Geomech. Geophys. Geo-energ. Geo-resour. (2024) 10:70
https://doi.org/10.1007/s40948-024-00787-5

CORRESPONDENCE

Data-driven lithofacies prediction in complex tight sandstone reservoirs: a supervised workflow integrating clustering and classification models

Muhammad Ali · Peimin Zhu · Ren Jiang · Ma Huolin · Umar Ashraf · Hao Zhang · Wakeel Hussain

Received: 4 January 2024 / Accepted: 2 April 2024

Abstract Lithofacies identification plays a pivotal role in understanding reservoir heterogeneity and optimizing production in tight sandstone reservoirs. In this study, we propose a novel supervised workflow aimed at accurately predicting lithofacies in complex and heterogeneous reservoirs with intercalated facies. The objectives of this study are to utilize advanced clustering techniques for facies identification and to evaluate the performance of various classification models for lithofacies prediction. Our methodology involves a two-information criteria clustering approach, revealing six distinct lithofacies and offering an unbiased alternative to conventional manual methods. Subsequently, Gaussian Process Classification (GPC), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Random Forest (RF) models are employed for lithofacies prediction. Results indicate that GPC outperforms other models in lithofacies identification, with SVM and ANN following suit, while RF exhibits comparatively lower performance. Validated against a testing dataset, the GPC model demonstrates accurate lithofacies prediction, supported by synchronization measures for synthetic log prediction. Furthermore, the integration of predicted lithofacies into acoustic impedance versus velocity ratio cross-plots enables the generation of 2D probability density functions. These functions, in conjunction with depth data, are then utilized to predict synthetic gamma-ray log responses using a neural network approach. The predicted gamma-ray logs exhibit strong agreement with measured data ($R^2 = 0.978$) and closely match average log trends. Additionally, inverted impedance and velocity ratio volumes are employed for lithofacies classification, resulting in a facies prediction volume that correlates well with lithofacies classification at well sites, even in the absence of core data. This study provides a novel methodological framework for reservoir characterization in the petroleum industry.

Keywords Facies classification · Machine learning · Well log · Two-information criteria clustering technique

M. Ali · P. Zhu (✉) · Ma. Huolin · H. Zhang · W. Hussain
Institute of Geophysics and Geomatics, China University of Geosciences, Wuhan, People's Republic of China
e-mail: zhupm@cug.edu.cn

R. Jiang
Research Institute of Petroleum Exploration and Development, Petro-China Company Limited, Beijing, People's Republic of China

U. Ashraf
School of Ecology and Environmental Sciences, Yunnan University, Kunming, People's Republic of China

U. Ashraf
Institute of International Rivers and Eco-Security, Yunnan University, Kunming 650500, China

Published online: 09 April 2024

1 Introduction

Machine learning (ML) is a branch of artificial intelligence (AI) that utilizes data analysis techniques such as classification, regression, and clustering to make predictions and identify patterns in large datasets (Ehsan and Gu 2020; Ashraf et al. 2024). The ML approach can be classified into two groups: supervised and unsupervised. Supervised ML involves using input parameters and desired outputs to train a model, while unsupervised ML identifies patterns without predefined outputs. In the oil and natural gas industry, machine learning has become a popular tool for solving geoscientific problems related to exploration, development, and production. Wire-line logs have become a commonly used tool for geoscientists in the oil and gas industry (Anees et al. 2022; Ali et al. 2024). With the development of machine learning, various neural networks have been widely used in oil exploration (Antariksa et al. 2022; Song, et al. 2021; Valentín et al. 2019). Chawshin et al. (2021) designed a convolutional neural network (CNN), that automatically predicted lithofacies from 2D core CT scan image slices. Alzubaidi et al. (2021) introduced a CNN-based method that utilized core images for automatic lithology prediction, although it exhibited poor performance in the subdivision of rock types. Al-Mudhafar et al. (2022) developed a novel technique using boosting algorithms to classify lithofacies in carbonate reservoirs, specifically in the Majnoon oil field in Iraq. They compared five machine learning algorithms and achieved high accuracy. Lithofacies predictions were validated against core data and compared with poro-perm interpretations, aiding in reservoir characterization and production optimization. Moghanloo et al., (2018) conducted pre-stack inversion analysis to extract P-wave and S-wave information from seismic data, offering advantages over post-stack inversion for reservoir fluid characterization. They determined key parameters, such as k, KC, m, and mC, and developed angle-dependent wavelets to derive acoustic impedance, shear impedance, and density sections. This method was applied successfully in identifying reservoir facies in an Iranian hydrocarbon field. Ghanbarnejad Moghanloo and Riahi (2023) developed an integrated workflow using recent geoscience data to assess reservoir characteristics and structural interpretation of the Burgan formation in SW Iran. They employed high-resolution SEM images for pore analysis, utilized a watershed segmentation algorithm, calibrated porosity logs, and employed supervised Bayesian classifiers for facies prediction. The approach was validated with seismic data, aiming to optimize drilling operations in similar geological settings. Zhang et al. (2021) employed convolutional neural networks to identify lithofacies from core images. Although these methods notably reduced the identification time, they still necessitated a considerable number of core sample images for network training and labeling, posing a challenge. To overcome this challenge, Zhang et al. (2021) used relatively low-cost well logs instead of core samples for lithofacies identification. This approach effectively provided a first-glance analysis of core data, even though the model's generalization required improvements. Despite the limitations, machine-led applications hold great promise in the oil and natural gas industry, enabling more efficient and accurate exploration and production.

Log data is widely used in lithofacies identification and evaluation due to its high vertical resolution and good continuity (Lai et al. 2018). The composition and structure of the reservoir will result in the division of various lithofacies, each exhibiting different logging responses (Hemmesch et al. 2014; Ozkan et al. 2011). Therefore, Bhattacharya et al. (2016) input five one-dimensional logs and other derived parameters into the lithofacies model using three machine learning algorithms, such as Artificial Neural Network (*ANN*), and proved that lithofacies identification could be modeled in that way. Similarly, Wu et al. selected deep resistivity (*LLD*), spontaneous potential (*SP*), Gamma (*GR*), Sonic (*DT*), Neutron (*NPHI*), and Density (*RHOB*) to summarize the logging response characteristics of five lithofacies based on the experimental results of core composition analysis, and successfully predicted the distribution of each lithofacies in a single well (Wu et al. 2020). He et al. (2016) optimized the identification model constructed by *DEN, AC, RT,* and other logs through the comparison of core observation, X-ray diffraction, and qualitatively identified lithofacies through the intersection diagram (He et al. 2016). Compared with *GR DEN, DT,* and other one-dimensional logs, resistivity images can directly observe formation changes and identify lithofacies boundaries. Its appearance improves the accuracy of lithofacies identification. Zhang and Pan (2011) utilized the Support

Vector Machine (SVM) algorithm to process conventional logs, such as natural gamma and photoelectric absorption cross-section index. They elucidated their findings using micro-resistivity images and ultimately conducted a more comprehensive analysis of the volcanic lithofacies (Zhang and Pan 2011). However the *SVM* is difficult to achieve large-scale training samples, and the neural network is easy to fall into the local optimum (LeCun et al. 2015). Therefore, Yu et al. (2021) established a lithology identification and classification model using the gradient boosting decision tree (GBDT) ensemble learning algorithm. The model correctly identified the lithofacies of the volcanic rocks using core and FMI-calibrated conventional logs as input (Yu et al. 2021). On this basis, to further improve the efficiency and accuracy of identification, research has been carried out. Lan et al. (2021) a positive and unlabeled machine learning (PU-learning) technique was developed for traditional log data, which only marked restricted log samples, and five carbonate logging lithofacies were effectively created; nevertheless, the accuracy of the results required to be improved (Yu et al. 2021). However, most of the above lithofacies identification methods require certain a priori judgment results for guidance. This kind of method is greatly influenced by manual subjective and has low precision and a huge workload. Therefore, it is necessary to identify lithofacies automatically. Tian et al. (2016) used the multi-resolution graph-based clustering (MRGC) method to automatically cluster the log of the Amu Darya basin without prior knowledge and finally obtained different lithofacies (Tian et al. 2016). Chai et al. designed an automatic lithofacies classification method for sedimentary facies of reef-shoal reservoirs (Chai et al. 2009). These above researchers once again proved the trend of design research of lithofacies automatic classification and identification method.

Therefore, the goal of this research is to investigate how supervised classification can improve the recognition of lithological facies in a dataset. The dataset consists of complex geometrically pro-gradational sequence environments that were formed during a significant tectonic event. This event not only affects the distribution of different facies but also leads to the presence of a large volume of volcaniclastic debris, which can negatively impact the quality of the reservoir. Accurately predicting facies is important because the Lower Goru formation in the Kadanwari gas field has the potential to produce significant amounts of natural gas. The process of facies categorization involves creating new logs, generating synthetic Gamma ray logs using machine learning regression models, and creating artificial data to fill in gaps in the dataset. The final step is to train the data samples using four different classification algorithms and select the most accurate facies classifier from the validation dataset. This study uses "ensemble learning," which combines multiple models to improve accuracy.

2 Geological setting and data analysis

In this section, we provide a comprehensive overview of the geological setting and data analysis for the Kadanwari and Sawan gas fields in the Lower Goru formation within the Central Indus Basin, Pakistan (Fig. 1a). The section is structured as follows:

2.1 Geographical and geological description

The study area encompasses the Kadanwari and Sawan gas fields, focusing on the conventional sands in D, E, and F, and the tight G sand layer. This region, situated in the Central Indus Basin, is characterized by a complex geometrically progradational sequence environment, formed during three significant tectonic events (Ahmad and Chaudhry 2002)–(Ali et al. 2020). The structural configuration of the field, shaped by these tectonic events, has a significant impact on the reservoir characteristics (Fig. 1a). The Lower Goru sands have been divided into seven sand-bearing intervals (Fig. 1b) from bottom B-Sand to top H-Sand (Ahmad and Chaudhry 2002). The primary producing sands in the area are E-Sand and G-Sand, while D-Sand and F-Sand have also yielded production from select wells (Ali et al. 2019, 2020, 2023). In Kadanwari, E-Sand, the main producer, is characterized as a conventional reservoir, forming an elongated body trending SW-NE parallel to the paleo shoreline of the Early Cretaceous time. However, B, C, D, G, and H exhibit tight characteristics. G-Sand has been productive post-hydraulic fracturing, and F-Sand exhibits hot sand characteristics in the field (Ashraf, et al. 2018; Ashraf et al. Oct. 2020).

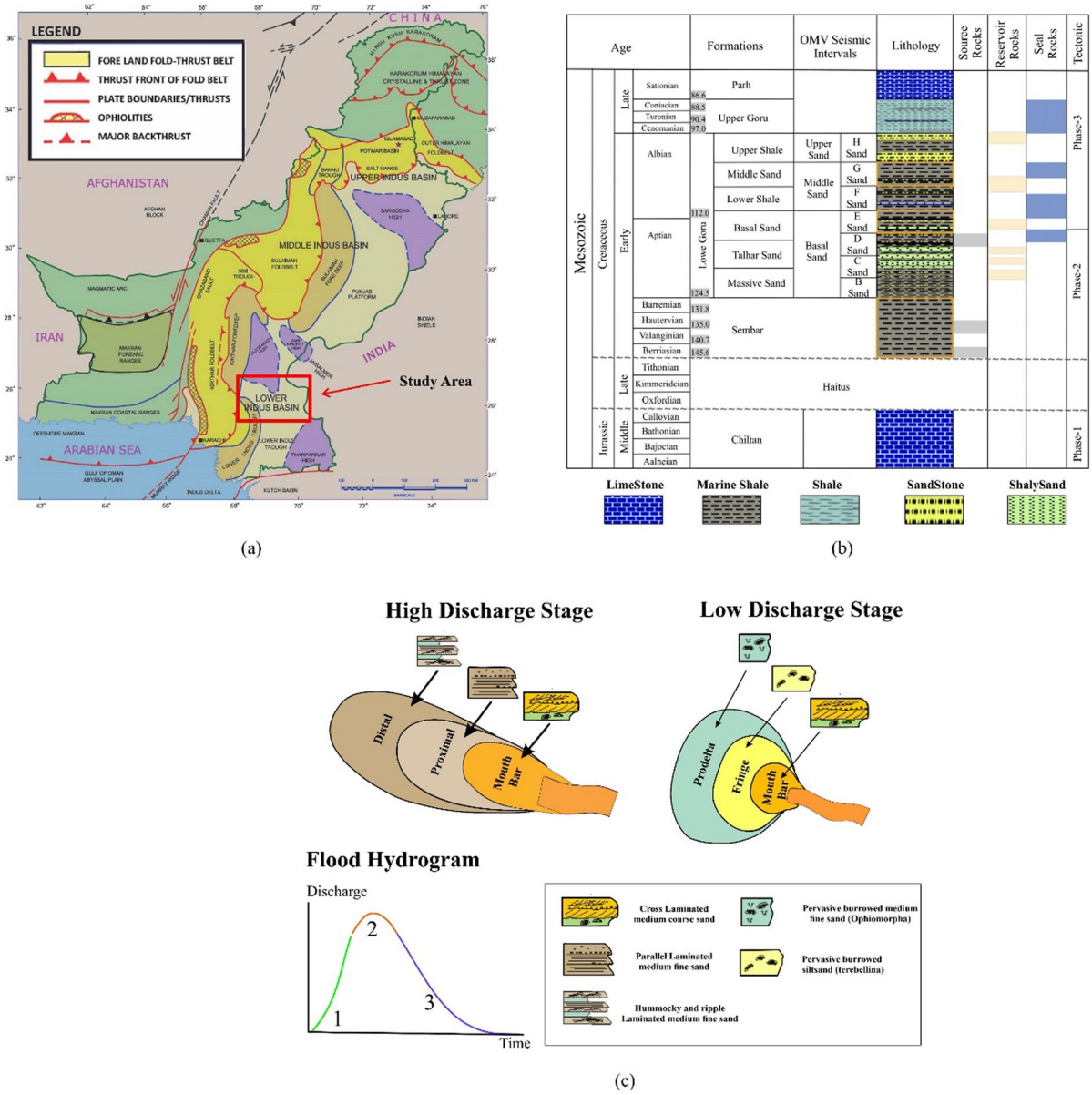

Fig. 1 Presents a clear visual representation, highlighting **a** the geographical positioning, **b** the lithological composition within the study region, and **c** the sedimentology model

2.2 Deltaic system characteristics

The Kadanwari and Sawan (from C to H layers) of the Lower Goru represent a clastic delta system characterized by a river-dominant regime with additional wave and tidal transformations. River dynamics leave their mark on both sand-prone "proximal" and fine-grained "distal" facies. Proximal facies exhibit cross-bedded medium to coarse sandstones, while distal facies are typified by hummocky cross-lamination, associated with hyperpycnal flow during massive seasonal storms and floods (Valzania, et al. 2011). Distinctive variations in the size and

shape of delta lobes deposited at different stages are evident (Fig. 1c).

The selection of the gas field stems from its complex geological features, particularly within the Lower Goru formation of the Central Indus Basin, Pakistan. The structural configuration, shaped by significant tectonic events, presents an ideal scenario for testing the efficacy of our proposed model. However, the data from this field indeed poses challenges, especially when conventional statistical approaches struggle to differentiate between facies solely based on logging data. This complexity underscores the need for more advanced analytical techniques, such as machine learning, to extract nuanced geological features from multivariate datasets effectively.

3 Data and methodology

This study utilized a dataset comprising wireline log measurements of six parameters, namely gamma ray (*GR*), laterolog deep resistivity (*LLD*), neutron porosity (*NPHI*), compressional wave velocity (*DT*), Photoelectric Effect (PEF), and bulk density (*RHOB*), as well as two petrophysical parameters estimated from the data: volume of shale and porosity. The wells are in the Lower Goru formation of the Kadanwari gas field block, which is situated in the central Indus Basin (Fig. 2).

K-15 and K-14 are wells that provide facies logs and facies descriptions derived from geological data, which we utilized as the targeted output for the

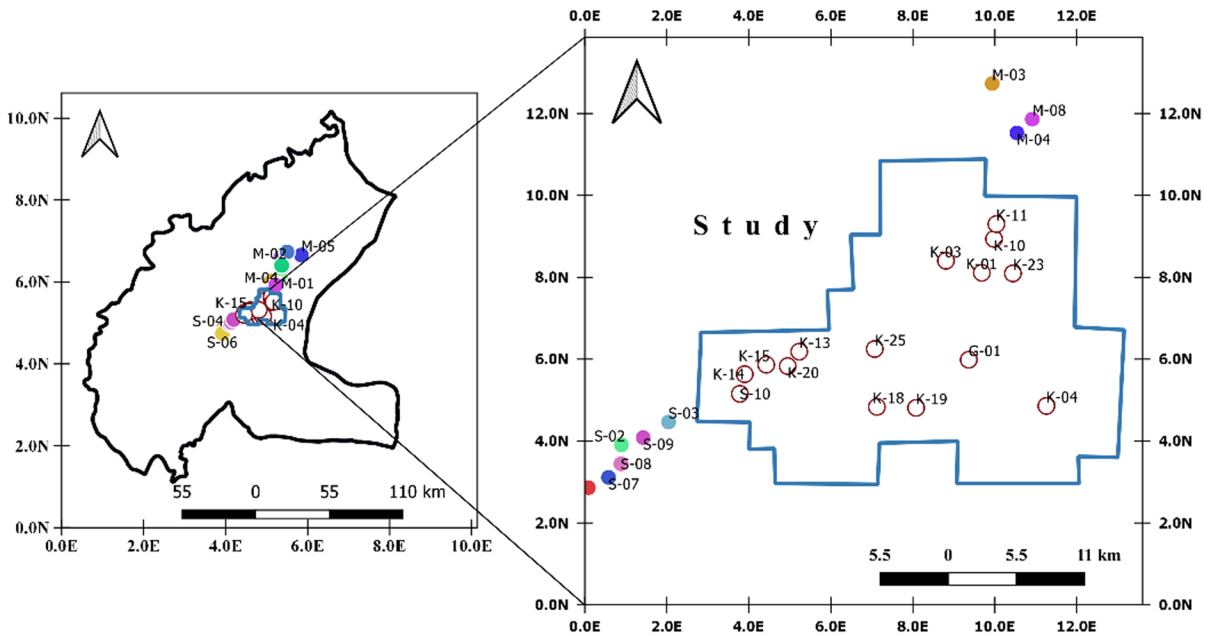

Fig. 2 The location of the wells is in the Lower Goru formation of the Kadanwari gas field block, situated in the central Indus Basin

Table 1 Facies classification of Lower Goru

Facies	Description
Sh	Shale to silty (Shelf deposits)
Slst	Silstone to silty-shaly sandstone (prodelta shales with turbiditic layers)
Css	Low-porosity, low permeability cemented sandstone (very distal mouth bar fringe)
Lss	Low-medium porosity, low permeability sideritic/chamositic sandstoe (shoreface to distal mouth bar)
Ss	High-porosity, high permeability sandstone (mouth bar)
Hs	Highly chamositic/siderite affected lithologies (chamositized mouth bar)

machine learning (ML) algorithms in our study. The lithology of the research area is classified into six categories, with each category identified through meticulous petrophysical analysis and core investigation. Table 1 contains the data and serves as a useful representation of the fluvial-based depositional system, using the mentioned facies' nomenclature.

For any machine learning approach to be effective data analysis and statistical representation of samples are essential, which includes visualizing the input–output correlation function. Figure 3 presents a flowchart outlining the process of generating synthetic GR logs from predicted rock facies and constructing a facies model.

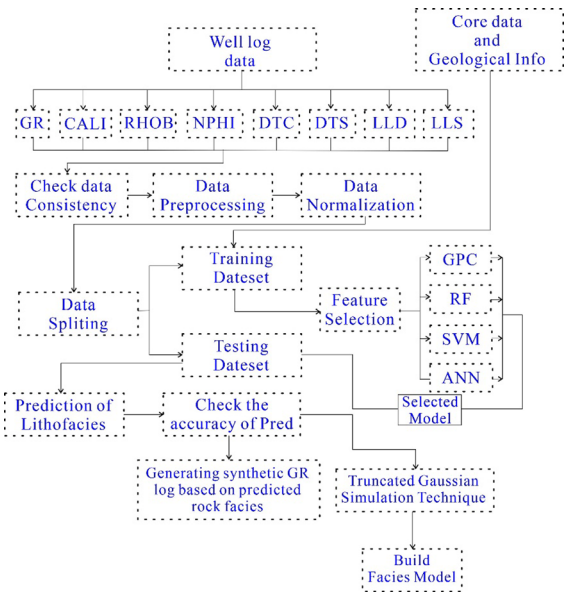

Fig. 3 Flowchart illustrating the method for generating synthetic GR logs based on predicted rock facies and constructing a facies model

3.1 Data cleaning

The Pauta criteria method is used to identify anomalies in logging data. The logging data often has imperfections resulting in significant deviations from the normal distribution (Zheng et al. 2021). However, the logging data still follows a standard distribution with some minor deviations (Zheng et al. 2021). To identify these anomalies, the Pauta criterion method is utilized. This method uses a formula that involves three standard deviations to determine the conviction probability of determining gross error, which is 99.7%. If a value is outside the range of uncertainty, it cannot be considered a mere statistical error but rather a significant one, and hence, it is considered a gross error. The Pauta criteria method is depicted in Fig. 4, and it is used to identify these gross errors in the logging data. Once these gross errors are identified, this study uses Lagrange interpolation to fill these outliers (Li et al. 2016).

3.2 Clusters selection criteria

The K-15 well in the study area has limited core facies data, as depicted in Fig. 5. Therefore, it is crucial to select an appropriate number of clusters that accurately represent the facies in the study area before utilizing machine learning algorithms for prediction. Overfitting or underfitting can introduce bias into the model, so two information criteria have been employed to maximize the number of clusters that best fit the data in the study area. The first criterion

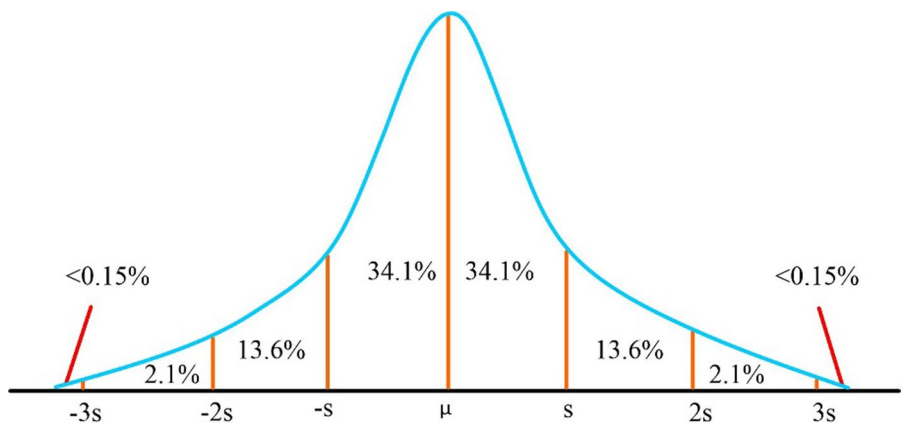

Fig. 4 Illustration of the Pauta criteria method utilized to identify gross errors in logging data, employing a formula based on three standard deviations to determine anomalies

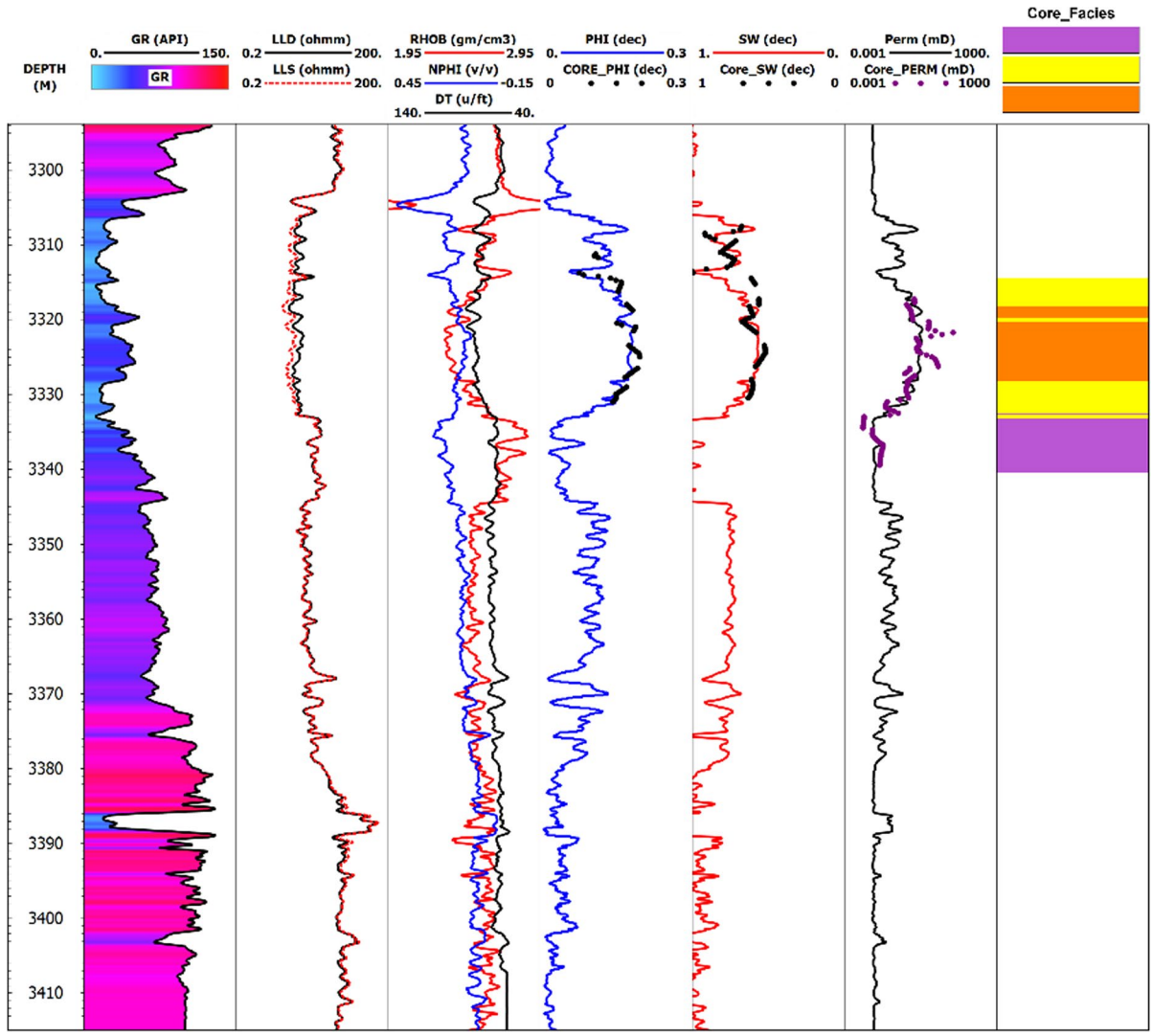

Fig. 5 Visualization of core facies data from a section of well K-15, indicating limited data availability

is the Akaike Information Criterion (AIC) (Akaike 1974), which assesses the probability of a model accurately predicting or estimating future values based on in-sample fit. A model with the lowest AIC is deemed to be the most appropriate. AIC is also helpful in selecting between the multiplicative and additive Holt–Winters models. Another criterion is the Bayesian Information Criterion (BIC) (Stone and Javid 1979), which balances the trade-off between model complexity and fit. Lower AIC or BIC values indicate a better fit. The study provides the equations to calculate the AIC and BIC values of a model.

$$AIC = -2 * ln(L) + 2 * k \tag{1}$$

$$BIC = -2 * ln(L) + 2 * ln(N) * N \tag{2}$$

where k refers to the number of parameters estimated, N denotes the recorded measurements and L is the likelihood value (Fig. 5).

3.3 One-class support vector machine (SVM)

The method is based on Soentpiet's (1999) extension of the original SVM algorithm developed by

Cortes and Vapnik (1995). Since their introduction, the SVMs have proven to be popular due to their good performance in capturing complex decision boundaries for supervised classification. In contrast to traditional SVM classification problems, where there may be multiple classes, the one-class extension is an unsupervised approach that seeks to trace the boundary of the training data in multidimensional space. What both methods have in common is the key idea of choosing a subset of training data samples, the so-called support vectors to define the decision boundary. The support vectors are chosen by an optimization routine, for further details see (Hastie et al. 2019). A vector of facies measurements at support vector i is denoted by x_i, and the optimization routine also provides weights α_i associated with each of the N support vectors. The support vectors define the SVM score for any test vector x as follows:

$$f(x) = \sum_{i=1}^{N} a_i K\left(x_i, x\right) + b \tag{3}$$

where b is a bias term that depends on the outlier fraction, and K is a Gaussian kernel function:

$$K\left(x_i, x\right) = e^{-\|x_i - x_j\|^{2/\sigma}} \tag{4}$$

Based on the SVM scores, we can define a decision function that provides a classification of a test vector x, as follows:

$$g(x) = sgn\left(\sum_{i=1}^{N} a_i K\left(x_i, x\right) + b\right) \tag{5}$$

The decision function equals 1 when the test vector is inside the boundary of the training data and the SVM score is positive; in this case, the test vector is classified as a normal measurement. The decision function equals -1 when the SVM score is negative, and the test vector is classified as an outlier (outside of the training data distribution). The decision function effectively partitions the multi-dimensional space of clusters into normal and outlier regions separated by a boundary, which we term the clusters of a training dataset. A well may have multiple clusters when partitioning the available logs into related groups such as quad combo, etc. A well can also have multiple clusters formed by partitioning its logs by stratigraphic zone, facies, or fluid type.

3.4 Gaussian process classifiers (GPCs)

Gaussian process classifiers (MGPCs) are a type of Bayesian method that is highly effective for non-parametric multiclass classification. This method produces probabilistic outputs that are useful for quantifying prediction uncertainty. GPCs are unique in that they are purely statistical models derived from Gaussian processes used in regression (Berczi et al. 2015; Gibbs and MacKay 2000). In GPCs, the value of a latent function at a given input is closely related to the likelihood of belonging to a particular class. Inference in MGPCs involves inferring a posterior over the latent function and the hyperparameters that govern it, given a prior over the latent function and the observed data. However, performing accurate Bayesian inference in MGPCs is often computationally prohibitive, making it necessary to use approximations such as Markov-chain Monte Carlo sampling, the Laplace approximation, or expectation propagation to achieve efficient and scalable solutions for Gaussian process classification. These approximations allow for the practical use of Gaussian process classification techniques in real-world applications.

Based on the GPC scores we can define an approximate inference that provides a classification of a test vector y*, as follows:

$$p(y* = +1/y) = Zdf * \sigma(f *)p(f * /y) \tag{6}$$

3.5 Random forest (RF)

Random forest is a popular machine-learning algorithm developed by Breiman (2001) and has been successfully used in various classification and regression problems (Akkurt et al. 2018). The algorithm works by creating a large number of decision trees from bootstrap samples of the training data. At each node of a tree, a random selection of variables is made to split on, and only a random subset of predictors is considered for splitting. The size of this subset is one of the few tuning parameters of the algorithm, along with the minimum number of samples in each node of the tree (Granitto et al. 2007). A significant advantage of random forests is their ability to perform well without extensive tuning of these parameters. The algorithm creates N decision trees using bootstrap sampling, and the values of each tree are aggregated

to obtain a final prediction. The figure below shows the basic structure of the random forest algorithm.

$$prob\left(y = y_k | x = x\prime\right) = \frac{1}{N}\sum_{n=1}^{N} I(\tau_n(x\prime) = y_k) \tag{7}$$

The building of the *mth* decision tree $\tau_n(x)$ is based on the S_m random sampling subset. The RF model $F(x)$ integrates the independent creation in parallel N base trees. The prediction of sample x_0 demonstrates that the predictions of N trees are initially established. The predicted value is the conditional probability for the kth class of X_0.

$$F(x\prime) = argmaxprob\left(y = y_k | x = x\prime\right) \tag{8}$$

The bagging algorithms have been demonstrated utilizing a variety of performance models, as well as the ensembles and robustness of group models. The unique form of suitcase trees that RF receives from nature. Due to the optimization of more random nodes, the variety is greatly increased. The RF is more accurate due to this characteristic than standard bagging trees without randomization.

3.6 Artificial neural networks (ANN)

The term 'artificial neural networks' refers to a category of numerical optimization algorithms that were initially conceived through research into the human brain and nervous system (Haykin 2011; Guresen and Kayakutlu 2011). Artificial neural networks (ANNs) are a type of nonlinear dynamical system that, through training, can improve their pattern recognition abilities. Training an ANN involves feeding it a series of inputs and target outcomes (training patterns). In practice, the ANN learns its complex predictive function during training. After initial training, the ANN can be used to predict the values of its output variables using trained data as input (Haykin 2011; Guresen and Kayakutlu 2011).

Assume neural networks with a hidden signal and an input layer n and output layer m, b j indicates the output of the hidden signals, θ_j is the value of the hidden layer's threshold, the value θ_k represents the threshold for the output signal, f_1 is indicated the transfer factor of the hidden signal, while represents the transfer function of the output signal, input layer to hidden layer weights of w_{ij}, while hidden layer to output layer weights w_{jk}. Subsequently, we can obtain the output of the network, which is denoted by y_k, while the output of the neuron of the hidden layer is denoted by t_k.

$$b_j = f_1\left(\sum_{i=1}^{n} w_{ij}x_i - \theta_j\right)(i = 1, 2, ..., n; j = 1, 2, ..., s) \tag{9}$$

Calculating the output of the output layer is:

$$y_k = f_2\left(\sum_{j=1}^{s} w_{jk}b_j - \theta_k\right)(j = 1, 2, ..., s; k = 1, 2, ..., m) \tag{10}$$

Defining the error function by the network's actual output, that is:

$$e = \sum_{k=1}^{m}\left(t_k - y_k\right)^2 \tag{11}$$

The purpose of network training is to decrease network error to a predetermined minimum or stop at a specific training step by continuously adjusting the weights and threshold. The prediction samples are then entered into the trained network, and the findings of the prediction are obtained.

3.7 Data split and cross-validation

In the field of machine learning, data partitioning is a crucial step to ensure that machine learning models are evaluated on previously unseen data (Thanh et al. 2022). In particular, data is typically split into a training set and a testing set (Alghazal and Krinis 2021), with the former used for training the models and the latter reserved for the final evaluation.

For small datasets, in particular, overfitting the training data is a common concern. To address this issue, a technique called cross-validation is often employed. During cross-validation, the training set is divided into several folds, and the machine learning estimator is sequentially evaluated on each fold while being trained on the remaining folds. This approach ensures that the models are evaluated on multiple partitions of the training data reducing the risk of overfitting.

At the end of the cross-validation process, the model with the best cross-validation score is selected and used to predict the testing set for final evaluation.

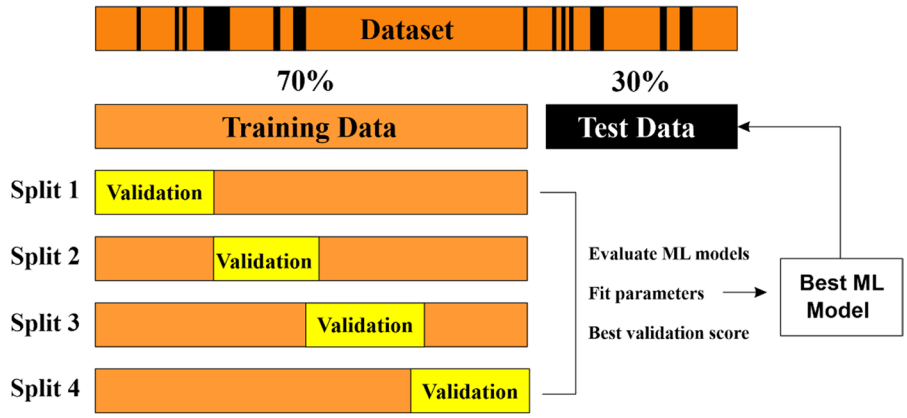

Fig. 6 Illustration of the division of the dataset into two sets: the first set is for training, and the second set is for testing the machine learning model

In this study, the available dataset consisted of multiple wells, which were randomly split into a 70% training set and a 30% testing set. To prevent overfitting and evaluate multiple candidate machine-learning models, K-fold cross-validation was used, with a four-fold size, chosen to resemble the number of samples in the final evaluation testing set. Figure 6 illustrates the overall machine-learning workflow and the adopted K-fold cross-validation technique.

3.8 Feature selection

Feature selection is a crucial pre-processing technique frequently used to minimize dataset dimensionality by systematically deleting pointless features from a set of available features. The use of feature selection in a machine-learning workflow has numerous advantages (Ali et al. 2021; Li et al. 2017; Guyon and Elisseeff 2003). These include reducing the model's complexity, which improves knowledge of the processes that produced the predictions, shortening model training timeframes, which reduces the computing cost associated with modeling, decreasing the risk of model overfitting, avoiding the "curse of dimensionality", and minimizing the effects of "garbage in, garbage out".

The univariate and Pearson's Correlation (r) feature selection techniques were selected for this latest study. The best features are found via univariate feature selection, which is based on univariate linear regression tests. This approach examines each feature independently and determines how it relates to the goal feature rather than considering all the features at once. A statistical method known as Pearson's Correlation assesses the strength of a linear relationship between two variables (a and b). The approach seeks to find the line of greatest fit through the data, with values ranging from -1 to $+1$. A value of 0 indicates that there is no correlation between the variables. A negative correlation is shown by a value between -1 and 0, whereas a positive correlation is indicated by a value between 0 and 1.

$$Pearson's\ correlation(r) = \frac{n\left(\Sigma_{ab}\right) - \left(\Sigma_a\right)\left(\Sigma_b\right)}{\sqrt{\left[n\Sigma a^2 - \left(\Sigma_a\right)^2\right]\left[n\Sigma b^2 - \left(\Sigma_b\right)^2\right]}} \tag{12}$$

3.9 Criteria for verifying model performance

The classification performance of the models was verified using a standardized confusion matrix, which detailed statistical results for both correctly and incorrectly classified lithofacies (Alghazal and Krinis 2021). Precision, recall, and F − 1 scores were all employed as verification metrics (Eqs. (13)–(15)).

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \tag{13}$$

$$Recall = \frac{True\ positive}{True\ positive + False\ negative} \tag{14}$$

$$\frac{1}{F_1} = \frac{1}{2}\left(\frac{1}{Precision} + \frac{1}{Recall}\right) \tag{15}$$

Additionally, the ROC AUC curve serves as a metric to evaluate the performance of classification tasks across various threshold settings. The ROC (Receiver

Operating Characteristic Curve) illustrates the relationship between sensitivity and specificity, while the AUC (Area Under the Curve) quantifies the separability between classes (Meyer-Baese and Schmid 2014). Essentially, it indicates the model's ability to differentiate between classes. A high AUC value, approaching 1, suggests excellent separability, whereas a low value near 0 indicates poor separability. An AUC of 0.5 implies that the model lacks the capacity to distinguish between classes altogether. TPR (16) and FPR (17) are categorized into two groups:

$$TPR = \frac{True\ positive}{True\ positive + False\ negative} \tag{16}$$

$$FPR = \frac{False\ positive}{False\ positive + True\ negative} \tag{17}$$

4 Result and discussion

This section contains multiple parts. The initial part explains the statistical summary of all variables and the correlation between them. The second part computes the number of clusters in the study area using a novel clustering algorithm as mentioned in Sect. 3.2. In the third part, feature selection techniques are employed to reduce the dimensionality of the dataset by systematically removing irrelevant features and correlating them with the target variable. We normalize each feature using the mean operator to ensure uniform scaling of input characteristics in machine learning applications, which aids in the convergence of the algorithm more quickly. In the fourth part, the results of each model's lithology identification are compared after establishing the model parameters. In the final step, the efficiency of the algorithm is examined, and finally, the final model is utilized to predict the facies of multiple/blind wells and check the prediction accuracy of facies similarity based on synchronization measures to predict synthetic logs.

4.1 Exploratory data analysis

A successful machine learning project must begin with the essential step of data exploration. Data exploration helps us to detect anomalies and patterns in the data at the initial investigation stage. A few summary statistics and graphical visualizations have been generated to better understand the input data (Table 2). The table contains descriptive statistics that summarize the count, mean, standard deviation, minimum (min), and maximum (max) as well as 25, 50, and 75 percentiles of the data except for undefined data points. It is evident that the data count of the LLS feature is lower than others because of some undefined values, and hence all feature data corresponding to these undefined points have been dropped for any future analysis. The training data hence comprises 720 data points only.

To analyze the bivariate measure of association between all the features, we have generated a heatmap of their correlation coefficients in Fig. 7. Among the feature pairs, GR-LLD, GR-NPHI, GR-DT, NPHI-PEF, NPHI-DT, NPHI-LLD, NPHI-GR, RHOB-SP, LLD-DT, LLD-NPHI, LLD-GR, and SP-RHOB have a high magnitude of correlation. Any pair with a very high correlation would have helped us identify feature redundancy and hence decrease the feature dimension to improve classification time.

Table 2 Statistical summary of the dataset

	GR	NPHI	RHOB	DT	PEF	LLD	CALI	LLS
Count	794	794	794	794	794	794	794	720
Mean	69.56371	0.163429	2.570256	71.42107	6.584696	12.67923	5.852819	11.3372
Std	31.38799	0.058181	0.115976	6.310227	2.810509	13.65903	0.336832	10.8894
Min	11.189	0.0166	2.3211	60.567	-0.9673	2.2065	5.011	2.20018
25%	43.9843	0.122975	2.498225	66.84135	4.653375	4.164275	5.6059	3.79210
50%	70.62605	0.1661	2.56235	70.22625	5.5659	6.61595	5.69055	5.51954
75%	99.04713	0.194725	2.63595	74.2924	7.385175	19.88548	6.1986	18.2483
Max	129.6746	0.4255	3.1343	89.0704	17.7259	123.8858	7.0775	100.254

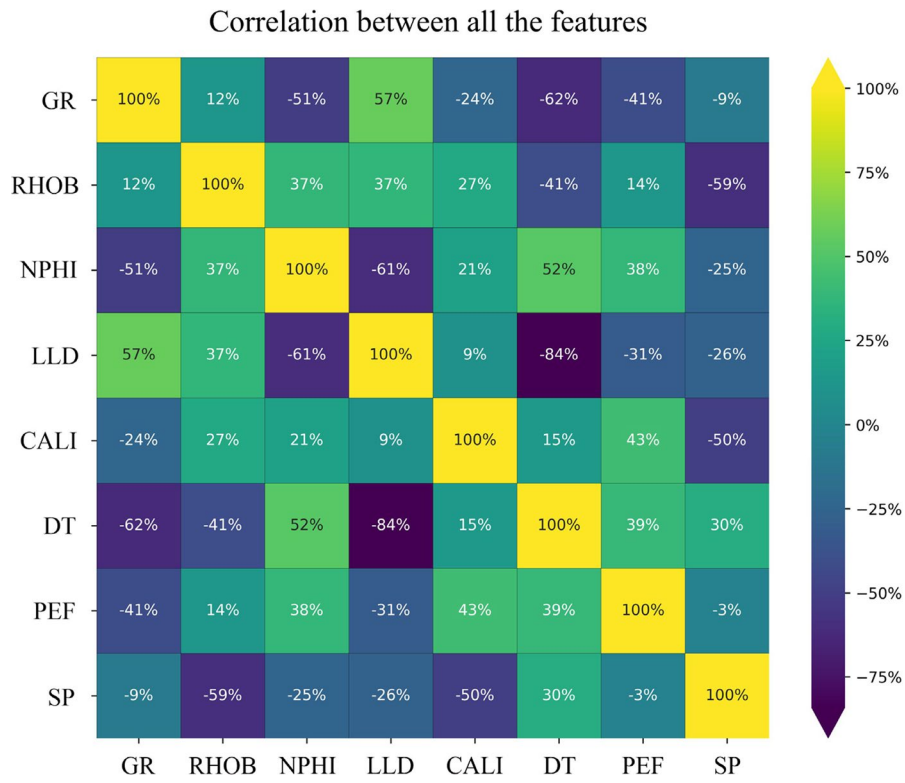

Fig. 7 Correlation between all the features

4.2 Cluster analysis

The most critical step in clustering analysis is the visual evaluation of the two-information criteria plot to establish the number of clusters (facies) for the dataset. The clusters are identified by selecting specific inflection points based on the Euclidean distance. Inflection points on the plot indicate where the advantages of having more clusters do not improve data characterization. The AIC/BIC curves leveling off at six clusters, as depicted in this plot (Fig. 8), are related to a model that accurately describes the clusters (facies) mentioned in the study area well logs.

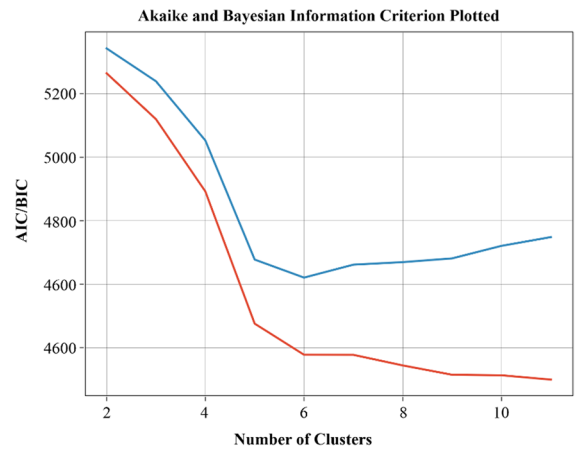

Fig. 8 Two-information criteria plot depicting AIC/BIC curves leveling off at six clusters, crucial for determining the optimal number of clusters (facies) in the dataset

4.3 Feature selection

Before proceeding with model training, it is crucial to select the appropriate feature curve that exhibits a strong correlation with the target curve. In this study, a set of eight logging curves was chosen from the dataset, representing commonly measured parameters within a well. For the purpose of evaluating the ranking mechanism, specific curves such as TVD, LLM, and LLS were excluded as they are not utilized in facies evaluation. The results obtained from the feature selection methods can be seen in Fig. 9. Among the selected features, the Gamma (GR) curve demonstrated the most significant impact, with an Influence Factor value of 36.21. The subsequent relevant feature

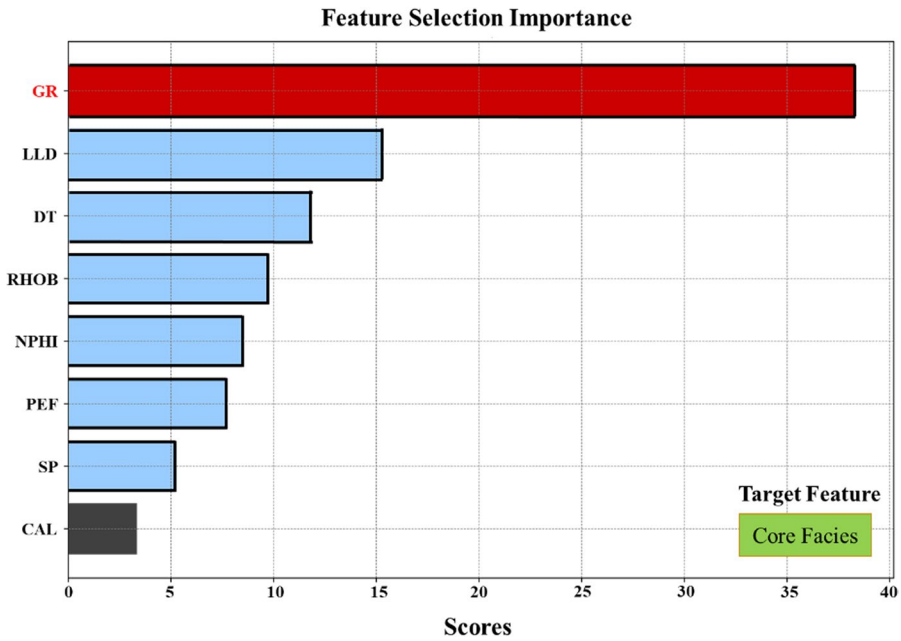

Fig. 9 Results of feature selection methods showing the Influence Factors of selected logging curves for facies evaluation

was the Deep Resistivity (LLD) with an Influence Factor of 16.04, followed by Compressional Sonic (DT) (13.17), Density (RHOB) (9.85), and Neutron (NPHI) (7.70), each showing a relatively smaller decrease in influence. On the other hand, features such as Photoelectric Factor (PEF), Spontaneous Potential (SP), and Caliper (CAL) inputs were ranked as the least impactful and relevant with Influence Factors of 6 and less than 6, respectively. Based on these findings, it can be suggested that the top four ranked inputs (GR, LLD, DT, and RHOB) are the most significant features to consider for facies evaluation.

Once the number of clusters is estimated from the novel two-information criteria plot, the core facies along with selected conventional logs from the feature selection method are inserted together into the training data, referring to the dataset without the blind/testing well. A set for one of the wells was removed from the training data and used to evaluate how well the algorithms performed. The distribution of each facies in the training dataset is shown in Fig. 10, revealing that the High-porosity and high permeability sandstone (Ss) facies are underrepresented in comparison to the other facies. Thus, we can search for more samples of these facies to enhance the effectiveness of the prediction models.

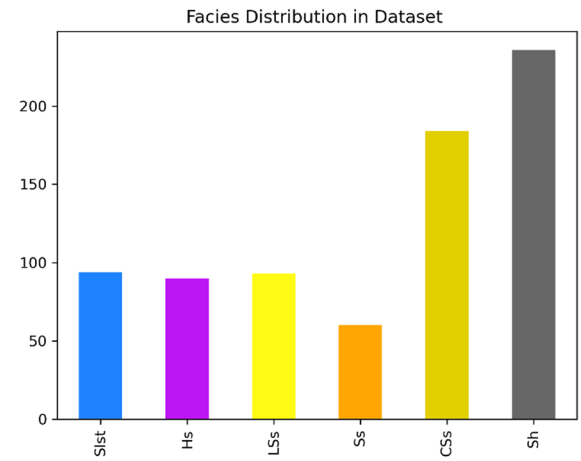

Fig. 10 Distribution of facies in the training dataset

In Fig. 11, we combined kernel-density plots of the good logs to compare the distribution of different features for each facies class. The probability density function (PDF) of a continuous random variable can be estimated using the nonparametric technique known as kernel density estimation (Koehrsen 2018). When features need to be compared between different classes, these plots provide a smooth representation of a histogram and scatter plot that is estimated from the data. The X and Y-axis show the value of

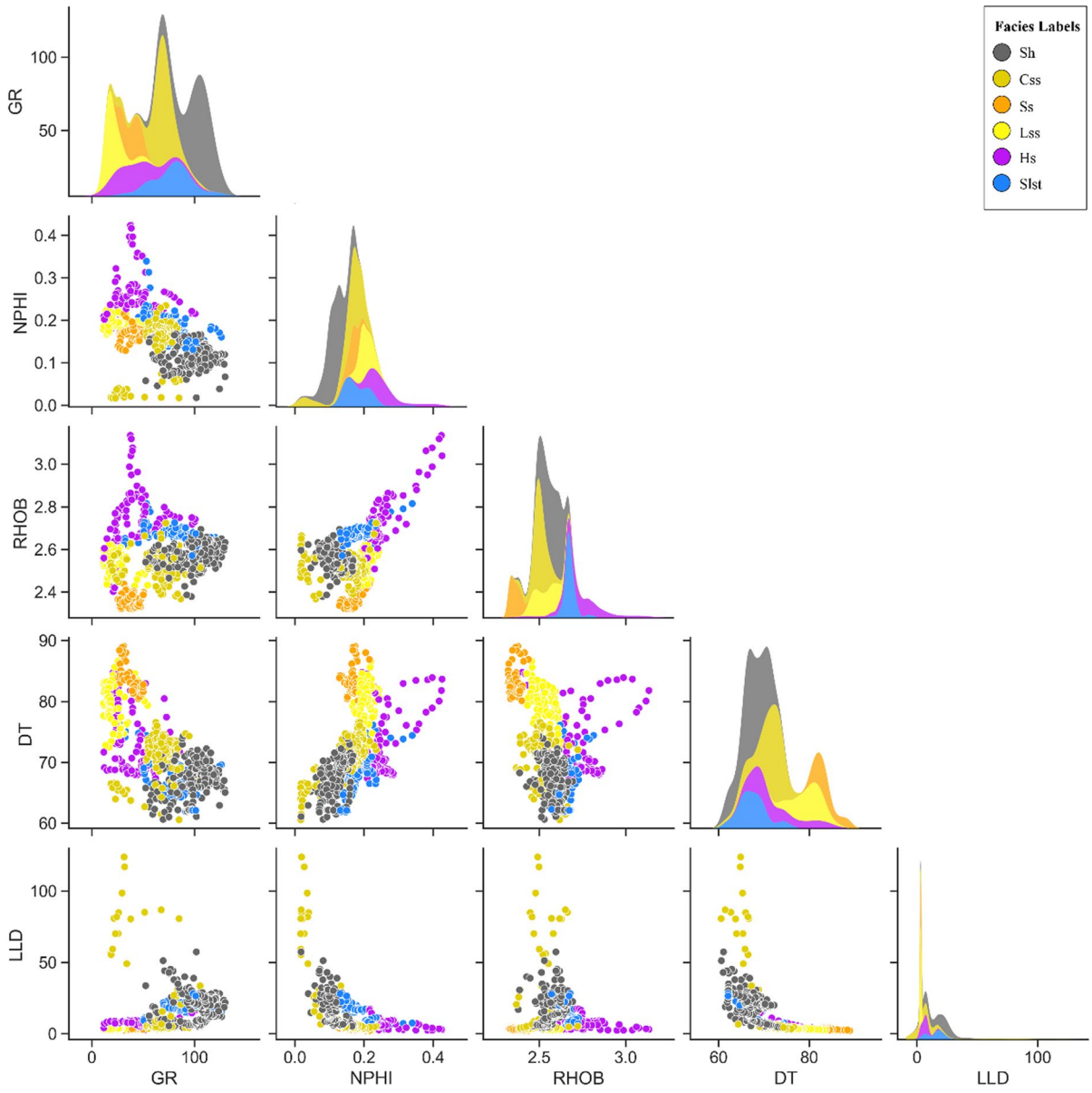

Fig. 11 Kernel-density plots comparing feature distributions for each facies class in logs

the features on the scatter plot, while the histogram shows the probability density function for each feature. Different features can differentiate the facies to varying degrees, and this inherent property in data will be learned by the classification algorithms.

However, the kernel-density plots demonstrate that all facies class overlap, which is indicative of the sub-surface data set and difficult to distinguish utilizing conventional statistical approaches. This visualization supports the use of machine learning to enhance the analysis of a multivariate set of data and extract complex geologically significant features. Figure 12 shows the data visualization distribution and facies classes of our labeled data in 2-dimensional space, which is obtained by radial coordinate visualization (RadViz). Radviz is a visualization technique that maps a series of large points in high-dimensional space and converts it into 2-dimensional space

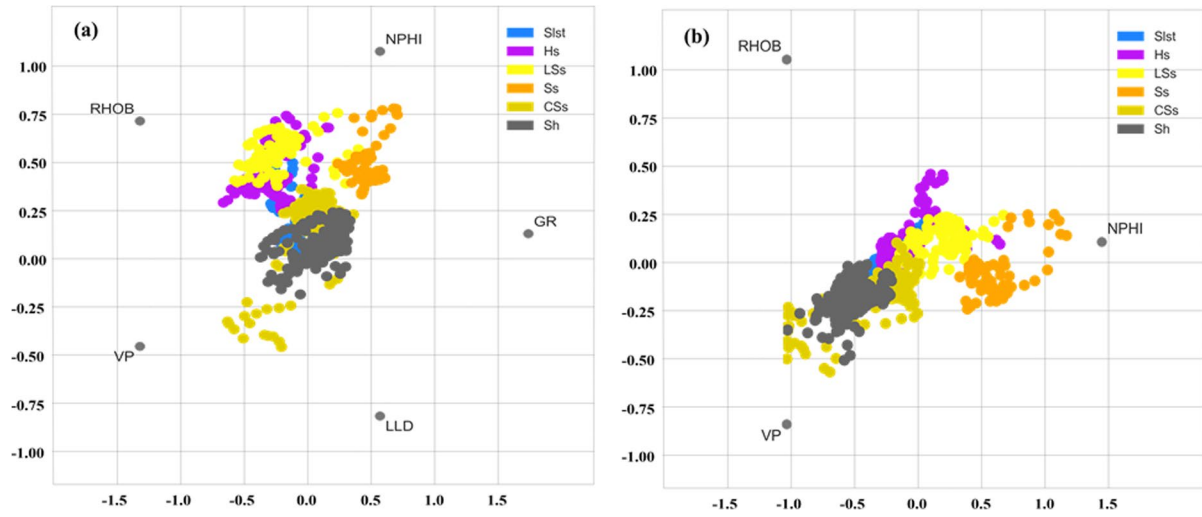

Fig. 12 RadViz demonstrates the distribution of facies classes in a two-dimensional space. **a** Shows the overlap of all attributes in the dataset, while **b** illustrates the predominant separation of classes based on three key features

through a nonlinear method. The RadViz visualization method distributes each attribute variable in the dataset equally on a unit circle. The *m* radius of the unit circle represents the m-dimensional space, and a point in the RadViz visualization graph represents a row of sample data. The design idea comes from the force balance theorem of objects. In Fig. 12a, it can be observed that all attributes from the existing dataset result in a non-informative representation for the reason that its demonstrations of all classes overlap. Nevertheless, in Fig. 12b, it can be observed that the three most appropriate features can be preponderantly separated classes. We can see from the positions of the data points in this diagram that some samples of *Ss* and *LSs* are close to the placements of the *NPHI* attribute, indicating that *SS* and *LSs* are more influenced by *NPHI* than by *RHOB* or *VP*, in contrast to *Sh* and *CSs*, which are more influenced by VP measurements with the remaining class presented between them all.

4.4 Facies predictions utilizing ML models and comparison

The next step in our study involved evaluating the accuracy of various machine learning (ML) models for predicting facies, as described in the methodology section. Initially, several default models were tested using the training dataset, and the results were evaluated using a confusion matrix. A confusion matrix is a matrix that summarizes the accuracy of predictions in a classification problem. The matrix contains counts of both accurate and inaccurate predictions, with each class being further subdivided. This tool is essential for evaluating the effectiveness of classification models and can be used to calculate several evaluation indicators, such as true positive rate, false positive rate, true negative rate, false negative rate, accuracy rate, and F1 index. Moreover, the confusion matrix can help estimate the predicted loss due to misclassification of the classification model by differentiating false positives from false negatives. Overall, the use of a confusion matrix is a vital step in evaluating the performance of machine learning models and provides valuable insights into the effectiveness of classification algorithms.

To measure the effectiveness of identification performance, various metrics such as precision, recall, and F1-score can be utilized. Figure 13 illustrates the identification performance of several models. The results indicate that the GPC model achieved the highest identification accuracy followed by SVM and ANN, while RF performed the worst. In the case of lithology identification, ensemble models are more suitable than RF, and better identification performance can be achieved using the GPC approach. Moreover, each model produces different identification results for different facies types. All models

Confusion Matrix Based on SVM Model							
Prediction	Slst	Hs	LSs	Ss	CSs	Sh	Total
Slst	0	5	1	0	2	11	21
Hs	1	16	2	0	0	0	19
LSs	0	0	18	0	0	0	18
Ss	0	0	10	7	0	0	17
CSs	0	1	0	0	37	0	38
Sh	0	0	0	0	6	33	39
Table: Prediction Accuracy Score							
Precision	0.67	0.73	0.58	1	0.82	0.75	0.76
Recall	0.1	0.84	1	0.41	0.97	0.85	0.74
F1	0.17	0.78	0.73	0.58	0.89	0.8	0.7

Confusion Matrix Based on ANN Model							
Prediction	Slst	Hs	LSs	Ss	CSs	Sh T	Total
Slst	9	9	0	0	0	3	21
Hs	1	16	2	0	0	0	19
LSs	0	0	16	2	0	0	18
Ss	0	0	0	17	0	0	17
CSs	0	1	0	0	35	2	38
Sh	1	0	0	0	3	35	39
Table: Prediction Accuracy Score							
Precision	0.82	0.62	0.89	0.89	0.92	0.88	0.85
Recall	0.43	0.84	0.89	1	0.92	0.9	0.84
F1	0.56	0.71	0.89	0.94	0.92	0.89	0.84

Confusion Matrix Based on GPC Model							
Prediction	Slst	Hs	LSs	Ss	CSs	Sh	Total
Slst	16	3	0	0	1	1	21
Hs	1	16	2	0	0	0	19
LSs	0	0	17	1	0	0	18
Ss	0	0	0	17	0	0	17
CSs	0	1	0	0	36	1	38
Sh	0	0	0	0	3	36	39
Table: Prediction Accuracy Score							
Precision	0.94	0.8	0.89	0.94	0.9	0.95	0.91
Recall	0.76	0.84	0.94	1	0.95	0.92	0.91
F1	0.84	0.82	0.92	0.97	0.92	0.94	0.91

Confusion Matrix Based on RF Model							
Prediction	Slst	Hs	LSs	Ss	CSs	Sh	Total
Slst	16	3	0	0	0	2	21
Hs	5	12	2	0	0	0	19
LSs	0	0	18	0	0	0	18
Ss	0	0	17	0	0	0	17
CSs	1	1	4	0	32	0	38
Sh	0	0	0	0	9	30	39
Table: Prediction Accuracy Score							
Precision	0.73	0.75	0.44	0	0.78	0.94	0.68
Recall	0.76	0.63	1	0	0.84	0.77	0.71
F1	0.74	0.69	0.61	0	0.81	0.85	0.68

Fig. 13 Confusion Matrix illustrating the facies prediction outcomes of SVM, ANN, GPC, and RF models. Specific facies are identified by their labels, which are explained in the accompanying table

demonstrate high accuracy when identifying LSs, Ss, CSs, and Sh. However, RF achieves the worst identification outcome with an F1 score of 0.00 while identifying Ss. On the other hand, the GPC model overall performs the best slightly outperforming ANN in identifying all classes. Additionally, SVM has a weaker identification ability for Hs, LSs, CSs, and Sh compared to ANN. These results highlight the significant advantage of ensemble models when dealing with non-uniformly distributed facies data.

Figure 13 displays the confusion matrices for the models, which show how the predicted facies compare to the actual facies for each lithofacies. Using the GPC model's confusion matrix as an example, the most common types of incorrect classifications are as follows: (1) 3% of samples were misclassified as Hs, CSs, and Sh, 1% for each Slst sample; (2) 3% of CSs samples were misclassified as Sh. Because the misclassified lithofacies have overlapping logging features, it can be challenging to identify these misclassifications. Nevertheless, the GPC model outperforms other models, and its prediction accuracy of lithofacies is high, as shown by the total amount of lithofacies in Fig. 13.

However, to gain deeper insights into the efficiency of different model performances, we constructed a multiclass ROC curve, as shown in Fig. 14. We analyzed the ROC curves individually for each of the four models. The results indicate that the GPC model achieved the highest identification accuracy followed by SVM and ANN, while RF performed the worst. In the case of lithology identification, ensemble models are more suitable classifiers for predicting lithofacies since they make fewer prediction errors, indicating that they performed well in the Lower Goru Formation. This confirms that the classifier is capable of perfectly separating the six clusters with a true positive rate. The GPC results for Slst, Hs, LSs, Ss, CSs, and Sh were strong with ROC scores of 0.88, 0.89, 0.96, 0.99, 0.93, and 0.98, respectively. It is evident that the GPC technique performs better than the other techniques because it can identify individual facies Slst, Hs, LSs, Ss, CSs, and Sh with reasonable accuracy, while SVM yields low accuracy in predicting facies Slst and Hs compared to ANN. However, RF

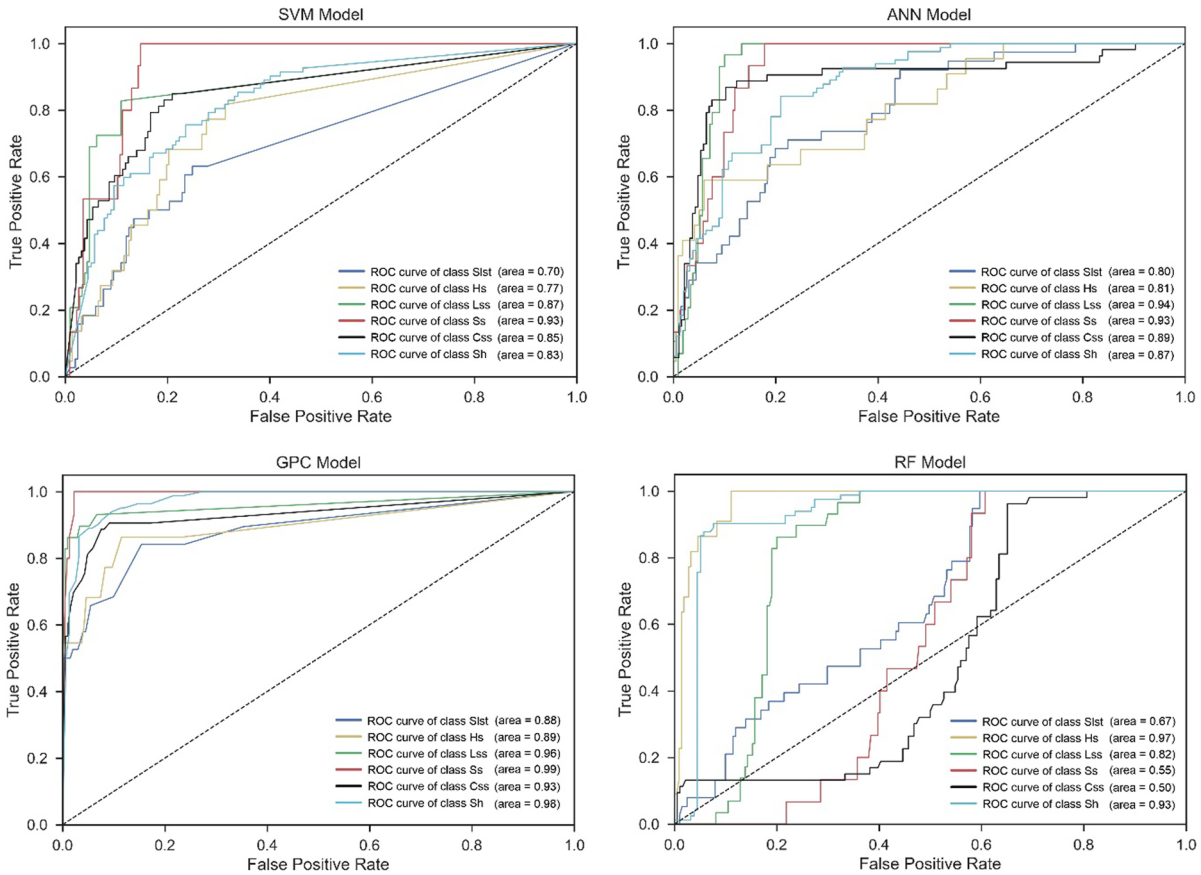

Fig. 14 ROC curves for each facies class score are depicted for SVM, ANN, RF, and GPC models are analyzed

achieves the worst identification outcome with ROC scores of 0.67, 0.55, and 0.50 while identifying Slst, Ss, and CSs. These results clearly reveal that the GPC model accurately distinguishes each facies class in the dataset.

Once the final model has been extracted from the previous comparative results, in the last step, we utilized the final model (GPC) to predict the lithofacies in the remaining/testing dataset and assess the prediction accuracy of facies similarity based on synchronization measures for predicting synthetic logs. Leveraging available core data from K-15 wells, we successfully extended facies predictions across the remaining wells, even in areas where core data were unavailable. This process not only allowed us to fill gaps in our dataset but also provided valuable insights into lithofacies distribution across the study area. The resulting facies distribution on a blind well using the final model (GPC) is depicted in Fig. 15. As depicted, the first four log tracks illustrate the measured logs as a function of depth (in meters), while the last track showcases the facies track predicted by GPC. This comprehensive analysis underscores the effectiveness of our approach in accurately predicting lithofacies and facilitating informed decision-making in exploration and production activities.

To evaluate the results obtained from the GPC in the absence of ground truth label data (e.g., core data) in Fig. 15, we utilized a novel technique to confirm the facies prediction accuracy in the blind well (K-14). Therefore, we employed elastic parameters, such as acoustic impedance and velocity ratio, which are directly associated with reservoir quality, lithofacies, and the corresponding petrophysical response. Subsequently, we constructed the acoustic impedance vs. velocity ratio cross-plot of the blind well based on predicted lithofacies to generate 2D probability density functions plotted on the top (Fig. 16a, b). These

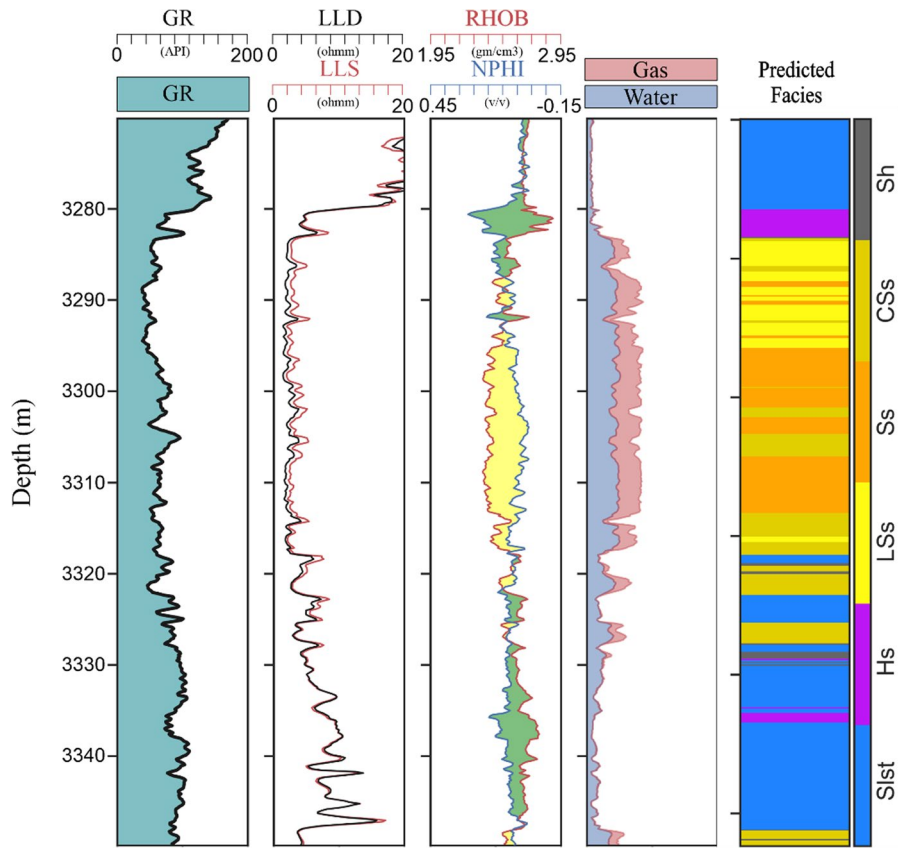

Fig. 15 Prediction results of the final model for the blind well (K-14) showcasing lithofacies distribution based on predicted logs and synchronization measures

probability density functions (PDFs) are defined for each sample indicating a point that is similar to the data points in its cluster and its dissimilarity to the data points not within its cluster. Once the PDFs were extracted by predicted lithofacies, we utilized them as input along with depth in the neural network to predict synthetic gamma-ray log responses (Fig. 16c). Since the gamma-ray log is a lithology indicator that plays a vital role in differentiating rock types, we predict the synthetic gamma-ray log response and assess the prediction accuracy of the facies similarity based on synchronization measures. In Fig. 16d, a comparison of the actual gamma log (GR) from these wells with the predicted gamma log (Syn_GR) from the neural network is presented. Qualitatively, the obtained results are visually satisfactory, and the average log trends are almost identical (Fig. 16d). Cross-plots between the measured GR and predicted Syn_GR from the machine learning algorithm at blind wells are shown in Fig. 16e. These cross-plots provide a quantitative measure of the predictive ability of the machine learning algorithm, with quite satisfactory R2 values (0.978), respectively.

Similarly, we have also applied the proposed final model to another well to predict lithofacies across the wells in the area where core data were unavailable. The result of the facies distribution is given in Fig. 17. As seen, core facies exhibit a fairly good match with predicted facies along the depth intervals with core facies description. This validation underscores the robustness and reliability of our predictive model in capturing lithofacies variations in diverse geological settings.

After predicting the lithofacies across all wells in the area with available and unavailable core data, the inverted Acoustic Impedance (AI) and Vp/Vs volumes were used to classify the predicted lithofacies. Subsequently, this information was leveraged to construct a facies prediction volume. Figure 18a displays a cross-section of wells in the area, exhibiting the predicted facies volume derived from an inversion with the predicted lithofacies obtained from the well log. Non-reservoir facies are depicted in grey and green,

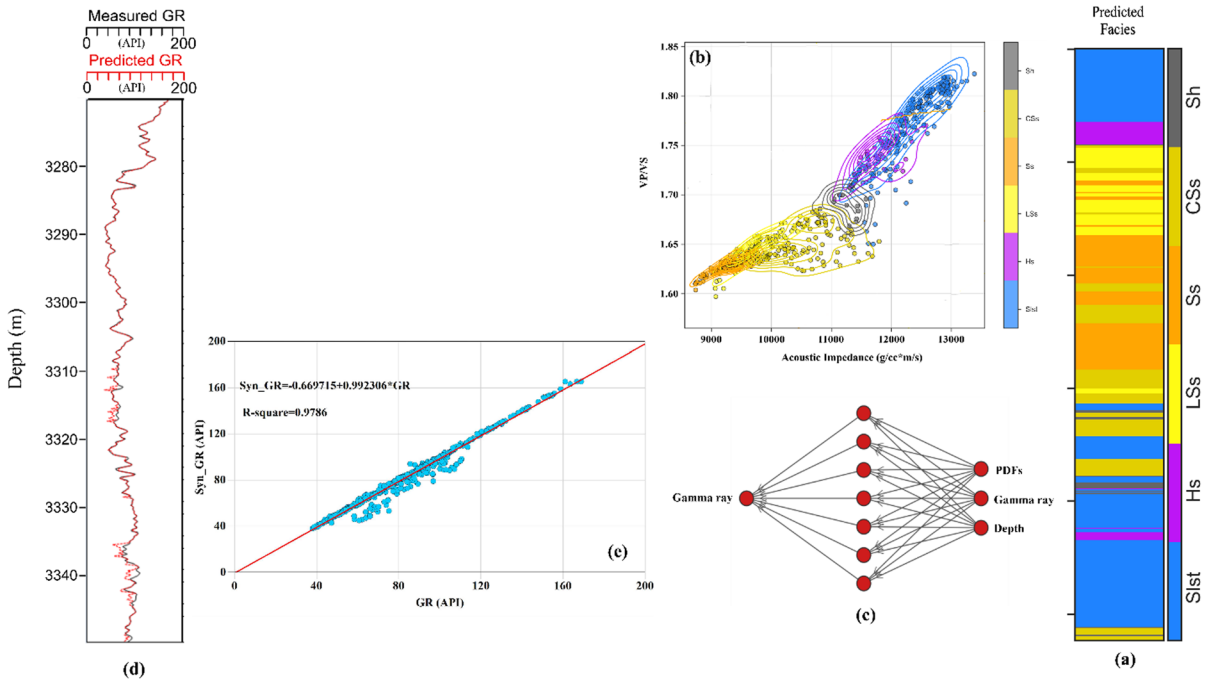

Fig. 16 Result of the blind well **a** predicted lithofacies **b** based on lithofacies extract PDFs **c** utilized neural network to predicted synthetic log **d** actual and predicted gamma-ray log response comparison trend **e** check the prediction accuracy result by the least square method

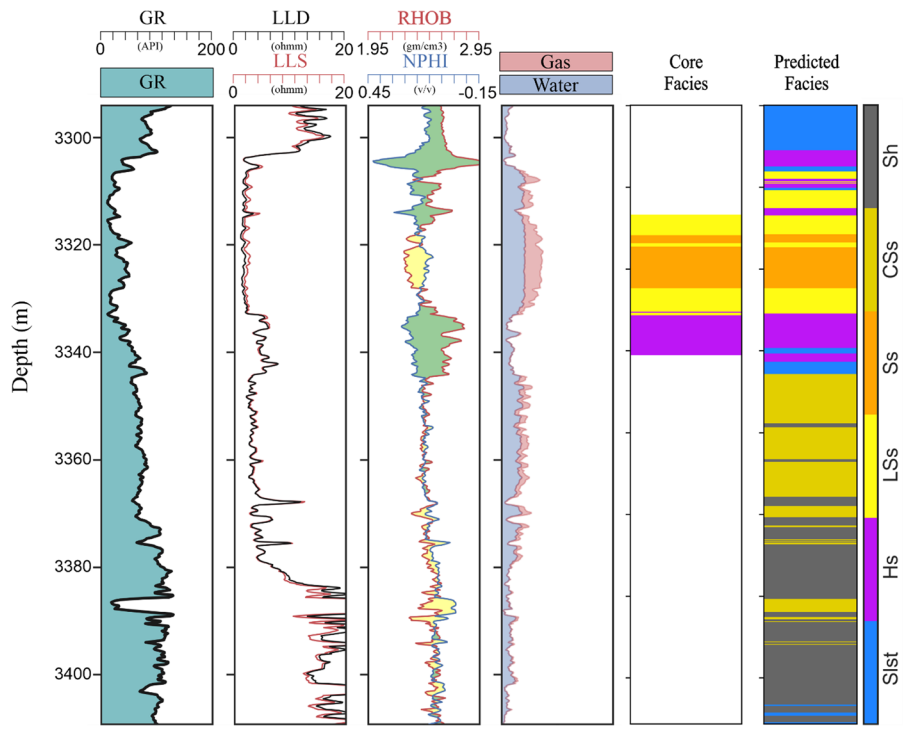

Fig. 17 Comparison of Core Facies with Predicted Distribution in Well: Validating the Accuracy of the Predictive Model

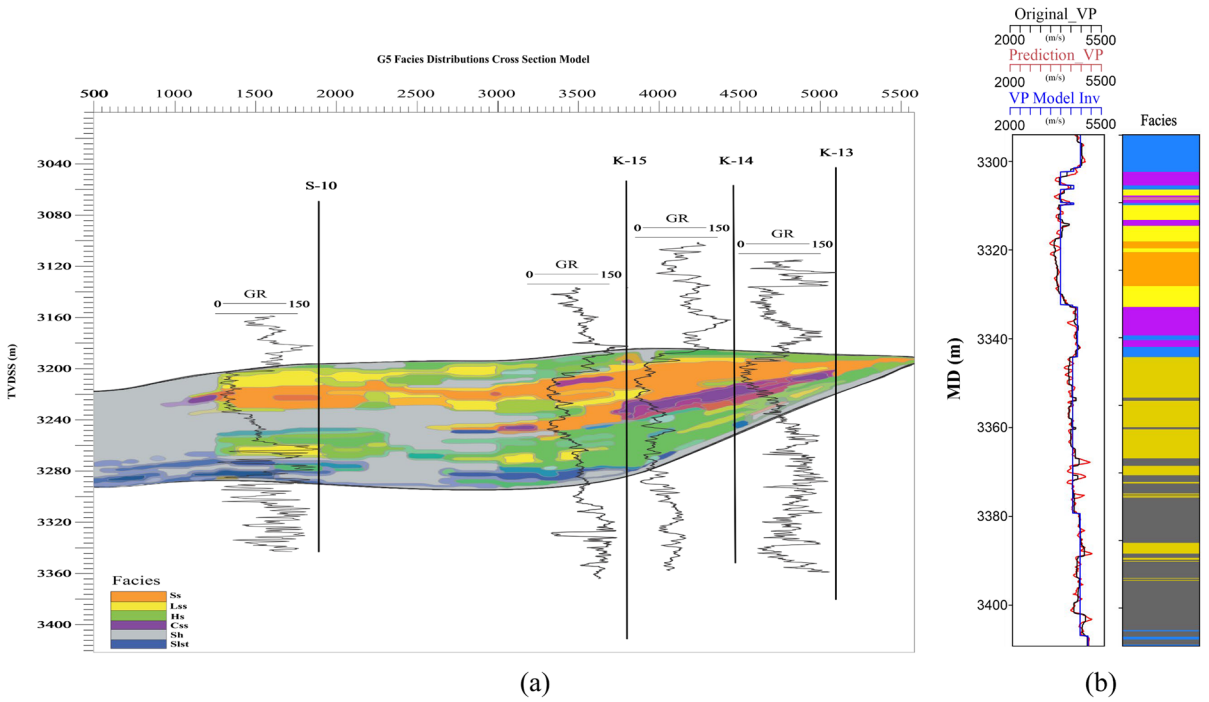

Fig. 18 Two panels are depicted: **a** shows the facies prediction volume created by classifying the predicted lithofacies using the inverted Acoustic Impedance and Vp/Vs volumes, and **b** presents a comparison of the original Vp log, Vp predicted from the facies model applied to the original porosity log, and modeled Vp obtained from the predicted porosity volume derived from inversion using the facies model

while good reservoir facies are represented in orange and yellow, and medium reservoir facies are shown in blue and purple. A notable correlation between the predicted classification of lithofacies and the predicted volume of facies can be observed at the well locations.

To ensure quality control, Fig. 18b provides a detailed illustration of the Vp log, the Vp modeled from the predicted porosity volume generated from inversion utilizing the facies model, and the Vp modeled from the original porosity log. The modeled Vp, calculated based on the porosity log, exhibits a strong correlation with the Vp log. Moreover, the predicted Vp from inversion demonstrates good alignment with the measured log, particularly in the upper section of the well where the reservoir facies are predominant. However, in the lower section of the well where non-reservoir facies are prevalent, the correlation is less significant. Despite this, overall, the model's performance remains quite high effectively distinguishing between different rock types.

5 Conclusion

This study presents a comprehensive workflow for predicting lithofacies using a combination of supervised machine learning algorithms and innovative data-driven clustering techniques. By employing a novel two-information-criteria clustering approach, bias from subjective human judgment in traditional manual approaches is minimized. Following lithofacies identification, four supervised machine learning classifiers, including a voting ensemble classifier were deployed. The drilling dataset was split into training and testing subsets with the former used to train models for target prediction. GPC exhibited the highest identification performance followed by SVM and ANN, while RF showed the lowest performance. Next, the GPC model was utilized to predict lithofacies in the testing dataset. The accuracy of facies similarity was assessed through synchronization measures to predict synthetic logs. Using the predicted lithofacies, a 2D probability density

function was generated from an acoustic impedance versus velocity ratio cross plot of a blind well. This, along with depth, was input into a neural network to predict synthetic gamma-ray log responses. The results from the neural network were visually satisfactory, showing nearly identical average log trends of the gamma-ray log and a high correlation between the measured GR and predicted Syn_GR from the machine learning algorithm at blind wells (R^2 of 0.978). Finally, the predicted lithofacies were used to create a facies prediction volume by employing inverted acoustic impedance and Vp/Vs volumes. These predicted facies volumes correlated well with the predicted lithofacies classification in both wells with and without core data. The comparative application and analysis of this workflow serve as a reference for sedimentary lithofacies logging identification in other study areas. It offers practical value in addressing challenges related to applying machine learning to well log data providing consistent, reliable, and efficient results while saving time and effort in data processing and interpretation. The methodology and results of this study are applicable to a wide range of earth science studies, facilitating more accurate lithofacies prediction and improving the understanding of subsurface geological formations.

Acknowledgements I am thankful to the Ministry of Petroleum, Pakistan, for providing the sample and research data. This research was funded by financial support from the National Natural Science Foundation of China (Grant No. 41774145).

Authors contribution Conceptualization, M.A; methodology, M.A; software, M.A and P.Z; validation, M.A and P.Z; formal analysis,, M.A and P.Z and M.A; investigation, M.A and W.K; resources, M.A; data correction, M.A; writing-original draft, M.A; and writing-review editing, M.A, M.H,P.Z, and H.Z; visualization, supervision, M.H, U.A and R.J; project administration, funding acquisition, P.Z.

Funding The research received financial support from the National Natural Science Foundation of China (Grant No. 41774145).

Availability of data and materials The data supporting the findings of this study can be obtained from the corresponding author upon a reasonable request.

Declarations

Competing interests The authors declare no competing interests.

Ethics approval and consent to participate No specific ethics approval was required for this text-based conversation as it does not involve the use of human subjects or data.

Non-clinical trial This study does not involve in any clinical trials.

Consent for publication All authors have given their explicit consent for this manuscript to be published in its current form and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interest The authors declare no conflicts of interest in relation to this work.

References

Ahmad N, Chaudhry S (2002) Kadanwari Gas Field, Pakistan: a disappointment turns into an attractive development opportunity. Pet Geosci. https://doi.org/10.1144/petgeo.8.4.307

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr 19(6):716–723

Akkurt R, Conroy TT, Psaila D, Paxton A, Low J, Spaans P (2018) Accelerating and enhancing petrophysical analysis with machine learning: a case study of an automated system for well log outlier detection and reconstruction. SPWLA 59th Annu. Logging Symp. 2–6 June, London, UK

Alghazal M, Krinis D (2021) A novel approach of using feature-based machine learning models to expand coverage

of oil saturation from dielectric logs. In: Soc. Pet. Eng. - SPE Eur. Featur. 82nd EAGE Conf. Exhib. EURO 2021, vol 2, p 10. https://doi.org/10.2118/205162-ms.

Ali M, Khan MJ, Ali M, Iftikhar S (2019) Petrophysical analysis of well logs for reservoir evaluation: a case study of 'Kadanwari' gas field, middle Indus basin, Pakistan. Arab J Geosci 12(6):215. https://doi.org/10.1007/s12517-019-4389-x

Ali M, Ma H, Pan H, Ashraf U, Jiang R (2020) Building a rock physics model for the formation evaluation of the Lower Goru sand reservoir of the Southern Indus Basin in Pakistan. J Pet Sci Eng 194:107461. https://doi.org/10.1016/j.petrol.2020.107461

Ali M et al (2021) Machine learning—a novel approach of well logs similarity based on synchronization measures to predict shear sonic logs. J Pet Sci Eng 203:108602. https://doi.org/10.1016/j.petrol.2021.108602

Ali M et al (2023) Quantitative characterization of shallow marine sediments in tight gas fields of middle indus basin: a rational approach of multiple rock physics diagnostic models. Processes 11(2):323. https://doi.org/10.3390/pr11020323

Ali M et al (2023) Reservoir characterization through comprehensive modeling of elastic logs prediction in heterogeneous rocks using unsupervised clustering and class-based ensemble machine learning. Appl Soft Comput 148:110843. https://doi.org/10.1016/j.asoc.2023.110843

Ali M, Zhu P, Jiang R, Huolin M, Ashraf U (2024) Improved prediction of thin reservoirs in complex structural regions using post-stack seismic waveform inversion: a case study in the Junggar Basin. Canadian Geotech J

Al-Mudhafar WJ, Abbas MA, Wood DA (2022) Performance evaluation of boosting machine learning algorithms for lithofacies classification in heterogeneous carbonate reservoirs. Mar Pet Geol 145:105886. https://doi.org/10.1016/j.marpetgeo.2022.105886

Alzubaidi F, Mostaghimi P, Swietojanski P, Clark SR, Armstrong RT (2021) Automated lithology classification from drill core images using convolutional neural networks. J Pet Sci Eng 197:107933. https://doi.org/10.1016/j.petrol.2020.107933

Anees A, Zhang H, Ashraf U, Wang R, Thanh HV, Radwan AE, Ullah J, Abbasi GR, Iqbal I, Ali N, Zhang X, Tan S, Shi W (2022) Sand-ratio distribution in an unconventional tight sandstone reservoir of Hangjinqi area, Ordos Basin: acoustic impedance inversion-based reservoir quality prediction. Front Earth Sci 10:1018105

Antariksa G, Muammar R, Lee J (2022) Performance evaluation of machine learning-based classification with rock-physics analysis of geological lithofacies in Tarakan Basin, Indonesia. J Pet Sci Eng 208:109250. https://doi.org/10.1016/j.petrol.2021.109250

Ashraf U et al (2019) Classification of reservoir facies using well log and 3D seismic attributes for prospect evaluation and field development: a case study of Sawan gas field, Pakistan. J Pet Sci Eng 175:338–351. https://doi.org/10.1016/j.petrol.2018.12.060

Ashraf U et al (2020) Controls on reservoir heterogeneity of a shallow-marine reservoir in Sawan gas field, SE Pakistan: implications for reservoir quality prediction using acoustic impedance inversion. Water 12(11):2972. https://doi.org/10.3390/w12112972

Ashraf U, Shi W, Zhang H, Anees A, Jiang R, Ali M, Mangi HN, Zhang X (2024) Reservoir rock typing assessment in a coal-tight sand based heterogeneous geological formation through advanced AI methods. Sci Rep 14(1):5659

Berczi L-P, Posner I, Barfoot TD (2015) "Learning to assess terrain from human demonstration using an introspective Gaussian-process classifier. IEEE Int Conf Robot Autom (ICRA) 2015:3178–3185

Bhattacharya S, Carr TR, 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. https://doi.org/10.1016/j.jngse.2016.04.055

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Chai H et al (2009) Automatic discrimination of sedimentary facies and lithologies in reef-bank reservoirs using borehole image logs. Appl Geophys 6(1):17–29. https://doi.org/10.1007/s11770-009-0011-4

Chawshin K, Gonzalez A, Berg CF, Varagnolo D, Heidari Z, Lopez O (2021) Classifying Lithofacies from Textural Features in Whole Core CT-Scan Images. SPE Reserv Eval Eng 24(02):341–357. https://doi.org/10.2118/205354-PA

Cortes C, Vapnik V (1995) Support vector machine. Mach Learn 20(3):273–297

Ehsan M, Gu H (2020) An integrated approach for the identification of lithofacies and clay mineralogy through Neuro-Fuzzy, cross plot, and statistical analyses, from well log data. J Earth Syst Sci 129:1–13

Ghanbarnejadmoghanloo H, Riahi MA (2023) Integrating watershed segmentation algorithm and supervised Bayesian classification for the assessment of petrophysical parameters, pore properties, and lithofacies: a case study from Abadan Plain, SW Iran. Earth Sci Informatics 16(4):3913–3930. https://doi.org/10.1007/s12145-023-01129-x

Gibbs MN, MacKay DJC (2000) Variational Gaussian process classifiers. IEEE Trans Neural Netw 11(6):1458–1464

Granitto PM, Gasperi F, Biasioli F, Trainotti E, Furlanello C (2007) Modern data mining tools in descriptive sensory analysis: a case study with a Random forest approach. Food Qual Prefer 18(4):681–689

Guresen E, Kayakutlu G (2011) Definition of artificial neural networks with comparison to other networks. Procedia Comput Sci 3:426–433. https://doi.org/10.1016/j.procs.2010.12.071

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3(null):1157–1182

Hastie GD et al (2019) Automated detection and tracking of marine mammals: a novel sonar tool for monitoring effects of marine industry. Aquat Conserv Mar Freshw Ecosyst 29:119–130

Haykin SO (2011) Neural networks and learning machines. Pearson Education. [Online]. https://books.google.com/books?id=faouAAAAQBAJ

He J, Ding W, Jiang Z, Li A, Wang R, Sun Y (2016) Logging identification and characteristic analysis of the lacustrine

organic-rich shale lithofacies: a case study from the Es3L shale in the Jiyang Depression, Bohai Bay Basin, Eastern China. J Pet Sci Eng 145:238–255. https://doi.org/10.1016/j.petrol.2016.05.017

Hemmesch NT, Harris NB, Mnich CA, Selby D (2014) A sequence-stratigraphic framework for the Upper Devonian Woodford Shale, Permian Basin, west Texas. Am Assoc Pet Geol Bull 98(1):23–47. https://doi.org/10.1306/05221312077

Koehrsen J (2018) Religious tastes and styles as markers of class belonging: a Bourdieuian perspective on pentecostalism in South America. Sociology 52(6):1237–1253

Lai J et al (2018) A review on the applications of image logs in structural analysis and sedimentary characterization. Mar Pet Geol 95:139–166. https://doi.org/10.1016/j.marpetgeo.2018.04.020

Lan X, Zou C, Kang Z, Wu X (2021) Log facies identification in carbonate reservoirs using multiclass semi-supervised learning strategy. Fuel 302:121145. https://doi.org/10.1016/j.fuel.2021.121145

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539

Li Y, Li T, Liu H (2017) Recent advances in feature selection and its applications. Knowl Inf Syst 53(3):551–577. https://doi.org/10.1007/s10115-017-1059-8

Li L, Wen Z, Wang Z (2016) Outlier detection and correction during the process of groundwater lever monitoring base on pauta criterion with self-learning and smooth processing. In: AsiaSim/SCS AutumnSim

Meyer-Baese A, Schmid V (2014) Foundations of neural networks. In: Pattern recognition and signal analysis in medical imaging. Elsevier, pp 197–243. https://doi.org/10.1016/B978-0-12-409545-8.00007-8.

Moghanloo HG, Riahi MA, Bagheri M (2018) Application of simultaneous prestack inversion in reservoir facies identification. J Geophys Eng 15(4):1376–1388. https://doi.org/10.1088/1742-2140/aab249

Ozkan A, Cumella S, Milliken K, Laubach S (2011) Prediction of lithofacies and reservoir quality using well logs, Late Cretaceous Williams Fork Formation, Mamm Creek field, Piceance Basin, Colorado. Am Assoc Pet Geol Bull 95:1699–1723. https://doi.org/10.1306/01191109143

Soentpiet R et al (1999) Advances in kernel methods: support vector learning. MIT Press, Cambridge

Song L et al (2021) Prediction and analysis of geomechanical properties of jimusaer shale using a machine learning approach. In SPWLA 62nd annual logging symposium

Stone WE, Javid MJ (1979) Quantitative evaluation of the actions of anticonvulsants against different chemical convulsants. Arch Int Pharmacodyn Ther 240(1):66–78

Tian Y et al (2016) Multi-resolution graph-based clustering analysis for lithofacies identification from well log data: Case study of intraplatform bank gas fields, Amu Darya Basin. Appl Geophys 13(4):598–607. https://doi.org/10.1007/s11770-016-0588-3

Thanh HV, Zamanyad A, Safaei-Farouji M, Ashraf U, Hemeng Z (2022) Application of hybrid artificial intelligent models to predict deliverability of underground natural gas storage sites. Renew Energy 200:169–184

Valentín MB et al (2019) A deep residual convolutional neural network for automatic lithological facies identification in Brazilian pre-salt oilfield wellbore image logs. J Pet Sci Eng 179:474–503. https://doi.org/10.1016/j.petrol.2019.04.030

Valzania S et al (2011) Kadanwari field: a tight gas reservoir study and a successful pilot well give new life to an exploited field. In: 73rd Eur. Assoc. Geosci. Eng. Conf. Exhib. 2011 Unconv. Resour. Role Technol. Inc. SPE Eur. 2011, vol 4, pp 2715–2744. https://doi.org/10.2118/143001-ms

Wu D et al (2020) Investigation and prediction of diagenetic facies using well logs in tight gas reservoirs: evidences from the Xu-2 member in the Xinchang structural belt of the western Sichuan Basin, western China. J Pet Sci Eng 192:107326. https://doi.org/10.1016/j.petrol.2020.107326

Yu Z et al (2021) Volcanic lithology identification based on parameter-optimized GBDT algorithm: a case study in the Jilin Oilfield, Songliao Basin, NE China. J Appl Geophys 194:104443. https://doi.org/10.1016/j.jappgeo.2021.104443

Zhang Y, Pan BZ (2011) The application of SVM and FMI to the lithologic identification of volcanic rocks. Geophys Geochemical Explor (in Chinese) 35(5):634–638

Zhang J, Ambrose W, Xie W (2021) Applying convolutional neural networks to identify lithofacies of large-n cores from the Permian Basin and Gulf of Mexico: the importance of the quantity and quality of training data. Mar Pet Geol 133:105307. https://doi.org/10.1016/j.marpetgeo.2021.105307

Zheng W, Tian F, Di Q, Xin W, Cheng F, Shan X (2021) Electrofacies classification of deeply buried carbonate strata using machine learning methods: a case study on ordovician paleokarst reservoirs in Tarim Basin. Mar Pet Geol 123:104720. https://doi.org/10.1016/j.marpetgeo.2020.104720