Global Journal of Computer Science and Technology: G Interdisciplinary
Volume 22 Issue 1 Version 1.0 Year 2022
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Machine Learning Algorithms for Predicting Reservoir Porosity using Stratigraphic-dependent Parameters

By Kachalla Aliyuda, Aliyuda Ali, Abdulwahab Muhammed Bello & Jerry Raymond

Gombe State University

Abstract- Predicting reservoir porosity, permeability and other reservoir parameters are very important but arduous task in formation evaluation, reservoir geophysics and reservoir engineering. Recent successes in machine learning and data analytics in different geoscience disciplines provides the opportunity to offer cheaper and faster techniques of predicting reservoir properties. This study used gross depositional environments, reservoir depth, diagenetic impact, permeability and stratigraphic heterogeneity from a database of 93 reservoir to predict reservoir porosity. The data for this study includes numeric and categorical descriptions of 93 reservoirs across the UK and Norwegian sector of the North Sea. Five models were trained using linear regression, support vector machine (SVM), boosted tree, bagged tree and random forest algorithms. The performance of the different models was evaluated using R-squared (R^2), root mean square error (RMSE) and mean absolute error (MAE). Model trained using random forest algorithm with R^2 score of 0.75, RMSE of 0.118 and MAE of 0.0028 outperformed other models. A comparison between predicted porosity and the actual porosity in training data and testing data show a good match, indicating the ability of the random forest model to make prediction on unseen data.

Keywords: *machine learning algorithms, reservoir porosity, sedimentology.*

GJCST-G Classification: *I.1.2*

Strictly as per the compliance and regulations of:

Machine Learning Algorithms for Predicting Reservoir Porosity using Stratigraphic-dependent Parameters

Kachalla Aliyuda [α], Aliyuda Ali [σ], Abdulwahab Muhammed Bello [ρ] & Jerry Raymond [ω]

Abstract- Predicting reservoir porosity, permeability and other reservoir parameters are very important but arduous task in formation evaluation, reservoir geophysics and reservoir engineering. Recent successes in machine learning and data analytics in different geoscience disciplines provides the opportunity to offer cheaper and faster techniques of predicting reservoir properties. This study used gross depositional environments, reservoir depth, diagenetic impact, permeability and stratigraphic heterogeneity from a database of 93 reservoir to predict reservoir porosity. The data for this study includes numeric and categorical descriptions of 93 reservoirs across the UK and Norwegian sector of the North Sea. Five models were trained using linear regression, support vector machine (SVM), boosted tree, bagged tree and random forest algorithms. The performance of the different models was evaluated using R-squared (R^2), root mean square error (RMSE) and mean absolute error (MAE). Model trained using random forest algorithm with R^2 score of 0.75, RMSE of 0.118 and MAE of 0.0028 outperformed other models. A comparison between predicted porosity and the actual porosity in training data and testing data show a good match, indicating the ability of the random forest model to make prediction on unseen data. The machine learning technique presented in this study represents a pragmatic approach to the classical log conversion problem that over the years has caused dilemmas to generations of geoscientists and petroleum engineers. The method requires no underlying mathematical models or costly assumptions of linearity among variables. Predicting porosity by using sedimentological parameters can effectively reduce the high cost of using petrophysical methods such as nuclear magnetic resonance and other logging methods.

Keywords: *machine learning algorithms, reservoir porosity, sedimentology.*

I. Introduction

Porosity, permeability, oil, water and gas saturation are commonly obtained from logging and core data, however, reservoir parameters obtained by logging or coring are limited in extent, such data are only valid a few centimetres away from the wellbore. Due to reservoir heterogeneity and the complexity of the geologic conditions, well logging data often exhibit a very strong nonlinear characteristic and the relative relation between different data is intricate (Chen et al., 2017). Different depositional facies and depositional environments ultimately controls reservoir character (Mathew et al 2008; William and Milne 1991; Larue and Legarre 2004; Jian et al 2004; Skorstad et al 2005; Skorstad et al 2008). The primary depositional fabric of the rock is modified during burial by compaction and cementation. Consequently, reservoir depth of burial is very critical in understanding the reservoir quality. (Aliyuda et al. 2021; Cade et al. 1994).

Accurate prediction of reservoir flow properties especially porosity and permeability are very vital in oil and gas recovery, production design, well placement and optimization, Co_2 sequestration, radioactive waste disposal, and management of water aquifer. Prediction of reservoir porosity and permeability is also crucial for basin-wide evaluation of fluid-migration and in mapping potential pressure seals to reduce drilling hazards.

Reservoir porosity is a function of many geological factors, these factors include depth of burial, structural complexity, sedimentary environment, lithology, and diagenetic impact. There is a general non-linear relationship between porosity and some petrophysical log properties such as density log, sonic log, and compensated neutron logs (Singh et al 2016; Zhong and Carr 2019). Several relationships which can relate porosity to wireline readings are available, common among such relationships are the sonic transit time and density logs. However, the conversion from density and transit time to equivalent porosity values is not straightforward. The common conversion formulae contain terms and factors that depend on the individual location and lithology of the well, for example, clay content, pore-fluid type, grain density and grain transit time for the conversion from density and sonic logs, that in general are unknowns and must be determined from rock sample analysis.

Geophysical well logs generally provide a good representation of the in-situ conditions in a lithological unit. However, as with most well-logging measurements, the sonic log does not provide a direct measurement of reservoir porosity, the parameter with which it has been traditionally associated with. In like manner, porosity

Author α ρ: Department of Geology, Gombe State University, Nigeria.
e-mails: aliyudakachalla@gmail.com, geomanaja@gmail.com
Author σ: Department of Computing and Data Science, Birmingham University, e-mail: Aliyuda.Ali@bcu.ac.uk
Author ω: Jerry Raymond: Department of Telecommunication Engineering, Air Force Institute of technology, Nigeria.
e-mail: j.raymond@afit.edu.ng

conversion from bulk density log requires that the grain density and fluid density be known (Vernik, 1997).

It seems obvious that no single log measurement is enough to obtain reliable values of porosity. Additional data would be required from the pore fluid and grain material, which normally are not at hand except for special studies in cored reservoir intervals.

Some conventional machine learning algorithms have been applied in predicting reservoir evaluation parameters, such as Back Propagation neural network (Leite and Vidal, 2011; Para et al, 2003; Shi et al, 2016; Wang et al, 2018), Support Vector Machine (Wang and Peng, 2018; Feng et al, 2020) and other shallow machine learning algorithms (Talkhestani, 2015; Wang and Peng, 2019; Haklidir, 2020; Mahmoud et al, 2020; He et al, 2020). Deep Learning methods specifically convolutional neural network (CNN), recurrent neural network (RNN), and stack auto encoder (SAE) were also successfully applied in predicting reservoir porosity (Zhang et al, 2021). However, most studies on predicting reservoir porosity were done using logs inputs. This study used sedimentological properties as inputs to predict porosity using a robust database of 93 reservoirs from the Norwegian continental shelf.

II. Database

The data for this study includes numeric and categorical descriptions of 93 reservoirs across the UK and Norwegian sector of the North Sea. 75 reservoirs from the Norwegian sector are from the Norwegian North Sea, Norwegian Sea, and the Barents Sea, while the remaining 18 reservoirsare from Viking graben on the UK sector. All the reservoirs were classified using the SAFARI schema into three gross depositional environments (Fluvial, Paralic/shallow marine and Deep marine). SAFARI is a Joint Industry Research Project between the University of Aberdeen and NORCE Research in Bergen, supported by a consortium of 16 companies, the Research Council of Norway and the Norwegian Petroleum Directorate. The goal of the SAFARI project is to develop a fully searchable repository of geological outcrop data from clastic sedimentary systems for reservoir modelling and exploration (www.safaridb.com). SAFARI uses a systematic hierarchical schema to classify sedimentary rocks into gross depositional environment (GDE), depositional environment (DE), sub-environment and architectural element (AE).

Further parameters which potentially influence permeability were also recorded for each of the reservoirs in the database that was used for this study. Parameters used for this study are gross depositional environments, reservoir depth, diagenetic impact, porosity, permeability, and stratigraphic heterogeneity (Table 1). Stratigraphic heterogeneity was defined on a scale of zero to eight, considering the vertical and horizontal heterogeneity of a given reservoir's depositional sub-environment following Tyler and Finley (1991), also summarized in Manzocchi et al. (2008). In this scheme, zero refers to a reservoir with no vertical and lateral heterogeneity while eight refers to a reservoir with high vertical and horizontal heterogeneity (Fig. 1).

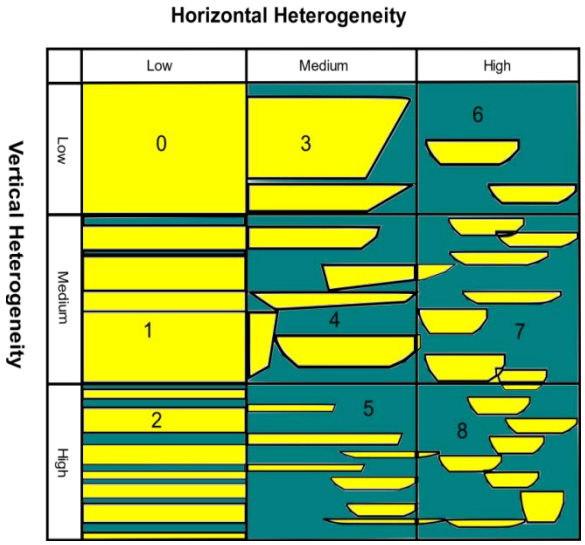

Figure 1: Depositional heterogeneity and flow unit diagram showing heterogeneity scale used for all the different Sub-environments, 0 = means low vertical and horizontal heterogeneity, highly connected sand reservoir with no clay or shale barriers. 8 = extremely heterogeneous, low net to gross reservoir sand bodies are isolated within clay or shale intervals (Modified after Tyler and Finley 1991; Aliyuda et al., 2020).

Table 1: Parameters used for the study, their range and definition.

Parameter	Description	Parameter range	Data source
Gross depositional environment	Specific environments of sediment deposition, reservoirs were classified further into Gross Depositional Environments (GDE) and Depositional-environment (DE) using the SAFARI classification Schema.	0.0 = Deep marine 0.5 Paralic/shallow marine 1 = Continental	NPD well reports, wireline logs, core images, literature
Diagenetic impact	Negative impact of reservoir sediments reconstitution and/or rearrangement resulting in a reduction of porosity and permeability only. It is classified into low, moderate or high impact. 0 = Low, 0.5 = Moderate, and 1 = High.	0 = Low 0.5 = Moderate 1 = High.	NPD well reports, literature
Stratigraphic heterogeneity	A measure of aerially extensive architecturally bounding surfaces that compartmentalize the reservoirs (after Tyler and Finley 1991). A scale of 0-8 was used with 0 = Very low heterogeneity and 8 = Extremely heterogeneous (Fig. 2.3)	0 = low vertical, low-horizontal heterogeneity 1 = low vertical, medium horizontal heterogeneity 2 = low vertical, high-horizontal heterogeneity 3 = medium vertical, low horizontal heterogeneity 4 = medium vertical, medium horizontal heterogeneity 5 = medium vertical, high horizontal heterogeneity 6 = high vertical, low horizontal heterogeneity 7 = high vertical, medium horizontal heterogeneity 8 = high vertical, high horizontal heterogeneity	NPD well reports, wireline logs, core images, literature
Porosity (%)	The average porosity of the reservoir interval	P10 porosity reported by field operators	NPD well reports, wireline logs, literature
Permeability (mD)	The average permeability of the reservoir interval	P10 permeability reported by field operators	NPD well reports, wireline logs, literature
Reservoir depth (m)	Highest point on reservoir units or interval		NPD well reports, literature

III. Methods

Five machine learning algorithms were used for the study; these are Linear Regression, Support Vector Machine (SVM), Boosted Tree, Bagged Tree and Random Forest. Boosted Tree and Bagged Tree are ensembles techniques of the Decision Tree methods.

Regressions are statistical technique that approximate the relationship between a dependent variable (the response) and one or more independent variables. Linear regression is mostly used for forecasting and finding out cause and response relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables. Linear regression models are often plagued by a significant bias (Seber 1977; Mann 1987), where the predictor variables are cross correlated with each other and with the response variables, this results into the models reporting high accuracy but do not make accurate prediction of the new data. Some alternatives to linear regression are regularised linear regression approaches such as LASSO regression, Ridge regression, Elastic Net and Non-parametric regressors, usually based on decision trees.

Decision tree regression takes multiple columns of potential predictor variables and finds a subset of predictor columns that best account for the variance of the target column values (Fig. 2). Boosted Decision Tree regression algorithms together with Bagged Decision Tree are ensembles of regression decision trees. In boosted regression, the algorithm learns by fitting the residual of the trees that preceded it, thereby improving accuracy with some small risk of less coverage. Bagged regression assumes a basic model structure as the one developed in a decision tree regression. Then, it divides the source data into several bags or groups and fits the same assumed model structure to each bag of data. Bagged regression aggregates the model estimates for each bag of data into one overall model.

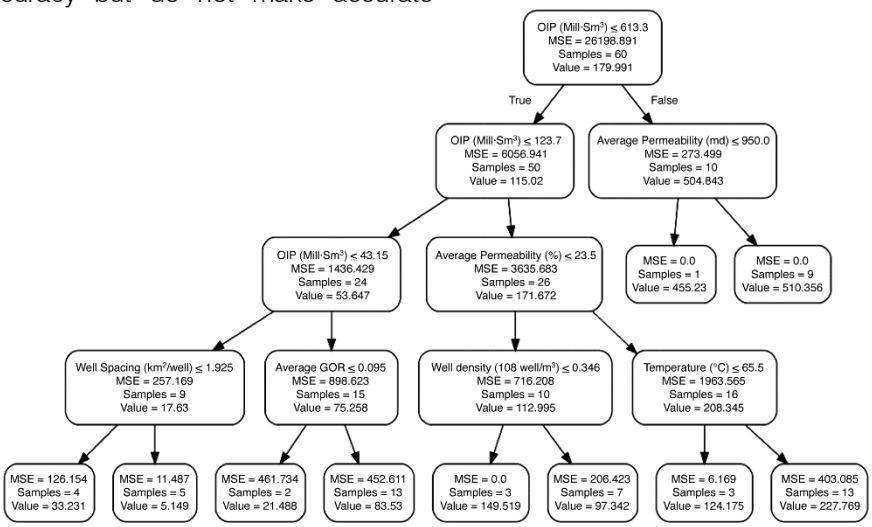

Figure 2: Decision tree regression schematic of a reservoir rate model, an example of a decision tree split at each node (Aliyuda et al., 2020).

Support vector machine (SVM) is a supervised machine learning algorithm that are commonly used to analyse data characteristic of both classification and regression problems. In SVM, each of the training data points is marked as one of two categories and then iteratively builds a region that will separate the data points in the space into two groups such that the data point in each region is well separated across the boundary with the maximum width. Support vector machine can generalize the characteristics that differentiate the training data that is provided to the algorithm. This is achieved by checking for a boundary that differentiates the two classes with the maximum margin. The boundary that separates the two classes is known as a hyperplane (Cortes and Vapnik 1995; Aliyuda and Howell, 2019; Ali et al., 2021a; Ali et al., 2021b).

Random forest is a common non-parametric regression approach which aggregates an ensemble of decision trees in order to arrive at a result. It predicts by taking the mean of the output from various trees. Increasing the number of trees increases the precision of the outcome.The decision trees are generated in parallel, and each split is made from random subsets of the dependent variables. Decision trees generated through taking random columns from the dependent variables are less prone to overfitting (Breiman, 2001). This technique allows random forests to be more robust than decision trees.

Data for this study were normalised using min-max method, other pre-processing techniques performed on the data include a split of the data into training and testing sets. These techniques prevent against over-fitting of the models. The training set is used to train the model, whereas the testing set is used to detect the accuracy of the model and output the predicted reservoir porosity.

Explained variance or R-squared (R2), square root of the mean squared error (RMSE) and mean absolute error (MAE) were used to estimate the performance and the accuracy of the trained models:

$$R^2 = \frac{N\sum xy - \sum x \sum y}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}, \quad \text{I}$$

$$RMSE = \frac{\sqrt{\sum_{i=1}^{N}(yi - pi)^2}}{N}, \quad \text{II}$$

$$MAE = \frac{\sum_{i=1}^{N}|yi - pi|}{N}, \quad \text{III}$$

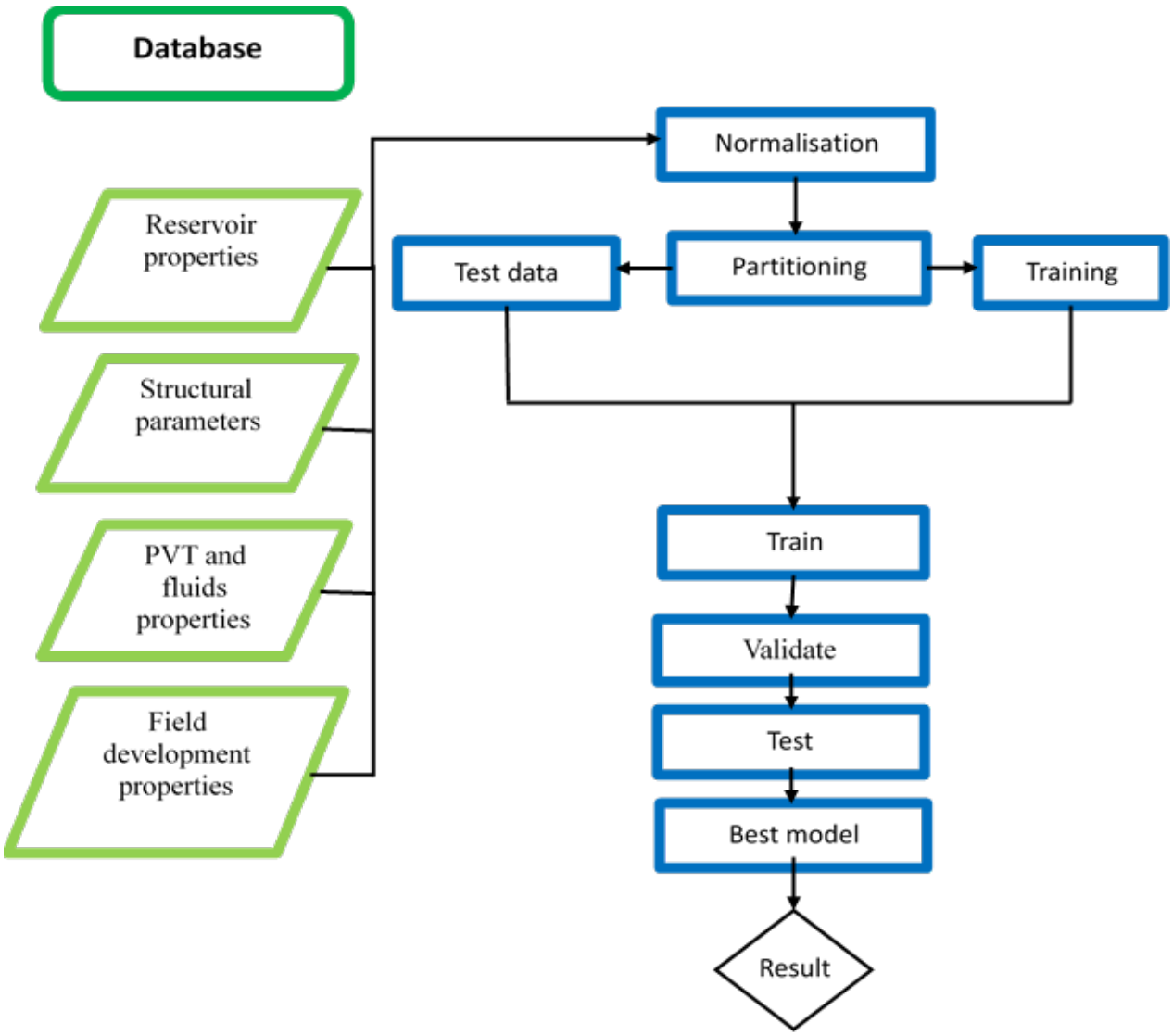

Figure 3: Workflow used to show the rundown of the procedure from building the database to training and testing of models (adopted from Aliyuda et al 2020).

IV. Results/Discussion

The distribution of some of the major predictors of the model is presented in Fig. 4, 5 and 6, these predictors are reservoir depth of burial (Fig. 4), gross depositional environments (Fig. 5) and reservoir stratigraphic heterogeneity (Fig. 6), as well as the response variable (Fig. 7).

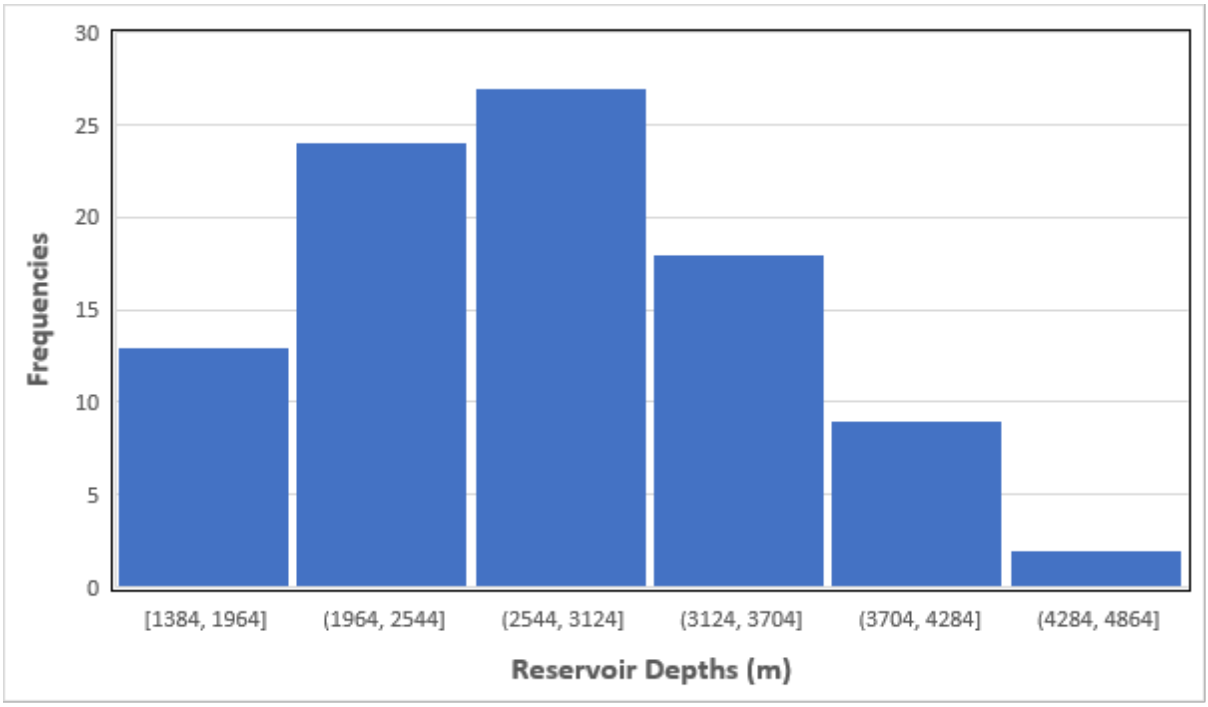

Figure 4: Distribution of reservoir depths of all the 93 reservoirs in the database. About half of the reservoirs are buried below 2,000 meters subsea.

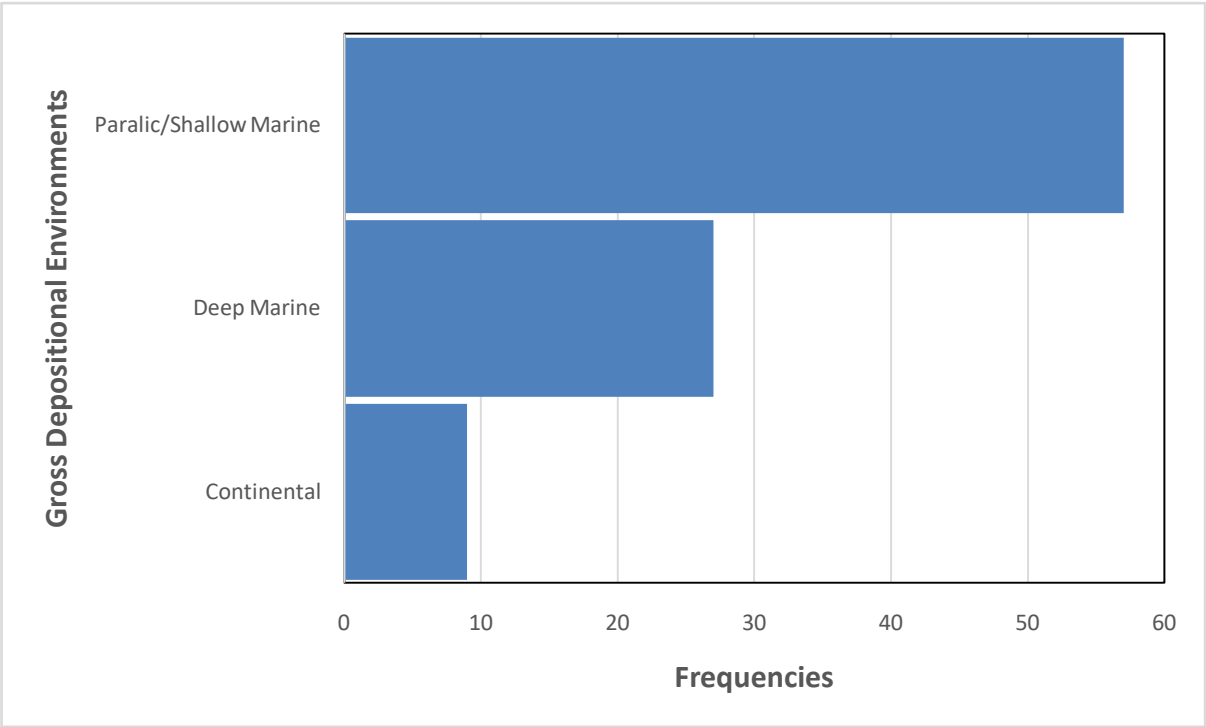

Figure 5: Proportion of the gross depositional environments of the reservoirs in the database.

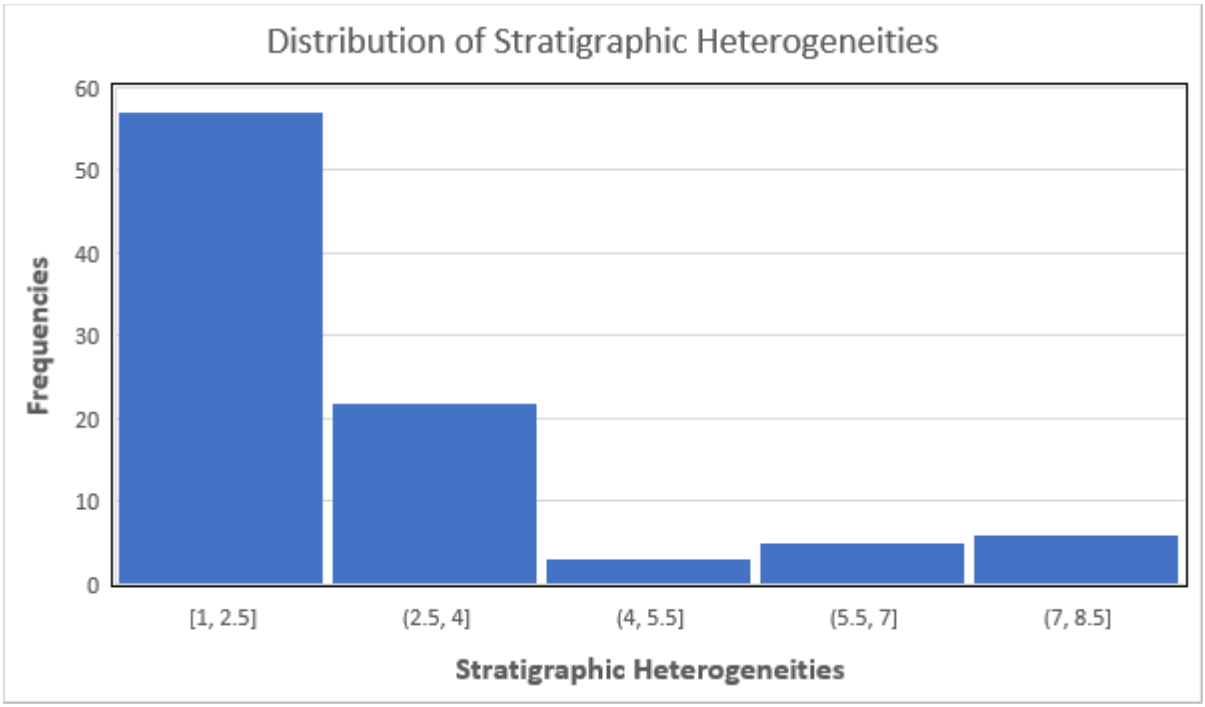

Figure 6: Distribution of stratigraphic heterogeneities in the database. Low values represent low heterogeneity, high values represent high heterogeneity.

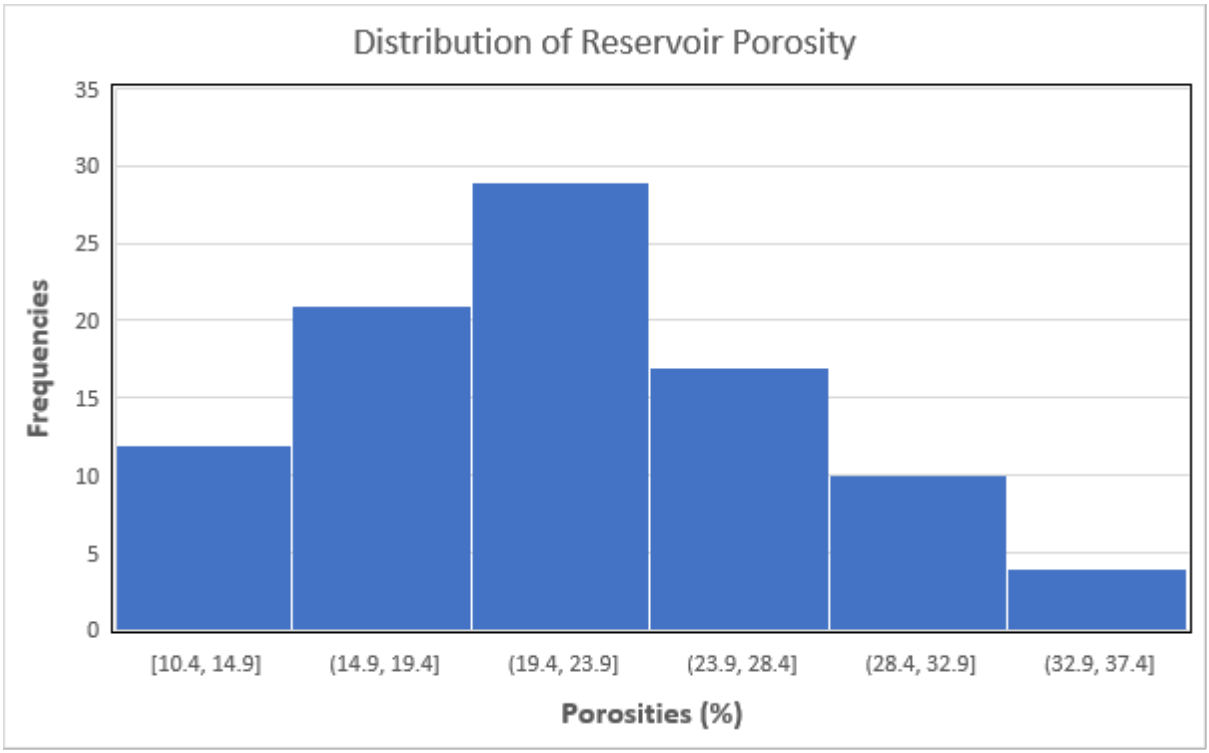

Figure 7: Porosity distribution of all reservoirs in the database. Reservoir porosity ranges from a minimum of 10% to a maximum of 37%.

Figures 8 and 9 demonstrate the correlation between two key predictors -reservoir depth of burial and stratigraphic heterogeneity with reservoir porosity. For the porosity against reservoir depth plot, it shows a slight decrease in porosity with increase in depth, except for a few outlier points which might indicate early migration of oil, halting reservoir porosity decline with increasing depth. The machine learning algorithms

learns from these data to make prediction. The relationship between porosity and reservoir stratigraphic heterogeneity (Fig. 9) is not as strong as the one between reservoir depth and porosity (Fig. 8), the plot still shows some level of correlation between the two variables.

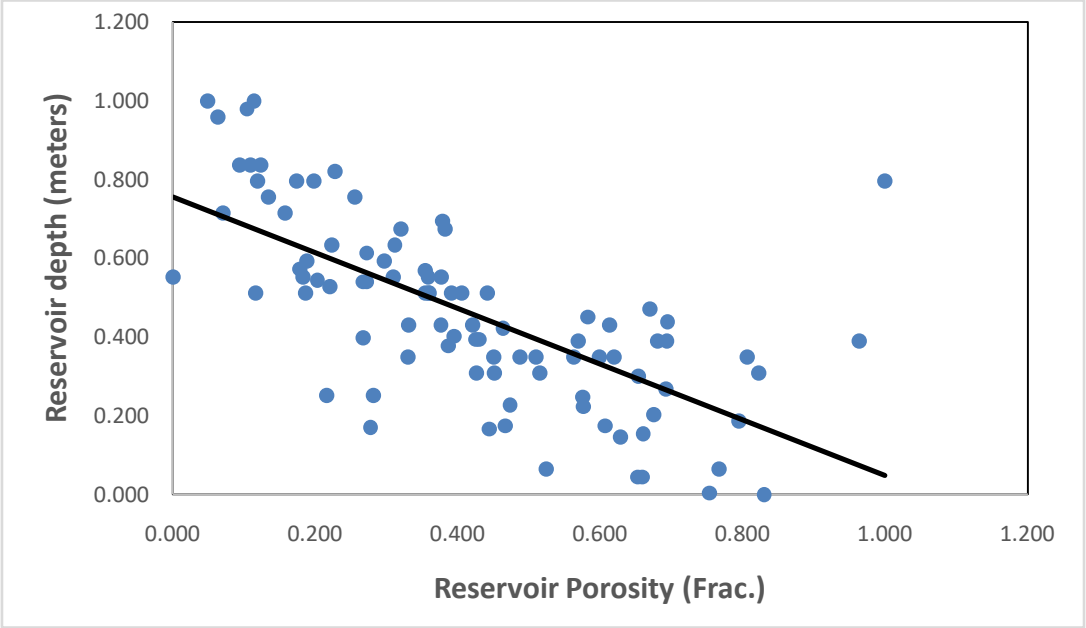

Figure 8: The relationship between reservoir depth and porosity, all measurements are in fraction (from 0 to 1).

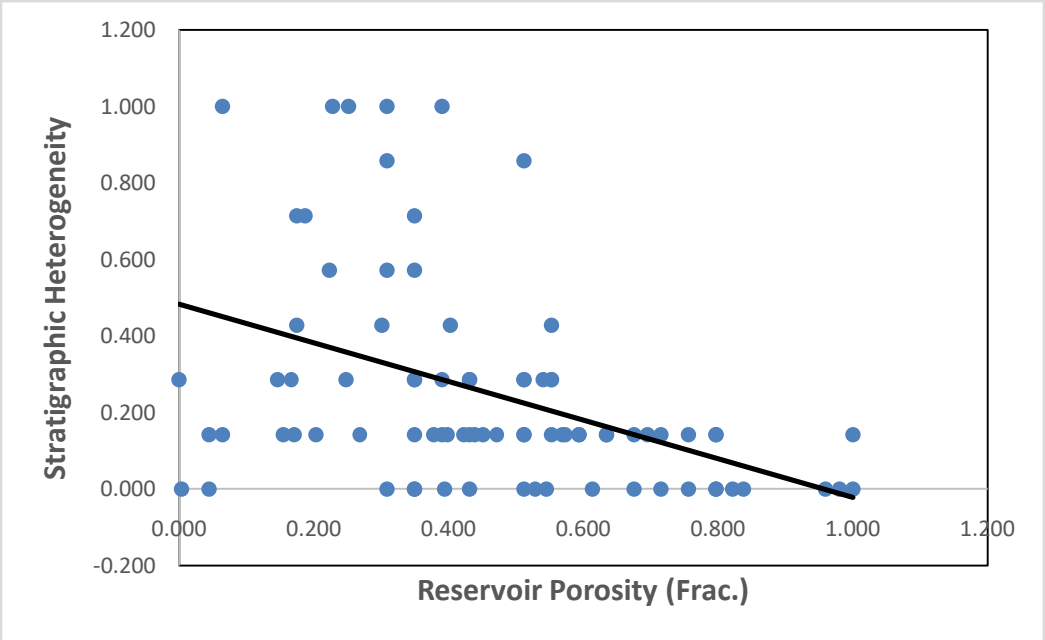

Figure 9: A plot of reservoir stratigraphic heterogeneity against porosity. Both measurements are in fraction.

We trained five different models using 5 different algorithms: Linear regression; support vector machine with a Gaussian kernel function, Boosted Tree with a minimum leaf size of 8, 30 number of learners and learning rate of 0.1; Bagged Tree with minimum leaf size of 8 and 30 number of learners; random forest regression with surrogate and 200 trees. The performance of the different models was compared using three metrics (Table 2), random forest model outperformed all other models. The comparison does not include model training time as no model took up to one minute to train.

Table 2: Performance of the different models trained compared using R-squared, root mean square error (RMSE) and mean absolute error (MAE).

Models	R2	RMSE	MAE
Linear Regression	0.57	0.155	0.116
Support Vector Machine	0.62	0.145	0.112
Boosted Tree	0.52	0.163	0.128
Bagged Tree	0.44	0.177	0.139
Random Forest	0.75	0.118	0.0028

Figures 10, 11 and 12 demonstrate the relationship between the predicted porosity and the actual porosity in the database for the random forest model. Fig. 12 shows a better match between the predicted porosity and the actual porosity in the test data with R^2 score of 0.87, compared to Fig. 10 and 11 with an R^2 of 0.75 and 0.71 respectively.

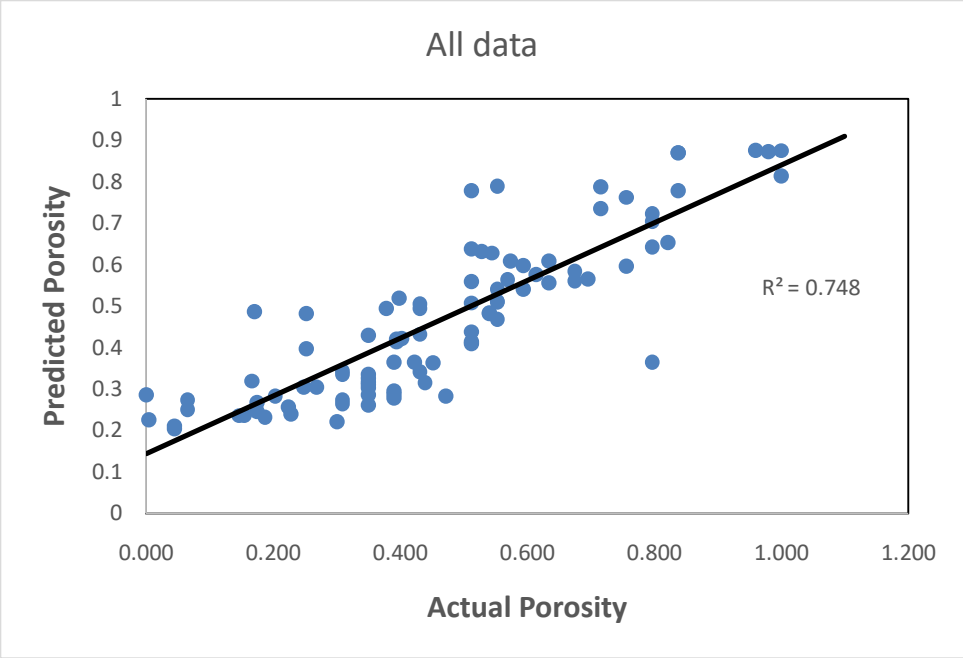

Figure 10: The relationship between predicted porosity and actual porosity for both training and testing data from the random forest model.

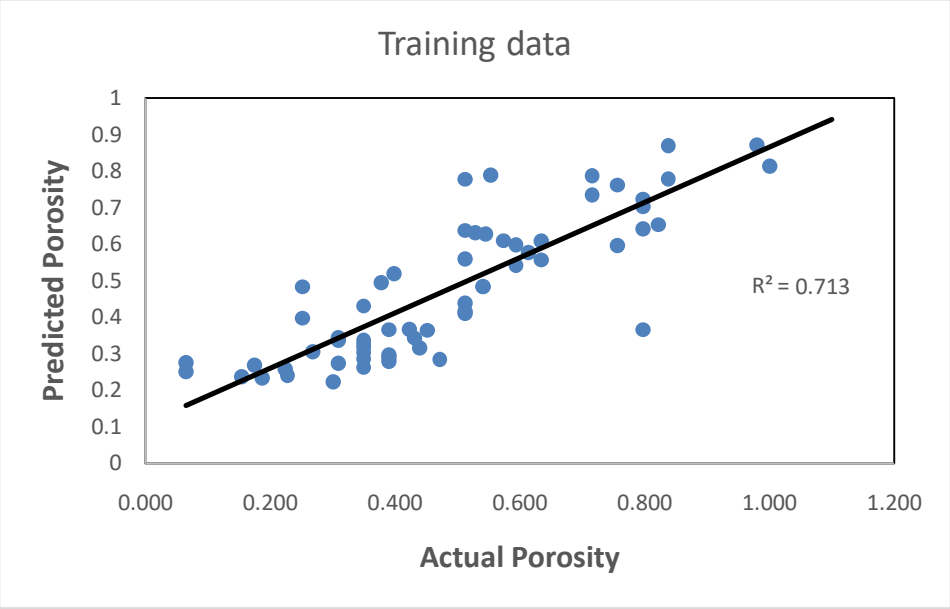

Figure 11: Cross plot of actual and predicted porosity for the training data.

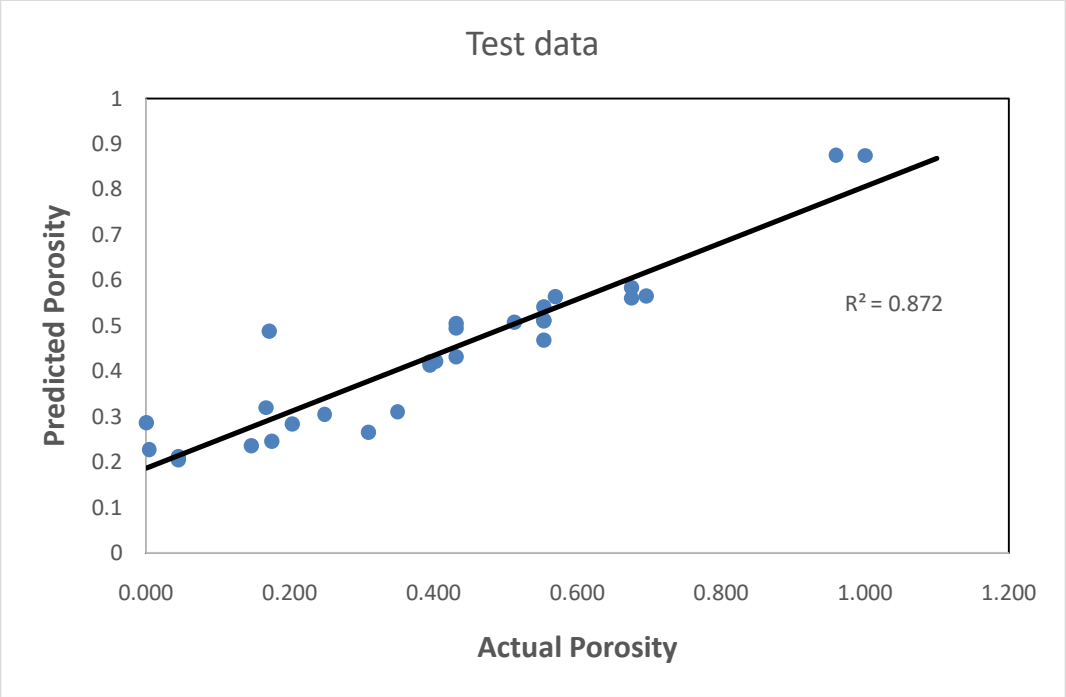

Figure 12: Relationship between the actual and predicted porosity for the test.

V. Conclusion

The machine learning technique of predicting porosity has numerous advantages over traditional techniques such as the empirical/semi-empirical formulae, Wyllie's equation and the density equation for porosity conversion where some suits of logs are used to predict porosity. The workflow shown in this study does not depend on any predetermined logs, it relays on a detailed characterization of the reservoir and its sedimentology. The machine learning approach represents a pragmatic approach to the classical log conversion problem that over the years has caused dilemmas to generations of geoscientists and petroleum engineers. The method requires no underlying mathematical models or costly assumptions of linearity among variables. Predicting porosity by using sedimentological parameters can effectively reduce the high cost of using petrophysical methods such as nuclear magnetic resonance and other logging methods.

The main limitation of the method is the amount of effort required to build a robust database, pre-processed the data and partition the data into training and testing sets, which is common for all models relying on real data, and the time to train and test the models. On the other hand, once established, the application of the models requires a minimum of computing time.

For the five porosity models trained, we find that models trained using random forest algorithm outperformed all the other models. The model has an R-squared score of 0.75 and MAE score of 0.0028.This study shows that machine learning has a strong potential to solve some important subsurface problems and could be an alternative to conventional methods of predicting porosity. This method can predict porosity not just around a wellbore but for some distance away from the well.

Acknowledgments

The authors wish to thank the University of Aberdeen for providing the software license of MATLAB used for this study.

Author Contributions

Conceptualisation: Kachalla Aliyuda

Data Curation: Aliyuda Ali, Kachalla Aliyuda

Data Analysis: Kachalla Aliyuda, Aliyuda Ali

Manuscript Writing: Kachalla Aliyuda

Manuscript Review and Editing: Kachalla Aliyuda, Jerry Raymond, Aliyuda Ali, Abdulwahab Muhammed Bello

Manuscript Proof-reading: Abdulwahab Muhammed Bello, Jerry Raymond

Competing Interests Statement

There are no financial conflicts of interest to disclose.

References Références Referencias

1. Ali. A, K. Aliyuda, M. K. Ahmed, and S. Saleh, (2021a). Data-driven-based pressure field decomposition and reconstruction for single-phase flow model. Science Forum (Journal of Pure and Applied Sciences), vol. 21, no. 2, pp. 357 – 364, 2021, doi: http://dx.doi.org/10.5455/sf.81326
2. Ali. A, Ahmed. M. K, Aliyuda. K, Bello. A. M, (2021b). Deep Neural Network Model for Improving Price Prediction of Natural Gas. 2021 International Conference on Data Analytics for Business and Industry (ICDABI) 978-1-6654-1656-6/21 ©2021 IEEEDOI: 10.1109/ICDABI53623.2021.9655885
3. Aliyuda Kachalla and John Howell. (2019). Machine-learning algorithm for estimating oil-recovery factor using a combination of engineering and stratigraphic dependent parameters. SEG Interpretation. Volume 7, Issue 3. 1A-T725pp.
4. Aliyuda Kachalla, John Howell and Humphrey Elliot. (2020) Impact of Geological variables in controlling Oil-reservoir performance: An insight from a machine-learning technique. SPE Reservoir evaluation and Engineering. 23 (04): 1314-1327 pp.
5. Breiman Leo. (2001) Random Forest. *Machine. Learning*. 45 (1). 5-32.https://doi.org/10.1023/A: 1 010933404324
6. Chen Zunde, Guo Aihua. (1998) Discussion on prediction permeability of seismic data. Oil Geophysical Prospecting, 33(S2): 86–90.
7. Chen, W., J. Xie, S. Zu, S. Gan, and Y. Chen, 2017, Multiple-reflections noise attenuation using adaptive randomized-order empirical mode decomposition: IEEE Geoscience and Remote Sensing Letters, 14, 18–22. doi: 10.1109/LGRS.2016.2622918.
8. Cortes Corinna and Vapnik Vladimir. (1995). Support-Vector Networks. Kluwer Academic Publishers, Boston. Machine Learning, 20, 273-297pp.
9. He Yan, Peng Wen, Yin Jun.(2001) Permeability prediction by seismic attribute data. Acta Petrolei Sinica, 22(6): 34–36.
10. Mann, C.J., (1987). Misuses of linear regression in earth sciences. In: B Size, W. (Ed.), Use and Abuse of Statistical Methods in the Earth Sciences. OUP, Oxford, pp. 74–106.
11. Seber G. A. F., (1977). Linear Regression Analysis. J. Wiley. New York.
12. Parra J. O., C. Hackert, M. Bennett, and H. A. Collier, (2003). Permeability and porosity images based on NMR, sonic, and seismic reflectivity: application to a carbonate aquifer, Ce Leading Edge, vol. 22, no. 11, pp. 1102–1108.
13. Leite E. P. and A. C. Vidal. (2011). 3D porosity prediction from seismic inversion and neural networks, Computers & Geosciences, vol. 37, no. 8, pp. 1174–1180.
14. Shi X., G. Liu, Y. Cheng et al.,(2016). Brittleness index prediction in shale gas reservoirs based on efficient network models, Journal of Natural Gas Science and Engineering, vol. 35, pp. 673–685.
15. Wang P., S. Peng, and T. He, (2018). A novel approach to total organic carbon content prediction in shale gas reservoirs with well logs data, Tonghua Basin, China. Journal of Natural Gas Science and Engineering, vol. 55, pp. 1–15.
16. Wang P. and S. Peng, (2018). 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.
17. Feng F, S., P. Wang, Z. Wei et al. (2020) A new method for predicting the permeability of sandstone in deep reservoirs, Geofluids, vol. 2020, Article ID 8844464, 16 pages.
18. Talkhestani A. A., (2015) Prediction of effective porosity from seismic attributes using locally linear model tree algorithm, Geophysical Prospecting, vol. 63, no. 3, pp. 680–693.
19. Wang P. and S. Peng, (2019). On a new method of estimating shear wave velocity from conventional well logs, Journal of Petroleum Science and Engineering, vol. 180, pp. 105–123.
20. Haklidir F. S. T. and M. Haklidir, (2020). Prediction of reservoir temperatures using hydrogeochemical data, western anatolia geothermal systems (Turkey): a machine learning approach, Natural Resources Research, vol. 29, no. 4, pp. 2333–2346.
21. Mahmoud A. A., S. Elkatatny, and D. Al Shehri, (2020). Application of machine learning in evaluation of the static young's modulus for sandstone formations, Sustainability, vol. 12, no. 5, p. 1880.
22. He M., H. Gu, and H. Wan, (2020). Log interpretation for lithology and fluid identification using deep neural network combined with MAHAKIL in a tight sandstone reservoir, Journal of Petroleum Science and Engineering, vol. 194, p. 107498.
23. Vernik L. (1997). Predicting porosity from acoustic velocities in siliciclastics: a new look. Geophysics 62, 118±128 pp.
24. Zhang Zhenhua, Yanbin Wang, and Pan Wang. (2011).On a Deep Learning Method of Estimating Reservoir Porosity Mathematical Problems in Engineering Volume 2021, Article ID 6641678, 13 pages.