International Journal of Innovative Computing, Information and Control
Volume **17**, Number **3**, June **2021**

ICIC International ©2021 ISSN 1349-4198
pp. **807–829**

EVALUATION OF TREE-BASED ENSEMBLE LEARNING ALGORITHMS TO ESTIMATE TOTAL ORGANIC CARBON FROM WIRELINE LOGS

Md Shokor A Rahaman[1], Pandian Vasant[1], Izhar ul Haq[2]
Eswaran Padmanabhan[2], Shiferaw Regassa Jufar[3]
Rajalingam Sokkalingam[1] and Junzo Watada[4]

[1]Department of Fundamental and Applied Sciences
[2]Department of Petroleum Geoscience
[3]Petroleum Engineering Department
Universiti Teknologi PETRONAS
Tronoh, Perak 31750, Malaysia
{ shokor103072; pvasant; for.izhar }@gmail.com
{ eswaran_padmanabhan; shiferaw.jufar; raja.sokkalingam }@utp.edu.my

[4]IPS Research Center
Waseda University
1-104 Totsukamachi, Shinjuku-ku, Tokyo 169-8050, Japan
junzo.watada@gmail.com

Received November 2020; revised March 2021

Abstract. *To evaluate the hydrocarbon generation potential, Total Organic Carbon (TOC) of source/reservoir rocks is of vital importance. TOC estimation from well logs is challenging and in laboratory from rock specimens is costly as well as time-consuming. TOC prediction from Passey method is low whereas AI techniques such as Artificial Neural Network (ANN), Support Vector Machine (SVM) get trapped in local optima, resulting in overfitting and are also considered ambiguous if the technique is not reasonable. In this paper, we proposed four efficient tree-based ensemble techniques that include Random Forest (RF), Extra Trees (ET), Gradient Boosting (GB), and eXtremely Gradient Boosting (XGB), capable of fitting highly non-linear data with minimum data pre-processing for TOC prediction. To evaluate the efficiency of these models, 205 data points and seven well logs from the Goldwyer Formation of the Canning Basin, Australia, were used for the training and testing purpose. Results validated that the accuracy of these tree-based ensemble techniques is at exemplary level for the TOC estimation, where the XGB model (for testing R^2 94.39%, MAE 0.0447, MSE 0.0039) outperformed the other techniques, i.e., RF (for testing R^2 90.59%, MAE 0.0549, MSE 0.0055), ET (for testing R^2 90.63%, MAE 0.0583, MSE 0.0058) and GB (for testing R^2 91.23%, MAE 0.0569, MSE 0.0053). These robust tree-based ensemble techniques have not only protected overfitting but also achieved better prediction results in dealing with the multidimensional data.*
Keywords: Total organic carbon, Well logs, Tree-based ensemble learning, Artificial intelligence

1. **Introduction.** Estimation of hydrocarbon generation potential is an important parameter in oil and gas exploration and requires precise estimation of the Total Organic Carbon (TOC) percentage of the rock. The quantity and quality of the organic matter, defined by TOC, are a significant and a fundamental index. Currently TOC of rocks is determined by two methods, i.e., by the combustion of rock powder in the laboratory and by using empirical correlations developed based on the linear regression analysis.

DOI: 10.24507/ijicic.17.03.807

Schmoker [1] developed the first empirical correlation for TOC estimation in Devonian shale. In this correlation the author assumed that TOC is a direct function of the formation density (Equation (1)). In 1979, Schmoker [1] developed the revised version of his earlier model, applied in the Bakken formation Equation (2). Passey et al. [2] developed a model for TOC estimation known as $\Delta \log R$ model (Equations (3) and (4)). Currently, $\Delta \log R$ is the most widely used model for evaluating the TOC of rocks. Wang et al. [3] suggested consideration of Gamma-Ray (GR) log with resistivity and sonic or density logs to improve the predictability of TOC for Devonian shale. The $\Delta \log R$ in this case is calculated using sonic log (Equation (5)) as well as density log (Equation (6)). To assess the $\Delta \log R$ method, Zhao et al. [4] calculated the TOC by overlaying clay content curve with gamma ray curve and found this method of calculating TOC better than the Passey et al.'s and Wang et al.'s model.

An important aspect observed in the recent studies is that for the non-linear implicit function, artificial intelligence proves to be more prominent. AI techniques have been used for estimating TOC from wireline logs by Huang et al. [5-8]. In their work, a relationship between well logs parameters and TOC has been established by Artificial Neural Network (ANN) which has shown that the prediction accuracy was closely related to the involved algorithm and kernel functions. A 3 layered BP-NN model has been used by Sfidari et al. [6] to predict TOC. Ouadfeul and Aliouane [9] used ANN to determine the TOC content using Schmoker's model in Barnett shale. Furthermore, Tan et al. [8] used three different regression algorithms and four different kernel functions to predict TOC content and Shi et al. [7] used MLP-NN and fuzzy logic to predict the TOC from the well logs to replace the Schmoker's model. Extreme Learning Machine (ELM) was introduced by Shi et al. [7] for TOC estimation where ELM was found quicker than ANN and Zhu et al. [10] introduced Integrated Hybrid Neural Network (IHNN) in Jaioshiba zone. SVM with RBF outperformed ANN in Bolandi et al.'s work [11] to predict TOC from well logs. Mahmoud et al. [12] and Johnson et al. [13] used ANN while Wang and Peng [14] proposed CNN which outperformed ANN to predict TOC using well logs.

Since the properties of organic-rich rocks vary extremely in different resource plays, $\Delta \log R$ model proposed by Passey et al. [2] could predict the TOC incorrectly when applied into a Formation different from the one used to build it. Another drawback of the $\Delta \log R$ model is that it was built based on a $1 : 50$ linear relationship between the porosity and logarithmic resistivity as indicated in Equation (3); this assumption restricts the applicability of the $\Delta \log R$ method to a limited range of formation porosity and resistivity. Charsky and Herron [15] evaluated the predictability of Schmoker and $\Delta \log R$ correlations in four various wells of different formations. The outcomes of their study showed that both models predicted the TOC with high average absolute differences of 1.6 wt.% and 1.7 wt.% from the actual TOC content. The model proposed by Wang et al. [14] removed the approximation of the linear relationship between porosity and resistivity and suggested estimation of these slopes based on the properties of the target formation. Due to the complicated non-linear function relation between the logging information and TOC content, as found in the Heidari [53] study, approximation of the real function relationship by simple linear regression is difficult and an alternative approach is required to estimate TOC content from well logs. The available empirical correlations developed based on the linear regression were made to learn and estimate TOC in a particular Formation. Therefore, to apply the same correlation in a different formation, the correlation must be modified according to the properties of the target formation. Decision trees, Artificial Neural Networks (ANNs) [16,17], Support Vector Machines (SVMs) [18,19] are well-known

AI techniques among data-driven methods but these algorithms are vulnerable to overfitting in front of high dimensional inputs, which will lead to deceptive diagnostic results [20,21].

To improve the precision in TOC estimation from well logs, it is essential to look for the high accuracy AI techniques. According to this quest, four tree-based ensemble techniques are proposed in this study as a novel research approach for TOC estimation from well logs. Ensemble learning techniques do not overfit easily in multi-dimensional classifier design and can aggregate multiple weight based learning models to obtain a combined model outperforming every single regression model in it [22]. Because of this reason, ensemble learning technique combines different machine learning methods to get a more reliable outcome [23]. In a broad spectrum of applications, ensemble learning is highly accurate and versatile [24-31]. Because, once a machine learning model has been trained, having carefully fine-tuned any hyper-parameters, it is extremely fast to obtain energy predictions for a given set of design inputs, and also understand the correlations between these parameters and energy consumption. Furthermore, tree-based ensemble learning techniques are from ensemble learning techniques that combine properties from machine learning and statistical approach [26]. Therefore, it is particularly popular among other ensemble learning techniques [24,27,29,30,32-34]. Besides obtaining high accuracy level, tree-based ensemble learning allows the interpretation of important features used in the prediction [30,35]. Moreover, these tree-based ensemble techniques are able to fit highly non-linear data and require minimum data pre-processing [23,28].

Given the high level of uncertainties in both the shale reservoir rocks properties and the model tuning parameters, neither the individual AI techniques nor their hybrid formulations could handle more than one hypothesis of the problem at a time [36]. The shale reservoir characterization problem is so complex and full of uncertainties that the existence of diversities of expert opinions that lead to diverse hypotheses needs a more cooperative and robust solution. Such solution should be able to incorporate and integrate existing diversities of expert opinions to solve the complex problems. Despite the success of the individual and HCI models, they are not robust enough to solve the complex problems and handle the uncertainties in the shale oil & gas industry. The tree-based ensemble techniques used in this study for TOC estimation have shown the ability to fit highly non-linear data with minimum data pre-processing. Moreover, the techniques have tackled the high level of uncertainties associated with organic rich source/reservoir rocks and the model tuning parameters at a time.

The approach followed in this study is summarized below.

- Four popular, robust and efficient tree-based ensemble techniques, namely (a) RF, (b) ET, (c) GB and (d) XGB have been investigated as a novel research approach in this paper study for the estimation of the TOC content from the wireline logs.
- Total 205 laboratory measured TOC data points and seven well logs namely GR, DT, RHOB, SP, NPHI, LLD, and LLS are used from the Canopus-1 well in the Goldwyer Formation of the Canning Basin for training and testing the tree-based ensemble models. It provided comparable results and evaluated the efficiency of these intelligent models' performance during the TOC content prediction process.
- The contribution of each feature on the trained models has been investigated to justify that even features with lower or non-statistically significant correlation with the target can be of use in tree-based ensemble techniques and improve their predictive power.

- A comprehensive comparative analysis has been established between four tree-based ensemble learning techniques to evaluate their accuracy and performance in TOC estimation from well logs.

This paper is organized as follows. Section 2 describes the geological background and stratigraphic setting of the Goldwyer Formation in the Canning Basin, Western Australia. Section 3 introduces the proposed data-driven tree-based ensemble learning techniques and their applications. Section 4 describes the methodology of proposed techniques. Numerical simulation results are given in Section 5, including both performance comparison studies and robustness tests, which are used to verify the effectiveness of the proposed method. Conclusions are drawn in the last section.

Equations.

$$TOC(\text{vol.}\%) = \frac{(\rho_B - \rho)}{1.378} \tag{1}$$

where ρ_B represents the bulk formation density (g/cm^3), and ρ denotes the organic matter free rock density (g/cm^3).

$$TOC(\text{wt.}\%) = \frac{[(100\rho_0) - (\rho - 0.9922\rho_{mi} - 0.039)]}{[R\rho(\rho_0 - 1.135\rho_{mi} - 0.675)]} \tag{2}$$

where ρ_0 represents the density of the organic matter (g/cm^3), R is the weight percentage ratio of the organic matter to organic carbon, and ρ_{mi} is the average density of the grain and pore fluid (g/cm^3).

$$\Delta \log R = \log_{10}\left(\frac{R}{R_{baseline}}\right) + 0.02 \times (\Delta t - \Delta t_{baseline}) \tag{3}$$

$$TOC = \Delta \log R \times 10^{(2.297 - 0.1688 \times LOM)} \tag{4}$$

where $\Delta \log R$ represents the separation in the resistivity and sonic transit time logs, R is the target formation resistivity (Ω.m), $R_{baseline}$ is the base formation resistivity corresponding to an organic lean shale (Ω.m), Δt represents the sonic transient time (μs/ft), $\Delta t_{baseline}$ is the base sonic transit time corresponding to an organic lean shale (μs/ft), and LOM denotes the level of maturity.

$$\Delta \log R = \log_{10}\left(\frac{R}{R_{baseline}}\right) + \frac{1}{\ln 10}\frac{m}{(\Delta t - \Delta t_m)} \times (\Delta t - \Delta t_{baseline}) \tag{5}$$

$$\Delta \log R = \log_{10}\left(\frac{R}{R_{baseline}}\right) + \frac{1}{\ln 10}\frac{m}{(\rho_m - \rho)} \times (\rho - \rho_{baseline}) \tag{6}$$

where m represents the cementation exponent, Δt_m and ρ_m denote the matrix sonic transit time and density in (μs/ft) and (g/cm^3), respectively. $\rho_{baseline}$ represents the baseline density corresponding to $R_{baseline}$ (g/cm^3).

Nomenclature

AI	Artificial Intelligence
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
EL	Ensemble Learning
ELM	Extreme Learning Machine
ET	Extra Trees
GA	Genetic Algorithm
GB	Gradient Boosting
GBRT	Gradient Boosting Regressor Trees

MAE	Maximum Absolute Error
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Square Error
R^2	Coefficient of Determination
RBF	Radial Basis Function
RF	Random Forest
SVM	Support Vector Machines
TOC	Total Organic Carbon
XGB	eXtremely Gradient Boosting
XGBRT	eXtremely Gradient Boosting Regressor Trees

2. **Geology of the Study Area.** The Canning Basin located in the north-western Australia is the largest sedimentary basin and comprises an area greater than 595,000 km^2 [37]. The basin is bounded in the north by the Precambrian Kimberley Block and in the south by Pilbara and Musgrave Blocks and structurally is an intra-cratonic depression developed in Early Paleozoic between the Pilbara and Kimberley Blocks (Figure 1) [36,39]. The Canning Basin underwent five major tectonic events. First event was subsidence due to an extension in the Early Ordovician [39] followed by a compression and erosion event in the Early Devonian. Third event was another extension and subsidence event in the Late Devonian. In the Middle and Late Carboniferous to Permian a sequence of compression and subsidence events took place followed by final erosional and transpressional uplift events in the Early Jurassic [37]. The thickness of the sedimentary pile in the Canning Basin is highly varied due to the movement of fault blocks. In the deepest troughs the basin hosts up to 15 km of sediments and in the structural highs the sediment thickness is as low as 1 km.

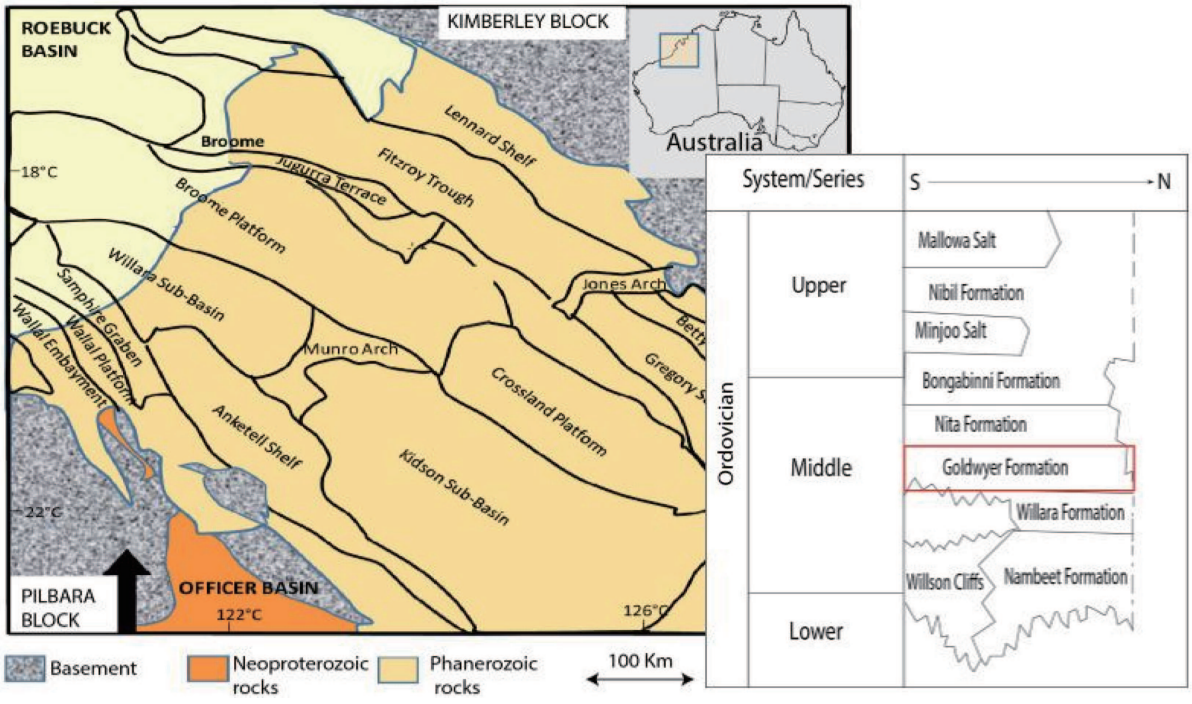


FIGURE 1. Map of the Canning Basin showing distribution of sub-basins and relevant stratigraphic column showing the lower and upper contacts of the Goldwyer Formation [modified after 39]

Sedimentary succession of the Canning Basin ranges in age from Ordovician to Cretaceous. Carbonates and fine-grained clastics of the Goldwyer Formation were deposited in the Middle Ordovician in a shallow epeiric sea and subtidal conditions as the sedimentary deposition slowed down [41]. The average thickness of the Goldwyer Formation is around 400 m in the Canning Basin while the thickest section, 739 m, is recorded in a Lennard Shelf Sub-basin well (Blackstone 1). The lithology of the Formation comprises mudstone and carbonate alternating with each other across the Basin. However, it tends to be mudstone dominated in the deeper parts of the Basin. Different subdivisions were made to the Goldwyer Formation, where in the Broome Platform multiple geological reports of some wells noted the presence of a lower shale, a middle limestone, and an upper shale unit. The Goldwyer Formation is believed to have excellent source rock potential [38] and is considered oil and gas prone with kerogen types II and III [42]. According to US EIA (Energy Information Administration) report of 2013 Goldwyer Formation has an estimated potential of 225 TCF of recoverable shale gas, the highest shale gas potential in Australia.

3. Data Set and Comparative Analysis.

3.1. **Data set.** Two sets of data were used to carry out this study i.e., laboratory measured TOC data (205 data points) and wireline log data, from the Goldwyer Formation of Canning Basin. The data was made open by Western Australia Department of Mines and Petroleum (WADMP). The main purpose of using both the data was to generate a relationship between the corresponding laboratory measured TOC value with that estimated from well logs. Depth matching was carried out by adjusting the depths to the most probable density log response with the measured TOC value within 1-3 m window, the standard range of error for cuttings depth (Guzmán [54]).

Table 1. Data set description from well Canopus-1

Feature	Target	Description	Unit	Minimum	Maximum	Mean	Std
X1	–	SP	[mV]	26.3125	72.14063	51.12851	11.31858
X2	–	GR	[api]	38.46875	198.375	113.2831	43.04463
X3	–	DT	[μs/ft]	53.1875	105.75	75.77485	14.21376
X4	–	LLD	[m.ohm]	1.679736	43.54575	9.466722	8.512348
X5	–	LLS	[m.ohm]	1.479652	45.43213	9.949067	9.166248
X6	–	RHOB	[g/cc]	1.614258	2.742188	2.485123	0.286654
X7	–	NPHI	[%]	3.857422	52.19727	20.69622	12.13644
–	Y1	TOC	[%]	0.1	1.5	0.441659	0.24671

3.2. **Relationship between well logs (input) and measured TOC (desired output).** This procedure has a significant role in the construction of the tree-based ensemble models. To classify the logs with strong relationship with each geochemical property simple regression plots are used. [43] stated that, the inputs having a stronger relationship with the output provide a more accurate prediction. For proper identification and elimination of noisy and potentially misleading data, cross-plots of well logs (input) and geochemical property (desired outputs) are used.

205 lab measured TOC data points and well log data are analyzed in this study. Before analyzing the data, depth matching was carried out. On the ground, Gamma-Ray (GR) validated the TOC values of core samples. Gamma-ray curve was compared with the measured TOC curve until it coincided.

Gamma-Ray (GR), Density (RHOB), Resistivity (LLD, LLS), Neutron Porosity (NPHI), Spontaneous (SP) and Sonic (DT) logs are the organic-sensitive wireline logs as shown in previous studies. Generally, an abnormality is observed in the logs with an increase in the organic matter content. Due to this, for predicting the TOC content multiple log parameters are required. For selecting the sensitive input, a Coefficient of Determination (R^2) was implemented between the well logs and lab measured TOC values. Following Equation (7) was used for calculating R^2:

$$R^2 = \frac{\sum_{i=1}^{n}(Y_{i,m} - Y_{i,e})^2}{\sum_{i=1}^{n}\left(Y_{i,m} - \bar{Y}_{i,m}\right)^2} \tag{7}$$

$\bar{Y}_{i,m}$, $Y_{i,e}$, $Y_{i,m}$, n represent average laboratory measured values, well logging parameters, laboratory measured values and number of samples respectively.

Table 2 shows the correlation coefficient matrix of Canopus-1 well obtained by calculation. It can be seen that the correlation coefficient between the gamma-ray curve and the TOC content was high (0.4278). Moreover, there is a decent coefficient of correlation between the TOC values and spontaneous potential, sonic, density, neutron porosity, and resistivity log curves. A good correlation of SP, GR, DT, LLD, LLS, RHOB and NPHI was found with the laboratory measured TOC data as shown in Figure 2. In summary, there is no one-to-one correspondence function relationship between the aforementioned well logs and the measured TOC as noticed from the analysis for the cross-plots (Figure 2) and coefficient of correlation (Table 2) but some well logs were significantly sensitive in different forms with measured TOC values.

Table 2. Correlation matrix of well Canopus-1

Parameters	SP	GR	DT	LLD	LLS	RHOB	NPHI	TOC
SP	1							
GR	−0.02976	1						
DT	−0.44986	0.739597	1					
LLD	0.174363	−0.64247	−0.74108	1				
LLS	0.165095	−0.63518	−0.73446	0.994352	1			
RHOB	0.735955	−0.32795	−0.69521	0.523373	0.512119	1		
NPHI	−0.57092	0.660468	0.943845	−0.71509	−0.70423	−0.77388	1	
TOC	0.191815	0.427844	0.049148	0.108673	0.122422	0.148619	−0.0123	1

3.3. **Tree based ensemble learning algorithms.**

3.3.1. *Base estimator (regression tree).* This method is not among the ensemble models but is tested because this is the base estimator for all four following tree-based ensemble techniques. A brief description of this method is presented below.

Through a series of hierarchical rule, a decision tree approximates a function as illustrated in Figure 3 which is also a supervised learning algorithm. From the example, preset points (Z_1-Z_4) sequence threshold the input variable (X_1 and X_2). Then the input variable divides the function domain into a partition set where each assigned the function value subsequently into an approximation. As example, if $X_2 > Z_2$ and $X_1 > Z_4$ then the tree will produce R_5 (the output shown in Figure 3).

In all ensemble techniques followed in this paper, this simple structure is used where multiple trees are being combined in different variety of ways. For knowing more details about this base estimator such as how the tree structure and threshold are determined, [44] provided good summary.

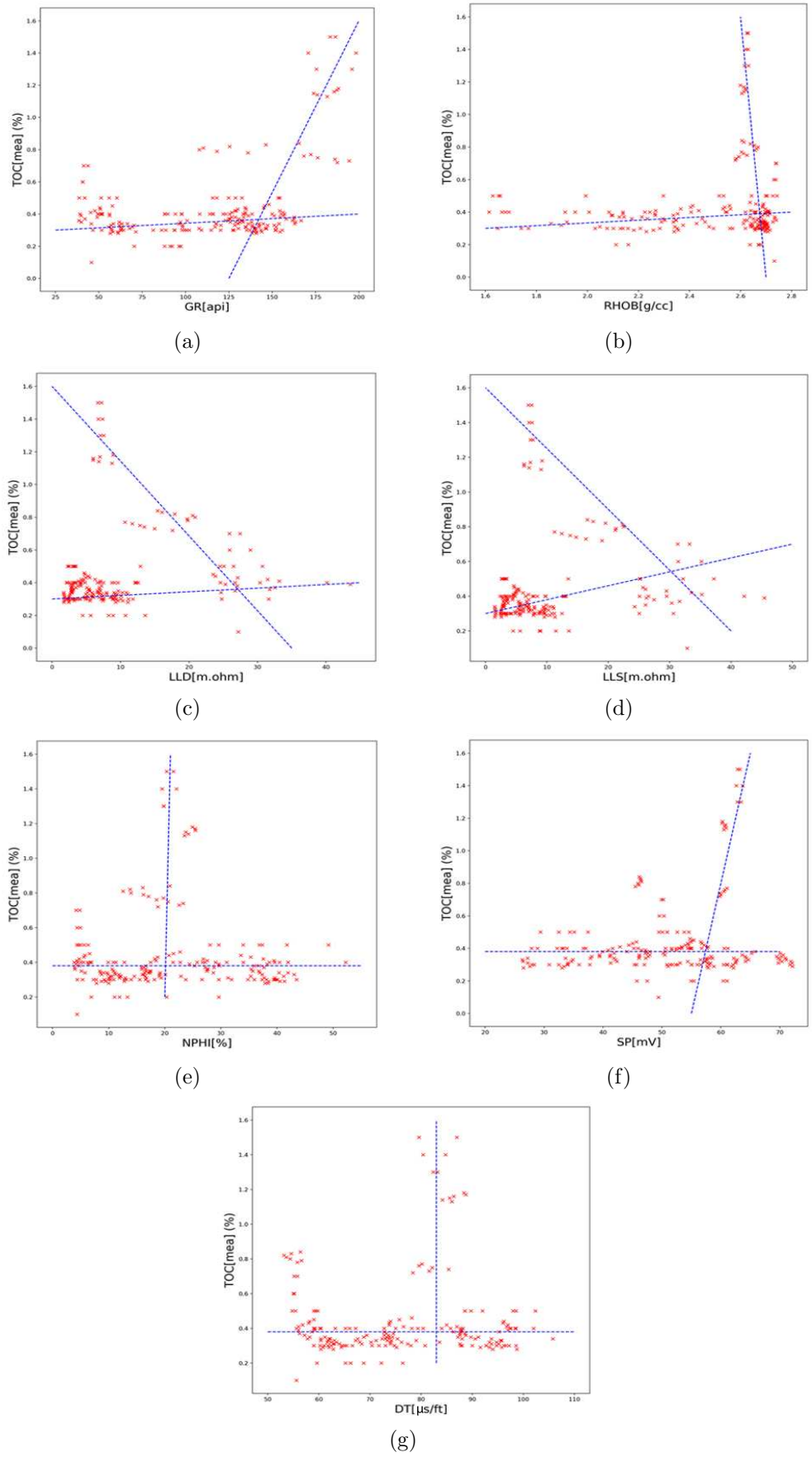


FIGURE 2. Plot showing correlation between SP, NPHI, RHOB, LLD, LLS, GR, DT logs and measured TOC

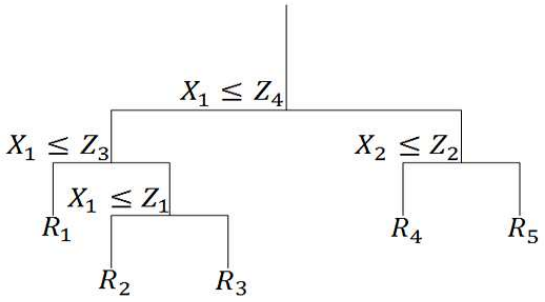


FIGURE 3. Base estimators (regression tree)

3.3.2. *Random Forest (RF).* Random Forest (RF) is an ensemble model which was proposed by Breiman [45]. RF has several instances of individual DT with the predictor variables randomly in each instance and used Breiman's "bagging" idea to ensemble a set of controlled decision tree variance. Firstly, RF starts with a single DT model then using a bootstrap strategy RF resampled the data which makes RF increase sequentially. After reaching the minimum number of nodes, the growth of tree models stops which helps to avoid overfitting. An excessive number of tree models are associated with RF. Random forest model has been accounted for to be effective and performing astoundingly well [46]; however, overfitting possibilities have also been reported [47,48]. For knowing more information about RF, Breiman paper [45] is a good read.

3.3.3. *Extremely randomized Trees (ET).* Extremely randomized trees (Extra trees) algorithm is first proposed by Geurts et al. [49] which shares RF several characteristics and relatively a recent approach by taking the randomness a step further in the tree splits. For training each base estimator, ET use a random feature subset similar to RF. The best among K randomly generated splits is picked by ET instead of choosing the most discriminative split in each node. The difference between ET and RF is that, whole training data set is used by ET for training each regression tree which is opposite to RF bootstrap sample. The split points explicit randomization in ET is expected to reduce other methods with randomization weaker schemes. The reduction in the model's bias motivates the utilization of full training data rather than a sample among them.

3.3.4. *Gradient Boosted Decision Trees (GBDT).* Gradient Boosted Decision Tree (GBDT) is first proposed by [44] which is widely used for regression and classification problems. Again, regression trees or decision stumps are used in GBDT as weak classifier. Error observed in each node is measured by the weak learners of GBDT and using test function GBDT split the node. Like the RF model, a set of weak learners is also combined in GBDT but GBDT trees are fit on formal trees residual for reducing the biases of GBDT where the variance is reduced by RF model. Because of this reason GBDT model cannot be trained in parallel where the RF model can easily be trained. Thus, the GBDT model is superior to the RF model in terms of computational costs and over-fitting. For getting more information about GBDT, [44] is a good read. In below the algorithm of GBDT is given:

(i) Initialize model: $F_0(x) = E[y]$
(ii) For $m = 1$ to M:
  (a) "pseudo-residuals" computation
  (b) Using the base regression tree fit pseudo-residuals
    i.e., set h_m to minimize $L(y, h_m(x))$
  (c) Find $\gamma_m = \arg\min_\gamma \{L(y_i, F_{m-1}(x_i) + \gamma_m h_m(x_i))\}$

(d) Update $F_m = F_{m-1}(x) + \gamma_m h_m(x)$
(iii) Model finalization is $F_m(x)$

3.3.5. *eXtremely Gradient Boosting Regressor Trees (XGBRT).* Chen and Guestrin [50] proposed eXtreme Gradient Boosting regression tree (XGBRT) in 2016 which is a machine learning scalable system for tree boosting. 17 solutions among 29 winning solutions in the machine learning competition Kaggle in 2015 were used by XGBRT.

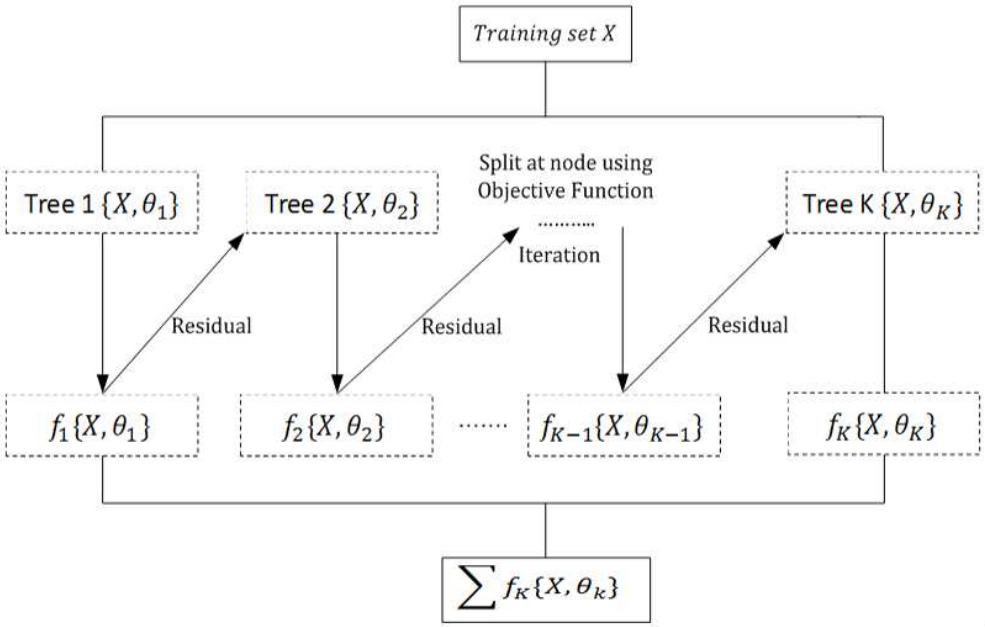


FIGURE 4. Flowchart of eXtreme Gradient Boosting (XGB)

XGBRT is from gradient boosting where weak base learning models are being combined in an iterative fashion for turning into a stronger learner [44]. From (Figure 4) each gradient boosting iteration, for correcting the previous prediction the residual will be used for being optimized the specified loss function. Again, for establishing the objecting function in XGBRT for the model performance measurements, regularization is added to the loss function which is for improvements as given in Equation (8):

$$J(\Theta) = L(\Theta) + \Omega(\Theta) \tag{8}$$

Θ is the parameter trained from given data and the training loss function is denoted as L. The regularization term such as $L1$ or $L2$ norm is Ω. Against the overfitting, the simpler models tend to have better performance. The model output $\widehat{Z}_i$ is averaged or voted by F of k trees since the base model is decision tree.

$$\widehat{Z}_i = \sum_{i=1}^{k} f_k(x_i), \quad f_k \in F \tag{9}$$

The $\widehat{Z_i^{(t)}}$ can be given as

$$\widehat{Z_i^{(t)}} = \sum_{k=1}^{t} f_k(x_i) = \widehat{Z_i^{(t)} + f_t}(x_i) \tag{10}$$

Regularization term $\Omega(f_k)$ for a decision tree is defined as

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{11}$$

where T, γ, λ and w are the number of leaves in a decision tree, the complexity of each leaf, a parameter to scale the penalty and the vector of scores on leaves respectively. Then, in XGBRT instead of first order in general gradient boosting, second order Taylos expansion is taken to the loss function. The objective function can be finally derived where the loss function is assumed as Mean Square Error (MSE) as

$$J^{(t)} \approx \sum_{i=1}^{n} \left[g_i w_{q(x_i)} + \frac{1}{2}\left(h_i w_{q(x_i)}^2\right) \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{12}$$

Here g_i is the first derivative and h_i is the second derivative of MSE loss function and $q(.)$ is a function of a leaf of assigned data point. In Equation (12), each data sample loss value summation determines the loss function because only one leaf is corresponded to each sample and by each leaf node of loss values summation also expresses the loss function. So,

$$J^{(t)} \approx \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \tag{13}$$

According to Equation (13), G_j and H_j can be defined as

$$G_j = \sum_{i \in I_j} g_i, \quad H_j = \sum_{i \in I_j} h_i \tag{14}$$

all data sample on j leaf nodes are represented by I_j. So, the problem for finding the quadratic minimum function can transform the objective functions optimization, which means, based on the objective function the change of model performance can be evaluated after a certain node split. This change will be adopted if the performance of decision tree model gets improved or the split will be stationary.

4. **Methodology.**

4.1. **Parameter tuning.** Parameter tuning or model selection in the context of machine learning is the identifying process for designing the parameters which optimize the performance of learning algorithms on a set of data. Using a grid search strategy, the important parameters of a wide range were evaluated for all four tree-based ensemble model discussed earlier. For increasing the selection of parameter's optimal probability, an exhaustive approach has been selected over randomized search.

Three categories can be sub-divided in the model parameters for the algorithm of boosting or boosting algorithm such as: (i) affecting parameters in boosting algorithm which can be called as a boosting parameters such as learning rate, (ii) each learner associated parameters which are called tree-specific parameters such as each trees maximum depth, and (iii) miscellaneous parameters such as the minimized cost function.

Three main features are affecting in RF and ET's performance. First is the number of maximum features that are allowing for trying in each learner. The second feature is the minimum sample number required for forming a leaf. And the third and the last parameters are the tree numbers which comprise the ensemble.

We randomly split the data into 70% for the training set and 30% for the testing set for the parameter tuning first step. Then we performed a grid search for being evaluated and 10-fold cross-validation for also being evaluated. Among 10-fold, we utilized 9-fold for

training the model and the rest was utilized for testing purpose. This process is repeated for 10 time until each fold gets validated and the scores get averaged.

4.2. **Model evaluation.** The performance and accuracy of four tree-based ensemble model studied here for the TOC content were evaluated. Again, by using three statistical indicators, i.e., Mean Absolute Error (MAE, Equation (15)), Mean Square Error (MSE, Equation (16)), Coefficient of Determination (R^2, Equation (17)) these four tree-based ensemble models were compared. These statistical indicators' mathematical equation has been described below:

$$\mathrm{MAE} = \frac{1}{N \times p} \sum_{i=1}^{p} \sum_{j=1}^{N} |T_{ij} - L_{ij}| \tag{15}$$

$$\mathrm{MSE} = \frac{1}{N \times p} \sum_{i=1}^{p} \sum_{j=1}^{N} (T_{ij} - L_{ij})^2 \tag{16}$$

$$\mathrm{R}^2 = \frac{\sum_{i=1}^{n} (Y_{i,m} - Y_{i,e})^2}{\sum_{i=1}^{n} \left(Y_{i,m} - \bar{Y}_{i,m}\right)^2} \tag{17}$$

where p, N represent the number of data set patterns and the number of output units. Again, T_{ij} are the target values and L_{ij} is the output value. Furthermore, $\bar{Y}_{i,m}$, $Y_{i,e}$, $Y_{i,m}$, and n represent average laboratory measured values, well logging parameters, laboratory measured values and number of samples respectively. The model will perform better if the value of MSE and MAE is low. Conversely, higher value of coefficient of determination (R^2) means its value is closer to 1 which makes the regression line fit the data well and better model performance.

The data of raw meteorological were normalized for the requirements of tree-based ensemble models (machine learning algorithm) which range between 0 and 1 by Equation (18):

$$Z^* = \frac{Z - Z_{\min}}{Z_{\max} - Z_{\min}} \tag{18}$$

where Z and Z^* represent raw data and normalized data and $Z_{\max}$ represents the maximum value and $Z_{\min}$ represents the minimum values of the original data respectively.

5. **Results & Discussion.** The main research findings have been presented in this section where in Section 4.1 it has been discussed by each technique's optimal hyper-parameters through 10-fold cross-validation. In Section 4.2 we discussed the studied model's performance of prediction and compared among them. Again, also analyze the algorithmic importance features.

5.1. **Model selection.** The performance of the tree-based ensemble models depends largely on the number of estimators, number of max depths, max number of features, max tree depth, learning rate, minimum sample leaf, validation friction, type of booster and others. These parameters reflect the range or distribution of the training data, which has a greater impact on the prediction effect of the model. There is no theoretical way to determine tree-based ensemble model parameters, and one of the commonly used methods for setting the parameters in the tree-based ensemble techniques is 'grid search', described in the following section.

The grid optimization algorithm is a large-scale point set search method. The determination of the search range needs to be set by the model builder. In order to determine the optimal value of the parameters, grid search is carried out in the tree-based ensemble

TABLE 3. Parameter search space

Model	Parameter	Range
RF	Number of estimators	200
	Max depth	4
	Min samples splits	[4, 5]
	Max tree depth	[2, 3, 4, 5]
ET	Number of estimators	100
	Max number of features	["all"]
	Min samples leaf	[1, 2, 3, 4, 5]
	Max depth	4
GB	Number of estimators	35
	Learning rate	0.1
	Validation fraction	0.1
	Max depth	2
XGB	Number of estimators	43
	Booster	'gbtree'
	Learning rate	0.1
	Max depth	3

techniques. Table 3 depicts the range of parameters for the four tree-based ensemble techniques.

From Table 3 we can see that in the case of RF, for TOC, the optimal number of estimators was 200. Again, for forming the best model a leaf node required 1 sample as a minimum, maximum tree depth of 4 and for looking to the best split all possible features were considered (max number of features = "all"). In ET, regarding the number of estimators, 100 was the optimal values for TOC. Again, ET also uses all possible feature to decide on the best split for TOC. Further, for a leaf node it requires 1 sample as the minimum and 4 for the maximum tree depth. GB required 2 max depth for TOC prediction. Again, for achieving the optimal estimators regarding TOC content (target), GB used 35 for TOC prediction. Further, for both learning rate and validation fraction GB required 0.1. Finally, for XGB, the combination of parameters is giving the most accurate results after cross-validation consists of 34 boosting stages (number of estimators), a learning rate of 0.1, and a maximum tree depth of 3 for TOC prediction. Again, the booster for XGB was 'gbtree'. Notice that among all four tree-based ensemble models, the cross-validated optimal models for RF for TOC prediction consist of a significantly higher number of estimators.

5.2. **Model performance & discussion of results.** Seven well logs namely natural Gamma-Ray (GR), Density (RHOB), Spontaneous Potential (SP), Deep Resistivity (LLD), Shallow Resistivity (LLS), Neutron Porosity (NPHI) and Density (DT) were used as an input variable and the TOC content was considered as an output variable for constructing four tree-based ensemble model which are Random Forest (RF), Extra Trees (ET), Gradient Boosting trees (GB), eXtreme Gradient Boosting trees (XGB) respectively. Core analysis data were used for all sample data as it is mentioned before. These data were stratified into two parts (training and testing data). For the training part, 70% of the total sample data were used while 30% were used for the testing part. Then, all of

the data were normalized in the range $[0, 1]$ for optimizing the model performance before modeling.

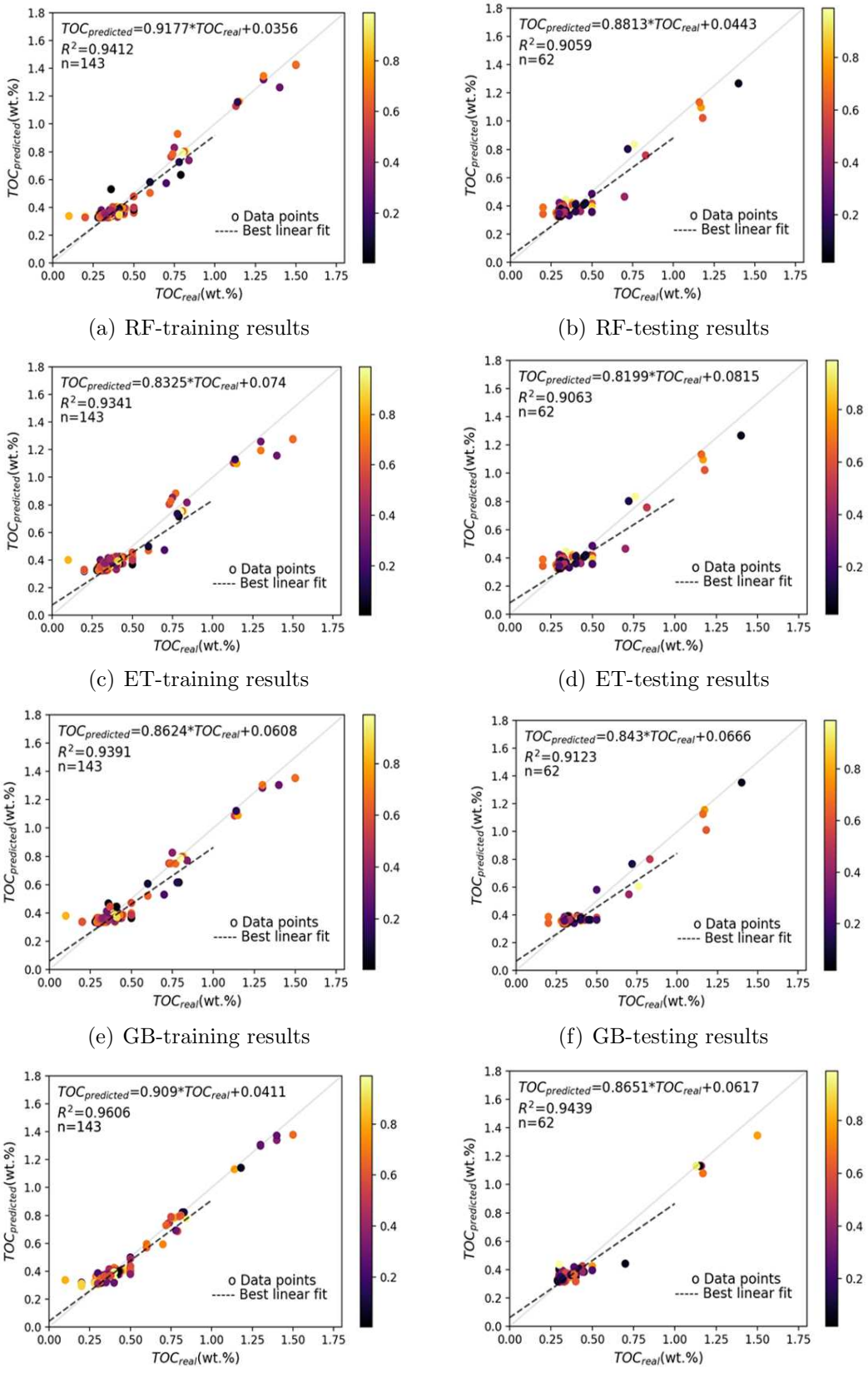
5.2.1. *Prediction results.* This section presents the results of testing the predictive accuracy of the trained models on previously unseen data. As mentioned earlier, the data set was randomly split into 10 folds, 9 of which were used for training and 1 for testing the algorithm. For generalization, the experiment was performed 100 times and the prediction errors were averaged. The predictions of the four algorithms were compared and the results are summarized in Figures 5, 6, 7, and 8. A first look at the bar plots from Figure 6 clearly shows that XGB improves the prediction of TOC, outperforming all other algorithms discussed in the literature. For instance, the XGB model outperformed other models by having the lowest MAE of 0.0347 for training, 0.0447 for testing and 0.0397 for all dataset and MSE of 0.0025 for training, 0.0039 for testing and 0.0032 for all dataset and highest R^2 of 0.9606 for training, 0.9439 for testing and 0.9523 for all dataset. Because of the superiority of XGB, better prediction performance can be gained through this model than the rest tree-based ensemble learning techniques studied in this paper.

An interesting discussion point is how different implementations of the same algorithm (i.e., RF) can lead to different accuracy levels, highlighting the need for a thorough parameter tuning when implementing machine learning algorithms. Because, once an ensemble technique has been trained, having carefully fine-tuned any hyper-parameters, it is extremely fast to obtain energy predictions for a given set of design inputs, and understand the correlations between these parameters and energy consumption. For this reason, an improvement of up to 70% can be realized using the optimized ensemble techniques compared with default ensemble techniques (without implementing any parameter tuning).

5.2.2. *Discussion.* Four tree-based ensemble models were chosen for evaluating and comparing the performance and employed to estimate TOC from well log data. Same input and output datasets were employed for reasonable comparison among the models in training and testing phase in TOC estimation.

Among 205 laboratory-measured TOC data, 143 data points were used for the training phase and 62 data points were used for testing phase for the four tree-based ensemble models. Again, in training and testing phases, the critical indicator for evaluating the four tree-based ensembles models' predictive performance between core/measure and predicted data was the correlation coefficient (R^2) as it is shown in Figure 5 where seven well logs were used as an input. Cross plots between measured TOC and RF, ET, GB and XGB model predicting TOC results are shown in Figures 5(a), 5(c), 5(e), 5(g) respectively through training samples. Among those derived TOC values with respect to real values, the correlation coefficient (R^2) of XGB is higher than other three tree-based ensemble models which are 0.9606 in training phase. Further, Figures 5(b), 5(d), 5(f) and 5(h) show the cross plots between measured TOC results and four tree-based ensemble models for the testing phase where correlation coefficient (R^2) between XGB-derived TOC value and real values were highest than others. The correlation coefficient (R^2) of XGB, GB, ET and RF were 0.9439, 0.9123, 0.9063, 0.9059 respectively which demonstrate the distinguished performance of the XGB over the other tree-based ensemble models employed in this study. Furthermore, there are some questions regarding the model with respect to their performance which relates to whether the performance of a predictive model is good or not.

In addition, the authors can evaluate the performance of these studied techniques by using statistical indicators such as correlation coefficient (R^2), Mean Absolute Error (MAE) and Mean Square Error (MSE) which are relative to the real value.



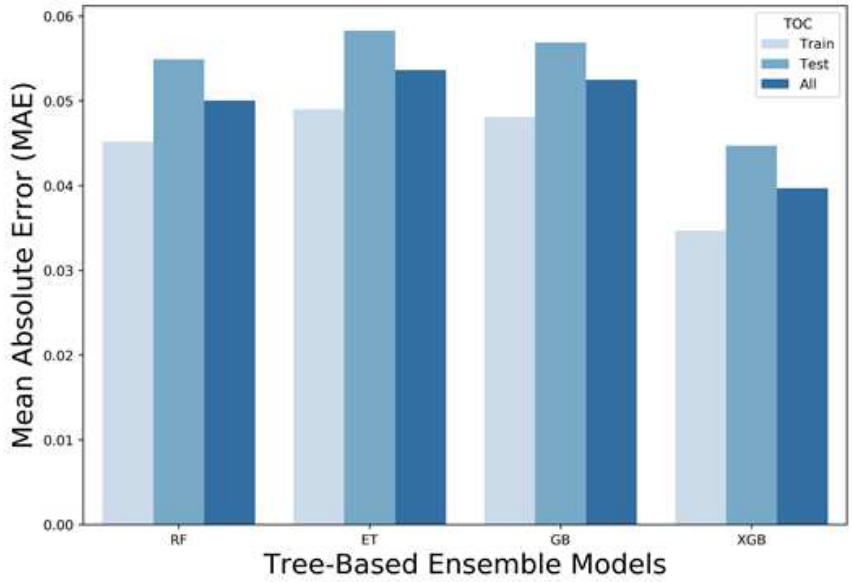
(a) RF-training results (b) RF-testing results

(c) ET-training results (d) ET-testing results

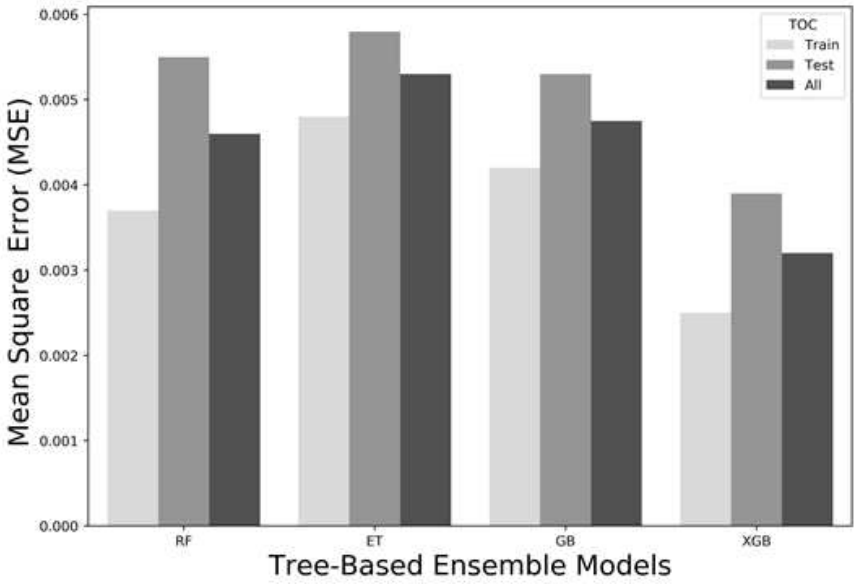
(e) GB-training results (f) GB-testing results

(g) XGB-training results (h) XGB-testing results

FIGURE 5. (color online) Cross-plot of measured and estimated TOC content using (a) RF, (c) ET, (e) GB and (g) XGB for training data sets and (b) RF, (d) ET, (f) GB and (h) XGB for testing data sets



(a) MAE of ensemble models



(b) MSE of ensemble models

FIGURE 6. Histograms comparing MAE, MSE of RF, ET, GB, and XGB models to predict TOC in training, testing and all data

In Figure 6, graphical presentation of the calculated MAE and MSE of training, testing samples and all dataset with four tree-based ensemble models (RF, ET, GB and XGB) is shown for having more intuitive comparison among the result and for better assessment. From the graphical results illustrated in Figure 6, it can be seen that among four tree-based ensemble models the XGB has lower MAE and MSE for the training phase, testing phase and all datasets.

Thus, through the comparison between RF, ET, GB and XGB, we can see that the XGB model was superior.

Table 4 has listed the results from Figures 5 and 6 of MAE, MSE and R^2 for training phase, testing phase and all dataset for four tree-based ensemble models. Among these

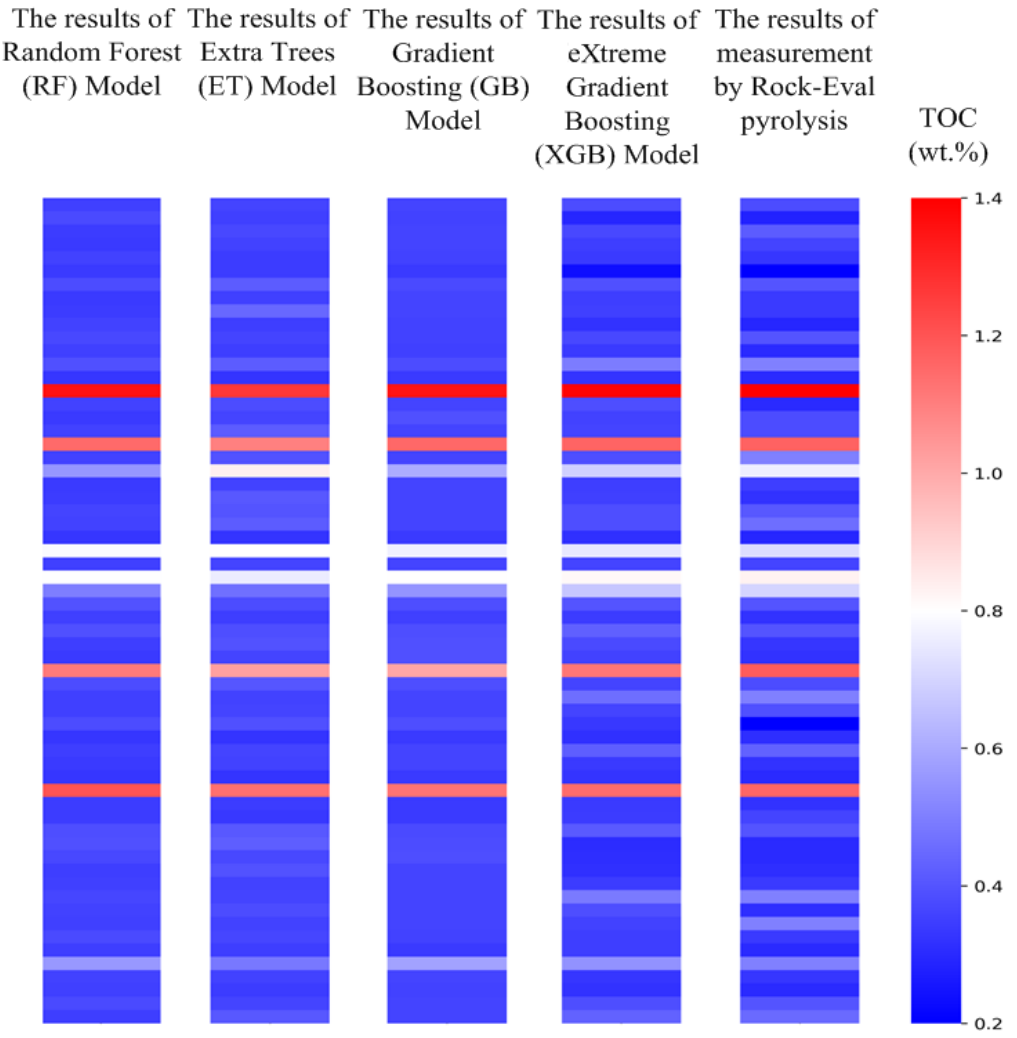


FIGURE 7. (color online) Image plots showing graphical predicted TOC content in testing phase versus measured TOC content

TABLE 4. Performance calculation for RF, ET, GB and XGB tree-based ensemble models in training phase, testing phase and all dataset

Model	Training			Testing			All		
	MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE	R^2
RF	0.0452	0.0037	0.9412	0.0549	0.0055	0.9059	0.05005	0.0046	0.9235
ET	0.049	0.0048	0.9341	0.0583	0.0058	0.9063	0.05365	0.0053	0.9202
GB	0.0481	0.0042	0.9391	0.0569	0.0053	0.9123	0.0525	0.00475	0.9257
XGB	0.0347	0.0025	0.9606	0.0447	0.0039	0.9439	0.0397	0.0032	0.9523

four tree-based ensemble models, the XGB model outperformed other models by having the lowest MAE of 0.0347 for training, 0.0447 for testing and 0.0397 for all dataset and MSE of 0.0025 for training, 0.0039 for testing and 0.0032 for all dataset and highest R^2 of 0.9606 for training, 0.9439 for testing and 0.9523 for all dataset. Because of the superiority of XGB, better prediction performance can be gained through this model than the rest tree-based ensemble learning techniques studied in this paper.

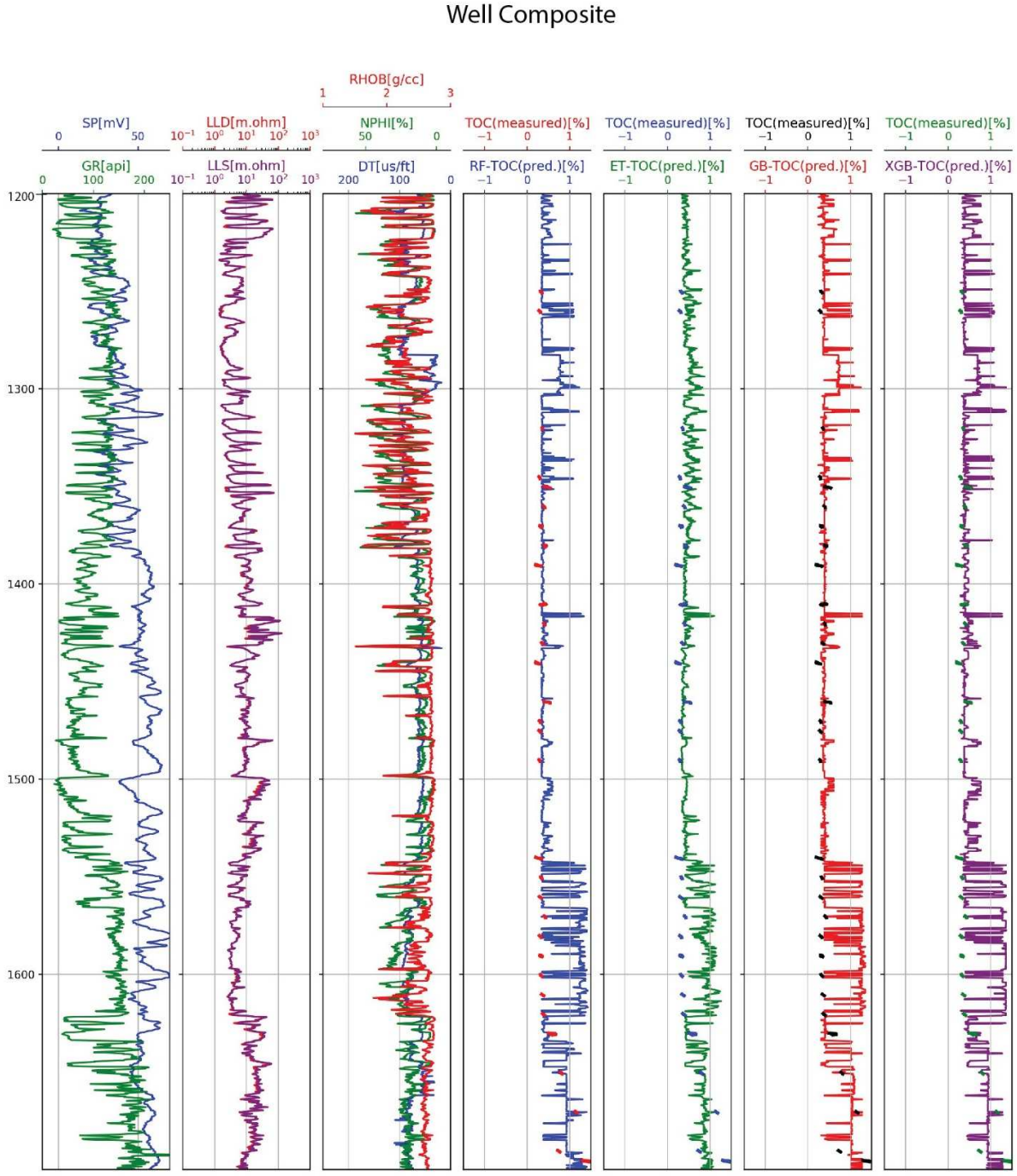


FIGURE 8. (color online) Wireline log curve with four tree-based ensemble techniques derived TOC content with measured TOC content

In this paper, a graphical comparison between predicted TOC results and the core measurements TOC through Rock-Eval pyrolysis is shown through tree-based ensemble models in both Figures 7 and 8. From Figure 7 it shows that the predictive performance of each tree-based ensemble model applied in the study is good enough. Further, among tree-based ensemble models, XGB is more successful in predicting TOC. Again, better consistency is observed between the TOC obtained from this model and actual tested TOC. Due to this reason, a better application prospect is held by XGB which should be

shown. From the different model, the predicted TOC values are compared with measurement for visualizing the prediction quality. As shown in Figure 8, the left three curves are well logs which acted as input for the intelligent models and the right four curves are model-derived TOC curved denoted as RF, ET, GB, XGB respectively. Between the right four curves illustrated in Figure 8, the continuous line is representing the predicted TOC content by the four tree-based ensemble techniques respectively and the dots are representing the actual/measured TOC content. Among the four model-derived curves, it is clear that the most consistent value with measured TOC results of core samples is XGB.

The trained models had been investigated for the contribution of each feature. In Figure 9, the relative importance of each feature is shown in the heatmap to each tested algorithm with the target variable. Starting from the relationship between feature and target, in the same degree all features are correlated with TOC. We noticed that the GR and DT parameters are strongly, positively, respectively correlated with TOC. Again, LLD and LLS parameters (features) seem to have reciprocal, similar relationship with the target. Further, the relationship between the compactness ration and the target feature is proportional which means the higher the compactness ratio is, the higher the target feature will be.

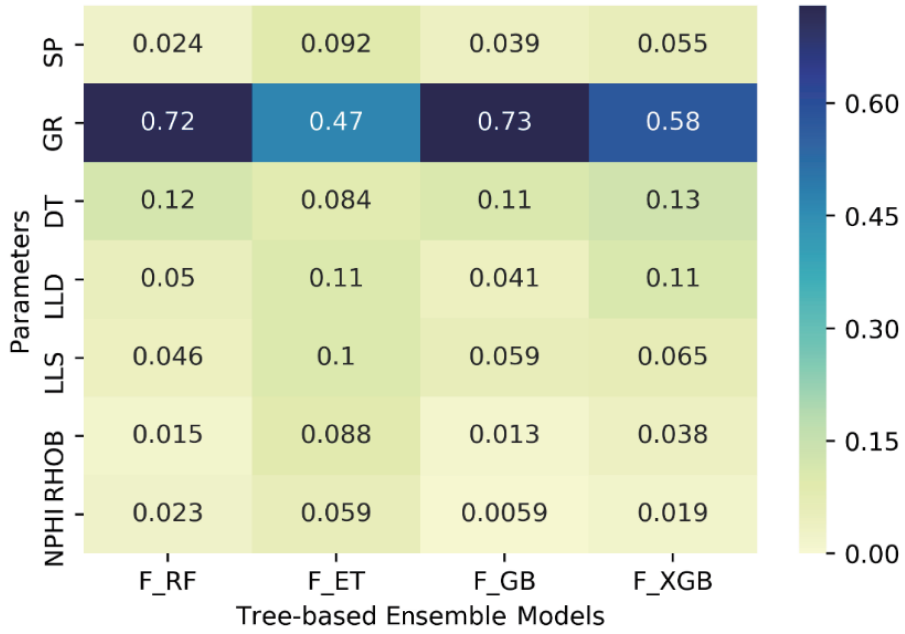


FIGURE 9. (color online) Tree-based ensemble model's feature importance for the TOC content prediction

Most of the information from the feature is drawn by ET models where the features are correlated with the target feature. On the other hand, the most important features of XGB are the less correlated ones (i.e., SP, RHOB, NPHI). From the XGB nature, the findings can be justified despite being counterintuitive. From Section 3.3.5, based on the errors of the previous model, XGB fits their sequential trees. Each constituent learner is focusing on the residual information rather than the target feature itself. For this reason, feature with non-statistically significant or lower correlation with target features can be used in XGB and improve the predictive power of XGB. Further, three features (i.e., SP, NPHI, RHOB) have a lower correlation with the rest of the predictors. Since those three features have a major contribution to the extreme boosting model, we can conclude

that the model is better suited for avoiding overfitting and handling the multicollinearity problem (correlation among feature). For the machine learning researcher, the correlation among the feature is a common issue [51], which often results in adversely affecting the model's predictive power and misleading conclusion.

Python 3.7.2 with libraries of scientific computing such as pandas 0.19.2 and NumPY 1.12.1 providing pre-processing methods and the efficient data structure has been used in this work for training the tree-based ensemble learning models. Again, Xgboost 0.6 and Scikitlearn have been imported for supporting the four tree-based ensemble models such as RF, ET, GB and XGB [52].

6. **Conclusions.** The first comprehensive performance evaluation of tree-based ensemble learning techniques has been presented in this work for the prediction of TOC content from well logs in a potential shale reservoir. Four popular, robust and efficient tree-based ensemble techniques, i.e., Random Forest (RF), Extra Trees (extremely randomized trees) (ET), Gradient Boosting (GB) and eXtremely Gradient Boosting (XGB) were studied that is a novel approach in the estimation of TOC. 205 laboratory measured TOC data points with seven well logs namely GR, DT, RHOB, SP, NPHI, LLD, and LLS were used for training and testing the tree-based ensemble models and evaluating the efficiency of these intelligent models' performance during the TOC content prediction process.

The results confirm the ability of tree-based ensemble learning models to accurately model and predict TOC content estimation from well logs in the shale reservoir as all the four tree-based ensemble techniques have achieved the exemplary level of accuracy. Among these models, the XGB predicted TOC values more accurately than the other three which means that the predicted TOC by XGB matches well with measured TOC content. In comparison to RF, ET, and GB, the XGB model has lower MSE and MAE and higher R^2 showing that XGB is most suitable to predict TOC in intervals having no core data. Furthermore, it is also noted that three logs, i.e., SP, NPHI, RHOB have a lower correlation with the rest of the predictors. Since these three logs have a major contribution to the extreme boosting model, it is concluded that the XGB model is better suited for avoiding overfitting and handling the multicollinearity problem (correlation among features). Moreover, these robust tree-based ensemble models can protect overfitting and have achieved better prediction results while dealing with the multidimensional data.

By proving the benefits and adequacy of utilizing the above-mentioned techniques has made the contribution of this paper significant for TOC prediction from well logs. These tested approaches can be reliably used in prediction of organic richness. The studied techniques can be generalized for additional elements with desirable accuracy in other research areas such as estimation of petrophysical and geochemical properties of hydrocarbon reservoirs.

Acknowledgment. The authors would like to thank Petroleum Research Fund (cost centre 0153AB-A33), Shale Gas Research Group (SGRG), and the project leader Assoc. Prof. Dr. Eswaran Padmanabhan for supporting this work.

REFERENCES

[1] J. W. Schmoker, Determination of organic content of Appalachian Devonian shales from formation-density logs: Geologic notes, *AAPG Bull.*, vol.63, no.9, pp.1504-1509, 1979.

[2] Q. R. Passey, S. Creaney, J. B. Kulla, F. J. Moretti and J. D. Stroud, A practical model for organic richness from porosity and resistivity logs, *AAPG Bull.*, vol.74, no.12, pp.1777-1794, 1990.

[3] P. Wang, Z. Chen, X. Pang, K. Hu, M. Sun and X. Chen, Revised models for determining TOC in shale play: Example from Devonian Duvernay shale, Western Canada sedimentary basin, *Mar. Pet. Geol.*, vol.70, pp.304-319, 2016.

[4] P. Zhao, H. Ma, V. Rasouli, W. Liu, J. Cai and Z. Huang, An improved model for estimating the TOC in shale formations, *Mar. Pet. Geol.*, vol.83, pp.174-183, 2017.
[5] Z. Huang and M. A. Williamson, Artificial neural network modelling as an aid to source rock characterization, *Mar. Pet. Geol.*, vol.13, no.2, pp.277-290, 1996.
[6] E. Sfidari, A. Kadkhodaie-Ilkhchi and S. Najjari, Comparison of intelligent and statistical clustering approaches to predicting total organic carbon using intelligent systems, *J. Pet. Sci. Eng.*, vol.86, pp.190-205, 2012.
[7] X. Shi, J. Wang, G. Liu, L. Yang, X. Ge and S. Jiang, 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.*, vol.33, pp.687-702, 2016.
[8] M. Tan, X. Song, X. Yang and Q. Wu, 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.*, vol.26, pp.792-802, 2015.
[9] S.-A. Ouadfeul and L. Aliouane, Shale gas reservoirs characterization using neural network, *Energy Procedia*, vol.59, pp.16-21, 2014.
[10] L. Zhu et al., Prediction of total organic carbon content in shale reservoir based on a new integrated hybrid neural network and conventional well logging curves, *J. Geophys. Eng.*, vol.15, no.3, pp.1050-1061, 2018.
[11] V. Bolandi, A. Kadkhodaie and R. Farzi, 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. Pet. Sci. Eng.*, vol.151, pp.224-234, 2017.
[12] A. A. A. Mahmoud, S. Elkatatny, M. Mahmoud, M. Abouelresh, A. Abdulraheem and A. Ali, Determination of the total organic carbon (TOC) based on conventional well logs using artificial neural network, *Int. J. Coal Geol.*, vol.179, pp.72-80, 2017.
[13] L. M. Johnson, R. Rezaee, A. Kadkhodaie, G. Smith and H. Yu, Geochemical property modelling of a potential shale reservoir in the Canning Basin (Western Australia), using Artificial Neural Networks and geostatistical tools, *Comput. Geosci.*, vol.120, pp.73-81, 2018.
[14] P. Wang and S. Peng, A new scheme to improve the performance of artificial intelligence techniques for estimating total organic carbon from well logs, *Energies*, vol.11, no.4, p.747, 2018.
[15] A. Charsky and S. Herron, Accurate, direct Total Organic Carbon (TOC) log from a new advanced geochemical spectroscopy tool: Comparison with conventional approaches for TOC estimation, *AAPG Annual Convention and Exhibition*, Pittsburg, PA, pp.19-22, 2013.
[16] A. E. Ben-Nakhi and M. A. Mahmoud, Cooling load prediction for buildings using general regression neural networks, *Energy Convers. Manag.*, vol.45, nos.13-14, pp.2127-2141, 2004.
[17] C. Turhan, T. Kazanasmaz, I. E. Uygun, K. E. Ekmen and G. G. Akkurt, Comparative study of a building energy performance software (KEP-IYTE-ESS) and ANN-based building heat load estimation, *Energy Build.*, vol.85, pp.115-125, 2014.
[18] B. Dong, C. Cao and S. E. Lee, Applying support vector machines to predict building energy consumption in tropical region, *Energy Build.*, vol.37, no.5, pp.545-553, 2005.
[19] Q. Li, Q. Meng, J. Cai, H. Yoshino and A. Mochida, Applying support vector machine to predict hourly cooling load in the building, *Appl. Energy*, vol.86, no.10, pp.2249-2256, 2009.
[20] S. U. Jan, Y.-D. Lee, J. Shin and I. Koo, Sensor fault classification based on support vector machine and statistical time-domain features, *IEEE Access*, vol.5, pp.8682-8690, 2017.
[21] M. S. A. Rahaman and P. Vasant, Artificial intelligence approach for predicting TOC from well logs in shale reservoirs: A review, *Deep Learn. Tech. Optim. Strateg. Big Data Anal.*, pp.46-77, 2020.
[22] A. Dutta and P. Dasgupta, Ensemble learning with weak classifiers for fast and reliable unknown terrain classification using mobile robots, *IEEE Trans. Syst. Man Cybern. Syst.*, vol.47, no.11, pp.2933-2944, 2016.
[23] C. Zhang and Y. Ma, *Ensemble Machine Learning: Methods and Applications*, Springer, 2012.
[24] L. Brillante et al., Investigating the use of gradient boosting machine, random forest and their ensemble to predict skin flavonoid content from berry physical-mechanical characteristics in wine grapes, *Comput. Electron. Agric.*, vol.117, pp.186-193, 2015.
[25] E. M. Burger and S. J. Moura, Gated ensemble learning method for demand-side electricity load forecasting, *Energy Build.*, vol.109, pp.23-34, 2015.
[26] J. Elith, J. R. Leathwick and T. Hastie, A working guide to boosted regression trees, *J. Anim. Ecol.*, vol.77, no.4, pp.802-813, 2008.
[27] A. Ellahyani, M. El Ansari and I. El Jaafari, Traffic sign detection and recognition based on random forests, *Appl. Soft Comput.*, vol.46, pp.805-815, 2016.

[28] J. Friedman, T. Hastie and R. Tibshirani, *The Elements of Statistical Learning*, Springer, New York, 2001.
[29] H. Sun et al., Assessing the potential of random forest method for estimating solar radiation using air pollution index, *Energy Convers. Manag.*, vol.119, pp.121-129, 2016.
[30] Y. Zhang and A. Haghani, A gradient boosting method to improve travel time prediction, *Transp. Res. Part C Emerg. Technol.*, vol.58, pp.308-324, 2015.
[31] M. S. A. Rahaman, P. M. Vasant, S. R. Jufar and J. Watada, Feature selection-based Artificial Intelligence techniques for estimating total organic carbon from well logs, *Journal of Physics: Conference Series*, vol.1529, no.4, 2020.
[32] R. Caruana and A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, *Proc. of the 23rd International Conference on Machine Learning*, pp.161-168, 2006.
[33] S. Ding, W. Song, D. Wang and H. Li, A classification method by using fuzzy neural network and ensemble learning, *International Conference on Fuzzy Information & Engineering*, pp.133-142, 2017.
[34] J. Islam et al., A modified niching crow search approach to well placement optimization, *Energies*, vol.14, no.4, p.857, 2021.
[35] A. Tsanas and A. Xifara, Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools, *Energy Build.*, vol.49, pp.560-567, 2012.
[36] F. Anifowose, S. Adeniye and A. Abdulraheem, Recent advances in the application of computational intelligence techniques in oil and gas reservoir characterisation: A comparative study, *J. Exp. Theor. Artif. Intell.*, vol.26, no.4, pp.551-570, 2014.
[37] *Summary of Petroleum Prospectivity: Canning Basin*, Department of Mines and Petroleum, Petroleum Division, Western Australia, 2014.
[38] S. J. Cadman, L. Pain, V. Vuckovic and S. R. Le Poidevin, Canning Basin, *WA Bur. Resour. Sci. Aust. Pet. Accumul. Rep.*, vol.9, p.81, 1993.
[39] S. A. Brown, I. M. Boserio, K. S. Jackson and K. W. Spence, *The Geological Evolution of the Canning Basin-Implications for Petroleum Exploration*, http://archives.datapages.com/data/petroleum-exploration-society-of-australia/conferences/009/009001/pdfs/85.htm, 1984.
[40] L. M. Johnson, G. Smith, R. Rezaee and A. Kadkhodaie, A 3D model of the unconventional play in the Goldwyer Formation: An integrated shale rock characterisation over the Broome Platform, Canning Basin, *SPE/AAPG/SEG Asia Pacific Unconventional Resources Technology*, Brisbane, Australia, 2019.
[41] G. Australie, Bureau of mineral resources, geophysics, in *Geological evolution of the Canning Basin, Western Australia*, D. J. Forman and D. W. Wales (eds.), Bureau of Mineral Resources, Geology and Geophysics, 1981.
[42] M. Alshakhs, *Shale Play Assessment of the Goldwyer Formation in the Canning Basin Using Property Modelling*, Ph.D. Thesis, Curtin University, 2017.
[43] A. Kadkhodaie-Ilkhchi, H. Rahimpour-Bonab and M. Rezaee, 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.*, vol.35, no.3, pp.459-474, 2009.
[44] J. H. Friedman, Greedy function approximation: A gradient boosting machine, *Ann. Stat.*, pp.1189-1232, 2001.
[45] L. Breiman, Random forests, *Mach. Learn.*, vol.45, no.1, pp.5-32, 2001.
[46] R. Caruana, N. Karampatziakis and A. Yessenalina, An empirical evaluation of supervised learning in high dimensions, *Proc. of the 25th International Conference on Machine Learning*, pp.96-103, 2008.
[47] M. K. Campbell, G. Piaggio, D. R. Elbourne and D. G. Altman, Consort 2010 statement: Extension to cluster randomised trials, *BMJ*, vol.345, DOI: 10.1136/bmj.e5661, 2012.
[48] L. Toloşi and T. Lengauer, Classification with correlated features: Unreliability of feature ranking and solutions, *Bioinformatics*, vol.27, no.14, pp.1986-1994, 2011.
[49] P. Geurts, D. Ernst and L. Wehenkel, Extremely randomized trees, *Mach. Learn.*, vol.63, no.1, pp.3-42, 2006.
[50] T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, *Proc. of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp.785-794, 2016.
[51] C. F. Dormann et al., Collinearity: A review of methods to deal with it and a simulation study evaluating their performance, *Ecography*, vol.36, no.1, pp.27-46, 2013.
[52] F. Pedregosa et al., Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.*, vol.12, no.10, pp.2825-2830, 2011.

[53] Z. Heidari, C. Torres-Verdin and W. E. Preeg, Quantitative method for estimating total organic carbon and porosity, and for diagnosing mineral constituents from well logs in shale-gas formations, *SPWLA the 52nd Annual Logging Symposium*, Colorado Springs, Colorado, 2011.

[54] J. Guzmán, *Formation Characterization in a Different Perspective: Drill Cuttings Analysis Revisited*, https://www.semanticscholar.org/paper/Formation-Characterization-in-a-Different-Drill-Guzm%C3%A1n/4c868b105dd967f712ca76c89c795d6c1dae14ef, 2003.