Petroleum Science 23 (2026) 3017–3036

Contents lists available at ScienceDirect

Petroleum Science

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

Original Paper

A data augmentation method for lacustrine shale lithofacies classification based on a conditional diffusion probabilistic model: A case study from the Dongying Depression, Bohai Bay Basin, China

Gui-Ang Li [a,b,c], Cheng-Yan Lin [a,b,c,*], Chun-Mei Dong [a,b,c,**], Li-Hua Ren [a,b,c,***], Peng-Jie Ma [d], Yu-Qi Wu [a,b,c], Guo-Yin Zhang [a,b,c], Xin-Yu Du [e], Zi-Ru Zhao [a,b,c]

[a] State Key Laboratory of Deep Oil and Gas, China University of Petroleum (East China), Qingdao, 266580, Shandong, China
[b] School of Geosciences, China University of Petroleum (East China), Qingdao, 266580, Shandong, China
[c] Shandong Provincial Key Laboratory of Reservoir Geology, China University of Petroleum (East China), Qingdao, 266580, Shandong, China
[d] Sanya Offshore Oil & Gas Research Institute, Northeast Petroleum University, Sanya, 572025, Hainan, China
[e] Wuxi Research Institute of Petroleum Geology, Petroleum Exploration and Production Research Institute, SINOPEC, Wuxi, 214126, Jiangsu, China

ARTICLE INFO

Article history:
Received 17 June 2025
Received in revised form
6 November 2025
Accepted 27 January 2026
Available online 7 February 2026

Edited by Xi Zhang and Jie Hao

Keywords:
Lacustrine shale
Lithofacies prediction
Conditional diffusion probabilistic model
Data augmentation
Class imbalance

ABSTRACT

Accurate identification of lithofacies is critical for shale hydrocarbon exploration and development. Although machine learning (ML) is one of the most effective approaches for predicting shale lithofacies, the inherent geological heterogeneity results in scarce training samples and severely imbalanced class distributions, leading conventional ML methods to experience overfitting and reduced accuracy. To address this issue, we introduced a conditional diffusion probabilistic model (CDPM) to address these challenges and developed a comprehensive data augmentation framework for shale lithofacies prediction. Applying this framework to the Upper Fourth Member of the Shahejie Formation (Es4s) in the Dongying Depression, we successfully generated 3,600 class-balanced augmented samples from 926 core-calibrated samples and eight conventional well log curves from well NY1, achieving an 878.3% increase in rare organic-rich fissile calcareous mudstone (L1) lithofacies. To ensure the reliability of the augmented data, a comprehensive quality assessment was conducted, demonstrating that augmented data effectively retained logging characteristics and petrophysical relationships, with a Fréchet Inception Distance (FID) of 22.9 and a maximum mean discrepancy (MMD) of 0.078. Building upon this high-quality augmented dataset, we evaluated the impact of data augmentation on lithofacies classification performance across random forest (RF), support vector machine (SVM), and gradient boosted decision tree (GBDT) algorithms. The results showed substantial improvements, with average accuracy and F1 score increases of 13.6% and 16.5%, respectively, and a 33.1% improvement in L1 recall. To further validate the practical applicability of our approach, blind-well validation on independent wells from different structural positions demonstrated robust generalization capability, achieving significant improvement over traditional ML methods. This study pioneers a conditional diffusion model for predicting shale lithofacies, providing a novel framework for characterising lacustrine shale oil reservoirs and predicting sweet spots.

1. Introduction

Shale lithofacies are critical hosts of unconventional hydrocarbon resources, directly influencing reservoir properties and development potential (He et al., 2023; Sun et al., 2024). However, shale lithofacies exhibit significant complexity in both mineralogical composition and spatial distribution, driven by multiple geological processes, including paleoclimatic fluctuations, redox

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail addresses: lincy@upc.edu.cn (C.-Y. Lin), lcydcm@upc.edu.cn (C.-M. Dong), renlh@upc.edu.cn (L.-H. Ren).

Peer review under the responsibility of China University of Petroleum (Beijing).

https://doi.org/10.1016/j.petsci.2026.01.042

variations, salinity changes, and terrigenous input (Liu et al., 2021; Wang et al., 2024). Such complexity obscures characteristic logging responses and creates a pronounced class imbalance in lithofacies datasets, thereby limiting prediction accuracy (Dong et al., 2022; Ellis and Singer, 2007).

ML methods have been extensively applied to shale lithofacies classification, demonstrating considerable effectiveness in capturing complex nonlinear relationships between well-log responses and lithofacies types. RF, as one of the most widely adopted algorithms, has shown robust performance in handling multi-dimensional logging data through ensemble learning strategies (Wang et al., 2023). SVM and artificial neural networks have also been successfully employed to identify lithofacies from conventional logs by constructing optimal decision boundaries in high-dimensional feature spaces (Wang and Carr, 2012). Recent advances in gradient boosting methods, such as XGBoost, have further enhanced classification accuracy (Wang et al., 2023). However, as discriminative models, these ML approaches fundamentally rely on learning decision boundaries from existing labeled samples rather than understanding the underlying data generation mechanisms. This characteristic renders them inherently vulnerable to issues with training data quality and distribution, mainly when geological sampling constraints result in insufficient or imbalanced datasets.

Obtaining high-quality labeled lithofacies datasets poses two fundamental challenges in field applications. First, the prohibitive costs of core drilling and laboratory analysis severely constrain sample acquisition (Geng et al., 2023; Xue et al., 2024). Second, the natural heterogeneity of lithofacies distribution creates an extreme class imbalance, with dominant lithofacies often outnumbering minor types by factors of several or more (Li et al., 2024a; Liang et al., 2018a). Recent advances in data-centric artificial intelligence paradigms have fundamentally reoriented research priorities from model-centric optimization to systematic data engineering, emphasizing that high-quality, representative training data often yields greater performance gains than architectural sophistication in data-constrained scenarios (Kuang et al., 2021; Zhao et al., 2024). The growing recognition of these data-centric challenges has catalyzed a paradigm shift in petroleum geoscience toward systematic data engineering approaches that prioritize training data quality, representativeness, and balance alongside model architecture optimization (Li et al., 2025). Recent studies in shale characterization have demonstrated that sophisticated data augmentation strategies can yield more substantial performance improvements than incremental model refinements when working with limited labeled samples (Wang et al., 2025; Xue et al., 2024). These combined factors—limited data availability and skewed distributions—cause traditional ML models to overfit the majority classes while failing to adequately capture the characteristics of minority classes. This bias negatively impacts model generalization and predictive performance (Dong et al., 2023; Liu et al., 2023). Researchers have investigated various strategies to address sample imbalance (Chawla et al., 2002; Engelmann and Lessmann, 2021). Traditional methods, such as oversampling, undersampling, and the synthetic minority oversampling technique (SMOTE), only provide limited relief for class bias, particularly when dealing with the high-dimensional, nonlinear nature of logging data (Barandela et al., 2004; Santoso et al., 2017). Generative models, especially generative adversarial networks (GANs), offer promising avenues for data augmentation. Nevertheless, training instability and mode collapse—persistent issues in geological GAN applications—significantly undermine the quality and diversity of augmented samples (Qian et al., 2024; Zha et al., 2020; Zhao et al., 2023).

Diffusion probabilistic models have recently emerged as powerful generative frameworks, demonstrating superior performance across diverse applications. These models simulate data generation through iterative forward diffusion processes that gradually transform real samples into Gaussian noise, followed by a reverse reconstruction that learns to generate high-fidelity, structurally coherent samples (Amit et al., 2021; Zhao et al., 2025). Unlike GANs, diffusion models exhibit enhanced training stability, improved sample diversity, and more robust conditional control mechanisms (Müller-Franzes et al., 2023). Conditional diffusion probabilistic models (CDPMs) further advance this framework by incorporating categorical or feature-based guidance mechanisms that enable targeted data augmentation for specific characteristics (Trabucco et al., 2023). While initially developed for continuous data modalities such as images and audio, recent methodological advances in applying diffusion models to tabular data have demonstrated their capability to handle heterogeneous features comprising both continuous and discrete variables (Li et al., 2025)—a critical methodological foundation for geoscience applications where well-log datasets exhibit similar mixed-type characteristics. The application of these tabular diffusion techniques to subsurface energy domains holds particular promise, as the high costs of wellbore data acquisition and the prevalence of class-imbalanced lithofacies distributions amplify the strategic value of generative augmentation methods in reservoir characterization workflows. Inspired by these developments in tabular data modeling (Kotelnikov et al., 2023; Villaizán-Vallelado et al., 2025), exploratory studies have begun investigating diffusion-based approaches for geological applications, including digital rock reconstruction (Lu et al., 2025), well-log imputation (Meng et al., 2024), and facies geomodeling (Li et al., 2024b; Xu et al., 2024). The application of these diffusion-based approaches to class-imbalanced shale lithofacies classification remains unexplored.

The study addresses this gap by developing a CDPM-based approach for well logging data augmentation, using the NY1 well from the Dongying Depression in the Bohai Bay Basin as a case study. We construct a comprehensive feature system based on eight conventional logging curves (GR, SP, CAL, AC, RT, DEN, CNL, and COND) and introduce wells N55-X1 and NY1-6HF from different structural positions as independent validation datasets to assess model generalization under limited sample conditions. Our primary contributions include: (1) Developing a CDPM architecture designed explicitly for well-logging data, integrating multilayer neural networks with adaptive noise scheduling mechanisms to effectively model complex nonlinear relationships among logging curves and generate high-fidelity augmented samples. (2) Conducting a systematic analysis of key hyperparameters, including diffusion steps, the number of augmented samples, and optimization strategies to enhance performance. This work provides an innovative technical pathway for identifying shale lithofacies. It expands the application of diffusion models in geoscience, offering a practical solution to the everyday challenges of small-sample class imbalance in geological datasets.

2. Geological background and lithofacies classification

2.1. Geological background

The Dongying Depression, located in the southeastern portion of the Jiyang Depression within the Bohai Bay Basin, constitutes a northeast-southwest trending fault-block sedimentary unit covering approximately 5,700 km^2 (Liang et al., 2018b; Shi et al., 2022). This region exhibits well-defined structural boundaries:

the northern margin is bordered by the Chenjiazhuang Uplift, the southern boundary is adjacent to the Luxi Uplift and the Qingcheng Uplift, the Lijin Fault controls the western edge, and the Qingtuozi Uplift defines the eastern limit. The overall structure exhibits a characteristic "uplift-depression" configuration, comprising four secondary depressions: Bozhong, Niuzhuang, Lijin, and Minfeng (Fig. 1(a)) (Zhang et al., 2016). The Paleogene Shahejie Formation serves as the primary hydrocarbon-bearing sequence in this area. The Es4s interval was deposited in deep lacustrine environments, resulting in the development of thick, organic-rich shales with diverse lithofacies assemblages (Liang et al., 2017) (Fig. 1(b)). Multiple factors, including paleoclimatic fluctuations, variations in water depth, salinity conditions, and redox environments, have created significant heterogeneity in shale lithofacies within this formation, providing an ideal geological target for lithofacies identification and data augmentation studies.

The NY1 well is situated in the gentle slope zone of the Niuzhuang Depression, within a structural-sedimentary transition zone near the lake basin center (Fig. 1(c)). High-quality continuous coring was conducted in the Es4s interval of this well, effectively capturing the primary lithofacies types in the region and providing highly representative geological samples for lithofacies classification modeling (Fang et al., 2023; Qiao et al., 2025). To comprehensively assess the method's generalization performance, we incorporated wells N55-X1 and NY1-6HF as independent test datasets. Well N55-X1, located in the gentle slope zone near the lake basin center, has continuous core coverage within the validation interval. Well NY1-6HF, situated in the steep slope zone, provides supplementary validation with 21 calibrated core samples. Both wells, located within the same lacustrine depositional system, enable comprehensive validation of the applicability and reliability of data augmentation methods across varying structural settings in shale oil exploration.

2.2. Lithofacies classification

The Es4s interval in the Dongying Depression exhibits significant vertical and lateral lithological heterogeneity, characterized by complex lithofacies assemblages and frequent spatial variations. Building upon previous research findings and systematic lithofacies classification frameworks established for lacustrine organic-rich mudstone (Liu et al., 2019), we employed multiple analytical approaches—including core observation, thin-section analysis, X-ray diffraction (XRD) testing, and total organic carbon (TOC) content determination—to systematically classify lithofacies (Li et al., 2022; Zhao et al., 2022). The classification integrated three key parameters: mineral compositions (three end-members of clay minerals, carbonates, and felsic minerals comprising quartz and feldspar), TOC contents (<2.0% as organic-poor, 2.0%–3.5% as intermediate-organic, and >3.5% as organic-rich), and sedimentary structure characteristics (Liang et al., 2018a; Ma et al., 2022a, 2022b). Considering the overlapping and transitional characteristics of lithofacies in their natural distribution, four primary lithofacies types were identified in the study area (Fig. 2).

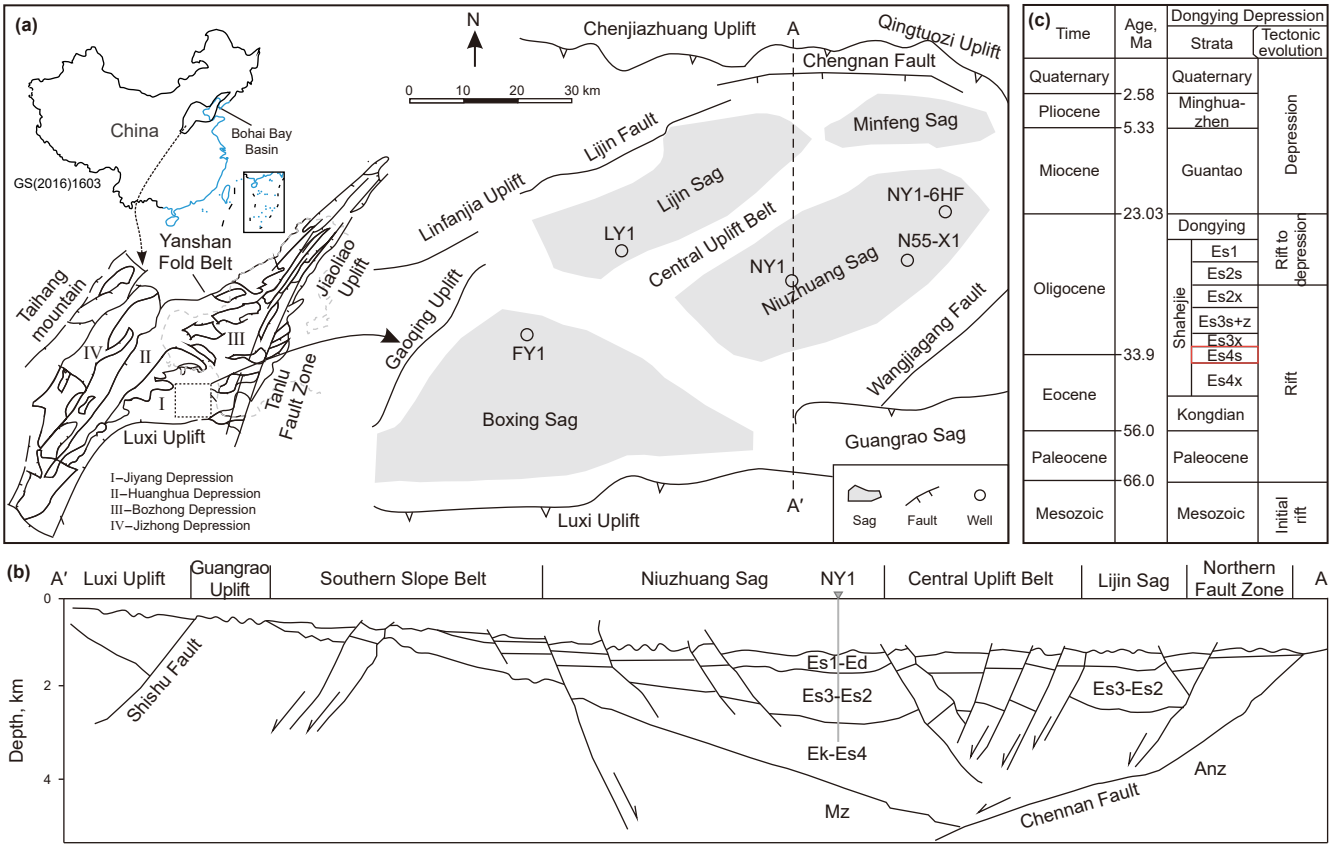

Fig. 1. Regional geological setting and stratigraphy of the study area. **(a)** Structural map of the Dongying Depression within the Bohai Bay Basin, illustrating major tectonic features and the distribution of representative shale oil wells (modified from Ma et al., 2022a, 2022b); **(b)** Geological cross-section showing the internal structural configuration of the Dongying Depression, with the section line indicated in panel A (adapted from Liang et al., 2018b); **(c)** Simplified stratigraphic column of the study area, illustrating the Paleogene succession, including the Kongdian Formation (Ek), the Shahejie Formation members (Es4 to Es1), and the overlying Dongying Formation (Ed).

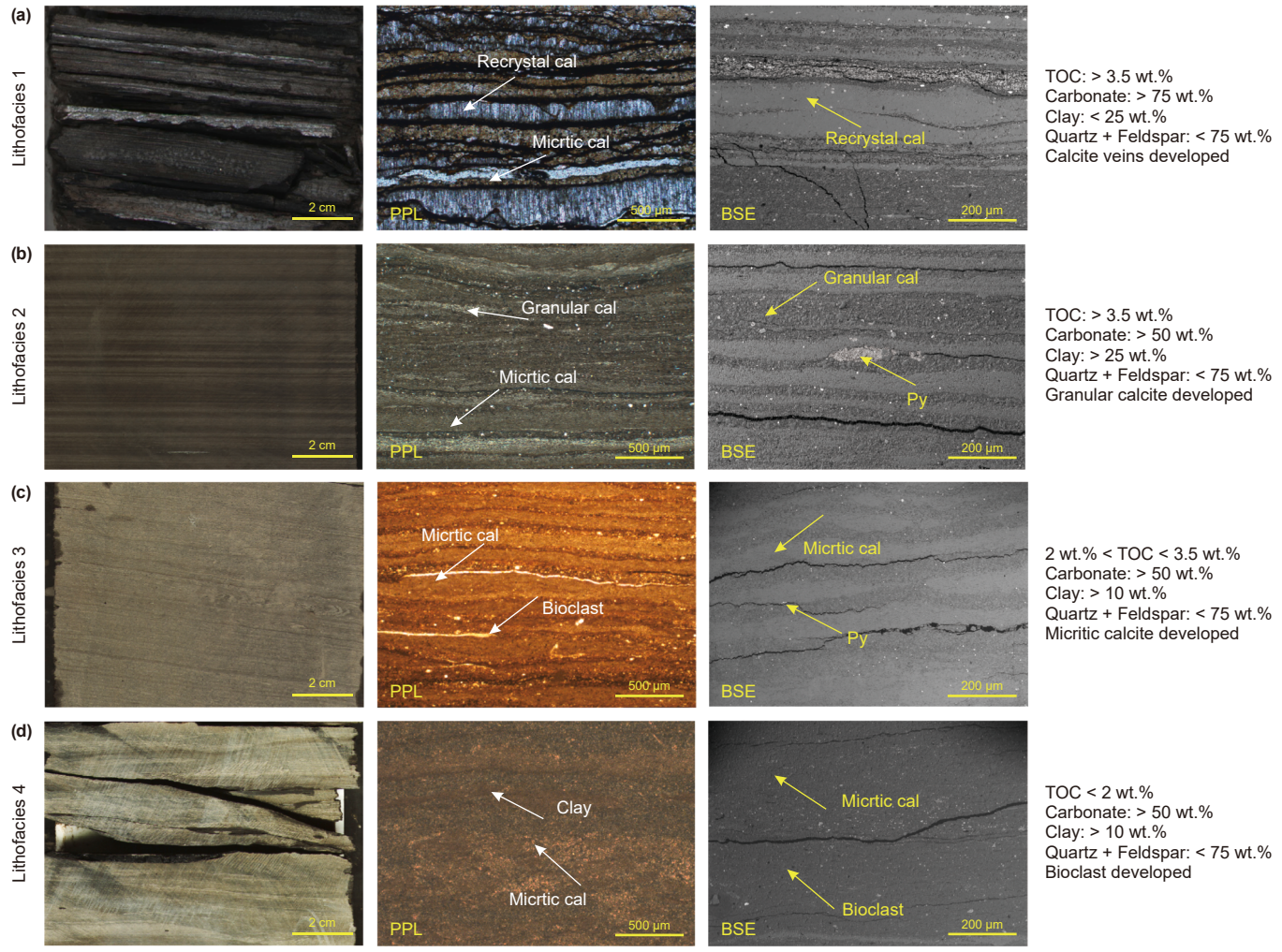

Fig. 2. Lithofacies within the Es4s shale (Li et al., 2022). Core image, plane-polarized light (PPL) transmitted light image, and backscattered electron (BSE) scanning electron microscopy images show examples of each lithofacies. **(a)** Organic-rich fissile calcareous mudstone. **(b)** Organic-rich laminated calcareous mudstone. **(c)** Intermediate-organic laminated calcareous mudstone. **(d)** Organic-poor calcareous mudstone. Abbreviations: cal, calcite; py, pyrite.

2.2.1. Organic-rich fissile calcareous mudstone (L1)

TOC content exceeds 3.5%, carbonate content typically above 75%, with well-developed calcite veins. Cores exhibit gray-black coloration with distinct lamination and abundant framboidal pyrite. This lithofacies exhibits a pronounced shaly texture with well-developed fissility, characterized by easy delamination along calcite vein planes and representing the highest potential for hydrocarbon generation (Liang et al., 2018a, 2023; Wu and Jiang, 2017). L1 formed under semi-deep, strongly reducing conditions with highly concentrated organic matter.

2.2.2. Organic-rich laminated calcareous mudstone (L2)

TOC content exceeds 3.5%, and carbonate content is below 60%. Characterized by granular calcite laminae development and, relative to other lithofacies, elevated quartz, feldspar, and clay mineral contents, with clay minerals reaching 30%. Organic matter laminae are highly concentrated, commonly intercalated with calcite laminae and clay layers, and display sharp lamination boundaries, as well as widespread framboidal pyrite. L2 was deposited under deep-water reducing conditions and exhibits slightly lower development potential, despite its favorable characteristics for organic matter enrichment compared to L1.

2.2.3. Intermediate-organic laminated calcareous mudstone (L3)

TOC content ranges from 2% to 3.5%, and carbonate content exceeds 65%. Well-developed, undulatory micritic calcite laminae indicate weak redox environmental conditions. This lithofacies is widely distributed throughout the study area, constituting the dominant lithofacies type regionally.

2.2.4. Organic-poor calcareous mudstone (L4)

TOC content below 2%, carbonate mineral content typically exceeds 50%. Dominated by micritic calcite with occasional bioclastic and micritic dolomite components. Lamination development is poor, with indistinct interfaces and a relatively uniform overall structure, indicating oxidizing depositional conditions.

The distribution of these four lithofacies categories exhibits an extreme imbalance across the study area, revealing distinct patterns. L3 lithofacies exhibit the most extensive distribution, representing the predominant lithofacies type in the region. In contrast, the organic-rich L1 lithofacies exhibit a relatively restricted distribution, primarily forming under specific, strongly reducing depositional environments. This imbalanced distribution pattern provides crucial geological constraints for subsequent ML model construction and performance evaluation.

3. Methods

3.1. Dataset preparation and lithofacies labeling

The study developed a CDPM-based framework for augmenting and evaluating lithofacies data. The workflow comprised four key stages (Fig. 3): (1) data acquisition and preprocessing to construct a lithofacies-labeled well log dataset; (2) conditional diffusion model training to learn deep characterization relationships between well log responses and lithofacies categories; (3) augmented data generation and quality assessment to achieve balanced category distribution; and (4) classification model training and generalization performance assessment to verify practical applicability.

For well log curve selection, the study chose eight conventional logging curves as input features, considering the high cost and limited applicability of specialized logging techniques (Zhao et al., 2019). These included GR, SP, CAL, AC, RT, DEN, CNL, and COND. These logging parameters are closely tied to the characteristics of shale lithofacies and provide a comprehensive reflection of the rocks' physical, electrical, acoustic, and mineralogical properties (Dong et al., 2022; Lu et al., 2023). Previous studies have demonstrated that this combination of conventional well logs effectively captures the primary differences in shale lithofacies characteristics, providing a comprehensive and robust foundation for lithofacies identification (Li et al., 2022; Qiao et al., 2025).

In terms of core-log depth matching, this study adopted the systematic depth matching methodology proposed by Larson et al. (2023). First, a baseline depth reference system was established using core box top depths and 5-cm marker positions. Subsequently, XRF element data were used to correct depth deviations in the fractured zone and sections with incomplete core recovery. Finally, the precise depth was determined by integrating log response characteristics. Building upon this approach, the study innovatively implemented a multi-parameter comparative method incorporating correlation analysis between core porosity and AC curves, consistency testing between core density and DEN curves, and high-resolution XRF scanning data to enable accurate depth offset determination. This method simultaneously accounted for cable stretching effects and wellbore deviation influences. Through multi-source data integration, it fully utilized multidimensional rock physical property information, significantly enhancing depth-matching accuracy and reliability.

During the data integration stage, one-dimensional linear interpolation technology is used to resample well-log data (with a 15-cm sampling interval) to corresponding core sampling point depths, and edge effects and outliers (such as "−999" missing data flags) are rigorously handled. The sample construction process

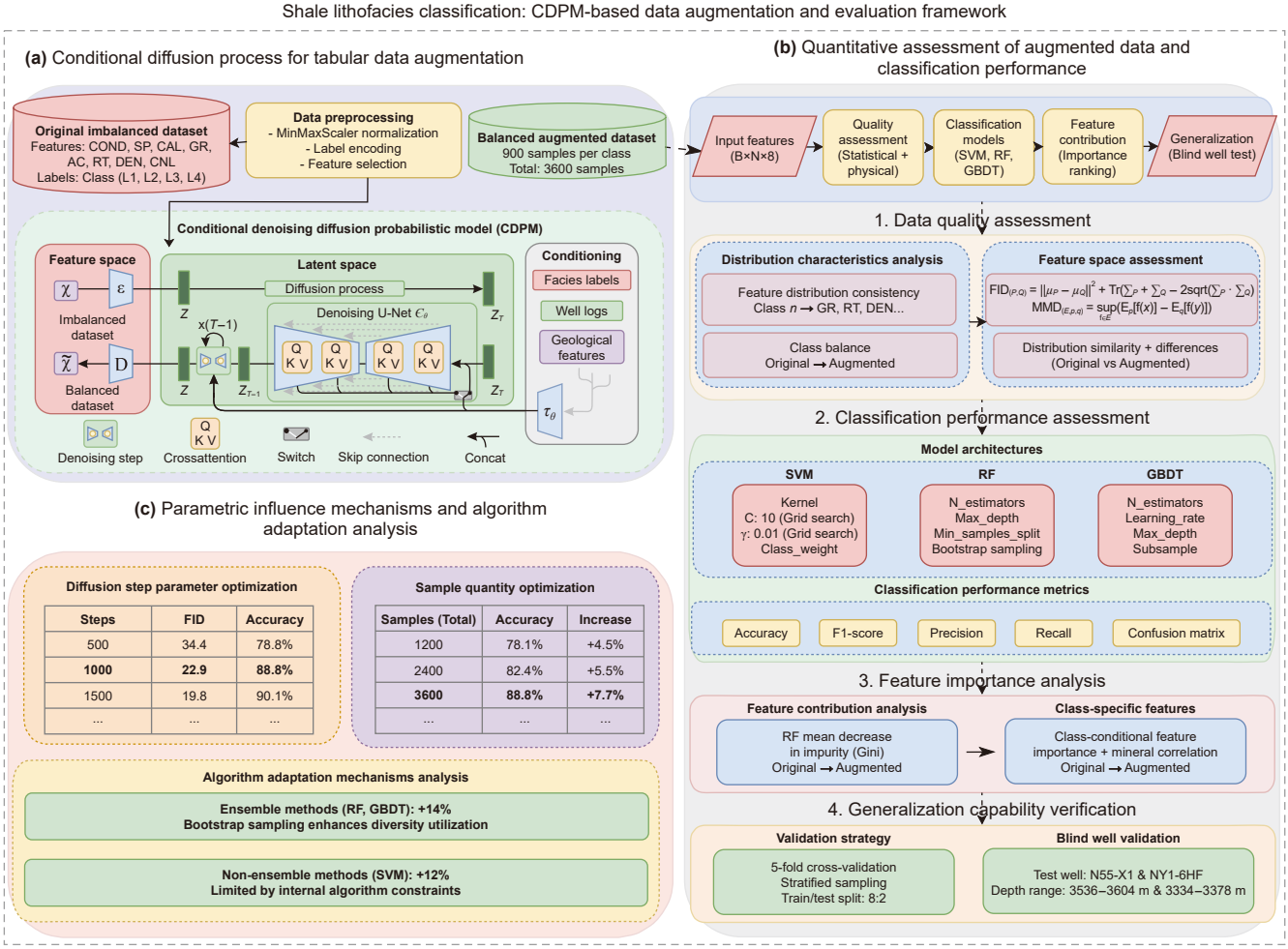

Fig. 3. Technical workflow of the CDPM-based data augmentation framework for shale lithofacies classification. The framework consists of three modules: **(a)** a conditional diffusion process for imbalanced data augmentation, **(b)** a quantitative assessment of augmented data quality and classification performance, and **(c)** a parametric influence analysis and model validation strategies.

followed three strict quality control guidelines: (1) prioritize relatively homogeneous lithofacies intervals to ensure that the spacing between adjacent sample points is ≥15 cm to minimize spatial correlation of well logging responses; (2) extract average value within 15 cm windows before and after each sample point to smooth random noise interference effectively; (3) combine macroscopic core features with microscopic thin-section analysis results to validate lithofacies label accuracy rigorously.

The data preprocessing workflow included outlier identification and removal, Z-score standardization, and Min–Max normalization (normalizing value ranges to the [0, 1] interval) to ensure balanced contributions of different logging features to model training. After rigorous selection and quality control procedures, we constructed a standardized dataset containing 926 high-quality samples. To avoid selection bias during model training, a stratified sampling strategy was employed to divide the dataset into training and testing datasets at an 8:2 ratio. This approach ensures a consistent distribution of lithofacies types across both subsets and provides a solid foundation for subsequent CDPM training and validation. It should be noted that despite rigorous quality control measures, certain subjective factors inherent to lithofacies labeling workflows may introduce uncertainties. Core-log depth matching, while employing systematic methodologies, involves interpretative decisions during core-log alignment, particularly in fractured zones or intervals with incomplete recovery (Chang et al., 2002). Even centimeter-scale depth offset uncertainties can introduce spatial misalignment between lithofacies labels and corresponding log responses. Additionally, lithofacies definition represents another source of subjectivity given the transitional and overlapping characteristics of shale lithofacies in natural settings (Wang et al., 2025; Xue et al., 2024). The inherent gradational boundaries between lithofacies types can lead to inconsistent labeling for samples exhibiting intermediate characteristics, potentially affecting the quality and generalization capability of models (Bressan et al., 2020; Li et al., 2024a). The aforementioned multi-parameter depth calibration, quantitative classification criteria, and preferential sampling strategies were specifically designed to minimize these subjective effects and enhance label consistency. Furthermore, the CDPM-based data augmentation approach inherently provides additional robustness against label uncertainty by preserving statistical characteristics while diluting the proportional influence of potentially mislabeled samples through balanced augmentation (Section 4.1). The multi-well blind validation results (Section 4.2.3) demonstrate that models trained on augmented datasets maintain strong generalization capability despite potential labeling uncertainties.

3.2. CDPM-based data augmentation method

CDPMs represent generative models that have achieved breakthrough performance in computer vision and natural language processing. Their theoretical foundations are rooted in stochastic diffusion processes from non-equilibrium statistical physics (Amit et al., 2021; Zhao et al., 2025). CDPMs were introduced to the geological well log data augmentation domain for the first time to address challenges of category imbalance in shale lithofacies classification. The conditional control mechanism enables targeted sample augmentation for specific lithofacies, effectively resolving category imbalance issues in shale lithofacies classification. The key principle of conditional diffusion models involves constructing a reversible bidirectional diffusion process (Lu et al., 2025; Trabucco et al., 2023). During forward diffusion, original logging data transforms into a pure noise distribution through gradual Gaussian noise injection:

$$q(x_t|x_0) = N\left(x_t; \sqrt{\bar{\alpha}_t}x_0, \left(1-\bar{\alpha}_t\right)I\right) \tag{1}$$

where N denotes a Gaussian distribution, x_t represents the noisy data at diffusion step t, x_0 is the original well-log data, I is the identity matrix, and $\bar{\alpha}_t$ is the cumulative noise coefficient defined as:

$$\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i) \tag{2}$$

where β_t represent preset noise scheduling parameters that control the amount of noise added at each step. Correspondingly, the reverse process learns denoising functions through deep neural networks and reconstructs data under lithofacie conditions on y with guidance, following the probability distribution:

$$p_\theta(x_{t-1}|x_t, y) = N\left(x_{t-1}; \mu_\theta(x_t, t, y), \sigma_t^2 I\right) \tag{3}$$

where θ denotes the neural network parameters, y represents the lithofacies condition label, μ_θ is the predicted mean, and σ_t is the standard deviation at step t. The mean function μ_θ is computed as:

$$\mu_\theta(x_t, t, y) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t, y)\right) \tag{4}$$

where ϵ_θ denotes the neural network-predicted noise term, this progressive augmentation strategy offers several advantages: (1) the variational inference framework ensures model training stability; (2) conditional control mechanisms enable augmented samples to match target lithofacies geological characteristics accurately; (3) multi-step diffusion processes achieve fine-grained control over data augmentation details, significantly enhancing augmented data quality.

The conditional diffusion network is designed to comprise three synergistic key modules based on the properties of logging data. The feature encoder employs multi-layer fully connected networks (dimensionality mapping: 8 → 512 →1024 → 512) for high-dimensional feature mapping of logging data, effectively preserving physical correlations between logging curves. The time embedding module converts discrete diffusion time steps into 512-dimensional continuous vectors through sinusoidal positional encoding, accurately capturing temporal dependencies in noise evolution processes. The lithofacies condition control module first maps discrete lithofacies labels to 64-dimensional continuous vectors, then expands them to 512 dimensions through fully connected layers and integrates them deeply with logging features to ensure precise category control during augmented sample generation. Network training employed noise prediction loss functions $L = E\left(\|\epsilon - \epsilon_\theta(x_t, t, y)\|^2\right)$, with parameter optimization achieved through AdamW optimizers (learning rate 1×10^{-4}, weight decay 0.01) combined with cosine annealing learning rate scheduling strategies.

Key model hyperparameter configurations were determined through systematic experiments and five-fold cross-validation, using the following settings: 1000 diffusion steps, a batch size of 64, a hidden layer dimension of 1024, and 900 augmented samples per class. Notably, diffusion steps determine the granularity of the

noise conversion process, while augmented sample quantities significantly influence achieving category balance. The specific impact mechanisms of these key parameters on model performance are discussed in Sections 5.1 and 5.2.

A multi-dimensional evaluation framework was established to comprehensively assess augmented data quality and effectiveness, encompassing three analytical approaches. For feature distribution consistency assessment, kernel density estimation (KDE) and well-log curve cross-plot analyses were employed to quantitatively examine distribution similarity between augmented and original data (Ashraf et al., 2024; Geng et al., 2023). For assessing spatial structure consistency, principal component analysis (PCA) evaluated the preservation of global linear structure, while t-distributed stochastic neighbor embedding (t-SNE) visualized nonlinear local structural features (Meyer et al., 2022; Zhou et al., 2018). At the quantitative assessment level, Fréchet inception distance (FID) measured overall distribution differences, maximum mean discrepancy (MMD) verified statistical moment consistency, and intra-class/inter-class variance ratios assessed separability between different lithofacies (Jayasumana et al., 2024; Yu et al., 2021). This comprehensive evaluation system thoroughly verifies the augmented data quality characteristics and geological consistency from multiple perspectives, ensuring that the generated augmented samples effectively improve the performance of the lithofacies classification model.

3.3. Classification model validation and performance evaluation

To systematically evaluate the impact of CDPM-augmented data on lithofacies classification performance, a multi-model validation framework was established using three well-known ML algorithms commonly employed for shale lithofacies identification: SVM, RF, and GBDT. These algorithms demonstrate strong predictive capabilities, are grounded in distinct theoretical foundations, and collectively offer comprehensive insights into the performance of augmented data across multiple dimensions (Dong et al., 2023; Tang et al., 2021).

SVM transforms original data into high-dimensional feature spaces via kernel functions to establish optimal hyperplanes for classification. The radial basis function (RBF) kernel was implemented, defined as:

$$K\left(x_i, x_j\right) = \exp\left(-\gamma \parallel x_i - x_j\parallel^2\right) \tag{5}$$

where γ represents the bandwidth parameter.

The corresponding decision function is:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b\right) \tag{6}$$

where α_i denotes the Lagrange multiplier, y_i is the training sample label, and b is the bias term. The bandwidth parameter γ and regularization coefficient C were optimized via grid search over ranges [0.001, 0.01, 0.1, 1] and [0.1, 1, 10, 100], respectively. The one-versus-rest strategy was employed for multi-class classification, previously validated for geological data by Al-Anazi and Gates (2010).

RF operates as an ensemble method that combines multiple decision trees through three key mechanisms: (1) bootstrap sampling to generate diverse training subsets, (2) random feature selection at each node split, and (3) majority voting for final predictions. The ensemble prediction function is expressed as:

$$H(x) = \mathrm{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\} \tag{7}$$

where h_t represents the prediction from the t-th tree, and T is the total number of trees. RF handles continuous and categorical data effectively and maintains stable performance across varied datasets (Zhao et al., 2025). Hyperparameters were optimized over ranges including n_estimators $\in$ [100, 200, 300], max_depth $\in$ [None, 10, 20, 30], and min_samples_split $\in$ [2, 5, 10].

GBDT implements forward stagewise additive modeling, iteratively improving predictions by fitting residuals from previous iterations (Tang et al., 2021). The ensemble after m iterations follows:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \tag{8}$$

where $F_m(x)$ is the previous ensemble, $h_m(x)$ is the current base learner, and η is the learning_rate. Hyperparameter optimization covered n_estimators $\in$ [100, 200, 300], learning_rate $\in$ [0.01, 0.1, 0.2], and max_depth $\in$ [3, 5, 7]. Five-fold cross-validation was used to determine the optimal configurations that balance model complexity with generalization performance.

Performance evaluation is centered on confusion matrix analysis derived from key classification metrics. For multi-class lithofacies identification, accuracy, precision, recall, and F1 score were calculated from true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN) (Tharwat, 2021). The F1 score, representing the harmonic mean of Precision and Recall, proves particularly valuable for imbalanced datasets. Off-diagonal confusion matrix elements revealed inter-class misclassification patterns, providing insights for refining augmentation strategies (Obi, 2023).

The model evaluation followed a three-stage protocol: First, each algorithm was trained on original (926 samples) and augmented (3600 samples) datasets, with performance improvements quantified through comparative analysis. Second, five-fold cross-validation was used to assess model stability and reliability, employing stratified sampling to maintain consistent lithofacies proportions across folds. Third, the best-performing augmented models underwent blind testing on independent wells N55-X1 and NY1-6HF from different structural positions to evaluate generalization to unseen data. This comprehensive framework validated the method's effectiveness across performance enhancement, stability, and generalization dimensions.

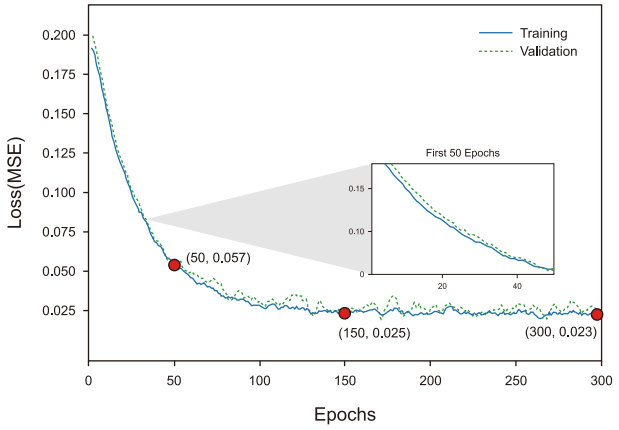

Fig. 4. Training and validation loss curves during the CDPM training process.

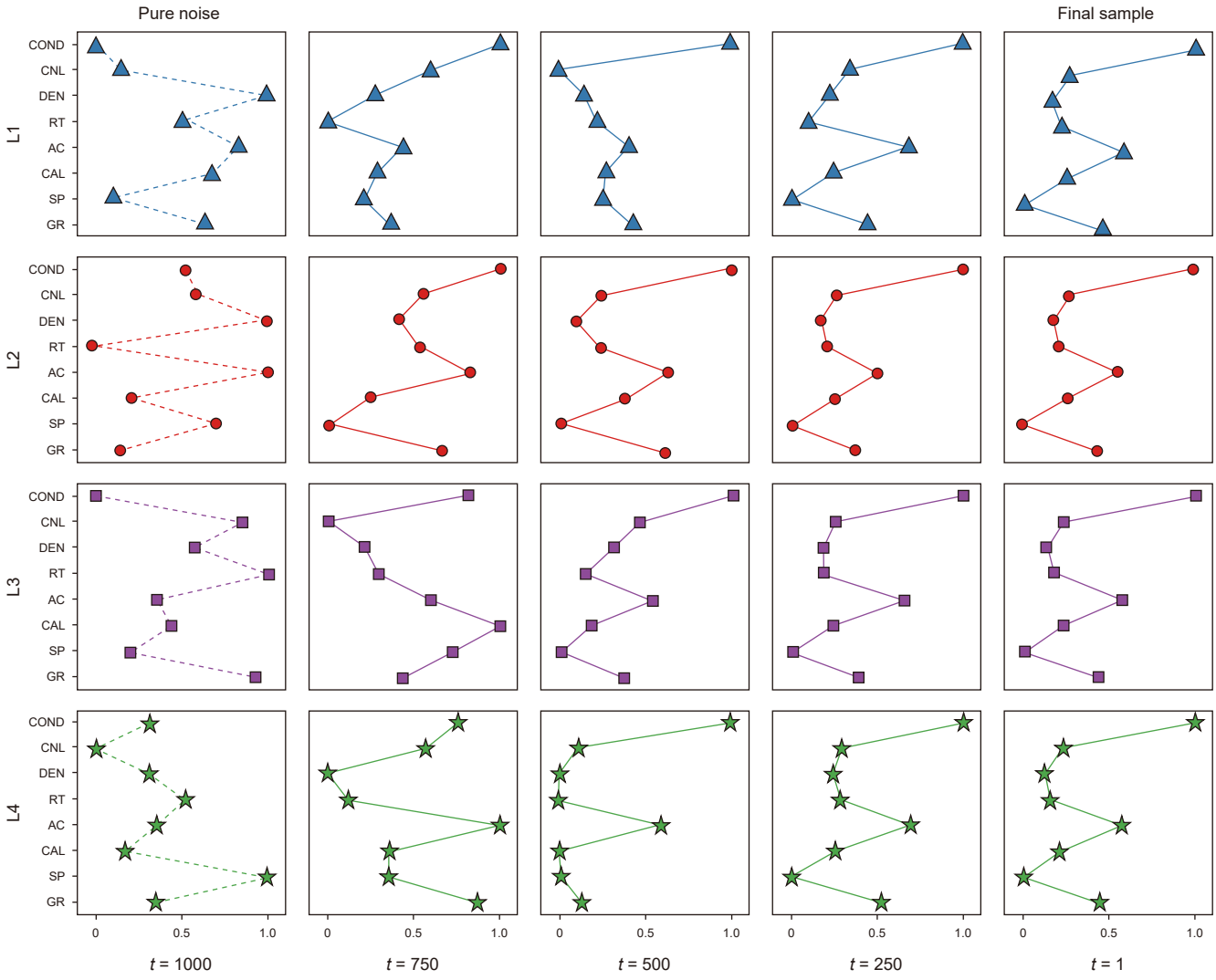

Fig. 5. Evolution of well-logging data for different lithofacies during the reverse diffusion process.

4. Results

4.1. Characteristics of CDPM training and augmented data

4.1.1. CDPM training process

A systematic CDPM training framework was established to learn the underlying distribution of well-log data and generate high-quality augmented samples. Using stratified sampling, the dataset was divided into training and validation sets at an 8:2 ratio, ensuring consistent lithofacies proportions across both subsets. Based on the network architecture described in Section 3.2, the model was trained using batch stochastic gradient descent with a batch size of 64 samples over 300 epochs.

The AdamW optimizer was used, with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, an initial learning rate of 1×10^{-4}, and a weight decay of 0.01. To ensure training stability, gradient norm clipping (with a threshold of 5.0) and adaptive learning rate scheduling were implemented. As shown in Fig. 4, the training loss decreased rapidly from 0.179 to 0.057 within the first 50 epochs, eventually converging smoothly to 0.023 at epoch 300. The validation loss exhibited a similar trend, stabilizing at 0.026. The gap between training and validation losses remained below 0.003, indicating good generalization without overfitting. Two strategies further improved training stability: cosine annealing reduced the learning rate to 85% of its original value every 75 epochs, suppressing gradient oscillations in later training stages. Additionally, a warm-up phase linearly increased the learning rate from 1×10^{-5} to 1×10^{-4} over the first 10 epochs, thereby reducing early training instability. The noise schedule parameter β increased linearly from 1×10^{-4} to 0.02 over 1000 diffusion steps, with mean squared error (MSE) as the loss function. Model parameters converged and stabilized after epoch 300. After training, the model generates augmented samples by starting from pure Gaussian noise and iteratively denoising over 1000

Table 1
Class distribution in original and augmented datasets.

Lithofacies	Original		Augmented		Increase, %
	Samples	Percentage, %	Samples	Percentage, %	
L1	92	9.9	900	25.0	878.3
L2	264	28.5	900	25.0	240.9
L3	355	38.3	900	25.0	153.5
L4	215	23.2	900	25.0	318.6
Total	926	100.0	3600	100.0	288.8

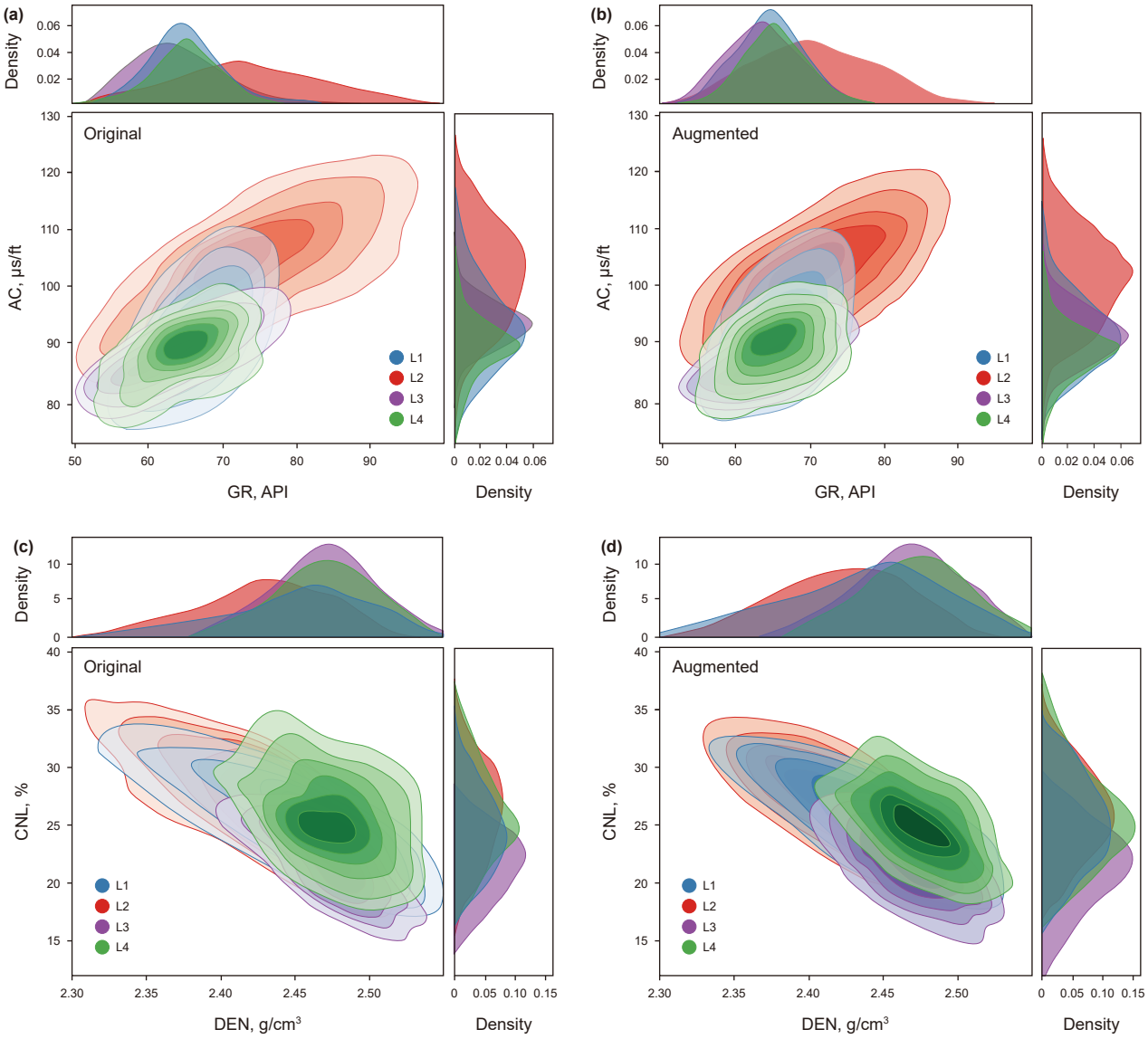

Fig. 6. Comparison of log curve distributions between original and CDPM-augmented data.

steps, guided by lithofacies condition labels. Fig. 5 illustrates the evolution of the data structure at various timesteps during reverse diffusion. The generation process unfolds in three phases: the early stage (t = 750–1000) builds basic data structure from noise; the middle stage (t = 250–500) develops characteristic differences between lithofacies categories; and the late stage (t = 1–250) refines physical relationships between logging parameters to ensure geological consistency.

4.1.2. Augmented data characteristics and quality assessment

CDPM demonstrated excellent performance in balancing class distributions for augmenting lacustrine shale lithofacies. The original dataset showed severe class imbalance, with L3 lithofacies dominating at 38.3%, while the critical L1 lithofacies represented only 9.9%—a nearly 4-fold difference (Table 1). Data augmentation adjusted each lithofacies class to 900 samples, achieving complete class balance and expanding the total dataset from 926 to 3,600 samples (an increase of 288.8%). This strategy used differentiated augmentation ratios: the rarest L1 lithofacies increased by 878.3%, while the dominant L3 lithofacies increased by only 153.5%. This "non-uniform augmentation" approach effectively resolved the class imbalance issue.

The augmented data also preserved the well-log response characteristics of each lithofacies. Fig. 6 shows that log parameter distributions for each lithofacies remained consistent with their geological characteristics. L1 lithofacies exhibited GR values (58.44–69.98 API), AC values (83.99–99.58 μs/ft), density values (2.39–2.50 g/cm^3), and CNL values (19.90%–29.69%) consistent with high carbonate and organic matter content. L2 lithofacies exhibited elevated GR values (60.23–80.11 API) and AC values (89.84–112.14 μs/ft), indicating a higher clay content. L3 lithofacies displayed GR values (57.50–68.70 API), AC values (85.93–97.87 μs/ft), and density values (2.42–2.52 g/cm^3) consistent with intermediate organic matter content. L4 lithofacies presented low AC values (84.39–94.92 μs/ft) and high density (2.43–2.51 g/cm^3), indicating low organic matter content. Other log parameters (SP, CAL, RT, COND) in the augmented dataset also maintained distributions consistent with the original data.

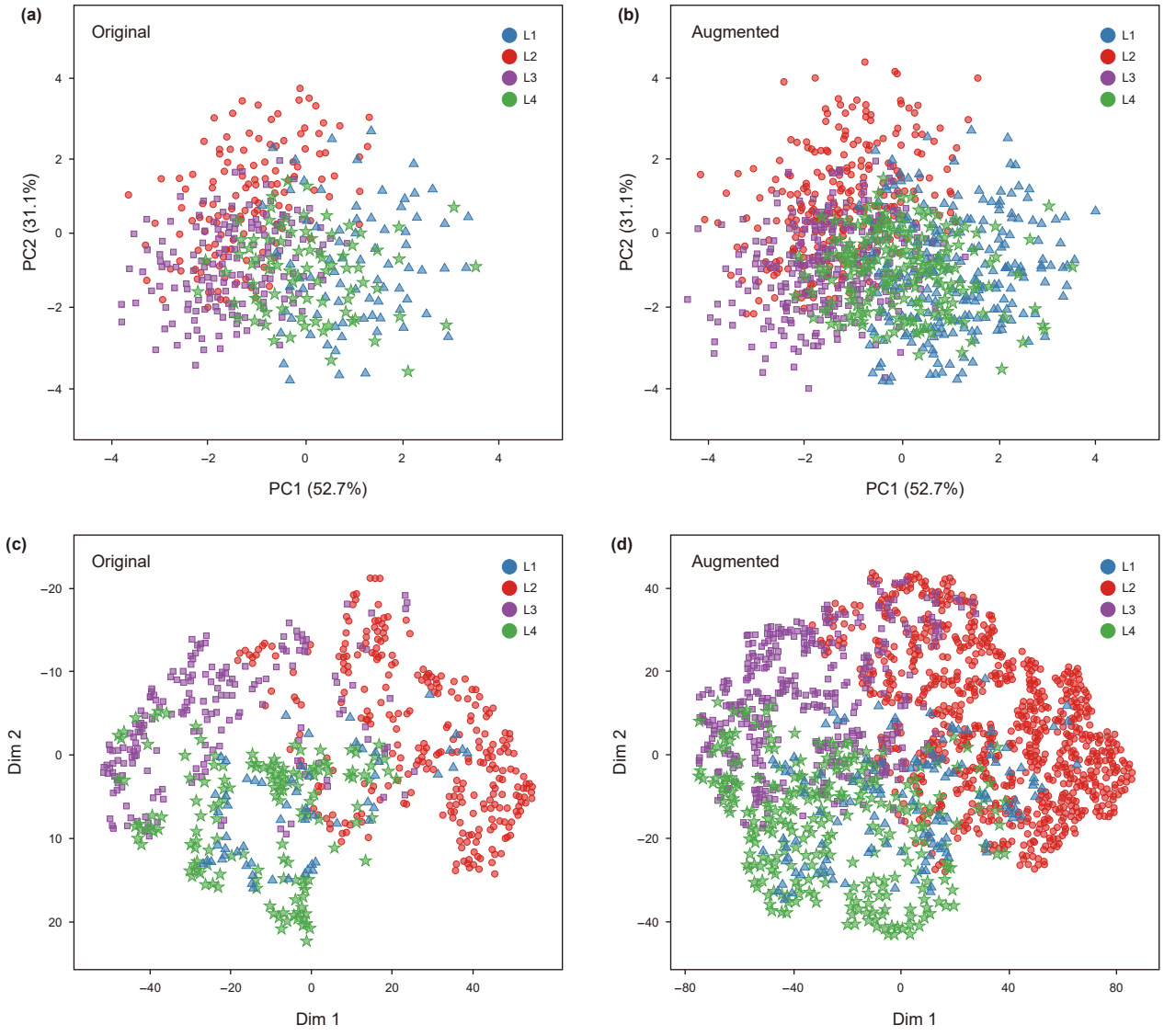

Fig. 7. Distribution comparison of original and augmented data in dimensionally reduced feature space.

The augmented data exhibited two key characteristics: first, more concentrated distributions that enhanced separability among lithofacies; second, preservation of relative positions among lithofacies in log response space, ensuring consistency in geological interpretation. Furthermore, physical correlations between log curves were well preserved, such as the negative correlation between CNL and DEN and the distribution patterns of lithofacies in AC–GR space. Areas where lithofacies partially overlap (such as L1 and L2) accurately reflect their geological similarities.

Table 2
Statistical quality metrics of augmented data.

Lithofacies	FID score	MMD score	Intra/Inter-class variance ratio
L1	26.3	0.085	0.5
L2	19.9	0.068	0.2
L3	23.7	0.081	0.4
L4	22.9	0.078	0.3
Average	22.9	0.078	0.4

Overall, the CDPM-generated augmented data achieved class balance while preserving physical characteristics and geological signatures, providing an optimal foundation for lithofacies classification.

4.1.3. Multidimensional distribution analysis

Quantitative analysis employed multidimensional feature space approaches to evaluate augmented data quality. Fig. 7 compares original and augmented data distributions in reduced-dimensional space, demonstrating CDPM's ability to preserve structural integrity.

PCA results showed similar distribution patterns between the augmented and original data in principal component space (Fig. 7(a) and (b)). The first two principal components accounted for 83.8% of the total variance (52.7% from PC1 and 31.1% from PC2), with apparent spatial clustering observed among lithofacies. L2 lithofacies clustered in the negative PC1/positive PC2 quadrant, L3 lithofacies concentrated in the negative PC1 region, L4 lithofacies occupied the negative PC2 region, and L1 lithofacies displayed transitional distributions at inter-lithofacies boundaries.

Notably, the augmented data achieved balanced sample density while preserving this spatial topology.

t-SNE analysis further confirmed that the augmented data preserved structure within nonlinear feature space (Fig. 7(c) and (d)). The original data exhibited distinct clustering patterns, with L2 lithofacies concentrated on the right, L3 lithofacies on the left, L4 lithofacies at the bottom, and L1 lithofacies distributed along interlithofacies boundaries. The augmented data preserved these clustering patterns while increasing sample density in boundary regions, filling the transition zones between L1–L2 and L3–L4 lithofacies.

Comparison of the original and augmented datasets revealed that data augmentation successfully addressed class imbalance issues. The initial problem of dense L3 and sparse L1 samples was significantly alleviated, resulting in a more balanced representation of lithofacies in the feature space. This improvement was particularly evident in the *t*-SNE representation, where the augmented data filled in sparse regions while maintaining inherent clustering properties. Notably, L1 lithofacies exhibited transitional distributions in both dimensionality reduction approaches, consistent with their complex geological origins.

The statistical quality metrics in Table 2 quantify the performance of augmented data. FID scores averaged 22.9 across lithofacies, with meaningful variation: L2 lithofacies scored lowest (19.9), reflecting tight clustering, while L1 lithofacies scored highest (26.3), consistent with dispersed transitional behavior. MMD scores ranged from 0.068 (L2) to 0.085 (L1), showing systematic progression. Intra-class/inter-class variance ratios further validated these patterns: L1 lithofacies exhibited the highest ratio (0.5), confirming extensive overlap; L2 lithofacies showed the lowest ratio (0.2), indicating distinct clustering; the ratios for L3 and L4 fell at intermediate levels. These metrics collectively confirm that augmented data maintains geological fidelity while achieving improved class balance.

4.2. Performance-based quality assessment of augmented data

4.2.1. Overall performance assessment

A quality assessment was conducted on the augmented data generated by the CDPM using three mainstream ML algorithms. Fig. 8 presents performance comparisons between RF, SVM, and GBDT models trained on original versus augmented datasets. These comparisons demonstrate the practical utility of the samples generated by the CDPM. The results indicate that CDPM significantly enhanced the performance of all classification algorithms, confirming the effectiveness of the augmentation approach.

Table 3
Classification performance metrics of the RF model on different lithofacies.

Lithofacies	Dataset	Precision, %	Recall, %	F1 score, %	Samples
L1	Original	66.7	52.6	58.8	19
	Augmented	87.0	85.7	86.4	363
L2	Original	79.5	81.7	80.6	71
	Augmented	88.2	88.3	88.2	256
L3	Original	73.3	83.0	77.9	53
	Augmented	89.5	87.6	88.5	250
L4	Original	78.9	69.8	74.1	43
	Augmented	87.7	87.2	87.4	250

The primary benefit of data augmentation was the effective correction of bias in classification outcomes. Models trained on the original data exhibited severe imbalances in precision and recall, with discrepancies reaching 10.6 percentage points for RF, 8.5 percentage points for SVM, and 4.6 percentage points for GBDT. In contrast, models trained on the augmented data significantly reduced these imbalances to 0.7 percentage points for Random Forest and 1.7 percentage points for SVM. This indicates that the augmented samples successfully corrected distributional bias and provided more balanced training examples.

The comprehensive performance improvements further validated the quality of the augmented data. The classification accuracy increased from 73.1% to 74.7% and then to 85.2%–88.8% across all algorithms, with F1 scores improving by an average of 15.7 percentage points. Notably, recall showed substantial enhancement, with an average improvement of 19.2 percentage points, significantly outpacing precision gains of 11.9 percentage points. This asymmetrical improvement pattern suggests that the augmented data greatly enhanced the model's sensitivity to minority classes while maintaining precision in the classification boundaries.

In summary, CDPM-generated augmented data demonstrated two critical quality characteristics: effective correction of class distribution bias and substantial enhancement of classification performance. This establishes a robust foundation for accurately identifying lithofacies.

4.2.2. Quality assessment for specific lithofacies

To evaluate the differential impact of data augmentation on individual lithofacies, the RF model, which demonstrated optimal performance, was employed to systematically compare lithofacies

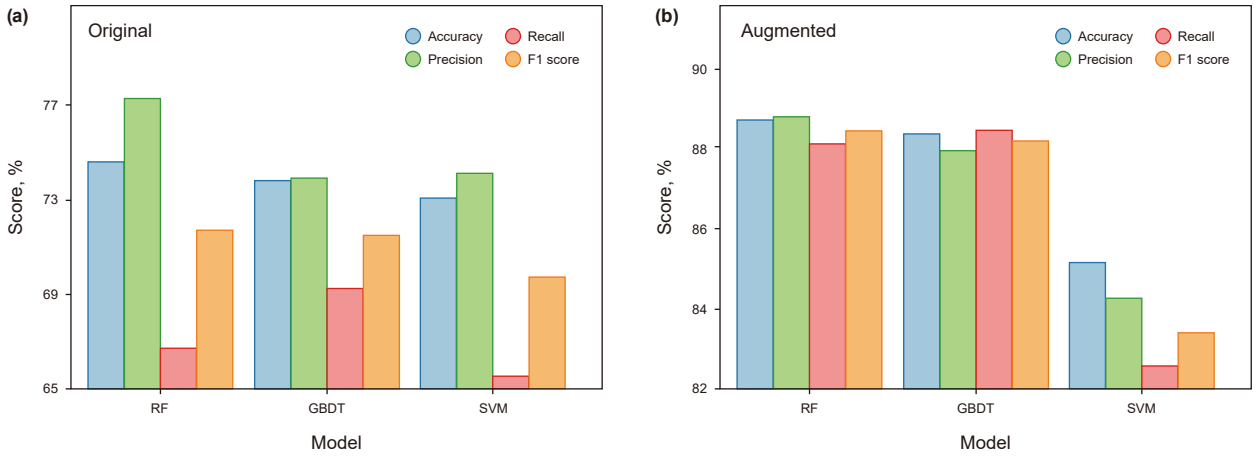

Fig. 8. Performance comparison of three classification models on original and augmented datasets.

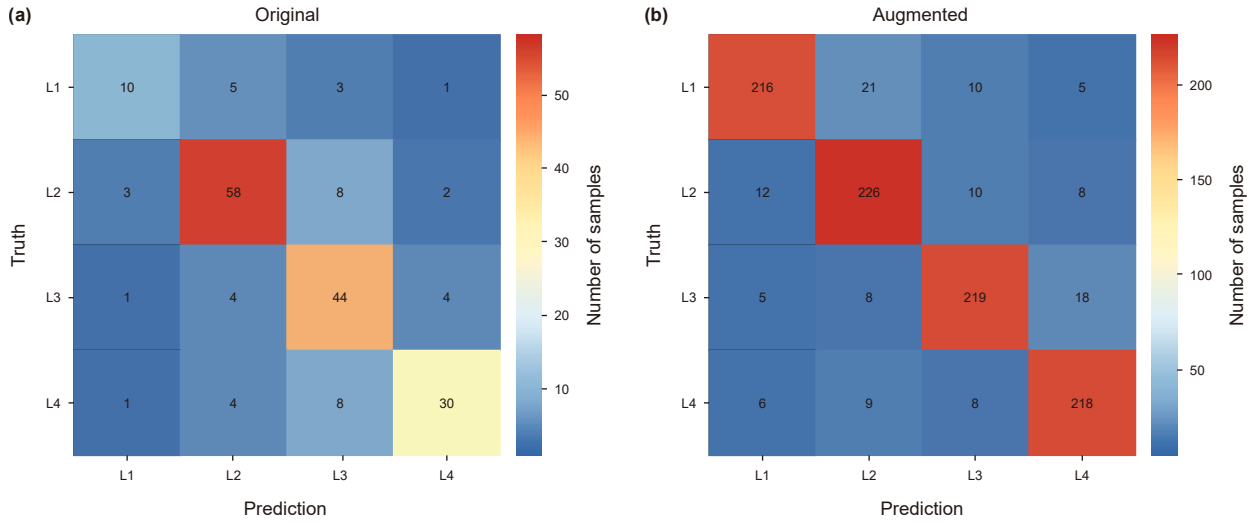

Fig. 9. Confusion matrices of the RF model on original and augmented datasets.

identification results between original and augmented datasets (Table 3).

Augmented data exhibited markedly different improvement patterns across lithofacies categories. The L1 lithofacies, representing the rarest category (9.9% of original samples), exhibited the most substantial enhancement: F1 score increased by 27.5 percentage points, and recall improved by 33.1 percentage points, demonstrating that augmentation effectively addressed the representation deficiency of rare samples. For the L3 lithofacies, originally the dominant category (38.3%), improvement manifested primarily through a 16.1 percentage point increase in precision, indicating that augmented data corrected model bias toward dominant categories and enhanced classification boundary accuracy. The L4 lithofacies achieved a significant 17.4 percentage point improvement in recall, while the L2 lithofacies demonstrated balanced enhancement across all metrics.

Data augmentation significantly improved the balance of lithofacies identification performance. After augmentation, the F1 score distribution narrowed from 58.8% to 80.6% (with a standard deviation of 9.7) to a tighter range of 86.4% to 88.5% (with a standard deviation of 0.9). This suggests that the augmented data effectively balanced the representational quality across different lithofacies, allowing the model to recognize all categories with equal capability.

Confusion matrix analysis (Fig. 9) revealed the exceptional performance of augmented data in preserving geological fidelity. L1 lithofacies misclassification decreased significantly from 47.4% to 14.3% while maintaining limited confusion with L2 lithofacies (8.3%), accurately reflecting the geological similarity between these organic-rich lithofacies. The approximately 7% mutual misclassification between L3 and L4 lithofacies similarly reflected geological characteristics associated with organic matter content gradients.

The CDPM-generated augmented data substantially improved the quality of rare lithofacies characterization and high-level performance balance across different lithofacies. It accurately preserves inter-lithofacies geological correlations, providing a high-quality training foundation for the precise identification of shale lithofacies.

4.2.3. Blind well validation

To comprehensively assess the practical applicability and generalization capability of the augmented dataset, external validation was conducted on two independent wells from different structural positions: well N55-X1 from the gentle slope zone and well NY1-6HF from the steep slope zone.

External validation was first performed on well N55-X1. The analysis focused on the 3536.25–3604.64 m interval, which contains continuous core coverage and complete well log responses, providing optimal conditions for evaluating augmented data quality (Li et al., 2022).

Independent well validation demonstrated the augmented data's superior generalization capability. The RF model trained on augmented data (CDPM-RF) achieved an average accuracy of 79.7%, representing a 10.2 percentage point improvement over the conventional MRGC method (69.5%) (Fig. 10) (Ma et al., 2022a, b). F1 scores across all lithofacies showed substantial enhancement, with average values increasing from 68.9% to 79.3% (Fig. 11), confirming robust performance on previously unseen samples.

The quality advantages of augmented data became particularly evident under complex geological conditions. Within the L1–L2 lithofacies transition zone, at depths of 3585–3595 m, the CDPM-RF model accurately captured critical transition interfaces and thin interlayers. In contrast, the MRGC method exhibited notable instability in comparable zones (Fig. 10(a)). In the intercalated lithofacies region at 3570–3580 m, the augmented data-trained model accurately delineated primary lithofacies distributions while successfully identifying multiple thin interlayers, demonstrating exceptional capability for characterizing rare geological features (Fig. 10(b)).

Geological continuity preservation was further validated through blind well predictions. Test interval predictions generally maintained reasonable geological coherence, particularly evident in the 3552–3554 m section, where L4 lithofacies distribution was accurately characterized, reflecting the augmented data's robust resistance to well log noise (Fig. 10(c)). This stability resulted from CDPM's precise learning of original data distribution characteristics and effective resolution of class imbalance issues.

Results from individual lithofacies identification further validated the quality of the augmented data. Although L1 lithofacies achieved a relatively modest F1 score of 77.0% in blind well testing, this still represented a substantial improvement over the MRGC method and remained consistent with its transitional distribution characteristics in feature space. L4 lithofacies enhanced F1 scores from 64.1% to 76.5%, confirming the effective enhancement in characterizing rare categories.

To further strengthen model validation across different structural positions, supplementary blind well testing was performed

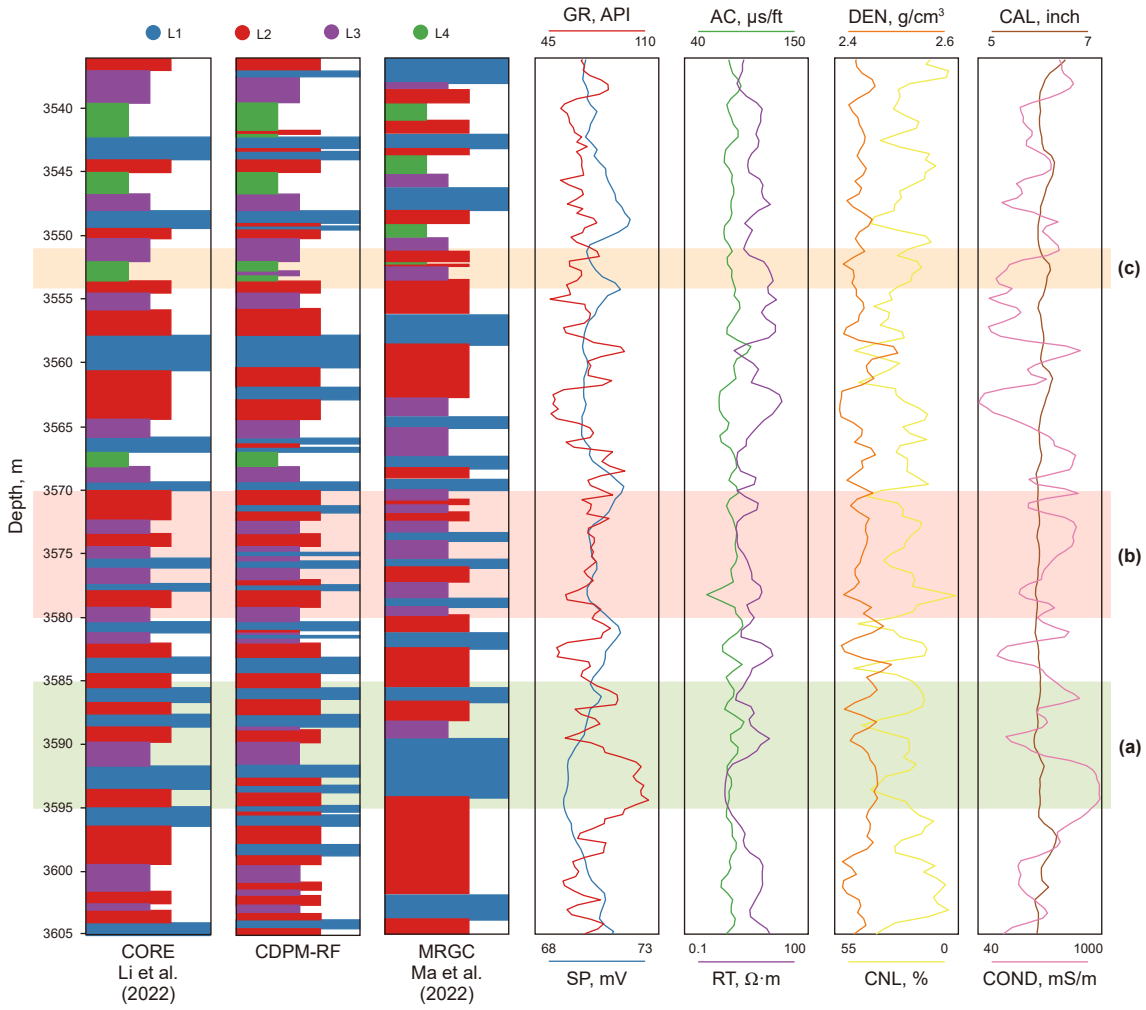

Fig. 10. Lithofacies prediction profile comparison of well N55-X1, showing core description (Li et al., 2022), MRGC method (Ma et al., 2022a,b), and CDPM-RF predictions.

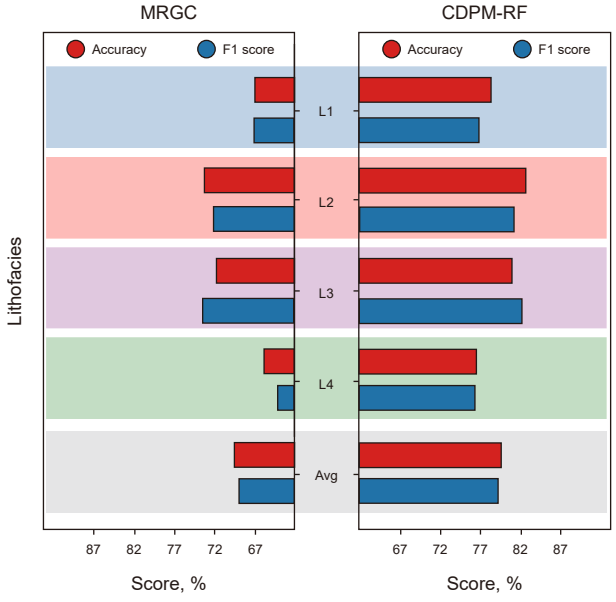

Fig. 11. Comparison of accuracy and F1 score between the MRGC method (Ma et al., 2022a, 2022b) and the CDPM-RF method on different lithofacies.

on well NY1-6HF, located in the steep slope zone. Given the limited core sample availability in this well, validation focused on the 3334.43–3377.70 m interval, where 21 core samples with calibrated lithofacies assignments and complete geochemical data were available. The CDPM-RF model demonstrated consistent predictive performance across this validation interval (Fig. 12). In the 3368.79–3374.25 m interval (Fig. 12(a)) and 3343.46–3347.51 m interval (Fig. 12(c)), the model successfully reproduced the interbedded patterns of L3 and L2 lithofacies observed in densely distributed core samples, with predicted lithofacies accurately capturing the complex interlayered distribution characteristics. Notably, the model successfully identified the economically critical L1 lithofacies at 3351.10–3353.90 m (Fig. 12(b)), the rarest yet most valuable lithofacies type, which represents only 9.9% of the original training dataset. These results demonstrate consistent model performance with the N55-X1 validation outcomes, confirming robust generalization capability. Independent multi-well validation across different structural positions demonstrated that CDPM-generated augmented data significantly enhanced the model's generalization capability for unseen samples, providing a reliable foundation for accurate shale lithofacies prediction.

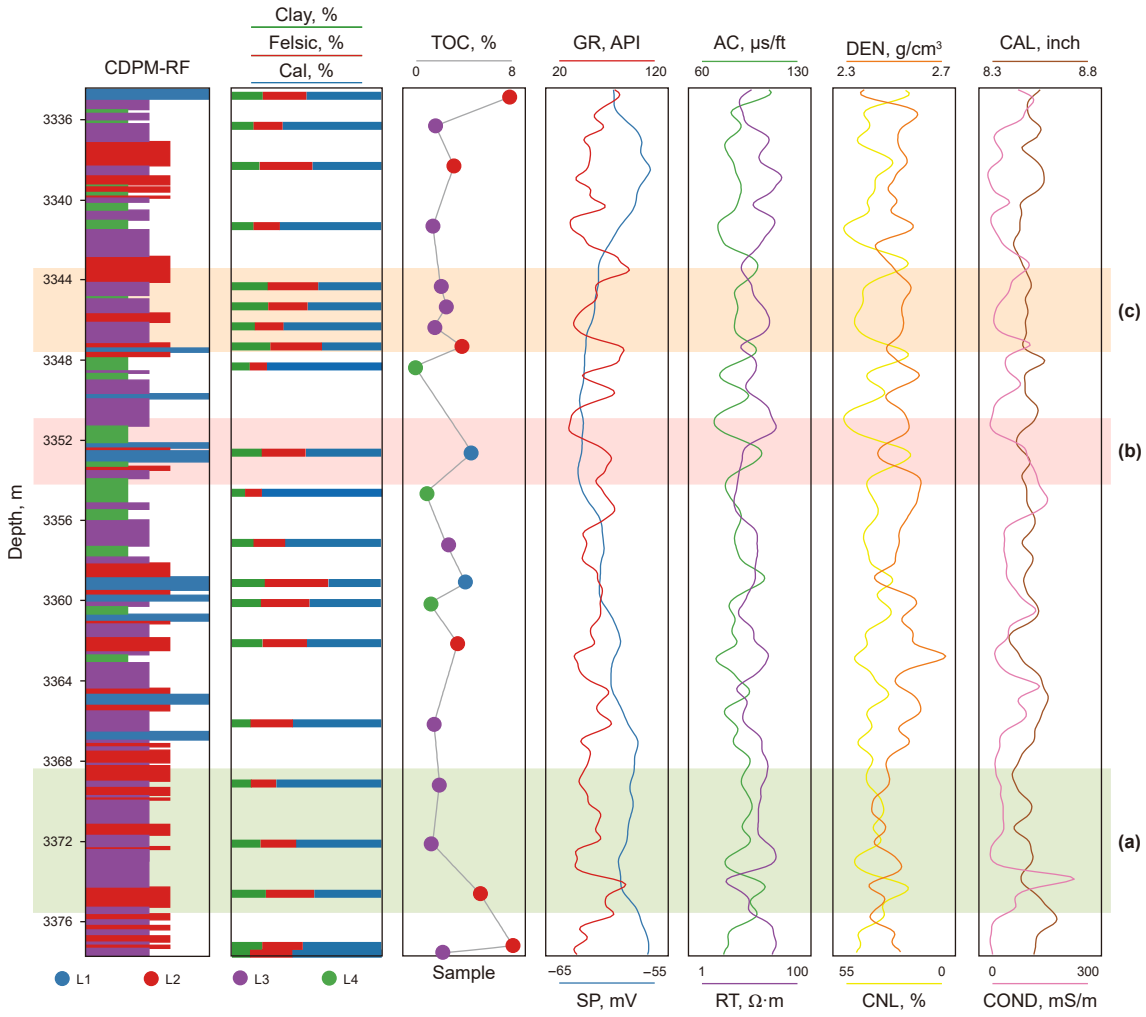

Fig. 12. Lithofacies prediction profile of well NY1-6HF showing CDPM-RF predictions, core samples with lithofacies assignments, geochemical and mineralogical data.

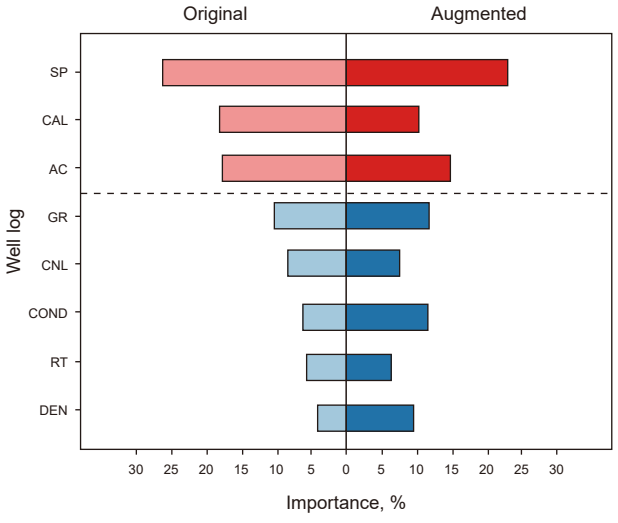

Fig. 13. Comparison of well-logging curve importance before and after CDPM data augmentation.

4.3. Feature importance analysis and data augmentation effects

Feature importance analysis based on the RF model revealed a significant impact of augmented data on optimizing the distribution of feature contributions. Fig. 13 illustrates the changes in the importance rankings of eight well-log curves between the original and augmented datasets, highlighting the advantages of CDPM in enhancing feature utilization balance.

The original dataset exhibited a pronounced bias that was dependent on specific features. Three curves—SP (27.5%), CAL (18.7%), and AC (18.3%)—accounted for a combined total of 64.5% of the overall importance. The high significance of the SP curve was due to its sensitivity to variations in siltstone content. In contrast, CAL curves effectively distinguished differences in stability among lithofacies, while AC curves indicated variations in rock acoustic transmission properties. On the other hand, the GR curve showed relatively low importance (10.7%), primarily because of the high carbonate content (50%–75%) across the lithofacies in the study area, which resulted in considerable overlap in GR values among lithofacies L1, L3, and L4 (with ranges of 58.44–69.98, 57.50–68.70, and 60.10–69.67 API, respectively). Only lithofacies L2 exhibited slight differentiation due to its elevated clay content. The CNL, COND, RT, and DEN curves contributed

Table 4
Impact of diffusion steps on data quality and classification performance.

Diffusion steps	FID score	Accuracy, %	F1 score, %	FID decrease, %	Accuracy gain, %
500	34.4	78.8	76.8	-	-
750	28.8	83.7	82.2	16.3	6.2
1000	22.9	88.8	88.5	20.5	6.0
1250	20.9	89.9	89.5	8.3	1.2
1500	19.9	90.1	89.9	5.3	0.3

minimally, with the DEN curves being particularly limited in distinguishing between lithofacies L1 and L4, as their density value ranges were similar (2.39–2.50 g/cm^3 versus 2.43–2.51 g/cm^3).

The introduction of augmented data significantly improved the balance in the distribution of feature importance. While SP retained its leading role, its contribution decreased to 24.3% (a 3.2 percentage point reduction). Similarly, CAL and AC contributions fell to 10.9% and 15.6%, respectively, with CAL experiencing the most substantial decline of 7.8 percentage points. Meanwhile, the importance of COND and DEN curves increased by 5.9 and 5.7 percentage points, rising to 12.2% and 9.9%, respectively, while GR and RT also showed moderate improvements. This redistribution indicated that augmented data effectively achieved a balanced optimization of feature contributions, reducing the model's over-reliance on individual features and promoting the comprehensive utilization of multidimensional information.

The benefit of augmented data manifests in its ability to facilitate synergistic mechanisms among multiple well-log curves. In the original dataset, the significant overlap between lithofacies L1 and L4 in GR and DEN values limited the ability to discriminate using a single curve. However, models trained with augmented data effectively integrated multi-curve information from SP, AC, and COND, leveraging their complementary strengths for the recognition of rare lithofacies. This synergistic effect was directly reflected in the substantial improvement of the F1 score for lithofacies L1, which increased from 58.8% to 86.4%. This confirms the advantages of augmented data in improving the identification of complex lithofacies.

5. Discussion

5.1. Impact of diffusion steps on data quality and performance

Diffusion steps constitute a critical hyperparameter in CDPM, directly governing the quality of augmented data and subsequent classification performance (Amit et al., 2021; Nichol and Dhariwal, 2021). This analysis systematically evaluates the effects of varying diffusion steps (500, 750, 1000, 1250, and 1500) on quality metrics and classification effectiveness, elucidating the complex relationship between diffusion step count, data fidelity, and model performance. The analysis reveals a dual impact of diffusion steps on augmented data quality and classification performance, characterized by a distinct "acceleration–saturation" progression (Table 4). Increasing the diffusion steps from 500 to 1000 resulted in a substantial 33.5% reduction in FID score, alongside a 9.9 percentage point improvement in classification accuracy. However, extending to 1500 steps produced diminishing returns, showing only a 13.1% improvement in FID score and a marginal 1.4 percentage point accuracy gain, demonstrating diminishing marginal benefits.

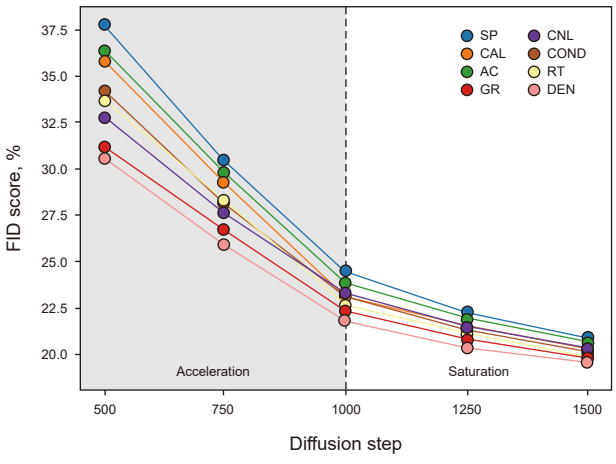

Fig. 14. Relationship between FID scores of different logging curves and diffusion steps.

Differential sensitivity to diffusion steps emerges across well-log curve types (Fig. 14). GR and DEN curves, characterized by relatively simple distributions among lithofacies, achieve satisfactory quality metrics (FID~31) at lower diffusion steps. Conversely, SP, AC, and CAL curves, exhibiting more complex distributions, require higher diffusion steps for substantial improvement. This phenomenon aligns with observations by Ho et al. (2020) in image generation: initial diffusion steps establish primary structural features, while subsequent iterations predominantly refine local details. The first 1000 diffusion steps primarily establish core distributional characteristics of well-log responses and principal inter-curve correlations (e.g., compensated neutron log-density negative correlation) (Fig. 6(c) and (d)), whereas subsequent steps optimize complex inter-curve relationships (e.g., subtle variations in acoustic-gamma ray parameter space) with relatively limited contribution to overall quality enhancement.

A strong correlation emerges between data quality and classification performance (Fig. 15(a)). Decreasing FID scores correspond to significant upward trends in accuracy and F1 scores, confirming data quality as a fundamental determinant of classification effectiveness. Notably, the log-transformation of diffusion steps reveals approximately linear growth trends for both accuracy and F1 scores, with coefficients of determination (R^2) reaching 0.94 and 0.93, respectively (Fig. 15(b)). This log-linear relationship corroborates findings by Nichol and Dhariwal (2021) in diffusion modeling research, providing theoretical guidance for optimal diffusion step selection.

This log-linear relationship reflects the complex mechanism of the diffusion process: initial diffusion phases primarily capture principal distributional characteristics of the data, and when quality reaches a threshold level (FID < 23.0), the model adequately represents core lithofacies features for effective classification. Subsequent iterations primarily optimize subtle boundary region differences, facilitating the discrimination of lithofacies with overlapping feature spaces, though they contribute only marginally to overall classification performance. Classification performance and data quality exhibit highly synchronized trends, demonstrating the characteristic "acceleration–saturation" behavior and confirming their fundamental interconnection.

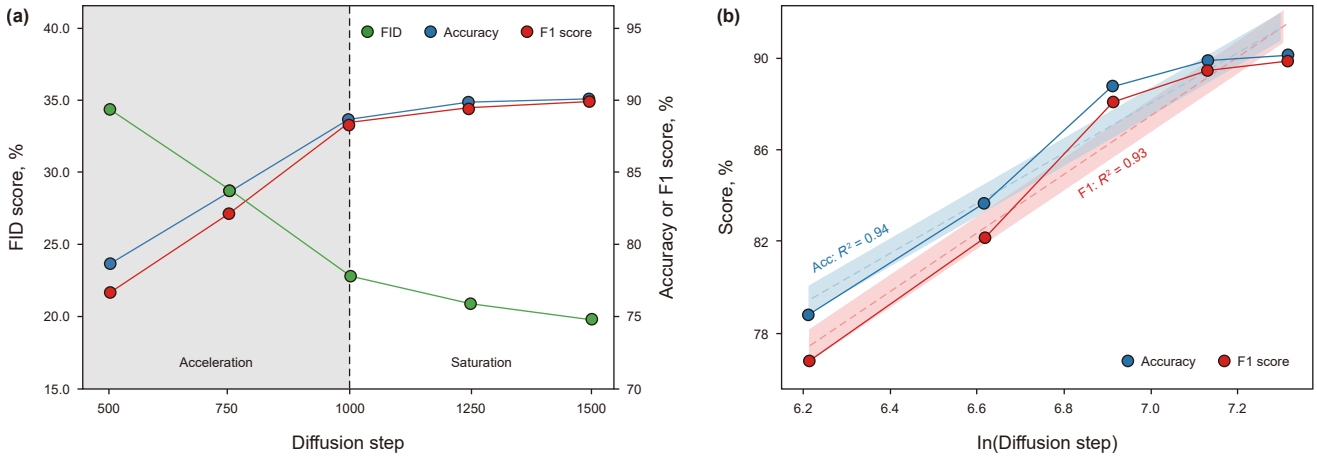

Fig. 15. Multi-perspective analysis of relationships among diffusion steps, data quality, and classification performance.

Table 5
Effect of augmented sample size on RF model performance.

Samples per class	Test accuracy, %	F1 score, %	Train accuracy, %	CV std dev, %	Accuracy gain, %
Original	74.7	71.7	98.1	3.6	-
300	78.1	77.1	91.2	3.3	4.5
600	82.4	83.1	86.4	2.2	5.5
900	88.8	88.5	89.7	1.4	7.7
1200	89.7	89.1	89.9	1.4	1.2
1500	89.9	89.3	90.1	1.3	0.1

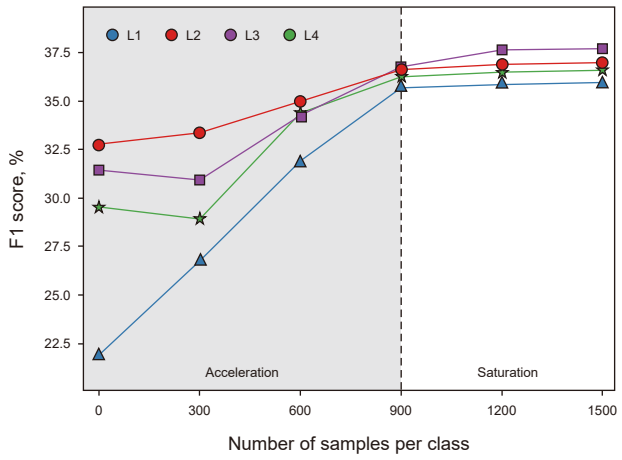

Fig. 16. Variation trends of F1 score for different lithofacies with increasing augmented sample size.

Considering data quality, classification performance, and computational efficiency collectively, 1000 diffusion steps represent the optimal configuration for this application: generating high-quality augmented data (FID = 22.9), supporting classification models achieving 88.8% accuracy and 88.5% F1 score—performance approaching that of 1500-step configurations while avoiding computational resource waste associated with excessive diffusion iterations.

5.2. Impact of augmented sample size on model performance

5.2.1. Effect on classification performance

The sample size is another critical parameter in CPDM implementation, and its influence on classification performance requires a systematic investigation (Yang and Ma, 2015). This analysis examined the impact of varying lithofacies sample quantities per class, ranging from 300 to 1,500, on classification performance while maintaining a diffusion step of 1,000.

Table 5 illustrates the evolution of the RF model's performance across various sample sizes. Consistent with diffusion step patterns, sample size increases exhibited a characteristic "acceleration–saturation" response: accuracy improved substantially by 10.6 percentage points (from 78.1% to 88.8%) when samples per class increased from 300 to 900, whereas further increases from 900 to 1,500 yielded only marginal improvements of 1.1 percentage points. F1 scores followed identical trends, rising from 71.8% in the original dataset to 88.5% at 900 samples before exhibiting diminished growth rates. This response pattern aligns with bias-variance decomposition theory (Geman et al., 1992): initial increases in sample size primarily reduce model variance, but as the sample size grows, the rate of variance reduction decelerates, and performance gains plateau.

Different lithofacies categories exhibited markedly distinct responses to sample augmentation (Fig. 16). The rarest lithofacies, L1 (original proportion: 9.9%), derived the maximum benefit from augmentation, with F1 scores increasing by 17.7 percentage points across the 300 to 900 sample range. Conversely, L3 lithofacies, representing the highest original proportion (38.3%), demonstrated initial performance degradation (declining from 77.9% to

Table 6
Effect of augmented sample size on data quality.

Samples per class	FID score	MMD score	FID decrease, %	MMD decrease, %
300	31.4	0.089	-	-
600	26.7	0.084	15.0	5.5
900	22.8	0.078	14.4	7.7
1200	21.9	0.077	3.9	1.7
1500	21.7	0.076	1.3	0.8

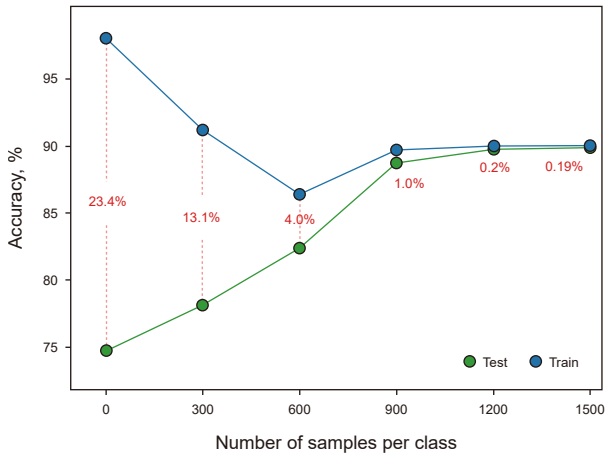

Fig. 17. Training-test performance gap with different augmented sample sizes.

76.8%) during early augmentation phases before recovering to 88.5%. This differential improvement pattern indicates that augmented data addresses class imbalance issues by preferentially enhancing recognition performance for underrepresented categories. Notably, all lithofacies categories demonstrated clear saturation trends beyond 900 samples per class, confirming the existence of optimal sample size ranges.

5.2.2. Effect on generalization capability

The quantity of augmented samples influences classification accuracy and determines the model's generalization capability. Table 6 demonstrates the evolution of data quality indicators across varying sample sizes. The FID score consistently decreased with increasing sample size, declining from 31.4 at 300 samples per class to 22.9 at 900 samples (27.3% reduction). However, further expansion from 900 to 1,500 samples yielded only a marginal decrease to 21.7 (5.3% reduction). The MMD score exhibited similar diminishing returns. This inflection point coincided precisely with the performance saturation observed in classification metrics, confirming that data quality is the critical mediating factor between sample quantity and generalization capability, consistent with findings reported by Karras et al. (2020) in generative model studies.

Fig. 17 illustrates the relationship between augmented sample size and model generalization performance. The original dataset exhibited a substantial accuracy gap of 23.4 percentage points between training and testing sets (98.1% vs. 74.7%), indicating severe overfitting. Progressive increases in augmented sample size systematically reduced this gap: 13.1, 4.00, and 1.00 percentage points at 300, 600, and 900 samples, respectively, with convergence to 0.2 percentage points at 1,200 samples, approaching the optimal bias-variance equilibrium.

Training set accuracy followed a characteristic pattern of initial decline followed by recovery, reflecting the model's transition from an overfitted to a balanced state (Goodfellow et al., 2016). Concurrently, the K-fold cross-validation standard deviation decreased from 3.6% on the original dataset to 1.4% with 900 augmented samples, indicating substantial improvement in model stability. The convergence of training-testing performance gaps and reduced variance collectively demonstrated that increased augmentation enabled the model to achieve superior bias-variance balance (Tantithamthavorn et al., 2016).

Notably, the marginal benefit of additional augmented samples on generalization improvement exhibited clear diminishing returns. Beyond 900 samples per class, improvements across all generalization metrics became negligible, indicating an optimal balance between augmentation quantity and generalization capability. Further sample increases provided no significant enhancement of generalization.

Based on a comprehensive evaluation of classification performance and generalization capability, 900 samples per class emerged as the optimal configuration for this study. This setting achieved high classification accuracy (88.8%) while maintaining robust generalization capability (training-testing gap of merely 1.0 percentage points) and superior data quality (FID = 22.9) while avoiding computational resource waste associated with excessive augmentation.

5.3. Geological limitations and cross-basin applicability

This study demonstrates the effectiveness of CDPM-based data augmentation for shale lithofacies classification within the Dongying Depression, achieving robust performance across wells from different structural positions. However, as a single-basin study, the model's generalization capability is inherently constrained by the geological representativeness of the training dataset. While our validation wells span different structural positions, they share the same basin-scale depositional framework. Lithofacies variations in unsampled depositional sub-environments within the basin remain to be characterized. Future work will focus on acquiring additional data from underrepresented settings to further validate and refine the augmentation framework.

A fundamental challenge for ML models trained on single-basin datasets is their transferability to basins with contrasting geological characteristics. Lacustrine lithofacies assemblages exhibit substantial variability controlled by complex interactions among depositional climate, lake water conditions, and sediment supply (Li et al., 2022; Liang et al., 2023). These factors are influenced by the basin-scale tectonic setting, paleoclimate, water chemistry, and sediment provenance, which collectively determine the mineralogical composition, type of organic matter, and geochemical signatures preserved in lacustrine shales. The quantity and geological diversity of training data critically influence model generalization performance. Limited data from a single geological setting may be insufficient to represent the full spectrum of variability encountered in other lacustrine systems (Meng et al., 2025; Zhao et al., 2024). For instance, saline lacustrine systems developed under arid paleoclimates typically accumulate carbonate-evaporite facies with elevated salinity indicators, contrasting markedly with freshwater systems that

accumulate clay-rich facies dominated by terrigenous input under humid conditions (Hou et al., 2022; Liang et al., 2018b). The Cretaceous Songliao Basin exemplifies such freshwater systems, characterized by clay-rich lithofacies with substantial terrigenous clastic input (Liu et al., 2019), which is fundamentally different from the carbonate-rich facies of saline rift basins, such as the Dongying Depression (Liang et al., 2018a). Although both systems can develop algal-derived organic matter under favorable conditions, their mineralogical frameworks differ markedly, reflecting distinct water chemistry and paleoclimate regimes that control lithofacies development.

These inter-basin geological variations highlight the importance of systematic cross-basin validation for evaluating model transferability. Future research directions include: (1) extending the CDPM framework to representative Chinese lacustrine basins (e.g., Songliao Basin, Subei Basin) with contrasting water chemistry and depositional settings; (2) identifying basin-independent feature engineering strategies that capture fundamental lithofacies-logging relationships; and (3) developing transfer learning approaches enabling efficient model adaptation with limited retraining data. By progressively incorporating multi-basin training datasets, we aim to construct increasingly generalized augmentation frameworks that can address the full spectrum of lacustrine shale diversity encountered in petroleum exploration and development practices.

6. Conclusion

Our study introduces the first application of CDPMs for classifying lacustrine shale lithofacies, thereby creating a systematic framework for data augmentation that addresses the ongoing issues of small sample sizes and class imbalance in geological datasets. By constructing a stochastic diffusion process with conditional control mechanisms for well-log data, the framework successfully transformed 926 severely imbalanced samples into 3,600 balanced samples, achieving an 878.3% increase in the representation of the rarest L1 lithofacies. The augmented dataset significantly improved the class distribution while preserving the original data's statistical characteristics (FID = 22.9, MMD = 0.078) and petrophysical relationships. The model's performance demonstrated the effectiveness of the augmentation approach, with average accuracy and F1 scores across three mainstream algorithms improving by 13.6 and 16.6 percentage points, respectively. Notably, the RF model achieved 88.8% accuracy. The approach notably reduced traditional model bias toward rare lithofacies, with L1 lithofacies F1 scores increasing from 58.8% to 86.4% and recall improving by 33.1 percentage points. Multi-well blind validation across different structural positions confirmed the method's generalization capability, with models trained on augmented data achieving 79.7% accuracy on well N55-X1 and maintaining consistent performance on well NY1-6HF— representing a 10.2 percentage point improvement over conventional approaches. These results establish CDPM-based data augmentation as a practical and effective tool for lithofacies classification in lacustrine shale reservoirs. The successful implementation of CDPMs in geological applications opens new opportunities for advancing ML techniques in earth sciences, particularly for hydrocarbon exploration and reservoir characterization. To further enhance the practical applicability of this methodology, future research should extend beyond the current single-basin framework. This includes acquiring additional data from underrepresented depositional settings within the Dongying Depression and systematically validating the approach across Chinese lacustrine basins with contrasting water chemistry and depositional characteristics. Such efforts will advance toward more robust and generalized augmentation frameworks capable of addressing the full spectrum of lacustrine shale diversity encountered in petroleum exploration and development practice.

CRediT authorship contribution statement

Gui-Ang Li: Writing - original draft. **Cheng-Yan Lin:** Writing - review & editing, Conceptualization. **Chun-Mei Dong:** Supervision, Writing - review & editing, Conceptualization. **Li-Hua Ren:** Project administration, Writing - review & editing, Conceptualization. **Peng-Jie Ma:** Writing - review & editing. **Yu-Qi Wu:** Writing - review & editing. **Guo-Yin Zhang:** Methodology. **Xin-Yu Du:** Data curation. **Zi-Ru Zhao:** Conceptualization.

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

This work is funded by the National Natural Science Foundation of China (Nos. 42372156; 42202146; 42102153; 42472194), the Key R&D Program of Shandong Province, China (Grant No. 2022CXPT048), and the Research Contract “Comprehensive Evaluation and Prediction of Shale Oil Development Sweet Spots” (Contract No. 30200018-19-ZC0613-0116). We acknowledge the Exploration and Development Research Institute of Shengli Oilfield Company, SINOPEC, for providing core samples, well-log data, and geological data for this study. Special thanks are extended to the editor and anonymous reviewers for their constructive comments and suggestions.

References

Al-Anazi, A., Gates, I.D., 2010. A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs. Eng. Geol. 114 (3-4), 267–277. https://doi.org/10.1016/j.enggeo.2010.05.005.

Amit, T., Shaharbany, T., Nachmani, E., et al., 2021. Segdiff: Image segmentation with diffusion probabilistic models. ArXiv preprint arXiv:2112.00390. https://doi.org/10.48550/arXiv.2112.00390.

Ashraf, U., Zhang, H., Thanh, H.V., et al., 2024. A robust strategy of geophysical logging for predicting payable lithofacies to forecast sweet spots using digital intelligence paradigms in a heterogeneous gas field. Nat. Resour. Res. 33 (4), 1741–1762. https://doi.org/10.1007/s11053-024-10350-4.

Barandela, R., Valdovinos, R.M., Sánchez, J.S., et al., 2004. The imbalanced training sample problem: under or over sampling?. In: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), pp. 806–814. https://doi.org/10.1007/978-3-540-27868-9_88.

Bressan, T.S., de Souza, M.K., Girelli, T.J., et al., 2020. Evaluation of machine learning methods for lithology classification using geophysical data. Comput. Geosci. 139, 104475. https://doi.org/10.1016/j.cageo.2020.104475.

Chang, H., Kopaska-Merkel, D., Chen, H., 2002. Identification of lithofacies using Kohonen self-organizing maps. Comput. Geosci. 28 (2), 223–229. https://doi.org/10.1016/S0098-3004(01)00067-X.

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

Dong, S., Zeng, L., Du, X., et al., 2022. Lithofacies identification in carbonate reservoirs by multiple kernel Fisher discriminant analysis using conventional well logs: A case study in A oilfield, Zagros Basin, Iraq. J. Petrol. Sci. Eng. 210, 110081. https://doi.org/10.1016/j.petrol.2021.110081.

Dong, S., Zhong, Z., Cui, X., 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.

Ellis, D.V., Singer, J.M., 2007. Well Logging for Earth Scientists. Springer, Dordrecht.

Engelmann, J., Lessmann, S., 2021. Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning. Expert Syst. Appl. 174, 114582. https://doi.org/10.1016/j.eswa.2021.114582.

Fang, Z., Zhang, L., Yan, S., 2023. Forecast of lacustrine shale lithofacies types in continental rift basins based on machine learning: A case study from Dongying

Sag, Jiyang Depression, Bohai Bay Basin, China. Front. Earth Sci. 11, 1047981. https://doi.org/10.3389/feart.2023.1047981.
Geman, S., Bienenstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilemma. Neural Comput. 4 (1), 1–58. https://doi.org/10.1162/neco.1992.4.1.1.
Geng, Z., Liu, J., Li, S., et al., 2023. Channel attention-based static-dynamic graph convolutional network for lithology identification with scarce labels. Geoenergy Sci. Eng. 223, 211526. https://doi.org/10.1016/j.geoen.2023.211526.
Goodfellow, I., Bengio, Y., Courville, A., et al., 2016. Deep Learning. MIT press, Cambridge, 1 (2).
He, X., Luo, Q., Li, X., et al., 2023. Characteristics of pore differences between different lithofacies of continental mixed shale and the influencing mechanism: An example from the Permian Lucaogou Formation in the Jimusar Depression. J. China Univ. Min. Technol. 125, 13247. https://doi.org/10.13247/j.cnkij.cumt.20230152 (in Chinese).
Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. Adv. Neural Inf. Process. Syst. 33, 6840–6851. https://api.semanticscholar.org/CorpusID:219955663.
Hou, H., Shao, L., Li, Y., et al., 2022. Effect of paleoclimate and paleoenvironment on organic matter accumulation in lacustrine shale: Constraints from lithofacies and element geochemistry in the northern Qaidam Basin, NW China. J. Petrol. Sci. Eng. 208, 109350. https://doi.org/10.1016/j.petrol.2021.109350.
Jayasumana, S., Ramalingam, S., Veit, A., et al., 2024. Rethinking FID: Towards a better evaluation metric for image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9307–9315. https://doi.org/10.1109/CVPR52733.2024.00889.
Karras, T., Aittala, M., Hellsten, J., et al., 2020. Training generative adversarial networks with limited data. Adv. Neural Inf. Process. Syst. 33, 12104–12114. https://doi.org/10.48550/arXiv.2006.06676.
Kotelnikov, A., Baranchuk, D., Rubachev, I., et al., 2023. Tabddpm: Modelling tabular data with diffusion models. In: International Conference on Machine Learning, pp. 17564–17579. https://doi.org/10.48550/arXiv.2209.15421.
Kuang, L., Liu, H., Ren, Y., et al., 2021. Application and development trend of artificial intelligence in petroleum exploration and development. Petrol. Explor. Dev. 48 (1), 1–14. https://doi.org/10.1016/S1876-3804(21)60001-0.
Larson, T.E., Sivil, J.E., Periwal, P., et al., 2023. A machine-learning workflow to integrate high-resolution core-based facies into basin-scale stratigraphic models for the Wolfcamp and Third Bone Spring Sand, Delaware Basin. Interpretation 11 (4), SC91–SC104. https://doi.org/10.1190/INT-2023-0009.1.
Li, G., Lin, C., Dong, C., et al., 2022. Sequence stratigraphy, depositional environment and associated lithofacies of lacustrine shale: A case from the upper fourth member of Shahejie Formation, Dongying Depression, Bohai Bay Basin. Front. Earth Sci. 10, 906987. https://doi.org/10.3389/feart.2022.906987.
Li, G., Lin, C., Ma, P., et al., 2024a. Chemofacies characterization of lacustrine shale based on machine learning classification: A case study from the Dongying Depression, Bohai Bay Basin, China. Geoenergy Sci. Eng. 241, 213154. https://doi.org/10.1016/j.geoen.2024.213154.
Li, G., Lin, C., Wu, Y., et al., 2025. Machine learning applications in tight porous media: challenges, advances, and future directions. Earth Sci. Rev., 105306 https://doi.org/10.1016/j.earscirev.2025.105306.
Li, H., You, Y., Zeng, D., et al., 2024b. 3D geological modeling of strongly heterogeneous dual-medium deep carbonate reservoirs: A case study of Feixianguan-Changxing formations in Puguang Gas Field. Petrol. Geol. Recov. Eff. 31 (1), 44–53. https://doi.org/10.13673/j.pgre.202305033.
Liang, C., Cao, Y., Jiang, Z., et al., 2017. Shale oil potential of lacustrine black shale in the Eocene Dongying depression: Implications for geochemistry and reservoir characteristics. AAPG Bull. 101 (11), 1835–1858. https://doi.org/10.1306/01251715249.
Liang, C., Cao, Y., Wu, J., et al., 2023. Water depth–terrigenous input dynamic equilibrium controls the Eocene lacustrine shale laminae records in Jiyang Depression, Bohai Bay Basin, East China. AAPG Bull. 107 (11), 1987–2016. https://doi.org/10.1306/07052321160.
Liang, C., Wu, J., Jiang, Z., et al., 2018a. Sedimentary environmental controls on petrology and organic matter accumulation in the upper fourth member of the Shahejie Formation (Paleogene, Dongying depression, Bohai Bay Basin, China). Int. J. Coal Geol. 186, 1–13. https://doi.org/10.1016/j.coal.2017.11.016.
Liang, C., Jiang, Z., Cao, Y., et al., 2018b. Sedimentary characteristics and origin of lacustrine organic-rich shales in the salinized Eocene Dongying Depression. Bulletin 130 (1-2), 154–174. https://doi.org/10.1130/B31584.1.
Liu, B., Sun, J.H., Zhang, Y.Q., et al., 2021. Reservoir space and enrichment model of shale oil in the first member of Cretaceous Qingshankou Formation in the Changling Sag, southern Songliao Basin, NE China. Petrol. Explor. Dev. 48 (3), 608–624. https://doi.org/10.1016/S1876-3804(21)60049-6.
Liu, B., Wang, H., Fu, X., et al., 2019. Lithofacies and depositional setting of a highly prospective lacustrine shale oil succession from the Upper Cretaceous Qingshankou Formation in the Gulong Sag, northern Songliao Basin, northeast China. AAPG Bull. 103, 405–432. https://doi.org/10.1306/08031817416.
Liu, R., Zhang, L., Wang, X., et al., 2023. Application and comparison of machine learning methods for mud shale petrographic identification. Processes 11 (7), 2042. https://doi.org/10.3390/pr11072042.
Lu, G., Zeng, L., Dong, S., et al., 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, 106168. https://doi.org/10.1016/j.marpetgeo.2023.106168.
Lu, H., Liao, G., Xiao, L., et al., 2025. Conditional diffusion models with integrated porosity fusion for 3D digital rock reconstruction. In: SPWLA Annual Logging Symposium. https://doi.org/10.30632/SPWLA-2025-0094. D041S019R002.
Ma, P., Lin, C., Li, G., et al., 2022a. Lithofacies characteristics and sweet spot distribution of lacustrine shale oil: A case study from the Dongying Depression, Bohai Bay Basin, China. Lithosphere 2022 (Special 12), 3135681. https://doi.org/10.2113/2022/3135681.
Ma, P., Dong, C., Lin, C., 2022b. Petrographic and geochemical characteristics of nodular carbonate-bearing fluorapatite in the lacustrine shale of the Shahejie Formation, Dongying Depression, Bohai Bay Basin. Sediment. Geol. 439, 106218. https://doi.org/10.1016/j.sedgeo.2022.106218.
Meng, H., Lin, B., Zhang, R., et al., 2024. A missing well-logs imputation method based on conditional denoising diffusion probabilistic models. SPE J. 1–16. https://doi.org/10.2118/219452-PA.
Meng, J., Zhao, L., Xu, M., et al., 2025. Deep learning-based seismic lithofacies prediction in sparse well areas via geology-informed pseudo-well construction and transfer learning. J. Appl. Geophys., 105891 https://doi.org/10.1016/j.jappgeo.2025.105891.
Meyer, B.H., Pozo, A.T.R., Zola, W.M.N., 2022. Global and local structure preserving GPU t-SNE methods for large-scale applications. Expert Syst. Appl. 201, 116918. https://doi.org/10.1016/j.eswa.2022.116918.
Müller-Franzes, G., Niehues, J.M., Khader, F., et al., 2023. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. Sci. Rep. 13 (1), 12098. https://doi.org/10.1038/s41598-023-39278-0.
Nichol, A.Q., Dhariwal, P., 2021. Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning, pp. 8162–8171. https://doi.org/10.48550/arXiv.2102.09672.
Obi, J.C., 2023. A comparative study of several classification metrics and their performances on data. World J. Adv. Eng. Tech. Sci. 8 (1), 308–314. https://doi.org/10.30574/wjaets.2023.8.1.0054.
Qian, H., Geng, Y., Wang, H., 2024. Lithology identification based on ramified structure model using generative adversarial network for imbalanced data. Geoenergy Sci. Eng. 240, 213036. https://doi.org/10.1016/j.geoen.2024.213036.
Qiao, L., He, T., Liu, X., et al., 2025. Utilizing conditional generative adversarial networks for data augmentation in logging evaluation. Phys. Fluids 37 (3), 036607. https://doi.org/10.1063/5.0255353.
Santoso, B., Wijayanto, H., Notodiputro, K.A., et al., 2017. Synthetic over sampling methods for handling class imbalanced problems: A review. In: IOP Conference Series: Earth and Environmental Science, vol. 58, 012031. https://doi.org/10.1088/1755-1315/58/1/012031 (1).
Shi, J., Jin, Z., Liu, Q., et al., 2022. Laminar characteristics of lacustrine organic-rich shales and their significance for shale reservoir formation: A case study of the Paleogene shales in the Dongying Sag, Bohai Bay Basin, China. J. Asian Earth Sci. 223, 104976. https://doi.org/10.1016/j.jseaes.2021.104976.
Sun, L., Zhu, R., Zhang, T., et al., 2024. Advances and trends of non-marine shale sedimentology: A case study from Gulong Shale of Daqing Oilfield, Songliao Basin, NE China. Petrol. Explor. Dev. 51 (6), 1367–1385. https://doi.org/10.1016/S1876-3804(25)60547-7.
Tang, J., Fan, B., Xiao, L., et al., 2021. A new ensemble machine-learning framework for searching sweet spots in shale reservoirs. SPE J. 26 (1), 482–497. https://doi.org/10.2118/204224-PA.
Tantithamthavorn, C., McIntosh, S., Hassan, A.E., et al., 2016. An empirical comparison of model validation techniques for defect prediction models. IEEE Trans. Software Eng. 43 (1), 1–18. https://doi.org/10.1109/TSE.2016.2584050.
Tharwat, A., 2021. Classification assessment methods. Appl. Comput. Inform. 17 (1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003.
Trabucco, B., Doherty, K., Gurinas, M., et al., 2023. Effective Data Augmentation with Diffusion Models. ArXiv preprint arXiv:2302.07944. https://doi.org/10.48550/arXiv.2302.07944.
Villaizán-Vallelado, M., Salvatori, M., Segura, C., et al., 2025. Diffusion models for tabular data imputation and synthetic data generation. ACM Trans. Knowl. Discov. Data 19 (6), 1–32. https://doi.org/10.1145/3742435.
Wang, G., Carr, T.R., 2012. Methodology of organic-rich shale lithofacies identification and prediction: A case study from Marcellus Shale in the Appalachian basin. Comput. Geosci. 49, 151–163. https://doi.org/10.1016/j.cageo.2012.07.011.
Wang, M., Yang, J., Wang, X., et al., 2023. Identification of shale lithofacies by well logs based on random forest algorithm. Earth Sci. 48 (1), 130–142. https://doi.org/10.3799/dqkx.2022.181 (in Chinese).
Wang, X., Zhu, X., Lai, J., et al., 2024. Paleoenvironmental reconstruction and organic matter accumulation of the paleogene shahejie oil shale in the Zhanhua Sag, Bohai Bay Basin, Eastern China. Pet. Sci. 21 (3), 1552–1568. https://doi.org/10.1016/j.petsci.2024.03.001.
Wang, Y., Xu, S., Hao, F., et al., 2025. Machine learning-based grayscale analyses for lithofacies identification of the Shahejie Formation, Bohai Bay Basin, China. Pet. Sci. 22 (1), 42–54. https://doi.org/10.1016/j.petsci.2024.07.021.
Wu, J., Jiang, Z., 2017. Division and characteristics of shale parasequences in the upper fourth member of the Shahejie Formation, Dongying Depression, Bohai Bay Basin, China. J. Earth Sci. 28 (6), 1006–1019. https://doi.org/10.1007/s12583-016-0943-6.
Xu, M., Song, S., Mukerji, T., 2024. DiffSim: Denoising diffusion probabilistic models for generative facies geomodeling. In: Fourth International Meeting for Applied Geoscience & Energy, pp. 1660–1664. https://doi.org/10.1190/image2024-4081304.1.

Xue, C., McBeck, J.A., Lu, H., et al., 2024. Classification of shale lithofacies with minimal data: application to the early Permian shales in the Ordos Basin, China. J. Asian Earth Sci. 259, 105901. https://doi.org/10.1016/j.jseaes.2023.105901.

Yang, J., Ma, J., 2015. A big-data processing framework for uncertainties in transportation data. In: 2015 IEEE International Conference on Fuzzy Systems. FUZZ-IEEE, pp. 1–6. https://doi.org/10.1109/FUZZ-IEEE.2015.7337843.

Yu, Y., Zhang, W., Deng, Y., 2021. Frechet Inception Distance (FID) for Evaluating Gans. China University of Mining Technology Beijing Graduate School (in Chinese).

Zha, W., Li, X., Xing, Y., et al., 2020. Reconstruction of shale image based on wasserstein generative adversarial networks with gradient penalty. Adv. Geo-Energy Res. 4 (1), 107–114. https://doi.org/10.46690/ager.2020.01.10.

Zhang, J., Jiang, Z., Liang, C., et al., 2016. Lacustrine massive mudrock in the Eocene Jiyang Depression, Bohai Bay Basin, China: Nature, origin and significance. Mar. Petrol. Geol. 77, 1042–1055. https://doi.org/10.1016/j.marpetgeo.2016.08.008.

Zhao, F., Han, Z., Fu, X., et al., 2025. LogDiffusion: A method of lithology identification based on diffusion probability model. Prog. Geophys. 40 (1), 106–120. https://doi.org/10.6038/pg2025II0075.

Zhao, F., Yang, Y., Kang, J., et al., 2023. CE-SGAN: Classification enhancement semi-supervised generative adversarial network for lithology identification. Geoenergy Sci. Eng. 223, 211562. https://doi.org/10.1016/j.geoen.2023.211562.

Zhao, P., Qin, R., Pan, H., et al., 2019. Study on array laterolog response simulation and mud-filtrate invasion correction. Adv. Geo-Energy Res. 3 (2), 175–186. https://doi.org/10.26804/ager.2019.02.07.

Zhao, T., Wang, S., Ouyang, C., et al., 2024. Artificial intelligence for geoscience: progress, challenges, and perspectives. Innovation 5 (5), 100691. https://doi.org/10.1016/j.xinn.2024.100691.

Zhao, Z., Dong, C., Ma, P., et al., 2022. Origin of dolomite in lacustrine organic-rich shale: a case study in the Shahejie Formation of the Dongying Sag, Bohai Bay Basin. Front. Earth Sci. 10, 909107. https://doi.org/10.3389/feart.2022.909107.

Zhou, H., Wang, F., Tao, P., 2018. t-Distributed stochastic neighbor embedding method with the least information loss for macromolecular simulations. J. Chem. Theor. Comput. 14 (11), 5499–5510. https://doi.org/10.1021/acs.jctc.8b00652.