Petroleum Science 23 (2026) 3854–3874

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

# A DBOKS-based hybrid machine learning method for carbonate rock classification

Wen-Chuang Li [a], Zhong-Xiang Zhao [a,b,c,*], You-Bin He [a,b,c], Lian-Hua Wu [a], Yu-Qing Zhang [a]

[a] School of Geosciences, Yangtze University, Wuhan, 430100, Hubei, China
[b] Hubei Engineering Research Center of Unconventional Petroleum Geology and Engineering, Wuhan, 430100, Hubei, China
[c] Hubei Key Laboratory of Complex Shale Oil and Gas Geology and Development in Southern China, Wuhan, 430100, Hubei, China

A R T I C L E I N F O

*Article history:*
Received 5 November 2025
Received in revised form
27 April 2026
Accepted 27 April 2026
Available online 29 April 2026

Edited by Xiu-Fang Hu

*Keywords:*
Lithology classification
Carbonate rocks
DBOKS algorithm
HMM framework
Machine learning

A B S T R A C T

Addressing the challenges in carbonate rock lithology identification, such as similar well-logging curve response characteristics, scarcity of transitional lithology samples leading to classification difficulties, and the inability of traditional models to capture dependencies in vertical lithological sequences, we propose a hybrid machine learning framework that integrates feature enhancement, sample balancing, and sequence correction. The method first employs the Dung Beetle Optimizer (DBO) to optimize *K*-means clustering, thereby enhancing the feature discriminability for similar logging responses. Subsequently, the SMOTE oversampling technique is applied to specifically address the issue of sparse transitional lithology samples while preserving the original data distribution. On this basis, the Hidden Markov Model (HMM), is introduced to vertically correct the preliminary identification results using prior knowledge of stratigraphic sequences, effectively modeling the inter-lithology dependencies. Experimental results show that tree-based ensemble models driven by this framework significantly outperform traditional methods, with all evaluation metrics exceeding 95%. This study demonstrates that by jointly addressing the three major challenges of feature separability, sample balance, and sequence continuity, the proposed framework can significantly enhance the accuracy and geological consistency of carbonate rock lithology identification, providing a reliable solution for intelligent well logging interpretation in complex reservoirs.


## 1. Introduction

Carbonate reservoirs, as one of the most important sources of global oil and gas resources, have long been a key focus of exploration. However, accurate lithology identification within these reservoirs remains a major challenge in both petroleum geology and well log interpretation (Katz and Everett, 2016; Li et al., 2018; Zhu et al., 2019; Ma et al., 2020; Mahdaviara et al., 2020; Shi et al., 2024). In structurally complex and sedimentologically heterogeneous regions, reliable carbonate lithology identification is crucial for accurate reservoir evaluation and directly influences exploration deployment and development decision-making (Lu et al., 2017; Tian et al., 2019; Sen et al., 2021). Therefore, precise lithology classification is a key yet difficult task in the evaluation and effective development of carbonate hydrocarbon reservoirs.

Traditionally, core analysis and geophysical well-logging data have been widely used to identify lithology types (Anifowose et al., 2015; Bai et al., 2016; Abbey et al., 2018; Chang et al., 2000). Classical lithology identification methods—such as thin-section analysis, core observation, and petrophysical measurements—largely rely on expert experience, resulting in low efficiency and strong subjectivity (Chang et al., 2000; Salehi and Honarvar, 2014; Zhang et al., 2017; Teillet et al., 2021). Moreover, due to the complex response characteristics of carbonate formations, these methods often fail to achieve satisfactory accuracy. Although core description remains the most direct and effective lithology identification method, it is limited by low coring rates and high costs. Imaging logging provides high vertical resolution and continuous geological information but is also costly and not always available for all wells (Tian et al., 2019; Lan et al., 2021; Zheng et al., 2021). In contrast, conventional well logs, which are

* Corresponding author.
*E-mail addresses:* Liwenchuang0914@163.com (W.-C. Li), zhaozhongxiang5@163.com (Z.-X. Zhao).
Peer review under the responsibility of China University of Petroleum (Beijing).

https://doi.org/10.1016/j.petsci.2026.04.043
1995-8226/

easy to acquire and cost-effective, are widely used for lithology identification. These logging curves contain abundant geological and lithological information about subsurface reservoirs, making them a practical and efficient means of lithology classification (Ozkan et al., 2011; Zhu et al., 2023; Liu, 2025). For instance, cross-plot techniques combine multiple log curves to differentiate lithologies. However, because carbonate lithologies often exhibit highly similar log responses, traditional cross-plot methods usually suffer from limited separability and low classification accuracy (Li et al., 2025). Consequently, lithology identification based on logging data is not a simple linear classification problem.

In recent years, the application of artificial intelligence technology (AI) in the field of petroleum geology has become increasingly widespread (Ghosh et al., 2016; Zhang et al., 2021; Liu et al., 2022; Shi et al., 2024). In carbonate lithology prediction, data-driven AI techniques play an essential role (Zhao et al., 2019; Mondal and Singh, 2022; Zhou et al., 2023). By analyzing large-scale logging datasets containing lithological information, machine learning (ML) methods can effectively predict lithologies in unexplored intervals. Furthermore, AI enables the integration and optimization of multiple prediction models, substantially improving both accuracy and computational efficiency (Delavar, 2022; Yu et al., 2022). However, an increasing number of studies have focused on constructing complex model structures, neglecting the inherent challenges of geology itself. During the sedimentary process, the number of samples for different lithologies often shows significant discrepancies due to factors such as sea level fluctuations and tectonic movements (He et al., 2020). This class imbalance issue can lead to a model's overall accuracy being biased towards the majority class, making it difficult to effectively learn the characteristics of minority lithologies. As a result, the ability to identify critical lithologies, such as thin layers and transition zones, significantly decreases, sometimes even leading to severe misclassifications (Luque et al., 2019; Thabtah et al., 2020). Conventional SMOTE methods have limited capability in recognizing class boundaries during the sample generation process. To achieve more accurate well log response boundary identification, this study combines SMOTE with the *K*-means clustering method. However, when selecting an unsupervised clustering method sensitive to parameters, one must face challenges such as local optimal solutions and the interaction of complex parameters (Yang et al., 2023; Bhandari et al., 2024; Kazemi et al., 2025). Therefore, we introduce a Dung Beetle Optimization (DBO) algorithm, merging its strengths with *K*-means clustering to overcome the limitations of traditional algorithms.

In addition, regarding classifier performance, traditional ML models have garnered significant attention due to their good interpretability and intuitive decision-making process. They are particularly well-suited for scenarios with limited data or when features have been carefully selected, as they tend to exhibit strong generalization ability under such conditions. Although various deep learning (DL) approaches have been successfully applied to carbonate lithology prediction, conventional ensemble learning models such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost) remain widely used due to their simplicity, interpretability, and robustness (Zhang et al., 2019; Liu et al., 2021; Shen et al., 2024; Zeng et al., 2024). However, these models typically treat each depth point as an independent sample and rely solely on static mappings between logging responses and lithology labels. This "pointwise classification" strategy neglects the inherent vertical continuity and depositional evolution of strata, often producing geologically inconsistent or isolated lithology transitions (Verma et al., 2014; Zaremba et al., 2014; Wang et al., 2022; Aftab et al., 2023; Carrasquilla, 2023; Dong et al., 2023; Song et al., 2023).

To address these challenges, this study introduces several methodological innovations for carbonate lithology identification:

(1) Based on the correspondence between carbonate lithology and well log responses, a hybrid method integrating the DBO algorithm, *K*-means clustering, and synthetic minority oversampling technique—collectively termed DBOKS—is proposed to mitigate issues such as the scarcity of transitional lithology samples and blurred logging response boundaries.
(2) A Hidden Markov Model (HMM)-based lithology sequence correction framework is introduced, which treats the entire stratigraphic profile as a dynamic sequence: the hidden lithology states are modeled as a latent sequence, and the observed well logs as the emission sequence. Compared with traditional sequence models, this approach not only alleviates overfitting and noise sensitivity but also produces lithology predictions that better conform to geological continuity and sedimentary logic.

Ultimately, the proposed DBOKS method achieves accurate lithology recognition while effectively addressing sample imbalance issues. The subsequent HMM-based correction further enhances the geological coherence and generalization capability of conventional ML models. Finally, the study summarizes the limitations, potential improvements, and main conclusions of this approach.

## 2. Geological background

The Sichuan Basin, located in southwestern China, is one of the four major onshore gas-rich basins in the country. Since the Mesozoic–Cenozoic, the basin has undergone complex tectonic evolution dominated by thrusting and nappe structures, forming a rhombic structural framework characterized by a west-high to east-low gradient and steeper slopes to the north than to the south (Tang et al., 2014). Based on current structural features, the Sichuan Basin can be subdivided into six tectonic units: Northern Sichuan Low-flat Structure, Western Sichuan Low-steep Structure, Eastern Sichuan High-steep Structure, Central Sichuan Low-flat Structure, Southwestern Sichuan Low-gentle Structure, and Southern Sichuan Low-steep Structure. The study area lies within the Eastern Sichuan High-steep Structure, which is one of the most tectonically active regions in the basin (Xu et al., 2005). The area is characterized by a depositional assemblage of carbonate and evaporite sequences developed in a restricted shallow-platform environment (Wen et al., 2023) (Fig. 1).

The Jialingjiang Formation in eastern Sichuan ranges in thickness from approximately 700 m to 1100 m and gradually thickens from southwest to northeast. Stratigraphically, it is subdivided into five members, namely, the first to the fifth, in an ascending order (Lei et al., 2006). Eight major lithologies are developed in the study area: argillaceous limestone (AL), argillaceous dolomite (AD), limestone (LS), dolomitic limestone (DL), calcareous dolomite (LD), dolomite (DM), gypsiferous dolomite (GD), and gypsum (GM) (Fig. 2). Among these, dolomitic limestone, calcareous dolomite, and gypsiferous dolomite represent typical transitional lithologies. These facies formed during early diagenetic to syn-diagenetic dolomitization or calcitization processes, exhibiting mineralogical similarity, pronounced heterogeneity in texture, and gradual compositional and structural transitions. They commonly occur as thin interbeds with rapid lateral variation and limited single-layer thickness, making core sampling less representative and making lithology identification by logging data particularly challenging.

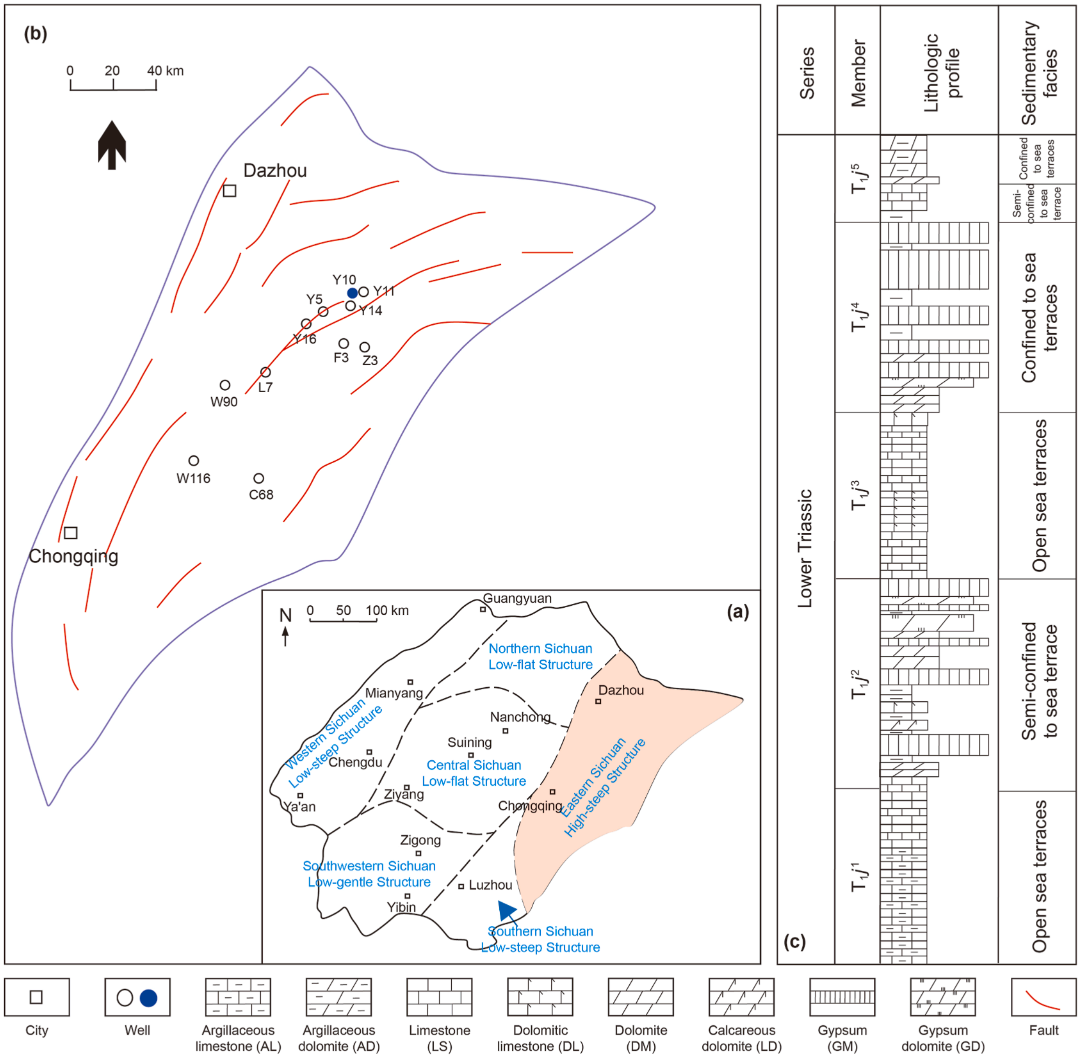


**Fig. 1.** Geological setting of the study area. **(a)** Tectonics of the study area in the Sichuan Basin, **(b)** well location map of the study area, **(c)** comprehensive lithologic-stratigraphic column of the Triassic Jialingjiang Formation (after Wen et al., 2023).

Although transitional lithologies constitute a relatively small proportion of the overall stratigraphic thickness, they frequently develop secondary porosity due to dolomitization-associated dissolution. As a result, they possess favorable petrophysical properties and act as the primary hosts for high-quality reservoirs within the Jialingjiang Formation, controlling local hydrocarbon enrichment (Fig. 3). Therefore, achieving efficient and accurate lithology classification is a key step in reservoir evaluation and plays a crucial role in the hydrocarbon exploration and development of the study area.

## 3. Methods

### 3.1. Workflow framework

All experiments were conducted on a personal workstation running Python 3.11. The computing environment was configured with an Intel® Core™ i9-14900HX processor (2.2 GHz), 16 GB RAM, and an NVIDIA GeForce RTX 4060 GPU. The implementation was primarily carried out using Python, with the machine learning models developed through widely used libraries such as scikit-learn.

A total of 11 wells were selected as the experimental dataset. Among them, 10 wells were used for model training and testing, while one well was reserved as an independent validation set to evaluate model generalization and accuracy (Fig. 1(b)). To systematically analyze the correlation between logging responses and lithology categories, six conventional logging curves were selected, representing three major logging series: (1) lithology log: natural gamma ray (GR); (2) porosity logs: acoustic travel time (AC), density (DEN), and compensated neutron porosity (CNL); and (3) resistivity logs: deep lateral resistivity (RT) and shallow resistivity (RXO).

Prior to lithology prediction, all data were preprocessed following the method proposed by Zheng et al. (2022), which includes three key steps: (1) Depth correction to ensure vertical alignment of multi-log data; (2) Removal of null and anomalous values to eliminate noise; and (3) Normalization of log measurements to ensure numerical stability and comparability across wells. After preprocessing, a total of 3934 valid data points were obtained from the 10 labeled wells. Lithology labels were determined primarily based on core descriptions and thin-section petrographic analyses, supplemented by log curve morphology and cuttings observations to ensure labelling reliability. Subsequently, key logging features were screened and cross-plot analyses were performed (Fig. 4) to visually examine the pairwise relationships among log responses and lithology categories. These

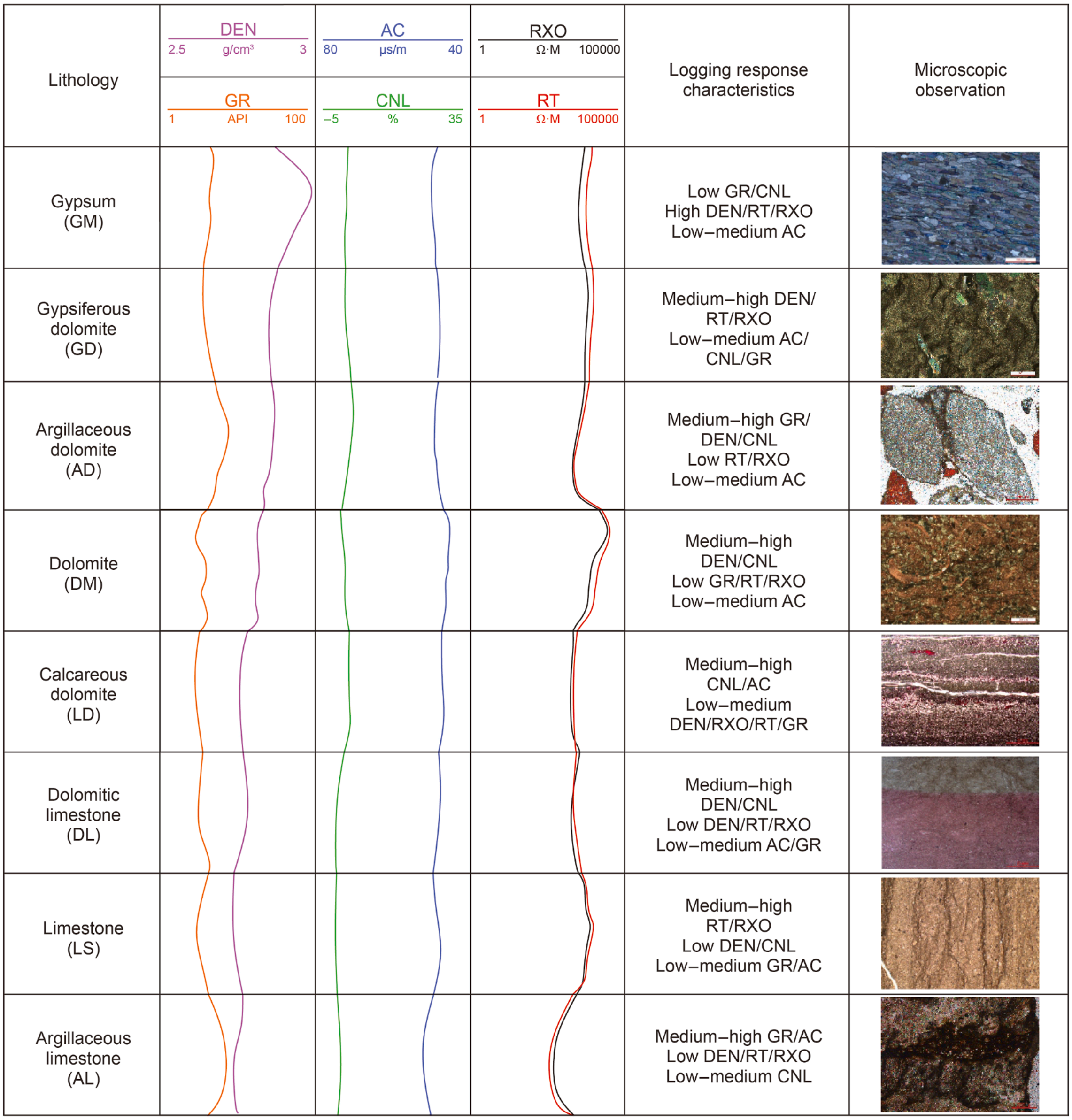


**Fig. 2.** Typical logging response characteristics of the eight lithologies.

cross-plots provided essential geological insight into the separability of carbonate lithologies and served as the foundation for subsequent machine learning model.

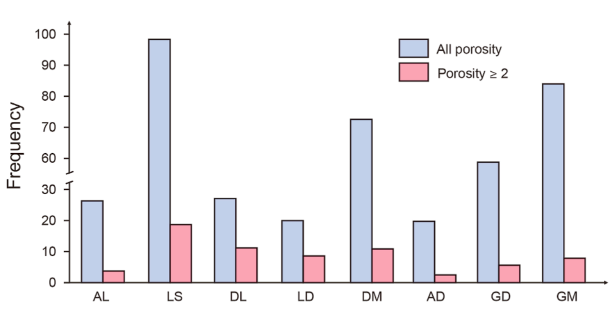


**Fig. 3.** Pore frequency distribution among different lithologies in the study area.

Furthermore, considering the sequential nature of well-logging data, an HMM-based sequence correction was applied following model training. This post-processing step not only preserved the feature enhancement introduced by DBOKS oversampling but also effectively captured vertical sequence dependencies, thus further enhancing the model's generalization capability and robustness. The workflow of this study is shown in Fig. 5.

### 3.2. DBOKS algorithm

DBOKS is an optimized clustering-based oversampling method that integrates the DBO algorithm with the *K*-means–SMOTE framework (Gao et al., 2024; Zeng et al., 2024). Fig. 6 illustrates the implementation process of the enhanced DBO–*K*-means clustering. Clustering analysis aims to partition a dataset into distinct groups that exhibit high intra-cluster similarity and low inter-cluster similarity (Javadi et al., 2017; Zhou et al., 2024). The conventional *K*-means algorithm, known for its simple structure,

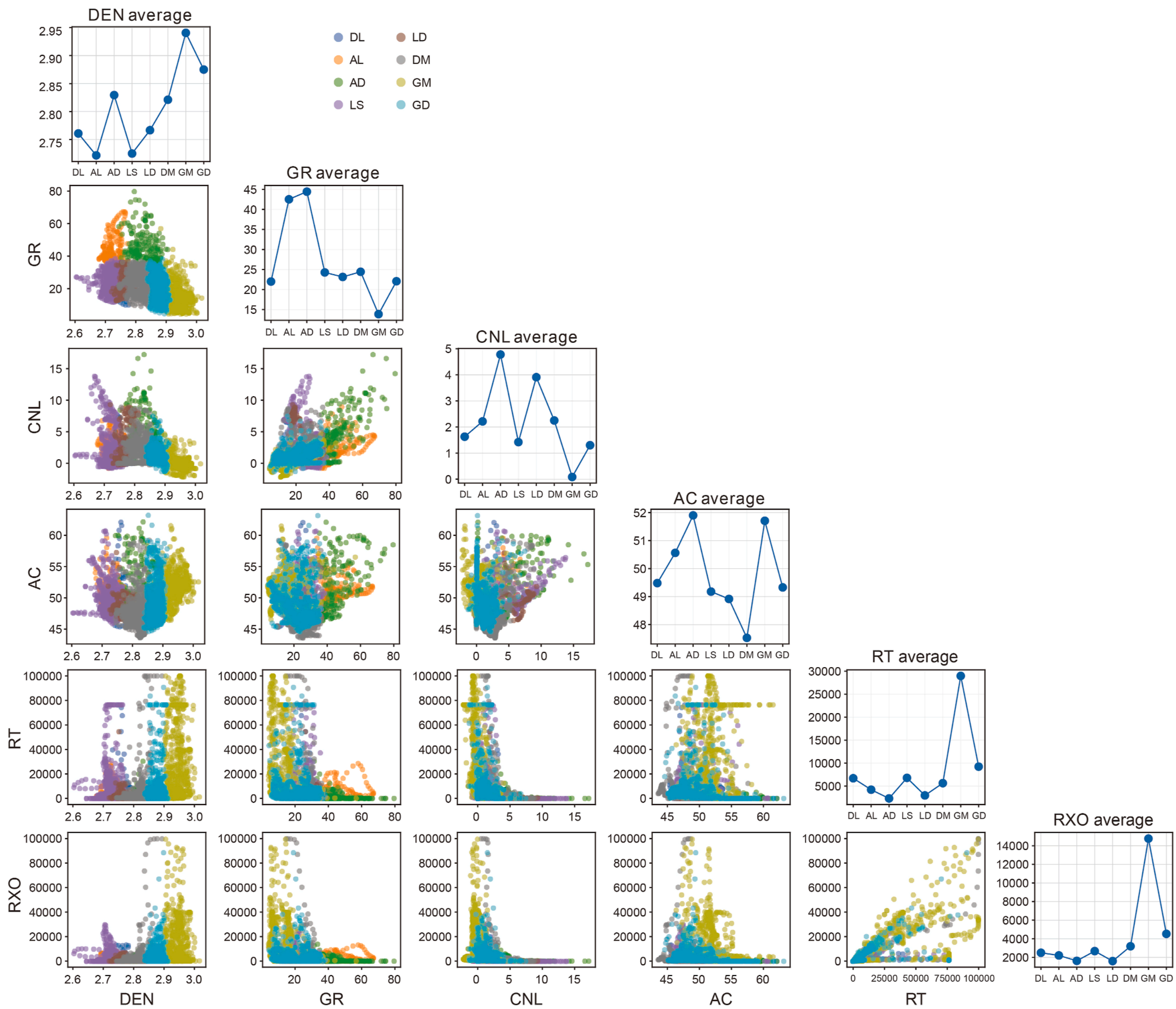


**Fig. 4.** Cross-plot of well logs and corresponding lithological labels.

interpretability, and fast convergence, classifies data based on feature similarity and distance metrics (MacQueen, 1967). However, its performance is highly sensitive to the initial centroid selection and can yield suboptimal results when dealing with imbalanced datasets, thereby limiting its effectiveness in practical applications.

To overcome these limitations, the DBO algorithm integrates local and global search strategies, achieving rapid convergence while avoiding entrapment in local optima (Xue and Shen, 2023). DBO not only possesses strong optimization capability but also exhibits efficient convergence behavior. By embedding DBO's dual search mechanism into *K*-means, the algorithm leverages its global search potential to significantly reduce the risk of local convergence. This innovation allows the model to automatically determine the optimal number of clusters and select suitable distance metrics (Zeng et al., 2025).

The DBO–*K*-means algorithm proceeds through a six-step iterative optimization process (Zeng et al., 2024): parameter initialization, fitness evaluation, position update of dung beetles, boundary condition check, optimization of the best solution, and termination criterion judgment. Once convergence criteria are met, the algorithm outputs the globally optimal clustering configuration and fitness value. As shown in Fig. 6(b)–(d), the algorithm simulates four distinct behavioral patterns of dung beetles, each corresponding to a unique position-update strategy, mathematically expressed in Equations (1)–(4).

$$\boldsymbol{x}_i(t+1)=\boldsymbol{x}_i(t)+\alpha k\boldsymbol{x}_i(t-1)+b\Delta\boldsymbol{x},\ \Delta\boldsymbol{x}=|\boldsymbol{x}_i(t)-\boldsymbol{X}^{\mathrm{w}}| \tag{1}$$

$$\boldsymbol{x}_i(t+1)=\boldsymbol{x}_i(t)+\tan(\theta)|\boldsymbol{x}_i(t)-\boldsymbol{x}_i(t-1)| \tag{2}$$

$$\begin{aligned}\boldsymbol{x}_i(t+1)&=\boldsymbol{x}_i(t)+C_1\left(\boldsymbol{x}_i(t)-\mathbf{Lb}^{\mathrm{b}}\right)+\mathbf{C}_2\left(\boldsymbol{x}_i(t)-\mathbf{Ub}^{\mathrm{b}}\right)\\ \mathbf{Lb}^{\mathrm{b}}&=\max\left((1-R)\boldsymbol{X}^{\mathrm{b}},\mathbf{Lb}\right)\\ \mathbf{Ub}^{\mathrm{b}}&=\min\left((1+R)\boldsymbol{X}^{\mathrm{b}},\mathbf{Ub}\right)\end{aligned} \tag{3}$$

$$\boldsymbol{x}_i(t+1)=\boldsymbol{X}^{\mathrm{b}}+S\mathbf{g}\left(\left|\boldsymbol{x}_i(t)-\boldsymbol{X}^{*}\right|+\left|\boldsymbol{x}_i(t)-\boldsymbol{X}^{\mathrm{b}}\right|\right) \tag{4}$$

Here, $t$ denotes the iteration step, $\boldsymbol{x}_i(t)$ represents the position

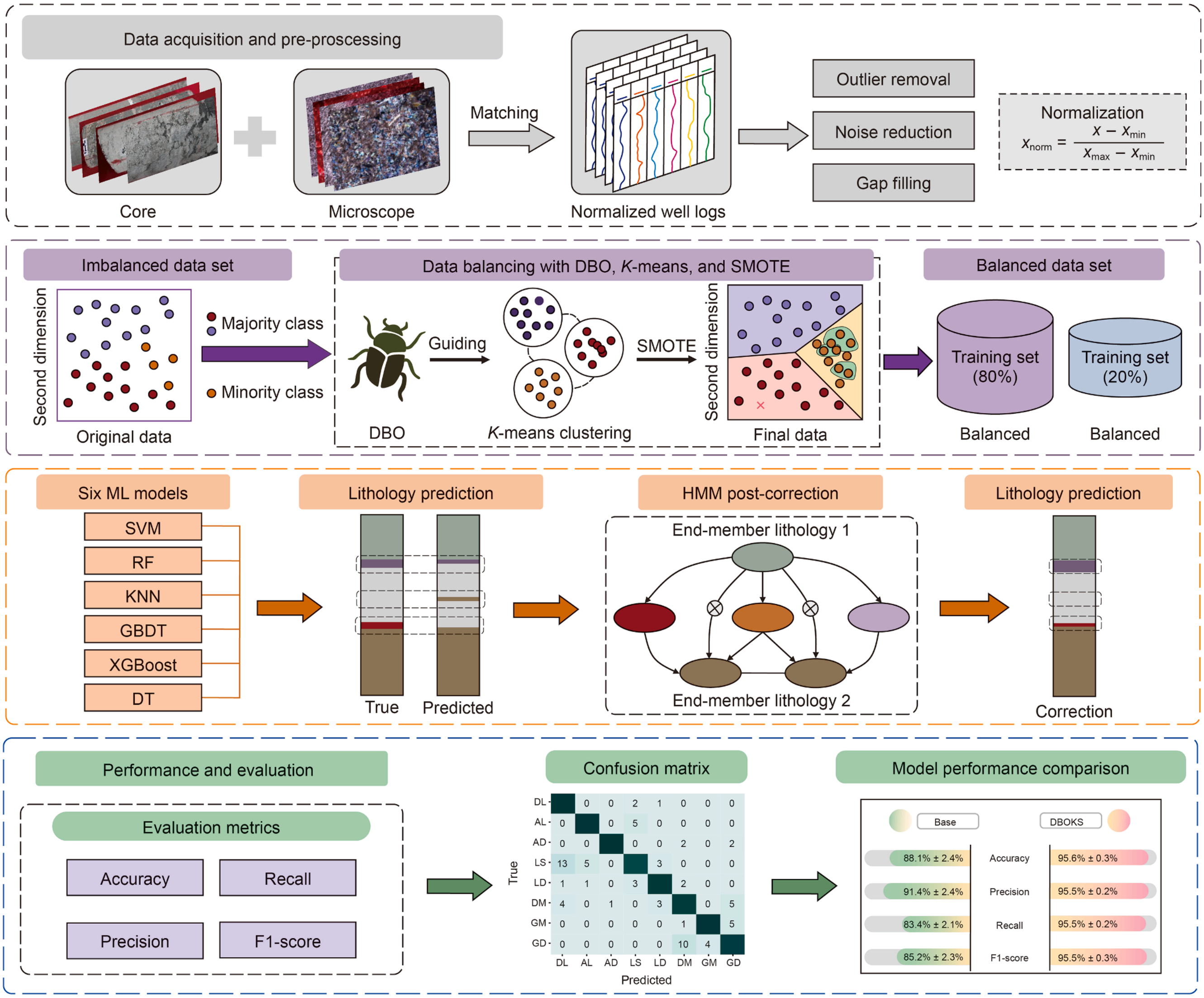


**Fig. 5.** Workflow of this study.

of the $i$-th beetle at iteration $t$, $k$ and $b$ are constants, and $\alpha$ is the bias coefficient ($\alpha = 1$ for unbiased movement, $\alpha = -1$ for reversed direction). $\boldsymbol{X}^{w}$ represents the global worst position, $\Delta \boldsymbol{x}$ models light intensity variation, and $\theta$ denotes the turning angle—when $\theta = 0$, $\pi/2$, or $\pi$, the beetle halts movement. $\mathbf{Lb}^{b}$ and $\mathbf{Ub}^{b}$ define the lower and upper bounds of the optimal foraging region, respectively. $C_1$ is a normally distributed random variable, $\boldsymbol{C}_2$ is a uniform random vector in $(0, 1)$, $\boldsymbol{X}^{b}$ denotes the best food source position, $S$ is a scaling constant, and $\boldsymbol{g}$ is a one-dimensional Gaussian random vector.

In carbonate lithology classification, transitional thin-layered carbonates often occur due to variations in depositional and tectonic environments. This leads to severe class imbalance, causing conventional ML algorithms to be biased toward majority classes and to degrade in performance. To address this issue, the present study incorporates a spatially constrained SMOTE oversampling strategy within the DBO–$K$-means subspace (Soltanzadeh and Hashemzadeh, 2021). Specifically, new synthetic samples are generated only between samples belonging to the same cluster (Chawla et al., 2002; Xu et al., 2019). The Euclidean distance between synthetic and real samples is strictly constrained, and physical thresholds are imposed on key logging parameters to ensure geological plausibility of the generated data. This approach increases the representation of transitional lithologies to levels comparable with dominant lithologies while minimizing noise introduction. The effectiveness of this technique has been validated in multiple domains. With only two parameters to optimize (k_neighbors and cluster_balance_threshold), a smaller population size is sufficient to effectively sample the search space while ensuring high computational efficiency. Therefore, we set the population size to 10. Additionally, preliminary experiments show that the DBO typically converges to a stable solution within 15 to 20 generations. Based on this observation, we set the maximum number of iterations to 20 to efficiently search for the optimal clustering parameters, thereby providing a reliable basis for subsequent oversampling within the safe cluster.

### 3.3. Machine learning framework

ML methods are widely recognized for their strong interpretability, low parameter dependence, real-time processing capability, and high accuracy in complex structural classification tasks. These advantages make ML models highly adaptable and generalizable in practical geoscientific applications (Sahin, 2020;

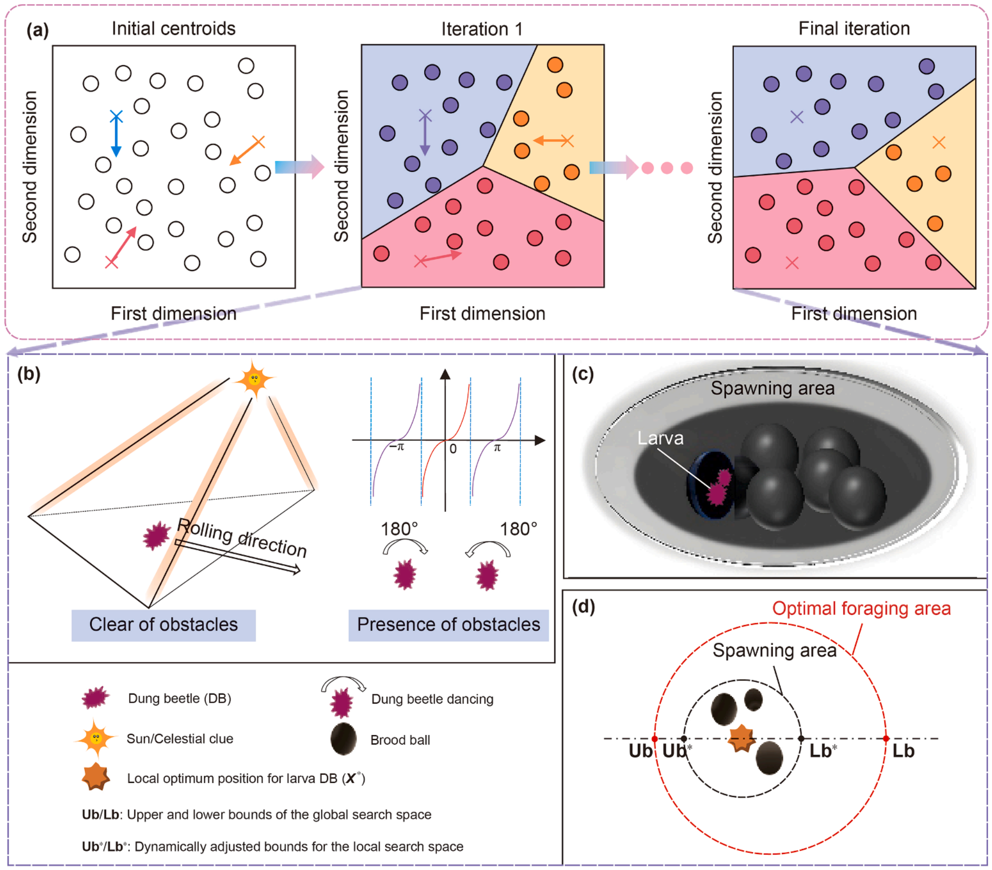


**Fig. 6.** Schematic diagram of the DBO principle (Zeng et al., 2024).

Nguyen et al., 2021; Hou et al., 2022; Shen et al., 2024). Based on the framework proposed by Zeng et al. (2024), six representative classifiers with superior performance and distinct learning mechanisms were selected as the baseline models in this study: Support Vector Machine (SVM; Fig. 7(a)), Decision Tree (DT; Fig. 7(b)), Random Forest (RF; Fig. 7(c)), Gradient Boosting Decision Tree (GBDT; Fig. 7(d)), Extreme Gradient Boosting (XGBoost; Fig. 7(e)), and K-Nearest Neighbors (KNN; Fig. 7(f)). It is noteworthy that Zeng et al. (2024) also evaluated three additional classifiers—Extra Trees (ExT), Backpropagation Neural Network (BPNN), and CatBoost. However, these were excluded from the present study due to their limitations under complex lithological conditions. Specifically, CatBoost requires substantial computation time; BPNN exhibits unstable convergence and is prone to overfitting when trained on small datasets; and ExT's random feature-splitting mechanism can occasionally assign undue importance to irrelevant features, thereby degrading model performance.

The final selection of six ML classifiers thus ensures both methodological diversity and practical robustness for lithology prediction. SVM is suitable for high-dimensional and nonlinear classification problems. Its core idea is to find an optimal hyperplane that maximizes the margin between classes, but it is sensitive to parameter selection and the size of the training data. It excels at identifying macroscopic boundaries between different lithologies in nonlinear lithology classification problems (Cortes and Vapnik, 1995; Çevik et al., 2015). Tree-based machine learning methods include DT, RF, GBDT, and XGBoost. These methods all use tree structures as the basic decision unit (Lundberg et al., 2020; Hou et al., 2022). DT provides intuitive inference, directly distinguishing lithology types through different logging response thresholds or decision paths. GBDT, through stepwise learning, better captures the subtle nonlinear boundaries between complex lithologies. RF is suitable for common local anomalies and curve jitter in logging data, offering stable evaluation of overall lithology types. Compared to RF, XGBoost typically provides higher lithology recognition accuracy for thin layers or strata with rapid sedimentary changes. KNN relies on local similarity for prediction, especially when lithology exhibits continuous and smooth distribution in the logging feature space. However, it has high computational complexity, is sensitive to noise and outliers, and requires significant storage space (Cover and Hart, 1967; Zhang et al., 2018).

### 3.4. Hidden Markov Model–based lithology sequence correction

The lithology sequence correction in this study is based on the Hidden Markov Model (HMM) framework, which conceptualizes the stratigraphic profile as a hidden state sequence representing the true lithology distribution, while the well-logging responses are treated as the corresponding observable sequence (Chen et al., 2025). Traditional machine learning models perform point-by-point classification by learning static mappings between log responses and lithology labels. However, such local predictions ignore the vertical evolution patterns and geological continuity inherent in sedimentary successions. To address this limitation, the proposed HMM-based method introduces a probabilistic transition mechanism that embeds geological prior knowledge into the classification process. Specifically, a state transition probability matrix is learned from the training data, capturing the

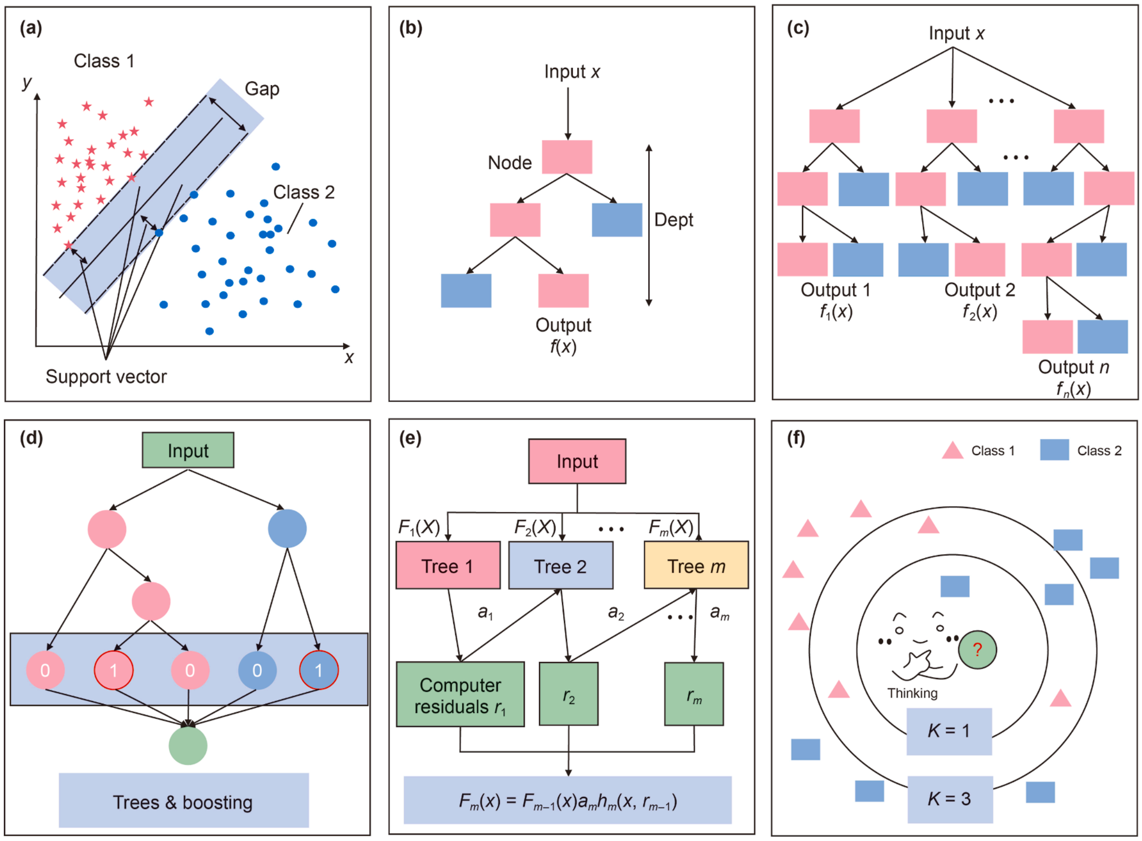


**Fig. 7.** Six machine learning methods: **(a)** SVM; **(b)** DT; **(c)** RF; **(d)** GBDT; **(e)** XGBoost; **(f)** KNN (after Zeng et al., 2024).

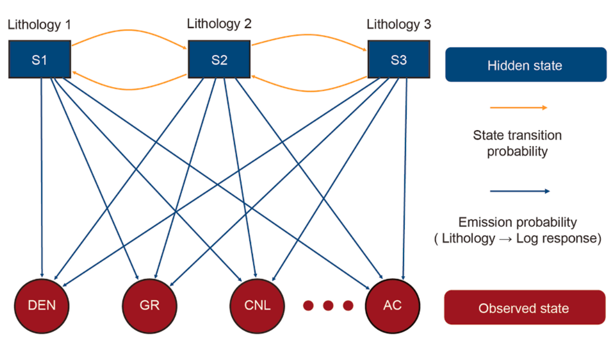


**Fig. 8.** Schematic diagram of HMM framework for lithology sequence correction.

likelihood of lithological transitions between adjacent depth intervals (Fig. 8).

By combining the posterior class probabilities predicted by the machine learning model with the lithology transition probabilities, a unified probabilistic graphical model is constructed. This framework aims to identify a lithological sequence that simultaneously maximizes the overall agreement with both the observed logging data and the expected geological continuity—thus achieving a shift from a local optimum to a global optimum in lithology interpretation.

To realize this global optimization, the Viterbi algorithm, a dynamic programming–based decoding algorithm, is employed. The Viterbi algorithm efficiently identifies the most probable sequence of lithological states that maximizes the joint probability of the entire depth series. For each depth point $t$ and each possible lithology state $j$, the algorithm recursively computes the maximum path probability $\delta_t(j)$ and records the most probable preceding state $\psi_t(j)$:

$$\delta_t(j) = \max_i \left[\delta_{t-1}(i) \cdot a_{ij}\right] \cdot b_j(O_t) \tag{5}$$

$$\psi_t(j) = \operatorname*{argmax}_i \left[\delta_{t-1}(i) \cdot a_{ij}\right] \tag{6}$$

where: $i$ and $j$ denote states (lithology types) at depth levels $t-1$ and $t$, respectively; $a_{ij}$ is the state transition probability, representing the likelihood of lithology $i$ at depth $t-1$ changing to lithology $j$ at depth $t$; $b_j(O_t)$ is the emission probability, the probability of observing log response $O_t$ given lithology $j$; $\delta_{t\text{-}1}(i)$ is the maximum probability of all paths leading to state $i$ at depth $t-1$; $\psi_t(j)$ records the most probable preceding state, enabling backtracking to reconstruct the optimal lithology sequence.

The study area shows a complete sequence of lithology that gradually transitions to a more arid sedimentary environment from bottom to top (Fig. 9). However, the transitional lithologies (such as DL and LD) have very thin intervals, and their logging response characteristics are highly similar to the lithologies above and below, making them prone to being overlooked or misclassified during identification (Fig. 9(b)). Moreover, in reservoir development sections or zones with higher clay content, the logging curves often exhibit local abrupt changes. If traditional machine learning models rely solely on simple thresholds for classification, they are highly likely to make incorrect judgments that do not align with geological principles (Fig. 9(a)). Therefore, introducing a sequence constraint mechanism is crucial, as it can effectively prevent unreasonable geological "jumps" in lithology

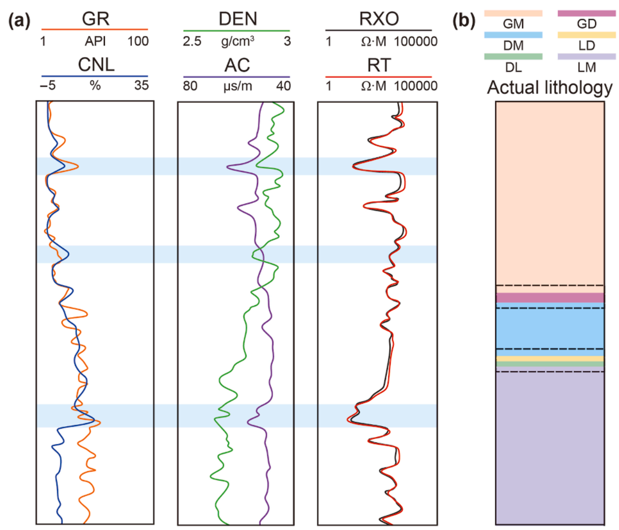


**Fig. 9.** Example of a typical well. The blue area in (a) represent regions prone to misclassification, and the dashed lines indicate transitional lithology zones.

prediction results, ensuring that the identified sequence is consistent with sedimentary evolution patterns.

### 3.5. Model training and evaluation metrics

To construct and evaluate the proposed models, both the original dataset and the DBOKS-balanced dataset were divided into training and testing subsets at a ratio of 8:2, which were used respectively for model training and performance validation. In addition, a five-fold cross-validation approach was adopted to further enhance the robustness of model assessment. This technique enables model ranking based on evaluation metrics and facilitates the selection of optimal hyperparameters. It has been widely recognized as a reliable and statistically sound model evaluation strategy (Antariksa et al., 2022). For the complex task of carbonate lithologies identification, six mainstream machine learning algorithms were employed to construct classification models. These algorithms, each possessing distinct advantages, have demonstrated strong performance across a wide range of classification problems. To further optimize model performance, the Bayesian optimization (BO) method was utilized—an efficient strategy for determining the optimal configuration of hyperparameters (Jones et al., 1998; Wu et al., 2024). This method iteratively searches for hyperparameter combinations that maximize model accuracy while leveraging prior evaluation results to guide the subsequent exploration.

Considering the trade-off between computational cost and optimization efficiency, the number of optimization iterations for each model was limited to 20, ensuring a balanced exploration–exploitation process. The results demonstrated that different hyperparameter settings had a significant impact on model accuracy. Table 1 summarizes the hyperparameter search ranges and the optimal configurations obtained for each classifier, along with their corresponding highest accuracies. The selection of optimal hyperparameters for each model was guided by the probabilistic surrogate model inherent to the Bayesian optimization framework, which estimates the likelihood of high-performance regions in the hyperparameter space. Through this systematic optimization process, each base classifier's potential was fully explored, thereby improving the overall performance of the integrated ensemble model for lithology classification.

**Table 1**
Key parameter values of six ML classifiers.

| Key parameter | Base model | Best parameter | DBOKS model | Best parameter |
|---|---|---|---|---|
| n_estimators | RF | 124 | RF | 209 |
| max_depth | | 20 | | 20 |
| min_samples_split | | 2 | | 5 |
| min_samples_leaf | | 1.0 | | 2 |
| n_estimators | GBDT | 145 | GBDT | 165 |
| learning_rate | | 9 | | 0.3 |
| max_depth | | 13 | | 10 |
| n_estimators | XGBoost | 207 | XGBoost | 165 |
| max_depth | | 11 | | 15 |
| learning_rate | | 0.11 | | 0.23 |
| subsample | | 0.83 | | 0.5 |
| colsample_bytree | | 0.63 | | 1 |
| $C$ | SVM | 6.7 | SVM | 7.8 |
| $\gamma$ | | 0.95 | | 1 |
| n_neighbors | KNN | 4 | KNN | 4 |
| weights | | 0.86 | | 1 |
| $p$ | | 1.6 | | 1 |

In classification tasks, the confusion matrix and its derived performance metrics (Fig. 10) serve as fundamental tools for comprehensive model evaluation. While classification accuracy provides a direct measure of the proportion of correctly predicted samples, its reliability declines significantly when applied to imbalanced datasets. In such cases, a classifier may achieve deceptively high accuracy simply by favoring the majority class, thereby obscuring deficiencies in identifying minority lithologies that are often geologically critical.

To address this, several complementary metrics are derived from the confusion matrix. Precision and recall provide different perspectives on model performance, revealing whether a classifier tends to overpredict (high false positives) or underpredict (high false negatives) for certain lithology types. The F1-score, defined as the harmonic mean of precision and recall, represents a balanced measure that penalizes both false alarms and missed detections. A high F1-score thus indicates that the model achieves a good trade-off between accuracy and coverage—an essential requirement in lithology classification where both correctness and completeness are critical. The mathematical definitions of the evaluation metrics are given as follows:

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

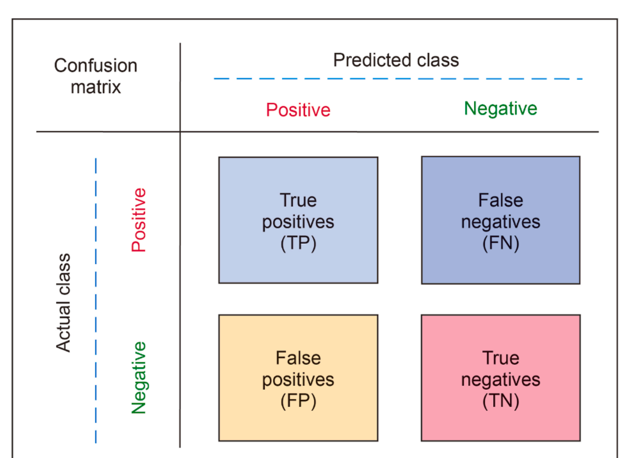


**Fig. 10.** Confusion matrix and performance indicators for classification.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

## 4. Results and discussion

### 4.1. Rationality analysis

To intuitively validate the rationale of the DBOKS oversampling method, this study employs *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE) to perform a visualization analysis of high-dimensional well log features. Fig. 11 sequentially presents the distribution of the original dataset, the dataset processed with traditional SMOTE, and the dataset processed with DBOKS in a two-dimensional feature space. In the original data (Fig. 11(a)), transition lithologies such as LD and DL are widely scattered and exhibit few samples in the two-dimensional plane. This is mainly due to severe well log response overlap and class imbalance issues, making it almost impossible to distinguish between different lithologies effectively. After processing with traditional SMOTE (Fig. 11(b)), although the sample numbers for each lithology become more balanced, the two-dimensional projection still shows mixed colors and blurred boundaries. This is because traditional SMOTE uses simple linear interpolation to generate samples, which can lead to the creation of pseudo-samples in noisy regions or near class boundaries. In contrast, the feature distribution after processing with the DBOKS method (Fig. 11(c)) shows significant improvement: most lithology classes exhibit good clustering and separability in the two-dimensional space. This improvement is attributed to the DBO algorithm integrated into DBOKS, which, with its global and local search capabilities and rapid convergence characteristics, effectively overcomes the limitation of *K*-means clustering being prone to local optima. This enables controlled and reasonable sample generation within clearly defined clustering groups. In conclusion, the DBOKS-based recognition framework not only performs robustly in balancing class distributions but also better extracts and preserves the internal structure of lithology features, thus validating the effectiveness of this method in enhancing lithology identification.

To further validate the geological plausibility of the DBOKS-generated samples, boxplot analyses (Figs. 12 and 13) were performed to compare the statistical distributions of six logging features across eight lithological classes before and after DBOKS processing. The results show that the median values of each logging feature exhibit no significant shift after oversampling (deviation <10%), and the response ranges remain nearly identical. This demonstrates that the synthetic samples do not deviate from the central tendency of real data while successfully preserving the extreme values that represent atypical geological conditions.

Overall, the validation results confirm that the DBOKS algorithm effectively balances class distributions without compromising the stratigraphic discriminability of the data. This property is particularly crucial for transitional lithologies such as LD and DL, where subtle gradients in logging features often correspond to geological boundaries. Moreover, the controlled preservation of outliers within a reasonable range ensures that the model retains sensitivity to diagenetic alterations and heterogeneity, which is essential for accurate carbonate reservoir characterization.

### 4.2. Classification performance analysis

The primary objective of this study is to propose an innovative carbonate lithology identification framework and to elucidate the intrinsic relationships between lithological types and their corresponding logging responses. To achieve this, a hybrid strategy integrating the DBOKS algorithm with sequential correction techniques was developed. As a baseline, the performance of six conventional machine learning classifiers without DBOKS preprocessing was first evaluated on the original dataset. The resulting confusion matrices for the training and testing datasets are shown in Figs. 14 and 15, respectively.

From the training confusion matrices (Fig. 14), it can be observed that the SVM model exhibits a highly dispersed prediction pattern, with numerous samples distributed across off-diagonal regions, indicating that it fails to adequately learn discriminative features from the training data. In contrast, the tree-based ensemble models—including DT, RF, GBDT, XGBoost, and KNN—display more concentrated diagonal patterns with darker shading, signifying higher classification accuracy on the training set. This contrast suggests that while linear classifiers struggle to capture the complex non-linear relationships inherent in geological–logging datasets, ensemble tree methods possess stronger representation learning capabilities and adaptability to multi-modal feature distributions. However, as shown in the testing confusion matrices (Fig. 15), the generalization performance of all classifiers decreased

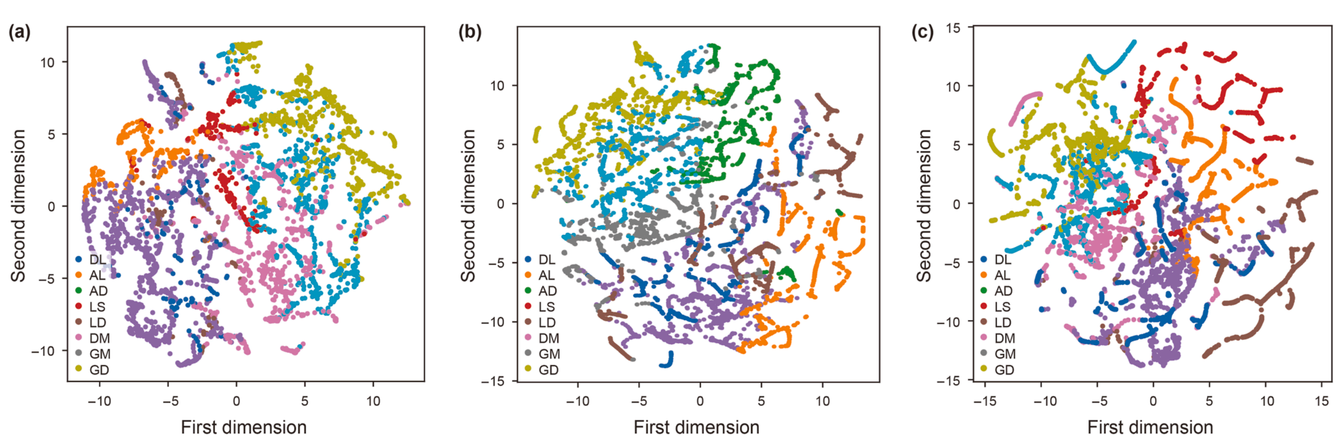


**Fig. 11.** *t*-SNE visualization of dataset: **(a)** original data, **(b)** traditional SMOTE, **(c)** DBOKS.

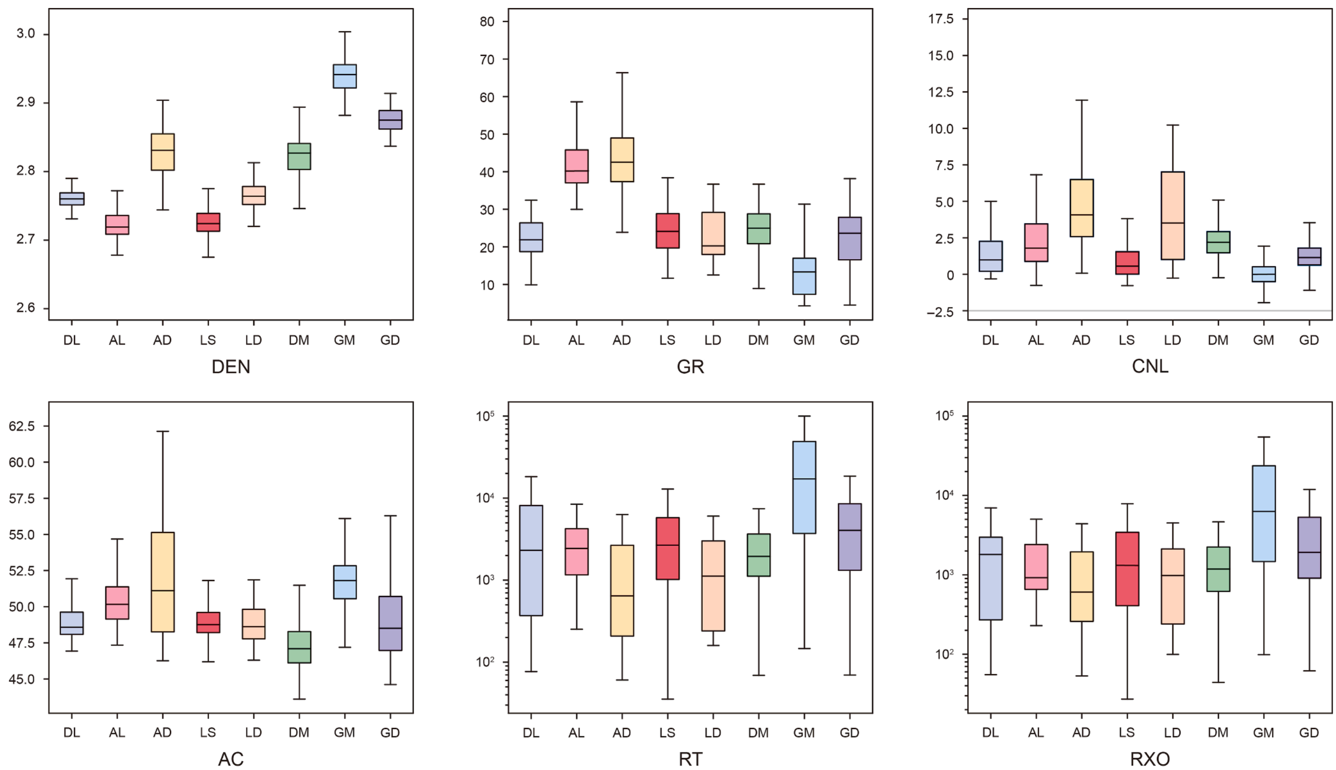


**Fig. 12.** Boxplot comparison of six logging features across eight lithologies before DBOKS processing.

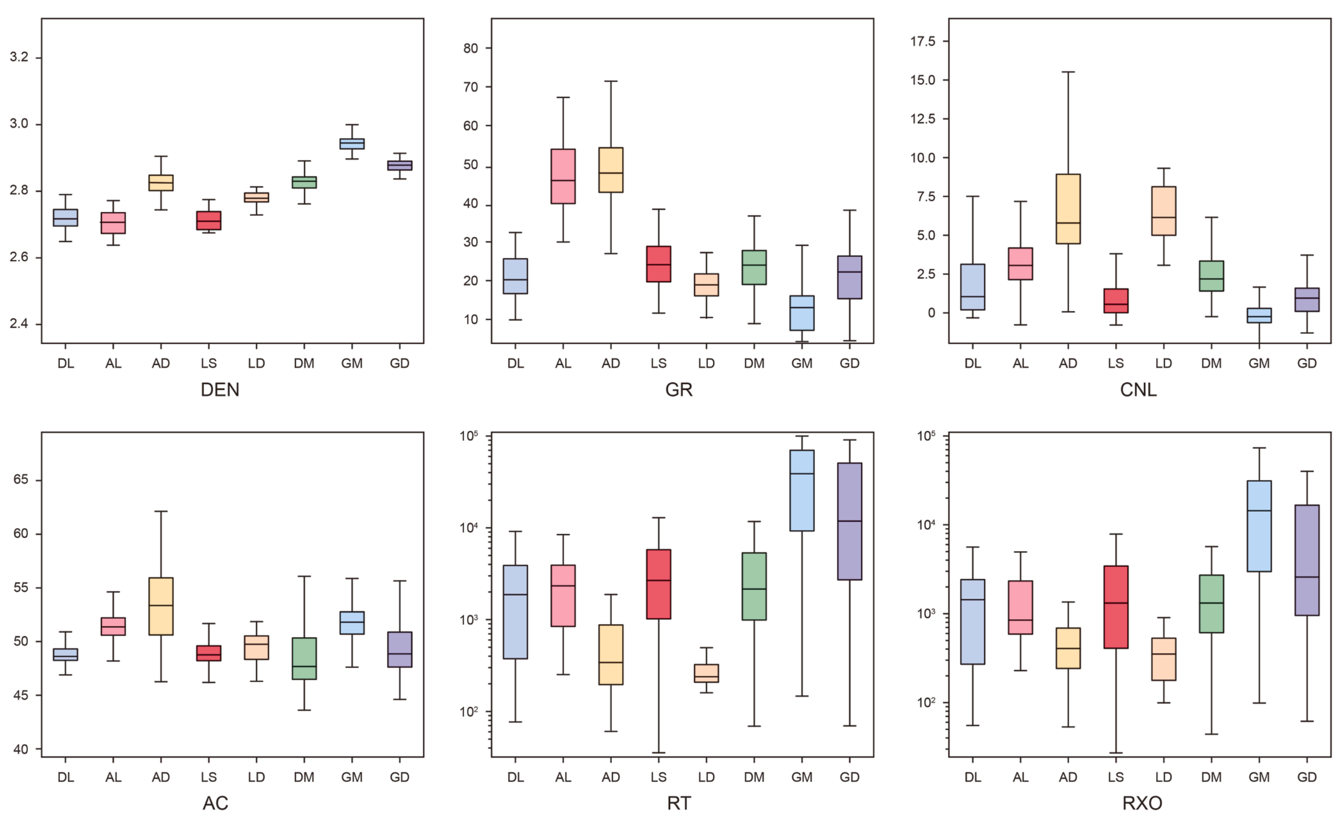


**Fig. 13.** Boxplot comparison of six logging features across eight lithologies after DBOKS processing.

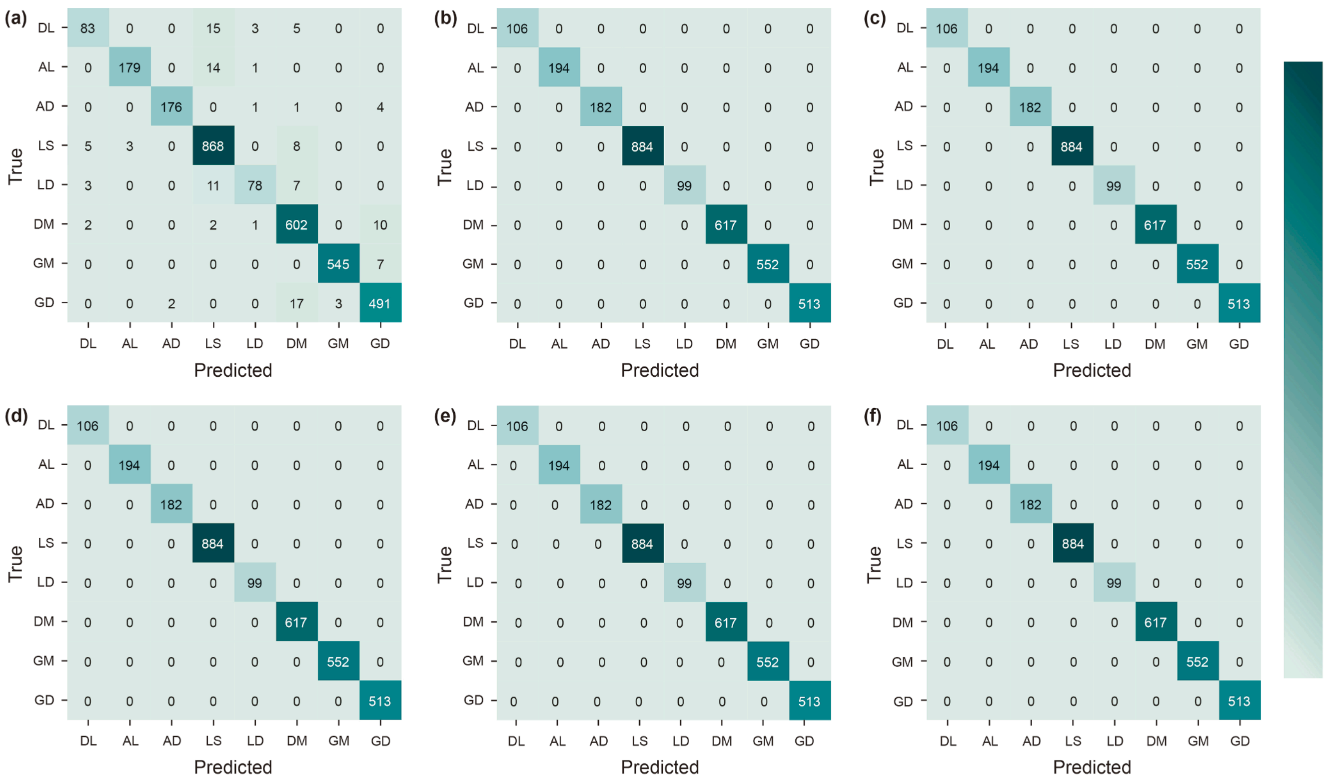


**Fig. 14.** Confusion matrices of six base classifiers on the training dataset. **(a)** SVM, **(b)** DT, **(c)** RF, **(d)** GBDT, **(e)** XGBoost, **(f)** KNN.

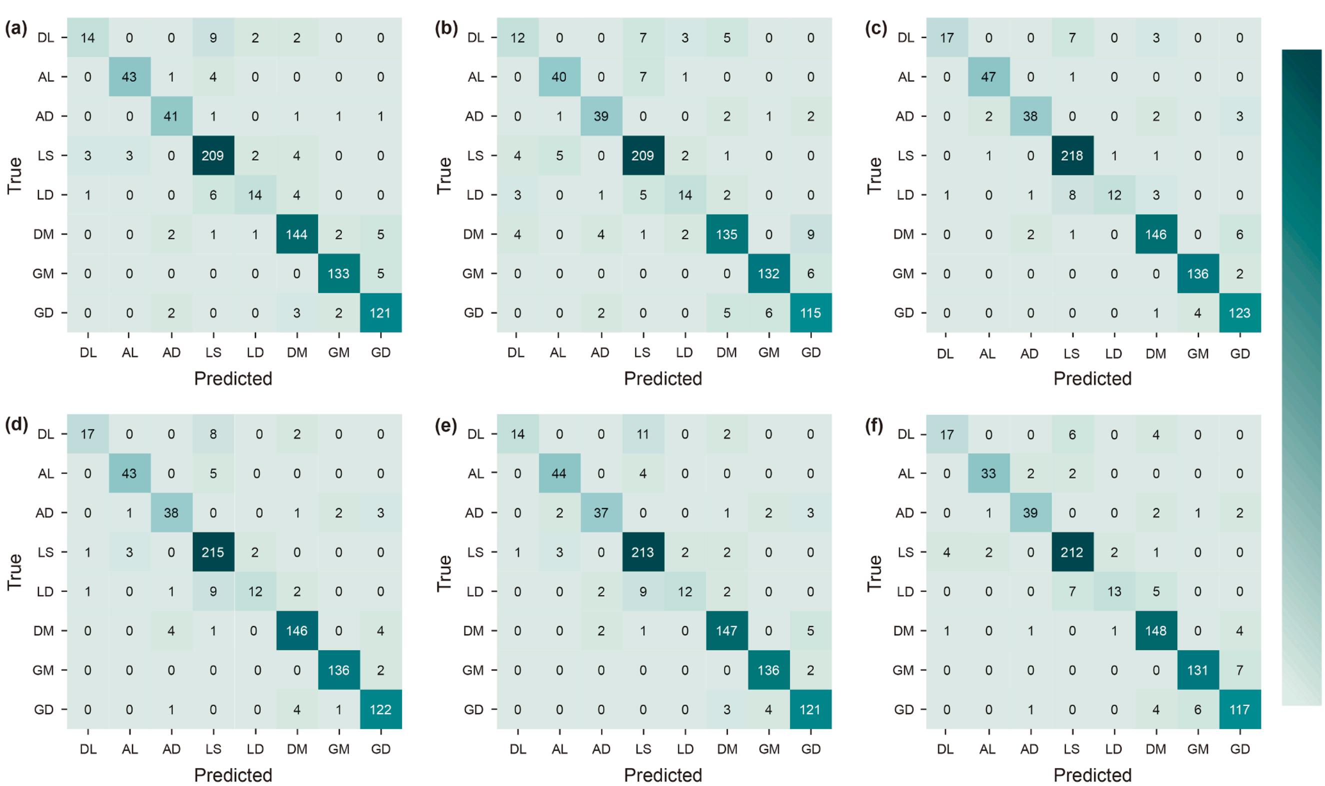


**Fig. 15.** Confusion matrices of six base classifiers on the testing dataset. **(a)** SVM, **(b)** DT, **(c)** RF, **(d)** GBDT, **(e)** XGBoost, **(f)** KNN.

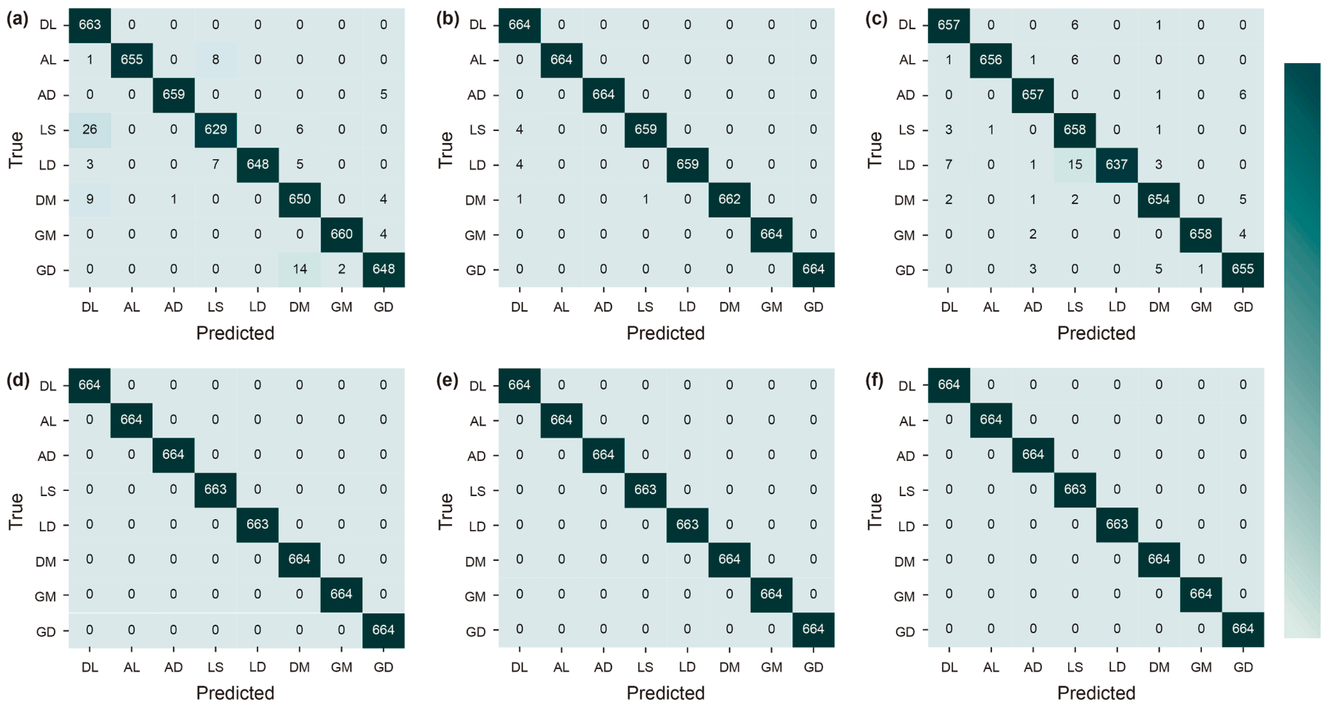

**Fig. 16.** Confusion matrices of DBOKS-enhanced classifiers on the training dataset. **(a)** SVM, **(b)** DT, **(c)** RF, **(d)** GBDT, **(e)** XGBoost, **(f)** KNN.

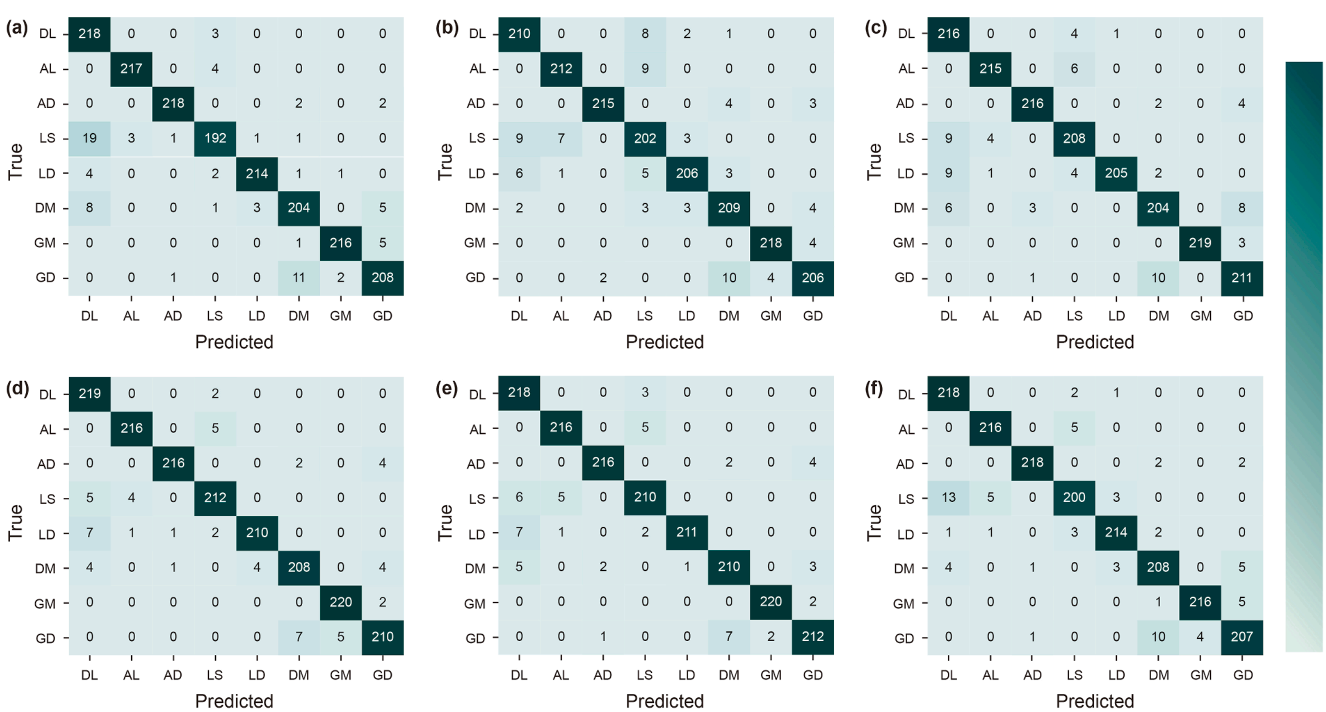

**Fig. 17.** Confusion matrices of DBOKS-enhanced classifiers on the testing dataset. **(a)** SVM, **(b)** DT, **(c)** RF, **(d)** GBDT, **(e)** XGBoost, **(f)** KNN.

markedly, particularly for minor lithological classes such as LD, DL, and AL, which suffered from severe misclassification. This performance degradation reveals that although ensemble methods can fit the training data effectively, their decision boundaries remain highly dependent on the quality and balance of the training samples. Consequently, the class imbalance and feature overlap in the original dataset substantially limit the model's generalization and practical applicability.

To evaluate the enhancement effect of DBOKS, each base classifier was retrained on the DBOKS-balanced dataset, and the corresponding confusion matrices for the training and testing sets are illustrated in Figs. 16 and 17, respectively. The results show that while SVM and RF still exhibited minor misclassifications—primarily among lithologies with similar mineralogical compositions and overlapping log responses (e.g., LS vs. DL, DM vs. DL, and GD vs. DM)—the remaining four models achieved near-optimal classification performance during training.

More importantly, compared with the unbalanced test results (Fig. 15), the DBOKS-enhanced models demonstrated substantial performance improvement across all lithology categories, particularly for transitional and minority lithologies such as AL, DL, and LD. This improvement highlights that DBOKS effectively

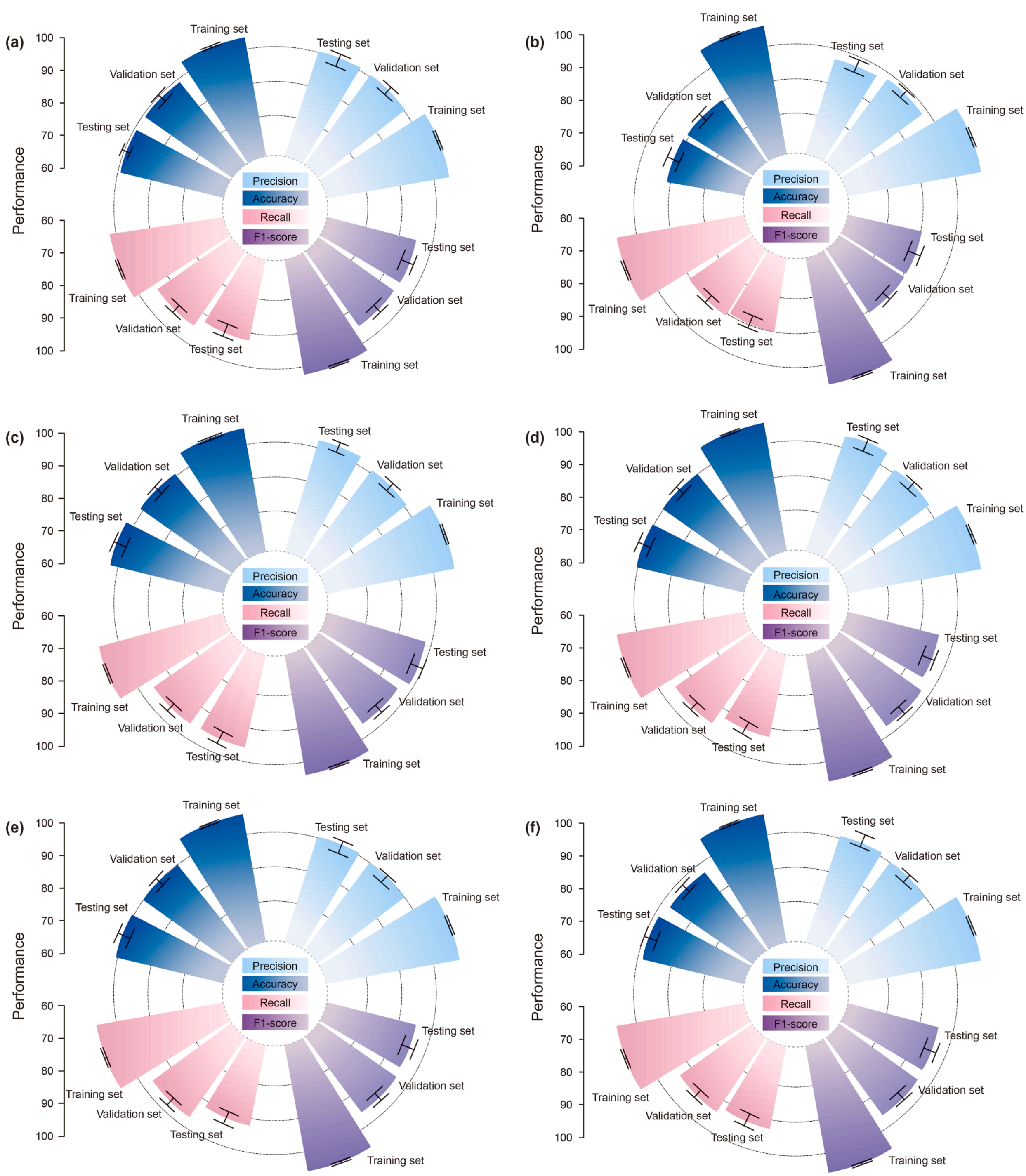


**Fig. 18.** Petal plots of baseline classifiers showing performance metrics across training, validation, and testing datasets.(**a**) SVM, (**b**) DT, (**c**) RF, (**d**) GBDT, (**e**) XGBoost, (**f**) KNN.

synthesizes minority class samples and optimizes the feature space distribution, thereby strengthening the discriminative power of the classifiers within fuzzy feature regions.

In summary, the integration of DBOKS with machine learning classifiers not only improves dataset balance but also enhances the model's feature representation and geological interpretability. This dual improvement substantially increases the recognition accuracy of complex carbonate lithological assemblages, providing a robust data foundation for high-resolution lithology interpretation in heterogeneous reservoirs.

### 4.3. Recognition performance analysis

While confusion matrices provide key insights into potential classification biases, a comprehensive evaluation of model fidelity requires additional metrics such as accuracy, precision, recall, and F1-score. To visualize and compare the overall performance consistency, Figs. 18 and 19 present petal plots illustrating the performance of the six baseline classifiers and their DBOKS-enhanced counterparts across the training, validation, and testing datasets. Each petal represents the magnitude of a single performance indicator, while the overall shape reflects the model's consistency and generalization ability.

As shown in Fig. 18, the radar plots of the six baseline classifiers—SVM, DT, RF, GBDT, XGBoost, and KNN—exhibit pronouncedly irregular and jagged contours, indicating large discrepancies in model behavior across datasets and insufficient generalization. Although most models maintained relatively high accuracy, their precision, recall, and F1-scores were consistently lower, implying that these models tended to rely excessively on majority-class samples while failing to capture minority lithology patterns. Such imbalance-induced bias can lead to misinterpretation of critical lithological boundaries, posing significant risks in geological decision-making.

Specifically, SVM and DT show markedly smaller radar areas compared with other models, reflecting inferior overall performance. Their metrics on the validation and testing sets remained below 85%, whereas RF, GBDT, XGBoost, and KNN achieved moderately better performance (85%–90%). This indicates possible overfitting or limited generalization capability in the baseline models. From a geological perspective, this discrepancy stems from the inherent overlap and nonlinearity among logging responses of transitional carbonate lithologies. Transitional types such as dolomitic limestone DL and LD share similar mineralogical compositions, leading to highly correlated or overlapping well-log features. Under these conditions, linear classifiers (e.g., SVM) relying on hyperplane separation, and simple split-based models (e.g., DT) constrained by univariate decision boundaries, struggle to capture subtle petrophysical variations. Consequently, these models exhibit high misclassification rates for transitional facies, underscoring their limitations in practical stratigraphic interpretation.

In contrast, Fig. 19 clearly demonstrates that the DBOKS-enhanced classifiers achieve substantially more uniform and symmetric radar shapes, indicating improved model consistency across all datasets. The previously observed jagged oscillations have largely disappeared, and the radar areas have expanded considerably—signifying a marked enhancement in overall classification performance. Under the DBOKS framework, all six classifiers achieved accuracy above 90%, with five (except DT) exceeding 95% on both validation and testing datasets.

This improvement stems from DBOKS's dual optimization strategy, which addresses class imbalance and feature sparsity simultaneously. By generating synthetic minority samples within safe clusters along the feature-space trajectories connecting neighboring minority instances, DBOKS avoids naive duplication of data and instead reinforces the representation of transitional lithologies such as DL and LD. This process preserves intra-class variance while mitigating overfitting to majority classes, thereby enhancing the model's ability to recognize subtle geological patterns and improving the geological interpretability and robustness of lithology classification.

A clear comparison between Figs. 18 and 19 reveals that the petal plots of the DBOKS-enhanced classifiers are markedly fuller and more symmetric, demonstrating substantial improvements in model generalization and robustness. To further quantify the performance gain brought by DBOKS, the results indicate a consistent and systematic improvement across all evaluation metrics, with particularly pronounced enhancements in recall and F1-score. These findings confirm the effectiveness of DBOKS in reconstructing sample distributions, optimizing feature-space structures, and improving performance in stratigraphic classification tasks.

Importantly, the improvement in the recognition of transitional lithologies holds significant geological implications. These lithologies typically exhibit thin bed thickness, limited sampling points, and complex mineral compositions with strongly overlapping well-log responses, resulting in severe class imbalance and sample scarcity. Moreover, transitional facies such as DL and LD often correspond to favorable hydrocarbon reservoirs, characterized by complex pore structures, high petrophysical heterogeneity, and ambiguous logging boundaries. Traditional machine learning models struggle to identify these lithologies due to blurred decision boundaries and insufficient feature representation. By generating synthetic minority samples and optimizing their spatial distribution, DBOKS significantly enhances the model's ability to distinguish these geologically critical lithologies with higher accuracy and reliability, highlighting its potential for real-world geological applications.

To gain deeper insight into the differential performance characteristics of the classifiers, it is essential to examine how lithological categories are represented in the geological feature space and how misclassification patterns emerge. By systematically comparing the spatial distributions of true and predicted lithology labels, one can not only evaluate the model's sensitivity to key geological features but also assess the degree of alignment between the learned decision boundaries and underlying geological processes. Figs. 18 and 19 present the lithology classification results before and after DBOKS enhancement, respectively, enabling a detailed analysis of systematic misclassifications and local confusion regions from a geological–machine learning perspective.

In Fig. 20, the baseline classifiers exhibit substantial misclassifications in transitional lithologies. Specifically, samples of DM are frequently misclassified as LS, while DL samples are systematically predicted as LS or DM. Additionally, lithologies such as GD, GM, and DM show varying degrees of confusion. These results suggest that traditional classifiers face significant limitations when confronted with highly overlapping feature responses and unclear interclass boundaries. Geologically, this can be attributed to the continuous dolomitization and calcitization processes that generate complex mineralogical compositions and heterogeneous pore structures, resulting in extensive feature-space overlap. Consequently, conventional algorithms struggle to form sufficiently discriminative decision boundaries. Moreover, the study area's arid to semi-arid paleoenvironment and high-frequency sea-level fluctuations promote the development of thin interbedded strata and limited core control, exacerbating sample scarcity and representation bias in minority classes.

As shown in Fig. 21, after DBOKS enhancement, the error rates of all classifiers are significantly reduced, and misclassifications are effectively suppressed—particularly for minority and transitional lithologies, where recall and F1-scores exhibit marked improvements. Nevertheless, it is worth noting that while DBOKS effectively

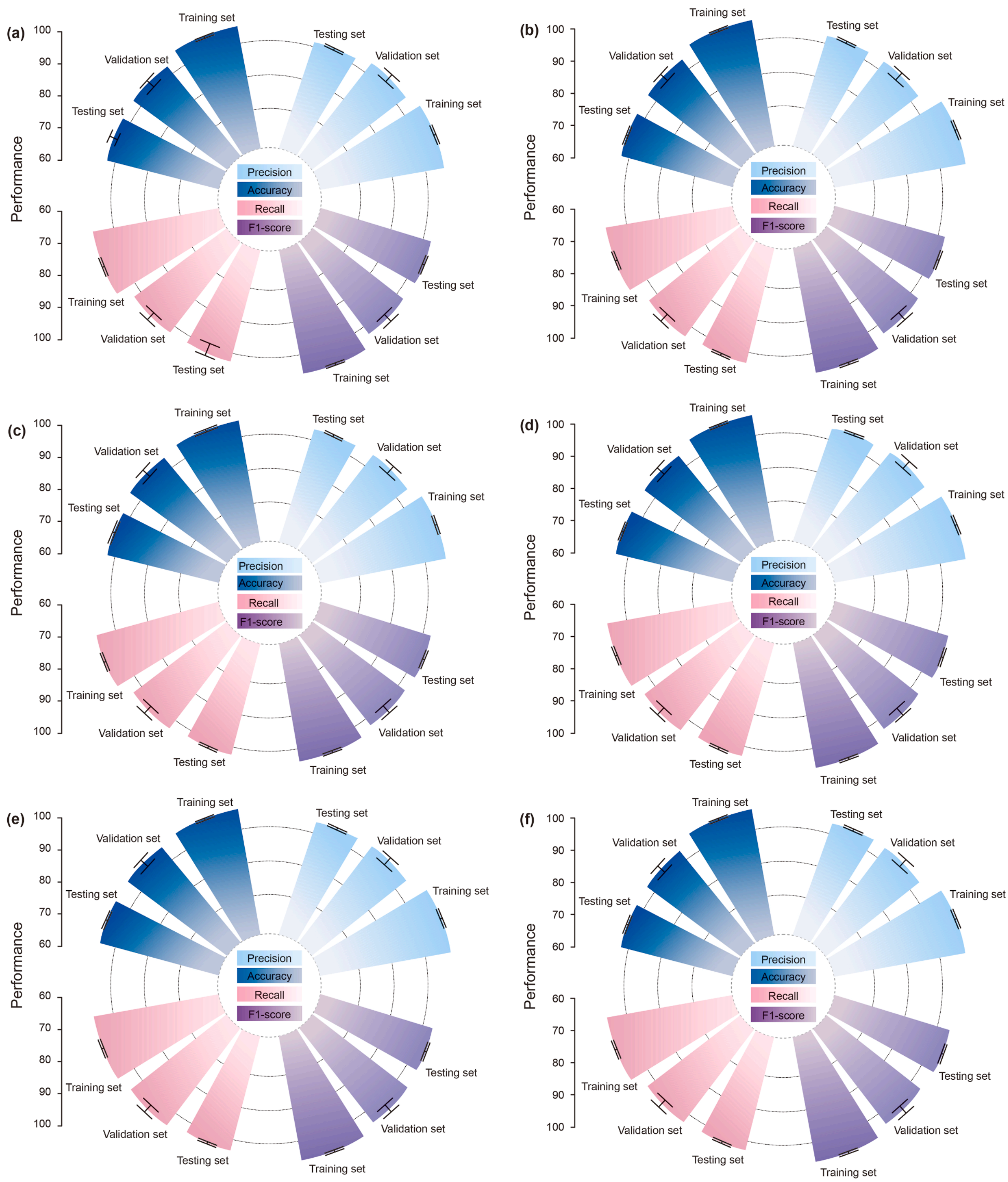


**Fig. 19.** Petal plots of DBOKS-enhanced classifiers showing improved consistency and overall classification performance. **(a)** SVM, **(b)** DT, **(c)** RF, **(d)** GBDT, **(e)** XGBoost, **(f)** KNN.

mitigates the representation bias caused by class imbalance, it does not entirely eliminate the classification challenges posed by compositional gradients and vertical continuity within stratigraphic sequences. Residual misclassifications still occur near lithological transition zones, primarily because conventional machine learning models rely on static feature vectors and fail to account for the temporal–spatial dependencies and geological continuity inherent in sequential stratigraphic data.

## 4.4. Sequence correction analysis

To fundamentally enhance the model's capability in lithological sequence discrimination, the HMM combined with the Viterbi decoding algorithm was introduced. This framework explicitly models both transition probabilities among lithological states and their sequential dependencies. By integrating the probabilistic outputs of multiple classifiers, the method captures vertical

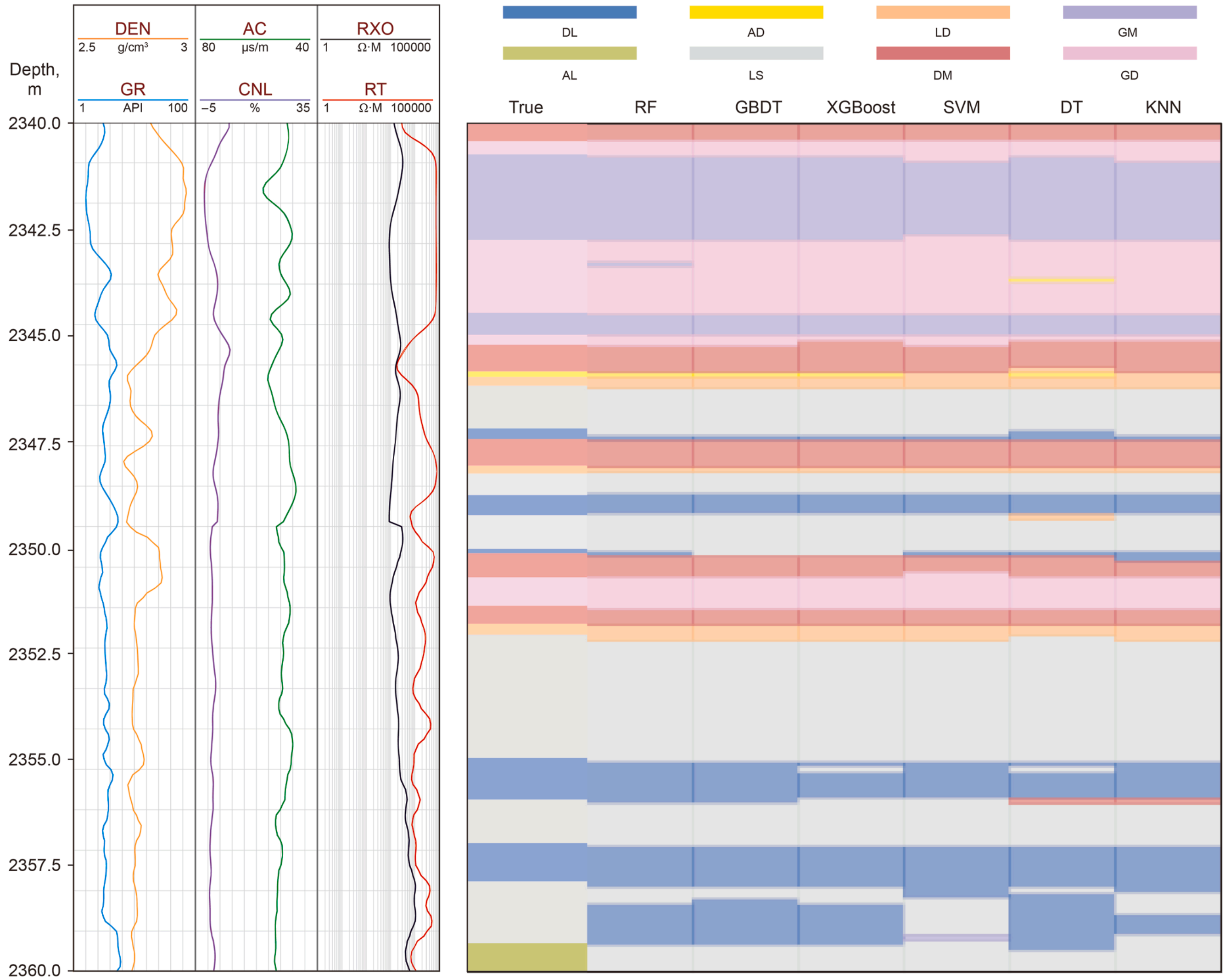


**Fig. 20.** Predicted lithology classification results from the baseline classifiers compared with the true lithology labels for Well Y10.

lithological evolution patterns within stratigraphic sequences rather than treating each depth point as an independent sample.

Among the six classifiers enhanced by DBOKS, significant variations in performance were observed under complex geological conditions. Based on preliminary experiments, three ensemble models (RF, GBDT, and XGBoost) were selected for sequence correction and visual comparison (Fig. 22). This selection was made for the following reasons: SVM and KNN models exhibited limited capability in handling highly nonlinear and feature-overlapping well-log data, making them inadequate for capturing subtle lithological variations; although DT model offers interpretability, its training time was considerably higher than that of ensemble methods, reducing scalability for large-scale datasets.

After sequence correction, the results demonstrated that the integration of HMM and Viterbi markedly strengthened the modeling of geologically constrained transition rules, effectively reducing misclassification between lithologies that are geologically non-contiguous. Specifically, the misclassification rate between LD and LM was substantially reduced, while the recognition accuracy of transitional lithologies such as LD and GD was significantly improved. This enhancement can be attributed to the fact that lithological evolution in carbonate sequences is typically continuous and gradual, governed by sedimentary continuity, progressive diagenesis, and stage-dependent tectonic deformation, whereas abrupt transitions are geologically rare. Consequently, the post-correction lithological predictions align more closely with actual geological stratigraphy.

The application of the Viterbi algorithm for sequence optimization not only preserves the advantages of DBOKS—namely, enhanced feature balance and minority class augmentation—but also introduces an explicit modeling of vertical sequence dependency. This integration allows the model to effectively capture the sequential characteristics of lithological transitions, thereby addressing a key theoretical limitation of traditional machine learning approaches that rely solely on static feature vectors. In essence, the HMM–Viterbi correction provides a mechanistic link between the temporal–spatial continuity of stratigraphic processes and the probabilistic structure of the classification model. The combined DBOKS–HMM framework thus establishes a new direction for achieving high-precision and high-robustness intelligent stratigraphic interpretation.

### 4.5. *Limitations and future work*

This study presents an empirical evaluation of a DBOKS-based hybrid machine learning and sequence correction framework for carbonate lithology classification. The results demonstrate that the

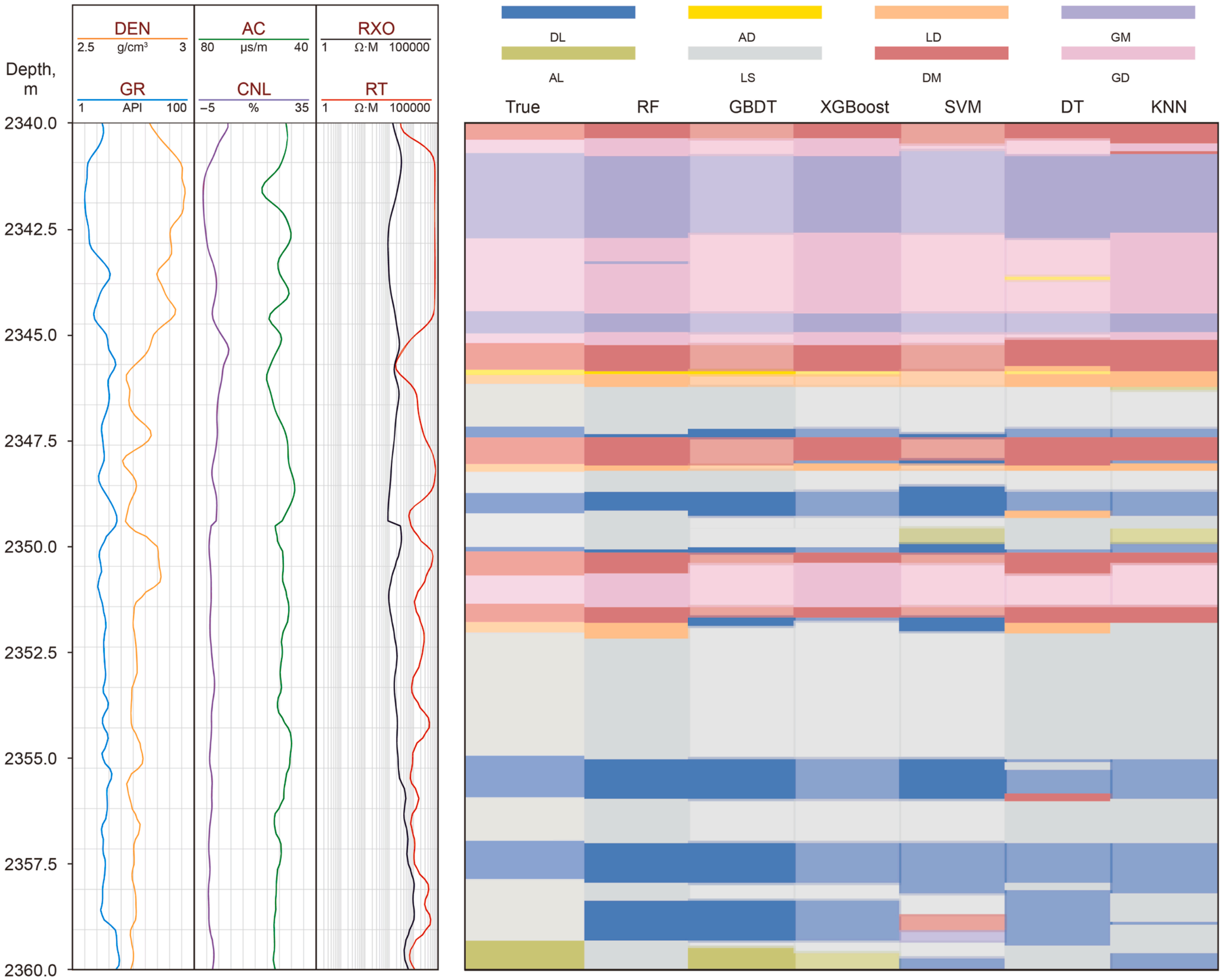


**Fig. 21.** Predicted lithology classification results from the DBOKS-enhanced classifiers compared with the true lithology labels for Well Y10.

proposed approach effectively mitigates class imbalance and enhances lithology identification accuracy. However, due to the limited data sources and regional scope, the generality and transferability of the method require further validation across broader geological settings and diverse datasets. Based on these findings, several promising directions for future research are proposed:

(1) Although the DBOKS method effectively preserves the statistical distribution characteristics of single variables, its performance in maintaining multivariate relationships still requires systematic evaluation. In complex carbonate rock lithology identification, the synergistic variations and correlations between multiple logging curves provide crucial geological information. Ensuring that synthetic samples retain these multivariate logging feature relationships is vital for the model's reliability. Future research will incorporate distribution consistency tests based on Mahalanobis distance to quantify the validity of synthetic samples in the multivariate feature space.
(2) This study has not yet conducted a systematic importance analysis of the logging features upon which the model relies. Local interpretability methods hold significant potential for analyzing the prediction results of lithology classification models, especially for layers with low classification confidence or conflicting geological understanding. Future research plans to integrate explainable artificial intelligence (XAI) methods, such as SHAP and LIME, to reveal the key logging responses and feature combinations upon which the model's decisions are based. This will provide intuitive and traceable evidence for geological experts, enhancing the model's credibility and practical utility in exploration and development.
(3) Although the framework proposed in this study has made progress in improving carbonate rock lithology identification accuracy, the current model is only trained and validated based on conventional logging curves. Limited by a single data source, the model's recognition ability and generalization performance still face significant bottlenecks. To further overcome this limitation, integrating multi-source heterogeneous data is considered a key path to improving the geological consistency and extrapolation capability of lithology identification models. By incorporating multimodal information, such as logging curves, seismic attributes, core scan images, and geochemical data, and leveraging cross-modal representation learning

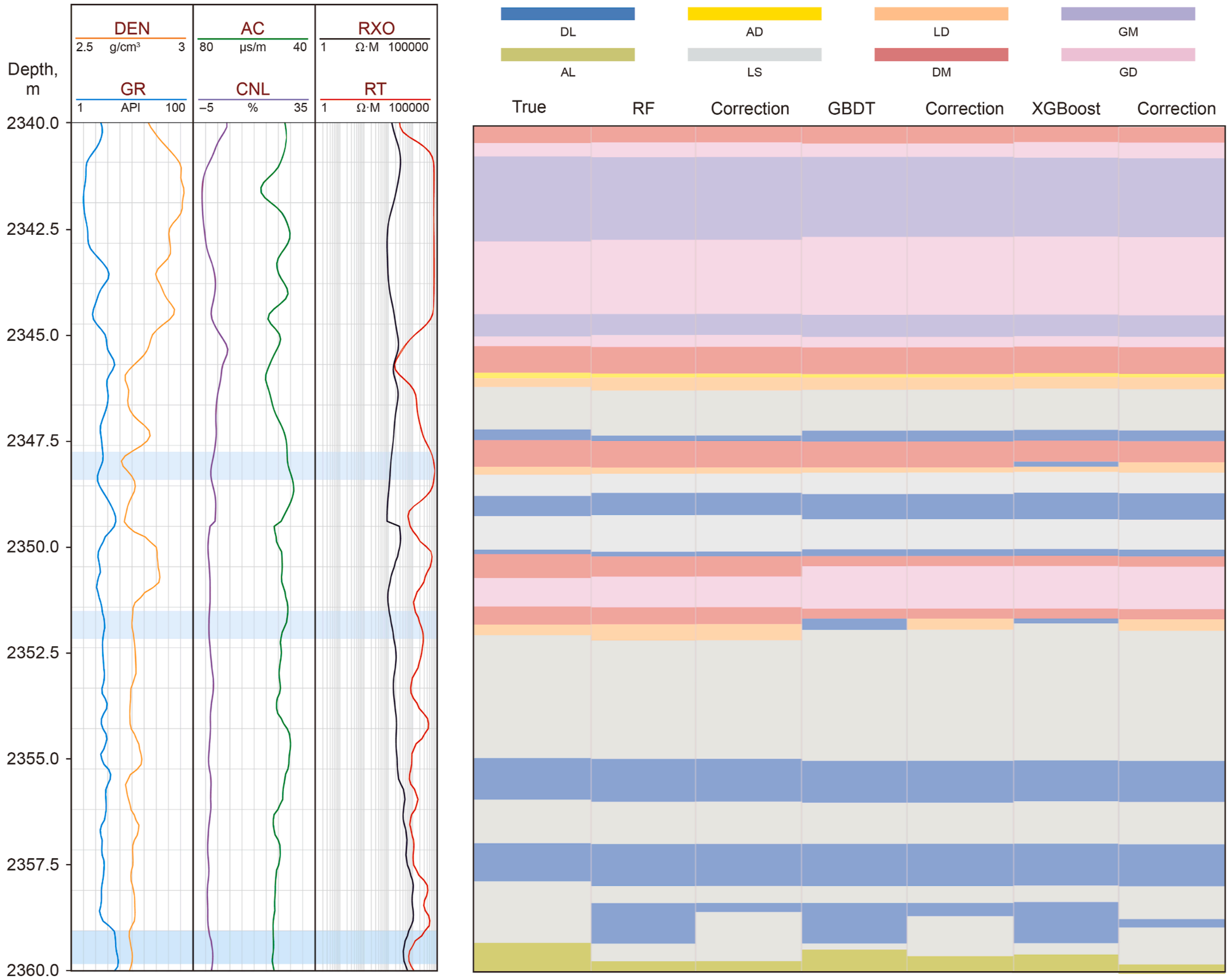


**Fig. 22.** Comparison chart of lithology correction application for Well Y10.

methods to extract complementary features, it is expected to develop a more robust and interpretable integrated recognition model. This direction not only helps overcome the limitations of single data sources but also promotes fine reservoir characterization through well-seismic synergy and lithology prediction at the oilfield scale, offering substantial practical value for exploration and development decisions in complex oil and gas reservoirs.

## 5. Conclusions

This study addresses the challenges of carbonate lithology classification, where highly similar mineral compositions, overlapping logging responses, rapid sedimentary changes, and thin transitional layers often lead to severe class imbalance and feature overlap. To enhance geological consistency and predictive reliability, we developed a hybrid machine learning framework that integrates the DBOKS oversampling technique with an HMM–Viterbi–based sequence correction. The proposed approach was validated using core descriptions and thin-section data, yielding the following main conclusions:

(1) The DBOKS method effectively mitigates model bias caused by the scarcity of transitional lithology samples through optimization of the clustering strategy. This method not only enhances the representation of minority class samples but also preserves the statistical distribution characteristics of the original logging response features, providing a more geologically meaningful and interpretable training data foundation for subsequent classification models.

(2) By introducing the HMM framework, this study significantly improves the model's ability to capture lithology transition patterns in the vertical stratigraphic sequence. This method effectively constrains lithology prediction results, prevents unreasonable stratigraphic jumps, and compensates for the shortcomings of traditional machine learning models that neglect sedimentary sequences and geological evolution processes, making the identification results more consistent with the gradual lithology transitions in actual geological environments. This improvement drives the evolution of lithology identification methods from "static point classification" to the paradigm of "dynamic sequence analysis."

(3) Among six mainstream machine learning models, tree-based ensemble models exhibit superior classification

stability and generalization capability. After joint optimization with the DBOKS method and sequence correction, the overall recognition accuracy and F1-score of these models increases to about 95%. This indicates that the coupling strategy of data augmentation and sequence modeling can more fully exploit the implicit nonlinear features and lithology transition information in logging data, significantly improving the reliability of lithology identification under complex reservoir conditions. This study also provides a reference for selecting appropriate machine learning models under different geological conditions.

## CRediT authorship contribution statement

**Wen-Chuang Li:** Writing – original draft, Validation, Software, Methodology, Investigation, Conceptualization. **Zhong-Xiang Zhao:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **You-Bin He:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Lian-Hua Wu:** Data curation, Formal analysis, Visualization. **Yu-Qing Zhang:** Data curation, Formal analysis, Investigation, Visualization.

## Code availability section

The relevant codes are written in Python. To access the source file of relevant data and codes, one can visit the repository on GitHub (https://github.com/20020914-bug/lithology-predict).

## Declaration of competing interest

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

## Acknowledgements

The authors would like to acknowledge the support by the National Natural Science Foundation of China with the award number 42002165 and 42272115.

## References

Abbey, C.P., Okpogo, E.U., Atueyi, I.O., 2018. Application of rock physics parameters for lithology and fluid prediction of 'TN' field of Niger Delta basin, Nigeria. Egypt. J. Petrol. 27 (4), 853–866. https://doi.org/10.1016/j.ejpe.2018.01.001.

Aftab, S., Moghadam, R.H., Leisi, A., 2023. New interpretation approach of well logging data for evaluation of Kern aquifer in South California. J. Appl. Geophys. 215, 105138. https://doi.org/10.1016/j.jappgeo.2023.105138.

Anifowose, F., Labadin, J., Abdulraheem, A., 2015. Improving the prediction of petroleum reservoir characterization with a stacked generalization ensemble model of support vector machines. Appl. Soft Comput. 26, 483–496. https://doi.org/10.1016/j.asoc.2014.10.017.

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

Bai, X.L., Zhang, S.N., Huang, Q.Y., et al., 2016. Origin of dolomite in the middle Ordovician peritidal platform carbonates in the northern Ordos basin, Western China. Pet. Sci. 13 (3), 434–449. https://doi.org/10.1007/s12182-016-0114-5.

Bhandari, K., Kumar, K., Sangal, A.L., 2024. DBOS_US: A density-based graph undersampling method to handle class imbalance and class overlap issues in software fault prediction. J. Supercomput. 80 (15), 22682–22725. https://doi.org/10.1007/s11227-024-06312-5.

Carrasquilla, A., 2023. Lithofacies prediction from conventional well logs using geological information, wavelet transform, and decision tree approach in a carbonate reservoir in southeastern Brazil. J. S Am N Earth Sci. 128, 104431. https://doi.org/10.1016/j.jsames.2023.104431.

Çevik, A., Kurtoğlu, A.E., Bilgehan, M., et al., 2015. Support vector machines in structural engineering: A review. J. Civ. Eng. Manag. 21 (3), 261–281. https://doi.org/10.3846/13923730.2015.1005021.

Chang, H.C., Kopaska-Merkel, D.C., Chen, H.C., et al., 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.

Chawla, N.V., Bowyer, K.W., Hall, L.O., et al., 2002. SMOTE: Synthetic minority oversampling technique. J. Artif. Intell. Res. 16 (1), 321–357. https://doi.org/10.1613/jair.953.

Chen, H.Y., Geoffrey Shen, Q., Skibniewski, M.J., et al., 2025. Dynamic prediction and optimization of tunneling parameters with high reliability based on a hybrid intelligent algorithm. Inf. Fusion 114, 102705. https://doi.org/10.1016/j.inffus.2024.102705.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297. https://doi.org/10.1007/BF00994018.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theor. 13 (1), 21–27. https://doi.org/10.1109/TIT.1967.1053964.

Delavar, M.R., 2022. Hybrid machine learning approaches for classification and detection of fractures in carbonate reservoir. J. Petrol. Sci. Eng. 208, 109327. https://doi.org/10.1016/j.petrol.2021.109327.

Dong, S.Q., Zhong, Z.H., Cui, X.H., et al., 2023. A deep kernel method for lithofacies identification using conventional well logs. Pet. Sci. 20 (3), 1411–1428. https://doi.org/10.1016/j.petsci.2022.11.027.

Gao, Y., Zhang, Y., Xiong, Z., et al., 2024. Optimal scheduling model of microgrid based on improved dung beetle optimization algorithm. Syst. Sci. Control Eng. 12 (1), 2385330. https://doi.org/10.1080/21642583.2024.2385330.

Ghosh, S., Chatterjee, R., Shanker, P., 2016. Estimation of ash, moisture content and detection of coal lithofacies from well logs using regression and artificial neural network modelling. Fuel 177, 279–287. https://doi.org/10.1016/j.fuel.2016.03.001.

He, M., Gu, H., Wan, H., 2020. Log interpretation for lithology and fluid identification using deep neural network combined with MAHAKIL in a tight sandstone reservoir. J. Petrol. Sci. Eng. 194, 107498. https://doi.org/10.1016/j.petrol.2020.107498.

Hou, S.K., Liu, Y.R., Yang, Q., 2022. Real-time prediction of rock mass classification based on TBM operation big data and stacking technique of ensemble learning. J. Rock Mech. Geotech. Eng. 14 (1), 123–143. https://doi.org/10.1016/j.jrmge.2021.05.004.

Javadi, S., Hashemy, S.M., Mohammadi, K., et al., 2017. Classification of aquifer vulnerability using *K*-means cluster analysis. J. Hydrol. 549, 27–37. https://doi.org/10.1016/j.jhydrol.2017.03.060.

Jones, D.R., Schonlau, M., Welch, W.J., 1998. Efficient global optimization of expensive black-box functions. J. Global Optim. 13 (4), 455–492. https://doi.org/10.1023/A:1008306431147.

Katz, B.J., Everett, M.A., 2016. An overview of pre-Devonian petroleum systems — unique characteristics and elevated risks. Mar. Petrol. Geol. 73, 492–516. https://doi.org/10.1016/j.marpetgeo.2016.03.019.

Kazemi, F., Asgarkhani, N., Shafighfard, T., et al., 2025. Machine-learning methods for estimating performance of structural concrete members reinforced with fiber-reinforced polymers. Arch. Comput. Methods Eng. 32 (1), 571–603. https://doi.org/10.1007/s11831-024-10143-1.

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

Lei, B.J., Zhou, Y.Z., Yang, J.L., et al., 2006. Sedimentary facies of the second member of the Jialingjiang formation in the central-southern Sichuan transition zone and its distribution characteristics. Geol. China 33 (5), 1149–1158. https://doi.org/10.3969/j.issn.1000-3657.2006.05.027 (in Chinese).

Li, Y., Kang, Z., Xue, Z., et al., 2018. Theories and practices of carbonate reservoirs development in China. Petrol. Explor. Dev. 45 (4), 712–722. https://doi.org/10.1016/S1876-3804(18)30074-0.

Li, K.R., Song, J.M., Wang, H., et al., 2025. Spectral graph convolution networks for microbialite lithology identification based on conventional well logs. Pet. Sci. 22 (4), 1513–1533. https://doi.org/10.1016/j.petsci.2025.02.008.

Liu, Y.L., 2025. Evaluation of original oil in-place of fracture pore type of carbonate rock oil-gas reservoir in the Middle East. Geol. Resour. 34 (2), 210–216. https://doi.org/10.13686/j.cnki.dzyzy.2025.02.008 (in Chinese).

Liu, M.B., Liao, S.M., Yang, Y.F., et al., 2021. Tunnel boring machine vibration-based deep learning for the ground identification of working faces. J. Rock Mech. Geotech. Eng. 13 (6), 1340–1357. https://doi.org/10.1016/j.jrmge.2021.09.004.

Liu, H., Ren, Y.L., Li, X., et al., 2022. Rock thin-section analysis and identification based on artificial intelligent technique. Pet. Sci. 19 (4), 1605–1621. https://doi.org/10.1016/j.petsci.2022.03.011.

Lu, X.B., Wang, Y., Tian, F., et al., 2017. New insights into the carbonate karstic fault system and reservoir formation in the Southern Tahe area of the Tarim basin. Mar. Petrol. Geol. 86, 587–605. https://doi.org/10.1016/j.marpetgeo.2017.06.023.

Lundberg, S.M., Erion, G., Chen, H., et al., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2 (1), 56–67. https://doi.org/10.1038/s42256-019-0138-9.

Luque, A., Carrasco, A., Martín, A., et al., 2019. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recogn. 91, 216–231. https://doi.org/10.1016/j.patcog.2019.02.023.

Ma, Y.S., Li, M.W., Cai, X.Y., et al., 2020. Mechanisms and exploitation of deep marine petroleum accumulations in China: Advances, technological bottlenecks and basic scientific problems. Oil Gas Geol. 41 (4), 655–672. https://doi.org/10.11743/ogg20200401 (in Chinese).

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, California, pp. 281–297.

Mahdaviara, M., Rostami, A., Shahbazi, K., 2020. State-of-the-art modeling permeability of the heterogeneous carbonate oil reservoirs using robust computational approaches. Fuel 268, 117389. https://doi.org/10.1016/j.fuel.2020.117389.

Mondal, I., Singh, K.H., 2022. Core-log integration and application of machine learning technique for better reservoir characterisation of Eocene carbonates, Indian offshore. Energy Geosci 3 (1), 49–62. https://doi.org/10.1016/j.engeos.2021.10.006.

Nguyen, H., Vu, T., Vo, T.P., et al., 2021. Efficient machine learning models for prediction of concrete strengths. Constr. Build. Mater. 266, 120950. https://doi.org/10.1016/j.conbuildmat.2020.120950.

Ozkan, A., Cumella, S.P., Milliken, K.L., et al., 2011. Prediction of lithofacies and reservoir quality using well logs, Late Cretaceous Williams Fork Formation, Mamm Creek field, Piceance basin, Colorado. AAPG Bull. 95 (10), 1699–1723. https://doi.org/10.1306/01191109143.

Sahin, E.K., 2020. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. SN Appl. Sci. 2 (7), 1308. https://doi.org/10.1007/s42452-020-3060-1.

Salehi, S.M., Honarvar, B., 2014. Automatic identification of formation lithology from well log data: A machine learning approach. J. Petrol. Sci. Res. 3 (2), 73–82. https://doi.org/10.14355/jpsr.2014.0302.04.

Sen, S., Abioui, M., Ganguli, S.S., et al., 2021. Petrophysical heterogeneity of the Early Cretaceous Alamein dolomite reservoir from North Razzak oil field, Egypt integrating well logs, core measurements, and machine learning approach. Fuel 306, 121698. https://doi.org/10.1016/j.fuel.2021.121698.

Shen, X., Chen, Y.F., Cao, L.Q., et al., 2024. Prediction of the slurry pressure and inversion of formation characteristics based on a machine learning algorithm during tunnelling in a fault fracture zone. Tunn. Undergr. Space Technol. 144, 105514. https://doi.org/10.1016/j.tust.2023.105514.

Shi, J.X., Zhao, X.Y., Zeng, L.B., et al., 2024. Identification of reservoir types in deep carbonates based on mixed-kernel machine learning using geophysical logging data. Pet. Sci. 21 (3), 1632–1648. https://doi.org/10.1016/j.petsci.2023.12.016.

Soltanzadeh, P., Hashemzadeh, M., 2021. RCSMOTE: Range-controlled synthetic minority over-sampling technique for handling the class imbalance problem. Inf. Sci. 542, 92–111. https://doi.org/10.1016/j.ins.2020.07.014.

Song, T., Zhu, W., Chen, Z., et al., 2023. A novel well-logging data generation model integrated with random forests and adaptive domain clustering algorithms. Geoenergy Sci. Eng. 231, 212381. https://doi.org/10.1016/j.geoen.2023.212381.

Tang, H., Tan, X.C., Liu, H., et al., 2014. Genesis and dolomitization of "Khali" powder crystal dolomite in Triassic Jialingjiang formation, Moxi gas field, central Sichuan basin, SW China. Petrol. Explor. Dev. 41 (4), 553–562. https://doi.org/10.1016/S1876-3804(14)60066-5.

Teillet, T., Fournier, F., Zhao, L., et al., 2021. Geophysical pore type inversion in carbonate reservoir: Integration of cores, well logs, and seismic data (Yadana field, offshore Myanmar). Geophysics 86 (3), B149–B164. https://doi.org/10.1190/geo2020-0486.1.

Thabtah, F., Hammoud, S., Kamalov, F., et al., 2020. Data imbalance in classification: Experimental evaluation. Inf. Sci. 513, 429–441. https://doi.org/10.1016/j.ins.2019.11.004.

Tian, F., Luo, X.L., Zhang, W., 2019. Integrated geological-geophysical characterizations of deeply buried fractured-vuggy carbonate reservoirs in Ordovician strata, Tarim Basin. Mar. Petrol. Geol. 99, 292–309. https://doi.org/10.1016/j.marpetgeo.2018.10.028.

Verma, A.K., Routray, A., Mohanty, W.K., 2014. Assessment of similarity between well logs using synchronization measures. IEEE Geosci. Remote Sens. Lett. 11 (12), 2032–2036. https://doi.org/10.1109/LGRS.2014.2317498.

Wang, J., Cao, J., Fu, J., et al., 2022. Missing well logs prediction using deep learning integrated neural network with the self-attention mechanism. Energy 261, 125270. https://doi.org/10.1016/j.energy.2022.125270.

Wen, H.G., Luo, B., Zhong, K.X., et al., 2023. Discovery of Linshui-Fengdu intra-platform depression of Triassic Jialingjiang Formation in eastern Sichuan Basin and its significance for oil and gas exploration. Lit. Res. 35 (1), 12–24. https://doi.org/10.12108/yxyqc.20230102 (in Chinese).

Wu, L.Y., Li, J.H., Zhang, J.W., et al., 2024. Prediction model for the compressive strength of rock based on stacking ensemble learning and Shapley Additive exPlanations. Bull. Eng. Geol. Environ. 83 (11), 439. https://doi.org/10.1007/s10064-024-03896-3.

Xu, G.S., Liu, S.G., Yuan, H.F., et al., 2005. Dynamics of natural gas pools of Carboniferous in the eastern Sichuan Basin. Acta Pet. Sin. 26 (4), 12–16. https://doi.org/10.3321/j.issn:0253-2697.2005.04.003 (in Chinese).

Xu, X.L., Chen, W., Sun, Y.F., 2019. Over-sampling algorithm for imbalanced data classification. J. Syst. Eng. Electron. 30 (6), 1182–1191. https://doi.org/10.21629/JSEE.2019.06.12.

Xue, J.K., Shen, B., 2023. Dung beetle optimizer: A new meta-heuristic algorithm for global optimization. J. Supercomput. 79 (7), 7305–7336. https://doi.org/10.1007/s11227-022-04959-6.

Yang, D.S., Xie, J.W., Yin, Z.Y., 2023. Improved *K*-means algorithm for fault diagnosis of vehicle. In: 2023 9$^{th}$ International Conference on Computer and Communications (ICCC). IEEE, pp. 1536–1540. https://doi.org/10.1109/ICCC59590.2023.10507334.

Yu, H.G., Tao, J.F., Qin, C.J., et al., 2022. A novel constrained dense convolutional autoencoder and DNN-based semi-supervised method for shield machine tunnel geological formation recognition. Mech. Syst. Signal Process. 165, 108353. https://doi.org/10.1016/j.ymssp.2021.108353.

Zaremba, W., Sutskever, I., Vinyals, O., 2014. Recurrent neural network regularization. https://doi.org/10.48550/arXiv.1409.2329 arXiv:1409.2329 [cs.NE].

Zeng, Y., Wei, Y.J., Yang, Y.Y., 2024. A novel identification technology and real-time classification forecasting model based on hybrid machine learning methods in mixed weathered mudstone-sand-pebble formation. Tunn. Undergr. Space Technol. 153, 106045. https://doi.org/10.1016/j.tust.2024.106045.

Zeng, Y., Wei, Y.J., Yang, Y.Y., et al., 2025. A novel real-time classification method of mixed weathered mudstone-sand-pebble formation ahead of an EPBS using DBOKM-SMOTE and stacking ensemble learning. Tunn. Undergr. Space Technol. 165, 106870. https://doi.org/10.1016/j.tust.2025.106870.

Zhang, J., Liu, S., Li, J., et al., 2017. Identification of sedimentary facies with well logs: An indirect approach with multinomial logistic regression and artificial neural network. Arabian J. Geosci. 10 (11), 247. https://doi.org/10.1007/s12517-017-3045-6.

Zhang, S.C., Li, X.L., Zong, M., et al., 2018. Efficient kNN classification with different numbers of nearest neighbors. IEEE Transact. Neural Networks Learn. Syst. 29 (5), 1774–1785. https://doi.org/10.1109/TNNLS.2017.2673241.

Zhang, Q.L., Liu, Z.Y., Tan, J.L., 2019. Prediction of geological conditions for a tunnel boring machine using big operational data. Autom. ConStruct. 100, 73–83. https://doi.org/10.1016/j.autcon.2018.12.022.

Zhang, Y.Y., Xi, K.L., Cao, Y.C., et al., 2021. The application of machine learning under supervision in identification of shale lamina combination types: A case study of Chang $7_3$ sub-member organic-rich shales in the Triassic Yanchang Formation, Ordos Basin, NW China. Pet. Sci. 18 (6), 1619–1629. https://doi.org/10.1016/j.petsci.2021.09.033.

Zhao, J.H., Shi, M.L., Hu, G., et al., 2019. A data-driven framework for tunnel geological-type prediction based on TBM operating data. IEEE Access 7, 66703–66713. https://doi.org/10.1109/ACCESS.2019.2917756.

Zheng, W.H., Tian, F., Di, Q.Y., et al., 2021. Electrofacies classification of deeply buried carbonate strata using machine learning methods: A case study on Ordovician paleokarst reservoirs in Tarim Basin. Mar. Petrol. Geol. 123, 104720. https://doi.org/10.1016/j.marpetgeo.2020.104720.

Zheng, D., Hou, M., Chen, A., et al., 2022. Application of machine learning in the identification of fluvial-lacustrine lithofacies from well logs: A case study from Sichuan Basin, China. J. Petrol. Sci. Eng. 215, 110610. https://doi.org/10.1016/j.petrol.2022.110610.

Zhou, C., Gao, Y.Y., Chen, E.J., et al., 2023. Deep learning technologies for shield tunneling: Challenges and opportunities. Autom. ConStruct. 154, 104982. https://doi.org/10.1016/j.autcon.2023.104982.

Zhou, J.W., Chen, J.L., Li, H.B., 2024. An optimized fuzzy *K*-means clustering method for automated rock discontinuities extraction from point clouds. Int. J. Rock Mech. Min. Sci. 173, 105627. https://doi.org/10.1016/j.ijrmms.2023.105627.

Zhu, G.Y., Milkov, A.V., Zhang, Z.Y., et al., 2019. Formation and preservation of a giant petroleum accumulation in superdeep carbonate reservoirs in the southern Halahatang oil field area, Tarim Basin, China. AAPG Bull. 103 (7), 1703–1743. https://doi.org/10.1306/11211817132.

Zhu, X., Zhang, H., Ren, Q., et al., 2023. An automatic identification method of imbalanced lithology based on Deep Forest and *K*-means SMOTE. Geoenergy Sci. Eng. 224, 211595. https://doi.org/10.1016/j.geoen.2023.211595.