Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Feature Selection-Based Artificial Intelligence Techniques for Estimating Total Organic Carbon from Well Logs

To cite this article: Md Shokor A Rahaman *et al* 2020 *J. Phys.: Conf. Ser.* **1529** 042084

Feature Selection-Based Artificial Intelligence Techniques for Estimating Total Organic Carbon from Well Logs

Md Shokor A Rahaman[1], Dr. Pandian M. Vasant[1], Dr Shiferaw R. Jufar[2], Junzo Watada[3]
[1]Department of Fundamental and Applied Sciences,
[2]Department of Petroleum Engineering,
[3]Department of Computer and Information Sciences,
Universiti Teknologi PETRONAS, 31750 Tronoh, Perak, Malaysia

shokor103072@gmail.com

Abstract: For shale oil and gas exploration total organic carbon (TOC) is the significant factors where TOC estimation considered as a challenges for geological engineers because direct laboratory coring analysis is costly and time consuming. Passey method and Artificial Intelligence (AI) technique have used on well logs extensively to determine TOC content. But, the prediction of Passey method is low and AI technique such as ANN, Support Vector Machine (SVM) trapped in local optima, overfitting and heavy computation work or error if the technique isn't reasonable. In this paper, for the first time in TOC prediction we propose three feature selection-based algorithm which are Decision Tree (DT), Gradient Boosting Regressor (GBR) and Random Forest (RF) respectively. This feature selection-based algorithm select the best attributes among the input parameters for TOC content prediction. Then those best attributes works as an input for AI models for training and testing the AI models which illustrates that making a correlation between well logs and TOC content for the prediction. Specifically, 2069 core shale sample and well logging sample data of the Texas University Lands of Kansas Geologic Society were divided into 1448 training sample and 621 validating sample to evaluate the proposed AI models. This proposed AI model and feature selection-based algorithm jointly allows TOC content to be accurately and continuously predicted based on conventional well logs.

Keywords: Feature selection, Random forest, Gradient Boosting Regressor, Decision Tree, Artificial Neural Network, Geophysical logs.

1. Introduction

In oil and gas exploration, the source rock which are evaluated correctly plays an essential role. Again, in the source rock the correct evaluation of organic matter is also an important part. Petrophysical properties such as thermally mature, rock mineralogy, total organic carbon (TOC) content et al. are the definition of productive shale quality. Some critical reservoir properties can judge the most productive shale oil and gas reservoir qualitatively. At least 2% TOC, Ro range within maturity windows makes a commercially potential shale oil and gas reservoir. Further, it also need to have more than 2% porosity, less than 40%

saturation and permeability (100 nanodarcy) which makes it a good flow capability and gas storage. Furthermore, carbonate in mineralogy or more than 40% quartz, a certain degree of natural fracture and low differential stress is also need to have for a good commercial shale oil and gas reservoir [1]. But TOC is the important indicator and fundamental among all the factors for explaining the reservoir potential. Thus, total organic carbon (TOC) content that reflect the source rocks physical characteristics is identified as an important and basic index [2-3]. The abundance of organic matter can be represented by this parameters. There are lots of direct method for obtaining the TOC content. Among them direct laboratory core analysis for obtaining the TOC content is costly and time consuming. Again this direct laboratory method can obtain limited amount of TOC data where the current demands for source rock evaluation is difficult. For this reason, the accurate and continuous prediction of TOC content is required for the unconventional natural oil and gas rapid development and exploration. On the other hand, well logging has the characteristics of the continuity of the data and high longitudinal resolution. So, from geochemical parameters to well logging information of source rock, a study has been started by researchers where TOC content prediction based on the logging parameters [4-6].

For predicting the TOC data content many milestone have been achieved by researchers with the continuous improvements. The direct laboratory method from where limited TOC data can be obtained is still important for the source rock evaluation. Special logging response characterize source rock that why there is some specific relation between organic matter and geophysical logging responses. Moreover, TOC content has some specific logging parameters such as neutron, natural gamma, resistivity, density and acoustic time difference. Natural gamma well log were used to calculate the TOC content in Beers et al. [7] and Schmoker J [8] work which found to be suitable for the source rock riches in radioactive elements of calculating TOC content. By using the acoustic time difference well log Autric and Dumesnil [9] calculated TOC content where a strong correlation has seen from TOC content to the acoustic time difference which showed a better prediction. Again, by combining the neutron porosity with resistivity well logs Guo et al. [10], Hu et al. [5], Wang et al.[11], and Zhao et al. [12] calculated TOC content. However, for the different researcher areas the background value of the TOC content and the source rock maturity were different and a significant impact were found on the prediction. Again, a complicated non-linear relationship is seen between the logging data and TOC content. Approximating the real function is hard through simple linear regression and it is impossible to predict the TOC content by using the well log. Recently, Artificial intelligence (AI) is being used by many researchers. So, by the current research results AI strategies has worked for predicting the TOC content. Sfidari et al [13] used 3 layered BP-NN model to predict TOC content. In Barnett shale Ouadfeula and Aliouane [14] usen ANN to determine the TOC content using Schmokker's model. Again, three different regression algorithms and four different kernel function were used to predict TOC content in Tan et al. [15] work. Shi et al.[16] used MLP-NN and Fuzzy logic to predict the TOC content data from wireline log to replace the Schmoke model. Extreme Learning Machine (ELM) and ANN were used to predict TOC in Shi et al [6] work where ELM was quicker than ANN. Integrated hybrid neural network (IHNN) were used to predict TOC in Zhu et al. [17] work in Jaioshiba zone. In Bolandi et al. [18] work, ANN was outperformed by SVM with RBF to predict TOC content from well logs. In Duvernay and Barnett shale formation ANN were used to estimate TOC by Mahmoud et al. [8]. Johnson et al.[19] used ANN where Wang et al. [20] used CNN to predict TOC content using well logs.

In our work we used three feature selection based algorithm Random Forest (RF), Decision Tree (DT) and Gradient Boosting Regressor (GBM) to select the best attributes from the input parameter well logs. Then for building a non-linear implicit function relationship between the selected well logs and TOC content 3 AI model Gradient Boosting Regressor (GBM), Random Forest Regressor (RFR) and ANN is used to predict TOC content from the selected well logs. Here the nobility of our work is Random Forest

Journal of Physics: Conference Series **1529** (2020) 042084 doi:10.1088/1742-6596/1529/4/042084

(RF) , Gradient Boosting Regressor (GBR) and Decision Tree (DT) have never been used for feature selection-based algorithm for TOC content prediction from well logs.

2. Methodology and Theory

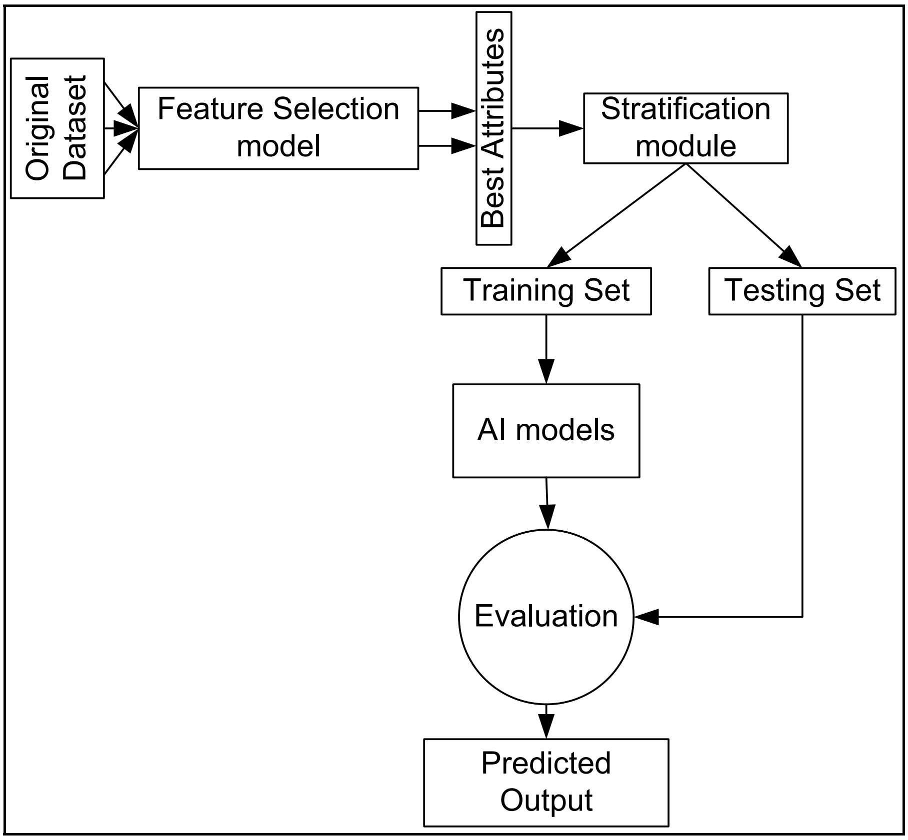

Figure 1. The conceptual framework of this study

The methodology of this study

1. Feature-Based selection applied on the original dataset to get the best attributes.
2. After getting the best attributes, stratification module had applied to divide the data into training set and test set.
3. Artificial technique applied on training set to fit the data.
4. Testing set would be applied on the fitted AI model to evaluate.
5. Predicted output of optimized AI would be achieved after evaluation and result of develop model and existing correlation will be compared.
6. Finally, best model for the prediction would be concluded.

2.1. Decision tree:

Decision tree model known as regression trees or classification trees is a type of supervised learning algorithm. For mapping observation, it uses a decision tree algorithm as a predictive technique about a problem to conclusion about the problems target value [21]. For understanding and interpreting its quite simple. They can handle both categorical and numerical data and required little data preparation. Its act as a white box model such that if situation can be observed in a model, Boolean logic gives the explanation for the condition. It performs well with the large data in a short time and its robust [22]. However, NP-complete is the learning problem of an optimal DT [21-22].

2.2. Random Forest Tree:
Random Forest (RF) is an ensemble model. RF has a several instances of individual DT with the predictor variables randomly in each instance. Firstly, RF start with a single DT model then using a bootstrap strategy RF resampled the data which makes RF increasing sequentially. After reaching the minimum number of nodes, the growing of Tree models stops which helps to avoid overfitting. Excessive number of tree models associated with RF. Random Forest model have been accounted for to be effective and performing astoundingly well [23] however overfitting possibilities have also been reported [24-25].

2.3. Gradient Boosting Algorithm:
Gradient Boosting Algorithm (GBM) is a boosting algorithm. GBM is utilized when plenty of data need to be predicted with high prediction power. Boosting is an ensemble of learning algorithms. Compare to single estimator, it combines the prediction of several base estimators to improve robustness. Further, for building strong predictor it consolidates average predictors or multiple weak.

2.4. Artificial Neural Network:
ANN is a supervised training intelligent system for solving nonlinear problems, which is developed under the category of artificial intelligence (AI) to provide a brain-like tool. An input layer, a hidden layer and an output layer consist in ANN model. Among those layers each layer is interconnected with many neurons with the specific function, such as sigmoid, purelin. A full connection is established between each node of the next layer. It can not only handle the difficult nonlinear problems in engineering research but also has a good effect on data classification, clustering and prediction.

3. Data Processing and Analysis:

3.1. Literature survey on feature selection
One of the novelties of this work is that Decision Tree (DT), Gradient Boosting Algorithm (GBA) and Random Forest (RF) have not been applied in shale gas reservoir characterization as a feature-based algorithm. However, Decision Tree (DT), Gradient Boosting Algorithm (GBA) and Random Forest (RF) have been used for decades as standalone predictive tools.

3.2. Description of data
For the design, testing and validation of our purpose AI model, 2069 data points for TOC prediction were used. These datapoints were obtained from a Texas University Lands of Kansas Geologic Society. The dataset has 75 petrophysical logs including TOC.

For fair comparison, the same datasets were used for all the feature selection algorithms as well as the proposed machine learning techniques. The datasets consist of laboratory measurements for oil and gas wells (with TOC as target) obtained from the Texas University Lands of Kansas Geologic Society. This site contains 75 predictor variables for TOC. This dataset is representative of the geological formations found in most parts of the shale gas producing world. Table 1 show the predictor variables for the TOC datasets where 9 logging parameters (VPYR and VOM= Matrix Volume Fraction, BVPYR and BVOM= Bulk Volume Fraction, DPHI and DPHI_N = DENSITY POROSITY -LIME-, RHOB and RHOB_N = BULK DENSITY and WTCLAY = Matrix Weight Fraction) were selected through feature selected-based algorithm from 74 input logging parameters. Further, Table 2 shows the logging parameters of selected logs with the TOC content data that were analysed in the core experiments.

Table 1. Input variables of ANN, GBR and RFR model.

Input	Well Logs
All logs	'CALI', 'DPHI', 'GR', 'NPHI', 'PE', 'RHOB', 'PHIX', 'C13', 'C24', 'DT', 'SPHI', 'SPHI_N', 'PE_N','PHIE_SUM', 'PAY_FLAG_1', 'GR3', 'ILD' , 'ILM' , 'SGRD' , 'SP' , 'CAL_N' , 'GR_N', 'RESMED_N', 'RESDEEP_N', 'NPHI_N', 'DPHI_N', 'VCLAY', 'VPYR', 'RHOM', 'RHOB_N' , 'DTC_N' ,'SP_N' ,'PORE_PRESS' , 'RES_TEMP' ,'NES', 'RW' , 'RMF' , 'RHO_HC', 'RHO_W', 'RHO_MF' ,'NPHI_HC', 'NPHI_W' , 'NPHI_MF' , 'OIP_SUM', 'BVH_SUM' ,'MU_HC', 'BO', 'BP', 'PHIE', 'SW' ,'SHC', 'BVH', 'BVW' ,'BVWI', 'BVWF' , 'BVOM' , 'BVCLAY' , 'BVPYR', 'VOM' , 'PC1' , 'PC2' , 'FACIES', 'WTCLAY' ,'WTPYR' ,'BVQTZ' ,'VQTZ', 'WTQTZ' , 'BVCLC', 'VCLC', 'WTCLC' , 'BVDOL', 'VDOL', 'WTDOL', 'OIP', 'PAY_FLAG_2', 'PAY_FLAG_3'
Selected Logs	DPHI, RHOB, DPHI_N, RHOB_N, BVOM, BVPYR, VOM, VPYR and WTCLAY

Table 2. Summary of selected recorded logging parameters

	DPHI	RHOB	DPHI_N	RHOB_N	BVOM	BVPYR	VOM	VPYR	WTCLAY	TOC
MEAN	0.109	2.524001	0.108767	2.528536	0.052845	0.010569	0.055805	0.011161	0.239481	0.029827
STD	0.03452	0.059038	0.034515	0.053238	0.017894	0.003579	0.01931	0.003862	0.067982	0.01052
MIN	-0.002	2.181	-0.002	2.365	0	0	0	0	0.043235	0
MAX	0.309	2.713	0.309	2.713	0.10497	0.020993	0.11258	0.022516	0.4882	0.061492

4. Evaluation criteria

In addition, this study conducted a correlation analysis viz. correlation coefficient (R-Square), root mean-squared error (RMSE), normalized root mean square error (NRMSE) and mean absolute error (MAE) for the well log and TOC content, to calculate the correlation between each well log and the TOC content. The calculation equation is as follows:

$$R - Square = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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

$$NRMSE = \frac{\sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}}}{y_{max} - y_{min}}$$

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

In which R-squared is the correlation coefficient, x_i and y_i are the corresponding logging observation values of the ith coring sample point, respectively and $\bar{x}$ and $\bar{y}$ are the average value of logging parameters. Further, y_{max} and y_{min} are the maximum and minimum value of y.

4.1. Input selection

In this paper, the logging parameters were selected through feature-based algorithm from 74 input logging parameters. From the Figure 2, the 5 most scored attributes in Decision Tree (DT) as a feature selection-based algorithm are DPHI, DPHI_N, RHOB_N, VOM, and WTCLAY. Again, from Figure 2, 7 most scored attributes in Gradient Boosting Regressor (GBR) as a feature selection-based algorithm are RHOB, RHOB_N, BVOM, BVPYR, VOM, VPYR and WTCLAY where Random Forest (RF) has 9 most scored attributes from fig 2 by its feature selection-based algorithm are respectively DPHI, RHOB, DPHI_N, RHOB_N, BVOM, BVPYR, VOM, VPYR and WTCLAY. Furthermore, TOC log was selected as an output layer for the prediction. Here we have used three feature algorithms to select the best attributes for the input logging parameters. Among those algorithms from Figure 1 we observed that Random Forest Tree algorithm has given the maximum number of best attributes of 9 from the logging parameters while the other two feature-based algorithm Decision Tree algorithm and Gradient boosting algorithm has selected 5 and 7 attributes respectively.

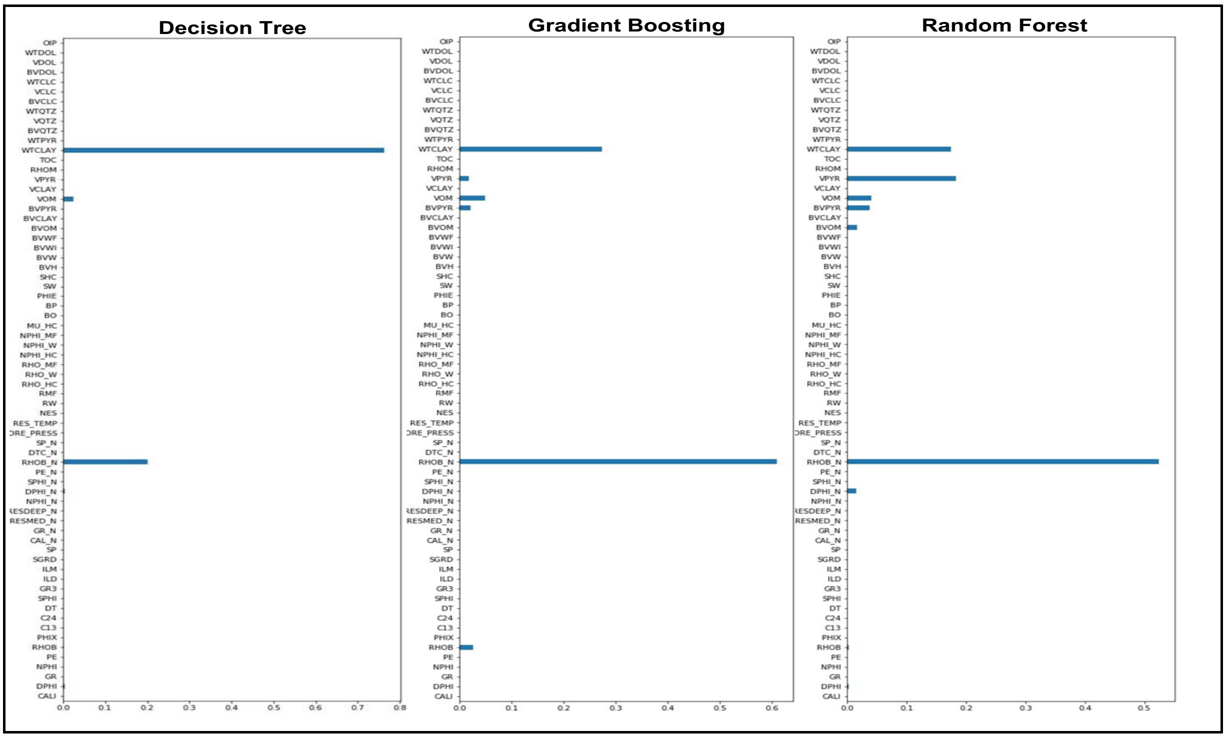

Figure 2. Feature Selection based algorithms performance

4.2. Data Standardization:
Each attribute of the logging data and geochemical index was normalized using the Gaussian initialization method.

$$x_{attr} = \frac{1}{N}\sum_{i=1}^{N} x_{i,attr}$$

$$s_{attr} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_{i,attr} - x_{attr})^2}$$

$$y_{i,attr} = \frac{x_{i,attr} - x_{attr}}{s_{attr}}$$

Where attr represents DPHI, RHOB, DPHI_N, RHOB_N, BVOM, BVPYR, VOM, VPYR and WTCLAY, $x_{i,attr}$ is the average value of attr for all samples, s_{attr} is the variance of attr for all samples, and $y_{i,attr}$ represents the normalized value of sample, N is the number of samples.

4.3. Model establishment
After applying the feature-based algorithm, as well as the optimization of test data, for correlation, this investigation set up three kinds of TOC content forecast models dependent on the sample data. These three models were the Artificial Neural Network (ANN), Gradient Boosting Regressor (GBR) and Random Forrest Regressor (RFR) respectively.

4.4. Model performance

4.4.1. Artificial Neural network Performance: In this study, the input variable (DPHI, RHOB, DPHI_N, RHOB_N, BVOM, BVPYR, VOM, VPYR and WTCLAY) were selected through Feature selection based algorithm where Random Forest (RF) act as a feature selection based algorithm and RF have selected 9 best attributes among 74 input parameters. For a superior evaluation, all input and output data were standardized. 1448 examples were gathered to train the ANN model while 621 examples were taken as testing samples to survey the generalization capacity of the model. The outcomes from Figure 3 demonstrate that the ANN model based on chosen log has R^2 of 0.996424 and NRMSE of 0.010363.Table 3 demonstrated the performance of three AI models.

Table 3. Model Evaluation

	MODEL	**NRMSE**	**MAE**	**RMSE**	**R2**	**EXPLAINED VARIANCE SCORE**
0	Gradient Boosting Regressor	0.00568	0.000208	0.000349	0.998926	0.998927
1	Artificial Neural Network	0.010363	0.000502	0.000637	0.996424	0.99649
2	Random Forest Regressor	0.001789	6.76E-05	0.00011	0.999893	0.999894

4.4.2. Gradient Boosting Regressor Performance: For Gradient Boosting Regressor (GBR), 9 input variables have been chosen from 74 input variable through using feature selection-based algorithm which are DPHI, RHOB, DPHI_N, RHOB_N, BVOM, BVPYR, VOM, VPYR and WTCLAY. For a sensible examination, same input and output is received by GBR model for training and testing. The result from Figure 3 demonstrates that the R^2 of GBR model is better than ANN model of 0.99893 and NRMSE is lower than ANN of 0.00586.

4.4.3. Random Forest Regressor Performance: For Random Forest Regressor (RFR), same input and output data for training and testing has applied as ANN and GBR. The result from Figure 3 demonstrates that the R^2 value of RFR model is better than the ANN and GBR model of 0.99989 and NRMSE is also lower than the ANN and GBR model as well.

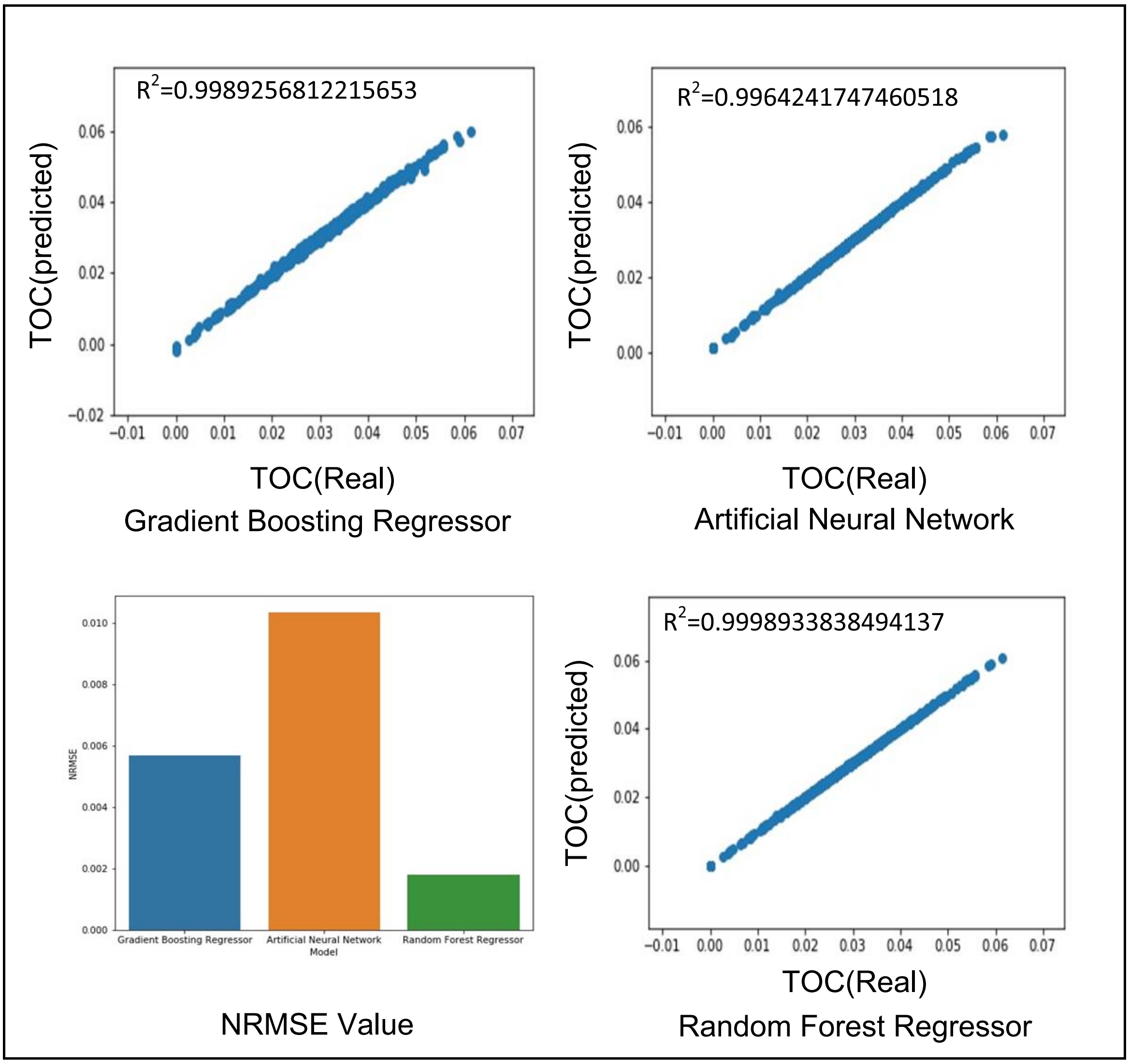

Figure 3. Cross plot of Real data and predicted data & bar plot of error

In general, from Figure 3 the RFR predicted TOC values are more accurate than the other two model which are ANN and GBR which means that the predicted TOC value by RFR match well with measured TOC. Compared to ANN and GBR, the RFR model has lower NRMSE and higher R^2 which means that RFR is most suitable to predict TOC in intervals without coring.

The main advantage of RFR as follows 1) Have very high accuracy and extremely flexible. 2) Preparation doesn't require for the input data so, there is no need to scale the data. 3) Even with the large proportion of missing data it maintains the accuracy. 4) The problem of overfitting is being overcome by combining or averaging the results of different decision trees. 5) Have laser variance than DT which means that for the large range of data items it works correctly than DT.

RFR has two major drawbacks: the first one is their complexity and compare to other algorithm the prediction process of it is time-consuming.

5. Conclusion

For the first time in TOC prediction we propose three feature selection-based algorithm which are Decision Tree (DT), Gradient Boosting Regressor (GBR) and Random Forest (RF) respectively. This feature selection-based algorithm selected the best attributes among 74 the input parameters for TOC content prediction. Here, Random Forest Tree algorithm has given the maximum number of best attributes of 9 from the 74 logging parameters while the other two feature-based algorithm Decision Tree algorithm and Gradient boosting algorithm has selected 5 and 7 attributes respectively. Then those best attributes works as an input for AI models Random Forest Regressor (RFR), Gradient Boosting Regressor (GBR) and Artificial Neural Network (ANN) for training and testing the AI models which illustrates that making a correlation between well logs and TOC content for the prediction. Among those AI models the RFR predicted TOC values are more accurate than the other two model which are ANN and GBR which means that the predicted TOC value by RFR match well with measured TOC. Compared to ANN and GBR, the RFR model has lower NRMSE and higher R^2 which means that RFR is most suitable to predict TOC in intervals without coring.

Reference

[1] C. H. Sondergeld, K. E. Newsham, J. T. Comisky, M. C. Rice, and C. S. Rai, "Petrophysical considerations in evaluating and producing shale gas resources," in *SPE Unconventional Gas Conference*, 2010.

[2] G. E. King, "Thirty years of gas shale fracturing: What have we learned?," in *SPE Annual Technical Conference and Exhibition*, 2010.

[3] 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.

[4] 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," in *SPWLA 52nd Annual Logging Symposium*, 2011.

[5] H. T. Hu *et al.*, "Models for calculating organic carbon content from logging information: Comparison and analysis," *Acta Sedimentol. Sin.*, vol. **29**, pp. 1199–1205, 2011.

[6] 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.

[7] R. F. Beers, "Radioactivity and organic content of some Paleozoic shales," *AAPG Bull.*, vol. **29**, no. 1, pp. 1–22, 1945.

[8] 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.
[9] A. Autric, “Resistivity, radioactivity and sonic transit time logs to evaluate the organic content of low permeability rocks,” *Log Anal.*, vol. **26**, no. 03, 1985.
[10] L. Guo, J. F. Chen, and Z. Y. Miao, “The study and application of a new overlay method of TOC content,” *Nat Gas Geosci*, vol. **20**, no. 6, pp. 951–956, 2009.
[11] 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.
[12] 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.
[13] 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.
[14] S.-A. Ouadfeul and L. Aliouane, “Shale gas reservoirs characterization using neural network,” *Energy Procedia*, vol. **59**, pp. 16–21, 2014.
[15] S.-A. Ouadfeul* and L. Aliouane, “Total Organic Carbon Prediction in Shale Gas Reservoirs using the Artificial intelligence with a comparative study between Fuzzy Logic and Neural Network,” in *14th International Congress of the Brazilian Geophysical Society & EXPOGEF, Rio de Janeiro, Brazil, 3-6 August 2015*, 2015, pp. 1390–1393.
[16] 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.
[17] 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.
[18] 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.
[19] 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.
[20] H. Wang, W. Wu, T. Chen, X. Dong, and G. Wang, “An improved neural network for TOC, S1 and S2 estimation based on conventional well logs,” *J. Pet. Sci. Eng.*, 2019.
[21] L. Breiman, J. Friedman, R. Olshen, and C. Stone, “Classification and regression trees. Wadsworth Int,” *Group*, vol. **37**, no. 15, pp. 237–251, 1984.
[22] Y. Yohannes and P. Webb, *Classification and regression trees, CART: a user manual for identifying indicators of vulnerability to famine and chronic food insecurity*, vol. 3. Intl Food Policy Res Inst, 1999.
[23] R. Caruana, N. Karampatziakis, and A. Yessenalina, “An empirical evaluation of supervised learning in high dimensions,” in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 96–103.
[24] M. K. Campbell, G. Piaggio, D. R. Elbourne, and D. G. Altman, “Consort 2010 statement: extension to cluster randomised trials,” *Bmj*, vol. **345**, p. e5661, 2012.
[25] L. Toloşi and T. Lengauer, “Classification with correlated features: unreliability of feature ranking and solutions,” *Bioinformatics*, vol. **27**, no. 14, pp. 1986–1994, 2011.