Marine and Petroleum Geology 116 (2020) 104347

Contents lists available at ScienceDirect

Marine and Petroleum Geology

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

Research paper

Prediction of total organic carbon at Rumaila oil field, Southern Iraq using conventional well logs and machine learning algorithms

Amna M. Handhal[a], Alaa M. Al-Abadi[a,*], Hussein E. Chafeet[b], Maher J. Ismail[c]

[a] Department of Geology, College of Science, University of Basrah, Basrah, Iraq
[b] Department of Oil and Gas Engineering, College of Oil and Gas Engineering, Basrah University for Oil and Gas, Basrah, Iraq
[c] Basra Oil Company, Basrah, Iraq

ARTICLE INFO

Keywords:
Total organic carbon
Rotation forest
Well logs
K nearest neighbors
Iraq

ABSTRACT

Total organic carbon (TOC) is an important parameter for assessing the hydrocarbon potential of source rocks. The standard method for analysis of TOC is the Rock-Eval pyrolysis on cutting and core samples. The coring process is always expensive and time consuming. Therefore, researchers around the world focused on developing techniques to estimate TOC and other organic parameters from readily available well logs data that are almost available in all wells. In this study, we evaluated the use of three machine learning models namely, random forest (RF), rotation forest (rF), k nearest neighbors (KNN) to estimate TOC based on conventional well logs data. The well logs involved gamma ray, acoustic, density, neutron, and deep resistivity. The efficacy of the models was tested against the most widely used backpropagation artificial neutral network (BPANN) and support vector regression (SVR) models. North Rumaila oilfield in southern Iraq was taken as a case study. The models were trained and tested using data from two wells in the field, namely R-167 and R-172. The number of TOC measurements used for training and testing were 40 (R-167) and 18 (R-172), respectively. The efficacy of the used algorithms was evaluated using mean absolute error (MAE), root mean squared error (RMSE), and correlation of determination (R^2). The models are also visually compared using Taylor diagram and violin plot to distinguish the best performance model. Results indicated the KNN was the best followed by RF and then rF. The worst performance models were BPANN and SVR models. This study confirmed the ability of machine learning models for building efficient model for estimating TOC from readily available borehole logs data without the need for very expensive coring process.

1. Introduction

TOC is a measure of the organic richness of a rock, that is, the amount of organic carbon in a rock sample (both kerogen and bitumen) (Jarvie, 1991; Peters and Cassa, 1994). This parameter is best measured on cutting and core dependent samples using Rock-Eval pyrolysis apparatuses. The process of obtaining core samples is expensive and time-consuming. In addition, results are incomplete, as a few exploration wells deliberately penetrate the source rock horizons so that a limited number of samples can be collected for laboratory analysis. Therefore, different mathematical relationships and empirical formulas are developed to predict TOC from readily available well logs data that are almost available in all wells (Passey et al., 1990; Huang and Williamson, 1996; Kamali and Mirshady, 2004). Due to the complexity of the relationship between the logs response and the geochemical data, most of the developed linear relationships are fail to attain appropriate accuracy (Wang et al., 2019). With recognition of its potential, artificial intelligent (AI) and machine learning (ML) techniques have been recently applied to model the relationship between TOC and well logs data with highly promising results (Abdizadeh et al., 2017; Bolandi et al., 2017; Farzi and Bolandi, 2016; Kadkhodaie-Ilkhchi et al., 2009; Ouadfeul and Aliouane, 2015; Sfidari et al., 2012; Wang et al., 2019). ML is a specific branch of AI that enables systems to learn from experience and improve from experience without explicit programming. The aim is to allow computers to learn automatically without human assistance to adjust subsequent actions accordingly (Bishop, 2006). Despite its great success in many applications in the fields of engineering, science, marketing, environment, and geospatial models, its use in prediction of TOC from well log data is still faltering, especially since many relatively new algorithms such as rotation forest (rF),

* Corresponding author.
E-mail addresses: amna.handhal@uobasrah.edu.iq (A.M. Handhal), alaa.atiaa@uobasrah.edu.iq (A.M. Al-Abadi), hussein.aliwi@buog.edu.iq (H.E. Chafeet), mahermji@gmail.com (M.J. Ismail).

https://doi.org/10.1016/j.marpetgeo.2020.104347
Received 12 January 2020; Received in revised form 11 March 2020; Accepted 11 March 2020
Available online 19 March 2020

Nomenclature

AC	Acoustic log
BPANN	Backpropagation artificial neural network
BRT	Boosted regression trees
Bk1	WEKA kNN package
CART	Classification and regression trees
DN	Density log
DT	Decision trees
GR	Gamma ray log
kNN	k-nearest neighbor
MAE	Mean absolute error
ML	Machine learning
NCL	Neutron log
NR	North Rumaila
PCA	Principle component analysis
r	Pearson's correlation coefficient
R^2	Correlation of determination
RF	Random forest
rF	rotation forest
RMSE	root mean squared error
RT	Resistivity log
SVR	Support vector regression
TOC	Total organic carbon

random forest (RF), k-nearest neighbors (kNN), boosted regression trees (BRT), classification and regression trees (CART), M5P, etc. have not been tested.

In this study, three ML techniques (rF, RF, and KNN) were used for predicting TOC (target variable) from a suite of conventional well logs data (gamma ray, resistivity, density, acoustic, and neutron) as regressors. The rF was chosen for its scarce use despite its efficiency and superiority to many other ML algorithms. The RF was chosen as the most algorithm used in the scientific and engineering fields that proved its efficiency in the fields of classification and regression. The KNN was selected for its simplicity and to compare its results with the most advantageous techniques (rF and RF). North Rumaila (NR) oilfield in southern Iraq was taken as case study. A comparison was also made with backpropagation artificial natural network (ANN) and support vector regression with radial basis function (SVR-radial) models to show the best technique to predict TOC.

2. Geological setting

2.1. Structural and tectonic

NR oilfield locates at Basrah Governorate, south of Iraq and encompasses an area of 1600 km^2 (Fig. 1). NR is the northern portion of the Rumaila supergiant oilfield, the largest producing field in Iraq since 1953. From the structural point of view, Rumaila oilfield is a doubly plunging simple and asymmetric anticline with N-E trend in the long axis. This system has two domes, North Rumaila (NR) and South Rumaila (SR). The length of this structure is 83 km and its width is 12 km (Handhal et al., 2019). The NR dome is an asymmetrical elongated anticline with a North-South trend. The dip on the flank is about 3°. It plunges gently towards the south to form a saddle separating it from the dome of SR. Given Iraq's tectonic divisions, NR is situated in the Zubair subzone of the Mesopotamian zone. It is located in the sagged basin of the Mesopotamian zone a part of the Arabian plate quasiplatform foreland. The Takadid Quarna and Basrah-Zubair faults have bounded this zone from north and south, respectively.

2.2. Stratigraphy

The stratigraphic column in NR comprises entirely of sedimentary rocks from the Jurassic to the Recent. The lithostratigraphic column consists of cycles of clastic, carbonate, and evaporite rocks. From the bottom to top, the stratigraphic column comprises Sargelu, Naokelekan, Najimah, Gotnia, Sulay, Yamama, Ratwai, Zubair, Shuaiba, Nahr Umr, Muddud, Ahamdi, Rumaila, Mishrif, Khasib, Tanuma, Saadi, Hardha, Shiranish, Tayarat, Umm Er-Radhuma, Rus, Dammam, Ghar, Lower Fars, Dibdibba, and Hammar formations (Fig. 2). The main reservoirs include Zubair and Mishrif formations, the source rocks are Sargelu, Sulaiy, Yamam, and Ratawi, and the cap rocks are Gotania, Tanuma, Shiranish, and Rus. The Zubair Formation is the most important formation of the Lower Cretaceous cycle in Iraq (Al-Sayyab, 1989). It is composed of fluvio-delatic and marine sandstone (Aqrawi et al., 2010) with an average thickness of 425 m. The age of the formation is Hauterivian until early Aptian (Bellen et al., 1959). Mishrif Formation, on the other hand, represents a heterogenous formation originally described as organic detrital limestone with beds of algal, rudist, and coral-reef limestones (Bellen et al., 1959). The age of formation is late Cenomanian. The Mishrif Formation succession indicates general shallowing from open-shelf to fore-reef slope, then reef flat and finally inner-shelf conditions (Aqrawi et al., 1998). Sargelu Formation in its type section (northern Iraq) comprises 115 m of thin-bedded, black, bituminous and dolomitic limestones, and black papery shales with streaks of thin black chert (Jassim and Goff, 2006). The depositional environment of this formation is the basinal euxinic marine environment. The age is of Bajocian-Bathonian. Sulaiy Formation, on other hands, is composed of limestone with some shale streaks at its base with an average thickness of 245 m. The age of the formation is late Tithonian-early cretaceous. It regarded as the best source rocks in southern Iraq, Kuwait, and Saudi Arabia (Al-Ameri et al., 1999). The depositional environment of this formation is offshore marine-shelf (Al-Ameri et al., 2011). Yamama Formation mainly comprises 12 m of specular and brown detrital limestone with thin shale beds overlain by 191 m of micritic limestone and oolitic limestone (Jassim and Goff, 2006). The formation is of Berriasian-Valanginian age (Bellen et al., 1959). The depositional environment was alternating oolitic shoal and deep inner shelf, probably controlled by subtle structural highs within a carbonate ramp (Sadooni, 1993). Gotania.

Formation comprises anhydride with subordinate beds of brown calcareous shales, thin black bituminous. The Tanuma Formation comprises 30 m of black, fissile shale with streaks of grey, macrocrystalline, argillaceous and detrital limestones with an oolitic limestone layer at the top (Jassim and Goff, 2006). The Tanuma Formation was deposited in a restricted shallow basin, in a partly euxinic environment. The age formation is Late Senonian (Bellen et al., 1959). The Shiranish Formation, in its type area, comprises thin-bedded argillaceous limestones overlain by blue pelagic marls (Owen and Nasr, 1958). The age is Late Campanian-Maastrichtian. The Rus Formation, in its type area, consists predominantly of anhydride with some unfossiliferous limestone, blue shale, and marl. The age of the formation is Early Eocene (Ypresian) (Al-Naqib, 1967). The formation was deposited in a lagoonal-sebkha environment on the stable shelf.

3. Methodology

The steps follow to achieve the objective of this study are presented in a flow chart (Fig. 3) and consists of: (1) collecting data concerning TOC (Rock-Eval pyrolysis of core samples) and conventional well logs data (2) test the relevance of well logs in predicting TOC using Pearson correlation coefficient method (3) Building ML models using R statistical software (4) Compare the performance of the models used via correlation of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE) in training and testing stages, and (5)

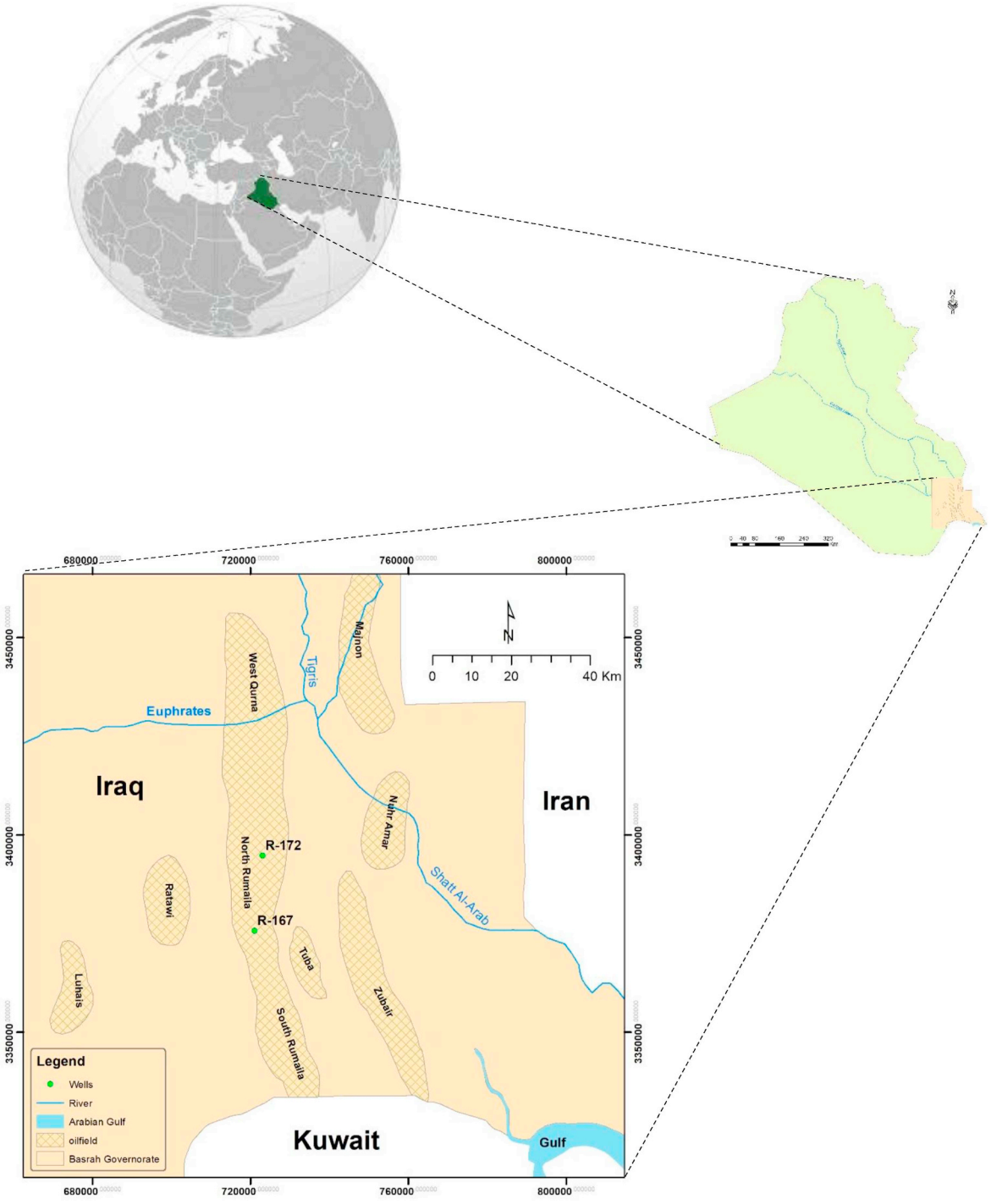

Fig. 1. Location of the study area.

compare across visualization the best of ML models using Taylor diagram and violin plot to select the best ML model.

3.1. TOC data

In this study, 58 core samples from two wells (R-167 and R-172) in NR oilfield were taken for Rock-Eval pyrolysis. The Rock-Eval analysis was conducted in the Geology Department of South Oil Company (SOC) Iraq following the method described by (Espitalie et al., 1977) and (Langford and Blanc-Valleron, 1990). The core samples were taken for the depth intervals (3200–4510 m) and (3036–4110 m) for the R-167 and R-172 wells, respectively. Forty (40) core samples were taken from Zubair, Ratawi, Yamama, and Suliay formations for the R-167. For the R-172, 18 core samples were taken from Zubair, Ratawi, and Yamama. The rock samples were first pulverized and then, 70 mg of each sample was weighted and placed in a clean crucible based on the depth of the samples (Rahmani et al., 2019). After that, the samples had been decontaminated to avoid device contamination. The Rock-Eval was run with a temperature schedule of 25 °C min^{-1}, where the final temperature in the pyrolysis oven exceeds 800 °C, and in the oxidation oven 850 °C. The Rock-Eval pyrolsis was designed to provide the following main parameters: TOC (wt. %), volatile hydrocarbon (S1 in mg HC (hydrocarbon)/g rock), hydrocarbon derived from kerogen pyrolsis (S2 in mg HC/g rock), the temperature at the highest yield of S2 (T_{max} in

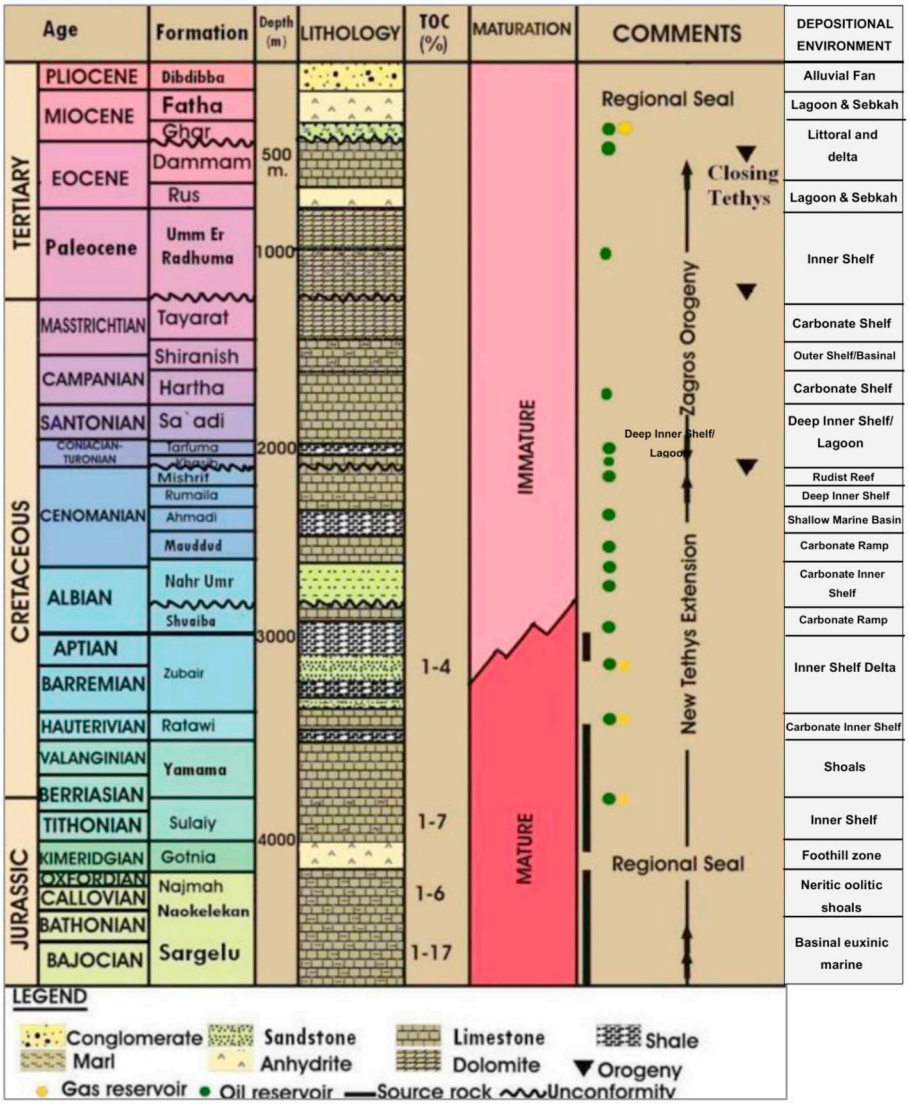

Fig. 2. Stratigraphic column of southern Iraq.

°C). A more detail on how this apparatus work and the derived parameters can be found in (Espitalie et al., 1977; Langford and Blanc-Valleron, 1990; Peters and Cassa, 1994). Table (1) showed the derived TOC for the study area's core samples.

3.2. Logs data

Previous studies have shown that the most sensitive logs for the presence of organic matter in the rocks are gamma ray (GR), deep resistivity (RT), density (DN), acoustic (AC), and neutron (NCL). The higher the organic content of rocks, the greater the anomalies in the response of these logs (Wang et al., 2019). Organic matter as a component of sedimentary rocks, has a relatively low density, slow velocity, and is high in hydrogen content, and often exhibits abnormally high uranium levels. The GR log is a record of a formation's radioactivity. The radiation emanates from uranium, thorium, and potassium which occur naturally (Rider, 2002). Rocks with a high content of organic matter have high concentrations of radioactive elements and thus increase the GR response (Bolandi et al., 2017). Resistivity logs are recording a formation's resistivity. They can be used to infer porosity, water saturation, and the presence of hydrocarbon (Evenick, 2008). RT is a resistivity log that measures the true formation resistivity and mainly used for estimating water saturation. The relationship between TOC content and the response of the RT log is not well understood; however, the recorded resistivity values against mature source rocks are high (Schmoker and Hester, 1989). The RT response in a source rock is influenced by the type of fluid and the maturity level of organic matter (Khoshnoodkia et al., 2011). When the immature formation has brine, the RT records low resistivity values; in contrast, when the matured source rocks are filled with hydrocarbon, the RT log measures high resistivity (Nixon, 1973). The DEN log measures the porosity of a formation based on the assumed density of the formation and drill fluid (Evenick, 2008). Formation bulk density is a function of matrix density, porosity, and fluids contained in the pore space. In general, the organic material has a density near 1.0 g/cm^3, while the average of shale mineral density is 2.7 g/cm^3. Therefore, this difference in density between the rock matrix and contained organic material causes significant changes in formation density and response of DEN log. Based on this fact, the organic content can be computed directly from DEN log if variations from other causes are taken into consideration (Schmoker, 1979). AC log, on the other hand, is a tool used to measure the travel time of elastic water through the formation. The AC records are mainly a function of lithology, porosity, and type of fluids. According to (Dellenbach et al., 1983), the AC is higher in immature source rock than the mature interval and therefore, the AC log offers an indirect method to quantify TOC content. Finally, NCL measures the number of neutrons

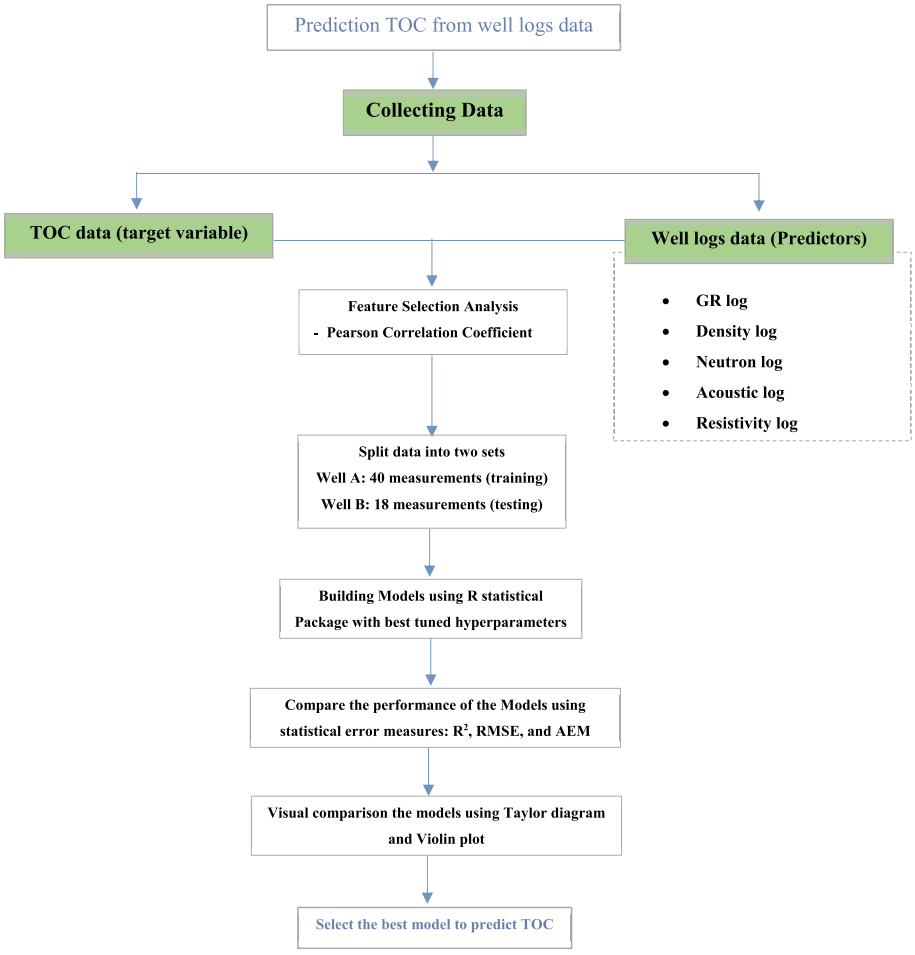

Fig. 3. Steps adopted in this study to predict TOC from well logs data.

Table 1
TOC data of Well R-167 and R-172 in NR oilfield.

Well	Formation	No. of Samples	Depth range (m)	Average TOC (%)	Source Rock generating potential	Percentage of Samples have < 1% TOC content
R-167	Zubair	11	3388–3775	1.01	Good	50
	Ratawi	2	3854–3360	0.89	Fair	50
	Yamama	11	3896–4210	1.23	Good	70
	Sulaiy	16	4220–4510	2.90	Very Good	70
R-172	Zubair	4	3100–3301	1.08	Good	50
	Ratawi	2	3400–3450	1.05	Good	50
	Yamama	12	3700–4110	3.16	Very Good	70

scattered from the formation after it is exposed to a neutron source. It is mainly used for estimating formation porosity. As the organic matter has a direct relationship with hydrogen atoms and porosity of the rock, the NCL curve increase against the organic-rich intervals.

3.3. The used techniques

3.3.1. Feature selection using pearson correlation coefficient

Feature selection (FS) is the process by which the irrelevant and redundant features are identified and removed from a training data set. This process reduces the dimensionality of the data and may enable ML algorithms to work more efficiently. FS decreases the complexity of an ML model and increases the efficiency of the features (Al-Abadi et al., 2019). Selecting only a minimum set of informative and relevant features could improve the robustness of models for learning parameters, classifying samples, or predicting response from a large amount of data (Saethang et al., 2008). The Pearson's correlation coefficient (r) was used in this study to test the relevance of the attributes (factors) for estimating the TOC. r is a measure of the linear correlation between two variables x and y. r is the covariance of the two variables divided by their standard deviations. It is calculated as:

$$r_{x,y} = \frac{cov(x,\ y)}{\sigma_x \sigma_y} \quad (1)$$

Where *cov* is the covariance, σ_x is the standard deviation of x, σ_y is the standard deviation of y.

3.3.2. rF algorithm

rF is an ensemble method, originally developed for classification problems. It is based on constructing each classifier with features obtained by rotating subspaces of the original dataset (Al-Abadi, 2018). The feature set is randomly divided into k subset to create the training data for a base classifier. The principal component analysis (PCA) is then applied to each subset. All principal components (PC) are maintained to keep the variability information in the data (Rodriguez et al., 2006). To create new features for a base classifier, the k is rotated to prompt both individualism and diversity within the ensemble. Diversity is achieved by using PCA to extract features for each base classifier, while accuracy is prompted by taking all the PC components and using all the datasets to train each base classifier (Zhang et al., 2008). Successful application of rF is dependent on the rotation matrix generated by the methods of transformation and the base classifiers chosen (Xia et al., 2017).

3.3.3. Random forest (RF)

RF is also an ensemble learning algorithm designed for handling both regression and classification problems with the use of multiple decision trees (DT) and a bootstrap aggregation (bagging) technique. Bagging builds multiple DT from resampled data and combined the predicted values through averaging and voting. Approximately, 2/3 of the learning samples are used for training, while the remaining 1/3 is used for validation (the out-of-bag OOB) (Al-Abadi and Shahid, 2016). The RF algorithm has the inherent ability to estimate a feature's importance by evaluating how much the prediction error increases when OOB is permuted for the feature while all others remain unchanged (Catani et al., 2013). RF is also capable to handle the missing values and maintains accuracy at the same time. It also resistance to the overfitting problem, and easily handle large dataset with higher dimensionality. Two hyperparameters need to be tune in the RF to get the best results: the number of regression trees (*ntree*; the default value is 500 trees) and the number of input features per node (*mtry*: the default value is 1/3 of the total number of features).

3.3.4. K-nearest neighbor (KNN)

KNN is an algorithm that is non-parametrically supervised and can be used for classification and regression problems. The KNN predicts a new sample from the training set using the K-closest samples (Kuhn and Johnson, 2013). To classify or predict a new case, the KNN relies on finding “similar” cases in the training data (Shmueli et al., 2016). These “neighbors” are then utilized to predict the new case by voting (for classification) or averaging (for prediction). The KNN is robust to noisy training data and is quite successful when a large training dataset is given (Mitchell, 1997). The main advantage of KNN is its simplicity and lack of parametric assumptions (Shmueli et al., 2016).

3.3.5. Artificial neural networks (ANNs)

ANN is a computing system composed of many simple computation elements integrated across a weighted connection (Isiyaka et al., 2019). ANN is a mimic of how data is synthesized by the biological nervous system. These systems “learn” to perform tasks by looking at examples, usually without programming the task-specific rules. The architecture of ANN is based on a collection of connected units (nodes) named artificial neurons that model the neurons in a biological brain. Each connection can convey a signal to other neurons, like the synapses in a biological brain. It is then processed by an artificial neuron that receives a signal and may signal associated neurons. In ANN applications, the “signal” at a connection is a real number, and some non-linear input sum function calculates the output of each neuron. Neurons and connections have a weight that adjusts as learning progresses. To train ANN, network architecture, weights, learning rate, and proper training algorithm should be carefully chosen and initialize (Wang et al., 2016). In this work, we used the Backpropagation neural network (BPANN) to estimate TOC. BP algorithms are a family of techniques utilized to train ANNs effectively following a gradient-based optimization algorithm. BP fine-tuning the weights of a neural net based on the error rate get in the previous iteration (epoch). Proper weight tuning leads to lower error rates and thus renders the model more effective by increasing its generalization.

3.3.6. Support vector regression (SVR)

Support vector machine (SVM) is a group of supervised kernel-based ML algorithm that can be applied to classification or regression problems. SVM is based on the statistical learning theory (SLT) and the Vapnik-Chervonenkis dimension (Vapnik and Chervonenkis, 1974). The SVM aims to form a hyperplane that gives the optimal separation within linearly separable classes in the space of decision boundary (Pal and Mather, 2005). SVM uses two concepts to optimize a solution: optimal hyperplane classification and kernel function (Handhal et al., 2019; Yao et al., 2008). The SVR is an SVM conversion for regression analysis. There are three conversions of SVM for regression problems: epsilon regression (ε-svr), nu regression (nu-svr), and bound-constraint SVM regression (eps-bsvr). In this work, ε-svr and nu-svr with radial basis function (RBF) kernel were used to estimate TOC. Three hyperparameters must be defined and optimized to optimize these algorithms using RBF kernel functions, namely the cost of constraint violation (C), epsilon (ε), and gamma (γ). The C specifies the trade-off between decision rule complexity and error rate (Cortes and Vapnik, 1995). A small value for C increases the number of training errors, while a large C results in similar behavior to that of a hard-margin SVM (Joachims, 2002). Epsilon (ε) has an effect on the smoothness of the SVM's and thus the complexity and generalization capability of the network (Horvath, 2003). On the other hand, γ is related to overfitting and underfitting problems (Rashidi et al., 2016).

3.3.7. Data normalization and model evaluation criteria

Data normalization is the process of organizing data to reduce redundancy and improve data integrity. There are various types of standardization methods such as min-max, decimal scaling and standard method of deviation. Choosing a normalization method depends on the application and the algorithm in which the normalized data will be used. In this study, data were normalized in the range [0, 1] using min-max scaling function according to the following equation:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

where x_{norm}, x_{min}, x_{max} are the normalized, minimum, and maximum of input data, respectively.

To evaluate the performance of the model used, three statistical measures were used: the mean absolute error (MAE), the root mean squared error (RMSE), and the correlation of determination coefficient (R^2). The measures are given by:

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{3}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - x_i)^2}{n}} \tag{4}$$

$$R^2 = \frac{explained\ variation}{total\ variation} = \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \right)^2 \quad (5)$$

where x_i and y_i are the measured and estimated values of TOC, and $\bar{x}$ and $\bar{y}$ are their arithmetic mean; n is the total number of measured TOC data.

To visually compare the performance of models used in this study, Taylor diagrams and violin plots were used. The Taylor diagram is a polar plot used to summarize multiple aspects of model performance in a single diagram (Taylor, 2001). It uses three statistics to evaluate the degree of correspondence between the estimated and measured values: the r, RMSE, and the standard deviation (σ). A violin plot is a plot similar to a boxplot with the addition of a rotated kernel density on each side (Hintze and Nelson, 1998).

3.3.8. Software used

The feature selection and the ML algorithms used in this study were implemented using WEKA 3.8 software. WEKA is an open source ML software that can be accessed through a graphical user interface, standard terminal applications, or a Java API. It is commonly used in education, research, and industrial applications and include a variety of advanced methods for standard ML tasks.

4. Results and discussion

4.1. TOC data

The derived TOC values in Table (1) indicated that more than 50% of the samples taken from Zubair Formation have TOC content less than 1%, representing a poor source rock according to the Peter's classification (Peters, 1986). Although this formation is a main clastic reservoir in the study are, it acts as a source rocks because the abundance of shale layers between the main sand units. In contrast, more than 70% of the samples taken from Yamama and Sulaiy formations have TOC content greater than 1% representing good source rocks according to Peters (1986).

4.2. Feature selection

WEKA supports correlation-based feature selection with the CorrelationAttributeEval technique that requires the use of a Ranker search method. Running this in our dataset suggested that all regressors used have average merits ≠ 0 (Table 2 and Fig. 4); therefore, they all have impacts on the estimating TOC. The GR with average merit (AM) equals to 0.431 was the most important feature in the analysis of TOC, followed by DN (AM = 0.379), NCL (AM = −0.276), and RT (AM = 0.250). The AC log with AM equal to −0.053 (very close to zero) confirmed that this log may play a minor role in determining TOC in the study area. Therefore, all factors were used in further analysis.

Table 2
Feature selection using *r* (10-fold cross validation).

Feature	AM	σ	Rank
GR	0.431	± 0.021	1
DN	0.379	± 0.028	2
RT	0.250	± 0.079	4
AC	−0.053	± 0.034	5
NCL	−0.276	± 0.033	3

AM: Average merit; σ: Standard deviation.

4.3. Applying the models

The ML models were constructed using the training dataset (well R-167) and validated by the testing dataset (well R-172) using WEKA 3.8 software. The rF was constructed using the rotation forest package and the base classifier used is RF. The default parameters of the algorithm was changed using trial and error procedure to get the best performance with minimal error. The final obtained results were gotten with maxGroup and minGroup equal to 3, the number of iteration equal to 20, and seed equals to 1. The obtained results of rF algorithm were shown in Table (3) for both training and testing phases It can be concluded from Table (3) that rF was a very good performance model because of the high obtained R^2 and relatively small values of MAE and RMSE.

The RF algorithm was fitted using the Random Forest package in WEKA software and the obtained results of RF for training and testing were presented in Table (3). A key hyperparameter for RF is the mtry (the number of attributes to consider in each split point). In WEKA, this hyperparameter can be governed by the number Features attribute, which is 0 by default, which automatically chooses the value based on a thumb rule. The other important hyperparameter is mtry (numIterations in WEKA, the default value is 100). The parameter is changed manually (Table 4) to judge the best value according to the error statistics used. The best performance model was with 1000 mtry. The high R^2 and low AEM and RMSE indicated that RF is a very good choice for modeling TOC. Investigate the importance of attributes (logs) based on average impurity decrease (Fig. 5) revealed that the most important attribute was GR, followed by DN, RT, NCL, and AC. This result confirmed the importance of GR, DN, and RT as powerful regressores in estimating TOC.

The kNN algorithm was implemented using IBk package. To configure IBk, two hyperparameters were tuned. These were the number of neighbors to query to make a prediction (k) and the distance metric (the way in which the neighbors are determined). Results of chaning k with kept distance matric on LinearNNSearch algorithm were shown in Table (5). It is obvious that the kNN performance dramatically getting worse with changing k from 1 to 3 and then to 7. When the k is 1, the kNN model was almost perfect in both training and testing stages (Table 4) which indicated that advanced ML models are not always the best. The algorithm may be simple, but it can yield results far superior to those of more complex (in our experiment, rF, RF, ANN, and SVM).

ANN model was fitted using Multi-Layer Perceptron algorithm after configured it to get the best performance. For the best model, the number of the hidden layer was set to be 2, the learning rate was 0.1, and the momentum was 0.3. The final architecture of the BPANN was 5: 3: 1. After successful the training of the network, the test dataset was passed to the network and the results were shown in Table 4. Overall, the performance of this model was somewhat poor (especially in the testing stage) due to the low R^2 and the high MAE and RMSE. The reason may be the small number of the training samples used (only 40 values) and associated overfitting problems.

To fit SVR, LibSVM library was used. The optimized performance of the ε-svr was obtained using 4 degrees of the RBF kernel, 0.001 for ε, 2.0 for γ, and 0.1 for the loss function (Table 4). In general, the performance of this model was better than the ANN, but it is less efficient than the rF, RF, and kNN models in terms of statistical error used. For the other version of the SVR algorithm, nu-svr, the recommended values for the best performance were 3 degrees of the RBF kernel, 0.001 for ε, 0.0 for γ, and 0.5 for nu. The results of implementing this version of SVR (Table 4) were better than ε-svr, but the SVR still low performance model if compare with other used ML algorithms.

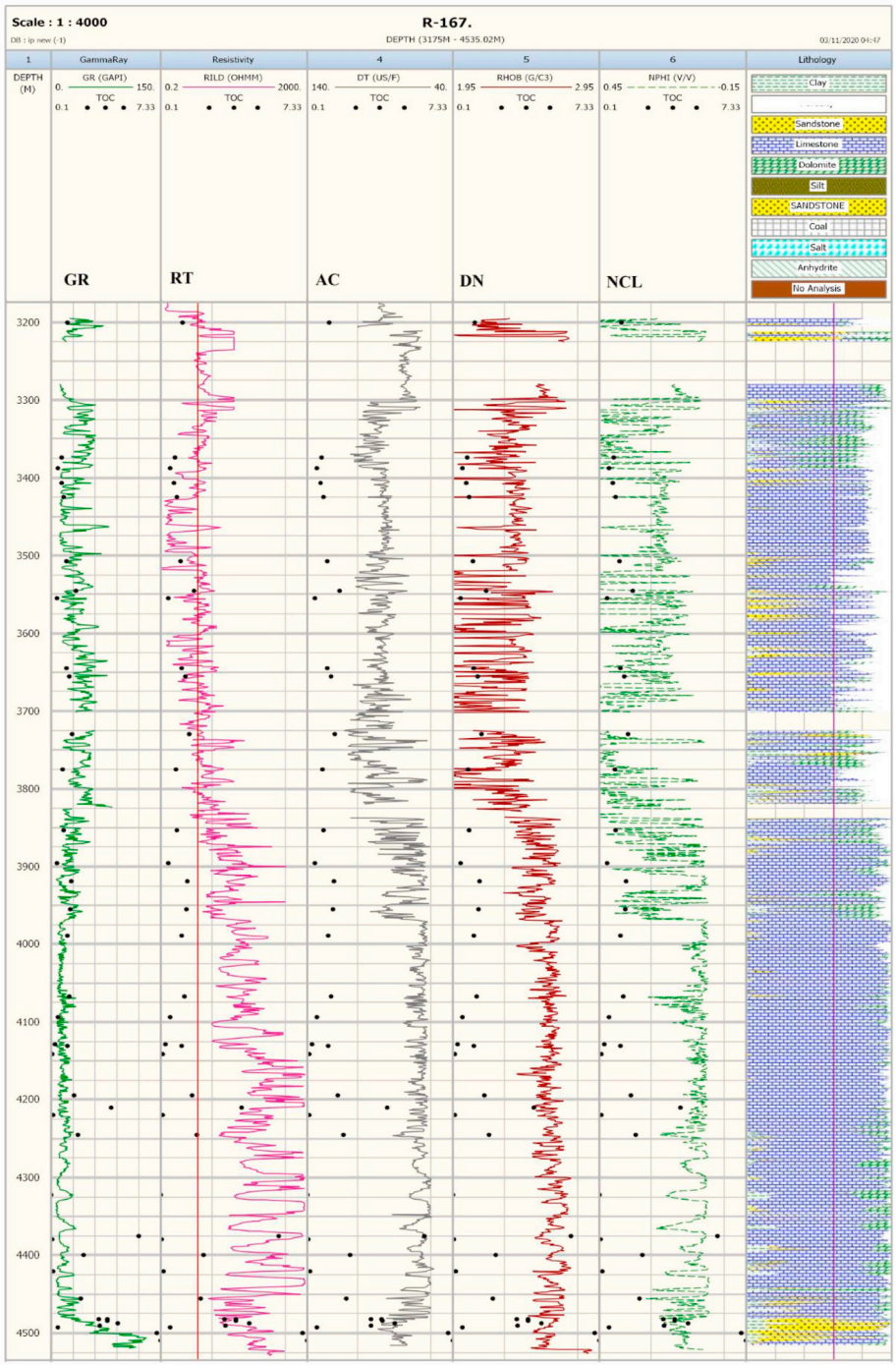

Fig. 4. The relationship between TOC, well logs data, and lithology.

4.4. Visual comparison of models used

A direct comparison of the model results in the testing stage was shown in Fig. (6). It is clear that TOC was predicted by kNN model with highest accuracy compared to other models used. RF and rF comes second, followed by the BPANN and SVR models. The viusal check of Fig. (6) confirms a high accuracy of the third ML models (kNN, RF, and rF) in estimating the TOC.

Comparing the performance of the ML model used using Taylor Diagram (Fig. 7) showed that the kNN was the best of all models used for high R^2 and low value of RMSE. The RF and rF models were almost similar in performance while the worst model was BPANN and ε-svr models. Using standard deviation (σ) for comparing the models, results showed that kNN, RF, and rF were more in agreement than the others and closer to the observed values. All remaining models have a lower standard deviation than the ones observed, and thus, have no ability to predict the observed data very well. The worst models were nu-svr and ε-svr.

Table 3
ML models performances in training and testing stages.

SEMs	ML models											
	rF		RF		KNN		BPANN		ε-svm		nu-svm	
	training	testing	Training	testing	training	testing	training	testing	training	testing	training	testing
MAE	0.313	0.367	0.336	0.407	0.000	0.001	0.833	0.807	0.631	0.617	0.848	0.902
RMSE	0.416	0.533	0.426	0.564	0.000	0.002	0.966	1.101	0.905	1.145	0.944	1.179
R^2	0.946	0.939	0.954	0.948	1.000	0.996	0.724	0.632	0.832	0.788	0.916	0.814

SEMs: statistical error measures.

Table 4
Effect of changing mtry on RF performance (training stage).

SEMs	mtry			
	100	500	1000	1500
MAE	0.375	0.358	**0.336**	0.351
RMSE	0.474	0.456	**0.426**	0.446
R2	0.925	0.945	**0.954**	0.951

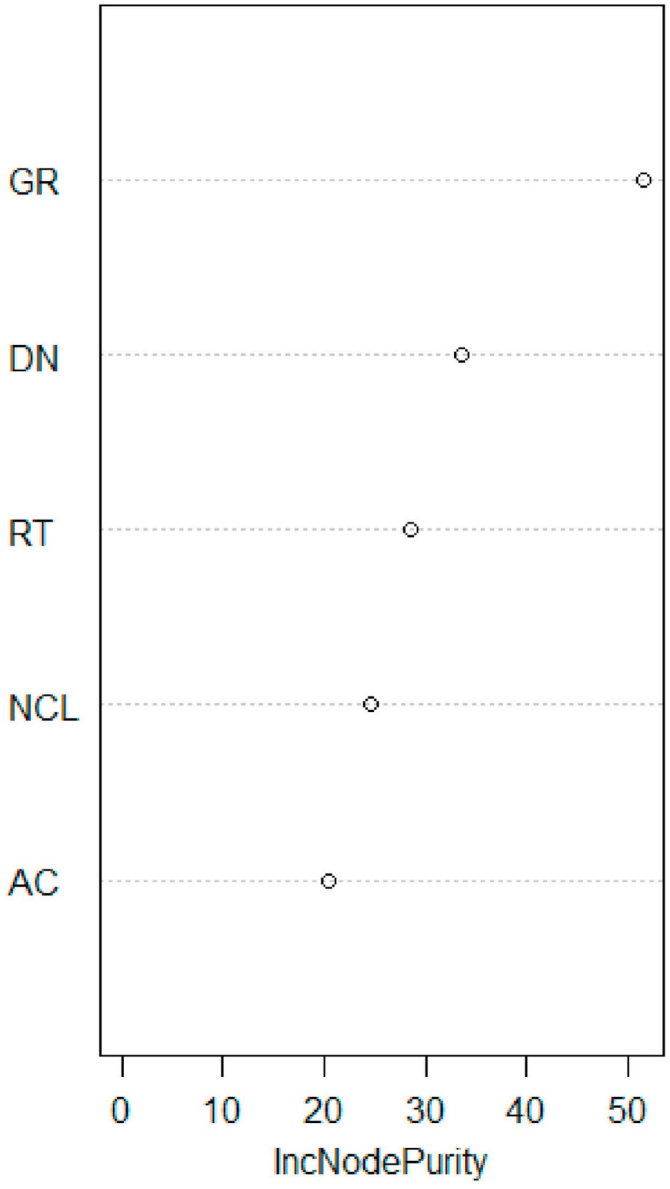

Fig. 5. Attribute importance based on the average impurity decrease.

Table 5
Effect of changing k on kNN performance (training stage).

SEMs	k		
	1	3	7
MAE	0	0.864	0.909
RMSE	0	1.008	1.282
R^2	1.0	0.536	0.465

Comparing the model by the violin plot (Fig. 8) indicated the high performance of kNN, RF, and rF (perfect fit of kNN and very good fits for RF and rF) in contrast to the BPANN and SVM models. The median (white points in violin plot) was very well predicted by kNN (Fig. 8), while the 5th and 95th percentiles (thin black lines in the plot) and the 25th (first quartiles Q1) and 75th (third quartiles Q3) percentiles (thick lines in the plot) in kNN model gave the highest perfect fit than the other used models. RF and rF models were similar in performance and they approached the kNN model in all statistical measures positions used. The BPANN overestimated the 5th and underestimated 95th percentiles, ranges of TOC respectively, while SVR models underestimate 5th percentile. In terms of the shape of the violin plot (the probability density function (PDF)), the PDF of the kNN model gave the closest fit to measurements TOC data, followed by RF and rF. The PDF of BPANN, ε-svr, and nu-svr models are totally different from the observed PDF of measurements TOC data. Therefore, these models have the worst fit with the measured TOC data.

Although the ANN and SVR models have been successfully utilized to predict TOC in the previous studies (Khoshnoodkia et al., 2011; Wang et al., 2019), This study showed that in estimating TOC, the kNN and ensemble ML algorithms (RF and rF) were superior to the BPANN and SVR models. Therefore, these undiscovered yet powerful algorithms in the oil industry can be important tools in TOC modeling. It was clear in this study that BPANN and SVR did not well capture the behavior of the measured TOC data as accurately as the kNN, RF, and rF models. The present study also indicated that nu-svm technique was slightly better than ε-svm and BPANN model. However, the nature of the problem and input data are the main criterion that determines which algorithm is the best (Al-Abadi et al., 2019).

5. Conclusions

In this study, five ML algorithms, specifically, rF, RF, kNN, BPANN, and SVR with radial kernel function were developed to estimate TOC (target variable) using five conventional logs namely, GR, RT, DN, AC, and NCL (regressors). The ML models were trained and tested using data from two wells in the field, namely R-167 and R-172. The performance of the developed models was compared using three error statistics criteria: MAE, RMSE, and R^2 and visually using Taylor and violin plots. Feature selection was firstly used for data screening using the Pearson correlation coefficient and, this stage of the analysis indicated that all logs used were relevant. The application of the models

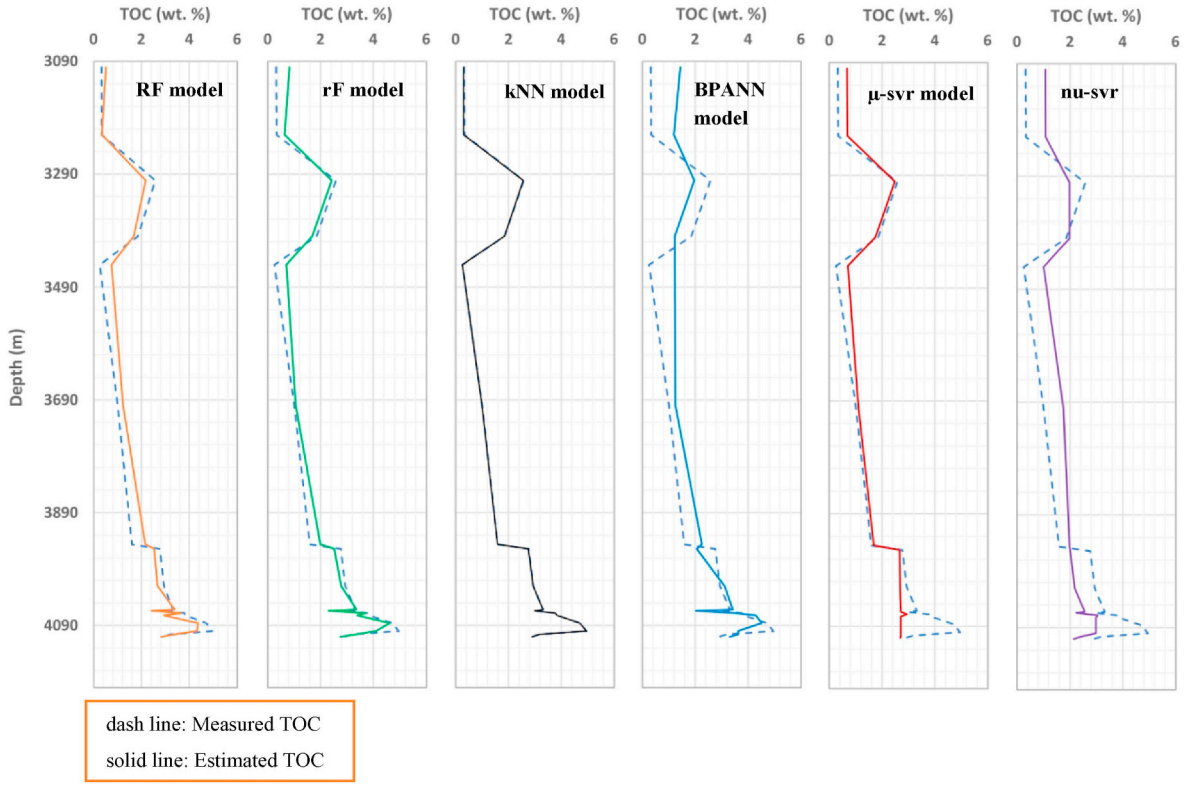

Fig. 6. The measured and estimated TOC for different ML models (testing stage).

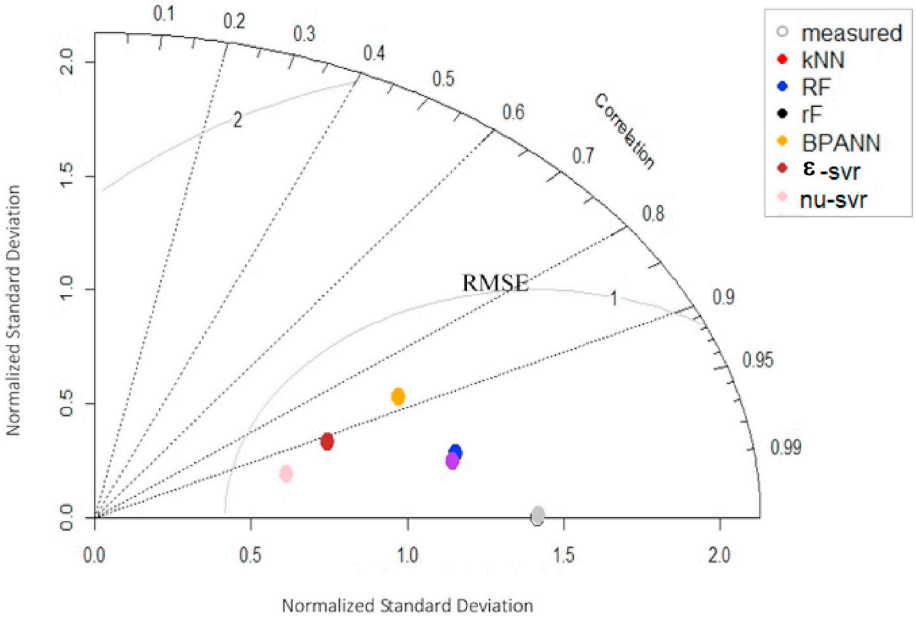

Fig. 7. Taylor diagram for visual comparison of ML models used (testing stage).

showed that kNN was the best model, followed by RF, and rF in terms of the error measures used. The low predictive ability models were BPANN and SVR. The applications of different ML generally result in models differing in their response as predictions depend on the nature of the data used and availability of the different methods and computing power. The reason behind why kNN, a simple ML algorithm, gave the best results than other models may be attributed to: (1) simplicity and lack of parametric assumptions (2) robust to noise (3) easily handle all types of data (categorical and continuous). This study indicated that advanced ML models are not always the best. The algorithm may be simple, but it can yield results far superior to those of more complex This study also confirmed the efficacy of machine learning models for building efficient models for estimating TOC from readily available borehole logs data without the need for very expensive coring process.

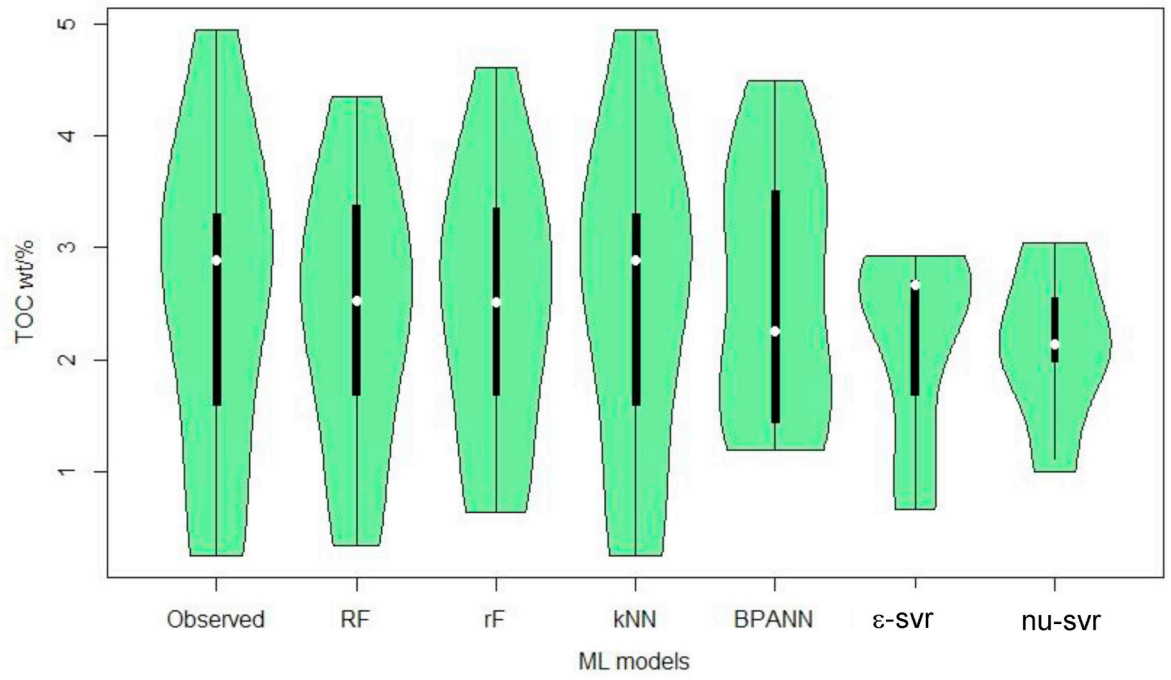

Fig. 8. Violin plot for visual comparison of ML models used (testing stage).

CRediT authorship contribution statement

Amna M. Handhal: Conceptualization, Supervision, Writing - original draft. **Alaa M. Al-Abadi:** Methodology, Writing - original draft, Supervision, Software, Visualization, Writing - review & editing. **Hussein E. Chafeet:** Validation, Formal analysis, Writing - original draft. **Maher J. Ismail:** Data curation, 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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.marpetgeo.2020.104347.

References

Abdizadeh, H., Kadkhodaie, A., Ahmadi, A., Heidarifard, M.H., 2017. Estimation of Total Organic Carbon from well logs and seismic sections via neural network and ant colony optimization approach: a case study from the Mansuri oil field, SW Iran. Geopersia 7 (2), 255–266.

Al-Abadi, A.M., 2018. Mapping flood susceptibility in an arid region of southern Iraq using ensemble machine learning classifiers: a comparative study. Arab. J. Geosci. 11. https://doi.org/10.1007/s12517-018-3584-5.

Al-Abadi, A.M., Handhal, A.M., Al-Ginamy, M.A., 2019. Evaluating the Dibdibba aquifer productivity at the Karbala–Najaf plateau (Central Iraq) using GIS-based tree machine learning algorithms. Nat. Resour. Res. https://doi.org/10.1007/s11053-019-09561-x.

Al-Abadi, A.M., Shahid, S., 2016. Spatial mapping of artesian zone at Iraqi southern desert using a GIS-based random forest machine learning model. Model. Earth Syst. Environ. 2 (2), 96.

Al-Ameri, T.K., Al-Musawi, F.A., Batten, D.J., 1999. Palynofacies and source potential for hydrocarbon, uppermost Jurassic-basal Cretaceous in Sulaiy Formation, southern Iraq. Cretac. Res. 20, 359–363.

Al-Ameri, Thamer Khazaal, Al-Musawi, F.A., 2011. Hydrocarbon generation potential of the uppermost Jurassic—basal Cretaceous Sulaiy formation, South Iraq. Arab. J. Geosci. 4 (1–2), 53–58.

Al-Naqib, K.M., 1967. Geology of the Arabian Peninsula. USGS Professional Paper No.

Al-Sayyab, A., 1989. Geology of Petroleum. University of Baghdad Press, Baghdad.

Aqrawi, A.A.M., Thehni, G.A., Sherwani, G.H., Kareem, B.M.A., 1998. Mid-Cretaceous rudist-bearing carbonates of the Mishrif Formation: an important reservoir sequence in the Mesopotamian Basin, Iraq. J. Petrol. Geol. 21 (1), 57–82.

Aqrawi, Adnan A.M., Goff, J.C., Horbury, A.D., Sadooni, F.N., 2010. The Petroleum Geology of Iraq. Scientific Press.

Bellen, R. C. Van, Dunnington, H.V., Wetzel, R., Morton, D., 1959. Lexique Stratigraphique International, vol. 3 Asie, Iraq 10a.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. springer.

Bolandi, V., Kadkhodaie, A., Farzi, R., 2017. Analyzing organic richness of source rocks from well log data by using SVM and ANN classifiers: a case study from the Kazhdumi formation, the Persian Gulf basin, offshore Iran. J. Petrol. Sci. Eng. 151, 224–234.

Catani, F., Lagomarsino, D., Segoni, S., Tofani, V., 2013. Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Nat. Hazards Earth Syst. Sci. 13 (11), 2815–2831.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

Dellenbach, J., Espitalie, J., Lebreton, F., 1983. Source Rock Logging: Transactions of the 8th European SPWLA Symposium. paper D.

Espitalie, J., Madec, M., Tissot, B., Mennig, J.J., Leplat, P., 1977. Source rock characterization method for petroleum exploration. Offshore Technology Conference. Offshore Technology Conference Estimation of Total Organic Carbon from well logs and seismic sections via neural network and ant colony optimization approach: a case study from the Mansuri oil field, SW Iran. (2017). Geopersia 7, 255–266 2.

Evenick, J., 2008. Introduction to Well Logs and Subsurface Maps. PennWell Books.

Farzi, R., Bolandi, V., 2016. Estimation of organic facies using ensemble methods in comparison with conventional intelligent approaches: a case study of the South Pars Gas Field, Persian Gulf, Iran. Model. Earth Syst. Environ. 2 (2), 105.

Handhal, A.M., Jawad, S.M., Al-Abadi, A.M., 2019. GIS-based machine learning models for mapping tar mat zones in upper part (DJ unit) of Zubair Formation in North Rumaila supergiant oil field, southern Iraq. J. Petrol. Sci. Eng. 178, 559–574.

Hintze, J.L., Nelson, R.D., 1998. Violin plots: a box plot-density trace synergism. Am. Statistician 52 (2), 181–184.

Horvath, G., 2003. CMAC neural network as an SVM with B-Spline kernel functions. In: Proceedings of the 20th IEEE Instrumentation Technology Conference (Cat. No. 03CH37412), vol. 2. IEEE, pp. 1108–1113.

Huang, Z., Williamson, M.A., 1996. Artificial neural network modelling as an aid to source rock characterization. Mar. Petrol. Geol. 13 (2), 277–290.

Isiyaka, H.A., Mustapha, A., Juahir, H., Phil-Eze, P., 2019. Water quality modelling using artificial neural network and multivariate statistical techniques. Model. Earth Syst. Environ. 5 (2), 583–593.

Jarvie, D.M., 1991. Total Organic Carbon (TOC) Analysis: Chapter 11: Geochemical Methods and Exploration.

Jassim, S.Z., Goff, J.C., 2006. Geology of Iraq. Dolin, Prague Moravian Mus. Brno 2006, 341.

Joachims, T., 2002. Learning to Classify Text Using Support Vector Machines, vol. 668 Springer Science & Business Media.

Kadkhodaie-Ilkhchi, A., Rahimpour-Bonab, H., Rezaee, M., 2009. A committee machine with intelligent systems for estimation of total organic carbon content from petrophysical data: an example from Kangan and Dalan reservoirs in South Pars Gas Field, Iran. Comput. Geosci. 35 (3), 459–474.

Kamali, M.R., Mirshady, A.A., 2004. Total organic carbon content determined from well

logs using ΔLogR and Neuro Fuzzy techniques. J. Petrol. Sci. Eng. 45 (3–4), 141–148.
Khoshnoodkia, M., Mohseni, H., Rahmani, O., Mohammadi, A., 2011. TOC determination of Gadvan Formation in South Pars Gas field, using artificial intelligent systems and geochemical data. J. Petrol. Sci. Eng. 78 (1), 119–130.
Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling, vol. 26 Springer.
Langford, F.F., Blanc-Valleron, M.-M., 1990. Interpreting Rock-Eval pyrolysis data using graphs of pyrolizable hydrocarbons vs. total organic carbon (1). AAPG (Am. Assoc. Pet. Geol.) Bull. 74 (6), 799–804.
Mitchell, T.M., 1997. 1997. Burr Ridge. Machine learning 45. IL: McGraw Hill, pp. 870–877 37.
Nixon, R.P., 1973. Oil source beds in Cretaceous Mowry Shale of northwestern interior United States. AAPG (Am. Assoc. Pet. Geol.) Bull. 57 (1), 136–161.
Ouadfeul, S.-A., Aliouane, L., 2015. Total organic carbon prediction in shale gas reservoirs from well logs data using the multilayer perceptron neural network with Levenberg Marquardt training algorithm: application to Barnett shale. Arabian J. Sci. Eng. 40 (11), 3345–3349.
Owen, R.M.S., Nasr, S.N., 1958. Stratigraphy of the Kuwait-Basra Area: Middle East.
Pal, M., Mather, P.M., 2005. Support vector machines for classification in remote sensing. Int. J. Rem. Sens. 26 (5), 1007–1011.
Passey, Q.R., Creaney, S., Kulla, J.B., Moretti, F.J., Stroud, J.D., 1990. A practical model for organic richness from porosity and resistivity logs. AAPG (Am. Assoc. Pet. Geol.) Bull. 74 (12), 1777–1794.
Peters, K.E., 1986. Guidelines for evaluating petroleumsource rock using programmed pyrolysis. Am. Assoc. Petrol. Geol. Bull. 70, 318–329.
Peters, Kenneth E., Cassa, M.R., 1994. Applied source rock geochemistry: chapter 5: Part II. Essent. Elem.
Rahmani, O., Khoshnoodkia, M., Kadkhodaie, A., Beiranvand Pour, A., Tsegab, H., 2019. Geochemical analysis for determining total organic carbon content based onΔ LogR technique in the south pars field. Minerals 9 (12), 735.
Rashidi, S., Vafakhah, M., Lafdani, E.K., Javadi, M.R., 2016. Evaluating the support vector machine for suspended sediment load forecasting based on gamma test. Arab. J. Geosci. 9 (11), 583.
Rider, M.H., 2002. Rogart. second ed. The Gamma Ray and Spectral Gamma Ray Logs. The Geological Interpretation of Well Logs, vol. 71. Rider-French Consulting Ltd, Whittle Publishing, pp. 74.
Rodriguez, J.J., Kuncheva, L.I., Alonso, C.J., 2006. Rotation forest: a new classifier ensemble method. IEEE Trans. Pattern Anal. Mach. Intell. 28 (10), 1619–1630.
Sadooni, F.N., 1993. Stratigraphic sequence, microfacies, and petroleum prospects of the Yamama Formation, Lower Cretaceous, southern Iraq. AAPG (Am. Assoc. Pet. Geol.) Bull. 77 (11), 1971–1988.
Saethang, T., Prom-On, S., Meechai, A., Chan, J.H., 2008. Sample filtering relief algorithm: robust algorithm for feature selection. In: International Conference on Neural Information Processing. Springer, pp. 260–267.
Schmoker, J.W., 1979. Determination of organic content of Appalachian Devonian shales from formation-density logs: geologic notes. AAPG (Am. Assoc. Pet. Geol.) Bull. 63 (9), 1504–1509.
Schmoker, J.W., Hester, T.C., 1989. Oil generation inferrred from formation resistivity–Bakken formation, Williston basin, north Dakota. In: SPWLA 30th Annual Logging Symposium. Society of Petrophysicists and Well-Log Analysts.
Sfidari, E., Kadkhodaie-Ilkhchi, A., Najjari, S., 2012. Comparison of intelligent and statistical clustering approaches to predicting total organic carbon using intelligent systems. J. Petrol. Sci. Eng. 86, 190–205.
Shmueli, G., Bruce, P.C., Patel, N.R., 2016. Data mining for business analytics. Retrieved from In: Data Mining for Business Analytics: Concepts, Techniques and Applications with XLMiner 3. https://edu.kpfu.ru/pluginfile.php/274079/mod_resource/content/2/DatMiningBusAnalytics.pdf.
Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res.: Atmospheres 106 (D7), 7183–7192.
Vapnik, V., Chervonenkis, A., 1974. Theory of Pattern Recognition. Nauka, Moscow.
Wang, H., Wu, W., Chen, T., Dong, X., Wang, G., 2019. An improved neural network for TOC, S1 and S2 estimation based on conventional well logs. J. Petrol. Sci. Eng. 176, 664–678.
Wang, L.-J., Guo, M., Sawada, K., Lin, J., Zhang, J., 2016. A comparative study of landslide susceptibility maps using logistic regression, frequency ratio, decision tree, weights of evidence and artificial neural network. Geosci. J. 20 (1), 117–136.
Xia, J., Falco, N., Benediktsson, J.A., Du, P., Chanussot, J., 2017. Hyperspectral image classification with rotation random forest via KPCA. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 10 (4), 1601–1609.
Yao, X., Tham, L.G., Dai, F.C., 2008. Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of Hong Kong, China. Geomorphology 101 (4), 572–582.
Zhang, C.-X., Zhang, J.-S., Wang, G.-W., 2008. An empirical study of using Rotation Forest to improve regressors. Appl. Math. Comput. 195 (2), 618–629.