Journal of Petroleum Science and Engineering 208 (2022) 109250

Contents lists available at ScienceDirect

Journal of Petroleum Science and Engineering

journal homepage: www.elsevier.com/locate/petrol

Performance evaluation of machine learning-based classification with rock-physics analysis of geological lithofacies in Tarakan Basin, Indonesia

Gian Antariksa [a], Radhi Muammar [b], Jihwan Lee [a,*]

[a] Department of Industrial and Data Engineering, Major in Industrial Data Science and Engineering, Pukyong National University, Busan, 48513, Republic of Korea
[b] Conrad Petroleum, Jakarta, 12430, Indonesia

ARTICLE INFO

Keywords:
Lithofacies
Rock-physics
Classification
Evaluation performance
Machine learning
Tarakan basin

ABSTRACT

This study aims to put a supervised learning method for automatically classifying lithofacies in well-logging dataset, where several machine learning algorithms were compared in this study that took place in the Tarakan Basin, Indonesia. The predicted lithofacies in this study including shale, shaly sandstone, sandstone, and coal, where coal is considered as the unique lithofacies in the study area. As training and testing datasets, we used two separate well log datasets from the Tarakan Basin. The first well, named Omnicron, was used to train the model, while the second well, named Kay, was used to test it. Random Forest and Gradient Boosting outperformed the other models in the experiment, with the accuracy of 87.49% and 87.01%, respectively. When it came to classifying coal, however, both approaches had issues. The Pr-Recall curve revealed that the coal score was under average precision in each facies, with values of roughly 0.52 and 0.38, respectively, which explaining why, even with high accuracy, the machine learning algorithm predicted poorly in one lithofacies class. In order to evaluate this coal misclassification, we used rock physics to analyze the machine learning prediction in this report. As result, we found that each facies is well-differentiated by physical properties, and the predicted lithofacies have a distribution that is close to the original facies however, coal may be potentially misclassified as other lithofacies as some of the coals have similar rock physical properties with the surrounding lithology (e.g. coal with a mixture of shale may have similar DT and GR responses). Based on this research, the use of machine learning in the Tarakan Basin effectively provides lithofacies data with a high degree of precision and accuracy in a much shorter time.

1. Introduction

The physical characteristics of geologic formations in the subsurface are represented by well log data. The formation's lithological composition, texture, and post-depositional processes all have an effect on these characteristics (e.g. diagenesis and fracturing). These physical properties can be used to calculate mineral volume, porosity, permeability, and fluid saturation, all of which were used to assess lithofacies. Such tasks formed the foundation for describing and comprehending reservoir architecture. The determination of lithofacies necessitates manual interpretation, which is usually time-consuming and labor-intensive. As a result, an effective method and computational tool for automatically identifying lithofacies by using known physical properties at a given area are required to provide an accurate and precise lithofacies result in a much shorter time, and substantially reduce the time to make confident decision-making. The use of applications that are capable to make more reasoned decisions than conventional interpretation methods will improve the efficiency of lithological classification (Hsieh et al., 2005; Jahdhami and Anboori, 2017).

In the petroleum industry, many machine learning-based techniques have recently emerged to increase the efficiency and accuracy of lithofacies classification. For example, in some studies, well log data was used in the subsurface analysis (Qi and Carr, 2006; Wang et al., 2014; Bhattacharya et al., 2016). Other studies used various machine learning models to classify the facies using well log data, including (a) random forest (Bressan et al., 2020; Adoghe et al., 2011; Al-Mudhafar, 2017; Male and Duncan, 2020) and (b) artificial neural networks (Tang et al., 2004; Baldwin et al., 1990; Rogers et al., 1992; Wong et al., 1995; Avseth and Mukerji, 2002; Dubois et al., 2007; Wood, 2019). (c) support vector machine (Al-Mudhafar, 2015, 2017; Ameur-Zaimeche et al., 2020; Bhattacharya and Carr, 2019; Bhattacharya and Mishra, 2018; Bressan et al., 2020) (d) density-based K-Nearest Neighbors (Dubois

* Corresponding author.
E-mail address: jihwan@pknu.ac.kr (J. Lee).

https://doi.org/10.1016/j.petrol.2021.109250
Received 24 December 2020; Received in revised form 5 July 2021; Accepted 14 July 2021
Available online 24 July 2021

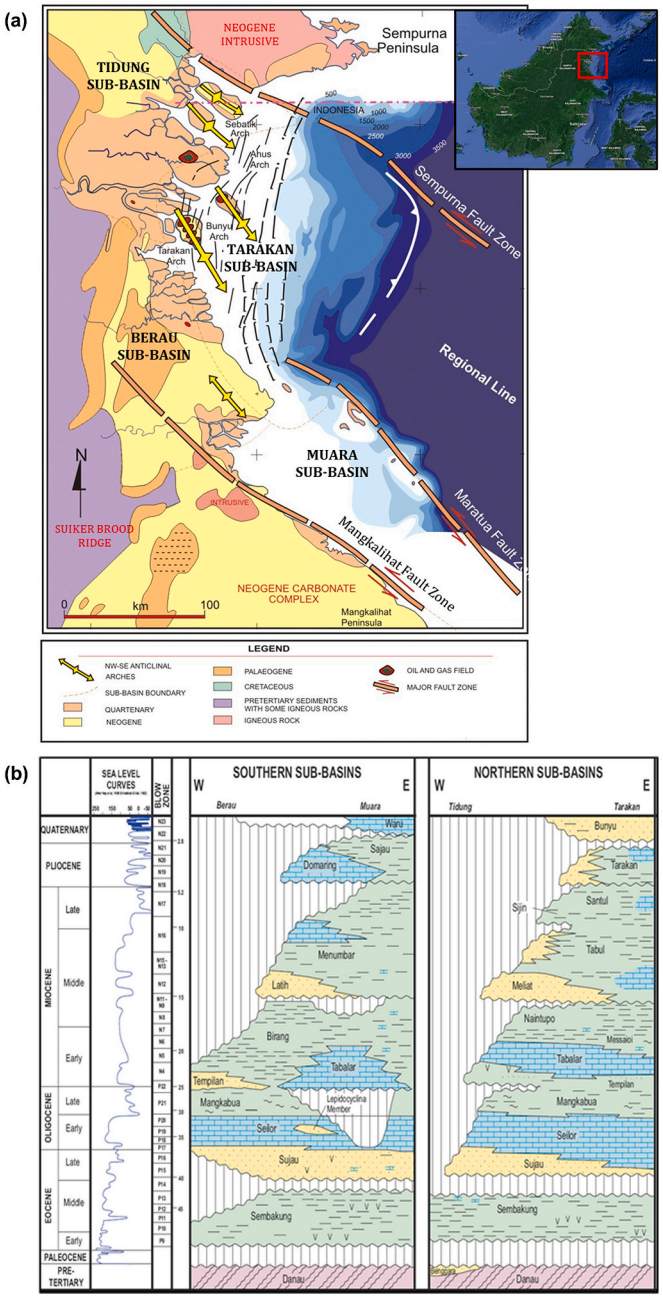


Fig. 1. (a) Structural elements map of the Tarakan Basin and (b) Regional Stratigraphic column of the Tarakan Basin (Lentini and Darman, 1996).

et al., 2007), (e) scaling algorithm for calibration process was developed in order to improve uncertainty analysis, on other hand, imputation of missing well log data based on machine learning were developed as trending issues (Feng, 2021; Feng et al., 2021).

This study aims to apply the machine learning-based approach to lithofacies classification using well log data from the Tarakan Basin in North Kalimantan, which is one of Indonesia's hydrocarbon-producing basins. The study of lithofacies analysis in the Tarakan Basin is still limited, and there has been no attempt to apply machine learning in the region to date. Formation heterogeneities in this basin make reliable machine learning results difficult to achieve because each formation has its unique characteristics, such as sedimentology and diagenetic processes, which lead to differences in physical properties. Although Tarakan Basin shows the heterogeneous distribution of lithofacies, this could be a challenge for us to predict lithofacies properly using the proposed machine learning methods. According to its variety, using machine learning implementation such as random forest (Al-Mudhafar, 2017) and multilayer perceptron (Bressan, 2020) could predict facies with high accuracy. The aim of this study is to validate whether the machine learning model which is trained by a well can be used to predict lithofacies of another well. Although both wells belong to Tarakan Basin, their statistical distribution may not be identical; They may be slightly different in their cut-off, and facies properties. To obtain reliable prediction accuracy even in the presence of the difference of the well locations, the size of the data set should be larger than the usual case when the training and testing dataset comes from the same well. Moreover, we also compared other supervised learning methods using a large training well log dataset to predict lithofacies of another well in the same area of Tarakan basin, followed by an analysis of several advanced statistical evaluations and geological interpretations on the validation well to establish the most reliable model for determining lithofacies on other wells in the basin. This model and analysis described in this study can be useful for future exploration activities in this area as this methodology substantially reduced the amount of interpretation time. Besides, while the demonstration of machine learning implementation is the study's main focus, we have also shown each lithofacies properties distribution to better understand the formation's and lithofacies' physical behaviors. Furthermore, the dataset which used for our analysis was very broad and high-resolution relative to other studies. As the size of the dataset increases, the ML-based model becomes resilient to noise and uncertainty, so the test dataset can yield accurate results. For example, Imamverdiyev and Sukhostat (2019), who presented findings of lithological facies classification using deep convolution neural networks, seem to use an approximately smaller dataset size. Another contribution of our research is the ability to predict coal, a recent result in machine-learning comparisons. Bressan et al. (2020) recently demonstrated the variety of implementation and evaluation of machine learning, however, have no additional details on the prediction of coal lithofacies. Predicting coal is important because it has rather distinctive properties relative to other facies. For example, coals are known to have strikingly lower RHOB but higher NPHI and DTC compared to the other lithofacies. This, however, is not always the case for some coal beds in the Tarakan Basin, where the properties of some coal beds are similar with other lithofacies, resulting in frequent perplexing of geoscientists and causing uncertainty when interpreting coal characteristics. A similar issue of coal misinterpretation has been reported by Fu et al. (2009) and Wang et al. (2003). Based on this study, the prediction of coal with all its uncertainties is essential to evaluate the differences between each lithofacies. To validate this discovery void, we set out to pursue new findings using the dataset in the study area.

We used several supervised learning methods for the lithofacies classification at Tarakan Basin using well log data, including decision tree, random forest, gradient boosting, logistic regression, k-nearest neighbors, and support vector machine. On the validation log well results, the best machine learning model had an overall accuracy of 87%, indicating that the trained model can be applied to other wells in the same area. Accuracy, precision, recall, F1-score, ROC curve, and precision-recall curve were used to test the outperformed methods. In contrast to previous studies, which concentrated solely on overall accuracy, an in-depth performance evaluation of each facies class was carried out using a variety of metrics. Surprisingly, the coal's PR-area-under-curve score was well below average when compared to other facies, approximately 0.52 for random forest and 0.38 for gradient boosting. This suggests that, despite the model's high overall accuracy, there is a significant difference in classification performance between lithofacies groups.

In addition, several other analyses were carried out to better understand the lithofacies cluster distribution, including log-facies classification using cross-plot analysis and rock physics analysis to compare the original lithofacies with the prediction result from the machine learning model and try to answer the potential cause of the difference

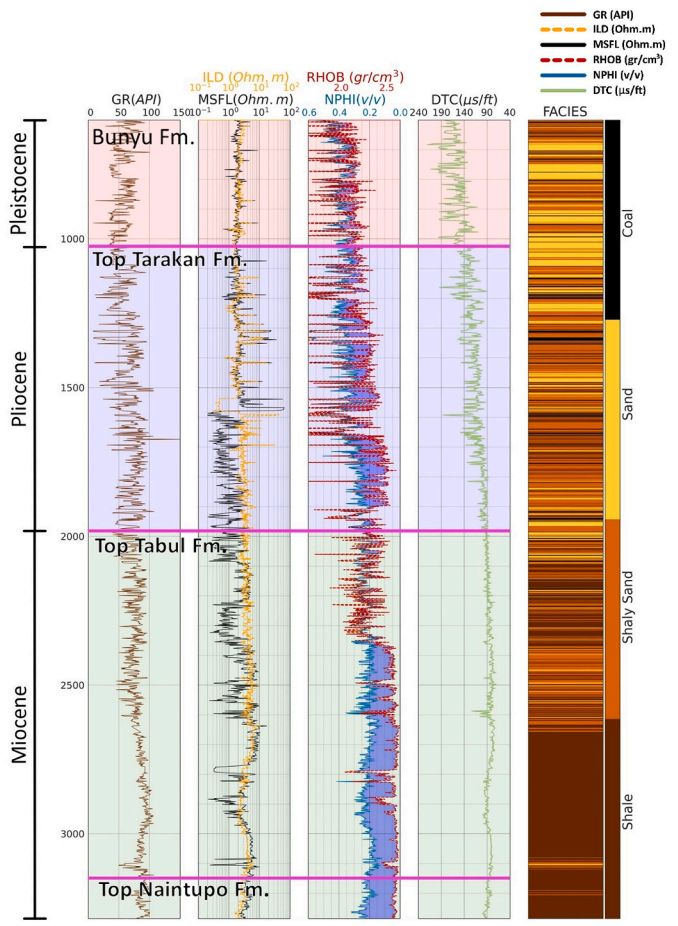


Fig. 2. Omnicron well log visualization (utilized as the training dataset).

between the original. Despite the differences in compaction trend and burial depth between the Pliocene and Miocene intervals, rock physics analysis revealed that the lithofacies predicted by random forest and gradient boosting methods have the best result among others, which also proven by the evaluation method of machine learning.

This is how the rest of the paper is organized. Following brief details on lithofacies, tectonics, and stratigraphy, Section 2 discusses the geological area of study. The Tarakan basin dataset is explained in Section 3 as well as the procedures for the methods used in this experiment. The experiment phase flow of implementing machine learning in the Tarakan basin is depicted in Section 4. In Section 5, detailed findings and a discussion of how the experiment worked and when it was carried out are presented. Finally, Section 6 discusses the study's conclusion.

2. Regional geology

Greater Tarakan Basin (Fig. 1(a)) is comprised of four sub-basins, namely the Tarakan, Muara, Tidung, and Berau Sub-basins. The basin was formed during the Middle Eocene, where the basin experienced an extensional regime possibly associated with the collision of Indian with Eurasian plates and consequently by the opening of the Makassar Strait to the east of the area (Hamilton, 1979).

In this basin, the extensional phase formed a wide basinal area filled by Middle Eocene-Early Oligocene transgressive clastics and carbonates in continental, shallow marine, and deep marine environments under an overall regressive phase. The deposition then continued from Late Oligocene-Early Miocene under the transgressive phase (Burollet and Salle, 1981; Situmorang, 1982, 1983, 1983).

The extensional phase then ceased during Middle Miocene-Pliocene, indicating a more tectonically stable phase. Deltaic sedimentation was initiated during this time with eastward sedimentation (towards the Tarakan sub-basin). The increase of overburden stress as a function of increasing sedimentary thickness induced the gravity-induced listric faults.

Finally, the latest tectonic phase (Pliocene-recent) is characterized by three major wrench faults, namely (from north to south) the Semporna, Maratua, and Mangkalihat Faults (Lentini and Darman, 1996; Baillie et al., 2004; Saputra and Wibisono. 2016) as shown on Fig. 1(a). These faults are transpressional in nature as shown by the positive flower structure formed in the area (Baillie et al., 2004). Fig. 1(a) shows that the Tarakan Basin is bounded to the north by the Sempurna Peninsula that is composed of Mesozoic age metamorphic rocks, volcanic, and ophiolite complex. To the south, this basin is bounded by the Mesozoic aged Mangkalihat Peninsula that is overlain by the Neogene carbonate complex (Achmad and Samuel, 1984; Lentini and Darman, 1996). To the east of Tarakan Basin is the present-day depocenter, where series of overthrust faults and deep-water deposits can be found. Finally, the western of this basin is bounded by the Kalimantan Central Range that is composed of pre-tertiary sediments, igneous, and mélange complex.

This study specifically took place on the onshore part of the Tarakan sub-basin. In this present-day deltaic environment, the Tarakan sub-basin offers quite a challenge to this machine learning study due to the various kinds of lithofacies in the area and formation heterogeneities, including the difference in sedimentology, compaction trend, and diagenetic processes underwent by each formation, leading to a variation in the rock physical properties.

Following the stratigraphic information from Lentini and Darman (1996), the study area is situated at the Miocene to Pliocene intervals of the Tarakan sub-basin (Fig. 1(b)). The explanation for each formation in the study area is as follow (Noon et al., 2003):

1) Naintupo = deposited during Early-Middle Miocene, this formation is composed of shale, marl, and occasional volcaniclastic in a fining upward pattern during a transgressive sequence.
2) Meliat = deposited during Middle Miocene unconformably on top of Naintupo Formation. This formation is composed of shale, sandstone, rare limestone, and rare coals deposited during a shift from predominantly transgressive to regressive system tract. Ellen et al. (2008) suggests that the lower part of Meliat is composed of thick sandstone bodies.
3) Tabul = deposited during Middle Miocene in a predominantly coarsening upward sequence in a tidal flat environment in the lower part and grading to amalgamated distributary channel in the upper part (Akuanbatin et al., 1984; Husein, 2017). This formation is composed of interbedded sandstone, siltstone, claystone/shale, and some limestone stringers at the top.
4) Santul = deposited during Late Miocene, this formation is composed of claystone, sandstone, limestone, and minor coals. Extensive turbidites were deposited during this time as unconfined, toe-of-slope fans of the outbuilding Tarakan deltas (Darman, 2001). This formation is generally absent on the onshore wells.
5) Tarakan = deposited during Pliocene unconformably overlying the older formations. This formation is composed of claystone/shale, sandstone, minor limestone, and coals in a predominantly deltaic environment.
6) Bunyu = this youngest formation was deposited during Pleistocene. Bunyu Formation is comprised of prograding deltaic sediments, including sandstone, claystone, and coal interbeds during a predominantly regressive system tract.

3. Methodology

3.1. Dataset

We used well logs data to train the proposed supervised learning

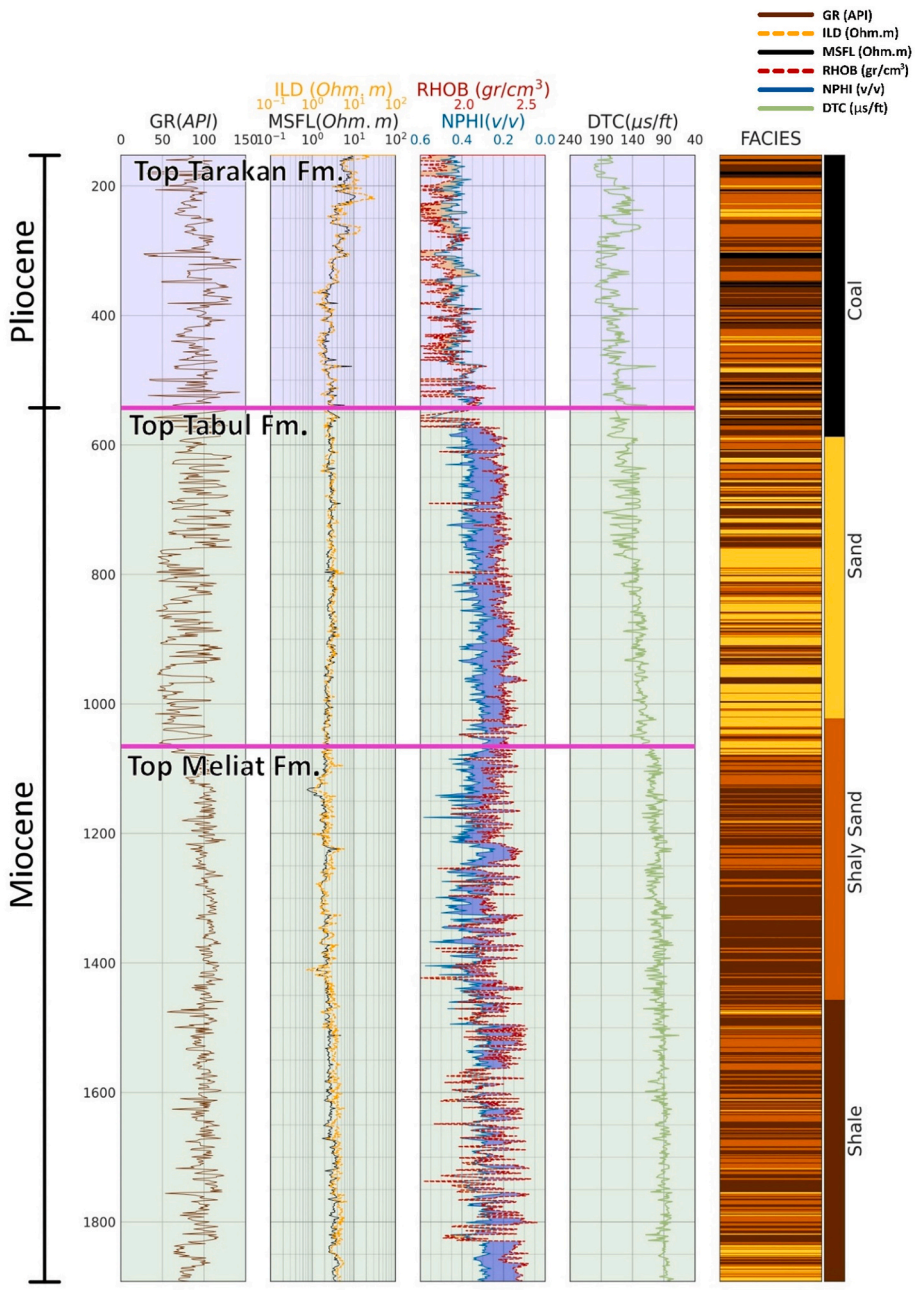


Fig. 3. Kay well log visualization (utilized as the validation dataset).

model to differentiate the lithofacies in this study. The entire dataset was built using data from two wells comprising 31,139 samples. The dataset consisted of well log data from the Omnicron (17,405 samples until a depth of about 2685 m) and Kay wells (13,734 samples until a depth of about 1740 m), which had their original lithofacies interpreted using well logs aided with the mud log and cutting sample data. GR, DTC, ILD, MSFL, NPHI, and RHOB logs were utilized to create the vector function. The vector's characteristics are the measured properties and their transformations. The log response of Omnicron well used for the training model is shown in Fig. 2, along with a plot of the original lithofacies. Fig. 3 depicts the log visualization and original lithofacies of the Kay well, which was used to test the model that had been trained from the previous well. As additional information related to the well logs, the mud log section is also shown in Fig. 4, in order to validate the original lithofacies of the Kay well. Fig. 5(a) and (b) display the distribution of each facies in the Omnicron and Kay wells, respectively. Each facies distribution was then recorded in the Omnicron well dataset Table 1 and the Kay dataset Table 2 for the exact value needed for comparison.

3.2. Machine learning models

3.2.1. Decision tree

A decision tree is a practical, easy, and comprehensive learning approach, according to Maimon and Rokach (2010), this is one of the supervised machine learning models. It is a useful method for discovering previously unknown information by analyzing a large amount of data. The decision tree is constructed by recursively partitioning the feature space of the training set. The aim is to find a set of decision-making rules that divide the feature space predictably and stably to produce a hierarchical classification model that is both informative and stable. Training can be used to define and improve models with important configurations such as the maximum depth of the tree,

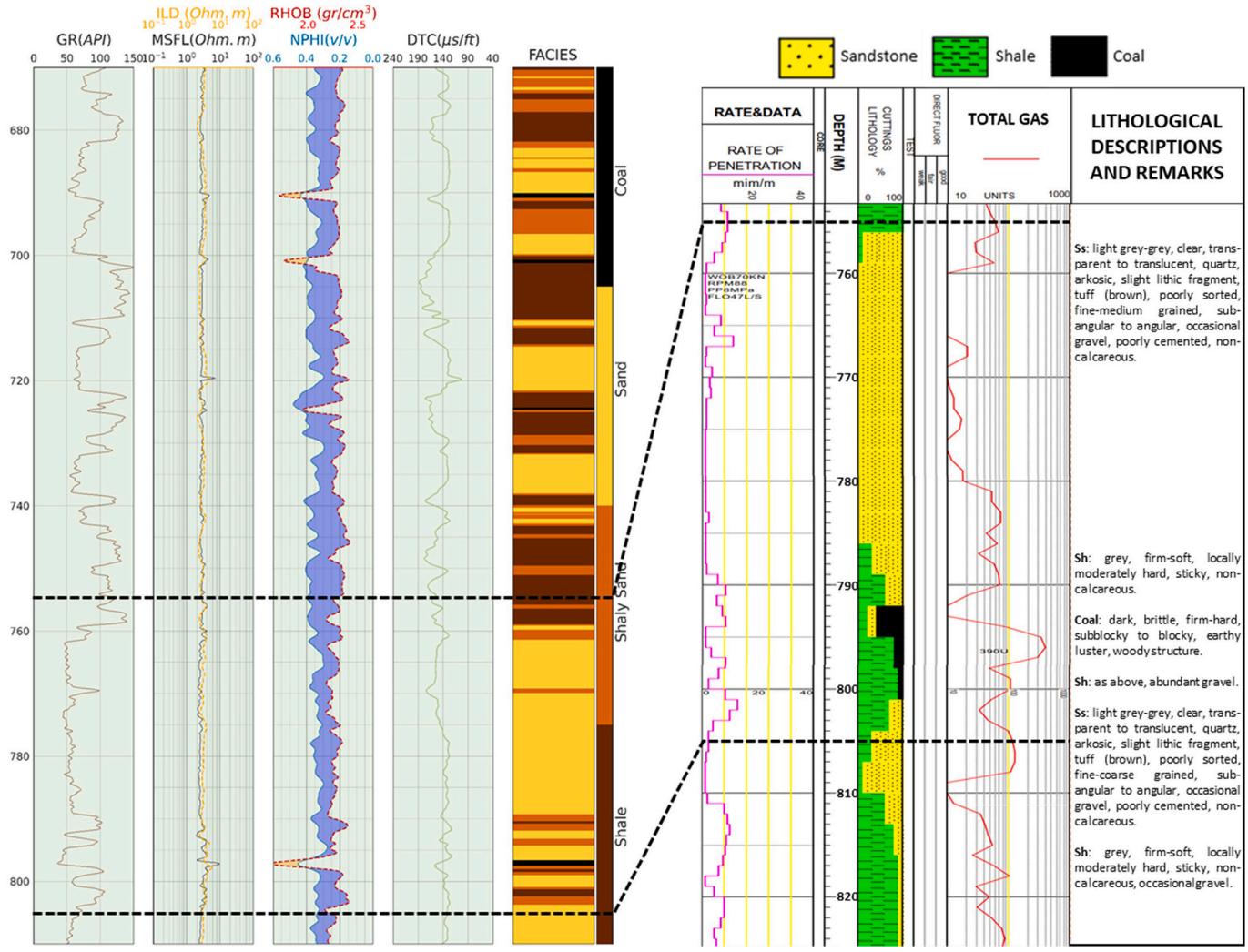


Fig. 4. Mud log of Kay well at depth 670–810 m.

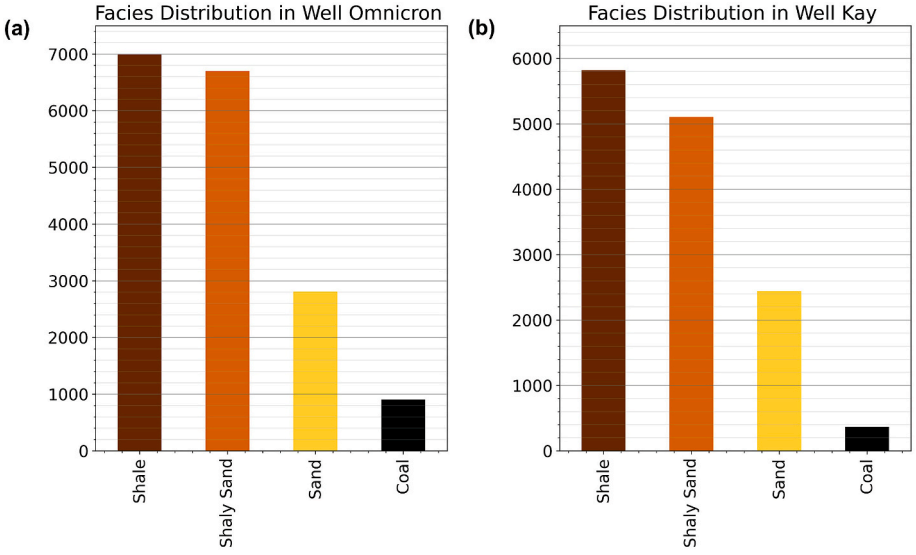


Fig. 5. (a) Omnicron facies distribution visualization (b) Kay facies distribution visualization.

Table 1
Facies distribution of Omnicron and Kay wells (by percentage).

Well name	Facies Type	Distribution	Percentage (%)
Omnicron	Shale	6991	40
	Shaly Sand	6700	38
	Sandstone	2809	16
	Coal	905	5
Kay	Shale	5819	42
	Shaly Sand	5104	37
	Sandstone	2441	17
	Coal	368	2

Table 2
Evaluation of the 80 % Training model dataset on the remaining 20 % of Testing dataset (Omnicron well).

ML - Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	98.48	98.47	98.47	98.47
Random Forest	98.85	98.85	98.85	98.85
Gradient Boosting	98.53	98.53	98.53	98.53
Logistic Regression	93.22	93.20	93.22	93.20
K-Nearest Neighbors	96.06	96.06	96.06	96.06
Support Vector Machine	95.89	95.89	95.89	95.87

Table 3
Comparison of testing model on validation dataset (Kay well) using supervised learning.

ML - Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	72 .91	79.55	72.90	74.84
Random Forest	87.49	89.58	87.48	88.12
Gradient Boosting	87.01	89.31	87.00	87.90
Logistic Regression	61.44	67.80	61.44	62.52
K-Nearest Neighbors	82.68	85.31	82.68	83.32
Support Vector Machine	82.73	87.31	82.72	84.33

Table 4
Comparison of Precision score on validation dataset (Kay well).

Facies	Precision				Precision average
	Shale	Shaly Sand	Sandstone	Coal	
Decision Tree	86.92	66.50	98.26	20.17	79.55
Random Forest	93.93	83.90	98.99	37.55	89.58
Gradient Boosting	93.14	84.72	98.34	33.81	89.31
Logistic Regression	78.59	52.20	82.28	17.46	67.80
KNearest Neighbor	92.48	74.99	97.09	37.06	85.31
Support Vector Machine	93.82	79.03	98.71	23.89	87.31

Table 5
Comparison of Recall score on validation dataset (Kay well).

Facies	**Recall**				**Recall average**
	Shale	**Shaly Sand**	**Sandstone**	**Coal**	
Decision Tree	64.39	76.18	85.37	79.62	72.90
Random Forest	86.63	90.56	84.39	79.08	87.48
Gradient Boosting	86.36	88.32	87.42	77.72	87.00
Logistic Regression	48.32	62.77	89.43	64.67	61.44
KNearest Neighbor	80.34	87.83	79.23	71.20	82.68
Support Vector Machine	81.15	84.11	84.56	76.36	82.72

Table 6
Comparison of F1-Score score on validation dataset (Kay well).

Facies	F1-Score				F1 Score average
	Shale	Shaly Sand	Sandstone	Coal	
Decision Tree	73.98	71.01	91.36	32.18	74.84
Random Forest	90.13	87.10	91.11	50.92	88.12
Gradient Boosting	89.62	86.48	92.56	47.12	87.90
Logistic Regression	59.85	57.00	85.71	27.50	62.52
KNearest Neighbor	85.98	80.91	87.25	48.74	83.32
Support Vector Machine	87.03	81.49	91.09	36.40	84.33

the number of features to better divide, the maximum number of nodes, the maximum number of sheets, and the choice of nodes function. In terms of scoring criteria for the partition and node selection, two methods are available: Gini (1) and entropy (2).

$$Gini(E) = 1 - \sum_{j=1}^{c} p_j^2 \tag{1}$$

$$H(E) = -\sum_{j=1}^{c} p_j \log p_j \tag{2}$$

The concept of Gini was described in Equation (1), Gini tests the reduction of class impurity from partitioning the feature space. Following Equation (2), Entropy is the details of the feature subspace, with *p(j)* is the prior probability that an instance is owned to class *j*, where it is the number of instances in node, divided by the number of class *j* instances in the node. The use of each parameter is determined by the quantity and format of the data available for training. In the case of massive volumes of data, the entropy parameter may be modified using logarithm, needing more statistical analysis and rendering it comparatively slower than the Gini criterion (Jing et al., 2017).

3.2.2. Random forest

A random forest is a hybrid approach to pattern categorization that relies on decision trees. Breiman (2001) proposed this method, which adds an extra layer of randomness to the bagging process. Random forests change the way classification or regression trees are built, in addition to using a different data bootstrap sample for each tree. In contrast to a decision tree, where each node is divided by the best partition between all variables, a random forest separates each node by the best of a set of predictors randomly chosen at that node. It is also a set of multiple decision trees, with the sum of all processed trees being calculated. In another way, it is a part of the configuration parameters for node division and selection (Gini and entropy), decision tree depth, and decision tree number, with a standard approach involving a bootstrap that calculates the average as a quantitative technique for measuring the element of error associated with a specific learning estimator. The random forest in this context includes many types of calculations, such as average calculations that combine many estimators and revert back the mean of their forecasts, reduce variances, and escalate calculations that combine a number of small and special estimators with correspondingly large returns and total outcomes. (Raschka, 2015; James et al., 2013).

3.2.3. Gradient boosting

Gradient Boosting algorithms were first created to solve classification problems by the machine learning group (Schapire, 1990; Freund, Y., 1995; Freund, Y. and Schapire, 1996). The theory method integrates several "ineffective learners" classification algorithms iteratively to create an "efficient learner" with improved predictive accuracy. Friedman et al. (2000) took a mathematical approach to boost, connecting the booster algorithm to loss function concepts. The gradient boosting machine (GBM) method was added by Friedman (2001) to improve the

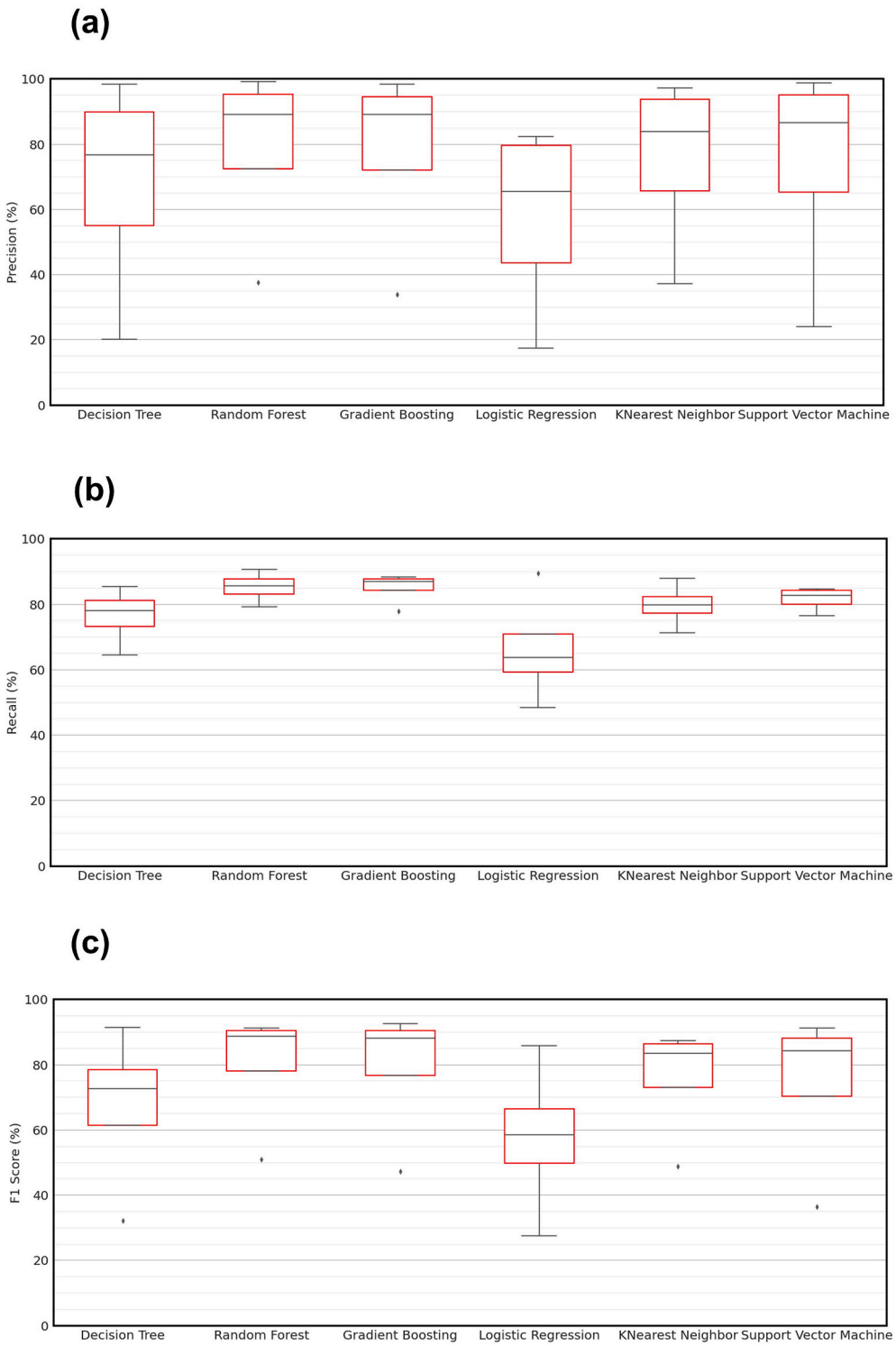


Fig. 6. (a) Boxplot diagram of the Precision evaluation (b) Recall evaluation (c) F1-Score evaluation of classification models within each facies.

regression boost. The GBM method can be thought of as a quantitative optimization algorithm that seeks to find an optimal loss function that minimizes the predicted loss. As a result, the GBM algorithm adds a new decision tree (i.e., "weak learner") at each stage to reduce the loss function the most. In regression, the algorithm starts by creating the model with a first approximation, which is usually a decision tree that minimizes the loss function (mean square error for regression), and then applying a new decision tree to the residual current and extending it to the previous model to keep updating the residual error. The algorithm will continue to iterate until the user-specified maximum number of iterations is reached. This approach is known as a step-wise process because the decision trees applied to the model in previous steps are not modified at each new step. The model is improved in areas where it does not perform well in order to apply the decision trees to the residues.

When the contribution of the extended decision tree is reduced at each iterative step using the deterioration parameter known as the learning rate, the GBM algorithm produces excellent results. The deterioration approach in the sense of GBM is based on the idea that a larger

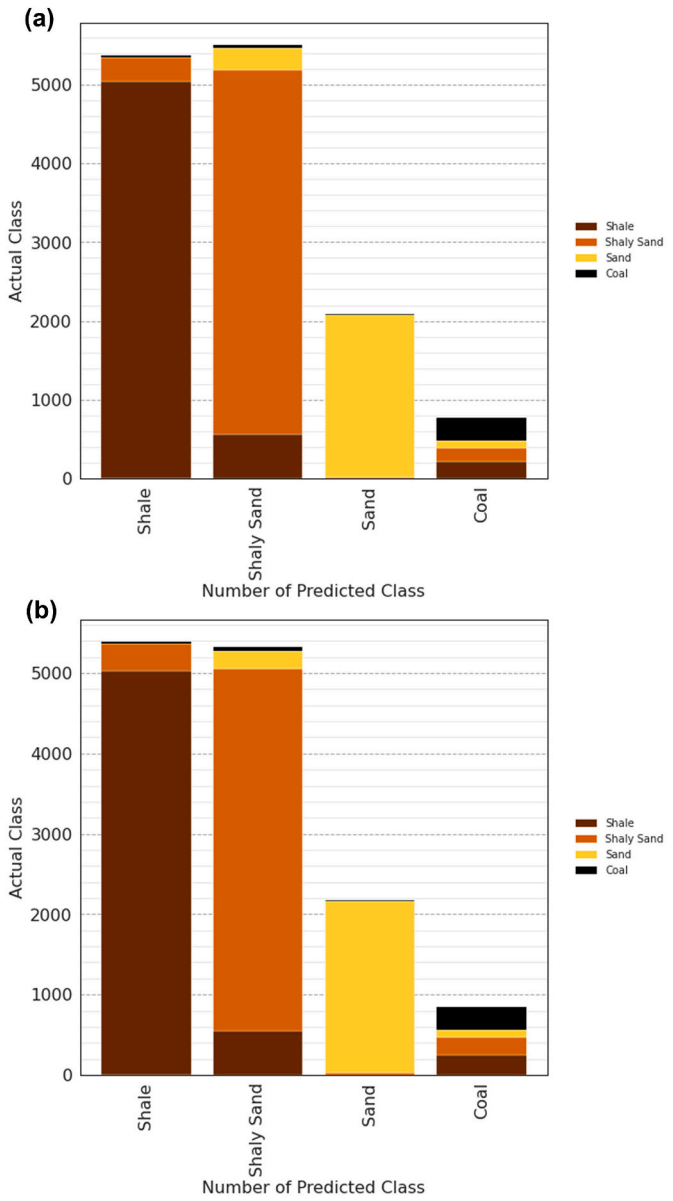


Fig. 7. Class prediction error for (a) Random forest (b) Gradient boosting.

number of minor steps has greater precision than a smaller number of major steps. The learning parameter can have a value between 0 and 1, with the lower the value, the more accurate the model. In contrast, since the value is inversely proportional to the number of iterations, choosing a better deterioration (smaller) requires a greater number of iterations to achieve convergence. Furthermore, where the error function was a traditional squared-error failure, the learning process resulted in a consecutive error correction to gain a better understanding. Normally, a researcher with different loss functions arising from the likelihood of a task-specific loss will be impacted by this option of the loss function (Johnson and Zhang, 2013). The GBMs' flexibility would make them extremely adaptable to any data-driven mission. This would give the platform a lot more flexibility, making it a much better substitute for the loss feature, but it would be a matter of trial and error. The GBMs, on the other hand, have extensive experience with a wide range of machine learning implementations, challenges, and data mining. As originally suggested, the entire gradient boosting algorithm has been developed (Friedman, 2001; Johnson and Zhang, 2013).

3.2.4. Support vector machine (SVM)

SVM is a supervised learning technique that uses a boundary (hyperplane) to describe data, making it easier to view, classify, and distinguish between different instances of data classes (Vapnik, 1998). Its implementation is based on the structure and association of such data, with linearly separable and linearly non-separable data being used. The formula is used to calculate the vector that best describes the hyperplane. The aid vectors are the points closest to the hyperplane axis, and the boundary is the distance between them. The framework favors the hyperplane with the largest margin, which is considered to be a more robust and error-prone model.

SVM's expression is fast and accurate, resulting in an excellent functional return on the classification's results. After applying the hyperplane representation to the group data, the model of linearly non-separable data includes a methodology for analyzing the data group that uses the kernel method to find similarities and associations.

3.2.5. Logistic regression

The relationships between a categorical contingent outcome and one or more independent explanatory variables are evaluated using logistic regression (Hosmer et al., 2013). The predicted likelihoods for each group will be calculated using logistic regression. Have a mark on each multi-class that can be represented using a logistic regression function (Houston and Woodruff, 1997). Logistic regression aims to find the best model for defining the relationship between the categorical characteristics of the dependent variable (the likelihood of occurrence, which is usually limited to 0 to 1) and the set of independent variables.

3.2.6. K-nearest neighborhood (KNN)

If the volume of data grows too large to evaluate in a reasonable amount of time, one alternative is to make decisions based on a small portion of the data (Neeb and Kurrus, 2016; Yong et al., 2009; Deng et al., 2016). The K-means clustering algorithm was used by Deng et al. (2016) to divide large training datasets into different clusters, and then the KNN algorithm was used to classify each test instance based on query instance neighbors in the nearest data cluster. The cluster with the shortest Euclidean distance between its center and the sample population was the closest.

3.3. Train test split for prediction

The results of the method-based practice tests were related to the variance of preparation, testing, and validation in templates 1, 2, and 3, followed by a percentage of 10 % for validation and 10–20 % for testing (Storkey, 2009; Korjus et al., 2016). The expedition classification was used to separate the functional template, preparation, validation, and testing. Following the training and forecasting, a table was created to compare the output of each template and each phase in the data set under consideration. For better structure and presentation, classification metrics and uncertainty matrix values were aggregated at each processing step.

3.4. Performance validation and evaluation

The confusion matrix is an example of successful and expected values that can be used in the proposed models to visualize the performance of the machine learning classifier process (Maria Navin and Pankaja, 2016). The confusion matrix is a table that is used to describe the performance of classification or procedure. It is structured to identify a binary dataset. The reasoned shows the accuracy of the evaluated documents by incorporating the true outcomes in the matrix's ordered form.

According to accuracy (Eq. (3)), it is established by dividing true positive and true negative values, as well as total positive and negative values:

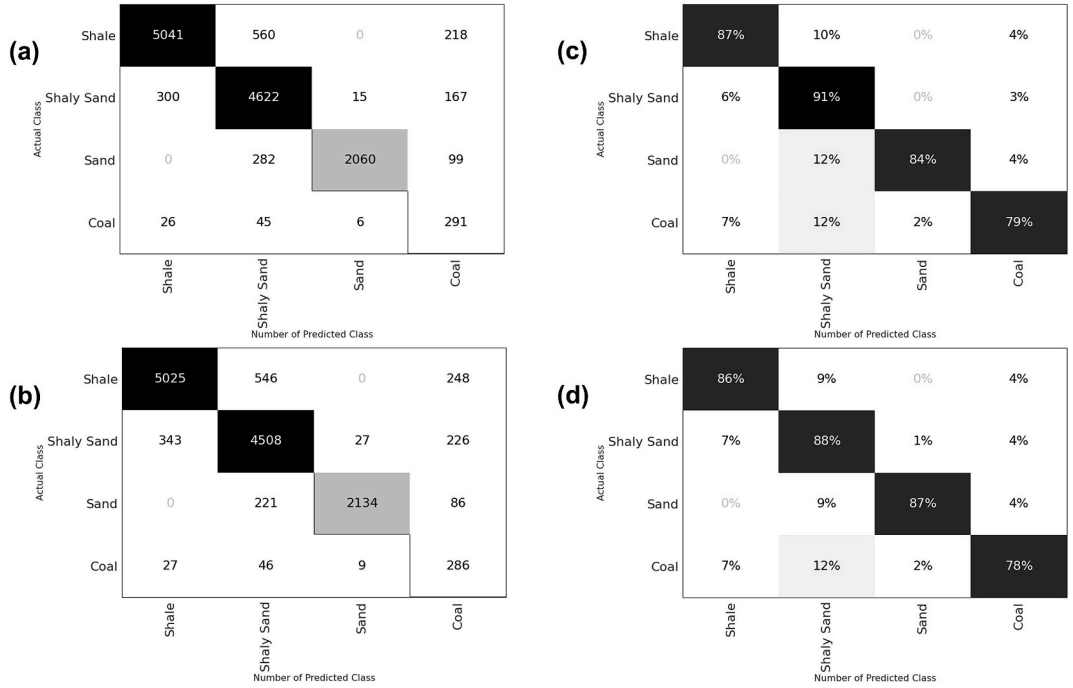


Fig. 8. Confusion matrix distribution of samples for (a) Random forest (b) Gradient boosting along with the distribution percentage for (c) Random forest (d) Gradient boosting.

$$Accuracy = \frac{TP + TN}{P + N} \tag{3}$$

It is important to note that this classification metric would produce a false method of results since it tests the viciousness between the dataset classes' returns (Hossin and Sulaiman, 2015).

The true positive values are divided by the number of the true positive and false positive values to calculate the precision. The recall is calculated by dividing all true positive and false negative values by the number of all true positive and false negative values. The F1-score metric includes precision (Eq. (4)) and recall (Eq. (5)). The following are the equations:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

F1-score (Eq. (6)) is the harmonic mean between precision and recall for the evaluation process. The F1-score is useful in the processing of datasets with a large number of different categories. The following is an example of an equation:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6}$$

The Receiver Operating Characteristic (ROC) is a diagram that can be used to evaluate, represent, and choose prediction systems (Tharwat, 2018). To determine the outcome, the uncertainty matrix is used, and two parameters of the probability of accuracy (True Positive Rate (TPR) and False Positive Rate (FPR)) are represented. TPR (Eq. (7)) and FPR (Eq. (8)) are divided into two categories:

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

At different classification points, the ROC curve shows TPR vs. FPR. The multidimensional adaptation aids in the visualization of the outcome variables around the graph spectrum. The descending diagonal (0,1) is a classification paradigm that is equally effective in both classes. The highest values are found in the upper left triangle of this diagonal, while the lowest values are found in the lower right triangle. The detection of signals and the assessment of the noise signal's propagation efficiency are at the heart of it. The analysis of major studies on the use of ROC for evaluation in medicine (Tilaki-Hajian, 2013), economics (Gajowniczek et al., 2014), climate forecasting (Zhao et al., 2011), and in geoscience (Vakhshoori and Zare, 2018; Chen and Wu, 2016; Airola et al., 2019) evaluating using this method.

3.5. Log-facies classification

Log-facies classification methods aimed to predict the lithofacies at a well site based on the rock's physical properties. Statistical approaches are one of the most advanced tools for facies classification. It usually improves the function that defines the probability of a collection of log responses belonging to a number of lithofacies, resulting in the most accurate identification of lithofacies along the borehole. Well logs, like neutron porosity, density, gamma-ray, resistivity, sonic log (P-wave velocity), and petrophysical properties including porosity, shale volume, and fluid saturations, were used as inputs for lithofacies classification (Darling, 2005; Ellis and Singer, 2007).

3.6. Rock physics diagnostics

Rock physics analysis was carried out to investigate the physical properties of each lithofacies with respect to their sedimentology and diagenetic processes such as compaction and cementation as these processes may severely affect the rock physical properties such as velocity and density (Avseth et al., 2005). Such changes in rock physical properties may deteriorate the lithofacies prediction results as the rocks that have been introduced to specific diagenetic processes have different rock physical properties and may ultimately lead to incorrect lithofacies prediction. To carry out the analysis, several rock physics cross-plots were constructed to observe the property distribution of each lithofacies and in this way, the potential cause of misprediction can be

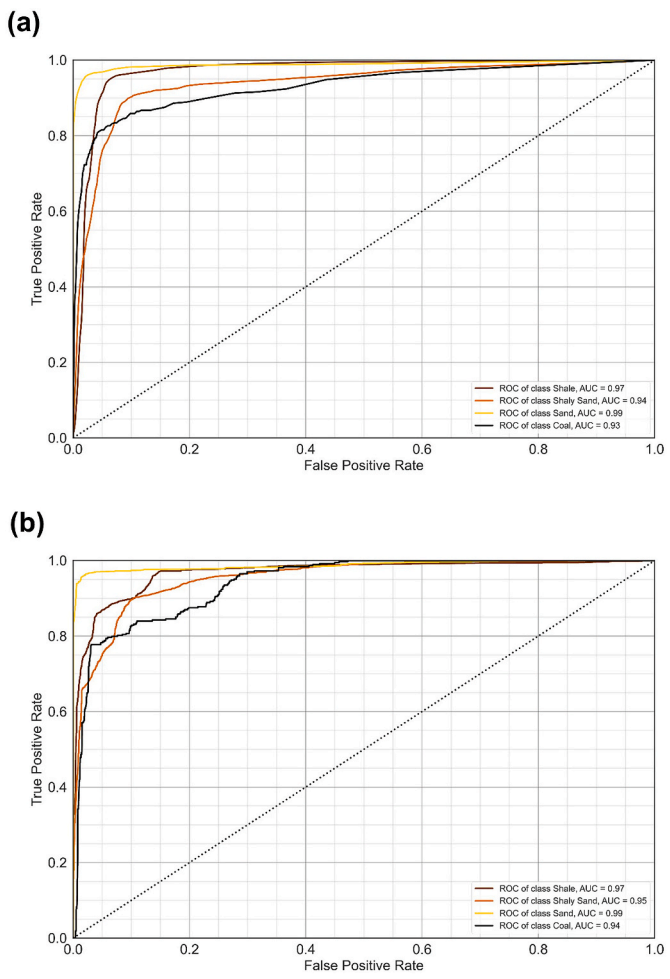


Fig. 9. ROC curve within score of each facies class for (a) Random forest (b) Gradient boosting.

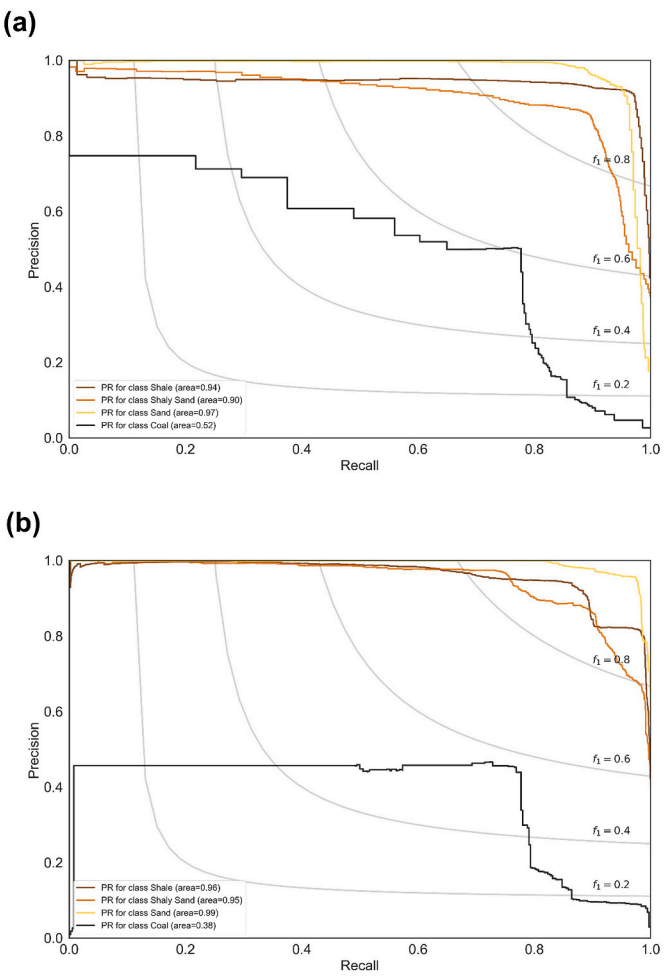


Fig. 10. Precision-recall curve and f1 curve validation within score of each facies class for (a) Random forest (b) Gradient boosting.

determined.

4. Experiments

The facies identified in this study are based on an analysis of approximately 31,137 total samples, with the results standardized to have a zero mean and unit variance. The input data for the first well (Omnicron) consisted of 17,405 samples, which were divided into training sets (about 13,732 samples, or 80 % of the dataset) and test sets (about 3481 samples, or 20 % of the dataset) to train the models. Furthermore, the second well (Kay) contains 13,734 samples for data validation. After that, supervised learning was used as a blind well test or to validate an assessment model. Six wireline log curves were used in machine learning, including gamma-ray (GR), sonic (DTC), resistivity (ILD) and (MSFL), neutron porosity (NPHI), and density (RHOB). Following the original interpretation carried out in the Tarakan Basin, the analyzed intervals were represented by shale, shaly sandstone, sandstone, and coal.

4.1. Workflow

The general workflow of a supervised learning classifier follows the steps in the classifier evaluation process:

1) Data preprocessing for NaN data removal and dataset standardization.
2) Obtaining a data set of training well logs and constructing a training set of about 80% and a testing set of about 20% to verify the model's results.
3). Applying the training model to the training dataset as a well facies classifier.
4). For the validation well dataset, run a training model on the testing well log dataset.
5). On each well facies and the entire testing well log dataset, evaluate the model output of the obtained classifier using accuracy, precision, recall, F1 score, ROC-AUC curve, and PR-Recall curve.
6). Comparison of machine learning results with real lithological data using log-facies classification (e.g. mud log and cutting samples).
7). Rock physics review to assess the possible cause of difference between the original lithofacies and machine learning by understanding the physical properties of each lithofacies.

4.2. Evaluation models

Each approach will be evaluated using standard assessment methods such as accuracy, precision, recall, and F1 score in order to determine the efficiency of the supervised learning classification. In addition, the output of each machine learning model for classifying each facies was visualized using the class prediction error evaluation, confusion matrix, ROC curve, and precision-recall curve.

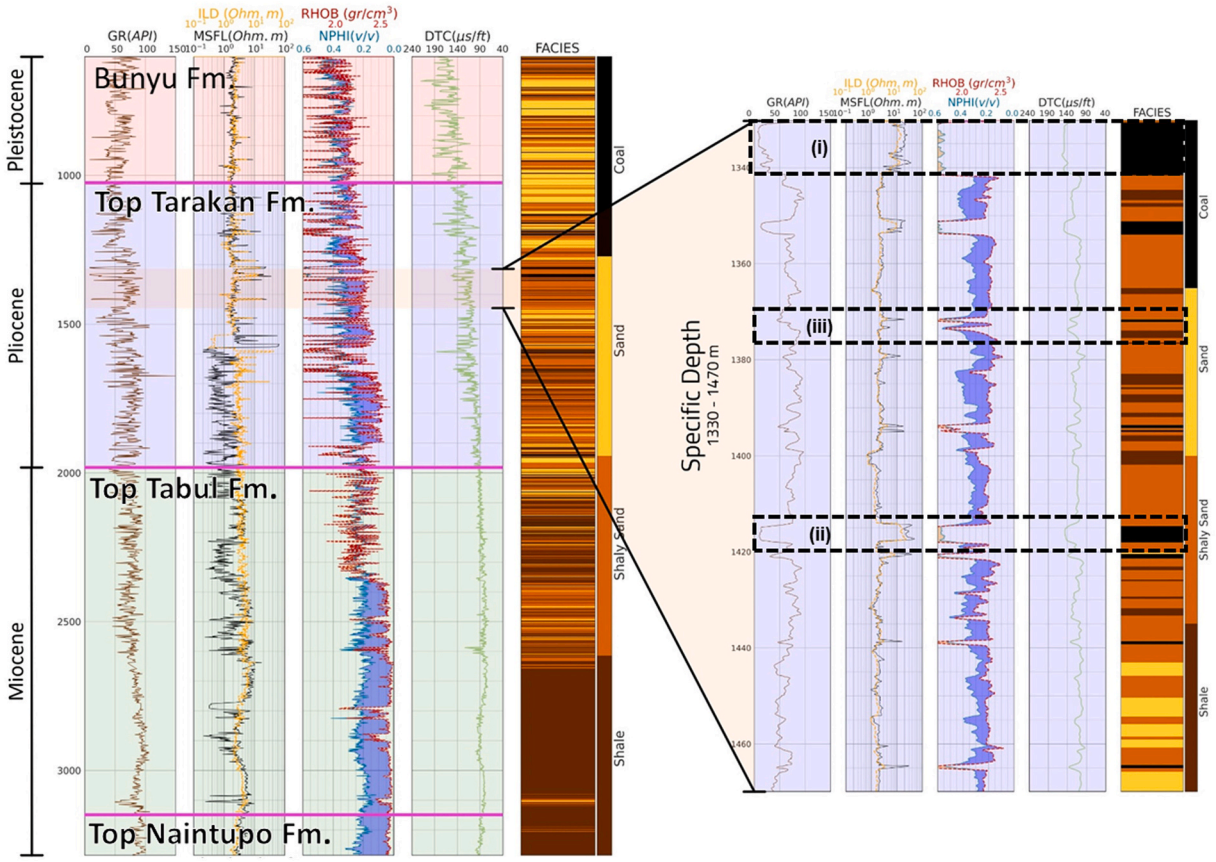


Fig. 11. Omnicron well log with an inset depth from 1330 to 1470 m showing the different log response between thick coal beds at (i) 1330–1340 m, (ii) 1415–1418 m, and the thin coal beds at 1372–1374 m.

4.3. Geological analysis

Machine learning models were then evaluated on rock physics cross-plots to observe the property distribution of each facies, this experiment used log-facies classification and rock physics analysis, in order to better understand the behavior of the model prediction. DTC vs. GR and RHOB vs. NPHI were the cross-plots for the log-facies classification used in this experiment. In addition, the DTC vs RHOB and PHIT vs Vp cross-plots were utilized to carry out rock physics analysis to determine the possible post-depositional processes.

5. Results and discussion

Shale, shaly sandstone, sandstone, and coal are among the interpreted lithofacies in the dataset. The main phase in the research was to fine-tune the machine learning model's configuration settings in order to achieve the highest possible classification accuracy in the dataset. The first well, Omnicron, was used to train the model, and the second well, Kay, was used to validate the experiment models. The majority of the evaluation's findings were validated on Kay well dataset samples, demonstrating how well many supervised learning algorithms performed in classifying the lithofacies in the Tarakan basin.

5.1. Machine learning scoring & evaluation

The decision tree, random forest, gradient boosting, logistic regression, support vector machine, and k-nearest neighbors methods were used to train the model. Table 2 shows the findings for the training dataset. Using Omnicron as the training dataset, it can be seen that the random forest has accuracy, precision, recall, and F1-score of about (98.85%, 98.85%, 98.85%, and 98.85% respectively) and gradient boosting has results of about (98.53%, 98.53%, 98.53%, and 98.53% respectively) which outperformed the other models. Furthermore, the results were compared and shown in Table 3 to validate the model, and it is shown that the accuracy of Kay well for validation dataset using random forest is about 87.49% and gradient boosting is about 87.01%, indicating that even with data validation, this result still dominates compared to other models. In comparison to the other models, logistic regression models yielded the lowest result of 61.44%. The random forest is the most reliable model in terms of accuracy as compared to the other models, with training and validation accuracy of 98.85% and 87.49%, respectively. Gradient boosting is the second stable model, with training and validation accuracy of 98.53% and 87.04%, respectively. As a result of these findings, it has decided to concentrate this work on random forest and gradient boosting for model evaluation performance analysis.

The precision results on Table 4 for the facies classification of shale, shaly sandstone, sandstone, and coal using random forest (93.93%, 83.90%, 98.99%, and 37.55%, respectively) and gradient boosting (93.14%, 84.72%, 98.34%, and 33.81% respectively). Because of the volatility of coal physical properties in the Tarakan Basin, both models still have a low precision result when forecasting coal.

Table 5 shows the recall results, where the random forest and gradient boosting overcome for shale (86.63% and 86.36%, respectively) and shaly sand (90.56% and 88.32%, respectively). In Table 6, the random forest method achieves the highest F1-score for shale, shaly sandstone, and coal (90.13%, 87.10%, and 50.92%, respectively), but the best method for sandstone was gradient boosting with an F1-score of around 92.56%.

For a visual comparison of the considered model results for each facies using the Kay well for validation dataset, see the boxplot diagrams in Fig. 6. The maximum and minimum evaluation score using (a)

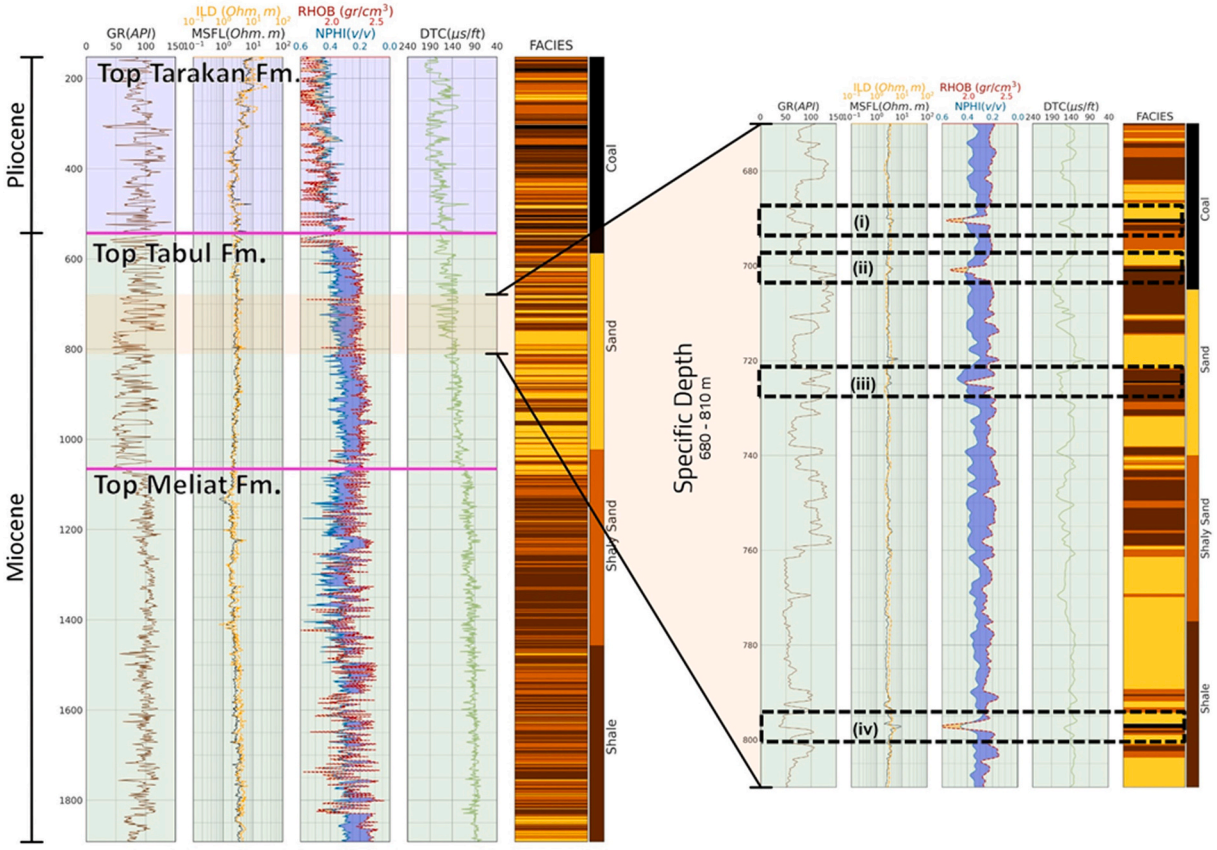
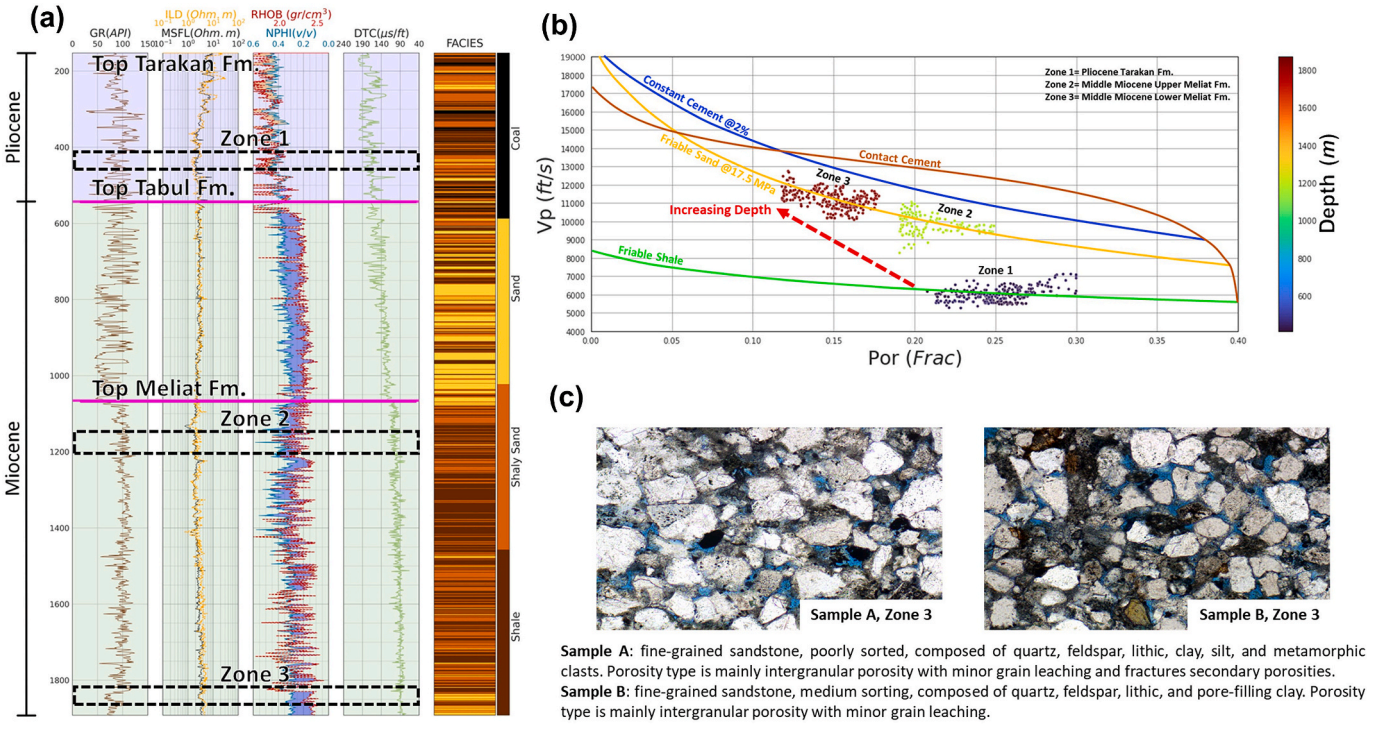


Fig. 12. Kay well log with an inset depth from 670 to 810 m showing the log response of thin coal beds at (i) 690 m, (ii) 701 m, (iii) 725 m, and (iv) 797 m.



Sample A: fine-grained sandstone, poorly sorted, composed of quartz, feldspar, lithic, clay, silt, and metamorphic clasts. Porosity type is mainly intergranular porosity with minor grain leaching and fractures secondary porosities.
Sample B: fine-grained sandstone, medium sorting, composed of quartz, feldspar, lithic, and pore-filling clay. Porosity type is mainly intergranular porosity with minor grain leaching.

Fig. 13. (a) Kay well log with the 3 zones of investigation (b) PhiT vs Vp (rock physics diagnostic) cross-plot of each zone of investigation and (c) petrography analysis at zone 3.

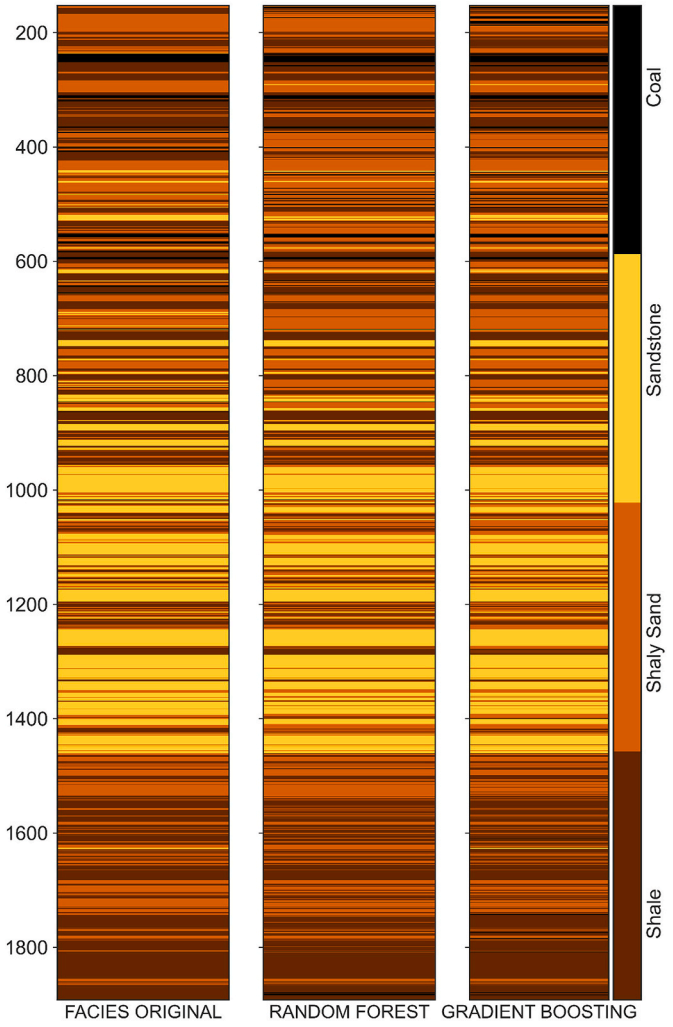


Fig. 14. Comparison of original facies (Kay well) with random forest and gradient boosting prediction.

precision, (b) recall, and (c) F1-score, as well as the evaluation information of each lithofacies compiled based on Tables 4–6.

Fig. 7 (a) depicted the class prediction error for random forest and (b) gradient boosting, and the results were further supported by a confusion matrix (Fig. 8) based on each instance of samples expected in each facies for (a) random forest and (b) gradient boosting. According to Fig. 8 (c) for the random forest, each true class for each facies is around (87%, 91%, 84%, and 79% respectively). Furthermore, as shown in Fig. 8(d) for the gradient boosting, the product of true class for each lithofacies is around (86%, 88%, 87%, and 78% respectively). As a result of this finding, predicting the behavior of shale, shaly sandstone, and sandstone is more stable to predict the behavior of lithofacies. However, by using random forest and gradient boosting to discern coal, the models still struggled because the true class result on coal was below average compared to other facies.

For model evaluation, the ROC curve is a rigorous sensitivity/specificity analysis. In the ROC curve, the true positive rate (TPR) is plotted against the false positive rate (FPR). The ROC curves of random forest are shown in Fig. 9 (a), which include the results of each lithofacies (shale = 0.97; shaly sandstone = 0.94; sandstone = 0.99; coal = 0.93). In addition, Fig. 9 (b) depicted the gradient enhancing ROC curves with the following results (shale = 0.97; shaly sandstone = 0.95; sandstone = 0.99; coal = 0.94). Random forest and gradient boosting are suitable classifiers for predicting sandstone since they make few prediction errors, with both showing 0.01 error, indicating that they performed well in the Tarakan Basin sandstone. This indicates that the classifier is capable of perfectly separating the sandstone groups, with a true positive rate of 0.99. The random forest results for shale, shaly sandstone, and coal were strong, with ROC scores of 0.97, 0.94, and 0.93, respectively. This case also worked for gradient boosting, with ROC scores of 0.97, 0.95, and 0.94, respectively. Both classifiers have performance ratings that are strongly correlated with the outcome as a result of these findings. With average scores of 0.96 and 0.96, respectively, both classifiers achieve a high TPR at the expense of a high FPR. Random forest and gradient boosting have similar performance for predicting facies on

Table 7
Summary of rock physical properties of each lithofacies (Kay well).

Lithofacies	GR	RHOB	NPHI	DTC	Remarks
Shale	95–150	1.80–2.10 1.95–2.50	0.24–0.53	145–190 72–180	Pliocene Miocene
Shaly Sandstone	70–105	1.70–2.10 2.00–2.53	0.21–0.50	145–188 69–174	Pliocene Miocene
Sandstone	40–80	1.79–2.29 2.15–2.56	0.19–0.42	133–170 65–155	Pliocene Miocene
Coal	15–120	1.50–1.75[a]	up to 0.60 0.36–0.54	164–199 152 - 185[b]	Thick coal bed Thin/mixed coal bed

[a] RHOB may be as high as 1.95.
[b] DTC May be as low as 135 at Miocene.

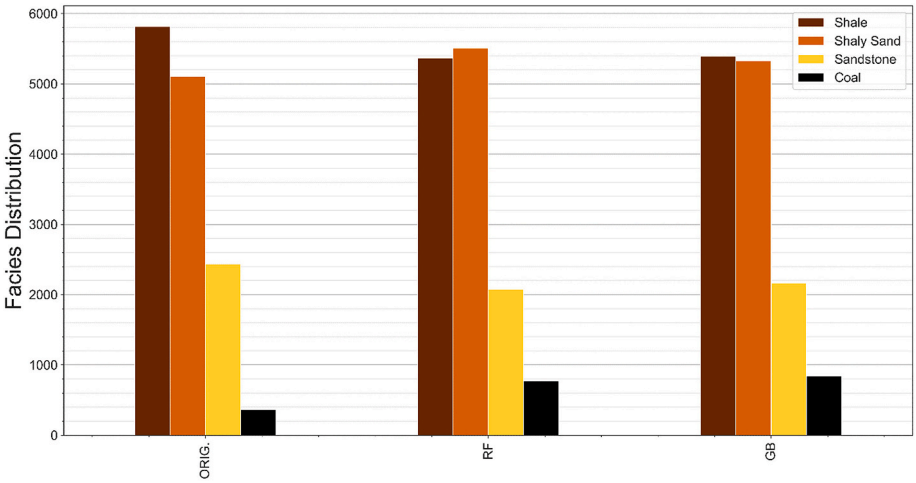


Fig. 15. Distribution of original facies (Kay well) with random forest and gradient boosting predictions.

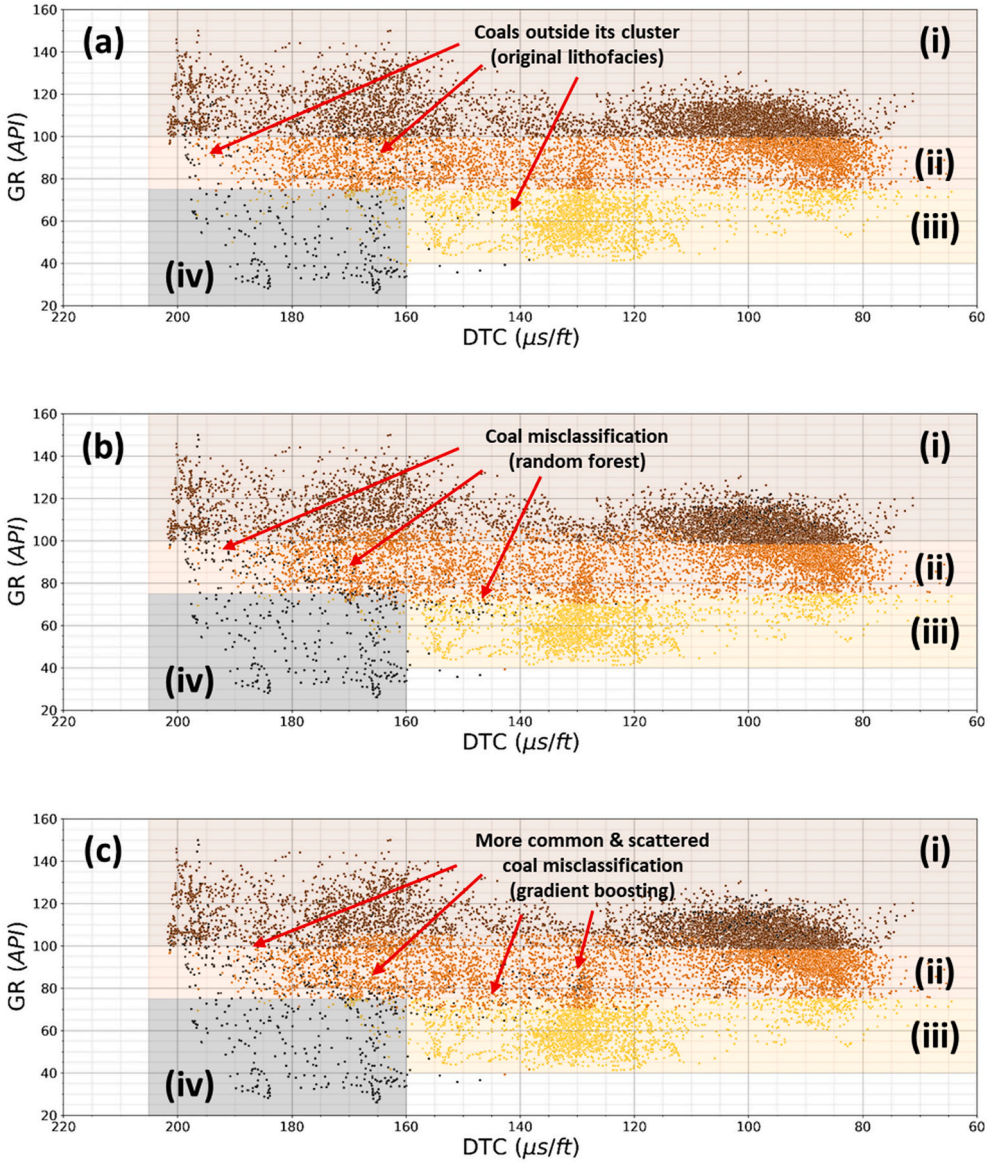


Fig. 16. Log-Facies classification using DTC vs GR cross-plot of (a) original lithofacies and the prediction results of (b) Random forest and (c) Gradient boosting showing the shale, shaly sandstone, and sandstone clusters.

the Tarakan basin, according to this ROC curve result.

The precision-recall curve often depicts how different thresholds trade-off precision and recall. High precision is associated with a low false-positive rate, and high recall is associated with a low false-negative rate, so the high region under the curve represents both high recall and high precision. The PR-Recall curve for random forest for each multi-class of lithofacies (shale = 0.94; shaly sandstone = 0.90; sandstone = 0.97; coal = 0.52) is shown in Fig. 10 (a), with an average PR-area-under-curve of 0.83 and a PR-area-under-curve of 0.83. The PR-Recall curve for gradient boosting is also shown in Fig. 10 (b), with an average PR-area-under-curve of 0.82 for each multiclass of lithofacies (shale = 0.96; shaly sandstone = 0.95; sandstone = 0.99; coal = 0.38) These findings indicate that shale, shaly sandstone, and sandstone can all be predicted with high accuracy. Both classifiers do a good job of separating each facies, but not perfectly. However, the precision-recall score using random forest and gradient boosting in the Tarakan basin revealed that coal performed below average. Both classifiers generated performance scores that are only tangentially related to the result, and only for the coal facies case did they achieve a high recall with low precision.

5.2. Geological analysis from the machine learning model

In order to understand the low precision issue of coal lithofacies shown in the inset of Omnicron and Kay well in Figs. 11 and 12, respectively. The well has peculiar subsurface characteristics, for example in Kay well at specific depths ranging from 680 to 810 m. The greatest uncertainty comes from the coal, which is denoted by (i) 690 m, (ii) 701 m, (iii) 725 m, and (iv) 797 m. The occurrences of these coals are supported by the mud log data (Fig. 4). Even though coals have distinctively different RHOB-NPHI values than the other lithofacies, however, they have similar/overlapping GR, DTC, and sometimes RHOB values as the other lithofacies. On the other hand, as shown by the inset of Omnicron well in Fig. 11, the thick coal beds at (i) 1330–1340 m and (ii) 1415–1418 m exhibit different log responses than the thin coal bed at (iii) 1372–1374 m. In comparison to thin coal, the thick coal has distinctively lower RHOB, but higher NPHI and DTC log reading. This may be due to the thin coal bed being mixed with shale or another lithology. Such well log response overlap can result in lithofacies misclassification. Due to the uncertainty involved with coal, it is one of the challenges associated with implementing machine learning in the

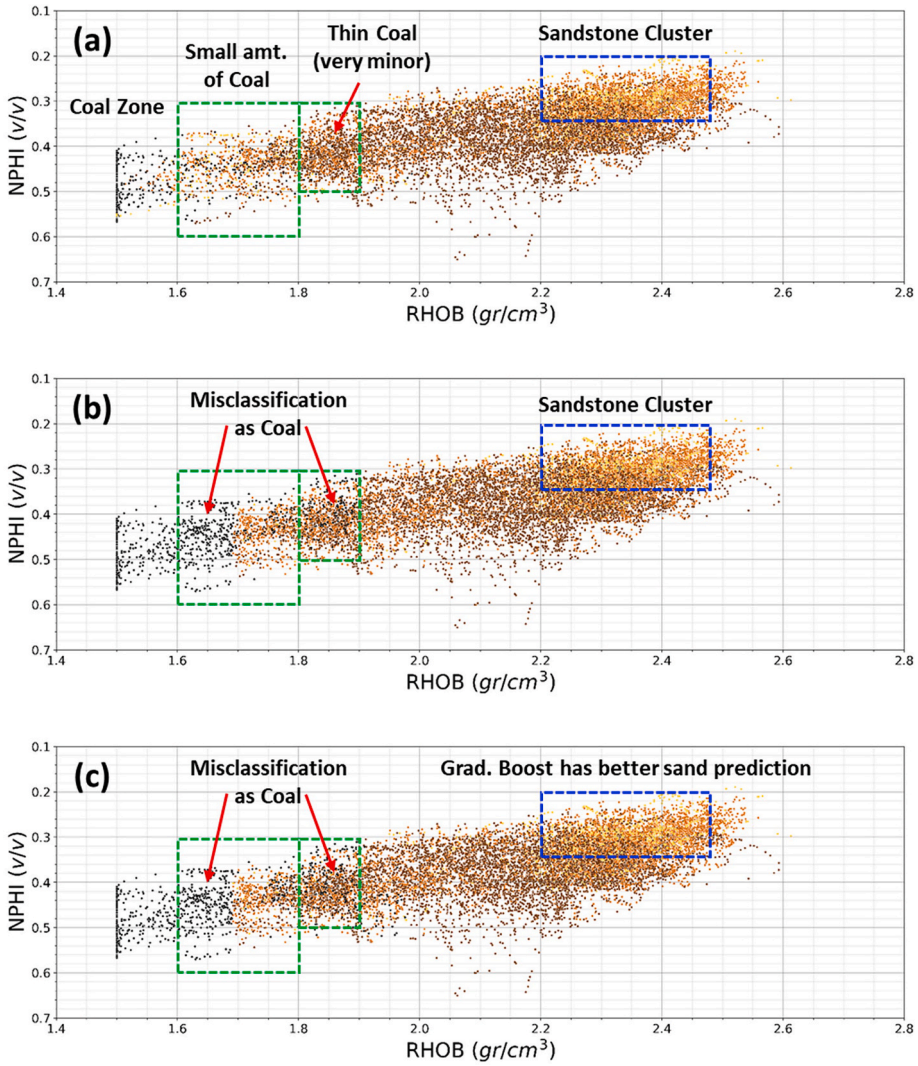


Fig. 17. Log-Facies classification using RHOB vs NPHI cross-plot of (a) original lithofacies and the prediction results of (b) Random forest and (c) Gradient boosting showing the coal cluster.

study area.

To understand the rock physical properties, cut-off distribution, and rock behavior, a rock physics diagnostic was performed. PHIT vs Vp cross-plot of Kay well (Fig. 13(a)) was constructed from data sampled in the Pliocene Tarakan (zone 1), Middle Miocene Tabul (zone 2), and Middle Miocene Meliat Formations (zone 3) to interpret the potential cause of variation in rock physical properties in the study area. Based on (Fig. 13(b)), It can be observed that the Pliocene and Miocene data distribution are clearly separated, where the Pliocene data distribution follows the Friable Shale trend, whereas the Miocene data distribution follows the Friable Sand trend. This analysis suggests that the change in rock physical properties between the Pliocene to Miocene is attributed to the difference in the rate of compaction with the absence of any significant cementation. According to (Fig. 13(c)), the thin section petrography of zone 3 validated the rock physics diagnostic, where the sandstone in zone 3 is generally fine-grained, has poor-medium sorting, mainly composed of quartz, feldspar, lithic, clay, silt, and clasts of metamorphic rocks with very poor or no cementation. This analysis indicates that the lithofacies prediction at deeper intervals may become more difficult to predict due to the smaller difference of rock physical properties between each lithofacies. Our results based on implementing the random forest and gradient boosting models, however, show an overall result of 87% accuracy, which is a relatively high score and has good performance.

In the next process, the evaluation of machine learning prediction based on geological was performed. Side by side comparison between the original lithofacies, random forest, and gradient boosting results on Kay well are shown in Fig. 14, whereas the bar chart is shown in Fig. 15 to observe the misclassification. It can be observed that shale is slightly decreasing from its original distribution, the shaly sandstone is slightly increasing from the original distribution, the sandstone does not have significant change, whereas the coal is increasing from the original distribution. Based on these results, it can be concluded that coal is the biggest contributor to misclassification in the study area when random forest and gradient boosting were performed. Table 7 shows the summary of the physical properties of each lithofacies in the Pliocene and Miocene sections of Kay well.

Following the result of relatively high accuracy of the models, the investigation of the potential cause of such incorrect prediction was carried out by utilizing the cross-plot analyses to observe the property distribution of each lithofacies. Fig. 16 shows the DTC vs GR of Kay well, which explains that prediction result of random forest in Fig. 16(b) and gradient boosting in 16(c) show similar behavior of the original facies distribution in Fig. 16(a). Each facies is denoted following the cluster separation of (i) shale (ii) shaly sandstone (iii) sandstone (iv) coal. This cross-plot is best used to discriminate the sandstone, shaly sandstone, and shale based on their gamma-ray response (lower GR for sandstone and higher GR for shale) however, the coal has a widely distributed

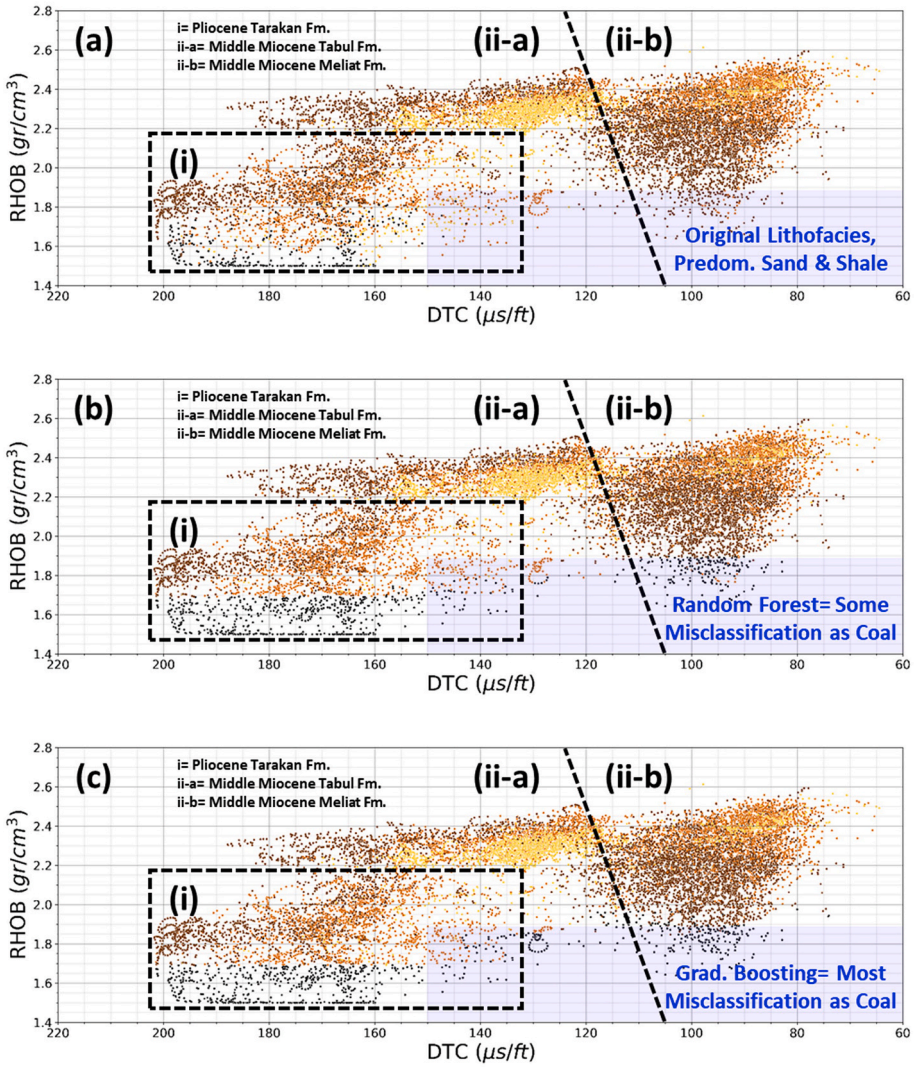


Fig. 18. Log-Facies classification using DTC vs RHOB cross-plot of (a) original lithofacies and the prediction results of (b) Random forest and (c) Gradient boosting showing the variation of rock physical properties between the Plicoene and Miocene sections.

gamma-ray response; this may lead to incorrect lithofacies prediction when utilizing machine learning methods. Based on Fig. 16(c), shows that the gradient boosting model is observably having a higher degree of coal misclassification compared to the random forest, this is also proven according to the previous evaluation of machine learning performance.

Fig. 17 shows the RHOB vs NPHI cross-plot of Kay well that includes the (a) original lithofacies, (b) random forest, and (c) gradient boosting to observe the behavior of each lithofacies. It can be observed that the RHOB and NPHI of shale are clustered around 1.8–2.5 gr/cm^3 and 0.24–0.53 respectively; for the shaly sandstone, the RHOB and NPHI are clustered around 1.7–2.53 gr/cm^3 and 0.21–0.5 respectively; for the sandstone, the RHOB and NPHI are clustered around 1.79–2.56 gr/cm^3 and 0.21–0.42 respectively; for the coal, the RHOB and NPHI are clustered around 1.5–1.75 gr/cm^3 and 0.36–0.60 respectively, however, some coal beds may have RHOB as high as 1.95 gr/cm^3 due to thin coal bed that is mixed with the surrounding shale beds. We could see a better cluster for coal that is commonly known to have low RHOB and high NPHI (lower left side of each cross plot). Some of the coals, however, having more similar RHOB and NPHI logs to that of the other lithofacies (e.g. thin coal or mixing of coal and shale). Based on this cross-plot, in the range of RHOB around 1.6–1.9 gr/cm^3 (highlighted by the green box), coal may be potentially misclassified as shale, shaly sandstone, and sandstone or vice versa. Despite the uncertainty of RHOB to classify the coal NPHI is observably has a more distinctive data range that is useful to discriminate the coal.

Fig. 18 shows the DTC vs RHOB cross-plot of Kay well to observe the variation of velocity (1/DTC) and density of each lithofacies. It can be observed that the DTC and density of shale are clustered around 145–190 μs/ft (Vp = 5263–6896 ft/s) and 1.8–2.1 gr/cm^3 respectively at the Pliocene interval or 72–180 μs/ft (Vp = 5555–13,888 ft/s) and 1.95–2.5 gr/cm^3 at the Miocene interval; for the shaly sandstone, the DTC and density are clustered around 145–188 μs/ft (Vp = 5319–6896 ft/s) and 1.7–2.1 gr/cm^3 respectively at the Pliocene interval or 69–174 μs/ft (Vp = 5747–14,492 ft/s) and 2.0–2.53 gr/cm^3 at the Miocene interval; for the sandstone, the DTC and density are clustered around 133–170 μs/ft (Vp = 5882–7518 ft/s) and 1.79–2.29 gr/cm^3 respectively at the Pliocene and uppermost part of the Miocene interval, it should be noted that at the deeper Miocene interval, sandstones were rarely encountered. However, a small sandstone cluster can be seen around DTC of 65–155 μs/ft (Vp = 6451–15,384 ft/s) and RHOB of 2.15–2.56 gr/cm^3; for the coal, the DTC and density are clustered around 164–199 μs/ft (Vp = 5050–6098 ft/s) and 1.5–1.75 gr/cm^3 respectively, however, small thin beds of coal were also encountered having more similar RHOB (as high as 1.95 gr/cm^3) and DTC to that of shale (152–185 μs/ft – may be as low as 135 μs/ft in the Miocene) due to the potential mixture of coal with shale as opposed to the thick “clean”

coal, this obviously may lead to an incorrect prediction between these lithologies. It should be noted that below an approximate depth of 1070 m, a sharp decrease in DTC can be observed, making the difference of DTC for each lithofacies becomes even smaller. At this depth, shale has DTC of 72–125 μs/ft (Vp = 8000–13,888 ft/s), shaly sandstone is about 69–117 μs/ft (8547–14,492 ft/s), whereas sandstone has DTC of 65–100 μs/ft (10,000–15,384 ft/s). Based on this cross-plot, it can be concluded that despite their good sensitivity towards the change in lithology, caution must be taken when utilizing DTC and RHOB data as these properties are controlled by the compaction and diagenetic processes, indicating that these properties and their corresponding cut-offs will change with increasing depth which denoted in different formation as (i) Pliocene and (ii) Miocene on Fig. 18.

6. Conclusion

An evaluation of supervised learning was used in this study to classify lithofacies in the Tarakan basin of Indonesia. As compared to other machine learning methods, random forest and gradient boosting generated more reliable results in lithofacies classification as both classifiers. These classifiers could accurately predict shale, shaly sandstone, and sandstone based on precision, recall, F1-score, ROC curve, and precision-recall curve. However, within PR-area-under-curve score for coal was below average compared to other facies, and this is clear evidence of machine learning misclassification in predicting one facies to another facies despite the generally high accuracy score. The presence of thin coal beds or a mixture of coal and shale has similar rock physical properties as opposed to the thick "clean" coals. According to the results of log-facies classification, sandstone, shaly sandstone, and shale can be accurately classified. Coal distribution, on the other hand, has been uneven. Furthermore, rock physics analysis showed that despite the difference in rock physics properties between the Pliocene and Miocene sections due to a different rate of compaction and diagenetic processes, the machine learning model was able to correctly predict the lithofacies. There is a chance that the sensitivity of such a result varies at deeper intervals due to the post-depositional processes that each formation underwent, but despite this, the machine learning model has an overall high result and has effectively aided the geological analysis in a much shorter time.

A simple approach for lithofacies classification using random forest and gradient boosting in the Tarakan sub-basin, evaluation of each multiclass of supervised learning methods, high accuracy results despite some coal misclassification, log-facies classification analysis, and rock physics analysis based on supervised learning were the main contributions of this work. Most importantly, this research has shown that the efficiency of random forest and gradient boosting can be evaluated from the standpoint of machine learning and geological views.

Credit author statement

Gian Antariksa: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft. Radhi Muammar: Conceptualization, Validation, Data curation, Writing – review & editing, Project administration. Ji Hwan Lee: Conceptualization, Validation, Methodology, Resources, Writing – review & editing, Funding acquisition, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Conrad Petroleum Ltd. and Directorate General of Oil and Gas in Indonesia for their permission to publish this work. We would also like to acknowledge Pusdatin ESDM for their time and dedication to compile all the oil and gas data in Indonesia in which, some of which were utilized by the authors to carry out this study. On the other hand, this research was financially supported by the Ministry of Trade, Industry, and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) through the National Innovation Cluster R&D program (Advancement of an open cloud platform for smart maritime convergence service_P0015306).

References

Achmad, Z., Samuel, L., 1984. Stratigraphy and depositional cycles in the NE kalimantan basin. In: Proceedings of Indonesian Petroleum Association 13th Annual Convention, pp. 109–120.
Adoghe, L.I., Aniekwe, O.S., Nwosu, C., 2011. Improving electrofacies modeling using multivariate analysis techniques: a deepwater turbidite case study. In: Nigeria Annual International Conference and Exhibition. Society of Petroleum Engineers.
Airola, A., Pohjankukka, J., Torppa, J., Middleton, M., Nykänen, V., Heikkonen, J., Pahikkala, T., 2019. The spatial leave-pair-out cross-validation method for reliable AUC estimation of spatial classifiers. Data Min. Knowl. Discov. 33 (3), 730–747.
Akuanbatin, H., Rosandi, T., Samuel, L., 1984. Depositional environment of the hydrocarbon bearing Tabul, Santul, and tarakan formations at Bunyu island, NE kalimantan. In: Proceedings of Indonesian Petroleum Association 13th Annual Convention, pp. 425–442.
Al-Mudhafar, W.J., 2015. Integrating component analysis and classification techniques for comparative prediction of continuous and discrete lithofacies distributions. In: Offshore Technology Conference (Houston, Texas).
Al-Mudhafar, W.J., 2017. Integrating well log interpretations for lithofacies classification and permeability modeling through advanced machine learning algorithms. Journal of Petroleum Exploration and Production Technology 7 (4), 1023–1033.
Ameur-Zaimeche, O., Aziez, Z., Heddam, S., Kechiched, R., 2020. Lithofacies prediction in non-cored wells from the Sif Fatima oil field (Berkine basin, southern Algeria): a comparative study of multilayer perceptron neural network and cluster analysis-based approaches. J. Afr. Earth Sci. 166, 103826.
Avseth, P., Mukerji, T., 2002. Seismic lithofacies classification from well logs using statistical rock physics. Petrophysics 43, 02.
Avseth, P., Mukerji, T., Mavko, G., 2005. Quantitative seismic interpretation. In: Applying Rock Physics Tools to Reduce Interpretation Risk, Tools for Seismic Analysis in Porous Media. Cambridge University Press.
Baillie, P., Darman, H., Fraser, T.H., 2004. Deformation of Cenozoic Basins of Borneo and West Sulawesi, IPA-AAPG Proceedings of Deep-Water and Frontier Exploration in Asia and Australasia Symposium.
Baldwin, J.L., Bateman, R.M., Wheatley, C.L., 1990. Application of a neural network to the problem of mineral identification from well logs. Log. Anal. 31, 05.
Bhattacharya, S., Carr, T.R., Pal, M., 2016. Comparison of supervised and unsupervised approaches for mudstone lithofacies classification: case studies from the Bakken and Mahantango-Marcellus Shale, USA. J. Nat. Gas Sci. Eng. 33, 1119–1133.
Bhattacharya, S., Mishra, S., 2018. Applications of machine learning for facies and fracture prediction using Bayesian Network Theory and Random Forest: case studies from the Appalachian basin, USA. J. Petrol. Sci. Eng. 170, 1005–1017.
Bhattacharya, S., Carr, T.R., 2019. Integrated data-driven 3D shale lithofacies modeling of the Bakken Formation in the Williston basin, North Dakota, United States. J. Petrol. Sci. Eng. 177, 1072–1086.
Bressan, T.S., de Souza, M.K., Girelli, T.J., Junior, F.C., 2020. Evaluation of machine learning methods for lithology classification using geophysical data. Comput. Geosci. 104475.
Burollet, P.F., Salle, C., 1981. A contribution to the geological study of sumba (Indonesia). In: Proceedings of Indonesian Petroleum Association 10th Annual Convention.
Chen, Y., Wu, W., 2016. A prospecting cost-benefit strategy for mineral potential mapping based on ROC curve analysis. Ore Geol. Rev. 74, 26–38.
Darling, T., 2005. Well Logging and Formation Evaluation. Elsevier.
Darman, H., 2001. Turbidite plays of Indonesia: an overview. Berita Sedimentologi 15, 2–21.
Deng, Z., Zhu, X., Cheng, D., Zong, M., Zhang, S., 2016. Efficient kNN classification algorithm for big data. Neurocomputing 195, 143–148.
Dubois, M.K., Bohling, G.C., Chakrabarti, S., 2007. Comparison of four approaches to a rock facies classification problem. Comput. Geosci. 33 (5), 599–617.
Ellen, H., Husni, M.N., Sukanta, U., Abimanyu, R., Feriyanto, Herdiyan, T., 2008. Middle Miocene Meliat Formation in the tarakan island, regional implications for deep exploration opportunity. In: Proceedings of Indonesian Petroleum Association 32nd Annual Convention and Exhibition.
Ellis, D.V., Singer, J.M., 2007. Well Logging for Earth Scientists, vol. 692. Springer, Dordrecht.
Feng, R., 2021. Improving uncertainty analysis in well log classification by machine learning with a scaling algorithm. J. Petrol. Sci. Eng. 196, 107995.
Feng, R., Grana, D., Balling, N., 2021. Imputation of missing well log data by random forest and its uncertainty analysis. Comput. Geosci. 152, 104763.
Freund, Y., 1995. Boosting a weak learning algorithm by majority. Inf. Comput. 121 (2), 256–285.

Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm. icml 96, 148–156.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.
Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann. Stat. 28 (2), 337–407.
Fu, X., Qin, Y., Wang, G.G., Rudolph, V., 2009. Evaluation of coal structure and permeability with the aid of geophysical logging technology. Fuel 88 (11), 2278–2285.
Gajowniczek, K., Ząbkowski, T., Szupiluk, R., 2014. Estimating the roc curve and its significance for classification models' assessment. Quantit. Methods Econ 15 (2), 382–391.
Hamilton, W., 1979. Tectonic of Indonesian Region. Geological Survey, US. Professional Paper 1078.
Hosmer Jr., D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied Logistic Regression, vol. 398. John Wiley & Sons.
Hossin, M., Sulaiman, M.N., 2015. A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process 5 (2), 1.
Houston, W.M., Woodruff, D.J., 1997. Empirical bayes estimates of parameters from the logistic regression model. ACT Research Report Series 97 (6). https://eric.ed.gov/?id=ED414311.
Hsieh, B.Z., Lewis, C., Lin, Z.S., 2005. Lithology identification of aquifers from geophysical well logs and fuzzy logic analysis: shui-Lin Area, Taiwan. Comput. Geosci. 31 (3), 263–275.
Husein, S., 2017. Lithostratigraphy of Tabul formation and onshore geology of nunukan island, North Kalimantan. Journal of Applied Geology 2 (1), 25–35.
Imamverdiyev, Y., Sukhostat, L., 2019. Lithological facies classification using deep convolutional neural network. J. Petrol. Sci. Eng. 174, 216–228.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning, vol. 112. springer, New York, p. 18.
Jahdhami, N.A., Anboori, A.A., 2017. The application of specific drilling Energy to identify overburden lithological boundaries and aid well operations-Oman khazzan field. In: Abu Dhabi International Petroleum Exhibition & Conference. Society of Petroleum Engineers.
Jing, S., Liu, C., Li, G., Yan, G., Zhang, Y., 2017. December. An efficient algorithm for parallel computation of rough entropy using cuda. In: 2017 13th International Conference on Computational Intelligence and Security (CIS). IEEE, pp. 1–5.
Johnson, R., Zhang, T., 2013. Learning nonlinear functions using regularized greedy forest. IEEE Trans. Pattern Anal. Mach. Intell. 36 (5), 942–954.
Korjus, K., Hebart, M.N., Vicente, R., 2016. An efficient data partitioning to improve classification performance while keeping parameters interpretable. PloS One 11 (8), e0161788.
Lentini, M.R., Darman, H., 1996. Aspects of the Neogene tectonic history and hydrocarbon geology of the Tarakan basin. In: Proceedings of Indonesian Petroleum Association, 25th Silver Anniversary Convention, pp. 241–251.
Maimon, O., Rokach, L., 2010. Data Mining and Knowledge Discovery Handbook, second ed.s. Springer, New York City, U.S.
Male, F., Duncan, I.J., 2020. Lessons for machine learning from the analysis of porosity-permeability transforms for carbonate reservoirs. J. Petrol. Sci. Eng. 187, 106825.
Maria Navin, J.R., Pankaja, R., 2016. Performance Analysis of Text Classification Algorithms Using Confusion Matrix.
Neeb, H., Kurrus, C., 2016. Distributed K-Nearest Neighbors.
Noon, S., Harrington, J., Darman, H., 2003. The Tarakan basin, east kalimantan: proven Neogene fluvio-deltaic, prospective deep-water and paleogene plays in a regional stratigraphic context. Proceedings of Indonesian Petroleum Association 29th Annual Convention and Exhibition 1, 1–14. https://archives.datapages.com/data/ipa/data/029/029001/1_ipa029aipa03-g-136.htm.
Qi, L., Carr, T.R., 2006. Neural network prediction of carbonate lithofacies from well logs, Big Bow and Sand Arroyo Creek fields, Southwest Kansas. Comput. Geosci. 32 (7), 947–964.
Raschka, S., 2015. Python Machine Learning. Packt Publishing Ltd, England, p. 454pp.
Rogers, S.J., Fang, J.H., Karr, C.L., Stanley, D.A., 1992. Determination of lithology from well logs using a neural network. AAPG Bull. 76 (5), 731–739.
Saputra, I., Wibisono, T., 2016. Strike-slip fault geometry and its significance for petroleum play in Tarakan basin: a perspective from onshore simenggaris area. In: Proceedings of Indonesian Petroleum Association 40th Annual Convention and Exhibition.
Schapire, R.E., 1990. The strength of weak learnability. Mach. Learn. 5 (2), 197–227.
Situmorang, B., 1982. the formation of the makassar basin as determined from subsidence curve. Proceedings of Indonesian Petroleum Association 11th Annual Convention, pp. 83–107.
Situmorang, B., 1983. Formation, Evolution, and Hydrocarbon Prospect of the Makassar Basin. Transaction of 3rd Circum Pacific Conference, Indonesia, pp. 227–232.
Storkey, A., 2009. When training and test sets are different: characterizing learning transfer. Dataset shift in machine learning 3–28.
Tang, H., White, C., Zeng, X., Gani, M., Bhattacharya, J., 2004. Comparison of multivariate statistical algorithms for wireline log facies classification. In: AAPG Annual Meeting Abstract, vol. 88, p. 13.
Tilaki-Hajian, K., 2013. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian journal of internal medicine 4 (2), 627.
Tharwat, A., 2018. Classification assessment methods. Appl Comput Inform 17 (1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003.
Vakhshoori, V., Zare, M., 2018. Is the ROC curve a reliable tool to compare the validity of landslide susceptibility maps? Geomatics, Nat. Hazards Risk 9 (1), 249–266.
Vapnik, V.N., 1998. Statistical Learning Theory. A Wiley—Interscience Publication, New York.
Wang, D., Yu, Y., Qin, S., Zhu, Q., 2003. A summary of the development of geophysical logging techniques for the coalbed methane reservoir. Acta Geosci. Sin. 24, 385–390 (4; ISSU 77).
Wang, G., Carr, T.R., Ju, Y., Li, C., 2014. Identifying organic-rich Marcellus Shale lithofacies by support vector machine classifier in the Appalachian basin. Comput. Geosci. 64, 52–60.
Wong, P.M., Jian, F.X., Taggart, I.J., 1995. A critical comparison of neural networks and discriminant analysis in lithofacies, porosity and permeability predictions. J. Petrol. Geol. 18 (2), 191–206.
Wood, D.A., 2019. Lithofacies and stratigraphy prediction methodology exploiting an optimized nearest-neighbour algorithm to mine well-log data. Mar. Petrol. Geol. 110, 347–367.
Yong, Z., Youwen, L., Shixiong, X., 2009. An improved KNN text classification algorithm based on clustering. J. Comput. 4 (3), 230–237.
Zhao, L.N., Tian, F.Y., Wu, H., Qi, D., Di, J.Y., Wang, Z., 2011. Verification and comparison of probabilistic precipitation forecasts using the TIGGE data in the upriver of Huaihe Basin. Advances in Geosciences 29, 95.