233 (2024) 212587

Contents lists available at ScienceDirect

Geoenergy Science and Engineering

journal homepage: www.sciencedirect.com/journal/geoenergy-science-and-engineering

Machine learning (ML) for fluvial lithofacies identification from well logs: A hybrid classification model integrating lithofacies characteristics, logging data distributions, and ML models applicability

Shiyi Jiang [a], Panke Sun [a,*], Fengqing Lyu [b], Sicheng Zhu [a], Ruifeng Zhou [a], Bin Li [a], Taihong He [b], Yujian Lin [c], Yining Gao [a], Wendan Song [a], Huaimin Xu [a]

[a] College of Geosciences, China University of Petroleum, Beijing, 102249, China
[b] CNPC Western Drilling Engineering Co., Ltd, Ordos, Inner Mongolia, 017300, China
[c] School of Economics and Management, China University of Petroleum, Beijing, 102249, China

ARTICLE INFO

Keywords:
Lithofacies types
Lithofacies thickness
Logging data distribution characteristics
GMM-BPNN
Lithofacies identification

ABSTRACT

Identifying lithofacies plays a central role in studying sandbody architecture and reservoir quality in fluvial reservoirs. Logging data is widely considered the most effective method for identifying subsurface lithofacies. Many machine learning methods have been developed to automatically identify lithofacies by analyzing the value or patterns of well logs. However, poor generalization of many classification models has resulted from a lack of exploration into the intrinsic relationship between lithofacies characteristics, data distribution characteristics, and classification model applicability. To address this problem, we conducted research on core description, logging curve sampling processing for layer data, and lithofacies identification using gaussian mixture model (GMM) and back-propagation neural network (BPNN) for a tight sandstone reservoir in the northern part of the Sulige gas field. We investigated the relationship between lithofacies characteristics, logging data distribution, and the performances of machine learning classification models. Based on this relationship, we developed a gaussian mixture model-backpropagation neural network hybrid classification model (GMM-BPNN). The results indicate that the logging curve sampling method reduced deviation caused by adjacent lithofacies influence, and made the lithofacies characteristics constrain the distribution characteristics of logging data, thus improving the application of GMM and BPNN. We observe that the distribution of logging data becomes more centralized as the thickness of certain lithofacies increases, thus improving the performance of the GMM applicable to the classification of centrally distributed data. Conversely, the distribution of logging data becomes more discrete as the thickness of certain lithofacies decreases, thus improving the performance of BPNN applicable to the classification of discretely distributed data. Furthermore, the GMM-BPNN (with an F1-score of 0.95) outperformed individual GMM (F1-score of 0.76) and BPNN (F1-score of 0.77). The hybrid classification model also shows better outcomes in the identification of complex lithofacies in other areas.

1. Introduction

Lithofacies, serving as a fundamental unit for characterizing the composition of sedimentary material and sedimentary structure under varying hydrodynamic conditions, profoundly influences the petrophysical properties of reservoirs. Therefore, the identification of lithofacies plays a central role in the study of sandbody architecture and reservoir quality in fluvial reservoirs (Allen, 1983; Miall, 1985; Avseth and Mukerji, 2002; Zengzhao et al., 2013; Colombera and Mountney, 2019; Aigbadon et al., 2022; Fu et al., 2022; Iraji et al., 2023a; Tan et al., 2023; Zhao et al., 2023; Soltanmohammadi et al., 2024). Currently, core observation and logging data analysis are the primary methods employed to determine subsurface lithofacies types. Typically, the most direct and effective approach involves drilling cores and conducting detailed visual inspections. However, this method can be costly and limited in terms of depth range, given that observations are confined to the core section, resulting in a lack of continuity across the entire well section (Chang et al., 2000; Li and Anderson-Sprecher, 2006; Sun et al.,

* Corresponding author.
E-mail address: sunpk@cup.edu.cn (P. Sun).

https://doi.org/10.1016/j.geoen.2023.212587
Received 21 September 2023; Received in revised form 16 November 2023; Accepted 11 December 2023
Available online 14 December 2023

2019; Lan et al., 2021; Iraji et al., 2023b). Hence, the most efficient approach for identifying subsurface lithofacies involves the utilization of logging data (Zhou et al., 2016; Wood, 2019; Li et al., 2022).

Numerous logging techniques have been extensively employed for the identification of subsurface lithofacies. These techniques include conventional logging methods, such as natural gamma logging, acoustic logging, and resistivity logging (Allen, 1975; Asquith and Krygowski, 2004; Yue et al., 2015; Nazeer et al., 2016; Xu et al., 2018; Shen et al., 2019; Shehata et al., 2021). Additionally, specialized logging techniques such as image logging and dipmeter logging are also utilized for this purpose. The identification of subsurface lithofacies using data from conventional logs and methods, including reconstructing lithofacies characteristics from multiple logs and utilizing histograms or cross plots, has gained wide acceptance and application in the field (McDowell, 1999; Liu et al., 2020; Li et al., 2022; Wang et al., 2023). Specialized logging techniques, including dipmeter log for identifying lithofacies types and image log for visualizing lithofacies, have been successfully utilized for the identification of subsurface lithofacies (Lai et al., 2018; El-Gendy et al., 2022; Hassan et al., 2022). Nevertheless, the methods above do have several shortcomings. They may exhibit lower accuracy when identifying complex types of lithofacies, require significant workload and time investment, incur high costs, and are susceptible to potential human errors resulting from subjective factors introduced by geologists (Rider, 1990; Radwan, 2020; Wang et al., 2023).

Machine learning (ML) algorithms have been widely studied for lithofacies identification in recent years, resulting in partial solutions to certain problems. Several machine learning algorithms, including artificial neural network (ANN), support vector machine (SVM), random forest (RF), adaptive boosting (AdaBoost), hidden markov models (HMMs), and convolutional neural network (CNN), have been extensively employed in lithofacies identification with successful outcomes (Harris and Grunsky, 2015; Bhattacharya et al., 2016; Bestagini et al., 2017; Imamverdiyev and Sukhostat, 2019; Al-Mudhafar, 2020; Lan et al., 2021). It is evident that as research continues, more and more machine learning algorithms are being used to improve the accuracy of complex lithofacies identification. However, such research trend gradually raises the computer technology barriers for geological researchers and reduces the practical usability of complex machine learning classification models. The increasing reliance on the superiority of machine-learning models for solving lithofacies identification problems has led to a lack of research on the intrinsic connection between lithofacies characteristics and the applicability of machine learning classification models. As a result, there are still numerous limitations in the application of machine learning classification models in the actual reservoir description work, which is limited by the differences in geological background (Jordan and Mitchell, 2015; Cherana et al., 2022). Two factors that may affect the effectiveness and generalizability of applied machine learning models are as follows: 1) The characteristics of lithofacies can influence the distribution of logging data. Each logging point on the logging curve is a weighted sum of the logging responses of neighboring sediments, and the acquisition of logging data may be affected by the thickness of the lithofacies, which leads to the deviation of the logging data collected by the instrument from its true value, and the noise data is introduced, which in turn leads to the large difference in the distribution characteristics of the logging data of different types of lithofacies (Lindberg et al., 2015; Chen et al., 2021; Tian et al., 2021; Li et al., 2022; Lu et al., 2023; Wang et al., 2023). Consequently, different lithofacies thicknesses can lead to distinct distributions of associated logging data. When the amount of labeled data is large, more noisy data will result in lower performance of a single classification model. However, when there is less labeled data, the classification model may not be able to fit adequately, which leads to a lower performance of the single classification model. 2) The performances of classification models is influenced by the distribution of data (Sancho et al., 2000; Gyori et al., 2022; Paiva et al., 2022; Banerjee et al., 2023; de Amorim et al., 2023; Ramírez et al., 2023). When the differences between classes are small and the data are centralized distributions, generative classification models rely on stable class-conditional probability distributions within the data and benefit from the smoothing effect of the model itself. Consequently, they tend to exhibit better classification performances compared to discriminative classification models (Theodoridis, 2015; Loog, 2018; Langer et al., 2020; González-Prieto et al., 2022). On the other hand, when the differences between classes are small and the data are discrete distributions, discriminative classification models excel by directly learning the classification decision function from the data and adapting to changes in the data distributions. As a result, they often demonstrate superior classification performances in such cases (Rubinstein and Hastie, 1997; Xue and Titterington, 2009, 2010; Cavalcanti and Soares, 2020). In practical research, different areas often exhibit different lithofacies types, lithofacies characteristics, and logging resolution, thus the data distribution characteristics are often different, which leads to the established machine learning classification model that may not apply to other areas, reducing the generalizability of the established model. Combining the above two factors, an understanding of how lithofacies characteristics influence the distribution of logging data can help identifying the distribution pattern of logging data. This enables the effective selection of an appropriate machine learning classification model andthe establishment of the model can be applied to complex lithofacies characteristics and is easily generalizable.

Considering the limitations of prior research, this study employs coring and logging data collected from an area located in the northern part of the Sulige gas field to establish a layer dataset using a logging data sampling method with feature engineering that reduces interference from the adjacent lithofacies. Based on the data in the layer data set, first, the influence of lithofacies characteristics on the distribution characteristics of logging data and the controlling factors are discussed in detail. Then, a basic generative model gaussian mixture model (GMM) and a basic discriminative model back propagation neural network (BPNN) are selected to study the influence of data distribution features on the performances of classification models. Finally, the relationship between lithofacies characteristics, data distribution characteristics and classification models' performances are further investigated. In addition, based on the results of the discussion, a hybrid classification model called gaussian mixture model-backpropagation neural network hybrid classification model (GMM-BPNN), which applies to reservoirs with different lithofacies characteristics, has been developed for the identification of lithofacies in the study area and other areas. This study places greater emphasis on the impact of geological characteristics on data distribution characteristics, and the application of different principles of machine learning algorithms to geological targets with different characteristics, rather than relying on the superiority of a single classification model, this way of establishing a hybrid classification model clearly articulates the intrinsic connection between the geological units to be identified and the identification method, which compensates for the shortcomings of the previous studies and provides an effective modeling thinking for the study of using machine learning to identify other geological units. Additionally, the study established a hybrid classification model that is not only applicable to the study area with more labeled data but also achieves good results in other study areas with less labeled data.

Section 2 describes the geological background of the study area and data sources. Section 3 describes all the methods used in the study. Section 4 shows the lithofacies types and characteristics of the study area, the distribution characteristics of various lithofacies layer data, and the performances of two machine learning classification models. Section 5 discusses the relationship between lithofacies characteristics, layer data distribution characteristics, and the performances of classification models, and establishes the GMM-BPNN, emphasizing its superiority and generalizability in lithofacies classification. Finally, Section 6 summarizes the study and make recommendations for further research and future work.

2. Data sources

2.1. Geological background

The core and logging data utilized in this study were collected from the northern area of the Sulige gas field, located in the Ordos Basin in central China (Fig. 1a and b). The key gas-producing strata in the study area consist of the Upper Paleozoic Permian He8 Member and Shanxi Formation (Fig. 2a). These formations represent near-source fluvial facies characterized by significant sedimentary hydrodynamics. The reservoir primarily comprises sandy sediments associated with meandering river and braided river systems (Yang et al., 2008; Guo et al., 2015; Liu et al., 2016; Wang et al., 2017). Within the coring section, the main sedimentary facies include the meandering river channel, braided river channel, point bar, and channel bar (Fig. 2b).

2.2. Data sources

The study area comprises a total of 900 completed wells, of which 63 core wells are evenly distributed across the area (Fig. 1c). The core and logging data are mainly from the main gas-producing strata within the study area, namely, the Upper Paleozoic Permian He8 member and Shanxi Formation. All wells are equipped with conventional logs, including the spontaneous potential log (SP), gamma ray log (GR), caliper log (CAL), acoustic log (AC), bulk density log (DEN), neutron porosity log (NPHI), deep resistivity log (RLLD), lateral resistivity log (RLLS), and photoelectric factor log (PE). The original resolution of each logging curve is detailed in Table 1.

3. Methods

The research and process of establishing the hybrid classification model in this study can be divided into three steps (Fig. 3): (1) Data set preparation. First, the actual core is described to obtain the lithofacies type, then the core corresponds to the logging curves based on the depth and is sampled to establish the layer data, and finally, the layers are normalized and screened for features. (2) Analyzing the lithofacies based on core description, the characteristics of layer data distribution based on data sampling, the performances of classification models based on feature engineering, and finally exploring the interrelationships among the three analytical results to conclude. (3) Based on the results of analysis and discussion, the GMM-BPNN is established.

3.1. Data sampling method

Each recorded value of logging curves is a weighted summation of the adjacent sediments' logging responses (Lindberg et al., 2015; Tian et al., 2021). Therefore, a data sampling method to minimize the influence of adjacent sediments on the logging data is built in this study. The data acquired by this sampling method are defined as "layer data". Fig. 4 illustrates the operational flow of this method, including the following steps: (1) Identifying the peaks and troughs of the GR logging curve of the target layer using the "find peaks" function. Recording the corresponding depth values; (2) Calculating the mean depth of the neighboring peaks and troughs. This mean depth represents the top-bottom depth of the layer to be identified; (3) Using the distance from the peaks to the top-bottom depth of a specific layer as a weighting factor. Calculating the weighted mean of each logging curve within the layer. This resulting weighted mean represents the layer data.

3.2. Feature engineering

3.2.1. Data normalization

To eliminate the influence of the order of magnitude and unit of the original logging data, the linear function normalization method is employed. This method linearly transforms the layer data, mapping the results to the range of 0–1, thereby achieving equal scaling of the layer data. The layer data from the linear normalization process are defined as GRn, SPn, CALn, Acn, RHOBn, NPHIn, RLLDn, RLLSn, and PEn, respectively. The normalization formula is as follows:

$$x_n = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

here, x_n is the normalized value of x, x is the original value of x, $x_{\min}$ is the smallest value of x in the dataset, and $x_{\max}$ is the largest value of x in the dataset.

3.2.2. Analysis of variance

To evaluate the level of differentiation among different types of logging data for each lithofacies, one-way analysis of variance (ANOVA) and effect size analysis (ESA) were conducted on each type of logging data (Cohen, 1977; Chen et al., 2022). One-way ANOVA is a kind of ANOVA that is used to help measure the impact of a feature on a target class in ML tasks. One-way ANOVA calculates a score for all features and then selects the features with the highest scores (Omer Fadl Elssied et al., 2014; Alassaf & Qamar, 2022). For the one-way ANOVA in this study,

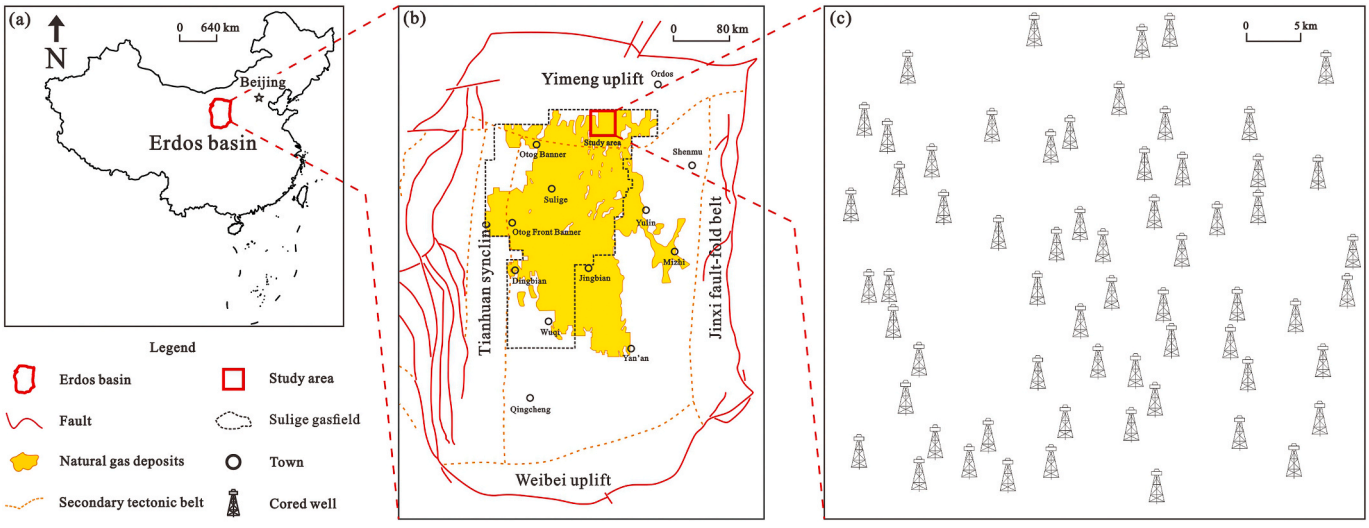

Fig. 1. Location of the study area and distribution of core wells. (a) Location of the Erdos Basin. (b) Location of the study area in the Erdos Basin. (c) Distribution of the core wells in the study area.

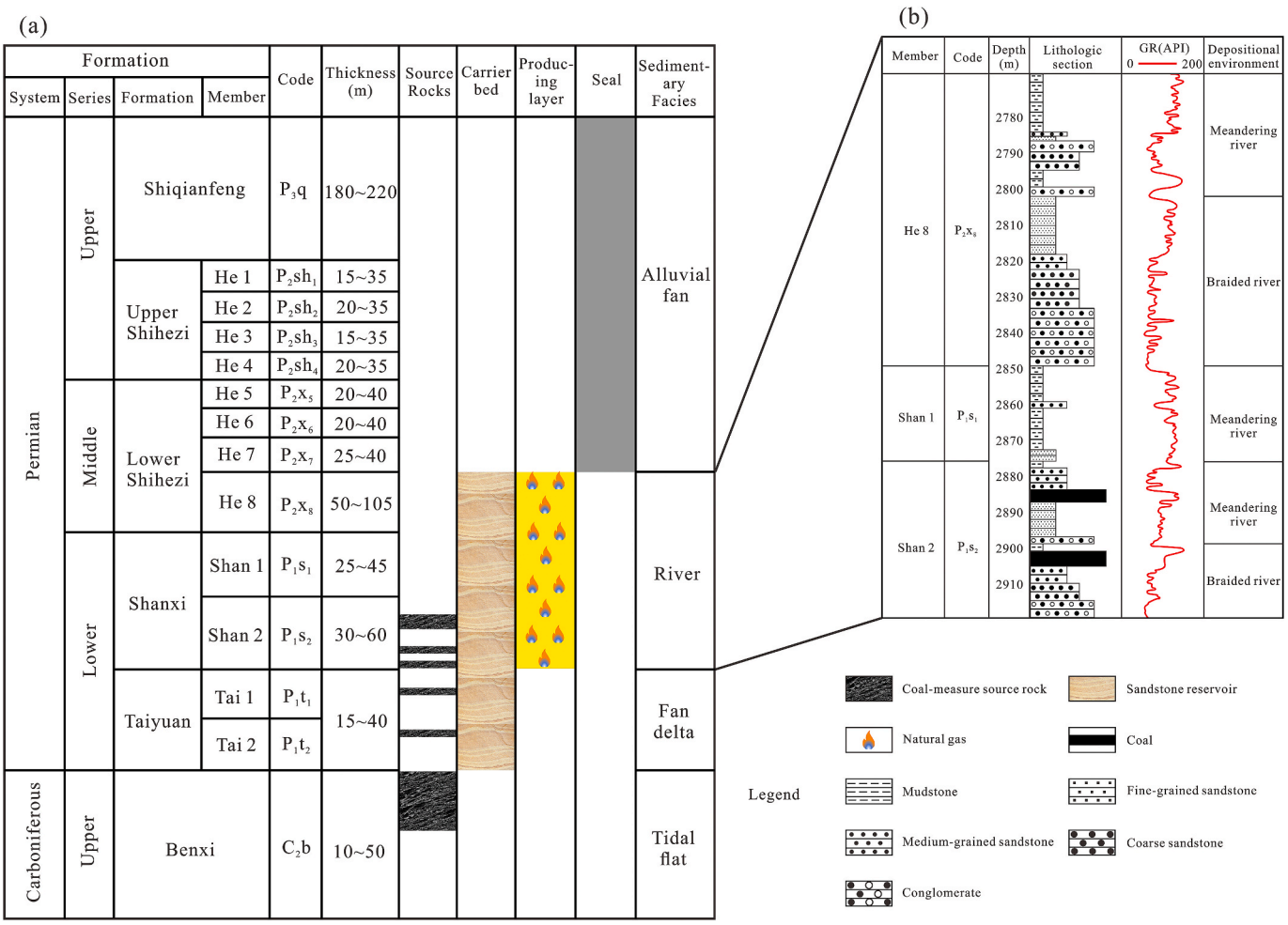

Fig. 2. Generalized stratigraphic column of the Sulige gas field and the studied sandstone group (modified from Wang et al., 2017). (a) Generalized stratigraphy of the Upper Paleozoic successions in the study area, showing the major natural gas system elements. (b) The lithologic sedimentary sequence and sedimentary environment of He8, Shan1, and Shan2 in the study area.

Table 1
Resolutions of all conventional logging curves in the study area.

Well logs	GR	SP	CAL	AC	RHOB	NPHI	RLLD	RLLS	PE
Resolution/m	1.00	2.00		1.00	1.00	1.00	0.60	0.60	1.00

the sum of squares between groups (SSB), the sum of squares within groups (SSE), degrees of freedom, F-value, and P-value were calculated. Based on the parameters calculated by one-way ANOVA, two effector parameters, Partial η^2 and Cohen's f, were further selected and calculated to analyze the differences among the normalized logging data of different types of lithofacies. To calculate Partial η^2 and Cohen's f for the layer data for each lithofacies, it is necessary to calculate the sum of SSB and the total sum of squares (SST). The formulas of SSB and SST are as follows:

$$SSB = \Sigma_{i=1}^{k} n_i \left(\overline{X}_j - \overline{X}_T\right)^2 \tag{2}$$

Here i is the serial number of a particular type of layer data (i=1, 2, …, k), n_i is the total number of layer data of a particular type, $\overline{X_j}$ is the mean of the jth set of lithofacies layer data in this type of layer data, $\overline{X_T}$ is the mean of all lithofacies data in this type of layer data.

$$SST = \Sigma_{i=1}^{k} \Sigma_{j=1}^{n_i} \left(X_{ij} - \overline{X}_T\right)^2 \tag{3}$$

here X_{ij} is the value of the jth lithofacies layer data in the ith type of layer data.

The formulas of Partial η^2 and Cohen's f are as follows:

$$Partial\ \eta^2 = \frac{SSB}{SST} \tag{4}$$

$$Cohen's\ f = Sqrt\left(\frac{Partial\ \eta^2}{1 - Partial\ \eta^2}\right) \tag{5}$$

When utilizing Partial η^2 to denote effect size, the thresholds for distinguishing small, medium, and large effect sizes are 0.01, 0.06, and 0.14, respectively; Cohen's f to denote effect size, the thresholds for distinguishing small, medium, and large effect sizes are 0.1, 0.25, and 0.40, respectively (Cohen, 1977; Rudstam et al., 2022; Zhou et al., 2022). These thresholds serve as guidelines to interpret the magnitude of the effect sizes observed in statistical analyses.

3.2.3. Pearson correlation analysis

In this study, to prevent the classification model from encountering issues related to dimensionality and to enhance the efficiency of model

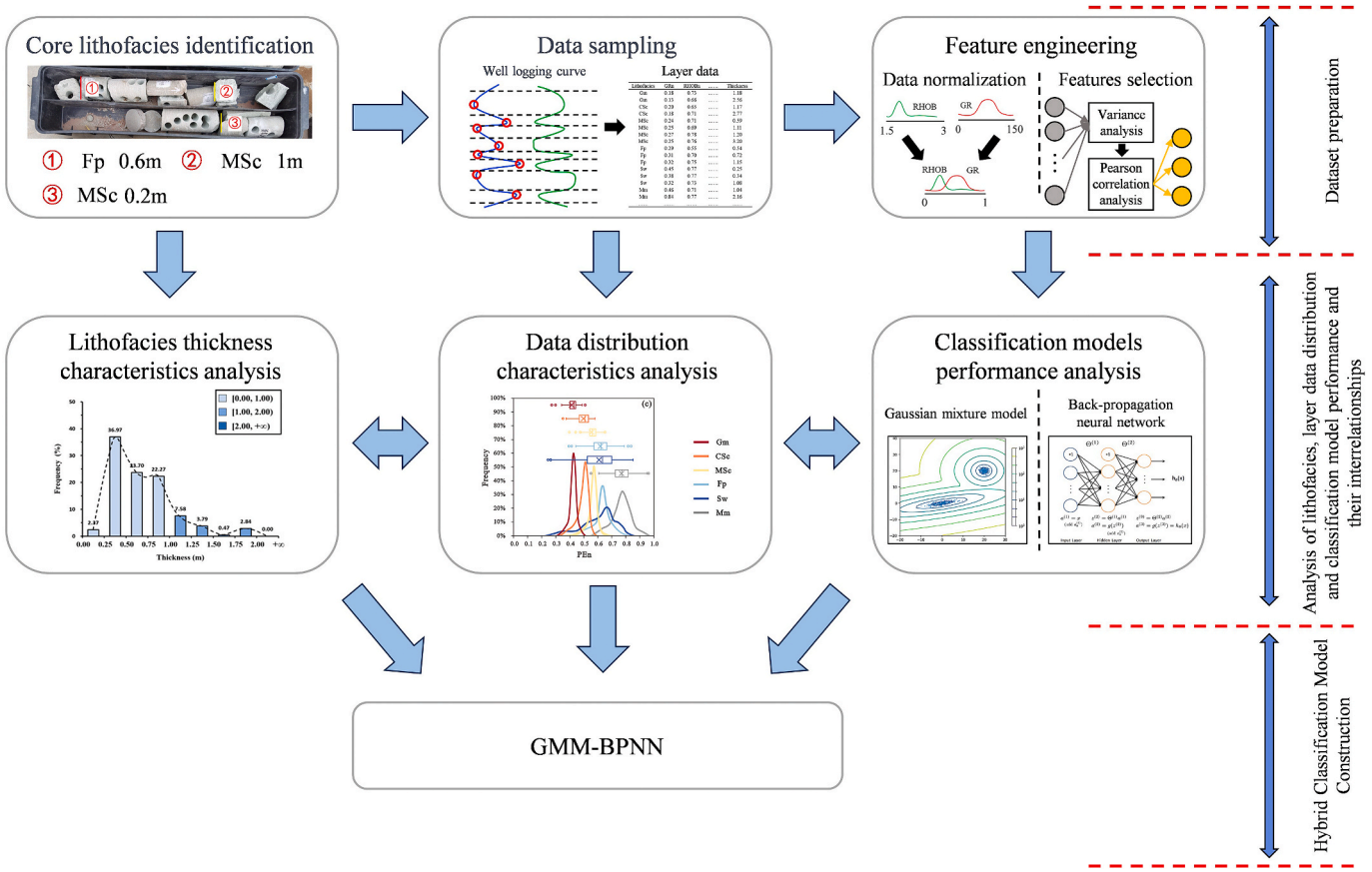

Fig. 3. Workflow diagram outlining the methodology. It consists of three steps: dataset preparation, relationship analysis between characteristics and models, and classification model establishment.

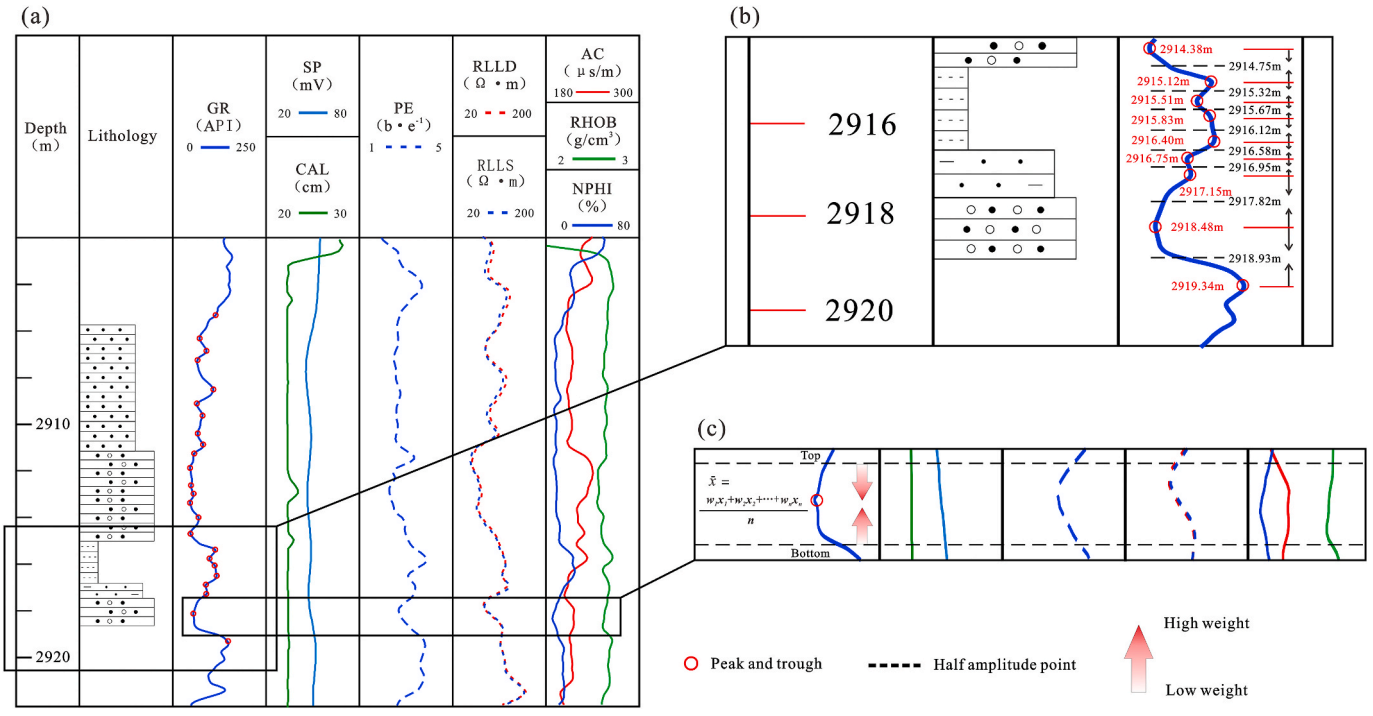

Fig. 4. The process of layer division and layer data calculation. (a) The "find peaks" function is applied to detect the peaks and troughs of the GR curve. (b) The depth of the half-amplitude point is determined by averaging the depths of the identified peaks and troughs. (c) The distance between the top and bottom depths from the peak point is used as the weight. Each logging curve is separately subjected to a weighted mean calculation using these weights. The resulting weighted mean represents the layer data of the respective layer.

training, pearson correlation analysis is employed to explore the relationships among the logging data (Pearson and Henrici, 1997; Rafik and Kamel, 2017). The formulas of the pearson correlation coefficient are as follows:

$$\rho(X,Y)=\frac{cov(X,Y)}{\sigma_x*\sigma_Y} \tag{6}$$

here $\rho(X, Y)$ is the pearson correlation coefficient between data X and data Y, $cov(X, Y)$ is the covariance of data X and data Y, σ_X and σ_Y are the standard deviations of data X and data Y, respectively.

3.3. Data visualization

In this study, frequency distribution histograms, frequency distribution density curves, and box plots were used to visualize the data and characterize the distribution of the data.

3.3.1. Frequency distribution histogram

A frequency distribution histogram is a graphical representation of data that divides it into intervals and shows the frequency of data points within each interval. It has the advantage of being able to visualize the distribution of the data at different intervals in a clear way.

3.3.2. Frequency distribution density curves

A probability density curve, which is a vital statistical tool, provides a visual representation of the probability distribution for continuous random variables through a smooth and uninterrupted line. This representation possesses the advantage of conveying complex probability structures with clarity and concision. The precise measurement of the probability that the random variable falls within a specified range is offered by the area beneath the curve.

3.3.3. Box plots

A box plot, which is also referred to as a box-and-whisker plot, is a succinct graphical diagram utilized in statistics to exhibit the distribution as well as the significant summary statistics of a given dataset. It consists of a rectangular box spanning the interquartile range (IQR) of the data, with a line inside the box representing the median. Box plots offer advantages, including the ability to provide a rapid and informative visual summary of a dataset's central tendency, spread, and outlier presence.

3.4. Statistical measures

In this study, the quantitative analysis of data dispersion was conducted using standard deviation (σ), kurtosis (K), group spacing (GS), and a newly developed parameter, K divided by GS (K/GS).

3.4.1. Standard deviation

Standard deviation (σ) is a statistical measure that quantifies the extent of variability or dispersion within a dataset. It calculates the mean deviation of individual data points from the dataset's mean, providing a numerical indicator of the data's spread. The advantages of using standard deviation for data analysis include its ability to offer a precise and quantitative measurement of dispersion.

3.4.2. Kurtosis

Kurtosis (K) is a statistical measure that quantifies the degree of peakedness or flatness of a probability distribution, indicating how data clusters around the mean and whether it has more or fewer extreme values (outliers) than a normal distribution. The advantages of using kurtosis for data analysis include its ability to provide insights into the shape of a distribution and the presence of outliers.

3.4.3. K/GS

Kurtosis may not precisely depict the traits of various data classes within the same data range because of its dependency on group spacing (GS) during calculation. However, to eliminate this influence and achieve a more precise reflection of the concentration and dispersion of each type of data within the same data range, K divided by GS (K/GS) was developed.

3.5. Evaluation metrics

The model's classification results for the lithofacies identification multi-classification problem were evaluated using a confusion matrix. This matrix includes the following metrics: true positive (TP), false positive (FP), false negative (FN), and true negative (TN). These metrics represent the number of positive samples correctly identified, the number of negative samples incorrectly identified as positive samples, the number of positive samples incorrectly identified as negative samples, and the number of negative samples correctly identified, respectively (Table 2) (Shen and Liu, 2019; Bressan et al., 2020). In this study, TP, FP, FN, and TN were used to calculate precision, recall, and F1-score. These metrics were employed to evaluate the classification results of the classification model.

Precision is an evaluation of the predicted results and represents the proportion of true positives out of all positive predictions. The formula is as follows:

$$\text{Precision}=\frac{TP}{TP+FP} \tag{7}$$

Recall is an evaluation of the original sample and represents the proportion of positive samples that are correctly identified. The formula is as follows:

$$\text{Recall}=\frac{TP}{TP+FN} \tag{8}$$

The F1-score (F1) is a metric that combines precision and recall into a single value, providing a balanced evaluation of the model's performance. It is a weighted mean of precision and recall, with a maximum value of 1 and a minimum of 0. The formula is as follows:

$$F_1=\frac{Precision\cdot Recall}{Precision+Recall} \tag{9}$$

Indeed, higher values of precision, recall, and F1-score indicate better classification results for the classification model.

3.6. Machine learning classification models

Generative and discriminative models exhibit varying degrees of applicability depending on data distribution characteristics (Rubinstein and Hastie, 1997; Xue and Titterington, 2010; Ramírez et al., 2023). Hence, in this study, a representative generative model, gaussian mixture model (GMM), and a typical discriminative model, back-propagation neural network (BPNN), are selected to classify lithofacies with diverse data distribution characteristics, aiming to attain optimal classification results.

3.6.1. Gaussian mixture model

The GMM is a commonly used generative model and clustering algorithm (Melnykov and Maitra, 2010; Rayens, 2012; Kiasari et al., 2017;

Table 2
Confusion matrix for multi-classification data.

	Results of classification	
	Positive class	Negative class
Positive class	*TP*	*FN*
Negative class	*FP*	*TN*

Jiao et al., 2022). It is widely employed for sample data classification. The GMM represents a linear combination of multiple single Gaussian model distribution functions (Dhanalakshmi et al., 2011; Kim et al., 2019). The probability density function of the single Gaussian model for GMM's multidimensional data can be expressed as follows:

$$P(x|\theta) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2} \right) \quad (10)$$

here μ is the expectation of the data, $\sum$ is the covariance and D is the dimension of the data.

GMM clustering involves assigning cluster members based on the calculation of cluster probabilities. When dividing the sample data into m clustering clusters, each cluster corresponds to a sub-model within the GMM. By randomly selecting observation data from the sample, each observation data point is associated with a weight coefficient αm representing the probability of belonging to the mth sub-model. Consequently, the probability density function of the GMM can be expressed as follows:

$$P(x|\theta) = \sum_{m=1}^{M} \alpha_m \varphi(x|\theta_m) \quad (11)$$

here α_m is the probability that the observed data belongs to the mth sub-model, $\alpha_m \geq 1, \sum_{m=1}^{M} \alpha_m = 1$; $\varphi(x|\theta_m)$ is the probability density function of the mth Gaussian sub-model,

In lithofacies identification using GMM, it is necessary to enhance the GMM by adapting it to a supervised learning model. In the current study, the combination of the expectation maximization algorithm (EM) with gaussian discriminant analysis (GDA) is employed to transform the GMM into a supervised learning approach (Dempster et al., 1977; Naim and Gildea, 2012). The fundamental concept is to utilize GDA to estimate the mean and covariance matrices for each type. Subsequently, the EM algorithm is applied to train the GMM while incorporating labeling information to constrain the types represented by each Gaussian distribution. The method consists of the following steps: (1) Assuming that the logging data for each lithofacies follows a Gaussian distribution, the number of Gaussian distributions is determined using GDA, based on the number of lithofacies types, and the mean (μ), covariance matrix (Σ), and prior probability (π) for each Gaussian distribution are derived. (2) The initialization of parameters for each Gaussian distribution in the GMM is carried out using the parameters determined by GDA. (3) The GMM parameters are optimized using the EM algorithm. Initially, labeled data is employed to restrict the classes represented by each Gaussian distribution. Subsequently, in Step E, the estimation of data point distribution is performed by calculating the posterior probability for each data point belonging to each Gaussian distribution, based on the current GMM parameters. Finally, in Step M, the parameters of each Gaussian distribution are re-estimated based on the posterior probabilities from Step E to enhance the fit of the GMM. The key to Step M is the improvement of the parameter estimation of the model by identifying parameters that maximize the expected value of the likelihood function through optimization algorithms such as gradient descent or newton's method. (4) Repeat step 3 until convergence, iteratively performing the E and M steps until the parameters of the GMM reach a stable state. To summarize, the collaboration between GDA, EM, and GMM can be described as follows: GDA offers initial parameters for GMM, while the EM algorithm iteratively refines the GMM parameters to better align with the data distribution, ultimately leading to the achievement of supervised lithofacies classification.

GMM's advantage lies in its ability to adaptively adjust the size and number of clusters when dealing with data that closely follows a Gaussian distribution or exhibits a more concentrated data distribution. This adaptability enables GMM to achieve quicker and more accurate data classification (Zhang et al., 2023). When working with a limited dataset, the presence of discrete values may prevent the data from displaying a typical Gaussian or clustered distribution. In response, GMM may escalate its complexity to better approximate the data distribution (Chen et al., 2023). It's also important to note that GMM may fail to converge when the data distribution is excessively discrete (Chen et al., 2015).

3.6.2. Back-propagation neural network

The BPNN is a widely used supervised learning algorithm in the field of machine learning. It is considered a typical discriminative model (Rumelhart et al., 1986; Rogers et al., 1992; Wu et al., 2006). The training process of BPNN involves an iterative procedure of forward propagation and error back propagation. This process continues until either the error falls below the predetermined threshold or the maximum number of iterations is reached, as specified in the design of the algorithm (Hush and Horne, 1993; Ren et al., 2019). The training process of BPNN with two hidden layers is illustrated in Fig. 5. During the forward propagation stage, the input data in the input layer is weighted and passed to the neurons in the first hidden layer. Each neuron in the first hidden layer computes its activation function and passes the result, weighted again, to the next hidden layer. In the second hidden layer, the data is once again computed by the activation function, weighted, and forwarded to the output layer. In the second hidden layer, the data is computed once more by the activation function, weighted, and then passed to the output layer. The activation function plays a crucial role in facilitating the forward propagation process of the BPNN. These functions perform a nonlinear mapping of inputs to neurons, enabling the network to learn and represent complex nonlinear relationships and data patterns. By facilitating this nonlinear transformation, activation functions play a pivotal role in mediating information transfer and feature extraction, allowing the neural network to abstract features from raw data (Hong, 2023).

If the output obtained from the BPNN differs from the expected result, the actual error is backpropagated through the network. This back propagation step adjusts the weights to minimize the error and improve the accuracy of the network (Yang et al., 2011). This process unfolds as follows: (1) Calculating the Loss Function. At the output layer, the discrepancy between the model's predicted value and the actual target value is computed, representing the loss function's value. (2) Backpropagating the Error. Starting from the output layer, the gradient of the loss function concerning each weight is determined. This is achieved by applying the chain rule to compute gradients at each layer, gradually propagating from the output layer to the input layer. These gradients reveal how the loss function changes when weights are slightly modified. (3) Weight Update. The weights are adjusted using the gradient descent method. The process of forward and backward propagation is iterated repeatedly, and this iterative cycle continues until the network's performance converges to the desired level.

BPNN's advantage lies in its flexibility with various data distributions, enabling effective classification even when working with limited data that doesn't adhere to Gaussian or clustered distributions but follows discrete patterns. However, when confronted with multiple Gaussian or clustered distributions, the high nonlinearity at the boundaries of various data types can reduce the model's fitting capability, leading to longer training times and diminished classification accuracy.

3.6.3. Hyperparameters for classification models

Based on the GMM constructed in this study, five key hyperparameters are listed in Table 3. The number of clusters in the GMM is determined by the number of lithofacies types in the study area. The initial parameters of the GMM, including μ, Σ, and π, are provided by GDA, which has the advantage of speeding up the convergence of the GMM model. A full covariance matrix is chosen because of the high correlation between different types of logging data, allowing the GMM model to effectively capture these correlations for more accurate data modeling. The convergence of the GMM model is evaluated based on the EM, which includes the maximum iterations and the parameter change

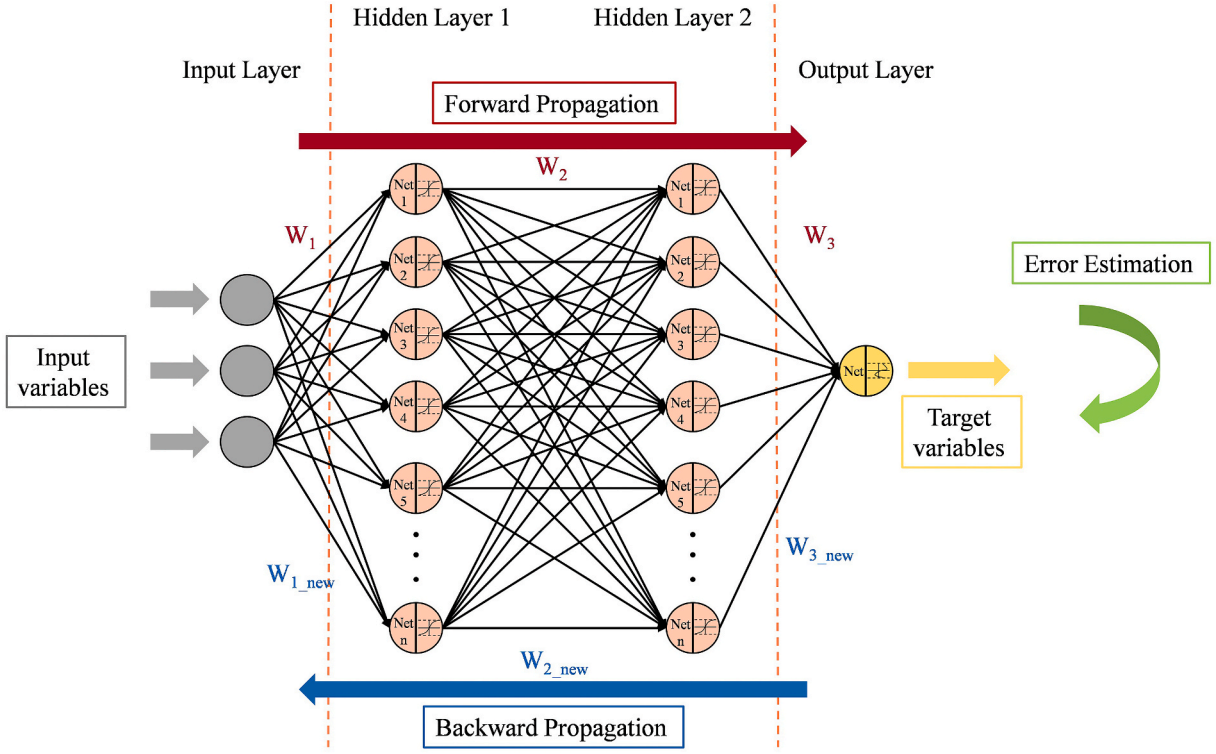

Fig. 5. Calculation process of the BPNN automatic identification method (Modified by Rumelhart et al., 1986; Rogers et al., 1992; Wu et al., 2006).

Table 3
GMM hyperparameters values.

Hyperparameter	Value
Number of Clusters	The number of lithofacies types
Initial Parameters	GDA
Covariance Matrix Type	Full Covariance Max
EM Algorithm Maximum Iterations	1500
EM Algorithm Parameter Change Threshold	0.001

threshold. Based on extensive previous research and experience, the maximum iterations to 1500 is set to ensure a sufficient number of iterations without excessive time consumption. The parameter change threshold is set to 0.001 to increase the precision of the data modeling and ensure that the model reaches a satisfactory level of accuracy.

To achieve the research objectives, the BPNN in this study is configured with hyperparameters suitable for multi-categorization tasks. The choice of categorical cross-entropy loss and gradient descent as the loss function and training function, respectively, is based on the effectiveness of categorical cross-entropy loss in providing gradient information for gradient descent optimization. This allows the neural network to learn the weights and biases more effectively, minimizing the loss function and improving the training efficiency of the model. Considering the inherent data nonlinearity in complex lithofacies, ReLU is selected as the activation function for the hidden layer. To output the probability of data belonging to each lithofacies, allowing for comparison with GMM results, softmax is chosen as the activation function for the output layer. In determining the number of hidden layers, the neurons in the hidden layer, and the learning rate, a series of experiments using the trial-and-error method are conducted. The number of neurons in the output layer depends on the number of lithofacies types. To enhance the model's generalization capability and prevent overfitting, the early-stop method is employed to limit the number of model iterations. Table 4 provides an overview of the essential hyperparameters used by the BPNN.

Table 4
BPNN hyperparameters values.

Hyperparameter	Value
Loss function	Categorical Cross-Entropy Loss
Training function	Gradient descent
Hidden layer activation function	ReLU
Output layer activation function	Softmax
Number of hidden layers	Trial-and-error method
Hidden layer neurons number	Trial-and-error method
Output layer neuron number	The number of lithofacies types
Learning rate	Trial-and-error method
Number of Epochs	Early stopping

4. Results

4.1. Lithofacies types and characteristics

Based on the observations and descriptions of the core samples and core photographs, the lithofacies in the study area have been classified into six types using Miall and Feng's criteria for fluvial systems (Miall, 1986; Feng, 2022). The lithofacies types and their corresponding naming rules are as follows: (1) Massive bedding mudstone (Mm): This lithofacies type is characterized by black mudstone with a prominent development of massive bedding. Refer to Fig. 6a for an illustration. (2) Wavy bedding siltstone (Sw): This lithofacies type consists of gray or black siltstone exhibiting distinct wavy bedding. See Fig. 6b for a visual representation. (3) Parallel bedding fine sandstone (Fp): This lithofacies type comprises fine sandstone with well-defined parallel bedding. Fig. 6c provides an example. (4) Cross-bedding medium sandstone (MSc): This lithofacies type is characterized by medium sandstone displaying plate cross-bedding (Fig. 6d) and trough cross-bedding (Fig. 6e). (5) Cross-bedding coarse sandstone (CSc): This lithofacies type consists of coarse sandstone exhibiting plate cross-bedding (Fig. 6f) and trough cross-bedding (Fig. 6g). (6) Massive bedding conglomerate (Gm): This lithofacies type is represented by a gray conglomerate with a

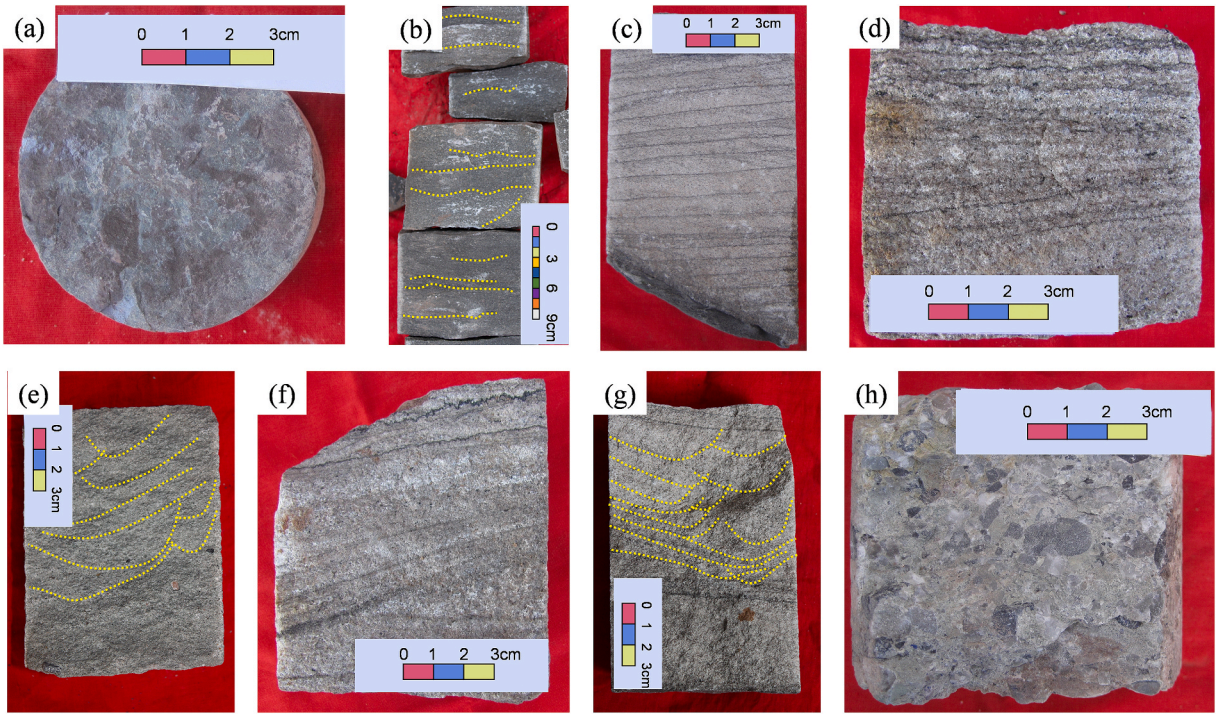

Fig. 6. Some photos of the core. The yellow line in Figure (b) marks the wavy bedding. The yellow lines in Figure (e) and Figure (g) mark the trough cross-bedding.

pronounced development of massive bedding. Refer to Fig. 6h for an example.

These lithofacies classifications are based on Miall and Feng's criteria, taking into account the characteristics observed in the study area.

In Fig. 7a, the frequency distribution of the thickness of Mm exhibits a bimodal pattern. The main peak of the distribution falls within the thickness interval of 0.50 m–0.75 m, while the secondary peak is observed in the interval of 2.00 m to +∞ m. Further analysis reveals that Mm with a thickness of less than 1 m accounts for 53.28% of the total, whereas Mm with a thickness between 1 m and 2 m represents 25.55%. Mm with a thickness exceeding 2 m makes up 21.17% of the total. These results indicate a significant variation in the thickness distribution of Mm, with the majority of occurrences having a small thickness and only a small proportion characterized by a larger thickness.

In Fig. 7b, the frequency distribution of the thickness of Sw displays a single-peaked distribution. The main peak of the distribution is observed within the thickness interval of 0.25 m–0.50 m. Further analysis reveals that Sw with a thickness of less than 1 m accounts for 85.31% of the total, while Sw with a thickness between 1 m and 2 m represents 14.69%. There are no occurrences of Sw with a thickness exceeding 2 m. These findings indicate that the thickness distribution of Sw is relatively less varied and generally thinner compared to other lithofacies. The majority of Sw occurrences have a thickness of less than 1 m, with only a small proportion falling within the range of 1 m–2 m in thickness.

In Fig. 7c, the frequency distribution of the thickness of Fp exhibits a multi-peak distribution. There are three distinct peaks observed, with the corresponding thickness intervals being 0.50–0.75 m, 1.00–1.25 m, and 2.00-+∞ m. Further analysis reveals that Fp with a thickness of less than 1 m accounts for 53.25% of the total, while Fp with a thickness between 1 m and 2 m represents 37.28%. Fp with a thickness exceeding 2 m makes up 9.47% of the total. These findings indicate that the thickness distribution of Fp displays small differences. The majority of Fp occurrences have small thicknesses, a larger portion exhibits medium thicknesses, and a small proportion has larger thicknesses.

In Fig. 7d, the frequency distribution of the thickness of MSc exhibits a multi-peak distribution. There are four distinct peaks observed, with the corresponding thickness intervals being 0.50–0.75 m, 1.00–1.25 m, 1.50–1.75 m, and 2.00-+∞ m. Further analysis reveals that MSc with a thickness of less than 1 m accounts for 38.46% of the total, while MSc with a thickness between 1 m and 2 m represents 47.93%. MSc with a thickness exceeding 2 m makes up 13.61% of the total. These findings indicate that the distribution of MSc thicknesses shows less variation. A larger portion of MSc occurrences has a small thickness, a significant proportion exhibits a medium thickness, and a small portion displays a larger thickness.

In Fig. 7e, the frequency distribution of the thickness of CSc displays a bimodal distribution. There are three distinct peaks observed, with the corresponding thickness intervals being 1.00–1.25 m, 1.50–1.75 m, and 2.00-+∞ m. Further analysis reveals that CSc with a thickness of less than 1 m accounts for 15.84% of the total, while CSc with a thickness between 1 m and 2 m represents 52.97%. CSc with a thickness exceeding 2 m makes up 31.19% of the total. These findings indicate that the thickness distribution of CSc exhibits significant variation. A small percentage of CSc occurrences have a small thickness, while the majority of occurrences have a larger thickness.

In Fig. 7f, the frequency distribution of the thickness of Gm exhibits a bimodal distribution. The main peak of the distribution falls within the thickness interval of 2.00 m to +∞ m, while the secondary peak is observed in the interval of 1.00 m–1.25 m. Further analysis reveals that Gm with a thickness of less than 1 m accounts for 14.35% of the total, while Gm with a thickness between 1 m and 2 m represents 47.22%. Gm with a thickness exceeding 2 m makes up 38.43% of the total. These findings indicate that the distribution of Gm thickness varies greatly. A small portion of Gm occurrences have a small thickness, while the majority exhibit a larger thickness.

4.2. Layer dataset

In this study, data from logging curves of 63 cored wells were sampled and processed for feature selection using normalization, one-way ANOVA, and pearson correlation analysis to construct the layer dataset. Table 5 displays the outcomes of the one-way ANOVA for each type of normalized treated layer data. Among the analyzed variables,

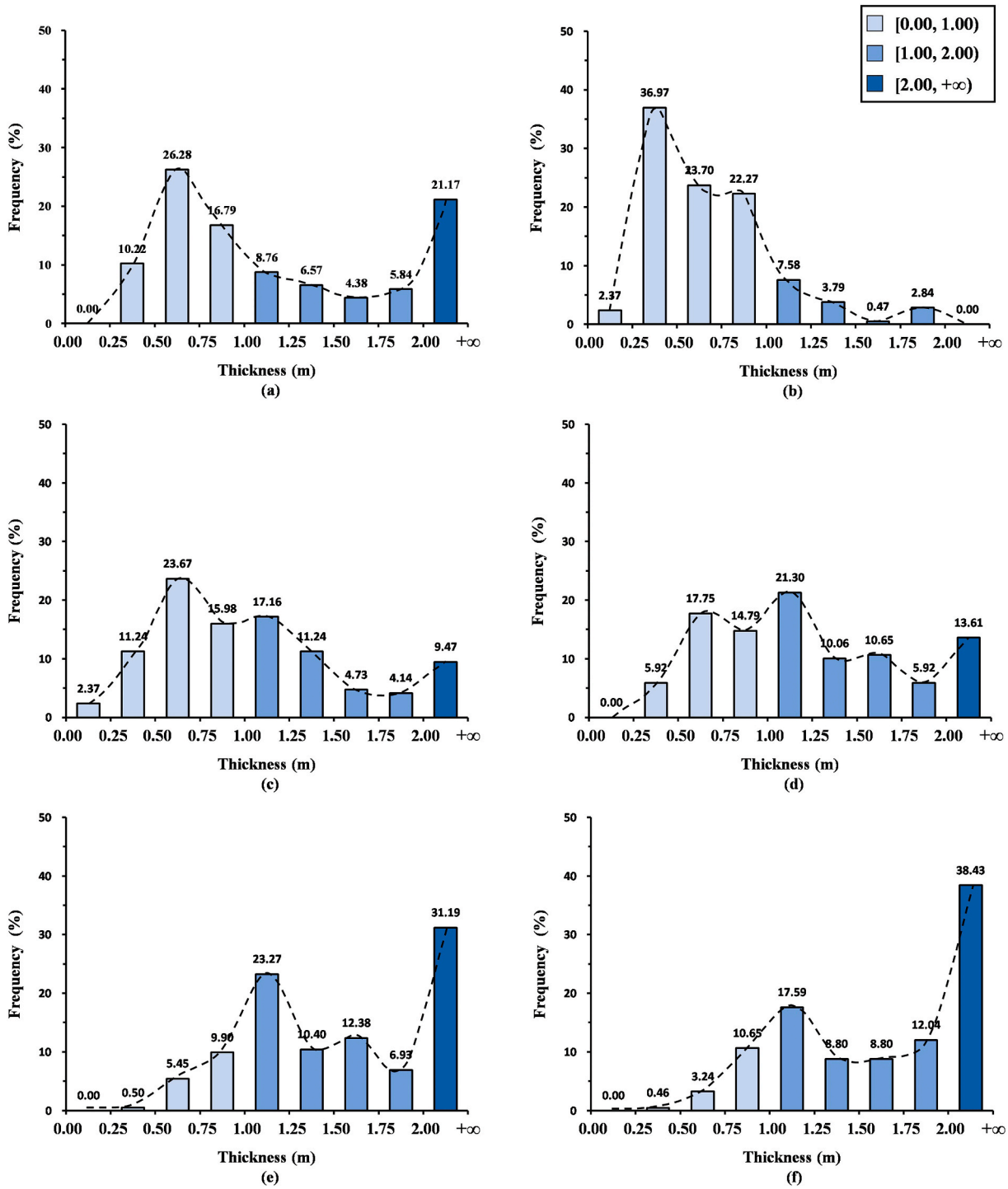

Fig. 7. Frequency distribution histograms representing the thickness of each lithofacies thickness. (a) Frequency distribution histogram of the thickness of Mm; (b) Frequency distribution histogram of the thickness of Sw; (c) Frequency distribution histogram of the thickness of Fp; (d) Frequency distribution histogram of the thickness of MSc; (e) Frequency distribution histogram of the thickness of CSc; (f) Frequency distribution histogram of the thickness of Gm. In this classification, thicknesses greater than or equal to double the logging resolution (2m) are considered to be in the large thickness group, thicknesses less than or equal to double the logging resolution (2m) and greater than the logging resolution (1m) are categorized as part of the medium thickness group, and thicknesses less than or equal to the logging resolution (1m) are classified as part of the small thickness group.

GRn, RHOBn, NPHIn, and PEn exhibited Cohen's f values greater than 0.4 and Partial η^2 values exceeding 0.14. Therefore, these four types of logging data were given preference in the subsequent analysis.

Fig. 8 illustrates the Pearson correlation coefficients between GRn, RHOBn, NPHIn, and PEn. The correlation coefficients reveal that NPHIn exhibits strong correlations with the other variables, particularly with GRn, displaying the highest correlation coefficient. Consequently, GRn, PEn, and RHOBn were selected due to their low correlation with each other.

The ultimate layer dataset consists of 1110 labeled data points, encompassing 182 GM data, 180 CSc data, 171 MSc data, 192 Fp data, 193 Sw data, and 192 Mm data. The data distribution across each type is

Table 5
Results of one-way ANOVA for various types of logging layer data.

Code	SSB	SST	Partial η^2	Cohen's f
GRn	27.968	31.647	0.884	2.757
SPn	3.391	32.824	0.103	0.339
CALIn	0.065	7.165	0.009	0.096
ACn	1.283	13.068	0.098	0.33
RHOBn	1.095	7.131	0.154	0.426
NPHIn	4.95	13.696	0.361	0.752
RLLDn	1.048	24.469	0.043	0.212
RLLSn	1.156	23.652	0.049	0.227
PEn	10.978	31.119	0.353	0.738

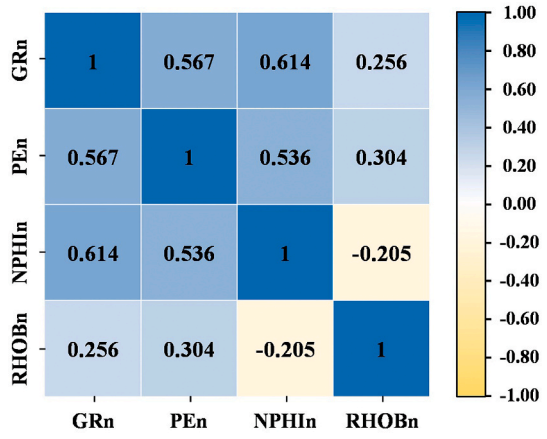

Fig. 8. Heat map of Pearson correlation coefficients for GRn, RHOBn, NPHIn, and PEn.

nearly balanced, mitigating the potential influence of an uneven sample distribution on the performances of the classification models. Additionally, it includes four types of parameters: GRn, RHOBn, PEn, and lithofacies thickness.

4.3. Distribution characteristics of layer data for each lithofacies

In this section of the study, frequency distribution density curves and box plots were employed to illustrate the distribution patterns of GRn, RHOBn, and PEn layer data for each lithofacies type in the layer dataset. Subsequently, K, σ, and K/GS are utilized to quantitatively assess the level of data concentration and dispersion. Based on the σ and K/GS values of GRn, RHOBn, and PEn for each lithofacies presented in Table 6, Table 7, and Table 8, respectively, as well as the data distribution characteristics shown in Fig. 9a and b, and Fig. 9c, the concentration of log layer data for each lithofacies type can be ranked as follows: For GRn: CSc > Gm > MSc > Fp > Mm > Sw; for RHOBn: CSc > Gm > MSc > Fp > Mm > Sw; for PEn: CSc > Gm > MSc > Fp > Mm > Sw.

In summary, the distribution of layer data for each lithofacies type can be characterized as follows: (1) The layer data for CSc and Gm is centrally distributed. (2) The layer data for MSc is relatively centrally distributed. (3) Thelayer data for Fp is relatively discretely distributed. (4) The layer data for Sw and Mm is discretely distributed.

Table 6
K, σ, GS, and K/GS for GRn layer data for each lithofacies.

Code	Mm	Sw	Fp	MSc	CSc	Gm
σ	0.093	0.114	0.089	0.044	0.036	0.039
K	2.723	3.256	3.810	4.149	2.688	2.088
GS	0.5	0.65	0.5	0.3	0.25	0.2
K/GS	5.446	5.009	7.62	13.83	10.752	10.44

Table 7
K, σ, GS, and K/GS for RHOBn layer data for each lithofacies.

Code	Mm	Sw	Fp	MSc	CSc	Gm
σ	0.108	0.145	0.082	0.069	0.049	0.050
K	3.257	3.014	4.054	4.610	4.078	4.440
GS	0.6	0.7	0.5	0.45	0.3	0.4
K/GS	5.428	4.306	8.108	10.244	13.593	11.100

Table 8
K, σ, GS, and K/GS for PEn layer data for each lithofacies.

Code	Mm	Sw	Fp	MSc	CSc	Gm
σ	0.081	0.132	0.100	0.065	0.059	0.047
K	3.765	3.122	4.516	4.226	4.190	4.509
GS	0.45	0.7	0.55	0.35	0.25	0.3
K/GS	8.367	4.460	8.211	12.074	16.760	15.030

4.4. Performance evaluation for classification models

Before beginning model training, it is necessary to provide a comprehensive description of the datasets utilized. The test set comprises layer data from 3 core wells with increased core lengths, while the labeled layer dataset consists of layer data from the remaining 60 core wells. To train both the GMM and BPNN models, we extracted the GRn, RHOBn, and PEn features from the labeled layer dataset. Randomly selected 70% of the layer data among the labeled layer dataset as the train set and the remaining 30% of the layer data as the validation set. This method of creating the training, validation, and test sets has multiple benefits. Following this methodology allows the model to make the most of an adequate amount of training data, resulting in a good fit without the possibility of underfitting. In addition, the considerable quantity of validation data helps avoid model overfitting while assisting the efficient iterative adjustment of model parameters. Additionally, when the model finishes its training and exhibits robust performance on the test set, it indicates its outstanding ability to generalize.

4.4.1. Gaussian mixture model training result

After convergence, the model provides precision, recall, and F1-score for the training, cross-validation, and test sets. Additionally, it presents these metrics for each lithofacies in the test set, along with the predicted probabilities of each lithofacies for each layer data. Based on the evaluation metrics presented in Table 9, it is evident that the overall training of the model is poor. Furthermore, analyzing the evaluation metrics for each lithofacies (Table 9 and Fig. 10), we observe that Gm and CSc exhibit the best training effects, MSc shows a comparatively better training effect, Fp and Mm demonstrate a weaker training effect, while Sw exhibits the poorest training effect.

4.4.2. Back-propagation neural network training result

Once the model converges, it generates precision, recall, and F1-score for the training, cross-validation, and test sets. Additionally, it provides these metrics for each lithofacies in the test set, along with the predicted probabilities of each lithofacies for each layer data. Based on the evaluation metrics presented in Table 10, it is evident that the overall training of the model is poor. Furthermore, analyzing the evaluation metrics for each lithofacies (Table 10 and Fig. 11), we observe that Mm demonstrates the best training effect, followed by Sw and Fp showing a comparatively better training effect. MSc and CSc, on the other hand, exhibit poorer training effects, while Gm displays the worst training effect.

4.5. Relationship between lithofacies characteristics and logging data distribution characteristics

Combining the results of Sections 4.1 and 4.3, the study revealed that

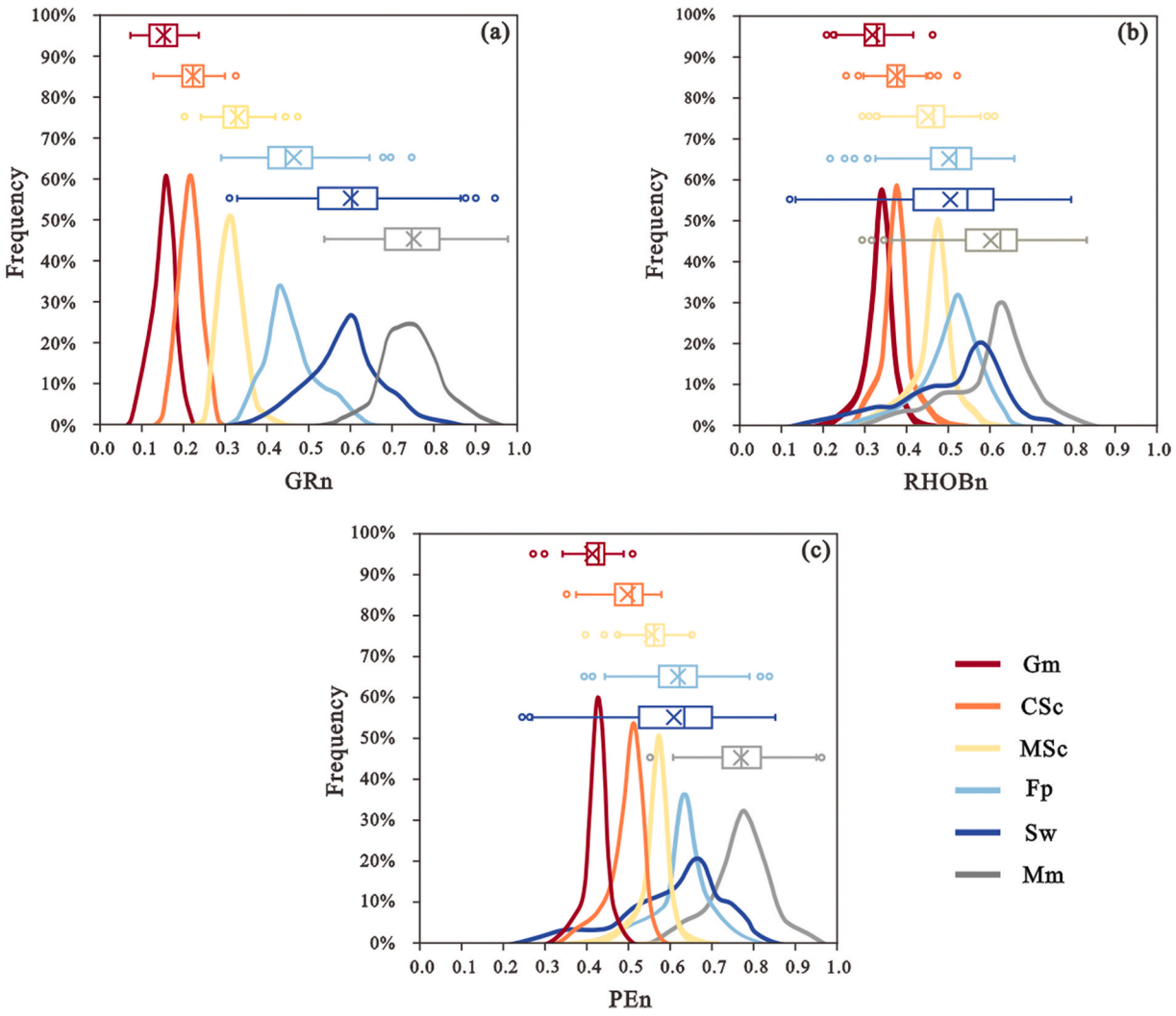

Fig. 9. Frequency distribution density curves and box plots of layer data for GRn, RHOBn, and PEn for each lithofacies. (a) Frequency distribution density curves and box plots of layer data of GRn for each lithofacies. (b) Frequency distribution density curves and box plots of layer data for RHOBn for each lithofacies. (c) Frequency distribution density curves and box plots of layer data for PEn for each lithofacies.

Table 9
Evaluation metrics of GMM training.

Evaluation metrics	Train set	Validation set	Test set	Mm	Sw	Fp	MSc	CSc	Gm
Precision	0.82	0.82	0.76	0.69	0.61	0.72	0.78	0.89	0.87
Recall	0.82	0.81	0.77	0.67	0.63	0.71	0.82	0.91	0.88
F1	0.82	0.81	0.76	0.68	0.62	0.71	0.80	0.90	0.87

lithofacies thickness influences layer data distribution. To clarify this relationship, correlations were established between the mean thickness of each lithofacies and the σ and K/GS of GRn, RHOBn, and PEn for each lithofacies. Fig. 10a illustrates a negative correlation between the mean thickness of each lithofacies and the σ of GRn, RHOBn, and PEn for each lithofacies. Fig. 10b illustrates a positive correlation among the mean thickness of each lithofacies and the K/GS of GRn, RHOBn, and PEn for each lithofacies. The relationship presented above indicates that the data sampling method developed in this study can effectively reduce deviations and regularize the data distribution to a certain extent (Figs. 7, 9 and 10).

4.6. Relationship between data distribution characteristics and performances of classification models

Building upon the findings of Section 4.3 and Section 4.4, the correlation between the characteristics of data distribution and the classification performances of the models was investigated.

Further counting the probabilities of each lithofacies in the GMM-trained test set and plotting the histograms of the probability distributions (Fig. 11) which show that Gm and CSc have a high probability of being single-peaked, MSc has a medium-to-high probability of being single-peaked, Fp and Mm have a medium-to-low probability of being single-peaked, and Sw has a low probability of being single-peaked.

Further counting the probabilities of each lithofacies in the BPNN-trained test set and plotting the histograms of the probability distributions (Fig. 12) which show that Mm has a high probability of being

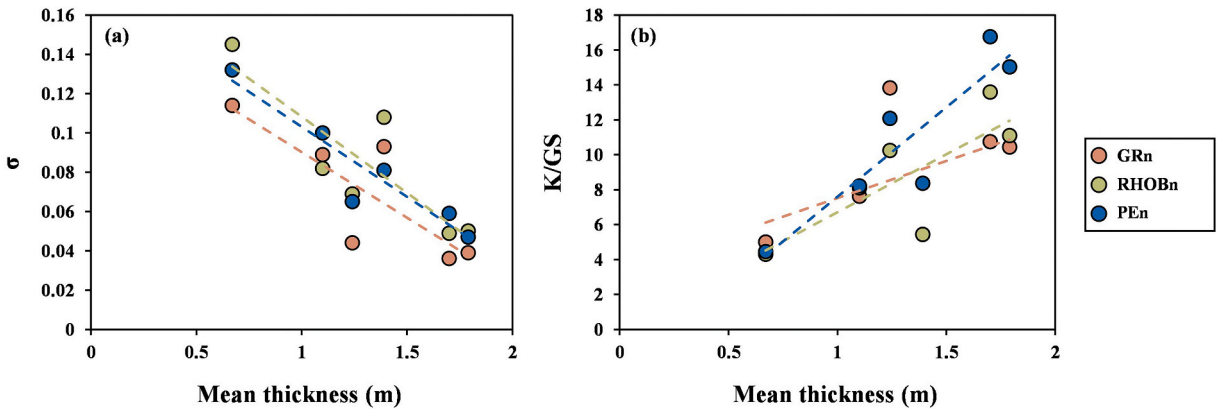

Fig. 10. Correlation of mean thickness with σ and K/GS. (a) Correlation of mean thickness with σ. (b) Correlation of mean thickness with K/GS.

Table 10
Evaluation metrics of BPNN training.

Evaluation metrics	Train set	Validation set	Test set	Mm	Sw	Fp	MSc	CSc	Gm
Precision	0.80	0.79	0.76	0.96	0.80	0.79	0.68	0.67	0.64
Recall	0.79	0.78	0.77	0.88	0.87	0.76	0.67	0.74	0.71
F1	0.79	0.78	0.76	0.92	0.83	0.77	0.67	0.70	0.68

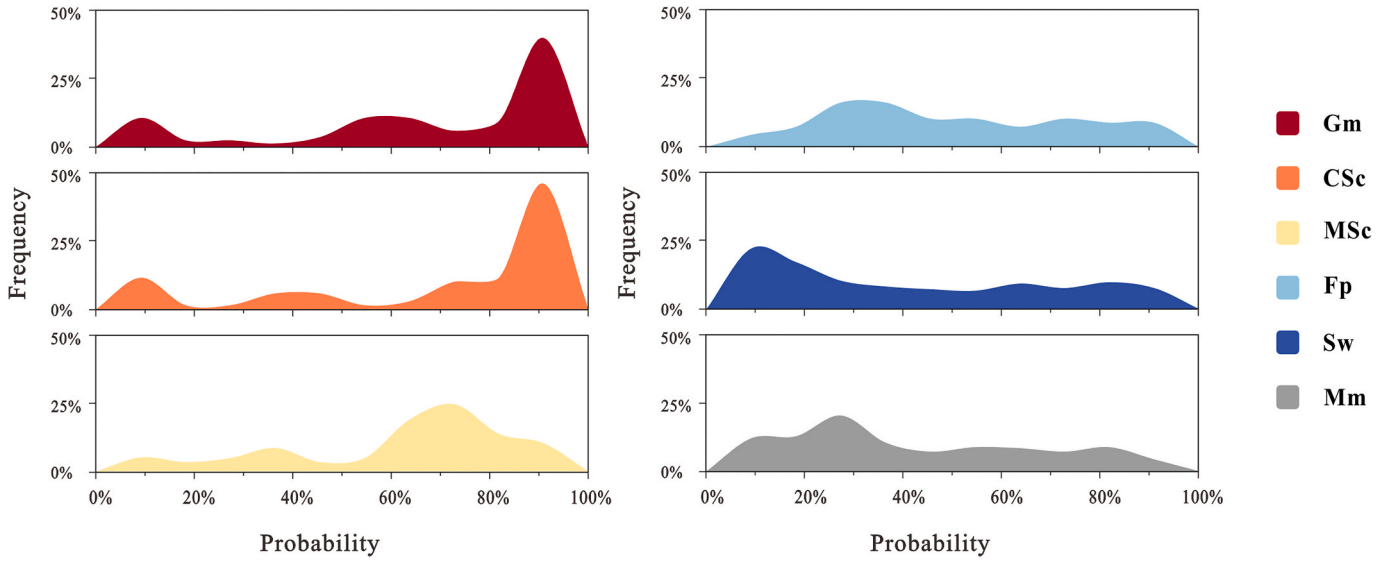

Fig. 11. Frequency distribution density curves of GMM's predicted probabilities for each lithofacies in the test set.

single-peaked, Sw has a high probability of being single-peaked, Fp has a medium probability of being single-peaked, MSc, CSc, and Gm have a medium-to-low probability of being single-peaked.

The aforementioned findings reveal that GMM is more appropriate for Gm and CSc in dataset distribution, whereas BPNN is better suited for Sw and Mm in discrete distribution. It should be noted that GMM and BPNN are not absolutely applicable to MSc and Fp. For MSc with more concentrated data distribution, GMM presents better classification performance than BPNN. In contrast, for Fp with a more discrete distribution, BPNN exhibits superior classification performance than GMM.

To make the relationship between the data distribution characteristics and the performances of the classification models more explicit, plots of the correlation between the mean σ and the mean K/GS of the logging data for each lithofacies and the mean F1 of the three datasets for the two models are created. Fig. 13a shows a good negative correlation between the mean σ and the mean F1 of GMM, and Fig. 13b shows a good positive correlation between the mean K/GS and the mean F1 of GMM. These findings indicate that the GMM has a good classification performance for centrally distributed data, while its performance becomes progressively worse as the data becomes more discrete.

Fig. 14a shows a good positive correlation between the mean σ and the mean F1 of BPNN, and Fig. 14b shows a good negative correlation between the mean K/GS and the mean F1 of BPNN. These findings indicate that the BPNN has a good classification performance for discretely distributed data, while its performance becomes progressively worse as the data becomes more central.

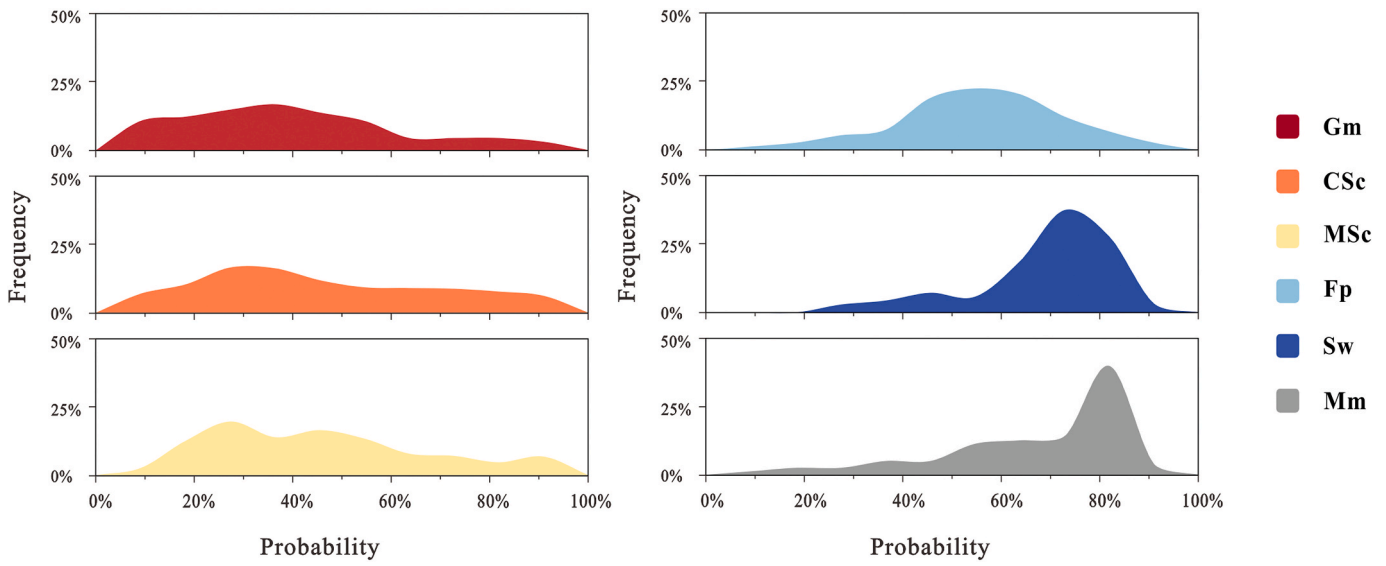

Fig. 12. Frequency distribution density curves of BPNN's predicted probabilities for each lithofacies in the test set.

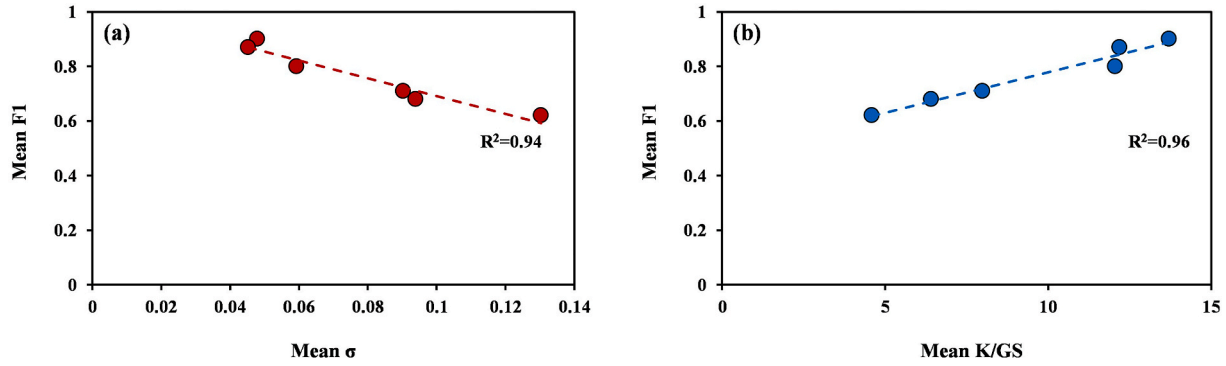

Fig. 13. Correlation of mean σ and mean K/GS with mean F1 for three datasets in GMM training. (a) Correlation of mean σ with mean F1. (b) Correlation of mean K/GS with mean F1.

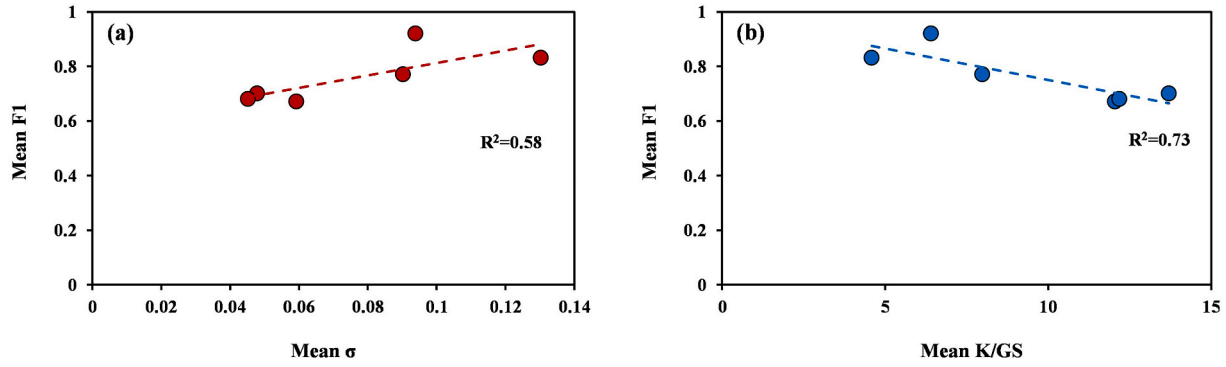

Fig. 14. Correlation of mean σ and mean K/GS with mean F1 for three datasets in BPNN training. (a) Correlation of mean σ with mean F1. (b) Correlation of mean K/GS with mean F1.

5. Discussion

5.1. The influence of sampling techniques and feature engineering on the performances of classification models

The original logging data were processed according to the dataset partitioning method described in Section 4.4, and the GMM and BPNN were trained separately. Tables 11 and 12 show the rather poor

Table 11
Evaluation metrics for GMM trained with original logging data.

Evaluation metrics	Train set	Validation set	Test set
Precision	0.58	0.32	0.27
Recall	0.54	0.30	0.28
F1	0.56	0.31	0.27

Table 12
Evaluation metrics for BPNN trained with original logging data.

Evaluation metrics	Train set	Validation set	Test set
Precision	0.62	0.44	0.19
Recall	0.57	0.46	0.18
F1	0.59	0.45	0.18

classification performances of the GMM and BPNN on the original logging data compared to the classification performances of the GMM and BPNN shown in Tables 9 and 10, where the F1 of the training, validation, and test sets is less than 0.6. The classification models input to the sampled and feature-engineered processed data are significantly better than the classification models input to the original logging curves for the following reasons: (1) Sampling processing minimizes the influence of adjacent lithofacies. The data collection from the logging instrument is impacted by adjacent lithofacies and can result in deviations, causing inconsistent distribution of logging data for lithofacies with varying thicknesses of the same type. This inconsistency further increases the difficulty of classification modeling. The sampling method proposed in this study can reduce the clutter of data distribution through the use of the weighted mean method. (2) The standardization process in feature engineering makes all types of logging data of the same order of magnitude. Different types of original logging data have different orders of magnitude, with GR, SP, CAL, AC, NPHI, RLLD, and RLLS data having relatively large orders of magnitude (0–1000), while RHOB and PE usually have smaller orders of magnitude (0–5). The existence of such a large order of magnitude difference in the input data can cause the loss function in the classification model to be sensitive to features with a larger order of magnitude while ignoring features with a smaller order of magnitude, which in turn leads to overfitting and affects the classification performance. This is the reason for the large difference in F1 among the training, validation, and test sets in Tables 11 and 12. (3) The ANOVA and Pearson correlation analysis in the feature engineering filtered out the interference of noisy data and screened out the logging data that responded well to the lithofacies. The RLLD and RLLS data mainly responded well to the formation fluids but poorly to the lithofacies, and if they were inputted into the models, they might mislead the model to learn the pattern of the fluid instead of the lithofacies. The SP data were greatly affected by formation water and The SP data are strongly influenced by formation water and mud mineralization, making it difficult to respond correctly to lithofacies. AC and NPHI data are impacted by formation gas and water saturation. The 'cycle skip' phenomenon of AC and the 'excavation effect' of NPHI affect the data accuracy. Conversely, GR, RHOB, and PE are less affected by other formation factors and can better respond to important parameters, such as shale content, rock density, and mineral composition, allowing for the distinction of different lithofacies (Iraji et al., 2023a).

5.2. Reasons for the influence of lithofacies characteristics on data distribution characteristics

To determine why lithofacies thickness impacts layer data distribution, additional research was conducted by examining logging resolution and lithofacies thickness from the perspective of previous researchers' recognized viewpoints and research findings (Lindberg et al., 2015; Iraji et al., 2023a). After synthesizing the characteristics of each lithofacies and the distribution characteristics of the logging data, that the thickness of the lithofacies has a significant effect on the distribution of the logging data is determined. This is because the data collected and processed by the logging instrument represents the cumulative logging response within a specific spatial area (Lindberg et al., 2015; Tian et al., 2021). The presence of adjacent lithofacies impacts the process, leading to deviations from the true values. The magnitude of these deviations is determined by the resolution of the logging instrument and the relative thickness of the lithofacies (Chen et al., 2021; Wang et al., 2023).

The frequency ratios of lithofacies with thicknesses less than 1 m (logging resolution), between 1 m and 2 m (double logging resolution), and greater than 2 m in the layer dataset is calculated. Subsequently, correlations among the frequency ratio of each lithofacies and the σ and K/CF of GRn, RHOBn, and PEn for each respective type of lithofacies is established. Fig. 15 illustrates that the frequency ratio of data with a thickness less than 1 m in each lithofacies shows a positive correlation with the σ of GRn, RHOBn, and PEn for that type of lithofacies (Fig. 15a, b, and c). It also shows a positive correlation with K/CF (Fig. 15d, e, and f). Conversely, the frequency ratio of data with thickness between 1m and 2m, as well as data with thickness greater than 2m in each lithofacies, shows a negative correlation with the σ of GRn, RHOBn, and PEn for that type of lithofacies (Fig. 15a, b and c). It also shows a negative correlation with K/CF (Fig. 15d, e, and f). The observed result can be explained as follows: The observed result can be explained as follows: When the thickness of a lithofacies is greater than the resolution of the logging instrument, the data acquired by the logging instrument is minimally affected by the adjacent lithofacies. This leads to smaller deviations between the acquired data and the actual lithofacies value. As a result, the layer data of that type of lithofacies will be close to the true value of that type of lithofacies. As a result, the distribution of layer data within the dataset of this type of lithofacies becomes more centralized as the proportion of layer data with thicknesses greater than the logging resolution increases (Fig. 15). However, as the lithofacies thickness gradually decreases and falls below the logging resolution, the logging instrument data acquisition becomes increasingly influenced by the adjacent lithofacies. This in turn leads to larger deviations between the collected data and the actual lithofacies value. The deviation of the layer data for a certain type of lithofacies from the actual value of that type of lithofacies also increases as a consequence. As a result, the distribution of layer data within the dataset of this type of lithofacies becomes more dispersed as the proportion of layer data with thicknesses less than the logging resolution increases (Fig. 15).

It can be seen that the distribution of lithofacies thickness affecting layer data is concentrated or discrete depending on the relative size of lithofacies thickness and logging resolution.

5.3. Relationship between lithofacies characteristics, layer data distribution characteristics, and performances of classification models

Based on the results in Sections 4.5 and 4.6 and the conclusions in Section 5.2, it can be seen that the lithofacies thickness affects the data distribution and the data distribution affects the classification models performance. It is necessary to further discuss the effect of facies thickness on the performances of the classification models and the relationship between lithofacies thickness, the distribution of layer data, and the performances of the classification models. Fig. 16 illustrates the relationship between lithofacies thickness and the probability of lithofacies identification for GMM and BPNN using layer data from three untrained well cores.

GMM shows a higher identification probability for each lithofacies type as thickness increases (Fig. 16a). As the lithofacies thickness increases beyond 1 m, the identification probability is greater than 80% and approaches 90%. As the lithofacies thickness increases beyond 1.8 m, the identification probability approaches 95% and remains stable. It is worth noting that the 1m mentioned above is the resolutions of the three types of logging data. The above phenomenon occurs because the GMM generative model achieves classification by learning the probability distribution of the input data.

Table 9 illustrates the GMM's training results: Gm and CSc perform very well, MSc is better, Fp and Mm have poor results, and Sw has the worst results. GMM is appropriate for data with a concentrated and stable distribution but not for discrete data (Blundell and Bond, 2000; Gm et al., 2020; Jiao et al., 2022). As stated in Section 5.2, the probability distribution of the lithofacies thickness control layer data is

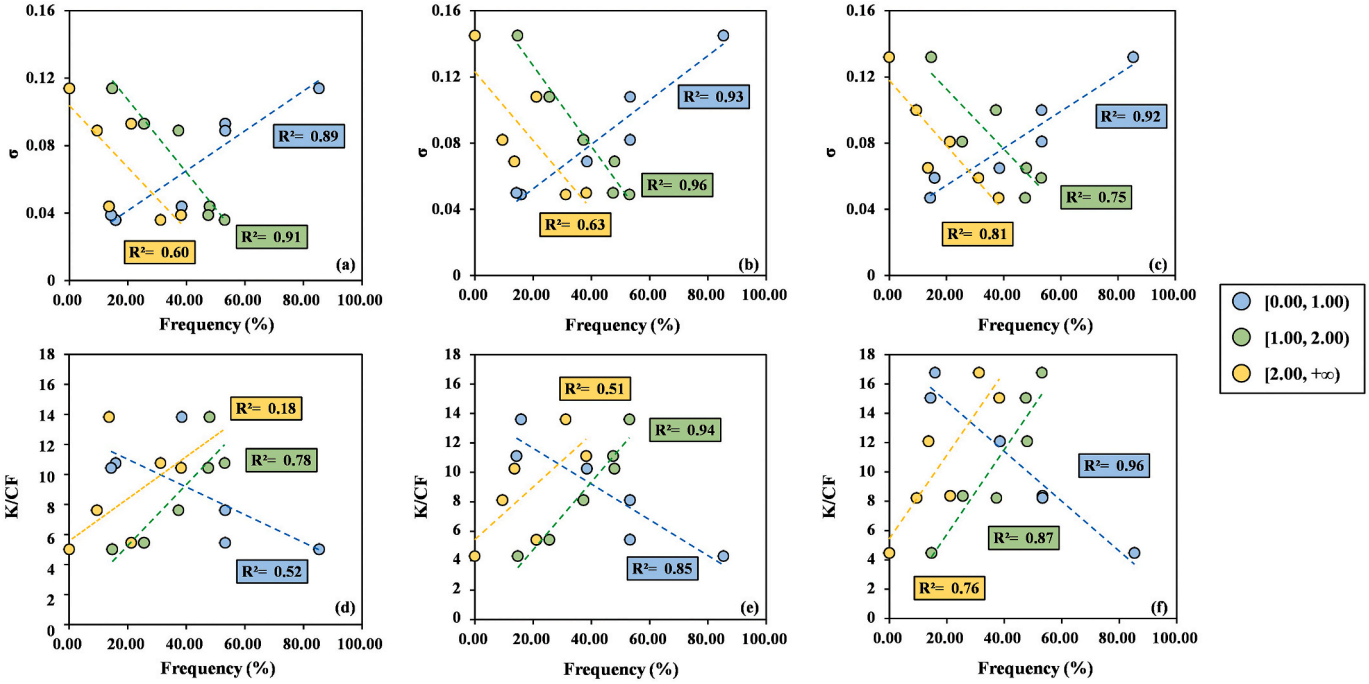

Fig. 15. Correlation of the percentage of the thickness of each lithofacies less than logging resolution, greater than logging resolution less than double logging resolution, and greater than double logging resolution with σ and K/GS for GRn, RHOBn, and PEn. (a) is the correlation between the frequency percentage of the thickness of each lithofacies and the σ of GRn. (b) is the correlation between the frequency percentage of the thickness of each lithofacies and the σ of RHOBn. (c) is the correlation between the frequency percentage of the thickness of each lithofacies and the σ of PEn. (d) is the correlation between the frequency percentage of the thickness of each lithofacies and the K/GS of GRn. (e) is the correlation between the frequency percentage of the thickness of each lithofacies and the K/GS of RHOBn. (f) is the correlation between the frequency percentage of the thickness of each lithofacies and the K/GS of PEn. Thickness intervals with less than logging resolution are 0.00–1.00m, thickness intervals with greater than logging resolution and less than double logging resolution are 1.00–2.00m, and thickness intervals with greater than double logging resolution are 2.00-+∞m.

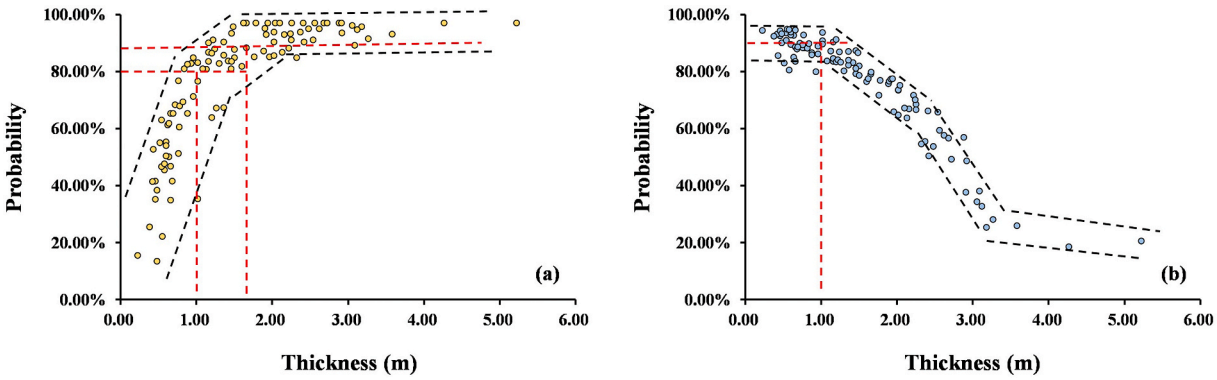

Fig. 16. Correlation between lithofacies thickness and the predicted probability of the classification model for lithofacies in the coring section of 3 wells that did not participate in the model training. (a) Correlation between lithofacies thickness and GMM's predicted probability of lithofacies in the coring section of three wells that did not participate in model training. (b) Correlation between lithofacies thickness and BPNN's predicted probability of lithofacies in the coring section of three wells that did not participate in model training.

primarily affected by logging resolution. The data distribution becomes more concentrated as the thickness exceeds the logging resolution. When the thickness is greater than the logging resolution and increases, the distribution of the layer data of each lithofacies becomes more and more concentrated and the probability distribution becomes more and more stable, and the higher the probability that the data in the test set meets the distribution of the layer data of the GMM trained that the data belongs to the type of its real lithofacies. The layer data for Gm and CSc typically have a thickness greater than 1 m (Fig. 7e and f), resulting in concentrated layer data distributions and stable probability distributions. Consequently, the GMM demonstrates a high probability identification effect for these layers (Fig. 11). The layer data for MSc mostly have a thickness of less than 1 m (Fig. 7d), leading to a relatively concentrated distribution and stable probability distribution. However, the layer data with a thickness of less than 1 m have a relatively discrete distribution and unstable probability distribution. As a result, the GMM shows a mostly medium-high probability identification effect and a small portion of low-probability identification effect for MSc (Fig. 11). For Fp and Mm, the majority of the layer data has a thickness of less than 1 m (Fig. 7a and c). Consequently, this part of the layer data has a relatively discrete distribution and unstable probability distribution. On the other hand, the layer data with a thickness greater than 1 m has a relatively centralized distribution and stable probability distribution. Therefore, the GMM exhibits mostly medium-low probability

identification effects and a small portion of medium-high probability identification effects for Fp and Mm (Fig. 11). Regarding Sw, the layer data predominantly has a thickness of less than 1 m (Fig. 7b). As a result, it displays discrete layer data distributions and unstable probability distributions. Consequently, the GMM performs low probability identification effects on Sw (Fig. 11).

BPNN shows a higher identification probability for each lithofacies as the thickness decreases (Fig. 16b). As the lithofacies' thickness decreases to less than 1 m, the identification probability tends to approach 90% and remains stable. The reason for the above phenomenon is that the discriminative model BPNN achieves classification by fitting boundaries between the input data (Gm et al., 2020; Otchere et al., 2021). which has the advantage of showing good classification performance even in the face of discretely distributed data, whereas according to the conclusion in 5.2, the thickness of the lithofacies is less than the logging resolution resulting in a discrete distribution of the layer data. When the thickness is less than 1 m and decreasing, the distribution of layer data for each lithofacies becomes more and more discrete, and the higher the probability that the data in the test set matches the BPNN trained to fall within the boundary of the layer data for its true lithofacies type. The majority of layer data for MSc, CSc, and Gm have a thickness greater than 1 m (Fig. 7d, e, and f), resulting in concentrated layer data distributions. However, due to the similarity in logging response among these lithofacies, the classification boundaries for these three types of data become highly complex. This complexity often leads the BPNN to encounter issues of overfitting or underfitting, resulting in low identification probabilities (Fig. 12). Nevertheless, some discretely distributed layer data enable the BPNN to fit appropriate classification boundaries more easily, leading to higher identification probability (Fig. 12). For Fp, most of the layer data have a thickness greater than 1 m (Fig. 7c), resulting in a relatively discrete distribution. Despite the similarity in logging response with other lithofacies, the sporadic distribution of layer data makes it comparatively easier for the BPNN to fit suitable classification boundaries and achieve medium-probability identification (Fig. 12). For Sw, primarily consists of layer data with a thickness less than 1 m (Fig. 7b). Consequently, its layer data are discretely distributed. Moreover, the logging response of Sw differs significantly from that of other sandstone lithofacies, making it easier for the BPNN to fit an appropriate classification boundary. As a result, the BPNN exhibits a high probability of identification for Sw. Although the logging responses of Sw and Mm are similar, the relatively discrete distribution of layer data makes it relatively easy for the BPNN to fit appropriate classification boundaries. Most of the layer data for Mm have a thickness of less than 1 m (Fig. 7a), resulting in a relatively discrete layer data distribution. The considerable disparity in logging response between Mm and other sandstone lithofacies further facilitates the BPNN in fitting appropriate classification boundaries. Consequently, the BPNN demonstrates a high probability of identification for Mm (Fig. 12). Despite the similar logging responses of Mm and Sw, their layer data are discretely distributed. This discrete distribution enables the BPNN to effectively identify both lithofacies with a high probability (Fig. 12).

It can be seen that the thickness of the lithofacies is based on the logging resolution that directly affects the distribution of the layer data and thus indirectly affects the performances of the classification models.

5.4. GMM-BPNN establishment and application effect analysis

Based on the effect of lithofacies thickness on the performances of the classification models, a gaussian mixture model-backpropagation neural network hybrid classification model (GMM-BPNN) was built to overcome the limitations of GMM and BPNN in lithofacies classification for layer data with varying distributions due to differences in thickness. This model combines the strengths of both approaches to improve the accuracy and robustness of lithofacies identification.

The effect of thickness on the performances of the classification models is discussed in Section 5.3, but no specific lithofacies thickness thresholds (T) are given that explicitly classify the layer datasets appropriately into centrally distributed datasets (C datasets) and discretely distributed datasets (D datasets). To determine the optimal T, a process is needed to achieve the best classification results for both GMM and BPNN models individually (Fig. 17). By determining this optimal threshold, the GMM-BPNN can achieve the best classification results overall.

The process of training GMM-BPNN is illustrated in Fig. 17, which is divided into the following steps: (1) By using 0 m as the starting T, the layer dataset into two datasets can be divided: C dataset for data greater than the threshold and D dataset for data less than the threshold. (2) The layer data of the C dataset is used as input for training the GMM model, with 70% of the layer data used for training and the remaining 30% used for validation. Similarly, the layer data of the D dataset is inputted into the BPNN model, with 70% of the layer data used for training and 30% for validation. Both models are trained until convergence. (3) The layer data from the three untrained wells is utilized as a test set to evaluate the trained model. The F1 is calculated based on the output of the test set. The model compare this F1 with the F1 obtained from the previous training. If the F1 from the current training is greater than or equal to the previous training, the model will increment the T from Step 1 by 0.1m and repeat Steps 1 and 2. Conversely, if the F1 is less than the previous training, the model will output the T from the previous iteration and consider the training of the hybrid classification model complete. According to Fig. 18, the output test set F1 varies with lithofacies thickness, peaking at 0.95 when the thickness is 1.8 m. This indicates that the model achieves its best classification performance at this lithofacies thickness.

Table 13 illustrates the identification results of three models: GMM, BPNN, and GMM-BPNN. It shows that GMM-BPNN outperforms both GMM and BPNN in terms of identification accuracy. Fig. 19 illustrates the lithofacies identification results of the three models (GMM, BPNN, and GMM-BPNN) for the destination strata core section of well A1, which has the longest core length among the three untrained wells. GMM-BPNN demonstrates more precise results for both thin and thick lithofacies. Moreover, the vertical stacking order of the lithofacies identification is more consistent with fluvial deposition. This improvement can be attributed to GMM-BPNN utilizing the strengths of both GMM and BPNN models to classify both centrally and discretely distributed data. The layer data probability distributions for the thicker lithofacies are stable and consistent with those given by the GMM trained on centrally distributed layer data. The layer data distribution of the thinner lithofacies is discrete but falls within the decision boundary fitted by the BPNN trained with discretely distributed layer data.

Besides, the GMM-BPNN was applied to the lithofacies identification of the fluvial deltaic reservoirs in the western part of the Sulige gas field in the Ordos Basin. Table 14 shows the results of training the model using 10 core wells and 2 coring wells used for blind testing. The limitation of data volume leads to the application effect in this area is not as good as in this study area, but the F1 of the blind test set reaches more than 0.8, which also indicates that the GMM-BPNN has good generalization. The limitation of data volume leads to the application effect in this area is not as good as in this study area, but the F1 of the blind test set reaches more than 0.8, which also indicates that the GMM-BPNN has good generalization. Such results further prove that the idea of establishing a classification model proposed in this study as well as the choice of the model are reasonable and can solve both the lithofacies prediction problem with a large amount of labeled data as well as the lithofacies prediction problem with a limited amount of labeled data.

6. Conclusion

In this study, a log sampling method has been developed to obtain layer data that effectively, mitigateing the effects of adjacent lithofacies. Based on the layer data, it is obseved that lithofacies thickness based on

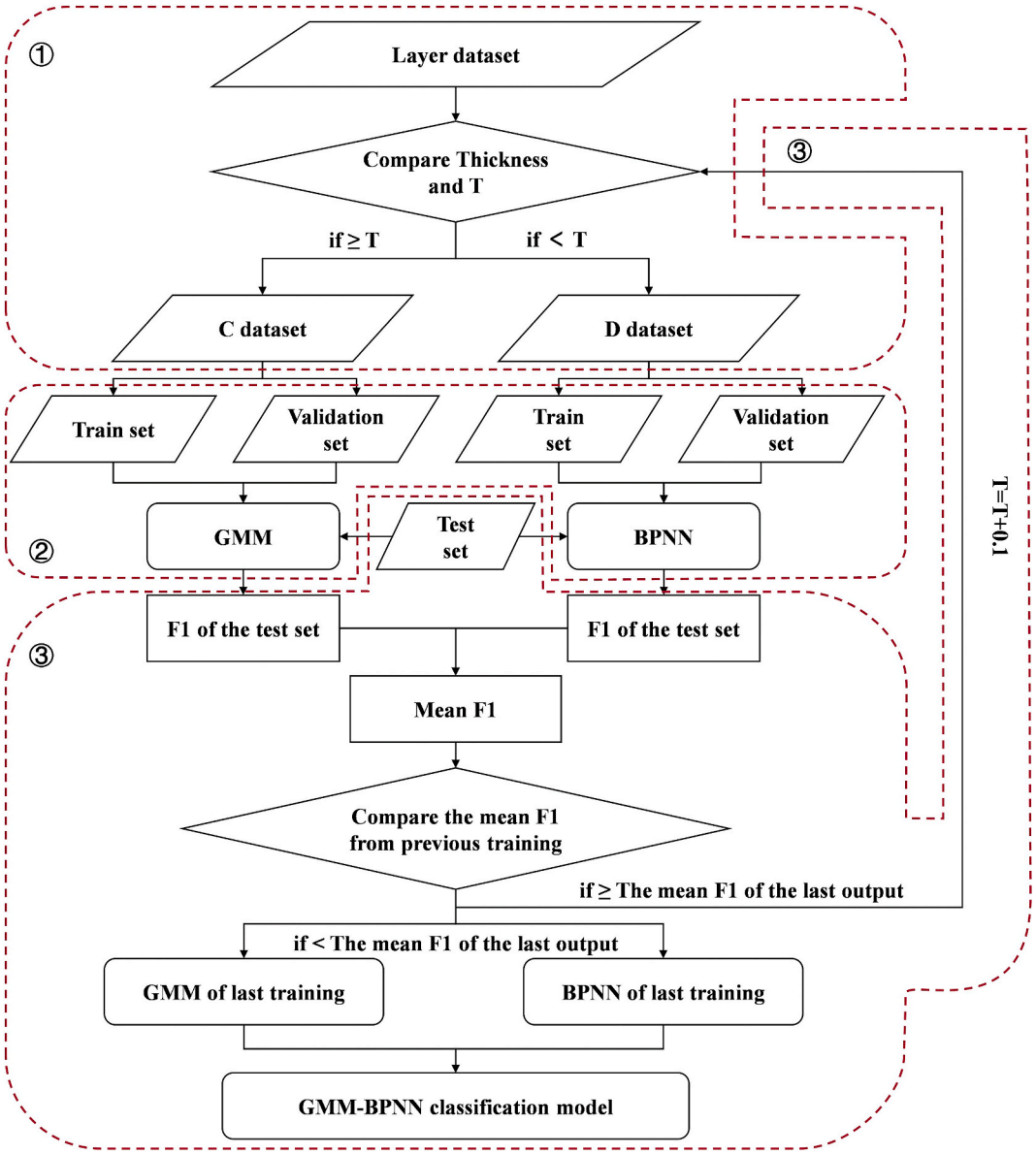

Fig. 17. The process of training GMM-BPNN.

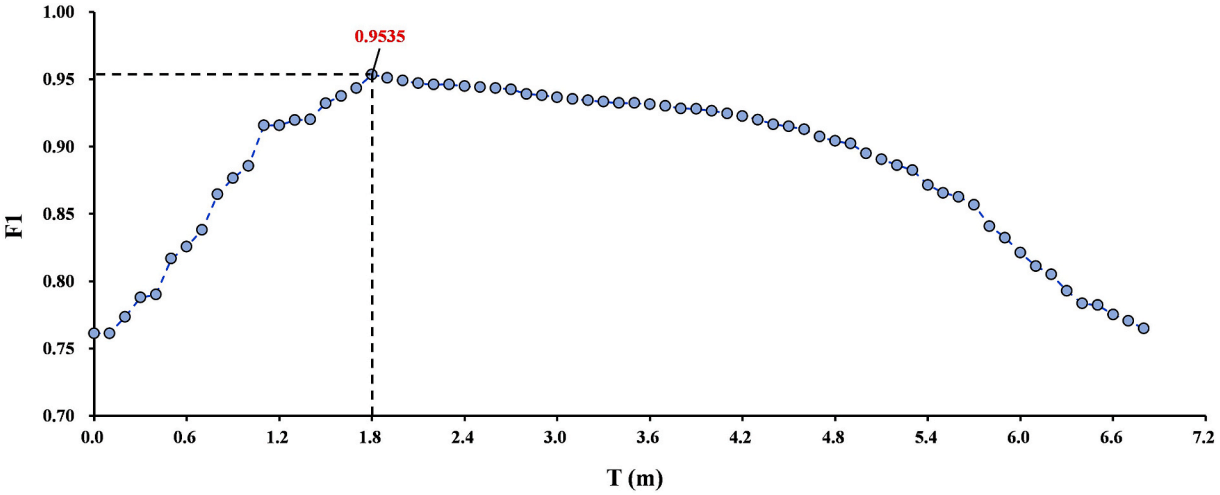

Fig. 18. Correlation of lithofacies thickness thresholds with F1 of the GMM-BPNN output test set.

Table 13
Evaluation metrics for GMM, BPNN, and GMM-BPNN.

Evaluation metrics	GMM	BPNN	GMM-BPNN
F1	0.76	0.77	0.95

logging resolution directly affects the distribution characteristics of the layer data; for a given lithofacies, the more the lithofacies thickness exceeds the resolutions of the logging curves, the more centralized the overall data for that lithofacies becomes. Conversely, when the lithofacies thickness is less than the resolutions of the logging curves, the data becomes more discrete. The thickness of lithofacies indirectly affects the performances of the classification models of GMM and BPNN based on the distribution of the layer data, and GMM is significantly better than BPNN in classifying thick lithofacies, while BPNN is significantly better than GMM in classifying thin lithofacies. Based on the above understanding, a GMM-BPNN is established for both thick and thin lithofacies classification. Compared with the single GMM and BPNN model, the identification performance of the hybrid model is greatly improved. Besides, the model has also been applied to lithofacies

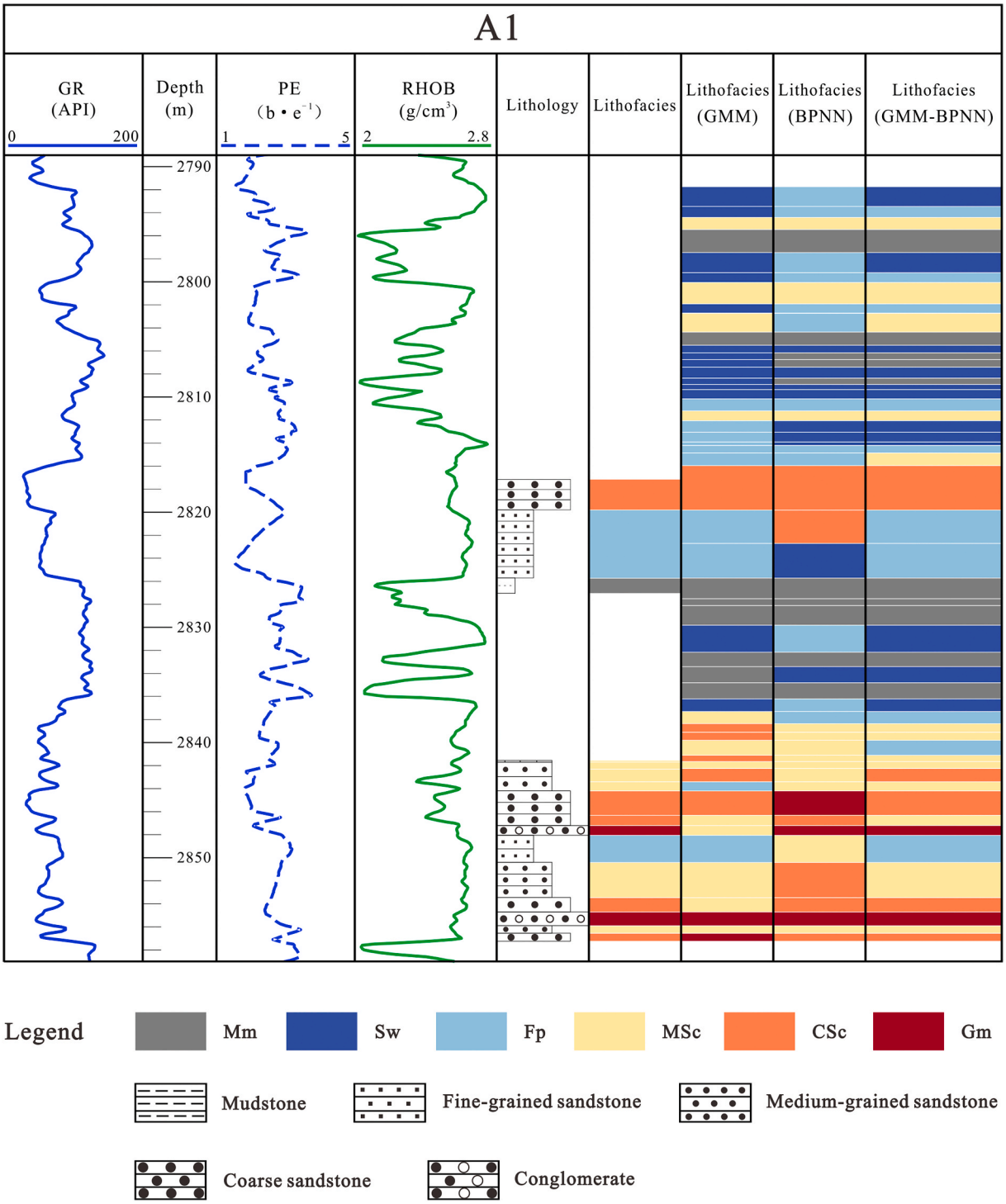

Fig. 19. GMM, BPNN, and GMM-BPNN single-well lithofacies identification results for the core section of the destination strata in well A1 of the three non-participating training wells.

Table 14
Evaluation metrics for GMM-BPNN in one other area.

Evaluation metrics	Train set	Validation set	Blind test set
Precision	0.88	0.84	0.82
Recall	0.87	0.81	0.80
F1	0.87	0.82	0.80

identification in the western part of the Sulige gas field, and good results have been achieved. This proves that the model applies to both lithofacies prediction work in the study area with a large amount of labeled data and a small amount of labeled data, and has good generalizability. It provides a valuable modeling approach for identifying geological units using machine learning.

In summary, when studying the use of machine learning to identify sedimentary and hydrocarbon geological units, it is important to balance the focus on algorithm sophistication with the characterization of the geological targets themselves and their integration with machine learning models. Last but not least, there are differences in lithofacies characteristics and well logging curve resolution in different study areas. In further studies, using further integration improving the sophistication of the algorithms and training the models with a large number of real datasets from different geological backgrounds and logging instruments will help to obtain more accurate and plausible conclusions as well as more generalized classification models.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, we used ChatGPT only to improve the language and readability. The content of the manuscript has been authored by a human being and the data appearing in the manuscript are real. After using this tool, we have reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

All authors disclosed no relevant relationships.

Data availability

Data will be made available on request.

Acknowledgment

This research is supported by the National Natural Science Foundation of China (No. 41902125) and the CNPC Western Drilling Engineering Co., Ltd. We express our gratitude to the staff of CNPC Western Drilling Engineering Co., Ltd., for their valuable assistance in collecting core samples and other data. We are grateful to all reviewers for their thorough and constructive reviews, which greatly helped to improve our manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoen.2023.212587.

References

Aigbadon, G.O., Christopher, S.D., Akudo, E.O., Akakuru, O.C., 2022. Sedimentary facies and textural characteristics of Cretaceous sandstones in the southern Bida Basin, Nigeria: implication for reservoir potential and depositional environment. Energy Geoscience 3 (3), 323–341. https://doi.org/10.1016/j.engeos.2022.05.002.

Al-Mudhafar, W.J., 2020. Integrating machine learning and data analytics for geostatistical characterization of clastic reservoirs. J. Petrol. Sci. Eng. 195 https://doi.org/10.1016/j.petrol.2020.107837.

Alassaf, M., Qamar, A.M., 2022. Improving sentiment analysis of Arabic tweets by one-way ANOVA. Journal of King Saud University - Computer and Information Sciences 34 (6), 2849–2859. https://doi.org/10.1016/j.jksuci.2020.10.023.

Allen, D.R., 1975. Chapter 7 identification of sediments - their depositional environment and degree of compaction—from well logs. In: Compaction of Coarse-Grained Sediments, pp. 349–401. https://doi.org/10.1016/s0070-4571(08)71089-6.

Allen, J.R.L., 1983. Studies in fluviatile sedimentation: bars, bar-complexes and sandstone sheets (low-sinuosity braided streams) in the brownstones (L. devonian), Welsh borders. Sediment. Geol. 33 (4), 237–293. https://doi.org/10.1016/0037-0738(83)90076-3.

Asquith, G.B., Krygowski, D., 2004. Basic Well Log Analysis, vol. 16. https://doi.org/10.1306/Mth16823.

Avseth, P., Mukerji, T., 2002. Seismic lithofacies classification from well logs using statistical rock physics. Petrophysics 43, 70–81.

Banerjee, P., Chattopadhyay, T., Chattopadhyay, A.K., 2023. Comparison among different clustering and classification techniques: astronomical data-dependent study. N. Astron. 100 https://doi.org/10.1016/j.newast.2022.101973.

Bestagini, P., Lipari, V., Tubaro, S., 2017. A machine learning approach to facies classification using well logs. In: SEG Technical Program Expanded Abstracts 2017.

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. https://doi.org/10.1016/j.jngse.2016.04.055.

Blundell, R., Bond, S., 2000. GMM Estimation with persistent panel data: an application to production functions. Econom. Rev. 19 (3), 321–340. https://doi.org/10.1080/07474930008800475.

Bressan, T.S., Kehl de Souza, M., Girelli, T.J., Junior, F.C., 2020. Evaluation of machine learning methods for lithology classification using geophysical data. Comput. Geosci. 139 https://doi.org/10.1016/j.cageo.2020.104475.

Cavalcanti, G.D.C., Soares, R.J.O., 2020. Ranking-based instance selection for pattern classification. Expert Syst. Appl. 150. https://doi.org/10.1016/j.eswa.2020.113269.

Chang, H.-c., Kopaska-Merkel, D.C., Chen, H.-C., Durrans, S.R., 2000. Lithofacies identification using multiple adaptive resonance theory neural networks and group decision expert system. Comput. Geosci. 26 (5), 591–601. https://doi.org/10.1016/S0098-3004(00)00010-8.

Chen, S., Liu, P., Tang, D., Tao, S., Zhang, T., 2021. Identification of thin-layer coal texture using geophysical logging data: investigation by Wavelet Transform and Linear Discrimination Analysis. Int. J. Coal Geol. 239 https://doi.org/10.1016/j.coal.2021.103727.

Chen, W.-H., Carrera Uribe, M., Kwon, E.E., Lin, K.-Y.A., Park, Y.-K., Ding, L., Saw, L.H., 2022. A comprehensive review of thermoelectric generation optimization by statistical approach: taguchi method, analysis of variance (ANOVA), and response surface methodology (RSM). Renew. Sustain. Energy Rev. 169 https://doi.org/10.1016/j.rser.2022.112917.

Chen, X., Xi, C., Cao, J., 2015. Research on moving object detection based on improved mixture Gaussian model. Optik 126 (20), 2256–2259. https://doi.org/10.1016/j.ijleo.2015.05.122.

Chen, Z., Zhang, G., He, R., Tian, Z., Fu, C., Jin, X., 2023. Acoustic emission analysis of crack type identification of corroded concrete columns under eccentric loading: a comparative analysis of RA-AF method and Gaussian mixture model. Case Stud. Constr. Mater. 18 https://doi.org/10.1016/j.cscm.2023.e02021.

Cherana, A., Aliouane, L., Doghmane, M.Z., Ouadfeul, S.-A., Nabawy, B.S., 2022. Lithofacies discrimination of the Ordovician unconventional gas-bearing tight sandstone reservoirs using a subtractive fuzzy clustering algorithm applied on the well log data: illizi Basin, the Algerian Sahara. J. Afr. Earth Sci. 196 https://doi.org/10.1016/j.jafrearsci.2022.104732.

Cohen, J., 1977. Chapter 8 - F tests on means in the analysis of variance and covariance. In: Cohen, J. (Ed.), Statistical Power Analysis for the Behavioral Sciences. Academic Press, pp. 273–406. https://doi.org/10.1016/B978-0-12-179060-8.50013-X.

Colombera, L., Mountney, N.P., 2019. The lithofacies organization of fluvial channel deposits: a meta-analysis of modern rivers. Sediment. Geol. 383, 16–40. https://doi.org/10.1016/j.sedgeo.2019.01.011.

de Amorim, L.B.V., Cavalcanti, G.D.C., Cruz, R.M.O., 2023. The choice of scaling technique matters for classification performance. Appl. Soft Comput. 133 https://doi.org/10.1016/j.asoc.2022.109924.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B 39 (1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Dhanalakshmi, P., Palanivel, S., Ramalingam, V., 2011. Classification of audio signals using AANN and GMM. Appl. Soft Comput. 11 (1), 716–723. https://doi.org/10.1016/j.asoc.2009.12.033.

El-Gendy, N.H., Radwan, A.E., Waziry, M.A., Dodd, T.J.H., Kh Barakat, M., 2022. An integrated sedimentological, rock typing, image logs, and artificial neural networks analysis for reservoir quality assessment of the heterogeneous fluvial-deltaic Messinian Abu Madi reservoirs, Salma field, onshore East Nile Delta, Egypt. Mar. Petrol. Geol. 145 https://doi.org/10.1016/j.marpetgeo.2022.105910.

Fu, Y., Luo, J., Shi, X., Cao, J., Mao, Q., Sheng, W., 2022. Implications of lithofacies and diagenetic evolution for reservoir quality: a case study of the Upper Triassic chang 6 tight sandstone, southeastern Ordos Basin, China. J. Petrol. Sci. Eng. 218 https://doi.org/10.1016/j.petrol.2022.111051.

Gm, H., Gourisaria, M.K., Pandey, M., Rautaray, S.S., 2020. A comprehensive survey and analysis of generative models in machine learning. Computer Science Review 38. https://doi.org/10.1016/j.cosrev.2020.100285.

González-Prieto, Á., Mozo, A., Gómez-Canaval, S., Talavera, E., 2022. Improving the quality of generative models through Smirnov transformation. Inf. Sci. 609, 1539–1566. https://doi.org/10.1016/j.ins.2022.07.066.

Guo, Z., Sun, L., Jia, A., Lu, T., 2015. 3-D geological modeling for tight sand gas reservoir of braided river facies. Petrol. Explor. Dev. 42 (1), 83–91. https://doi.org/10.1016/s1876-3804(15)60009-x.

Gyori, N.G., Palombo, M., Clark, C.A., Zhang, H., Alexander, D.C., 2022. Training data distribution significantly impacts the estimation of tissue microstructure with machine learning. Magn. Reson. Med. 87 (2), 932–947. https://doi.org/10.1002/mrm.29014.

Harris, J.R., Grunsky, E.C., 2015. Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data. Comput. Geosci. 80, 9–25. https://doi.org/10.1016/j.cageo.2015.03.013.

Hassan, S., Darwish, M., Tahoun, S.S., Radwan, A.E., 2022. An integrated high-resolution image log, sequence stratigraphy and palynofacies analysis to reconstruct the Albian – cenomanian basin depositional setting and cyclicity: insights from the southern Tethys. Mar. Petrol. Geol. 137 https://doi.org/10.1016/j.marpetgeo.2021.105502.

Hong, W.-K., 2023. 4 - forward and backpropagation for artificial neural networks. In: Hong, W.-K. (Ed.), Artificial Intelligence-Based Design of Reinforced Concrete Structures. Woodhead Publishing, pp. 67–116. https://doi.org/10.1016/B978-0-443-15252-8.00006-6.

Hush, D.R., Horne, B.G., 1993. Progress in supervised neural networks. IEEE Signal Process. Mag. 10 (1), 8–39. https://doi.org/10.1109/79.180705.

Imamverdiyev, Y., Sukhostat, L., 2019. Lithological facies classification using deep convolutional neural network. J. Petrol. Sci. Eng. 174, 216–228. https://doi.org/10.1016/j.petrol.2018.11.023.

Iraji, S., Soltanmohammadi, R., Matheus, G.F., Basso, M., Vidal, A.C., 2023a. Application of unsupervised learning and deep learning for rock type prediction and petrophysical characterization using multi-scale data. Geoenergy Science and Engineering 230. https://doi.org/10.1016/j.geoen.2023.212241.

Iraji, S., Soltanmohammadi, R., Munoz, E.R., Basso, M., Vidal, A.C., 2023b. Core scale investigation of fluid flow in the heterogeneous porous media based on X-ray computed tomography images: upscaling and history matching approaches. Geoenergy Science and Engineering 225. https://doi.org/10.1016/j.geoen.2023.211716.

Jiao, L., Denœux, T., Liu, Z.-g., Pan, Q., 2022. EGMM: an evidential version of the Gaussian mixture model for clustering. Appl. Soft Comput. 129 https://doi.org/10.1016/j.asoc.2022.109619.

Jordan, M., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. Science (New York, N.Y.) 349, 255–260. https://doi.org/10.1126/science.aaa8415.

Kiasari, M.A., Jang, G.-J., Lee, M., 2017. Novel iterative approach using generative and discriminative models for classification with missing features. Neurocomputing 225, 23–30. https://doi.org/10.1016/j.neucom.2016.11.015.

Kim, H.-K., Kim, K.-H., Yun, S.-T., Oh, J., Kim, H.-R., Park, S.-H., Kim, M.-S., Kim, T.-S., 2019. Probabilistic assessment of potential leachate leakage from livestock mortality burial pits: a supervised classification approach using a Gaussian mixture model (GMM) fitted to a groundwater quality monitoring dataset. Process Saf. Environ. Protect. 129, 326–338. https://doi.org/10.1016/j.psep.2019.07.015.

Lai, J., Wang, G., Wang, S., Cao, J., Li, M., Pang, X., Han, C., Fan, X., Yang, L., He, Z., Qin, Z., 2018. A review on the applications of image logs in structural analysis and sedimentary characterization. Mar. Petrol. 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. https://doi.org/10.1016/j.fuel.2021.121145.

Langer, H., Falsaperla, S., Hammer, C., 2020. Supervised learning. In: Advantages and Pitfalls of Pattern Recognition, pp. 33–85. https://doi.org/10.1016/b978-0-12-811842-9.00002-9.

Li, Y., Anderson-Sprecher, R., 2006. Facies identification from well logs: a comparison of discriminant analysis and naïve Bayes classifier. J. Petrol. Sci. Eng. 53 (3), 149–157. https://doi.org/10.1016/j.petrol.2006.06.001.

Li, Z., Li, P., Liu, Z., Cui, Y., 2022. Single-well lithofacies identification based on logging response and convolutional neural network. J. Appl. Geophys. 207 https://doi.org/10.1016/j.jappgeo.2022.104865.

Lindberg, D.V., Rimstad, E., Omre, H., 2015. Inversion of well logs into facies accounting for spatial dependencies and convolution effects. J. Petrol. Sci. Eng. 134, 237–246. https://doi.org/10.1016/j.petrol.2015.09.027.

Liu, J., Liu, K., Huang, X., 2016. Effect of sedimentary heterogeneities on hydrocarbon accumulations in the Permian Shanxi Formation, Ordos Basin, China: insight from an integrated stratigraphic forward and petroleum system modelling. Mar. Petrol. Geol. 76, 412–431. https://doi.org/10.1016/j.marpetgeo.2016.05.028.

Liu, J., Liu, Z., Xiao, K., Huang, Y., Jin, W., 2020. Characterization of favorable lithofacies in tight sandstone reservoirs and its significance for gas exploration and exploitation: a case study of the 2nd Member of Triassic Xujiahe Formation in the Xinchang area, Sichuan Basin. Petrol. Explor. Dev. 47 (6), 1194–1205. https://doi.org/10.1016/s1876-3804(20)60129-5.

Loog, M., 2018. Supervised classification: quite a brief overview. In: Machine Learning Techniques for Space Weather, pp. 113–145. https://doi.org/10.1016/b978-0-12-811788-0.00005-6.

Lu, G., Zeng, L., Dong, S., Huang, L., Liu, G., Ostadhassan, M., He, W., Du, X., Bao, C., 2023. Lithology identification using graph neural network in continental shale oil reservoirs: a case study in Mahu Sag, Junggar Basin, Western China. Mar. Petrol. Geol. 150 https://doi.org/10.1016/j.marpetgeo.2023.106168.

McDowell, G., 1999. In-site Nickel Assay by Prompt Gamma Neutron Activation Wireline Logging, vol. 17. Seg Technical Program Expanded Abstracts. https://doi.org/10.1190/1.1820589.

Melnykov, V., Maitra, R., 2010. Finite mixture models and model-based clustering. Stat. Surv. 4 (none) https://doi.org/10.1214/09-ss053.

Miall, A.D., 1985. Architectural-element analysis: a new method of facies analysis applied to fluvial deposits. Earth Sci. Rev. 22 (4), 261–308. https://doi.org/10.1016/0012-8252(85)90001-7.

Naim, I., Gildea, D., 2012. Convergence of the EM algorithm for Gaussian mixtures with unbalanced mixing coefficients. In: Proceedings of the 29th International Conference on Machine Learning, ICML 2012, p. 2.

Nazeer, A., Ahmed, S., Solangi, S., 2016. Sedimentary facies interpretation of gamma ray (GR) log as basic well logs in central and lower indus basin of Pakistan. Geodesy and Geodynamics 7. https://doi.org/10.1016/j.geog.2016.06.006.

Omer Fadl Elssied, N., Ibrahim, O., Hamza Osman, A., 2014. A novel feature selection based on one-way ANOVA F-test for E-mail spam classification. Res. J. Appl. Sci. Eng. Technol. 7 (3), 625–638. https://doi.org/10.19026/rjaset.7.299.

Otchere, D.A., Arbi Ganat, T.O., Gholami, R., Ridha, S., 2021. Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: comparative analysis of ANN and SVM models. J. Petrol. Sci. Eng. 200 https://doi.org/10.1016/j.petrol.2020.108182.

Paiva, P.Y.A., Moreno, C.C., Smith-Miles, K., Valeriano, M.G., Lorena, A.C., 2022. Relating instance hardness to classification performance in a dataset: a visual approach. Mach. Learn. 111 (8), 3085–3123. https://doi.org/10.1007/s10994-022-06205-9.

Pearson, K., Henrici, O.M.F.E., 1997. III. Contributions to the mathematical theory of evolution. Proc. Roy. Soc. Lond. 54 (326–330), 329–333. https://doi.org/10.1098/rspl.1893.0079.

Radwan, A.E., 2020. Modeling the depositional environment of the sandstone reservoir in the middle miocene sidri member, badri field, gulf of suez basin, Egypt: integration of gamma-ray log patterns and petrographic characteristics of lithology. Nat. Resour. Res. 30 (1), 431–449. https://doi.org/10.1007/s11053-020-09757-6.

Rafik, B., Kamel, B., 2017. Prediction of permeability and porosity from well log data using the nonparametric regression with multivariate analysis and neural network, Hassi R'Mel Field, Algeria. Egyptian Journal of Petroleum 26 (3), 763–778. https://doi.org/10.1016/j.ejpe.2016.10.013.

Ramírez, J.M., Díez, F., Rojo, P., Mancuso, V., Fernández-Anta, A., 2023. Explainable machine learning for performance anomaly detection and classification in mobile networks. Comput. Commun. 200, 113–131. https://doi.org/10.1016/j.comcom.2023.01.003.

Rayens, W., 2012. Discriminant analysis and statistical pattern recognition. Technometrics 35, 324–326. https://doi.org/10.1080/00401706.1993.10485331.

Ren, X., Hou, J., Song, S., Liu, Y., Chen, D., Wang, X., Dou, L., 2019. Lithology identification using well logs: a method by integrating artificial neural networks and sedimentary patterns. J. Petrol. Sci. Eng. 182 https://doi.org/10.1016/j.petrol.2019.106336.

Rider, M., 1990. Gamma-ray log shape used as a facies indicator: critical analysis of an oversimplified methodology. Geological Society, London, Special Publications 48, 27–37. https://doi.org/10.1144/GSL.SP.1990.048.01.04.

Rogers, S.J., Fang, J.H., Karr, C.L., Stanley, D.A., 1992. Determination of lithology from well logs using a neural Network1. AAPG (Am. Assoc. Pet. Geol.) Bull. 76 (5), 731–739. https://doi.org/10.1306/BDFF88BC-1718-11D7-8645000102C1865D.

Rubinstein, Y.D., Hastie, T., 1997. Discriminative vs informative learning. In: Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, Newport Beach, CA.

Rudstam, G., Elofsson, U.O.E., Söndergaard, H.P., Bonde, L.O., Beck, B.D., 2022. Trauma-focused group music and imagery with women suffering from PTSD/Complex PTSD: a randomized controlled study. European Journal of Trauma & Dissociation 6 (3). https://doi.org/10.1016/j.ejtd.2022.100277.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536. https://doi.org/10.1038/323533a0.

Sancho, J.-L., Pierson, W.E., Ulug, B., Figueiras-Vidal, A.b.R., Ahalt, S.C., 2000. Class separability estimation and incremental learning using boundary methods. Neurocomputing 35 (1), 3–26. https://doi.org/10.1016/S0925-2312(00)00293-9.

Shehata, A.A., Osman, O.A., Nabawy, B.S., 2021. Neural network application to petrophysical and lithofacies analysis based on multi-scale data: an integrated study using conventional well log, core and borehole image data. J. Nat. Gas Sci. Eng. 93 https://doi.org/10.1016/j.jngse.2021.104015.

Shen, C., Asante-Okyere, S., Yevenyo Ziggah, Y., Wang, L., Zhu, X., 2019. Group method of data handling (GMDH) lithology identification based on wavelet analysis and dimensionality reduction as well log data pre-processing techniques. Energies 12 (8). https://doi.org/10.3390/en12081509.

Shen, Y., Liu, F., 2019. An Approach for Semantic Web Discovery Using Unsupervised Learning Algorithms, pp. 56–72. https://doi.org/10.1007/978-981-15-1922-2_4.

Soltanmohammadi, R., Iraji, S., Rodrigues de Almeida, T., Basso, M., Ruidiaz Munoz, E., Campane Vidal, A., 2023. Investigation of pore geometry influence on fluid flow in heterogeneous porous media: a pore-scale study. Energy Geoscience 5 (1). https://doi.org/10.1016/j.engeos.2023.100222.

Sun, J., Li, Q., Chen, M., Ren, L., Huang, G., Li, C., Zhang, Z., 2019. Optimization of models for a rapid identification of lithology while drilling - a win-win strategy based on machine learning. J. Petrol. Sci. Eng. 176, 321–341. https://doi.org/10.1016/j.petrol.2019.01.006.

Tan, X., Huang, Y., Lei, T., Wang, J., Cao, T., Zhang, Z., Hao, T., Gao, Z., Luo, L., Zhu, C., Mo, S., 2023. Sedimentary characteristics of sandy braided river deposits and factors controlling their deposition: a case study of the lower Shihezi Formation in the northern ordos basin, China. Geoenergy Science and Engineering. https://doi.org/10.1016/j.geoen.2023.211932.

Theodoridis, S., 2015. Classification. In: Machine Learning, pp. 275–325. https://doi.org/10.1016/b978-0-12-801522-3.00007-0.

Tian, M., Omre, H., Xu, H., 2021. Inversion of well logs into lithology classes accounting for spatial dependencies by using hidden markov models and recurrent neural networks. J. Petrol. Sci. Eng. 196 https://doi.org/10.1016/j.petrol.2020.107598.

Wang, M., Tang, H., Zhao, F., Liu, S., Yang, Y., Zhang, L., Liao, J., Lu, H., 2017. Controlling factor analysis and prediction of the quality of tight sandstone reservoirs: a case study of the He8 Member in the eastern Sulige Gas Field, Ordos Basin, China. J. Nat. Gas Sci. Eng. 46, 680–698. https://doi.org/10.1016/j.jngse.2017.08.033.

Wang, Z., Xie, K., Wen, C., Sheng, G., He, J., Tian, H., 2023. Multi-scale spatiotemporal feature lithology identification method based on split-frequency weighted reconstruction. Geoenergy Science and Engineering 226. https://doi.org/10.1016/j.geoen.2023.211794.

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. https://doi.org/10.1016/j.marpetgeo.2019.07.026.

Wu, D., Yang, Z., Liang, L., 2006. Using DEA-neural network approach to evaluate branch efficiency of a large Canadian bank. Expert Syst. Appl. 31 (1), 108–115. https://doi.org/10.1016/j.eswa.2005.09.034.

Xu, Z., Zhang, B., Li, F., Cao, G., Liu, Y., 2018. Well-log decomposition using variational mode decomposition in assisting the sequence stratigraphy analysis of a conglomerate reservoir. Geophysics 83 (4), B221–B228. https://doi.org/10.1190/geo2017-0817.1.

Xue, J.-H., Titterington, D.M., 2009. Interpretation of hybrid generative/discriminative algorithms. Neurocomputing 72 (7–9), 1648–1655. https://doi.org/10.1016/j.neucom.2008.08.009.

Xue, J.-H., Titterington, D.M., 2010. On the generative–discriminative tradeoff approach: interpretation, asymptotic efficiency and classification performance. Comput. Stat. Data Anal. 54 (2), 438–451. https://doi.org/10.1016/j.csda.2009.09.011.

Yang, H., Fu, J., Wei, X., Liu, X., 2008. Sulige field in the Ordos Basin: geological setting, field discovery and tight gas reservoirs. Mar. Petrol. Geol. 25 (4–5), 387–400. https://doi.org/10.1016/j.marpetgeo.2008.01.007.

Yang, Y., Zhu, J., Zhao, C., Liu, S., Tong, X., 2011. The spatial continuity study of NDVI based on Kriging and BPNN algorithm. Math. Comput. Model. 54 (3–4), 1138–1144. https://doi.org/10.1016/j.mcm.2010.11.046.

Yue, D., Jiagen, H., Yuming, L., Ye, W., Jing, Z., Yanqing, S., Jingyun, Z., 2015. 15-17 Aug. 2015). A back propagation artificial neural network application in lithofacies identification. In: 2015 11th International Conference on Natural Computation (ICNC).

Zengzhao, F., Hongping, B., Jinhua, J., Yigang, W., Xiuqin, D., Yuan, W., Min, L., 2013. Lithofacies palaeogeography as a guide to petroleum exploration. J. Palaeogeogr. 2 (2), 109–126. https://doi.org/10.3724/SP.J.1261.2013.00021.

Zhang, X., Song, C., Zhao, J., Xia, D., 2023. Gaussian mixture continuously adaptive regression for multimode processes soft sensing under time-varying virtual drift. J. Process Control 124, 1–13. https://doi.org/10.1016/j.jprocont.2023.02.003.

Zhao, D., Hou, J., Sarma, H., Guo, W., Liu, Y., Xie, P., Dou, L., Chen, R., Zhang, Z., 2023. Pore throat heterogeneity of different lithofacies and diagenetic effects in gravelly braided river deposits: implications for understanding the formation process of high-quality reservoirs. Geoenergy Science and Engineering 221. https://doi.org/10.1016/j.petrol.2022.111309.

Zhou, M., Wei, P., Deng, L., 2022. Research on the factorial effect of science and technology innovation (STI) policy mix using multifactor analysis of variance (ANOVA). Journal of Innovation & Knowledge 7 (4). https://doi.org/10.1016/j.jik.2022.100249.

Zhou, Z., Wang, G., Ran, Y., Lai, J., Cui, Y., Zhao, X., 2016. A logging identification method of tight oil reservoir lithology and lithofacies: a case from Chang7 Member of Triassic Yanchang Formation in Heshui area, Ordos Basin, NW China. Petrol. Explor. Dev. 43 (1), 65–73. https://doi.org/10.1016/s1876-3804(16)30007-6.