J Afr Earth Sci 2019;150:546–65.
[61] Zhou Z, Tao Y, Li S, Ding W. Hydrocarbon potential in the key basins in the East coast of Africa. Petrol Explor Dev 2013;40(5):582–91.
[62] Smelror M, Fossum K, Dypvik H, Hudson W, Mweneinda A. Late Jurassic–early cretaceous palynostratigraphy of the onshore Mandawa Basin, southeastern Tanzania. Rev Palaeobot Palynol 2018;258:248–55.
[63] Á Berrocoso, Huber B, MacLeod K, Petrizzo M, Lees J, Wendler I, et al. The Lindi Formation (upper Albian–coniacian) and Tanzania drilling project sites 36–40 (lower cretaceous to Paleogene): lithostratigraphy, biostratigraphy and chemostratigraphy. J Afr Earth Sci 2015;101:282–308.
[64] Sun Z, Han Y, Wang Z, Chen Y, Liu P, Qin Z, et al. Detection of voltage fault in the battery system of electric vehicles using statistical analysis. Appl Energy 2022;307: 118172.
[65] Ordoñez L, Vogel H, Sebag D, Ariztegui D, Adatte T, Russell JM, et al. Empowering conventional Rock-Eval pyrolysis for organic matter characterization of the siderite-rich sediments of Lake Towuti (Indonesia) using End-Member Analysis. Org Geochem 2019;134:32–44.
[66] Su X, Sun B, Wang J, Zhang W, Ma S, He X, et al. Fast capacity estimation for lithium-ion battery based on online identification of low-frequency electrochemical impedance spectroscopy and Gaussian process regression. Appl Energy 2022;322:119516.
[67] Zhang Z, Ye L, Qin H, Liu Y, Wang C, Yu X, et al. Wind speed prediction method using shared weight long short-term memory network and Gaussian process regression. Appl Energy 2019;247:270–84.
[68] Wang Z, Peng X, Xia A, Shah AA, Yan H, Huang Y, et al. Comparison of machine learning methods for predicting the methane production from anaerobic digestion of lignocellulosic biomass. Energy 2023;263:125883.
[69] Niu W, Lu J, Sun Y, Guo W, Liu Y, Mu Y. Development of visual prediction model for shale gas wells production based on screening main controlling factors. Energy 2022;250:123812.
[70] Chen H, Zhang C, Yu H, Wang Z, Duncan I, Zhou X, et al. Application of machine learning to evaluating and remediating models for energy and environmental engineering. Appl Energy 2022;320:119286.
[71] Shen C, Asante-Okyere S, Yevenyo Ziggah Y, Wang L, Zhu X. Group method of data handling (GMDH) lithology identification based on wavelet analysis and dimensionality reduction as well log data pre-processing techniques. Energies 2019;12(8):1509.
[72] Najafzadeh M, Azamathulla HM. Group method of data handling to predict scour depth around bridge piers. Neural Comput Appl 2013;23(7):2107–12.
[73] Anastasakis L, Mort N. 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; 2001.
[74] Armaghani DJ, Momeni E, Asteris PG. Application of group method of data handling technique in assessing deformation of rock mass. Metaheuristic Computing Applied 2020;1(1):1–18.
[75] Ivakhnenko AG. The group method of data of handling; a rival of the method of stochastic approximation. Soviet Automat Control 1968;13:43–55.
[76] Dembicki H. Practical petroleum geochemistry for exploration and production. Elsevier; 2022.
[77] Clayton JL, Ryder RT. Organic geochemistry of black shales and oils in the minnelusa formation (permian and pennsylvanian). Wyoming: Powder River Basin; 1984.
[78] Espitalié J, Deroo G, Marquis F. La pyrolyse Rock-Eval et ses applications. Rev Inst Fr Pétr. 1985;40:563–79.
[79] Waples DW. Maturity modeling: thermal indicators, hydrocarbon generation, and oil cracking. Identification and Characterization; 1994 [Chapter 17]: Part IV.
[80] Cruz Luque M, Aguilera R. Eagle ford and pimienta shales in Mexico: a case study. SPE Reservoir Eval Eng 2019;22:1305–22. 04.
[81] Bordenave M, Espitalié J, Leplat P, Oudin J, Vandenbroucke M. Screening techniques for source rock evaluation. Appl Petrol Geochem 1993:217–78.
[82] Hunt J. Petroleum geochemistry and geology (textbook). In: Petroleum geochemistry and geology (textbook). 2 nd Ed. WH Freeman Company; 1995.
[83] Jackson MJ, Powell TG, Summons RE, Sweet IP. Hydrocarbon shows and petroleum source rocks in sediments as old as 1.7 × 109 years. Nature 1986;322 (6081):727–9.