PETROVIETNAM JOURNAL
Volume 6/2022, pp. 27 - 35
ISSN 2615-9902

SUPERVISED MACHINE LEARNING APPLICATION OF LITHOFACIES CLASSIFICATION FOR A HYDRODYNAMICALLY COMPLEX GAS - CONDENSATE RESERVOIR IN NAM CON SON BASIN

Nguyen Ngoc Tan, Tran Ngoc The Hung, Hoang Ky Son, Tran Vu Tung
Bien Dong Petroleum Operating Company (BIENDONG POC)
Email: sonhk@biendongpoc.vn
https://doi.org/10.47800/PVJ.2022.06-03

Summary

Conventional integration of rock physics and seismic inversion can quantitatively evaluate and contrast reservoir properties. However, the available output attributes are occasionally not a perfect indicator for specific information such as lithology or fluid saturation due to technology constraints. Each attribute commonly exhibits a combination of geological characteristics that could lead to subjective interpretations and provides only qualitative results. Meanwhile, machine learning (ML) is emerging as an independent interpreter to synthesise all parameters simultaneously, mitigate the uncertainty of biased cut-off, and objectively classify lithofacies on the accuracy scale.

In this paper, multiple classification algorithms including support vector machine (SVM), random forest (RF), decision tree (DT), K-nearest neighbours (KNN), logistic regression, Gaussian, Bernoulli, multinomial Naïve Bayes, and linear discriminant analysis were executed on the seismic attributes for lithofacies prediction. Initially, all data points of five seismic attributes of acoustic impedance, Lambda-Rho, Mu-Rho, density (ρ), and compressional wave to shear wave velocity (VpVs) within 25-metre radius and 25-metre interval offset top and base of reservoir were orbitally extracted on 4 wells to create the datasets. Cross-validation and grid search were also implemented on the best four algorithms to optimise the hyper-parameters for each algorithm and avoid overfitting during training. Finally, confusion matrix and accuracy scores were exploited to determine the ultimate model for discrete lithofacies prediction. The machine learning models were applied to predict lithofacies for a complex reservoir in an area of 163 km^2.

From the perspective of classification, the random forest method achieved the highest accuracy score of 0.907 compared to support vector machine (0.896), K-nearest neighbours (0.895), and decision tree (0.892). At well locations, the correlation factor was excellent with 0.88 for random forest results versus sand thickness. In terms of sand and shale distribution, the machine learning outputs demonstrated geologically reasonable results, even in undrilled regions and reservoir boundary areas.

Key words: Lithofacies classification, reservoir characterisation, seismic attributes, supervised machine learning, Nam Con Son basin.

1. Introduction

Sand30 is a major gas - condensate reservoir in Hai Thach field. This reservoir has one exploration well and three production wells with very different production performance [1]. Many studies have been conducted to better understand, characterise and model Sand30 [1 - 4]. Reservoir extent and lithofacies distribution are the main focus of the current study.

Date of receipt: 15/5/2022. Date of review and editing: 15/5 - 23/6/2022.
Date of approval: 27/6/2022.

Machine learning has been shown to be capable of complementing and elevating human analysis by objectively examining input data and automatically repeating the calculation until the best output is determined. Because of this benefit, machine learning has been widely used in recent years in the oil and gas business, such as for lithofacies classification [5 - 7], depositional facies prediction [8, 9], well log correlation [10, 11], seismic facies classification [12, 13], and seismic facies analysis [14].

In this study, supervised machine learning was used to predict lithofacies using classification techniques in-

cluding decision tree, support vector machine, and random forest, etc. There are five steps in the overall workflow for this investigation, as shown in Figure 1. First, all seismic data from 5 inversion cubes, including acoustic impedance (AI), Lambda-Rho (LR), Mu-Rho (MR), density, and compressional wave to shear wave velocity ratio (VpVs), were recovered from within 25 m of 4 drilled holes. They were also classified into two groups based on well log data: reservoir and non-reservoir. To ensure that data was labelled correctly, seismic well ties were meticulously conducted. Second, those seismic data were thoroughly examined in order to determine whether or not they were related to facies data. Only seismic data with a good correlation with facies was employed as a training dataset for machine learning. Third, the supervised machine learning was used to determine the best models from the data. Fourth, those models were applied to predict lithofacies for the whole reservoir. Finally, the anticipated facies were retrieved from the map or raw data and compared to the well or present inversion seismic data to assess their quality and reliability.

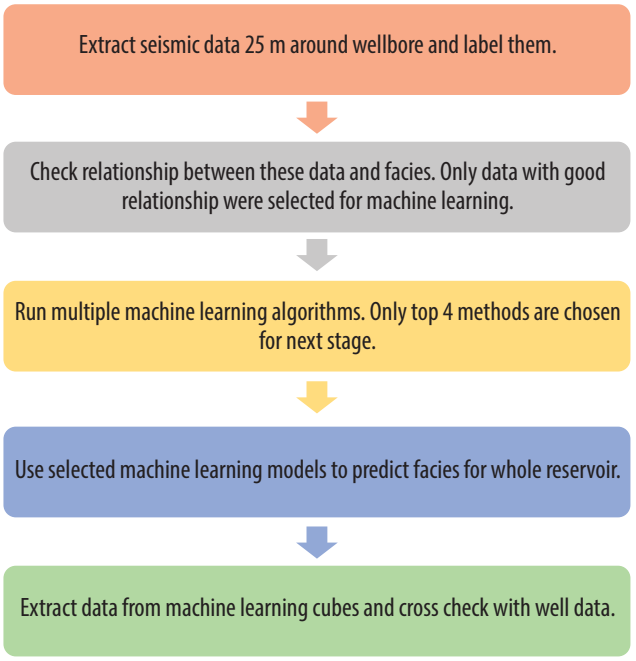

Figure 1. *Overall workflow.*

2. Data generation and visualisation

The input data included available well logs from four drilled holes and five seismic inversion cubes. Well logs included gamma ray, interpreted facies logs used for zonation and facies classification, density and sonic used for seismic well tie. All well data were carefully checked before making the seismic well tie. The purpose of this step was to ensure that all the seismic data and well logs were consistent, as shown in Figure 2.

Five seismic inversion cubes were then exported using orbital extraction (Figure 3) with radius of 25 m, which corresponds to the minimum seismic bin size and therefore the best input for obtaining the most reasonable correlation between well log data and seismic data. Because

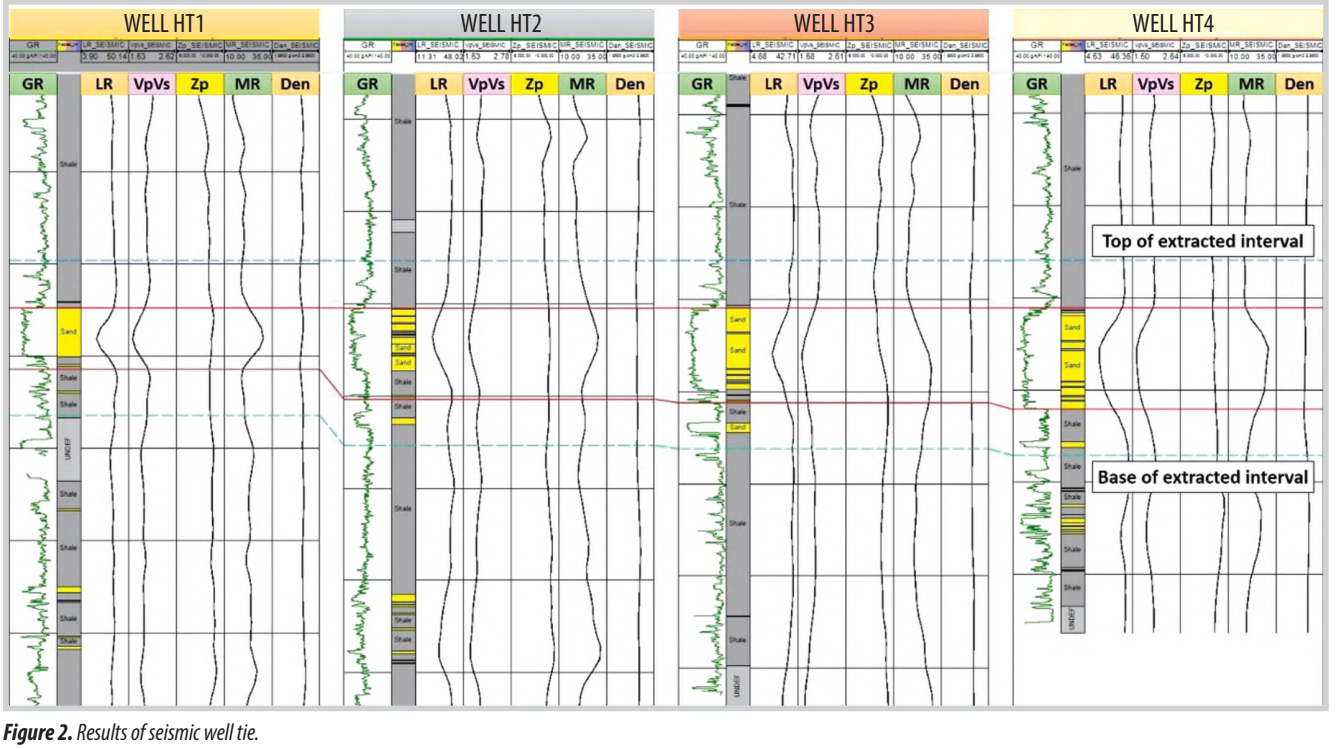

Figure 2. *Results of seismic well tie.*

the extraction takes the average of nearby grid values, the extraction radius should not be less than the minimum bin size in order to avoid skipping the surrounding wellbore information. On the other hand, the depth of investigation of well logging tools is very close to the wellbore wall, only a few centimetres to metres beyond the wall; thus, the smaller the extraction radius, the better the correlation. Some trials with extraction radius larger than 25 m were also carried out; however, the achieved correlation was degraded. The studied interval included reservoir interval and 25 m above the top and below the base of reservoir (half of average reservoir thickness of 50 m) which is considered the best representative for facies ratio of reservoir/non-reservoir samples. Before being used for machine learning, these data were conditioned and tagged with facies (reservoir and non-reservoir) using the seismic well tie results (Figure 2). The extracted dataset comprised of a total of 5,515 valid samples, and reservoir to non-reservoir facies ratio was approximately 3:4.

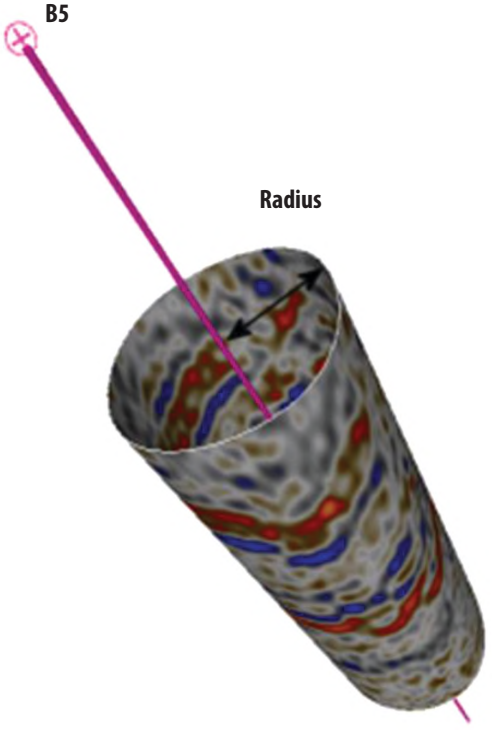

Figure 3. *Orbital extraction.*

Density curve histograms and heat map were used to determine which qualities were the most related to facies. The best markers for facies indication in this study were Lambda-Rho, VpVs, and Mu-Rho. There was relatively clear separation between reservoir and non-reservoir facies in those curves but not for acoustic impedance (Zp) and density (Den) (Figure 4). Similarly, the heat map results which showed correlation between seismic properties and facies also revealed the same conclusion by correlation factor (0.7 for Lambda-Rho and VpVs, and 0.47 for Mu-Rho) (Figure 5). For those reasons, only 3 properties Lambda-Rho, VpVs and Mu-Rho were used as inputs for machine learning in the next step.

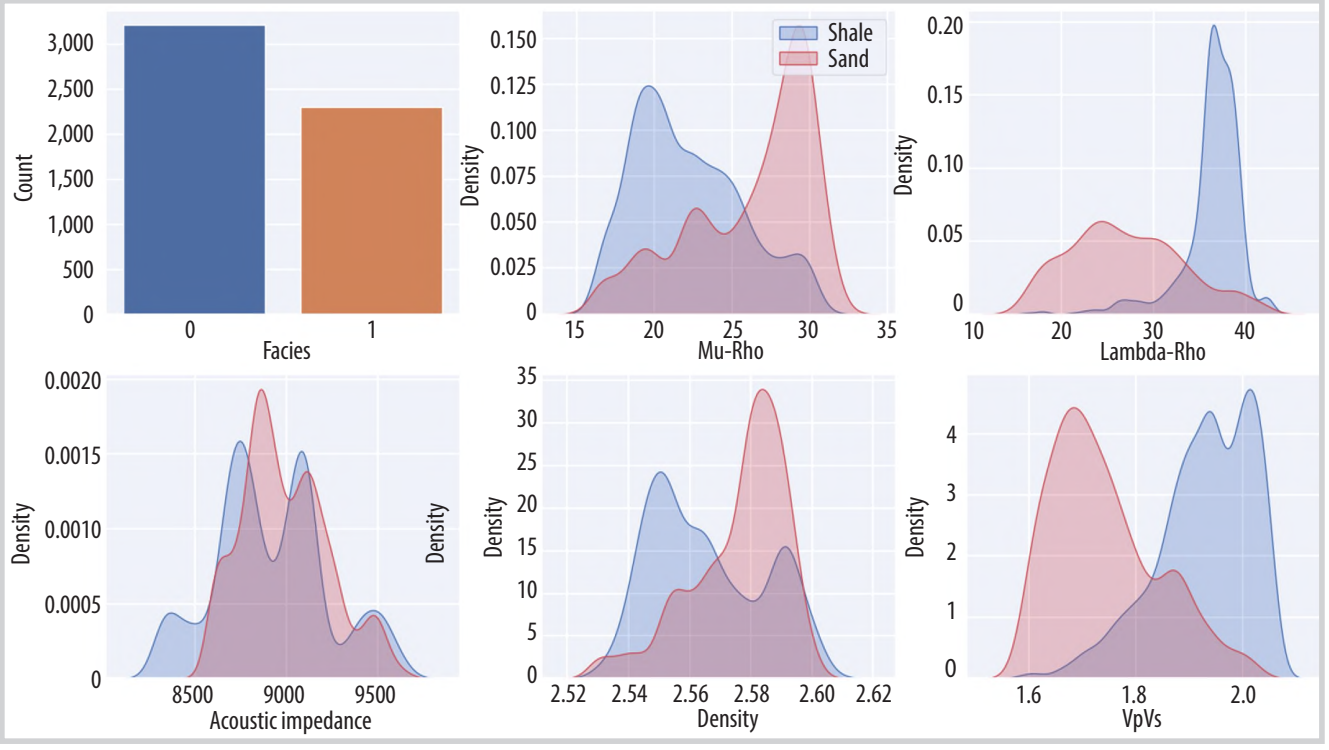

Figure 4. *Density curve histogram for seismic attributes.*

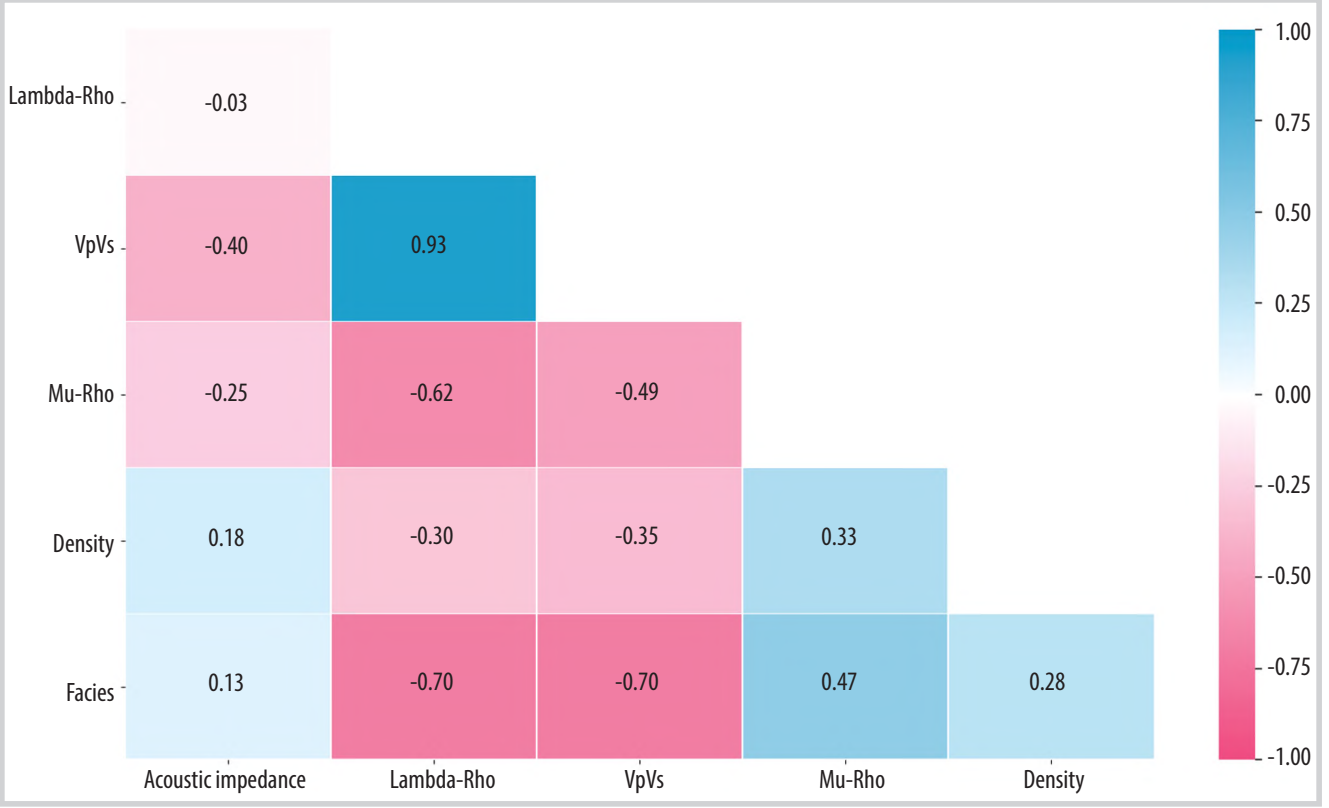

Figure 5. *Heat map for 5 seismic properties versus facies.*

Table 1. *Accuracy score of facies prediction*

Method	Accuracy on training set	Accuracy on test set
K-nearest neighbours	0.94	0.92
Decision tree classifier	1.00	0.90
Support vector machine	0.90	0.90
Random forest	0.88	0.87
Logistic regression classifier	0.87	0.86
Bernoulli classifier	0.87	0.86
Linear discriminant analysis	0.87	0.86
Gaussian Naïve Bayes	0.86	0.86

3. Machine learning approach

True positive (TP), true negative (TN), false positive (FP), and false negative (FN) are the four categories of prediction outcomes used in this study. True negative denotes that models correctly predict non-reservoir facies, while true positive says that reservoir facies are accurately predicted. On the other hand, there are two kinds of errors that could be encountered: false positive and false negative. False positive means facies that are predicted to be reservoirs but are actually non-reservoirs, whereas false negative represents facies that are predicted to be non-reservoirs but are actually reservoirs. Both error types reduce model accuracy, but in terms of HIIP calculation, the false positive type error is more severe than the false negative type because it can result in an overestimation of reservoir facies, which is the main contributor to HIIP. As a result, low false positive error is one of the most important factors for model selection. The following formula was used to compute the accuracy score:

$$Accuracy\ score=(True\ positive+True\ negative)/Total$$

At the beginning of the study, many supervised classification algorithms were investigated, including logistic regression, Gaussian Naïve Bayes, Bernoulli Naïve Bayes, multinomial Naïve Bayes, linear discriminant analysis, support vector machine, K-nearest neighbours, decision tree, and random forest, as shown in Table 1, to find the best four algorithms based on the accuracy score for latter stage.

At the second stage, only the top four algorithms were selected to build the model. At this stage, cross

validation and GridSearchCv technique were used to optimise hyper-parameters and avoid overfitting.

For cross validation, the test data would be kept separate and reserved for the final evaluation step to check the "reaction" of the model when encountering completely unseen data. The training data would be randomly divided into K parts (K is an integer, usually either 5 or 10). The model would be trained K times, each time one part would be chosen as validation data and K-1 parts as training data. The final model evaluation results would be the average of the evaluation results of K training times. With cross validation, the evaluation is more objective and precise.

In addition, one of the important things about machine learning is optimising parameters, called hyper parameters, which cannot be learned directly. Each model can have many hyper parameters and finding the best combination of parameters can be considered a search problem. In this study, GridSearchCv was used to find the optimal combination.

4. Machine learning results and validation

The average accuracy score of K training times is listed in Table 2. Random forest achieved the highest score, followed by support vector machine, K-nearest neighbours, and decision tree.

Similarly, the confusion matrix report system was also used in this study to evaluate the performance of each model. The confusion matrix is as follows:

$$\begin{bmatrix} True\ negative & False\ positive \\ False\ negative & True\ positive \end{bmatrix}$$

According to the confusion matrix, random forest had

Table 2. *Average accuracy score*

Machine learning algorithm	Average accuracy score
Random forest	0.907
Support vector machine	0.896
K-nearest neighbours	0.895
Decision tree	0.892

Table 3. *Confusion matrix*

Method	Confusion matrix for test set
Random forest	$\begin{bmatrix} 593 & 43 \\ 53 & 414 \end{bmatrix}$
K-nearest neighbours	$\begin{bmatrix} 588 & 48 \\ 60 & 407 \end{bmatrix}$
Support vector machine	$\begin{bmatrix} 593 & 43 \\ 76 & 391 \end{bmatrix}$
Decision tree	$\begin{bmatrix} 585 & 51 \\ 73 & 394 \end{bmatrix}$

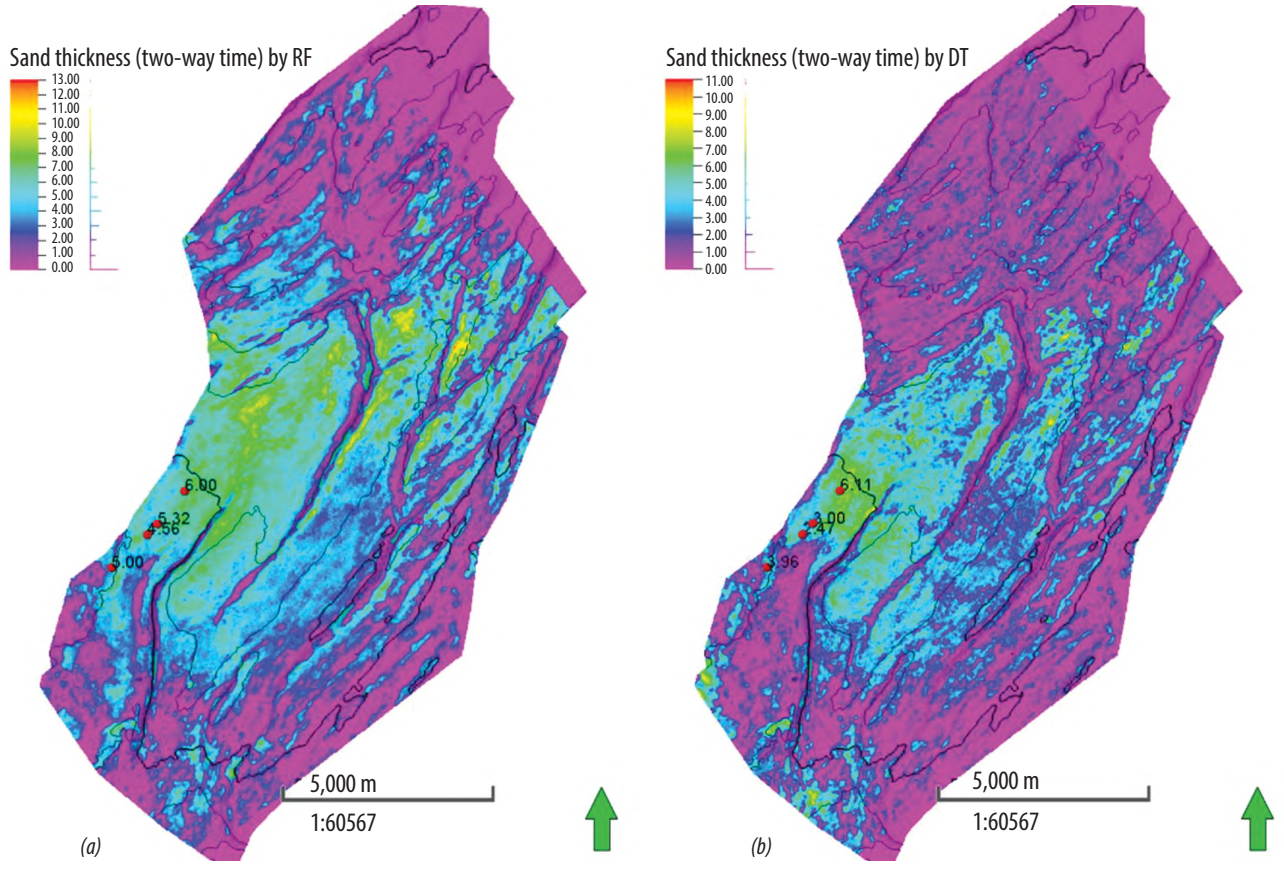

Figure 6. *Sand thickness (two-way time) map by random forest (a) and decision tree (b).*

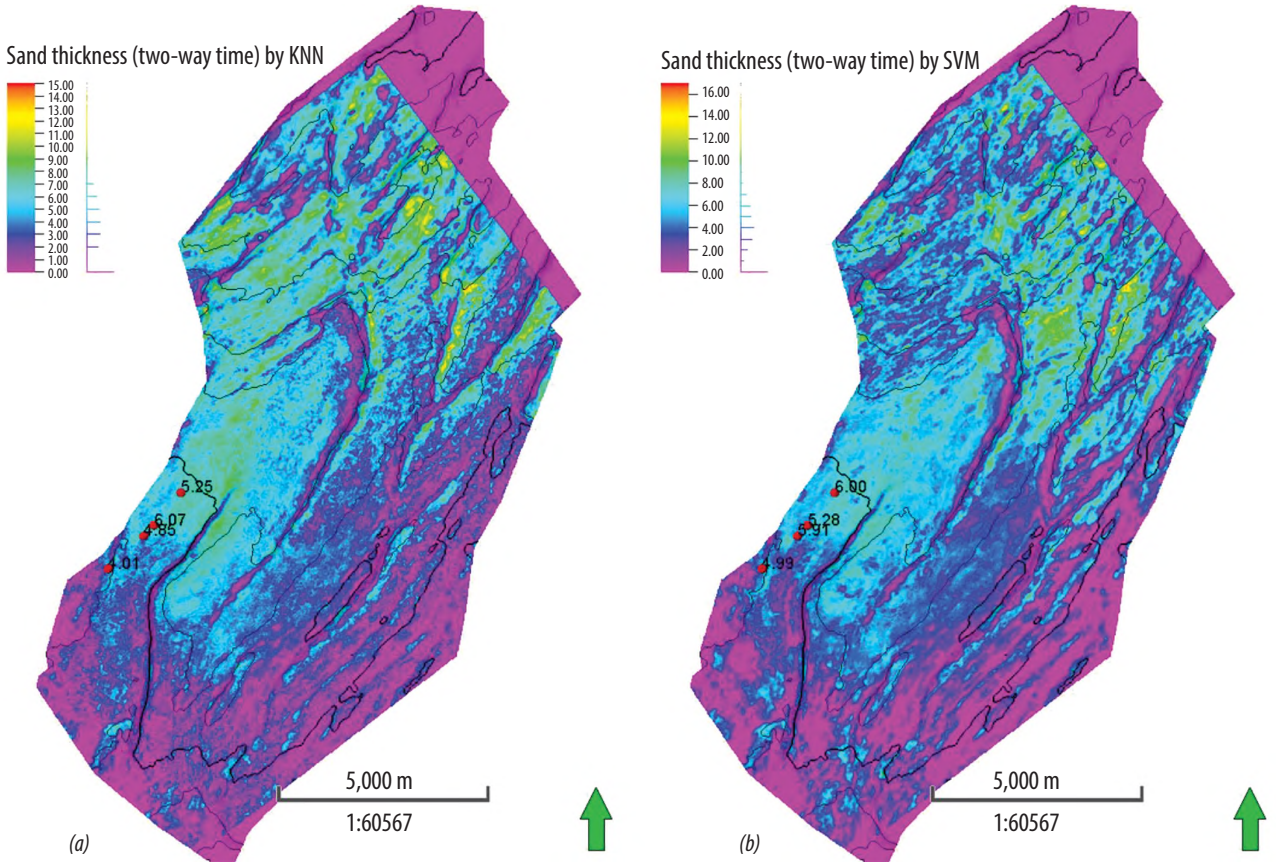

Figure 7. *Sand thickness (two-way time) map by K-nearest neighbours (a) and support vector machine (b).*

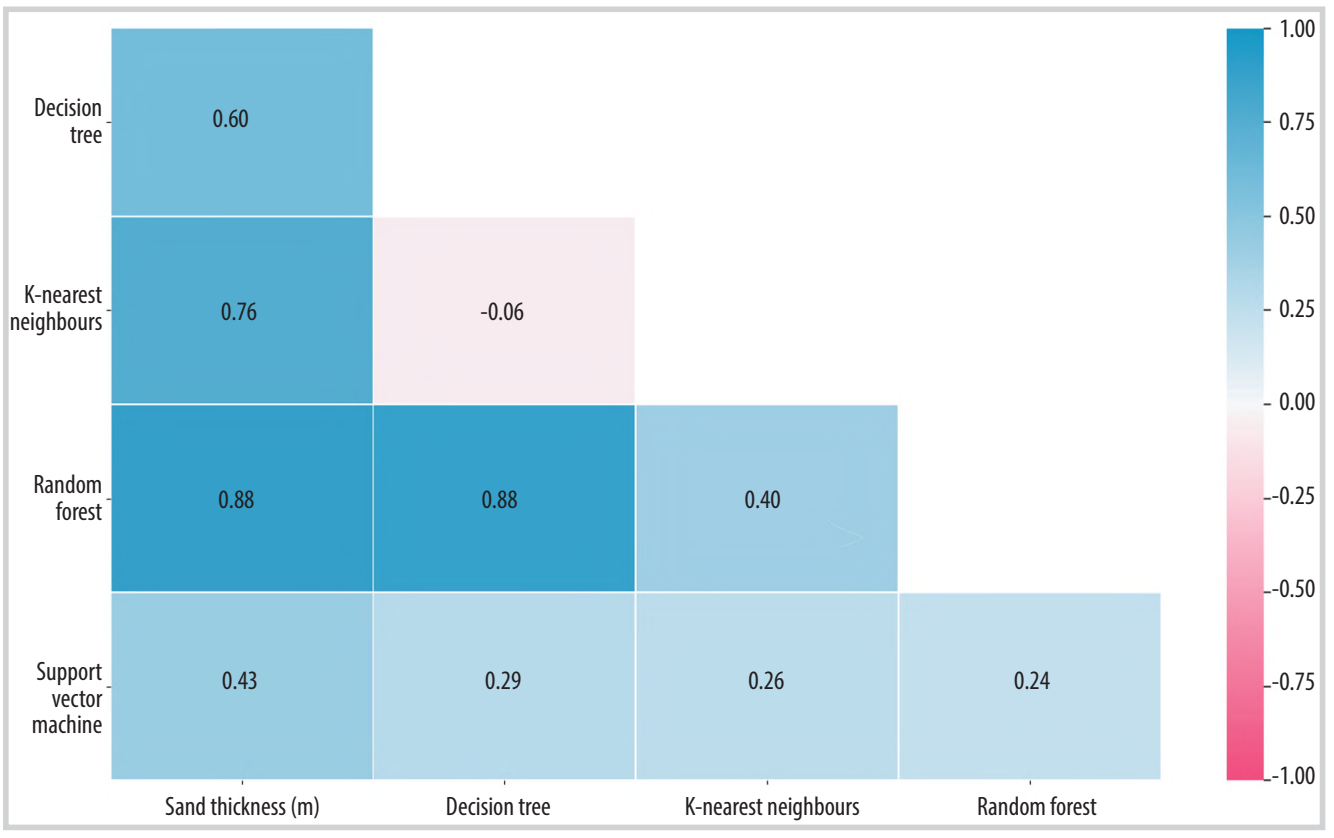

Figure 8. *Correlation between machine learning cubes versus sand thickness at well location.*

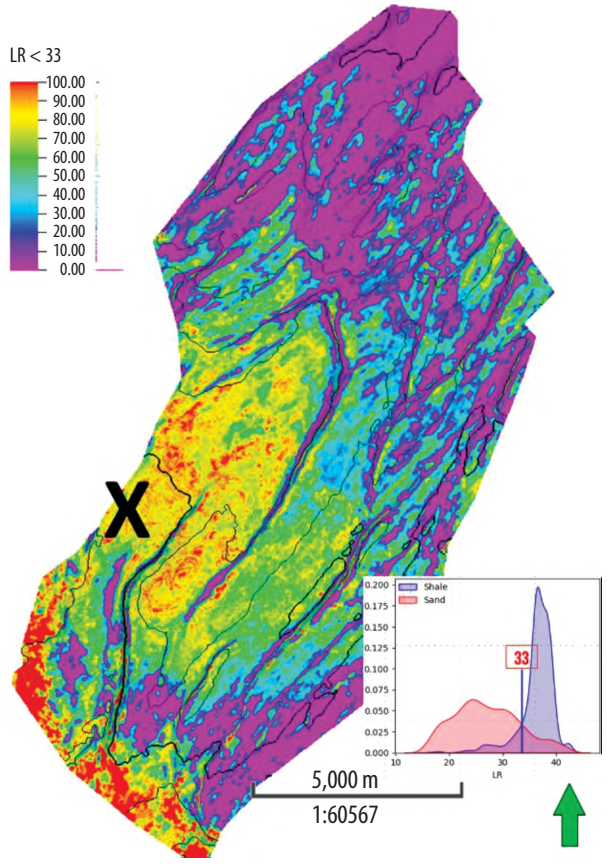

Figure 9. *Lambda-Rho attribute with threshold below 33 (as defined by seismic histogram).*

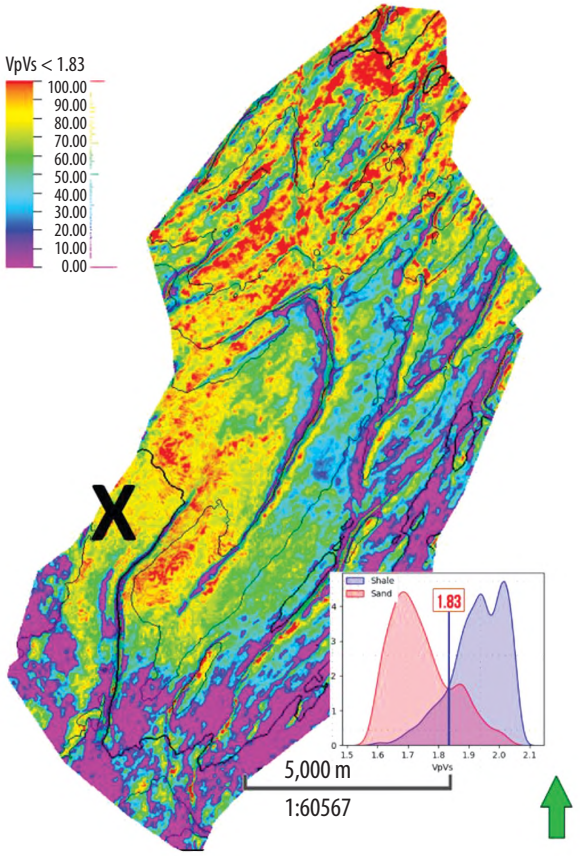

Figure 10. *VpVs attribute with threshold below 1.83 (as defined by seismic histogram).*

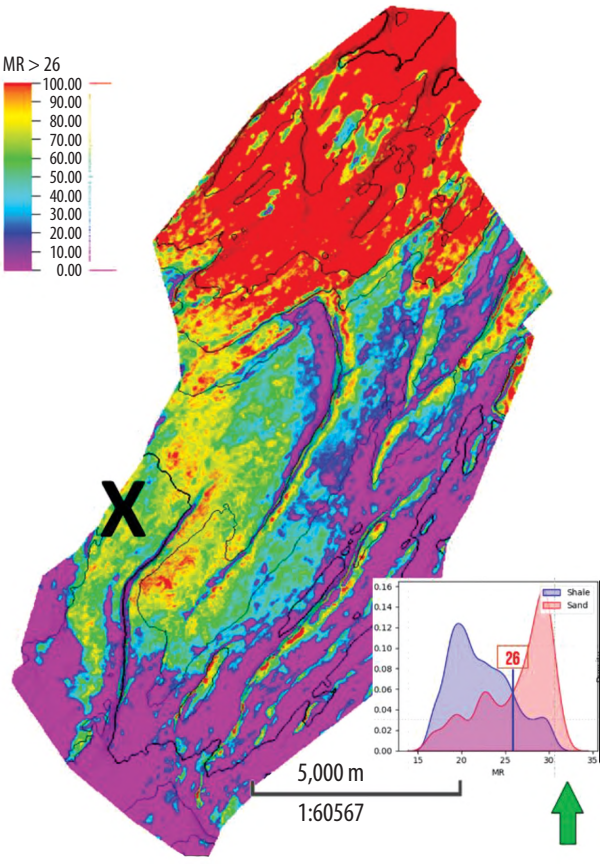

Figure 11. *Mu-Rho attribute with threshold above 26 (as defined by seismic histogram).*

the lowest total false prediction (false positive + false negative) results (96 errors), followed by K-nearest neighbours (108 errors), support vector machine (119 errors), and decision tree (124 errors). Regarding, false positive, the most serious errors, random forest had the fewest number of errors (43 errors) and decision tree had the highest (51 errors).

Properties and maps from four machine learning cubes (Figures 6 and 7) were also extracted at well locations to determine the relationship between actual well sand thickness and reservoir thickness from machine learning using a heat map based on Pandas correlation function (Figure 8). The correlation between well data and random forest cube was the highest (0.88) on the heat map, followed by K-nearest neighbours (0.76), decision tree (0.60), and support vector machine (0.43). It is likely that the random forest algorithm is the most dependable approach for this investigation.

5. Discussions and application

Attribute maps, which may be utilised as guidelines for property populations in 3D model, are one of the most notable contributions of seismic data. Normally, single

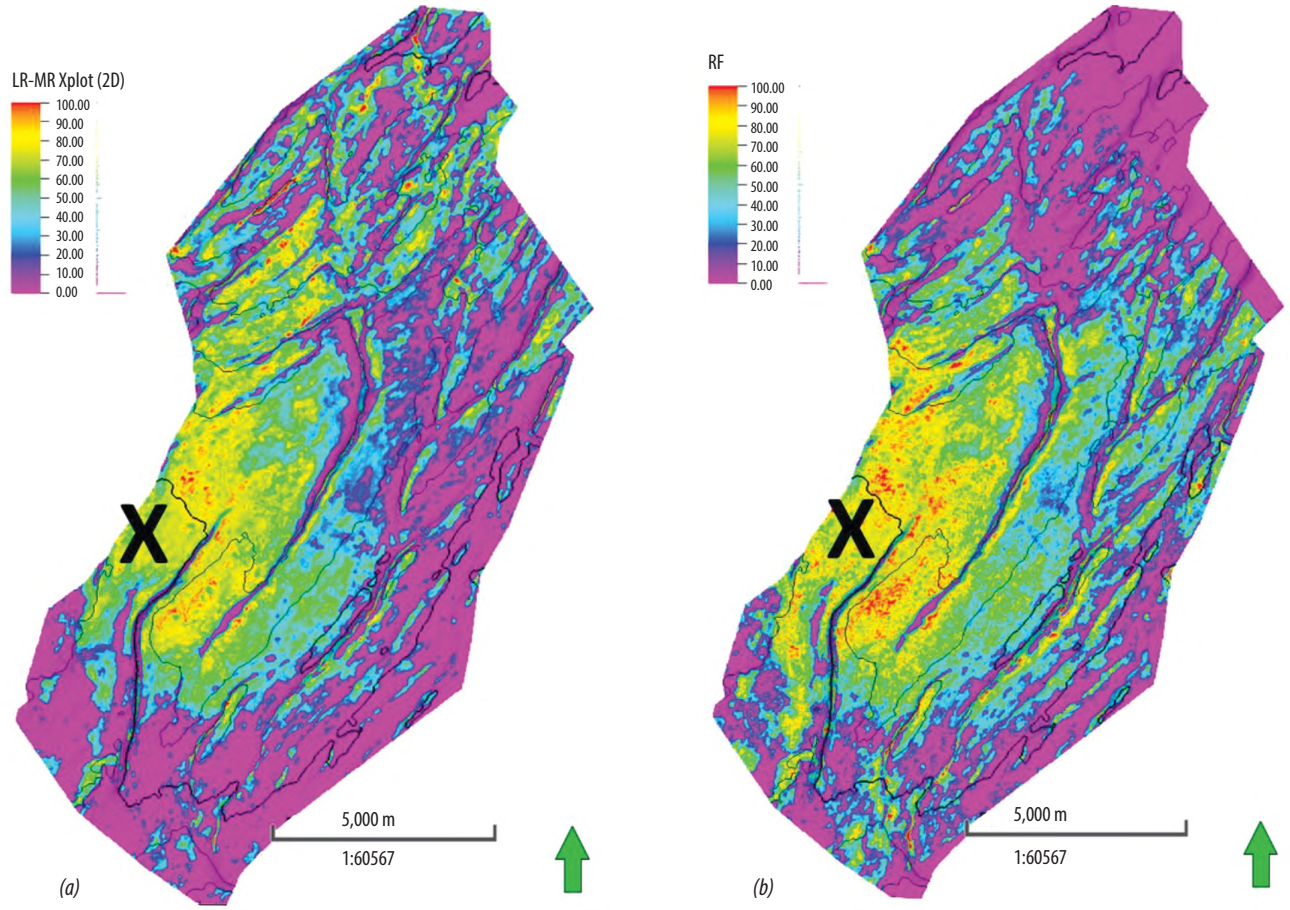

Figure 12. *Lambda-Rho - Mu-Rho cross-plot attribute (a) and random forest results (b).*

seismic attributes may give reasonable results around the drilled well areas but questionable for far away areas. For example, in our location, Lambda-Rho cut-off attributes (Figure 9) showed good results in the drilled area (X area), but lots of non-geological anomalies in far away areas, especially in the southern area. VpVs and Mu-Rho also had similar performance (Figures 10 and 11). Consequently, selecting the best attribute for further study in this case is very challenging and risky. Therefore, in our location, the combination between Lambda-Rho and Mu-Rho was used to minimise the potential risks (Figure 12a). However, this procedure itself requires high experience from the interpreters so the results seem to be very subjective. Furthermore, only Mu-Rho and Lambda-Rho were used in this combination while VpVs was not even though it could be very valuable in terms of geological meaning.

With machine learning workflow, the number of integrated attributes can be more flexible, as long as data are correlated with each other. There are no subjective parameters used such as threshold cut-offs which seem to be very sensitive. On top of that, results from machine learning is very promising and reliable, for example, most of non-geological anomalies in the southern part (as shown on Lambda-Rho map), and northern part (as shown on Mu-Rho and VpVs maps) were not present in random forest map and the results in the drilled area (X) are still of high quality (Figure 12b).

6. Conclusions

The main conclusions of this study can be summarised as follows:

- The traditional approach of using single seismic attribute such as Lambda-Rho, VpVs, or Mu-Rho for facies prediction leads to potential risks especially for remote areas without wells. Moreover, it highly depends on the experience of interpreters in selecting cut-off parameters;

The more advanced approach of combining seismic attributes can improve prediction accuracy but highly depends on the experience of interpreters and sometimes cannot use all available data;

- Machine learning techniques such as random forest, decision tree, K-nearest neighbours, and support vector machine were used to overcome the disadvantages of

traditional approaches by analysing all input parameters objectively. The study successfully classified facies from each other;

- Random forest was found to be the most dependable method for the study area;

- The results from machine learning are of very high quality and can be used for HIIP calculation and 3D static modelling.

Acknowledgment

The research work described herein was part of Research Project 077.2021.CNKK.QG/HDKHCN, Order 196/QD-BCT of the Vietnam Ministry of Industry and Trade.

References

[1] Phạm Hoàng Duy, Hoàng Kỳ Sơn, Trần Ngọc Thế Hùng, và Trần Vũ Tùng, "Kết quả đo độ thấm bằng nhiều phương pháp khác nhau cho vỉa turbidite mỏ Hải Thạch, bể Nam Côn Sơn", *Tạp chí Dầu khí*, Số 2, trang 35 - 44, 2019.

[2] Pham Hoang Duy, Hoang Ky Son, Trinh Xuan Vinh, and Tran Vu Tung, "Condensate banking characterization and quantification of improvement from different mitigations using pressure transient analysis: A case study in Hai Thach field offshore Vietnam", *Offshore Technology Conference Asia, Kuala Lumpur, Malaysia, 2 - 6 November 2020*. DOI: 10.4043/30142-MS.

[3] Hoang Ky Son, Tran Vu Tung, Nguyen Ngoc Tan, Truong Tu Anh, Pham Hoang Duy, Tran Ngoc Trung, Trinh Xuan Vinh, and Ngo Tuan Anh, "Successful application of machine learning to improve dynamic modeling and history matching for complex gas-condensate reservoirs in Hai Thach field, Nam Con Son basin, offshore Vietnam", *SPE Symposium: Artificial intelligence - Towards a Resilient and Efficient Energy Industry*, 18 - 19 October 2021. DOI: 10.2118/208657-MS.

[4] Hoang Ky Son, Tran Vu Tung, Nguyen Ngoc Tan, Truong Anh Tu, Pham Hoang Duy, Tran Ngoc Trung, Trinh Xuan Vinh, and Ngo Tuan Anh, "Successful case study of machine learning application to streamline and improve history matching process for complex gas-condensate reservoirs in Hai Thach field, offshore Vietnam", *SPE Middle East Oil & Gas Show and Conference*, 29 November 2021. DOI: 10.2118/204835-MS.

[5] Leila Aliouane, Sid-Ali Ouadfeul, Nouredine Djarfour, and Amar Boudella, "Lithofacies prediction from well logs data using different neural network models", *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods (PRG-2013)*, 2013. DOI: 10.5220/0004380707020706.

[6] Paolo Bestagini, Vincenzo Lipari, and Stefano Tubaro, "A machine learning approach to facies classification using well logs", *SEG Technical Program Expanded Abstracts 2017*. DOI: 10.1190/segam2017-17729805.1.

[7] Jing Jing Liu and Jian Chao Liu, "Integrating deep learning and logging data analytics for lithofacies classification and 3D modeling of tight sandstone reservoirs", *Geoscience Frontiers*, Vol. 17, No. 1, 2022. DOI: 10.1016/j.gsf.2021.101311.

[8] Randall S. Miller, Skip Rhodes, Deepak Khosla, and Fernando Nino, "Application of artificial intelligence for depositional facies recognition - Permian basin", *Unconventional Resources Technology Conference, Denver, Colorado, USA, 22 - 24 July 2019*. DOI: 10.15530/urtec-2019-193.

[9] Tran Vu Tung, Ngo Huu Hai, Hoang Ky Son, Tran Ngoc The Hung, and Joseph J. Lambiase, "Depositional facies prediction using artificial intelligence to improve reservoir characterization in a mature field of Nam Con Son basin, offshore Vietnam", *Offshore Technology Conference Asia, Kuala Lumpur, Malaysia, 2 - 6 November 2020*. DOI: 10.4043/30086-MS.

[10] Hiren Maniar, Srikanth Ryali, Mandar S. Kulkarni, and Aria Abubakar, "Machine learning methods in geoscience", *SEG Technical Program Expanded Abstracts 2018*. DOI: 10.1190/segam2018-2997218.1.

[11] Seth Brazell, Alex Bayeh, Michael Ashby, and Darrin Burton, "A machine-learning-based approach to assistive well-log correlation", *Petrophysics*, Vol. 60, No. 4, pp. 469 - 479, 2019. DOI: 10.30632/PJV60N4-2019a1.

[12] Satinder Chopra and Kurt J. Marfurt, "Seismic facies characterization using some unsupervised machine learning methods", *SEG Technical Program Expanded Abstracts 2018*. DOI: 10.1190/segam2018-2997356.1.

[13] Vladimir Puzyrev and Chris Elders, "Unsupervised seismic facies classification using deep convolutional autoencoder", *Geophysics*, Vol. 87, No. 4, 2022. DOI: 10.1190/geo2021-0016.1.

[14] Thilo Wrona, Indranil Pan, Robert L. Gawthorpe, and Haakon Fossen, "Seismic facies analysis using machine learning", *Geophysics*, Vol. 83, No. 5, 2018. DOI: 10.1190/geo2017-0595.1.