Iraqi Geological Journal

Journal homepage: https://www.igj-iraq.org

K-Mean Clustering Analysis and Logistic Boosting Regression for Rock Facies Characterization and Classification in Zubair Reservoir in Luhais Oil Field, Southern Iraq

Mohammed Albuslimi[1,*], Yasir Alkalby[2] and Tabarak Al-Taweel[3]

[1] Basra Oil Company, Basra, Iraq
[2] Anton Oilfield Services (Group) Ltd, Basra, Iraq
[3] University of Basrah, Basra, Iraq
* Correspondence: mohammed.a.abbas@majnoon-ifms.com

Received: 11 February 2021; Accepted:23 June 2021; Published: 31 August 2021

Abstract

Identifying rock facies from petrophysical logs is a crucial step in the evaluation and characterization of hydrocarbon reservoirs. The rock facies can be obtained either from core analysis (lithofacies) or from well logging data (electrofacies). In this research, two advanced machine learning approaches were adopted for electrofacies identification and for lithofacies classification, both given the well-logging interpretations from a well in the upper shale member in Luhais Oil Field, southern Iraq. Specifically, the K-mean partitioning analysis and Logistic Boosting (Logit Boost) were conducted for electrofacies characterization and lithofacies classification, respectively. The dataset includes the routine core analysis of core porosity, core permeability, and measured discrete lithofacies along with the well-logging interpretations include (shale volume, water saturation and effective porosity) given the entire reservoir interval. The K-Mean clustering technique demonstrated good matching between the vertical sequence of identified electrofacies and the observed lithofacies from core description as attained 89.92% total correct percent from the confusion matrix table. The Logit Boost showed excellent matching between the recognized lithofacies from the core description and the predicted lithofacies through attained 98.26% total correct classification rate index from the confusion matrix table. The high accuracy of the Logit Boost algorithm comes from taking into account the non-linearity between the lithofacies and petrophysical properties in the classification process. The high degree of lithofacies classification by Logit Boost in this research can be considered in a similar procedure at all sandstone reservoirs to improve the reservoir characterization. The complete facies identification and classification were implemented with the programming language R, the powerful open-source statistical computing language.

Keywords: Clustering analysis; Logistic boosting regression; Lithofacies classification; Electrofacies characterization; Luhais Oil Field; Zubair Formation

1. Introduction

Proper identification of reservoir rock facies is a crucial process for reservoir characterization and development, especially in the lack of extensive core information. The discrete rock facies are vitally required to enhance the porosity-permeability relationship in non-cored wells and preserve the

DOI: 10.46717/igj.54.2B.6Ms-2021-08-26

heterogeneity of the reservoir. The vertical distribution of rock facies affects entire uncertainties in the reservoir as it's an essential process of areal facies distribution in the geological model. Thus, it is necessary to find the most realistic facies prediction to reach representative reservoir characterization, modeling, better productivity estimation, and investment economic decisions through the field life cycle (Al-Mudhafer & Bondarenko, 2015).

It is challenging to obtain the vertical lithofacies sequence for the whole reservoir interval for all wells due to the high expense and the needed time for their acquisition. Therefore, a variety of Advanced Data-Driven approaches have been proposed to estimate electrofacies or lithofacies distribution in non-cored wells. It can be summarized into two types: 1) The unsupervised machine learning techniques to partition well log response into different clusters and estimate electrofacies such as Model-Based Clustering (Woan et al., 2012), k-Mean Clustering (Abbas & Al Lawe, 2019) and Ward's hierarchical clustering (Al-Jafar & Al-Jaberi, 2019). 2) The supervised machine learning algorithms are used for lithofacies classification given well-logging and observed lithofacies from cored wells to predict their distribution in other non-cored wells. in the last few years, several supervised machine learning algorithms have been published in many works of literature to obtain the lithofacies classification such as Naive Bayes Classifier (Murphy, 2006), Logistic Boosting Regression (Marc, 2017), Linear Discriminant Analysis (Pires & J, 2010), Multinomial Logistic Regression (Long, 2006), Tree-Based Classification Models (Breiman et al., 1984) and Kernel Support Vector Machine (Al-Mudhafer, 2017). Moreover, there are several empirical methods for rock typing which are highly dependent on core analysis data such as flow zone indicator (Amaefule et al., 1993), Winland's R-35 (Pittman, 1992), and Leveret's J-function (Leverett, 1940).

The Luhais Oil Field was discovered in 1961 and located in a structural region known as the Mesopotamian Basin between 47° 14'- 47° 19' and (30° 13'-30° 24') latitude and longitude, respectively, south of Iraq (Habeeb & Al-Dulaimi, 2018). The field was started producing in 1971 with an API gravity of 32° (Alher et al., 2018). The length of the Luhais oil field is around 20 km and the width range from 5 km at the northern part to10 km in the central and southern portion of it. The clastic formation (Zubair) is the principal producing sandstone reservoir in the oil fields, southern Iraq. Zubair formation in the field under study comprised mainly of sandstone, shale, and interbedded siltstone, which can be concluded from the core description and the behavior of petrophysical properties.

In this study, the K-Mean clustering analysis and Logit Boost were implemented for electrofacies estimation and lithofacies modeling given well-logging interpretations and routine core analysis from a cored-well in the upper shale member in Luhais oil field in the south of Iraq. The Machine Learning techniques rigorously were setup to take into consideration the non-linearity between the core measurements and well logging data. We noticed that the Logit Boost algorithm performed much higher classification accuracy than the K-Mean algorithm through attained 98.26% total correct classification rate index from the confusion matrix table. The high-quality estimation of lithofacies distribution is required for the next step as include as an additional predictor in the core permeability prediction to produce separate regression lines for each facies type and attaining a representative geological and simulation model.

2. Geological Description

The Luhais oil field is a gentle anticline fold situated in the south of Iraq (around 100 km in the northwest of Basra) as demonstrated in Fig. 1. The main producing reservoirs in this field are Zubair and Nahr-Umar (Al-Yasri & Al-Baldawi, 2015). Zubair is an oil reservoir that belongs to the Lower Cretaceous age deposited in the fluvial-deltaic, deltaic and marine environments (Al-Mudhafer, 2017). Zubair Formation is extended in the Arabian Plate including northern Saudi Arabia, Kuwait, and Iraq (Mohammed & Al-Zaidy, 2018). The Zubair Formation in southern Iraq is divided into five members:

Upper Shale; Upper Sand; Middle Shale; Lower Sand, and Lower Shale. The petrophysical analysis of Zubair Formation in the Luhais oil field refers to only three members.

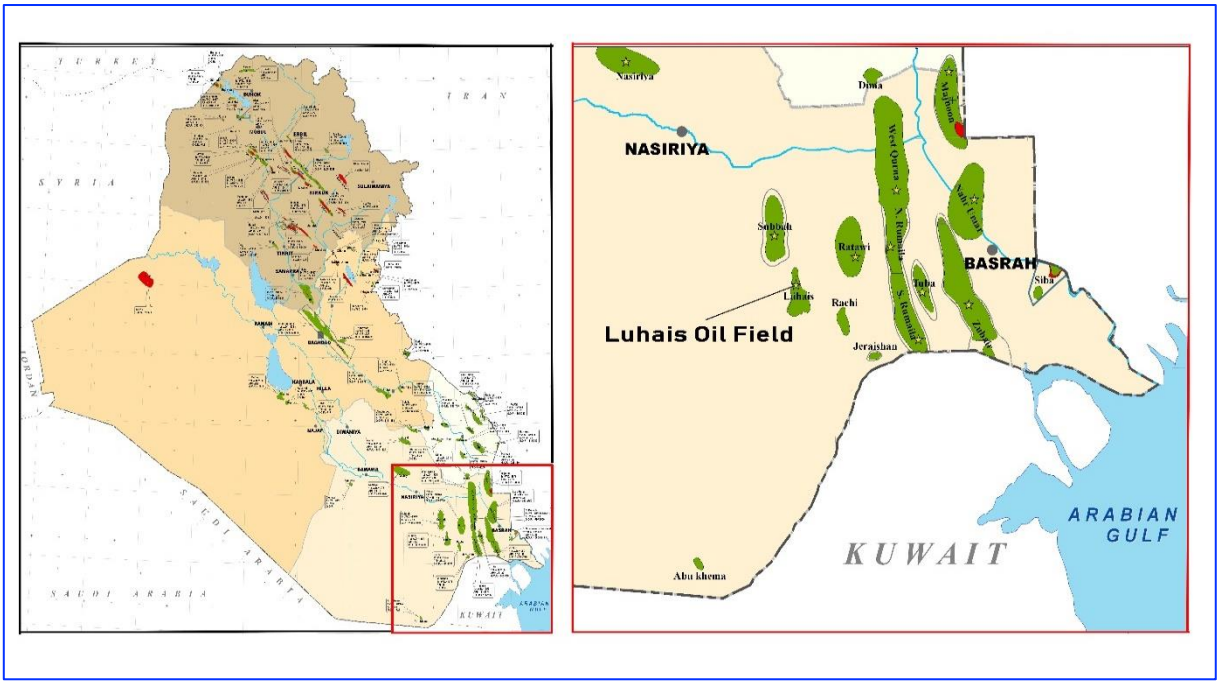

Fig.1. Geographical locations of Luhais Oil Field (Abbas & Al Lawe, 2019)

The nonexistence of the middle shale member was lead to incorporating upper and lower sand members to form the Middle Sand (Al-Yasri & Al-Baldawi, 2015; Al-Zaidy & Mohammed, 2017; Saqer et al., 2019). The upper shale member is the focus of this research, is an oil-bearing zone located above water-oil contact, while the other two are situated in the water zones. The core description in the upper shale showed three main lithofacies: shale, siltstone and sandstone. In this regard, the well-logging interpretations (shale volume, effective porosity, and water saturation) were incorporated along with conventional core experiments (core porosity, core permeability, and discrete lithofacies). The core to logs combination was performed by depth shifting of core porosity to the reference log derived porosity to correlate each core sample to the proper depth position. As shown in Fig. 2, the well-logging interpretations are similar to petrophysical behavior obtained from core analysis. Table 1 and Fig. 3 summarize the statistical indexes and the pairwise scatterplot of integrated log and core data for the well under study.

Table 1. Statistical parameters of the integrated log and core data of a reference well

Index	Depth	PHIE	PHIE Core	SW	VSH	Facies
Min	2798	0.0001	0.0150	0.0733	0.0014	Sh. :20
1st Qu	2819	0.08765	0.1370	0.7723	0.0372	Slt. :14
Median	2828	0.19000	0.2140	0.9268	0.1641	Sst. :85
Mean	2824	0.15046	0.1818	0.8265	0.2461	
3rd Qu	2832	0.22000	0.2420	0.9900	0.3785	
Max	2845	0.26100	0.2860	1.0000	0.9676	

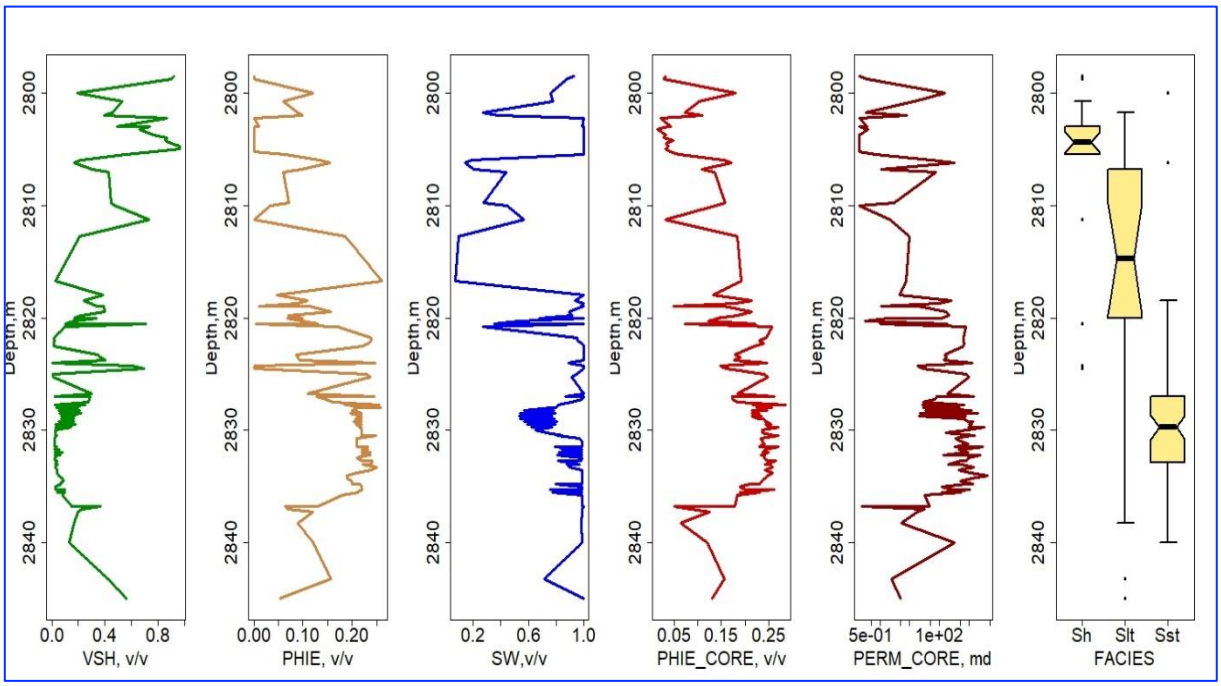

Fig.2. Log view of vertical sequence for well-logging interpretations and core analysis of a reference well

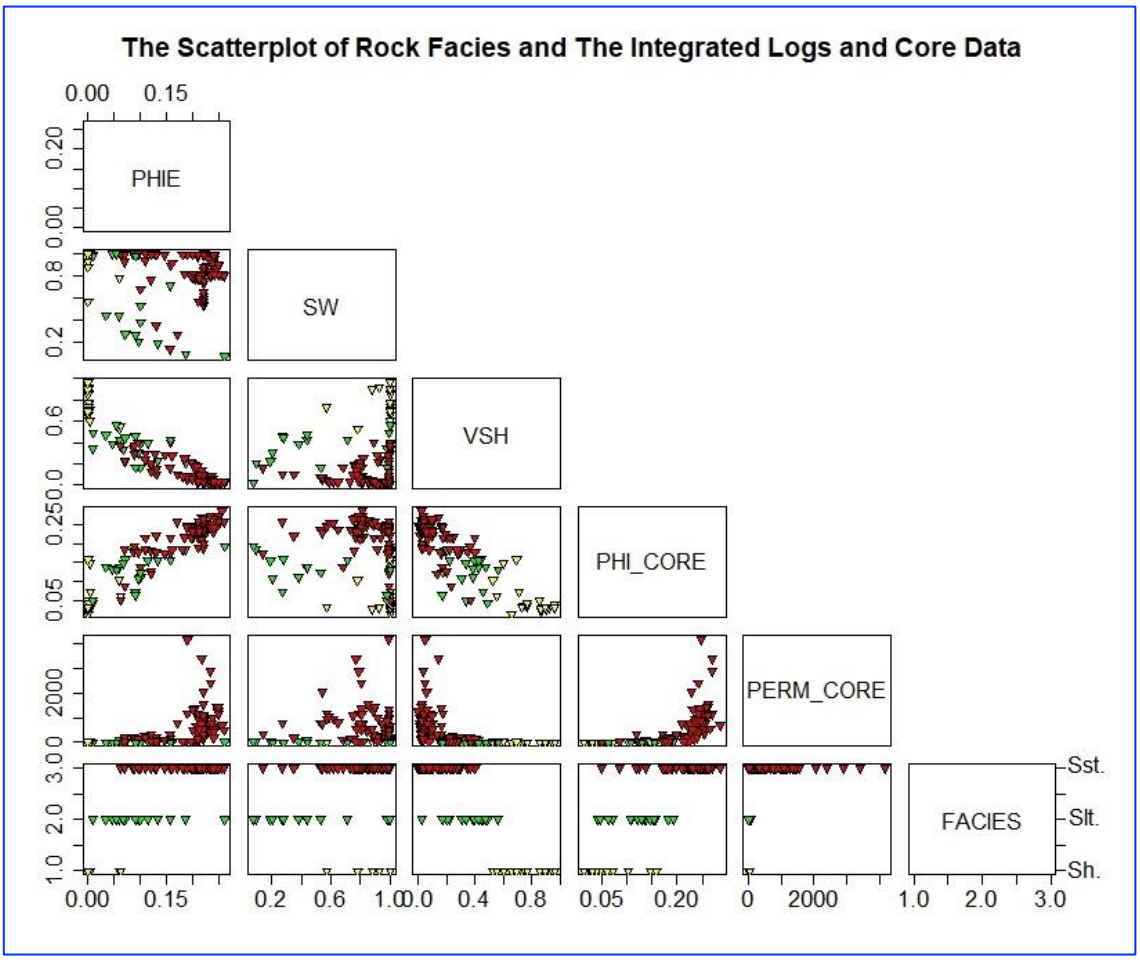

Fig. 3. The pairwise scatterplot of integrated log and core data of a reference well

3. Materials and Methods

3.1. Electrofacies Characterization

During the last sixty years, many statistical techniques have been developed to categorize the reservoir rocks. Data clustering is one of the most important and widespread methods. The main objective of clustering is to group data points into a specific number of clusters due to their similarities. In general, there are various types of clustering that involve: density-based, prototype-based, graph-based, well-separated, and shared property (conceptual clusters) (Al-Mudhafer & Al Lawi, 2019).

3.1.1. K-mean clustering analysis

K-means clustering is the simplest and most common unsupervised machine learning technique. It simply tends to cluster the data points and found the underlying pattern by looking for a fixed number of (k) clusters in the data, which are referred to the number of centroids in the database. Once the optimal K clusters in the data have been determined, they will be used randomly as centroids. Then distances have to be computed between each data point and the centroids. Afterward, the points would be allocated to the nearest cluster. This process will be repeated till constant cluster centroids will be achieved (Subbalakshmi, 2019). The main objective of the K-means algorithm is to decrease the the squared error function, objective function (Pang Ning et al., 2006) that can be formulated mathematically as shown below:

$$J = \sum_{\mathrm{j}=1}^{\mathrm{k}} \sum_{i=1}^{n} \| \mathrm{X}_{\mathrm{i}}^{(\mathrm{j})} - c_j \|^2 \qquad (1)$$

Where:
J= objective function
K= number of clusters
n=number of cases
X: case
C: centroid

3.1.2. Optimal number of clusters

Prior data partitioning, the number of clusters must be predicted and utilized as input for the K-Mean algorithm (Abbas & Al Lawe, 2019). Several statistical approaches can be used for determining clusters number. In this paper, the Silhouette analysis k-medoids (K-PAM) have been applied. K-PAM is similar to the K-means, it used the medoids for data partitioning, not the mean values (Hothorn & Everitt, 2009).

3.2. Lithofacies Classification

The classification represents forecasting the observed discrete lithofacies in a well as the function of well-logging in order to predict their distribution in other wells that have no observed facies. (Al-Muthafer, 2020). In this research, the Logistic Boosting Regression (LogitBoost) was used as an efficient classification method to mimic the measured lithofacies distribution in the well of sandstone reservoir under study.

3.2.1. Logistic boosting regression

Logistic Boosting is one of the most recent developments in supervised machine learning technique that is an ensemble method of combing the weak misclassified samples to produce powerful classifiers

and ameliorating the classification. The previous weak classifiers samples will be boosted and added sequentially to the model to reduce the residual loss (Al-Mudhafar, 2016). The ambiguity of this methodology can be understood by describing the well-known statistical tools, namely additive modeling and maximum likelihood (Friedman, 2000). Stump decision trees are adopted in LogitBoost that are flexible, interpretable and computationally efficient. Each decision tree tends to be shallow models that do not overfit but can be biased. With stumps, LogitBoost uses considerably less data than the others algorithms and is thereby correspondingly faster (Marc, 2017). However, LogitBoost is befitted for imbalanced datasets but unsatisfactory in performance within the presence of noise (Siddharth et al., 2020). The Logit term refers to the logarithm of odds wherever odds define as the ratio of the probability of success to the probability of failure. The logit function used to model the probability p(x) using curve where the predictor X ϵ R, p(x) ϵ (0,1) as following:

$$Log\left(\frac{p(x)}{1-p(x)}\right) = \beta(x) \quad (2)$$

The goal of learning is estimate β using maximum the following likelihood function:

$$l(\beta) = \sum_{i=1}^{n} y_i \beta_i - \log\left(1 + e^{\beta_{xi}}\right) \quad (3)$$

The likelihood function is a non-algebraic (transcendental) equation that needs to use Quasi-Newton as numerical tool for approximation the solution (Friedman, 2000)

4. Results and Discussion

4.1. K-means Partitioning

The analysis of well-logging interpretations (effective Porosity, water saturation, and shale volume) using the K-PAM algorithm led to three groups represent the ideal number of clusters as shown in Fig. 4.

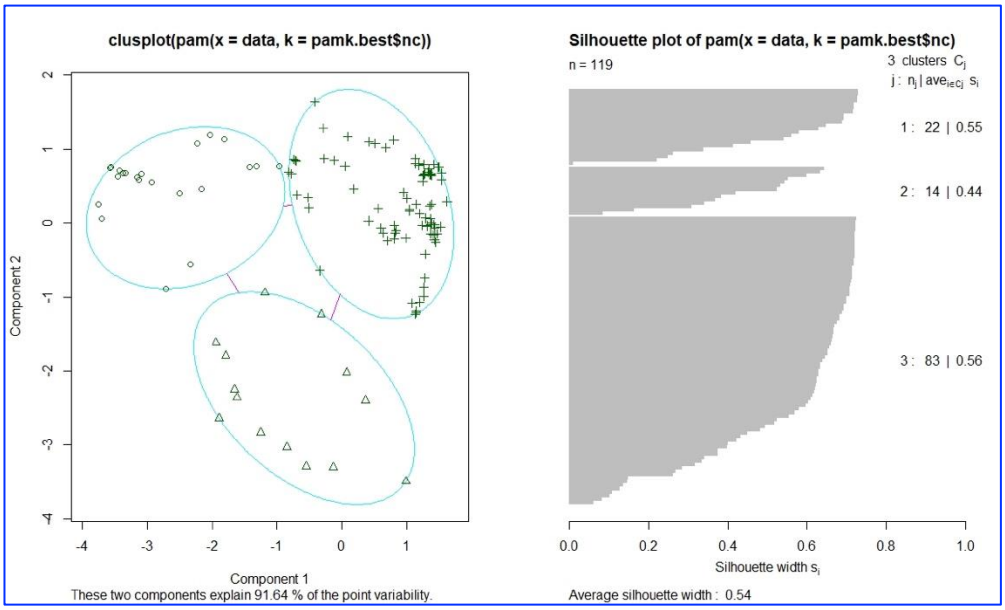

Fig.4. The optimal number of clusters suit the data determined by the k-pam algorithm

The three Clusters are used as input for the K-Mean clustering algorithm to partition the data. In order to evaluate the clustering as well as classification accuracy, the matrix confusion table between the observed lithofacies and predicted litho/electro distribution was constructed to compute the total correct percent, which represents the count of success and fail points. As demonstrated in Fig. 5, the K-Mean clustering analysis showed good matching between the vertical distribution of identified electrofacies and the observed lithofacies from the core description through attaining 89.92% total correct percent. The numerical results showed that the k-mean algorithm mismatched a few observations that can be identified from the confusion matrixes in Table 2.

Table 2. Confusion accuracy matrix of the k-mean algorithm for the entire dataset

	Sh.	**Slt.**	**Sst**
Sh.	18	0	0
Slt.	2	10	6
Sst.	0	4	79

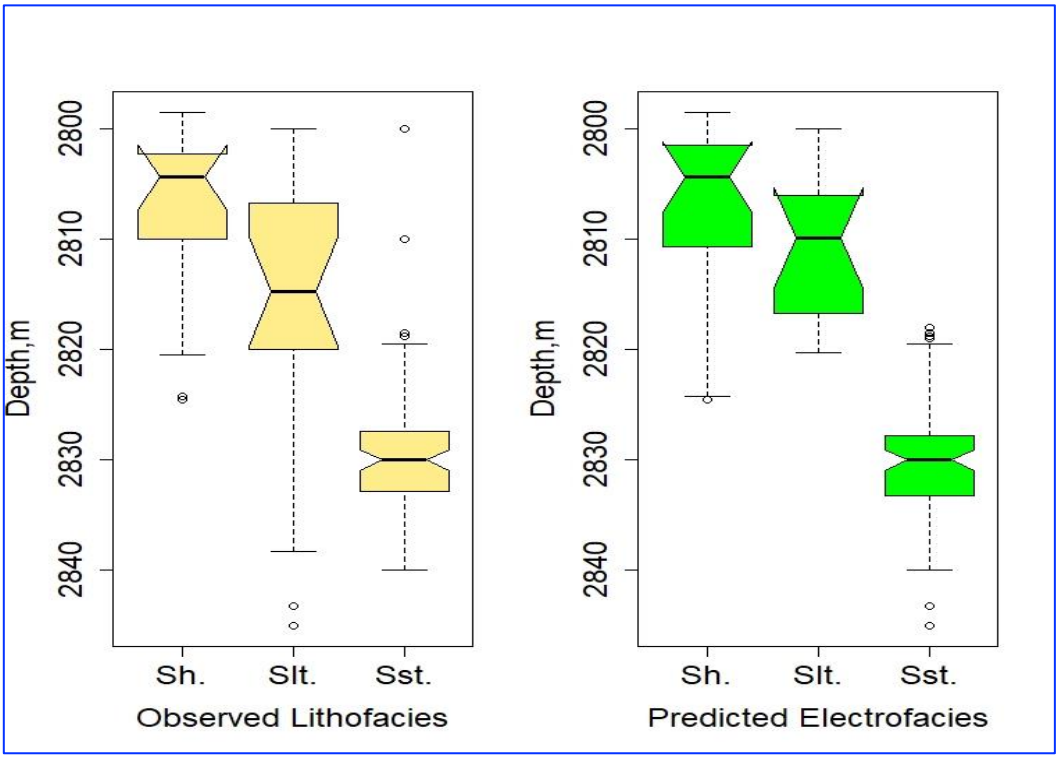

Fig.5. Comparison between the observed lithofacies and k-means-predicted electrofacies for the entire dataset

4.2. Logistic Boosting Regression

The classification was conducted using the Logistic Boosting Regression algorithm by caTools package to classify and estimate the discrete lithofacies given well-logging interpretations. The classification was first implemented given the entire data set and then using cross validation. In the first approach, the modeling and prediction was adopted as a function of the full data set. Fig. 6 illustrates the matching between the measured and forecasted lithofacies distribution for the full dataset. Table 3 represents the LogitBoost resulting confusion table that shows an accuracy rate of 98.26%.

Table 3. Confusion accuracy matrix of the logit boost algorithm for the entire dataset

	Sh.	**Slt.**	**Sst**
Sh.	18	0	0
Slt.	0	16	2
Sst.	0	0	83

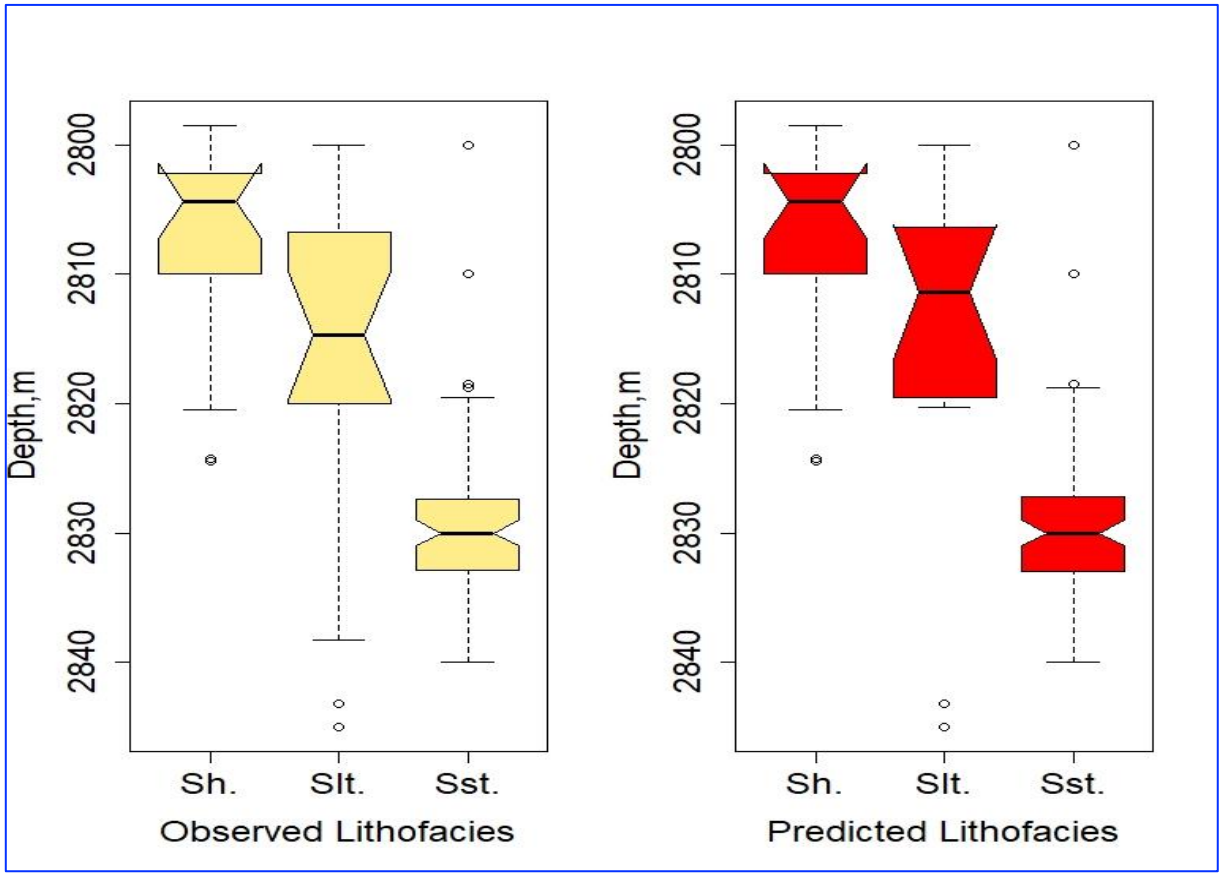

Fig.6. Matching between the observed and logitboost-predicted lithofacies for the entire dataset

The second approach of lithofacies classification by LogitBoost was implemented through the random subsampling cross-validation. The full data set was sub-divided into two parts: 70% for training and 30% for testing. The modeling was conducted on the training subset and estimation was obtained given the training and testing subsets, respectively. The total correct present of the predicted lithofacies given the training subset was 98.8%; however, the accuracy given the testing subset was 94.12%. Fig. 7 and 8. decorate the boxplots matching between the measured and predicted facies given the training and testing subsets, respectively. Tables 4 and 5. represents the total percent correct of LogitBoost classification given the training and testing subsets, respectively.

Table 4. Confusion accuracy matrix of the logitboost algorithm for the train-dataset

	Sh.	**Slt.**	**Sst**
Sh.	14	0	0
Slt.	0	12	1
Sst.	0	0	56

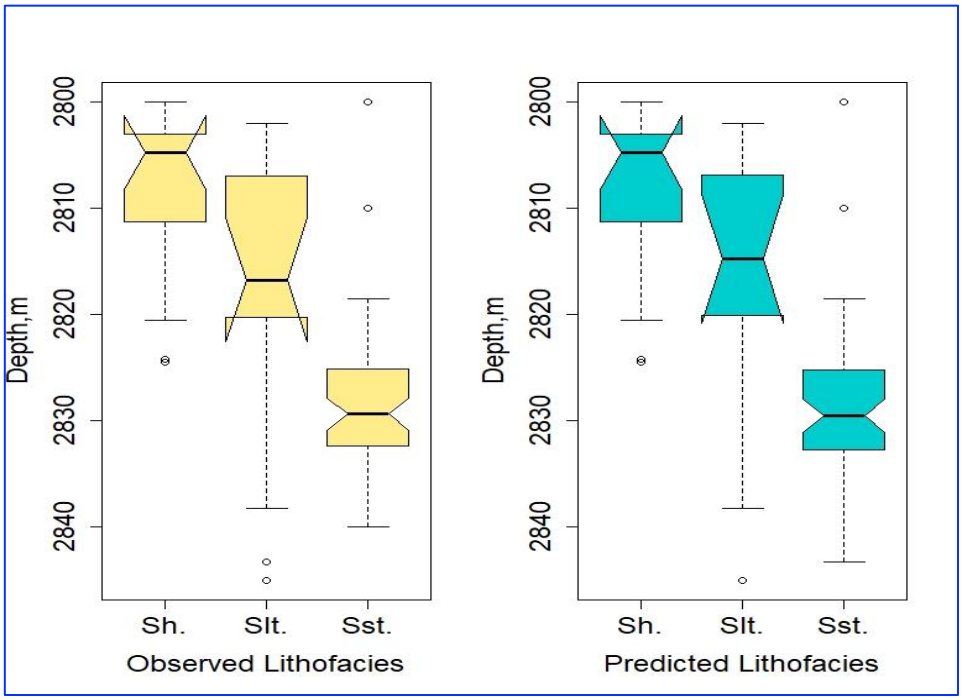

Fig. 7. Matching between the observed and logitboost-predicted lithofacies for the train-dataset

Table 5. Confusion accuracy matrix of the logitboost algorithm for the test-dataset

	Sh.	**Slt.**	**Sst**
Sh.	4	0	0
Slt.	0	5	0
Sst.	0	2	25

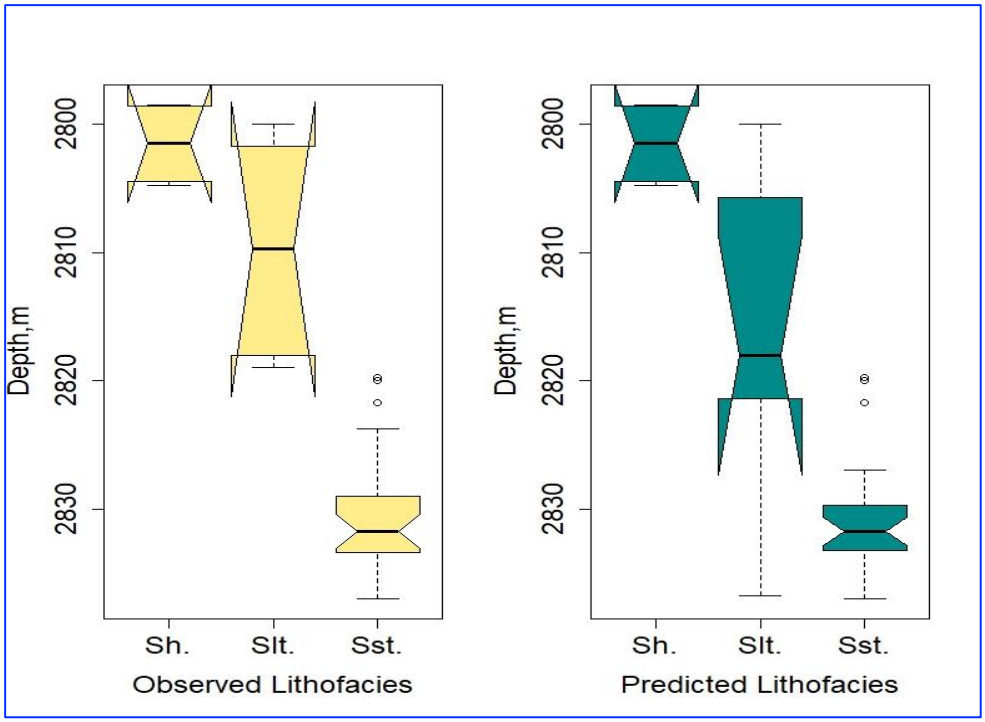

Fig. 8. Matching between the observed and logitboost-predicted lithofacies for the test-dataset

5. Conclusions

Two advanced unsupervised and supervised machine-learning approaches were adopted for electrofacies identification and lithofacies classification, respectively. The K-mean clustering analysis was used to identify the electrofacies from the well logging interpretations (CPI). Since the clustered electrofacies can be conducted for all the wells in the reservoir, especially when the well-logging interpretations are available in most cases. Since the availability of lithofacies distribution was only for one well, it was needed to model the lithofacies as a function of the well-logging interpretations of that well to be predicted lithofacies for all other wells. The Logistic Boosting Regression was used for lithofacies classification and it achieved an exact accuracy of 98.26% total correct percent of facies classification. In contrast, k-mean clustering analysis led to good matching between observed lithofacies and identified electrofacies as attained 89.92 of the total-correct percent. The results reflect the applicability of machine learning to overcome the challenge of reservoir description with a lack of data. The R scripts that applied in this research can be used in other reservoirs to characterize and predict electrofacies and lithofacies.

Acknowledgements

The authors would like to thank Dr. W.J. Al-Mudhafar, the Research Advisor at the University of Texas at Austin & Chief Reservoir Engineer at Basrah Oil Company, for his guidance and supports. The authors are very grateful to the Editor in Chief Prof. Dr. Salih M. Awadh, the Secretary of Journal Mr. Samir R. Hijab and the Technical Editors Dr. Heba S. Al-Mimar for their great efforts and valuable comments.

References

Abbas, M. & Al Lawe, E., 2019. Clustering Analysis and Flow Zone Indicator for Electrofacies Characterization in the Upper Shale Member in Luhais Oil Field, SouthernIraq. Abu Dhabi,UAE, SPE.

Alher, A. A., Aljawad, M. S., & Ali, A. A., 2018. Static model of Zubair Reservoir in Luhais Oil Field. Iraqi Journal of Chemical and Petroleum Engineering, 19(1), 57-60.

Al-Jafar, M. . K. & Al-Jaberi, M. . H., 2019. Well logging and electrofacies of Zubair Formation for Upper sandstone member in Zubair Oil Feld, Southern Iraq. Iraqi Geological Journal, 52, 101- 124.

Al-Mudhafar, W. J., 2016. Applied Geostatistical Reservoir Characterization in R: Review and Implementation of Rock Facies Classification and Prediction Algorithms-Part I. Houston, Texas, SPE.

Al-Mudhafar, W. J., 2017. Integrating kernel support vector machines for efficient rock facies classification in the main pay of Zubair formation in South Rumaila oil field, Iraq. Modeling Earth Systems and Environment, 3(1), 12.

Al-Mudhafer, W. & Al Lawi, E., 2019. Clustering Analysis for Improved Characterization of Carbonate Reservoirs. Houston,Texxas ,USA.

Al-Mudhafer, W. & Bondarenko, M., 2015. Integrating K-Means Clustering Analysis and Generalized Additive Model for Efficient Reservoir Characterization. Madrid, Spain.

Al-Muthafer, W. J., 2020. Advanced Supervised Machine Learning Algorithms for Efficient Electrofacies Classification of a Carbonate Reservoir In a Giant Southern Iraqi Oil Field. Houston,USA, SPE.

Al-Yasi, A. I., & Al-Baldawi, B. A., 2015. Using geophysical well logs in studying reservoir properties of Zubair Formation in Luhais oil field, Southern Iraq. Iraqi Journal of Science, 56(3C), 2615-2626.

Al-Zaidy, A. A., & Mohammed, K. S., 2017. Petrophysical evaluation and reservoir characterization of the Zubair Formation in the Luhais and Rachi oil fields, Southern Iraq. International Journal of Advanced Engineering Research and Science, 4(12), 237332.

Amaefule, J. O., 1993. Enhanced Reservoir Description: Using Core and Log Data to Identfy Hydraulic Flow Units and predict Permeability in uncored Intervals, Wells. Houston,Texas, USA.

Breiman, L., Friedman, J., Olshen, R. & Stone, C., 1984. Classification and regression trees. In: Boca Raton, FL: Chapman and Hall, CRC Press.

Subbalakshmi, C., Rao, S. K. M., & Rao, S. K. M. 2014. Performance issues on K-mean partitioning clustering algorithm. International Journal of Computer, 14(1), 41-51.

Friedman, J. T. H. a. R. T., 2000. Additive Logistic Regression: a Statistical View of Boosting (with Discussion and a Rejoinder by the Authors). In: The Annals of Statistics.

Habeeb, A. J. & Al-Dulaimi, S. . I., 2018. Reservoir facies of Nahr Umr Formation in Luhais Oilfield, southern Iraq. International Journal of Science and Research, 7(5), 31-36.

Hothorn, T. & Everitt, B., 2009. A Handbook of Statistical Analyses Using R. CRC Press.

Leverett, M., 1940. Capillary Behavior in Porous Solids. In: Petroleum Technology, 152-169.

Long, J. S. A. J. F., 2006. Regression Models for Categorical and Limited Dependent Variables Using Stata. Texas: Stata Press.

Marc , G., 2017. Logit Boost autoregressive networks. Computational Statistics and Data Analysis, 88-98.

Mohammed, K. & Al-Zaidy, A., 2018. Facies Analysis and Sequence Stratigraphy of the Barremian Succession in the Luhais and Rachi Oil Fields, Southern of Iraq. Journal of University of Babylon for Engineering Sciences, 26.

Murphy, K. P., 2006. Naive Bayes Classifiers:University of British Columbia.

Pang Ning, T., Steinbach, M. & Kumar, V., 2006. Introduction to Data Mining. Boston, USA.

Pires, A. & J, A., 2010. Projection-pursuit approach to robust linear discriminant analysis. Journal of Multivariate Analysis, 101, 2464-2485.

Pittman, E., 1992. Relationship of porosity and permeability to various parameters derived from mercury injection-capillary pressure curves for sandstone. American Association Petroleum Geologists,191-198.

Saqer, M. H., AL-Shahwan, M. F. & AL-Yasiri, A. A., 2019. The standerd model of oil charachteristics of Zubair Reservoir in Luhais Field Sothren Iraq. Biochemical and Cellular Archives, 19, 1849-1855,.

Siddharth , M., Hao, L. & Jiabo , H., 2020. Machine Learning for Subsurface Characterization.

Woan, . J. T., G Paul , W. & John , H. D., 2012. Improved Reservoir Characterization using Petrophysical Classifiers within Electrofacies. Oklahoma, USA, SPE.