257 (2026) 214266

Contents lists available at ScienceDirect

Geoenergy Science and Engineering

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

Impact of sample size and pore structure on machine learning prediction of petrophysical properties in low-permeability sandstone reservoirs

Yuxi Sun [a], Liang Chen [b], Yuan Qi [a], Yiping He [c], Hancheng Ji [a,*], Yanqing Shi [a], Shuangqi Feng [a]

[a] College of Geosciences, China University of Petroleum, Beijing, 102249, China
[b] College of Science, China University of Petroleum, Beijing, 102249, China
[c] No. 5 Oil Production Plant, Changqing Oilfield Company, PetroChina, Xi'an, 710200, Shaanxi, China

ARTICLE INFO

Keywords:
Machine learning
Pore structure
Porosity prediction
Permeability prediction
Low-permeability sandstone
Tight sandstone

ABSTRACT

Data-driven machine learning (ML) methods have demonstrated considerable advantages in predicting reservoir rock properties. However, the high cost of data collection and processing means that systematic studies on the optimal training sample size for ML remain scarce. Meanwhile, the strong heterogeneity and complex pore structures in low-permeability sandstone reservoirs often limit predictive accuracy and generalization ability. In this study, we utilized approximately 10,000 sets of core petrophysical data and well logs from 64 wells in the Triassic Yanchang Formation of the Ordos Basin. We evaluated the effects of different ML methods, sample sizes, and pore structures on the performance of porosity and permeability prediction models. The results indicate that the ability to model nonlinear relationships is particularly important and that incorporating discrete geological variables further enhances model performance. When trained on small sample sizes, models tend to underfit and exhibit substantial predictive uncertainties. As the sample size increases, performance steadily improves but exhibits a saturation trend. For highly heterogeneous pore systems, conventional well logs cannot effectively capture differences in pore structure, implying that larger training datasets or supplementary geological constraints are necessary to reveal underlying patterns. Through a sample size scaling experiment, we revealed a power-law decay relationship between sample size and model performance. This finding suggests that effective prediction does not require an excessively large dataset. It also provides a quantitative basis for the economical and efficient development of low-permeability sandstone reservoirs.

1. Introduction

Low-permeability sandstone reservoirs have become key targets for the development of unconventional oil, gas, and other subsurface energy resources. (Zou et al., 2018; Lai et al., 2018a, 2022; Feng et al., 2019). Such reservoirs typically feature complex pore structures and poor pore connectivity, which pose significant challenges for quantitative reservoir evaluation (Lai et al., 2018b; Sharifi et al., 2023; Xu et al., 2025). Porosity and permeability are fundamental physical parameters in reservoir evaluation, as they jointly determine the storage capacity and fluid mobility of the reservoir (Pittman, 1992; Katz and Arango, 2018). Currently, core analysis and well logs are widely used in the oil and gas industry to obtain these parameters (Pittman, 1992; Zhao et al., 2022; He et al., 2025). Core measurements are accurate and reliable but are also time-consuming and expensive. In contrast, estimating porosity and permeability from well logs reduces costs and improves efficiency to a certain extent (Wood, 2020; Fang et al., 2024). However, such predictions often suffer from limited accuracy owing to model simplifications and the choice of input well logs.

In recent years, with rapid advances in data science and artificial intelligence (AI), machine learning (ML) methods have been widely applied in geosciences and petroleum engineering (Baraboshkin et al., 2020; Song et al., 2022; Hu et al., 2023; Jiang et al., 2024b; Bione et al., 2024; Lai et al., 2024). Compared with traditional methods, ML is more effective at uncovering complex nonlinear relationships and provides data-driven solutions. This provides new approaches for reservoir property characterization based on well logs (Zhao et al., 2022; Koray et al., 2024; He et al., 2025). Common supervised learning algorithms, such as k-nearest neighbors (KNN), support vector machines (SVM), extreme gradient boosting (XGBoost), and deep neural networks (DNN), have been applied to predict petrophysical parameters (Wood, 2020; Iraji et al., 2023; Ehsan et al., 2024; Shehata et al., 2025). Meanwhile,

* Corresponding author. 18 Fuxue Road, Changping District, 102249 Beijing, China.
E-mail address: jhch@cup.edu.cn (H. Ji).

https://doi.org/10.1016/j.geoen.2025.214266
Received 2 July 2025; Received in revised form 12 September 2025; Accepted 28 October 2025
Available online 30 October 2025

unsupervised learning methods such as hierarchical clustering (HC) and k-means clustering are frequently used for lithofacies classification and reservoir evaluation (Maldar et al., 2022; Koray et al., 2024; Vera-Arroyo and Bedle, 2025).

Despite these advancements, applying ML to low-permeability sandstone reservoirs still presents several challenges. First, due to the high cost of core property measurements (Zhao et al., 2024; Abid et al., 2025), most studies are limited by small sample sizes. For example, Zhang et al. (2021) used 253 core samples from five wells for training, while Zhao et al. (2022) used 347 samples from three wells. Although these studies yielded favorable prediction results, they often did not assess model generalizability when applied to larger, unseen datasets. Second, low-permeability reservoirs typically exhibit multiscale pore systems (Qiao et al., 2020), whereas conventional well logs can only indirectly reflect pore structures (Lai et al., 2023; Abid et al., 2025). It remains unclear whether the overlooked microscopic pore features limit the prediction accuracy of ML models.

The low-permeability sandstone reservoirs of the Triassic Yanchang Formation in the Ordos Basin contain abundant oil and gas resources (Yao et al., 2013). Currently, porosity and permeability predictions in this region rely mainly on traditional linear regression models (Fang et al., 2024). In this study, we collected approximately 10,000 core property data points from the region. For the first time in low-permeability sandstone reservoirs, we conducted a large-scale sample-size scaling experiment to quantify the effects of sample size and pore structure on the performance of prediction models. The results show that: (1) nonlinear modeling capability is of great significance for predicting low-permeability sandstone reservoirs; (2) incorporating discrete geological variables can further enhance model performance; (3) the relationship between sample size and model performance follows a power-law decay; and (4) for highly tight and heterogeneous reservoirs, more training samples are needed to reveal underlying patterns. These findings provide strong support for implementing data-driven approaches in the characterization of low-permeability sandstone reservoirs.

2. Geological setting

The Ordos Basin is located in the inland region of central-western China and constitutes a large Mesozoic oil- and gas-bearing basin in the western part of the North China Craton. It is underlain by an Archean-Lower Proterozoic crystalline basement and covers a total area of approximately 3.2×10^5 km^2 (Fig. 1a) (Li et al., 2021; Bai et al., 2022). During its geological history, the basin has undergone multiple orogenic events (including the Luliang, Indosinian, Yanshan, and Himalayan movements), resulting in an asymmetric structural configuration that extends generally north-south, with gentle dips in the east and steep dips in the west (Zhao et al., 2024). Based on structural style and basement characteristics, the interior of the basin is divided into six tectonic units: the Yimeng Uplift, the Western Thrust Zone, the Tianhuan Depression, the Yishan Slope, the Jinxi Flexure Zones, and the Weibei Uplift (Qin et al., 2025). To date, the Ordos Basin has experienced four major tectono-sedimentary evolutionary stages: Early Paleozoic shallow-marine platform stage, Late Paleozoic basin-margin rifting stage, Mesozoic intra-continental depression stage, and Cenozoic basin-margin fault-depression subsidence stage (Guo et al., 2018; Lu et al., 2024).

Influenced by the Indosinian orogeny, the basin gradually transformed from a Paleozoic cratonic basin into a Mesozoic depression basin, with sedimentary environments shifting from marine to continental settings (Fu et al., 2020; Lin et al., 2024). During the sedimentation of the Triassic Yanchang Formation, a terrigenous clastic system dominated by fluvial-deltaic-lacustrine facies developed across the basin interior (Wang et al., 2024). Based on sedimentary evolution

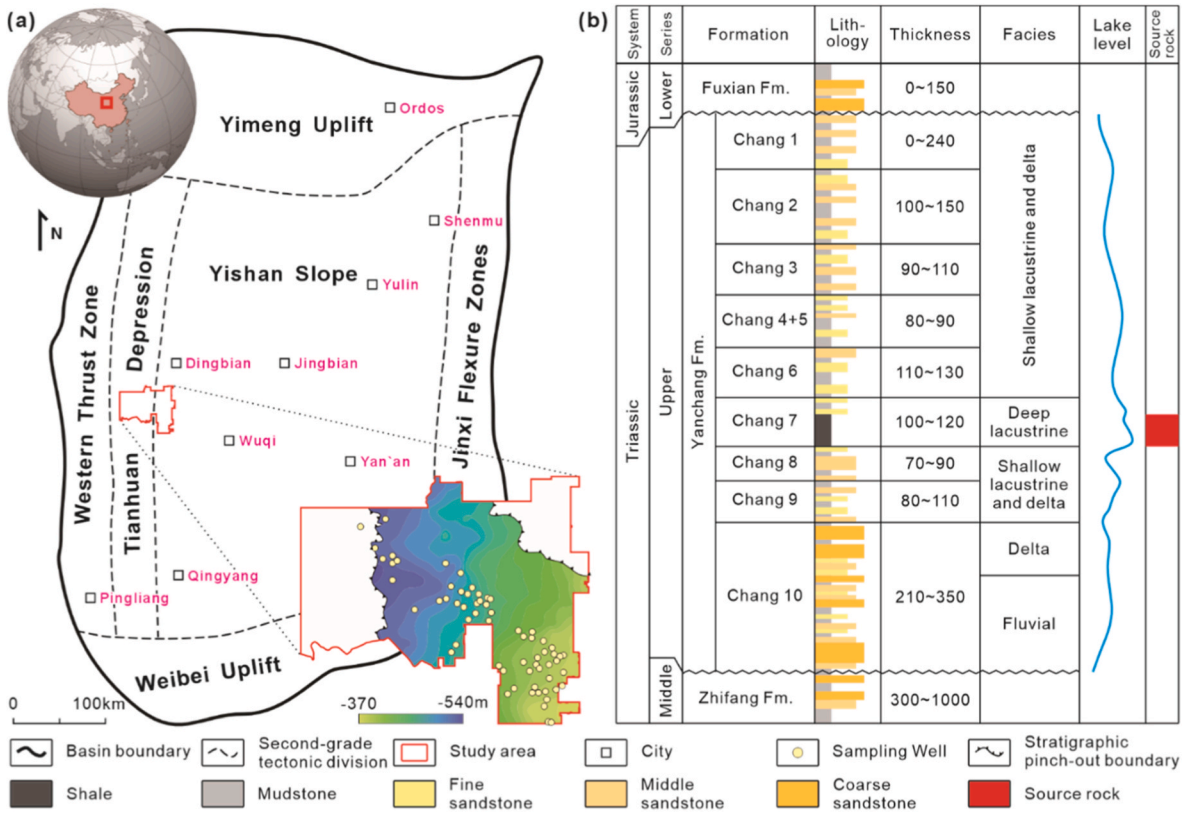

Fig. 1. (a) Overview of the Ordos Basin. The red box highlights the Jiyuan area, showing the top surface structure of the Chang 2 and sampling well locations; (b) Generalized stratigraphic column of the Upper Triassic Yanchang Formation. It includes system, series, formation, member, lithology, thickness, sedimentary facies, lake level, and source rock. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

characteristics, the Yanchang Formation can be subdivided from bottom to top into ten members, ranging from Chang 10 at the base to Chang 1 at the top (Fig. 1b). These members document the full depositional history of the basin, from rapid subsidence and lake formation to gradual shrinkage and eventual termination. (Liu et al., 2025). The Chang 7 represents the maximum lake-transgression phase. During this stage, the lake reached its greatest extent, and widespread sedimentation of dark mudstones and oil shales occurred in deep-lacustrine settings. This member constitutes the most important source rock of the Yanchang Formation (Lu et al., 2022). At the end of the Triassic, a tectonic uplift caused differential rise of the basin, resulting in varying degrees of weathering and fluvial erosion at the top of the Yanchang Formation (Deng et al., 2020).

The Jiyuan area is situated in the central-western Ordos Basin, within the transitional zone between the Tianhuan Depression and the Yishan Slope (Fig. 1a). This area is underlain by two major oil-bearing stratigraphic intervals: the Triassic Yanchang Formation and the Jurassic Yan'an Formation. It is one of the key hydrocarbon accumulation zones within the basin (Li et al., 2017). In this area, the reservoir rocks are primarily fine-grained arkose and lithic arkose (Yang et al., 2020). The sand bodies in these reservoirs are mainly of deltaic distributary-channel or fluvial channel origin (Wang et al., 2017). The reservoir porosity is dominated by secondary dissolution pores and residual primary pores, and the microscopic pore structure is characterized by small pore-throat radii and poor connectivity (Zhang et al., 2025). Due to the combined influence of high clay content and strong pore-structure heterogeneity, the overall reservoir exhibits the typical low-porosity, low-permeability characteristics of tight sandstones (Yang et al., 2020; Zhang et al., 2025).

3. Data and methods

3.1. Dataset

The dataset used in this study was collected from 64 wells in the Jiyuan area of the Ordos Basin (Fig. 1a). It comprises stratigraphic member codes (e.g., Chang 1 and Chang 2) encoded as integers and seven types of well logs as input features, along with 9,845 sets of core petrophysical measurements as target variables. In addition, 112 high-pressure mercury injection (HPMI) tests were used for pore structure analysis. The detailed data types are listed in Table 1. All data were provided by the No. 5 Oil Production Plant of PetroChina Changqing Oilfield Company. As shown in Fig. 2, the overall study was structured into a workflow for reference.

3.2. Data processing

3.2.1. Gaussian smoothing

Well logs often contain high-frequency noise fluctuations. To enhance the stability of subsequent analysis and modeling, Gaussian smoothing was applied to the raw well logs (Fig. 3a). This method effectively reduces noise interference (Aftab and Moghadam, 2022) and helps mitigate the risk of model overfitting.

Table 1
Overview of the dataset.

Data category	Quantity	Data type/Parameters
Well logs	64 wells	Acoustic (AC), Caliper (CAL), Compensated neutron (CNL), Density (DEN), Gamma ray (GR), Deep lateral resistivity (RILD), Spontaneous potential (SP)
Stratigraphic members	64 wells	Chang 1, Chang 2, Chang 3, Chang 4 + 5, Chang 6, Chang 7, Chang 8, Chang 9
Petrophysical samples	9,845 cores	Porosity(Por), Permeability(Perm)
HPMI samples	112 cores	Median pore-throat diameter(D_M), Median displacement pressure(P_{50}), Displacement pressure(P_d), Maximum mercury saturation (S_{Hg}), Sorting coefficient(S_p)

3.2.2. Core-to-log calibration

To accurately integrate well logs with core data, it was necessary to correct depth discrepancies between core samples and well logs. Laboratory-measured core density was manually aligned with the corresponding density log (Fig. 3a). This correction ensured reliable analyses of porosity, permeability, and pore structure characteristics.

3.2.3. Log_{10} transformation

Permeability and deep lateral resistivity (RILD) typically exhibit lognormal distributions with significant right skewness and extreme values. To normalize the data distribution and reduce the influence of outliers, a base-10 logarithmic (log_{10}) transformation was applied. This preprocessing step enhanced both model accuracy and interpretability.

3.2.4. Outlier detection and removal

Well logs often contain outliers caused by equipment errors or borehole conditions. To address this issue, Robust Principal Component Analysis (Robust PCA) was used to decompose the well logs matrix into low-rank and sparse components (Yaniv and Beck, 2024). This method effectively identifies and removes significant outliers while preserving essential data trends (Fig. 3b).

3.2.5. Standardization

To eliminate the influence of scale differences among features, the data were standardized. This procedure transformed all features to a distribution with zero mean and unit variance. This ensured their comparability on a common scale during model training and enhanced the stability of the training process.

3.2.6. Dataset split

The dataset was split into training (90 %) and testing (10 %) sets. The testing set was kept fixed within each experiment to ensure a fair comparison of model performance under different models or sample sizes. Additionally, 5-fold cross-validation was conducted on the training set to optimize model parameters and improve the stability of performance evaluation.

3.3. Model construction and evaluation

We selected eight widely used ML methods: multiple linear regression (MR), ridge regression (Ridge), support vector machine (SVM), k-nearest neighbors (KNN), random forest (RF), extreme gradient boosting (XGBoost), light gradient boosting machine (LGBM), and deep neural networks (DNN) (Hoerl and Kennard, 1970; Coker, 1995; Breiman, 2001; Noble, 2006; LeCun et al., 2015; Chen and Guestrin, 2016; Ke et al., 2017; Taunk et al., 2019). The DNN was implemented using *PyTorch*, while the other models were implemented with *scikit-learn* in *Python 3.12*. Hyperparameter optimization for all models was conducted using *Optuna* to ensure optimal performance (Mao et al., 2025). Further details of the modeling procedures are provided in Supplementary Text S1 and Table S1.

To comprehensively evaluate model accuracy, stability, and fit, we used the coefficient of determination (R^2), root mean squared error (RMSE), mean squared error (MSE), and mean absolute error (MAE).

3.4. Experiment design

The experimental design comprised three parts: (1) comparing the performance of different ML models in predicting porosity and permeability to identify the optimal model; (2) evaluating the impact of sample size on prediction accuracy using the optimal model. Sample sizes were increased in gradients (10, 20, 40, 80, 160, etc.), with each size randomly sampled five times to ensure statistical stability; and (3)

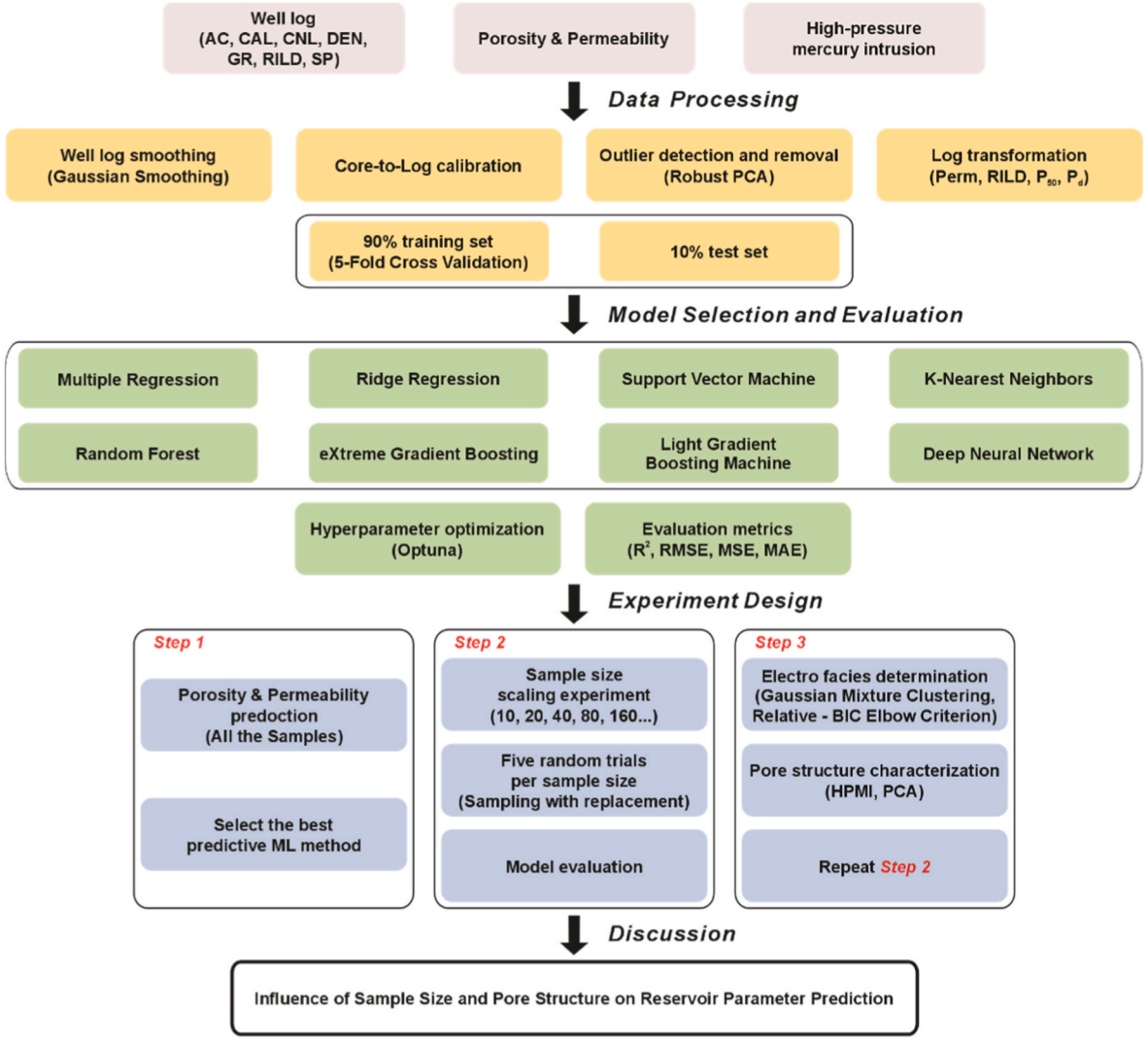

Fig. 2. Workflow of the study, including data processing, model selection and evaluation, experimental design, and discussion.

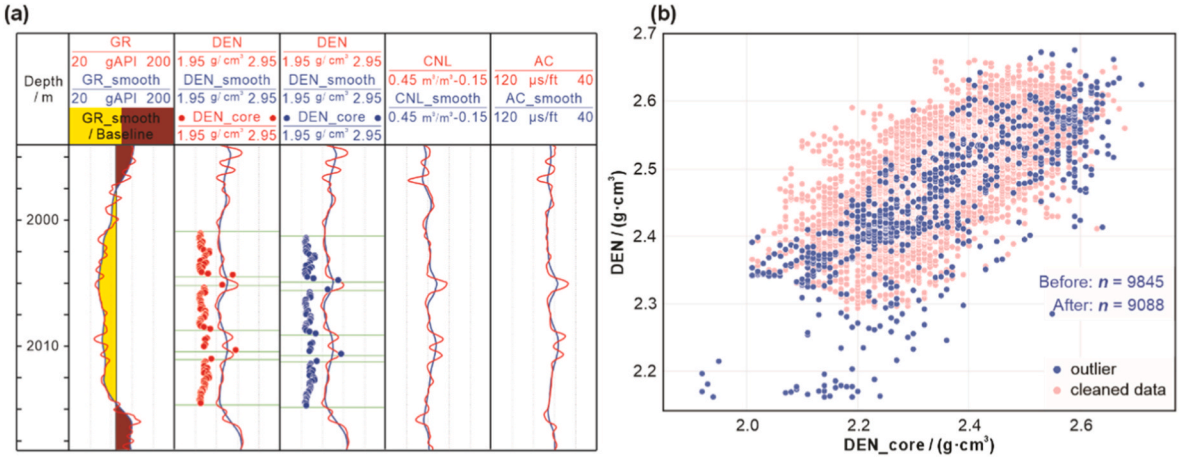

Fig. 3. (a) Effects of gaussian smoothing and core-to-log calibration; (b) Changes in the dataset after outlier detection and removal.

analyzing the impact of pore structures on model predictions. A Gaussian Mixture Model (GMM) was used to cluster the well logs, and principal component analysis (PCA) was applied to reduce the dimensionality of the HPMI data. Geological information of different electrical clusters was summarized, and sample size scaling experiments were conducted.

3.4.1. Gaussian mixture clustering

GMM assumes that the data consist of multiple Gaussian distributions. It employs the expectation-maximization algorithm to estimate parameters and perform soft clustering (Jiao et al., 2022; Jiang et al., 2024a). Given that well logs often exhibit multimodal and non-spherical characteristics, GMM effectively captures transitions between formations or lithologies and quantifies clustering uncertainty.

3.4.2. Relative-BIC elbow criterion

We applied the Relative-BIC Elbow Criterion, based on the relative decline of the BIC curve, to objectively determine the optimal number of GMM clusters. First, the seven types of well logs were standardized using Z-score normalization and reduced to four principal components via PCA, retaining 90 % cumulative variance. This step reduced model complexity and multicollinearity. GMM models were then fitted for K = 1 to K_{max}, and the ΔBIC was calculated (Eq (1)):

$$\Delta BIC(k) = BIC(k\text{-}1)\text{-}BIC(k) \quad (1)$$

The optimal cluster number (K) was identified at the first point where ΔBIC dropped below 1 % of the initial BIC. This method avoids overfitting by stopping the monotonic decrease of BIC, retaining only components with significant informational gain, particularly useful with highly correlated features or large sample sizes.

4. Results

4.1. Overview of petrophysical properties

Based on the feature-engineered dataset (Fig. 4a), sandstone porosity in the Yanchang Formation of the study area mainly ranges from 10 % to 15 %, with permeability predominantly between 0.1 and 1 mD. These characteristics define typical low-permeability sandstone reservoirs. Overall, porosity and permeability exhibit a strong positive correlation (R^2 = 0.703), indicating higher permeability in samples with greater porosity. However, the data points show noticeable dispersion. This implies considerable heterogeneity in reservoir properties, as permeabilities significantly vary even at similar porosity levels.

Different well logs exhibit distinct correlations with reservoir properties (Fig. 4b). DEN shows the strongest negative correlation, indicating lower porosity and permeability with higher densities, consistent with rock physics principles. AC, which reflects acoustic wave propagation, correlates positively with porosity, effectively indicating pore development. RILD and SP are influenced by lithology, fluid types, and formation water chemistry, but they still partially reflect reservoir characteristics.

Fig. 4c and d show the distributions of core porosity and permeability in different members (Chang 1 to Chang 9) of the Yanchang Formation. The analysis reveals that reservoirs in the Chang 1 to Chang 3 exhibit relatively higher porosity and permeability. These members are mainly associated with delta plain or shallow-water sedimentary environments. Conversely, the Chang 6 to Chang 8 mainly represent subaqueous fans or deep-water sedimentary environments. These members are characterized by fine-grained sediments. Porosity and permeability in the remaining members are intermediate. The observed vertical variation in these properties closely correlates with sedimentary environments and diagenetic evolution.

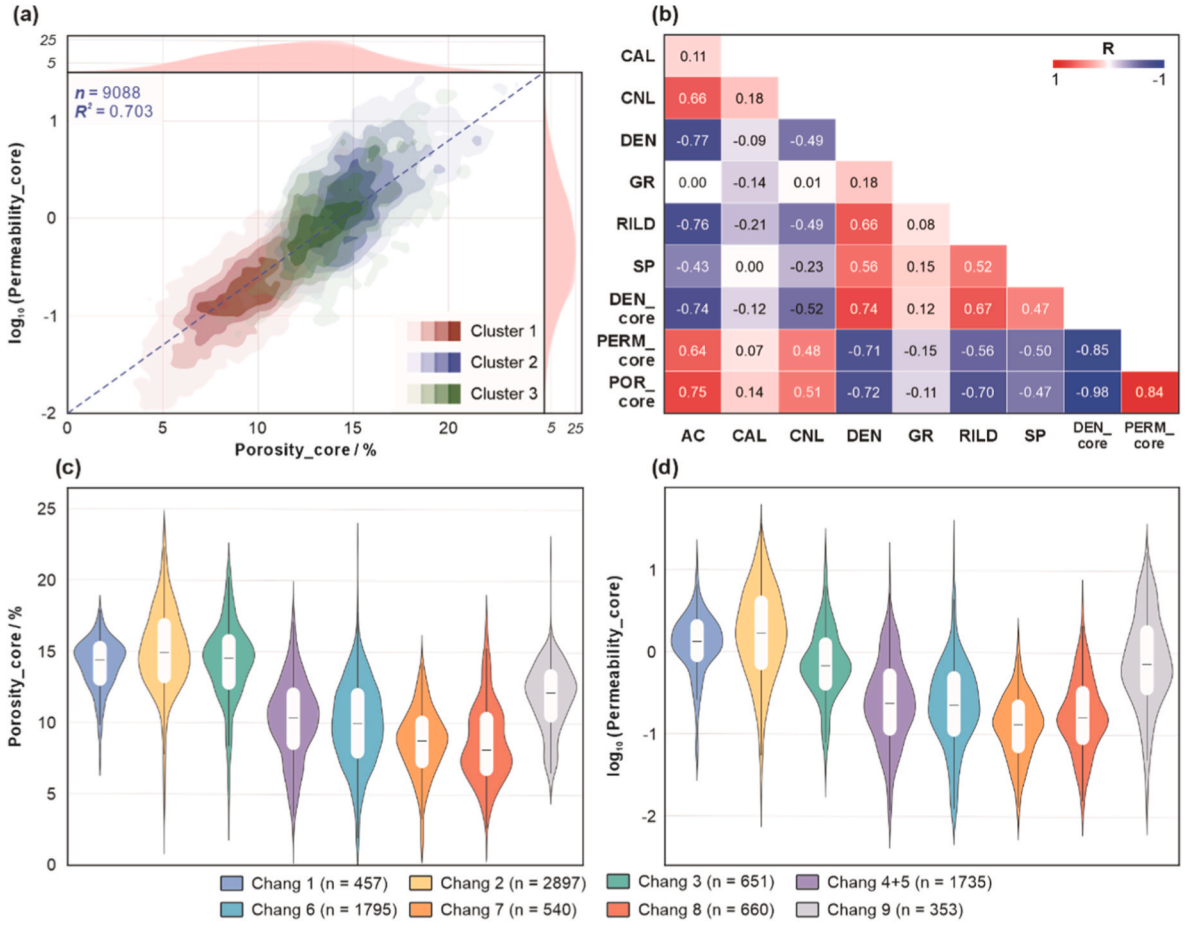

Fig. 4. (a) Crossplot showing the correlation between core porosity and permeability; (b) Cross-correlation matrix of variables in the dataset; (c) and (d) Distribution differences in porosity and permeability across different members of the Yanchang Formation.

4.2. Performance comparison of multiple ML methods

In conventional log-based reservoir evaluation, linear regression models are widely used. They typically achieve good prediction results in reservoirs with homogeneous lithology and low diagenesis levels (Chen et al., 2021; Safavi and Maldar, 2024). However, our results indicate that two linear regression models (MR and Ridge) yield relatively high prediction errors (Fig. 5a and b; Table 2). Their results were almost identical, with RMSE values of 2.13 for porosity and 0.42 for permeability on the test set. The corresponding R^2 values are 0.68 and 0.57, respectively. Scatter plots of predicted versus actual values (Figure S1 and Figure S2) show significant dispersion, deviating considerably from the diagonal line. Permeability even exhibits a more diffuse scatter pattern. These findings suggest that linear models struggle to capture the complexity of low-permeability sandstone reservoirs, highlighting the importance of nonlinear modeling capabilities.

To address this, we further employed additional ML models: SVM, KNN, RF, XGBoost, LGBM and DNN. For porosity prediction (Fig. 5a), the tree-based ensemble models (XGBoost, RF, and LGBM) achieved superior performance, with R^2 values of 0.840, 0.835, and 0.830, respectively. The DNN model performed slightly less effectively but remained close to the ensemble models, while the relatively simple SVM and KNN models formed a second tier of performance. For permeability prediction (Fig. 5b), the overall performance was lower than for porosity prediction, but the three ensemble models still achieved the best results.

Overall, these machine learning models demonstrated comparable predictive performance, characterized by lower errors and higher goodness of fit (Table 2). Most scatter points align closely along the diagonal, showing a clear improvement over the linear models (Figure S1 and Figure S2). These results indicate that ML models can more effectively capture the nonlinear relationships between well logs and reservoir properties. Considering both performance and computational cost, XGBoost was selected for prediction in the subsequent analysis.

4.3. Impact of sample size on model performance

Sample size is a crucial factor influencing the performance and generalization ability of ML models (Ma et al., 2024). To investigate the impact of sample size on the accuracy of reservoir property predictions, we conducted a scaling experiment. Ten sample sizes were tested: 10, 20, 40, 80, 160, 320, 640, 1280, 2560, and 5120. For each level, five rounds of random sampling with replacement were performed, and both porosity and permeability were predicted.

The results show a steady improvement in prediction accuracy as the training sample size increased from 10 to 5120 (Figs. 6a and 7a; Tables S2 and S3). Specifically, RMSE consistently decreased, while R^2 increased with larger sample sizes. Scatter plots of predicted versus actual values further confirmed this trend (Fig. 6b–k and 7b–k). At smaller sample sizes, the scatter points were widely dispersed and deviated significantly from the diagonal line, indicating underfitting. As the sample size grew, the points gradually converged toward the diagonal. These findings demonstrate that with sufficient training samples, the model achieves higher accuracy and becomes more robust to random sampling variability, resulting in more stable and reliable predictions.

4.4. Impact of pore structure on model performance

4.4.1. Electrical-based clustering characteristics

To evaluate the impact of different pore structures on model performance, we used the Relative-BIC Elbow Criterion to determine the optimal number of clusters. As shown in Fig. 8a, the ΔBIC first dropped below the relative threshold of 1 %$|BIC_1|$ at K = 3, indicating that adding more clusters beyond this point yielded less than 1 % gain. Based on this, the dataset was clustered into three electrical facies (Cluster 1, Cluster 2, Cluster 3) using a GMM. These clusters show significant vertical distribution differences (Fig. 8b). Cluster 1 is mainly located in dense lower and middle members of the Yanchang Formation. Cluster 2 appears predominantly in the upper members with relatively better reservoir quality. Cluster 3 is mostly in transitional middle members. This distribution suggests different sedimentary facies or diagenetic histories for each cluster.

To further investigate the pore structure differences among the three clusters, we performed PCA on the pore structure parameters obtained from HPMI. In the PCA projection (Fig. 8c), Cluster 1 is primarily distributed in the second and third quadrants, while Clusters 2 and 3 occupy the first and fourth quadrants. Based on the loading coefficient plot (Fig. 8d), Cluster 1 is characterized by small pore-throat sizes, poor connectivity, and high displacement pressure. Cluster 2 shows higher porosity, permeability, and larger pore-throat radii, while Cluster 3 exhibits intermediate pore structure characteristics.

Based on reservoir properties, well logs, and pore structure parameters (Figs. 4a and 9; Table 3), the characteristics of the three electrical clusters are summarized as follows:

Cluster 1 (Poor-Quality Tight Sandstone Facies): Porosity mostly below 11 %, permeability less than 1 mD, indicating typical tight reservoirs. Logging responses indicate higher clay content and tighter lithology. The Rock Quality Index (RQI) is extremely low, while both median displacement pressure (P_{50}) and displacement pressure (P_d) are the highest among the clusters, indicating the dominance of small pore throats. The maximum mercury saturation (S_{Hg}) is relatively low, indicating poor pore connectivity and a proportion of unfillable pores.

Cluster 2 (High-Quality Reservoir Facies): The average porosity is 15.13 %, and the average permeability is 1.68 mD, indicating a low-permeability yet high-quality reservoir. GR suggest clean sandstones, while porosity-related logs (AC, CNL, DEN) indicate well-developed porosity. The RQI is relatively high, and both P_{50} and P_d suggest dominance of large pore throats. High S_{Hg} values reflect good pore

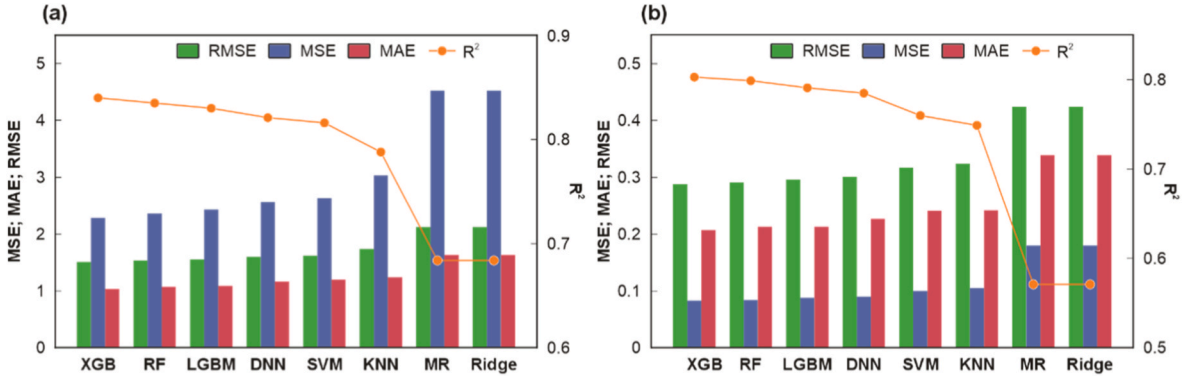

Fig. 5. Performance of different ML methods for porosity (a) and permeability(b) prediction, including MSE, MAE, RMSE, and R^2.

Table 2
Performance metrics of different ML methods for porosity and permeability prediction.

Model	Porosity				Permeability			
	R^2	RMSE	MSE	MAE	R^2	RMSE	MSE	MAE
MR	0.684	2.128	4.527	1.639	0.571	0.424	0.180	0.339
Ridge	0.684	2.128	4.527	1.639	0.571	0.424	0.180	0.339
SVM	0.816	1.624	2.638	1.203	0.760	0.317	0.100	0.241
KNN	0.788	1.743	3.038	1.245	0.749	0.324	0.105	0.242
RF	0.835	1.539	2.368	1.076	0.799	0.291	0.084	0.213
XGB	0.840	1.514	2.291	1.040	0.803	0.288	0.083	0.207
LGBM	0.830	1.561	2.436	1.094	0.791	0.296	0.088	0.213
DNN	0.821	1.603	2.569	1.171	0.785	0.301	0.090	0.227

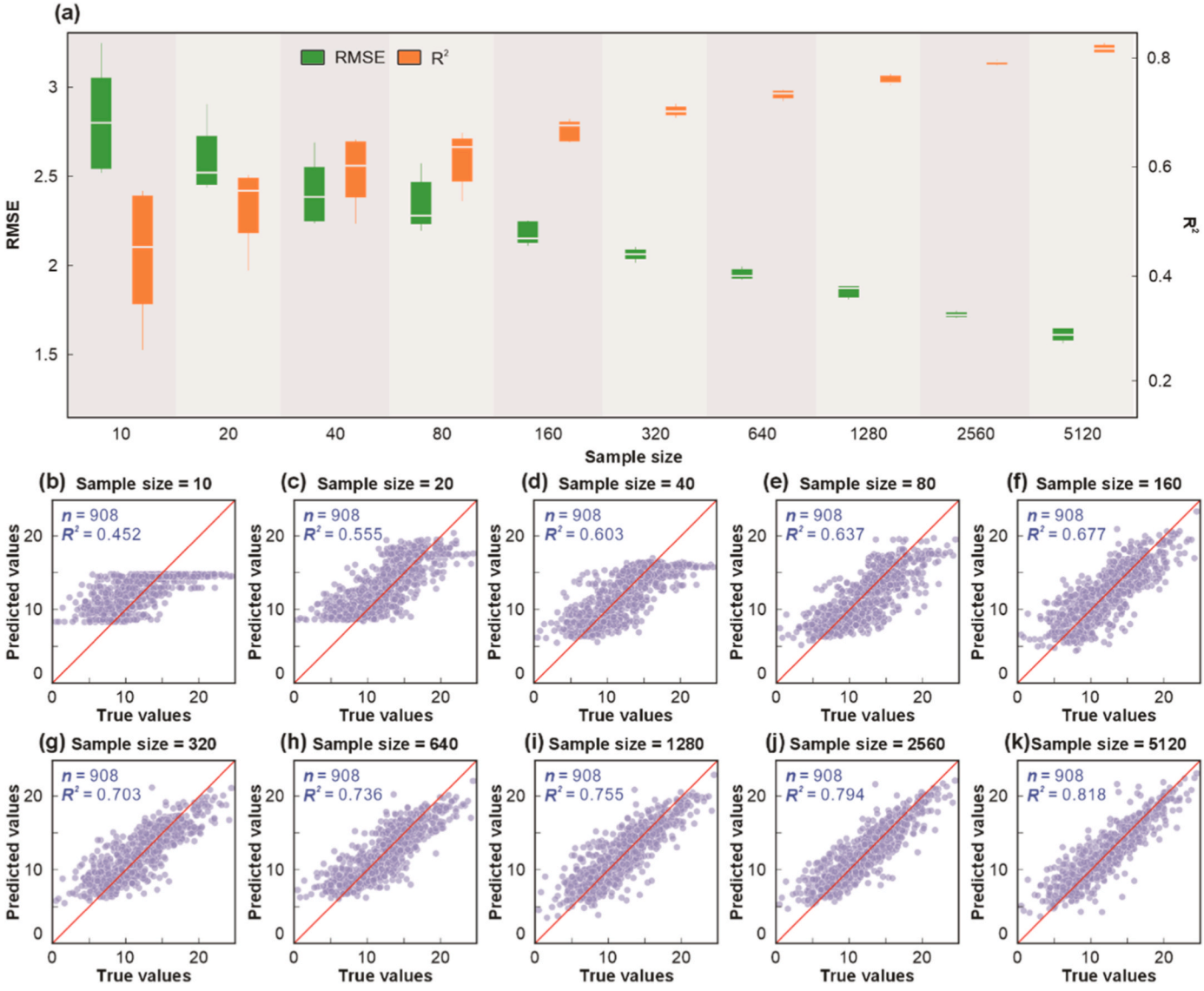

Fig. 6. Performance of the XGBoost model in porosity prediction under different training sample sizes. (a) Evaluation metrics including RMSE and R^2; (b)–(k) Scatter plots of predicted vs. actual values for sample sizes of 10, 20, 40, 80, 160, 320, 640, 1280, 2560, and 5120.

connectivity.

Cluster 3 (Transitional Reservoir Facies): This cluster has an average porosity of 13.74 % and a permeability of 0.89 mD. Logging responses display intermediate characteristics between Clusters 1 and 2. The RQI indicates moderate reservoir quality. S_{Hg} is slightly lower than that of Cluster 2, suggesting that some micropores remain unfilled. Overall, Cluster 3 exhibits significantly better reservoir quality than Cluster 1, though slightly inferior to Cluster 2.

4.4.2. *Model performance of different clusters*

Fig. 10 and Fig. S3 show the performance of ML models trained on the three electrical clusters, based on the scaling experiment (sample sizes ranging from 10 to 2560). For porosity prediction, all clusters exhibit a positive correlation between model performance and training

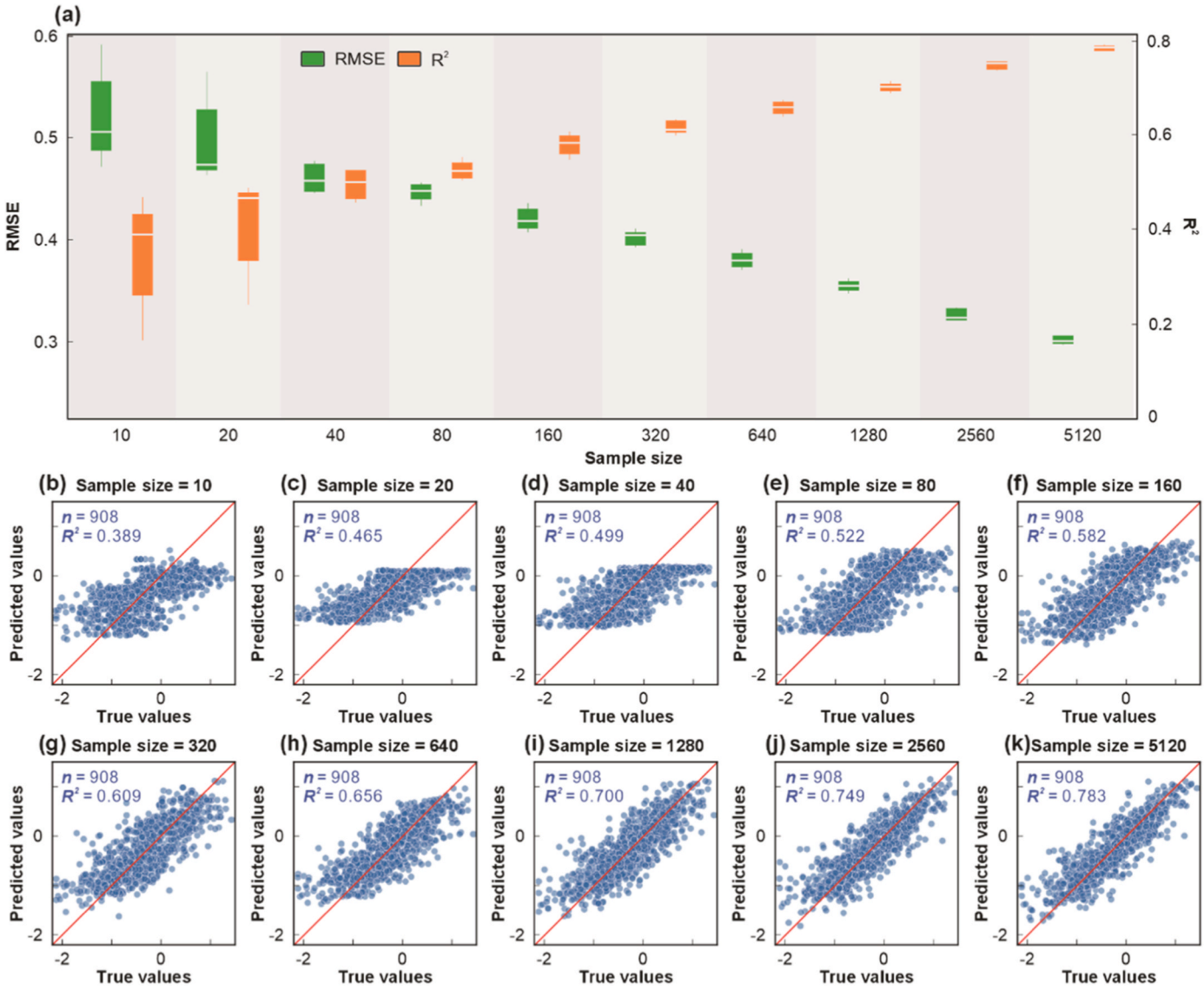

Fig. 7. Performance of the XGBoost model in permeability prediction under different training sample sizes. (a) Evaluation metrics including RMSE and R^2; (b)–(k) Scatter plots of predicted vs. actual values for sample sizes of 10, 20, 40, 80, 160, 320, 640, 1280, 2560, and 5120.

sample size (Fig. 10a). Cluster 1 showed minimal learning capability with small sample sizes. Even with all 2560 samples, the average R^2 and RMSE were only 0.598 and 1.822, respectively. In contrast, when using all 2316 samples, Cluster 2 achieved an average R^2 of 0.665 and an RMSE of 1.534. Cluster 3 performed poorly at 10 samples but showed higher data efficiency (i.e., greater performance gain per additional training sample) as the sample size increased, ultimately reaching an average R^2 of 0.680 with all 1801 samples.

Permeability predictions generally show lower accuracy than porosity (Fig. 10b), but follow similar patterns. All three clusters showed signs of severe underfitting with small sample sizes. Models begin to capture effective signals at 40–160 samples, and the performance indicators gradually improve. Among them, Cluster 1 consistently performed worse than Clusters 2 and 3 and was more sensitive to training sample size. Overall, in relatively high-quality reservoirs, the coupling relationship between well logs and reservoir properties is more clearly defined, allowing better model performance. Conversely, in tight reservoirs with complex pore structures, model learning becomes more challenging and requires substantially larger datasets to achieve acceptable prediction accuracy. Detailed modeling results for each cluster are provided in Supplementary Table S4–S6.

5. Discussion

5.1. ML method and feature variable

In this study, we confirmed the critical role of nonlinear modeling capabilities of ML for predicting petrophysical parameters in low-permeability sandstone reservoirs. Controlled by both depositional and diagenetic processes, such reservoirs typically exhibit high heterogeneity (Lai et al., 2018a; Luo et al., 2024), leading to complex nonlinear relationships between well logs and reservoir properties. Traditional linear regression models struggle to capture these multi-factor interactions (Zhao et al., 2024). In contrast, tree-based ensemble models improve prediction accuracy by integrating multiple decision trees, enabling segmented identification of interactions among logging features. DNN, through multilayer nonlinear transformations, progressively extract and integrate high-level features such as macroscopic lithofacies and microscopic pore structures. From a geological perspective, these models are equivalent to constructing separate sub-models for different facies, thereby improving overall prediction accuracy.

Feature importance is another topic worthy of discussion during the

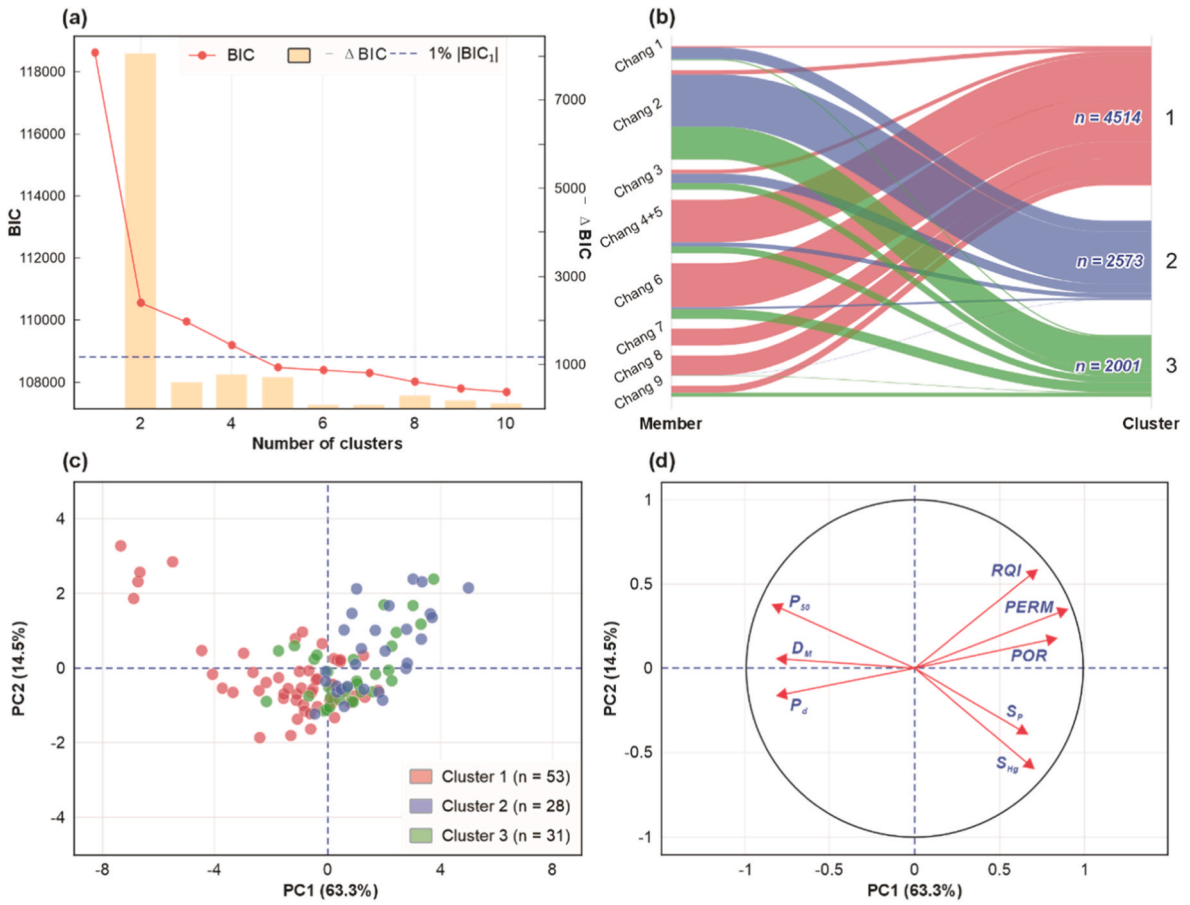

Fig. 8. (a) Optimal number of clusters based on the Relative-BIC Elbow Criterion; (b) Vertical distribution of the three electrical clusters; (c) PCA projection of pore structure parameters obtained from HPMI; (d) Variable loading plot.

prediction process. Using the optimal porosity and permeability prediction models established in Section 4.2, we employed SHapley Additive exPlanations (SHAP) to quantify feature importance (Fig. 11a and b). For porosity prediction, the top four contributing features were RILD, DEN, AC, and Member. For permeability prediction, the top features were DEN, Member, SP, and AC. From a geophysical standpoint, AC, DEN, and RILD are indicative of pore volume and serve as effective indicators of reservoir properties. SP reflects formation water resistivity and clay content, serving as a secondary indicator of lithology and permeability. Therefore, the high rankings of these features have clear physical significance.

Notably, the integer-encoded stratigraphic member (Member) contributed 16.12 % and 16.51 % to porosity and permeability prediction, respectively. To verify whether the performance improvement introduced by the Member variable was stable and significant, we performed ten random subsampling experiments at a sample size of 1,280 using the XGBoost model. The results indicate that including the Member variable improved model performance, with R^2 increasing by approximately 1.82 % for porosity and 3.45 % for permeability (Fig. 11c–f). Although the magnitude of improvement is modest, the effect is stable and statistically significant.

Stratigraphic members typically encapsulate shared information on sedimentary environment, burial history, and diagenetic processes. Integer coding roughly preserves their stratigraphic order, enabling the model to capture depth- or age-related trends in reservoir properties. Previous studies have also shown that incorporating geological information, such as lithofacies, sedimentary facies, and diagenetic facies, into predictive models can significantly improve performance (Zhang et al., 2021; Zhao et al., 2024; Tao et al., 2025). This finding highlights the critical role of discrete geological attributes in reservoir property prediction.

5.2. Sample size

In recent years, data-driven ML methods have demonstrated promising results in predicting reservoir properties (Tan et al., 2020; Fang et al., 2024; He et al., 2025). However, the acquisition and processing of large-scale datasets often entail significant economic costs (Zhao et al., 2022). Although small-sample learning has gained attention in various research fields (Moreno-Barea et al., 2020; Li et al., 2025), its application in low-permeability sandstone reservoirs remains limited. This study confirms through a sample size scaling experiment that a sufficient number of training samples not only improves prediction accuracy but also enhances model generalization and stability. Limited sample sizes fail to capture the full extent of reservoir heterogeneity, while a larger sample size enhances predictive reliability on unseen data.

It is worth noting that improvements in model performance do not increase linearly with sample size but exhibit a clear saturation trend. We quantitatively described the relationship between sample increment and model performance using first derivative analysis. This approach draws on the concept of learning curves in ML theory (Mohr and van Rijn, 2024). Details of the calculation method are provided in Supplementary Text S2.

The results show that as the sample size increases, the marginal gain in performance diminishes (Fig. 12a and b), which is consistent with the principles of statistical learning theory (Vapnik, 2000). This suggests that the collected dataset has already captured the main geological characteristics of the reservoir, and additional samples contribute limited new information. This observation aligns with prior studies reporting the existence of a “sample saturation point” or “performance

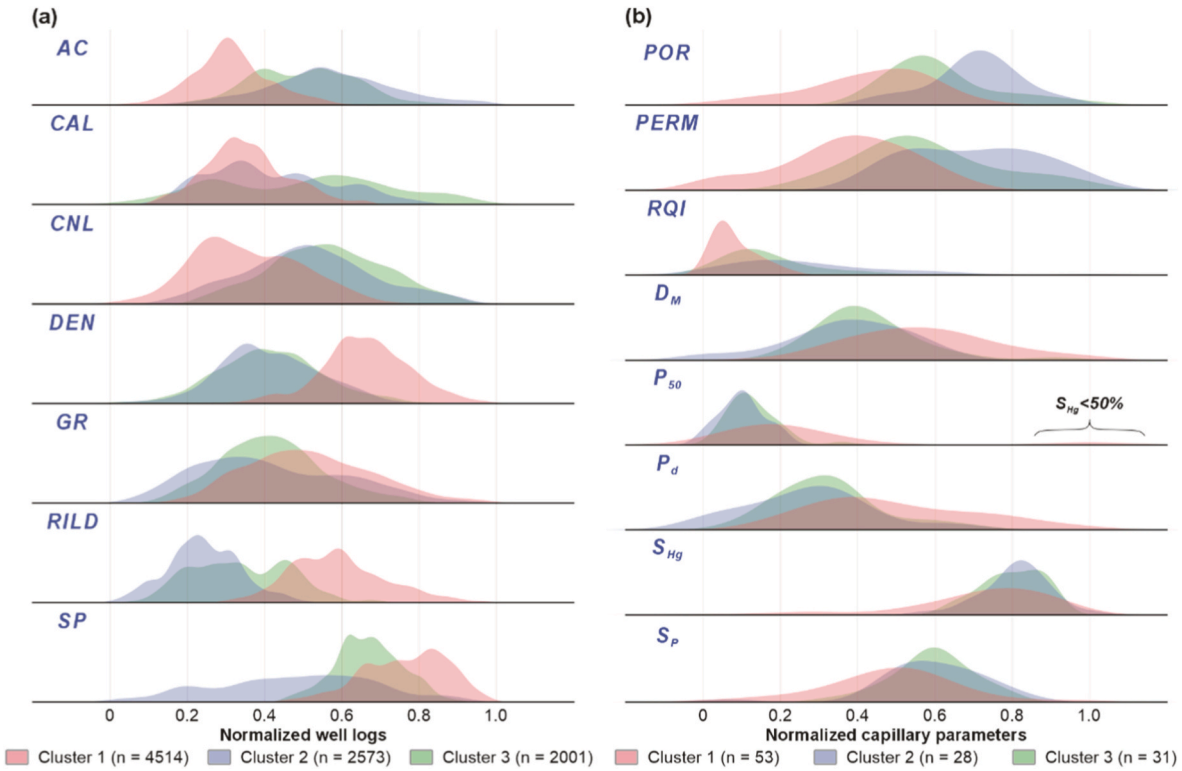

Fig. 9. (a) Kernel density distributions of well logs (AC, CAL, CNL, DEN, GR, RILD, and SP) for the three electrical clusters; (b) Kernel density distributions of pore structure parameters (POR, PERM, RQI, D_M, P_{50}, P_d, S_{Hg}, and S_P) for the three electrical clusters.

Table 3
Statistical summary of HPMI-derived pore parameters by cluster.

Cluster	Statistics	Por, %	RQI	Log_{10}(Perm)	D_M, μm	S_p	$Log_{10}(P_{50})$	$Log_{10}(P_d)$	S_{Hg}, %
1 (n = 53)	Mean	9.623	1.420	−0.817	12.451	1.811	1.307	0.214	76.157
	Std	2.888	0.632	0.491	1.186	0.645	1.271	0.543	15.596
	Min	1.882	0.553	−1.921	9.650	0.062	0.266	−0.672	19.800
	Max	15.270	3.424	0.074	15.174	3.659	5.000	1.524	95.538
2 (n = 28)	Mean	14.580	3.579	0.150	11.230	2.185	0.478	−0.367	84.127
	Std	2.075	1.925	0.496	0.984	0.517	0.285	0.424	6.702
	Min	9.890	1.482	−0.618	8.883	0.280	−0.030	−1.084	65.900
	Max	18.730	7.969	1.034	12.996	2.999	0.992	0.634	99.191
3 (n = 31)	Mean	12.964	2.561	−0.209	11.573	2.197	0.619	−0.232	83.274
	Std	2.505	1.597	0.529	0.878	0.377	0.333	0.356	6.242
	Min	9.450	0.786	−1.092	10.160	1.246	0.155	−0.861	70.300
	Max	19.900	7.594	0.986	14.623	2.890	1.773	0.740	91.475

plateau" in ML (Rajput et al., 2023; Ahmadisharaf et al., 2024; Schmidinger et al., 2024). While training sample size is crucial for achieving accuracy and stability, there exists an optimal range that balances performance and economic cost. Exceeding this range yields minimal benefit, while staying below it risks underfitting and limits practical application.

For under-explored blocks with limited data availability, determining the optimal sample size is a critical challenge. To address this, we propose a dynamic progressive sampling strategy. Specifically, the training set was expanded exponentially, while the test set was incrementally supplemented with a sample size equivalent to 10 % of the current training set, simulating progressive sampling process in new areas. For each sample size, we perform five random samples with replacement to build models, and then compute the performance derivative between adjacent sample sizes. The results show a strong correlation between performance derivatives derived from progressive sampling and those from the full dataset (Fig. 12c and d). This indicates that the strategy enables modeling of performance trends relative to sample size without requiring access to large-scale datasets. Consequently, it provides a practical means to identify the optimal sample range for new exploration blocks.

5.3. Pore structure

Pore structure exerts a decisive influence on key petrophysical parameters such as porosity and permeability (Lai et al., 2018b). This study demonstrates that sandstone reservoirs with complex pore systems typically require larger datasets for accurate modeling. Based on feature importance rankings, four key well logs (AC, DEN, RILD, and SP) were selected for PCA (Fig. 13a). Scatter plots were then constructed using the first principal component (PC1) of the well logs and the PC1 of the pore structure to evaluate the explanatory power of well logs in capturing pore structure variability (Fig. 13b).

The results reveal that the 90 % confidence ellipse for Cluster 1 is significantly larger than those for Clusters 2 and 3, indicating a greater variance and higher heterogeneity in pore structure. Specifically, the variance of PC1 for well logs in Cluster 1 is 0.860, while that of pore structure PC1 reaches 4.734 (Table 4). This implies that well logs

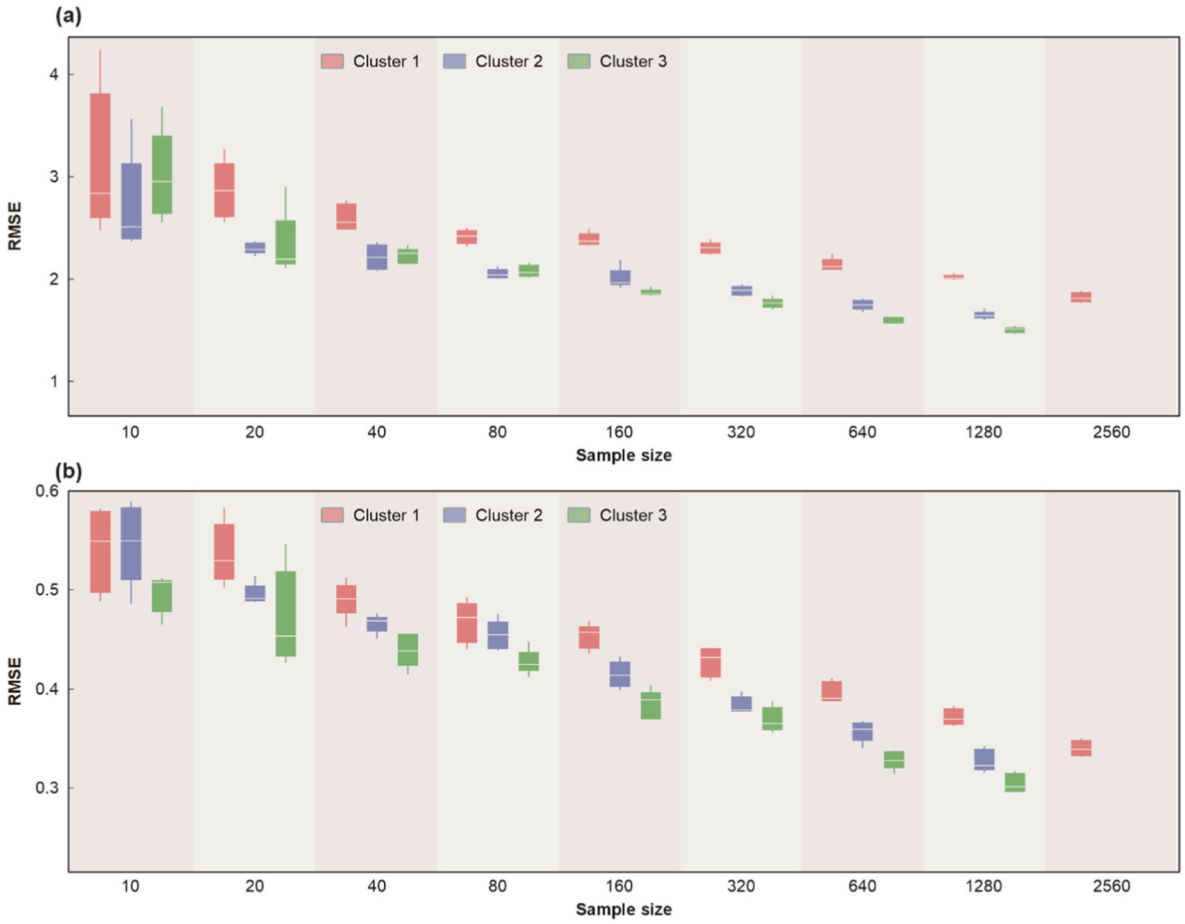

Fig. 10. Performance under varying training sample sizes for the three electrical clusters, evaluated using RMSE. (a) Porosity and (b) permeability prediction using the XGBoost model.

explain only approximately 18.17 % of the pore structure variation. In contrast, the explanatory power for Cluster 2 and Cluster 3 reaches 31.55 % and 42.39 %, respectively. From a geological perspective, tight sandstones (such as those in Cluster 1) often undergo multiple diagenetic processes, including compaction, cementation, dissolution, and fracture infilling, resulting in highly heterogeneous pore systems. Conventional logging methods are generally insufficient to capture these subtle pore structure differences. Previous studies have also confirmed the limitations of conventional logging in characterizing the pore systems of tight sandstones (Tan et al., 2020; Jiao et al., 2024; Su et al., 2024; Lai et al., 2024).

Performance derivative curves further illustrate the differences in learning difficulty across the three clusters (Fig. 13c and d). In porosity prediction, at a sample size of 20, Cluster 3 achieved the highest marginal gain, with each additional sample increasing R^2 by 0.057 and decreasing RMSE by 0.068. By contrast, Cluster 1 exhibited minimal improvement, with ΔR^2 of only 0.023 and ΔRMSE of −0.026, indicating that reservoirs with complex pore structures improve most slowly in low-sample regimes. As the training sample size increased, all derivative curves decayed according to a power-law trend and eventually approached zero, although their convergence rates varied. Cluster 3 exhibited the steepest decay, with slopes of −1.41 for R^2 and -1.26 for RMSE. Cluster 1 showed the slowest decay, with slopes of −1.20 and −1.09, respectively, indicating a delayed performance plateau. These findings highlight that for highly heterogeneous reservoirs like Cluster 1, achieving reliable predictions requires a combination of larger training datasets, more sophisticated algorithms, and stronger geological constraints.

5.4. Limitations of this study

The ML models developed in this study were not optimized to achieve the absolute best evaluation metrics. Rather, the objective was to investigate how different algorithms, sample sizes, and pore structures influence model performance. All models underwent hyperparameter tuning within a consistent experimental framework. Specifically, the *Optuna* sampler was employed for 200 optimization trials, each using fixed 5-fold cross-validation to minimize RMSE. Although this approach may not fully exploit the performance potential of each model, it ensures both the comparability and reproducibility of the results.

Furthermore, we found that complex pore structures negatively affect model performance. This observation may raise concerns about whether it merely results from improperly defined electrical cluster boundaries. To address this, we conducted multiple clustering runs with different initialization parameters to validate the stability of the clustering results. The phenomenon observed in Cluster 1—*homogeneous logging responses but heterogeneous petrophysical properties*—persisted across all runs. This indicates that the pattern is not an artifact of random clustering, but rather an inherent pattern within the data. This suggests that even when logging responses appear similar, significant differences in microscopic pore structures may still exist. Consequently, this underscores a key limitation in the feature variables used in this study. Future research should aim to integrate geological domain knowledge and incorporating pore-scale experimental data (such as HPMI and NMR) to improve the precision and reliability of reservoir property predictions in low-permeability sandstone reservoirs.

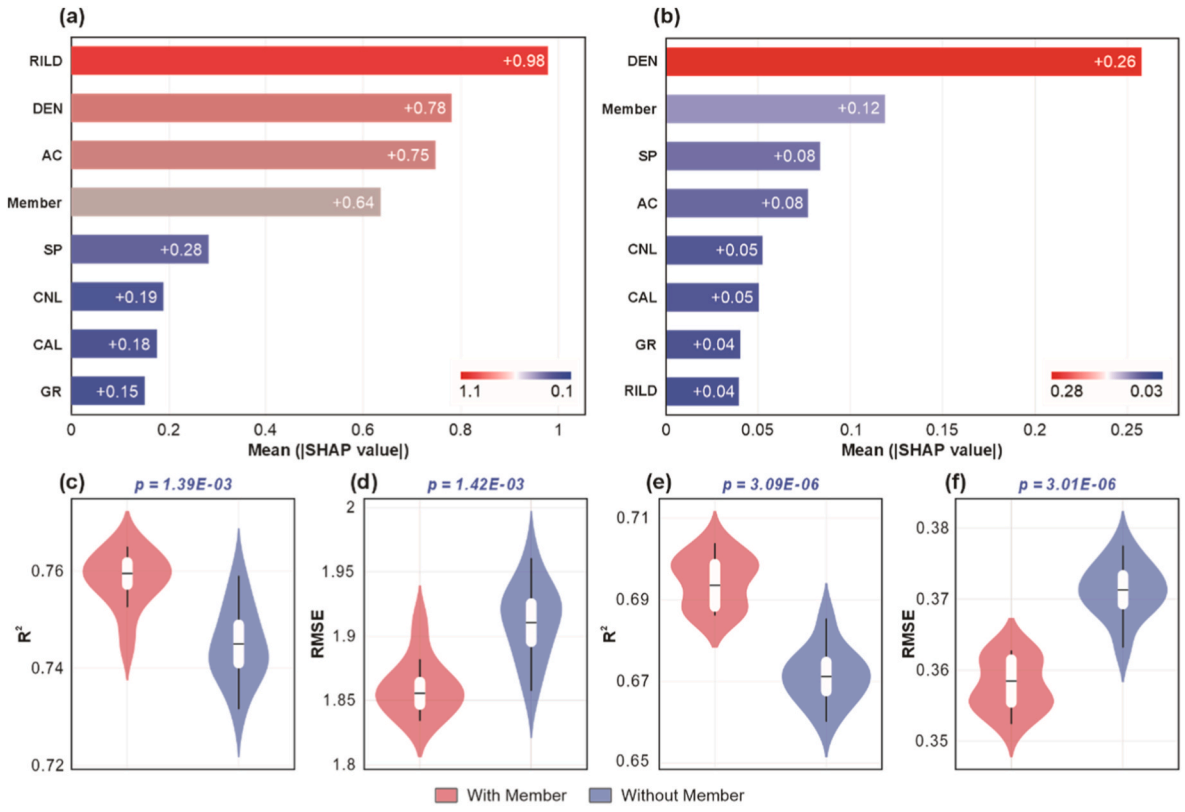

Fig. 11. Feature importance analysis and the effect of the Member variable on model performance. (a), (c), and (d) Porosity; (b), (e), and (f) Permeability.

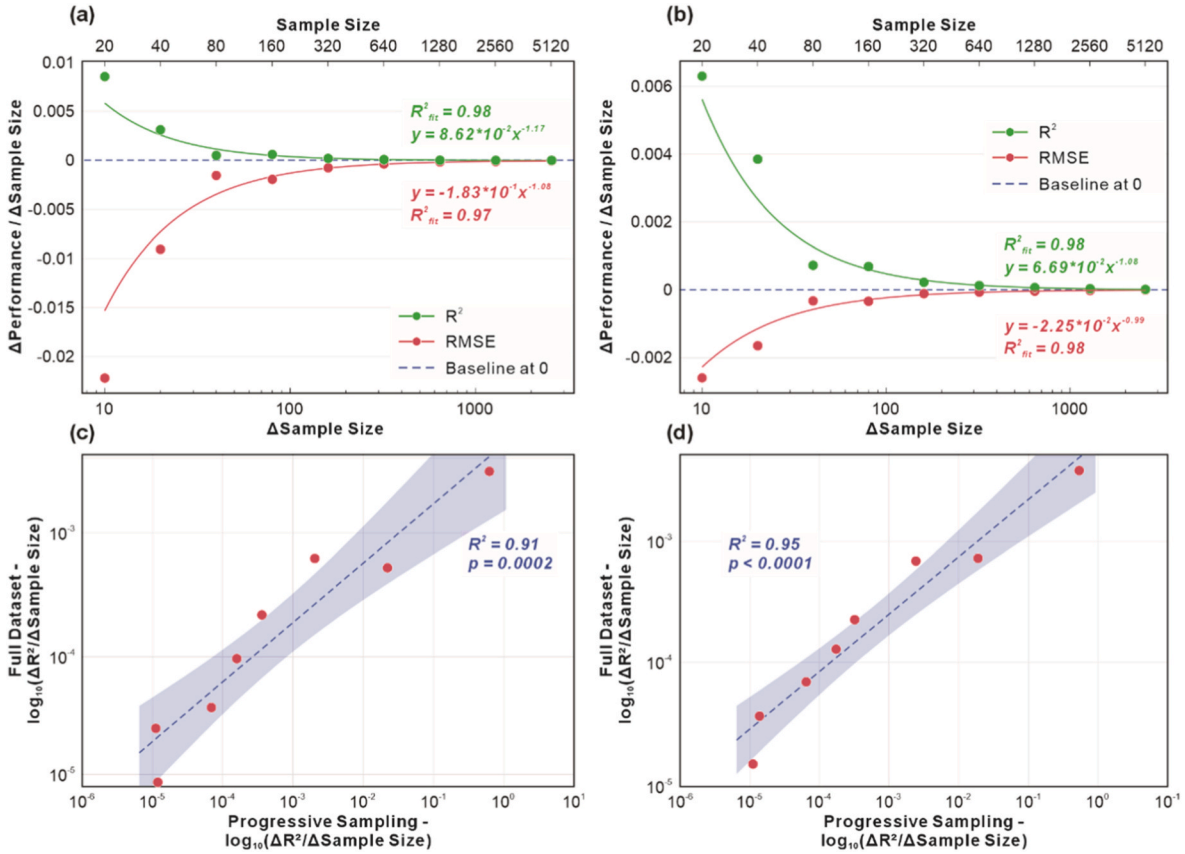

Fig. 12. (a) and (b) Power-law decay of first-order derivative curves (R^2 and RMSE) for porosity and permeability prediction; (c) and (d) Correlation between first-order derivative values of performance obtained from progressive sampling and from the full dataset, for porosity and permeability prediction.

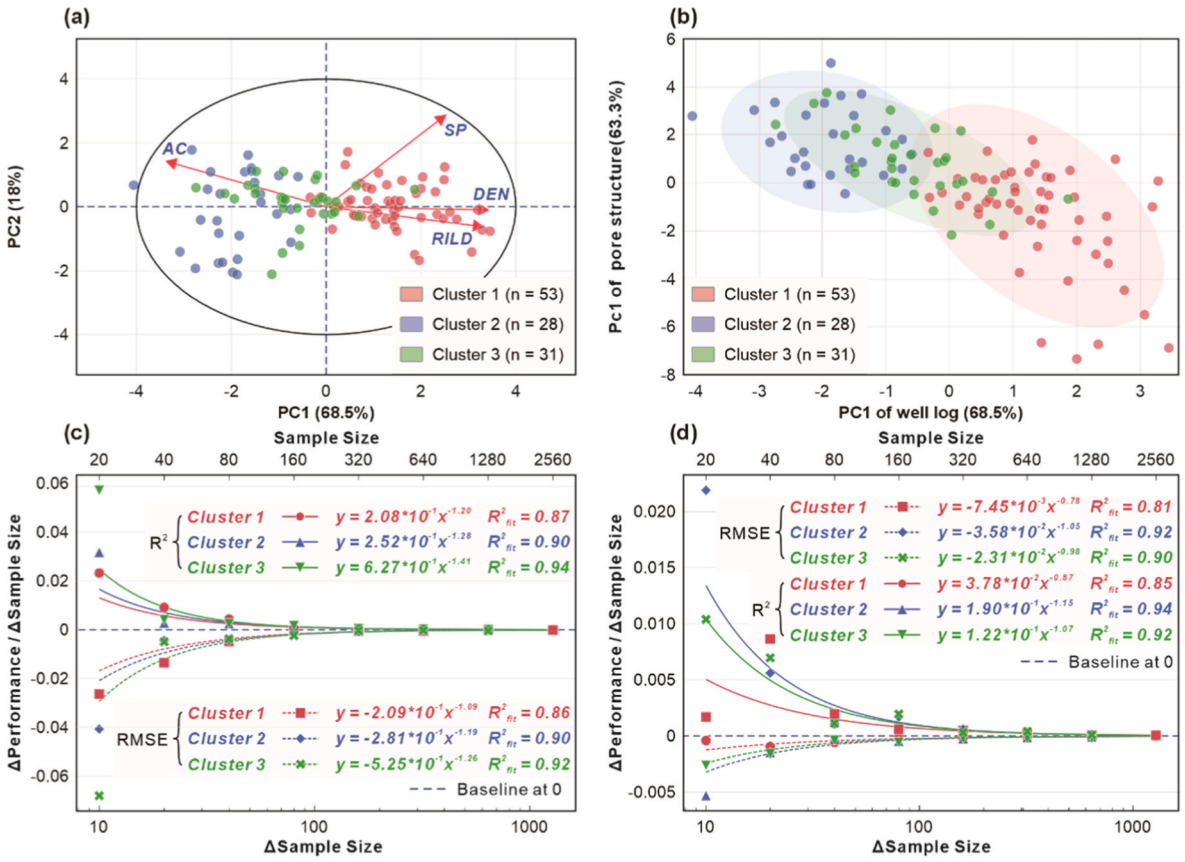

Fig. 13. (a) Projection of key logging variables in PCA projection; (b) Correlation between PC1 of well log and PC1 of pore structure across different electrical clusters, shown with 90 % confidence ellipses; (c) and (d) Performance derivative curves (R^2 and RMSE) for porosity and permeability prediction across clusters.

Table 4
Statistical summary of PCA results for pore-structure parameters and well logs by cluster.

Cluster	Statistics	Pore Structure		Well Log	
		PC1	PC2	PC1	PC2
1 (n = 53)	Mean	−1.382	−0.204	1.383	0.006
	Var.	4.734	1.211	0.860	0.457
2 (n = 28)	Mean	1.707	0.468	−1.927	−0.138
	Var.	1.912	1.204	0.603	1.487
3 (n = 31)	Mean	0.822	−0.073	−0.625	0.113
	Var.	2.014	0.823	0.854	0.504

6. Conclusion

This study demonstrates that accurately predicting petrophysical properties in low-permeability sandstone reservoirs does not necessarily require extremely large datasets. As the number of training samples increases, the marginal gains in model performance diminish significantly. Within a certain range, there exists an economically efficient optimal sample size that ensures accuracy and stability, exceeding this threshold offers limited additional benefit.

By selecting ML methods with strong nonlinear modeling capabilities and incorporating discrete geological variables such as lithofacies and sedimentary facies, prediction accuracy can be further improved. Moreover, in tight sandstones with highly complex pore structures, conventional well logs often fail to capture subtle structural differences, thus limiting model performance. To enhance the model's sensitivity to such complex pore networks, either a larger training dataset or the integration of geological and experimental data that directly characterize pore structure is required.

In the context of widespread AI adoption, these findings underscore the importance of rational sample size determination and detailed pore structure assessment in improving cost-efficiency and predictive accuracy. Future work should also explore the integration of multi-source variables to mitigate the impact of limited sample sizes. These variables should not be restricted to well-based data but should incorporate large-scale planar or spatial grid data. Such an approach is expected to enhance the model's predictive capability by providing spatial context beyond individual well locations.

CRediT authorship contribution statement

Yuxi Sun: Writing – original draft, Visualization, Software, Investigation, Formal analysis, Data curation, Conceptualization. **Liang Chen:** Writing – review & editing, Methodology, Conceptualization. **Yuan Qi:** Visualization, Validation, Investigation. **Yiping He:** Resources. **Hancheng Ji:** Supervision, Project administration, Funding acquisition. **Yanqing Shi:** Formal analysis. **Shuangqi Feng:** Visualization.

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

During the preparation of this work, the author(s) used ChatGPT (OpenAI) for the purpose of improving the readability and language clarity of the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

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 study was financially supported by China University of Petroleum, Beijing (2462023BJR011), the CNPC Innovation Fund (2021D002-0102). The authors sincerely thank PetroChina Changqing Oilfield Company for providing samples and data access.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoen.2025.214266.

Data availability

Data will be made available on request.

References

Abid, M., Ba, J., Markus, U.I., Tariq, Z., Ali, S.H., 2025. Modified approach to estimate effective porosity using density and neutron logging data in conventional and unconventional reservoirs. J. Appl. Geophys. 233, 105571. https://doi.org/10.1016/j.jappgeo.2024.105571.

Aftab, S., Moghadam, R.H., 2022. Robust data smoothing algorithms and wavelet filter for denoising sonic log signals. J. Appl. Geophys. 206, 104836. https://doi.org/10.1016/j.jappgeo.2022.104836.

Ahmadisharaf, A., Nematirad, R., Sabouri, S., Pachepsky, Y., Ghanbarian, B., 2024. Representative sample size for estimating saturated hydraulic conductivity via machine learning: a proof-of-concept study. Water Resour. Res. 60, e2023WR036783. https://doi.org/10.1029/2023WR036783.

Bai, J.L., Zhao, J.F., Ren, Z.L., Li, W.H., Wang, K., Li, X., 2022. Paleogeographic and sedimentary evolution of meso–neoproterozoic strata in the Ordos Basin, western North China Craton. J. Pet. Sci. Eng. 215, 110600. https://doi.org/10.1016/j.petrol.2022.110600.

Baraboshkin, E.E., Ismailova, L.S., Orlov, D.M., Zhukovskaya, E.A., Kalmykov, G.A., Khotylev, O.V., Baraboshkin, E.Yu, Koroteev, D.A., 2020. Deep convolutions for in-depth automated rock typing. Comput. Geosci. 135, 104330. https://doi.org/10.1016/j.cageo.2019.104330.

Bione, F.R.A., Venancio, I.M., Santos, T.P., Belem, A.L., Rangel, B.R., Souza, I.V.A.F., Spigolon, A.L.D., Albuquerque, A.L.S., 2024. Estimating total organic carbon of potential source rocks in the Espírito Santo Basin, SE Brazil, using XGBoost. Mar. Pet. Geol. 162, 106765. https://doi.org/10.1016/j.marpetgeo.2024.106765.

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

Chen, L., Lin, W.B., Chen, P., Jiang, S., Liu, L., Hu, H.Y., 2021. Porosity prediction from well logs using back propagation neural network optimized by genetic algorithm in one heterogeneous oil reservoirs of Ordos Basin, China. J. Earth Sci. 32, 828–838. https://doi.org/10.1007/s12583-020-1396-5.

Chen, T.Q., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. Association for Computing Machinery, New York, NY, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Coker, A.K., 1995. Chapter 1 - numerical computation. In: Coker, A.K. (Ed.), Fortran Programs for Chemical Process Design, Analysis, and Simulation. Gulf Professional Publishing, Houston, pp. 1–102. https://doi.org/10.1016/B978-088415280-4/50002-9.

Deng, X.Q., Cheng, D.X., Zhou, X.P., Shi, Z.W., Guo, Y.X., 2020. Formation hydrochemical characteristics and genesis of the Lower Jurassic, Ordos Basin. Acta Sedimentol. Sin. 38, 1099–1110. https://doi.org/10.14027/j.issn.1000-0550.2019.080 (in Chinese).

Ehsan, M., Manzoor, U., Chen, R., Hussain, M., Abdelrahman, K., Radwan, A.E., Ullah, J., Iftikhar, M.K., Arshad, F., 2024. Pore pressure prediction based on conventional well logs and seismic data using an advanced machine learning approach. J. Rock Mech. Geotech. Eng. https://doi.org/10.1016/j.jrmge.2024.09.049.

Fang, Z.J., Ba, J., Carcione, J.M., Xiong, F.S., Gao, L., 2024. Permeability prediction using logging data from tight reservoirs based on deep neural networks. J. Appl. Geophys. 229, 105501. https://doi.org/10.1016/j.jappgeo.2024.105501.

Feng, R.M., Chen, S.N., Bryant, S., Liu, J., 2019. Stress-dependent permeability measurement techniques for unconventional gas reservoirs: review, evaluation, and application. Fuel 256, 115987. https://doi.org/10.1016/j.fuel.2019.115987.

Fu, X.G., Wang, J., Wen, H.G., Wang, Z.W., Zeng, S.Q., Song, C.Y., Chen, W.B., Wan, Y.L., 2020. A possible link between the Carnian Pluvial Event, global carbon-cycle perturbation, and volcanism: new data from the Qinghai-Tibet Plateau. Glob. Planet. Change 194, 103300. https://doi.org/10.1016/j.gloplacha.2020.103300.

Guo, H.J., He, R.L., Jia, W.L., Peng, P.A., Lei, Y.H., Luo, X.R., Wang, X.Z., Zhang, L.X., Jiang, C.F., 2018. Pore characteristics of lacustrine shale within the oil window in the Upper Triassic Yanchang Formation, southeastern Ordos Basin, China. Mar. Pet. Geol. 91, 279–296. https://doi.org/10.1016/j.marpetgeo.2018.01.013.

He, Y.J., Zhang, H.J., Wu, Z.Y., Zhang, H.B., Zhang, X., Zhuo, X.J., Song, X.L., Dai, S., Dang, W., 2025. Porosity prediction of tight reservoir rock using well logging data and machine learning. Sci. Rep. 15, 13124. https://doi.org/10.1038/s41598-025-95578-7.

Hu, F., Wu, C.L., Shang, J.W., Yan, Y.M., Wang, L.Q., Zhang, H., 2023. Multi-condition controlled sedimentary facies modeling based on generative adversarial network. Comput. Geosci. 171, 105290. https://doi.org/10.1016/j.cageo.2022.105290.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67. https://doi.org/10.1080/00401706.1970.10488634.

Iraji, S., Soltanmohammadi, R., Matheus, G.F., Basso, M., Vidal, A.C., 2023. Application of unsupervised learning and deep learning for rock type prediction and petrophysical characterization using multi-scale data. Geoenergy Sci. Eng. 230, 212241. https://doi.org/10.1016/j.geoen.2023.212241.

Jiang, S.Y., Sun, P.K., Lyu, F.Q., Zhu, S.C., Zhou, R.F., Li, B., He, T.H., Lin, Y.J., Gao, Y.N., Song, W.D., Xu, H.M., 2024a. Machine learning (ML) for fluvial lithofacies identification from well logs: a hybrid classification model integrating lithofacies characteristics, logging data distributions, and ML models applicability. Geoenergy Sci. Eng. 233, 212587. https://doi.org/10.1016/j.geoen.2023.212587.

Jiang, S.J., Sweet, L., Blougouras, G., Brenning, A., Li, W.T., Reichstein, M., Denzler, J., Shangguan, W., Yu, G., Huang, F.N., Zscheischler, J., 2024b. How interpretable machine learning can benefit process understanding in the geosciences. Earths Future 12, e2024EF004540. https://doi.org/10.1029/2024EF004540.

Jiao, L.M., Denœux, T., Liu, Z.G., Pan, Q., 2022. EGMM: an evidential version of the Gaussian mixture model for clustering. Appl. Soft Comput. 129, 109619. https://doi.org/10.1016/j.asoc.2022.109619.

Jiao, S.X., Zhao, J., Ren, X.F., Wen, X.F., Liu, A.P., Cai, F., Yu, B.C., Lai, Q., 2024. Parameter evaluation method of tight carbonate reservoir using electrical imaging pores diameter spectrum. Geomech. Geophys. Geo-Energy Geo-Resour. 10, 34. https://doi.org/10.1007/s40948-024-00757-x.

Katz, B.J., Arango, I., 2018. Organic porosity: a geochemist's view of the current state of understanding. Org. Geochem. 123, 1–16. https://doi.org/10.1016/j.orggeochem.2018.05.015.

Koray, A.-M., Bui, D., Kubi, E.A., Ampomah, W., Amosu, A., 2024. Machine learning based reservoir characterization and numerical modeling from integrated well log and core data. Geoenergy Sci. Eng. 243, 213296. https://doi.org/10.1016/j.geoen.2024.213296.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17. Curran Associates Inc., Red Hook, NY, USA, pp. 3149–3157. https://doi.org/10.5555/3294996.3295074.

Lai, J., Su, Y., Xiao, L., Zhao, F., Bai, T.Y., Li, Y.H., Li, H.B., Huang, Y.Y., Wang, G.W., Qin, Z.Q., 2024. Application of geophysical well logs in solving geologic issues: past, present and future prospect. Geosci. Front. 15, 101779. https://doi.org/10.1016/j.gsf.2024.101779.

Lai, J., Wang, G.W., Fan, Q.X., Pang, X.J., Li, H.B., Zhao, F., Li, Y.H., Zhao, X., Zhao, Y.D., Huang, Y.Y., Bao, M., Qin, Z.Q., Wang, Q.Q., 2022. Geophysical well-log evaluation in the era of unconventional hydrocarbon resources: a review on current status and prospects. Surv. Geophys. 43, 913–957. https://doi.org/10.1007/s10712-022-09705-4.

Lai, J., Wang, G.W., Fan, Q.X., Zhao, F., Zhao, X., Li, Y.H., Zhao, Y.D., Pang, X.J., 2023. Toward the scientific interpretation of geophysical well logs: typical misunderstandings and countermeasures. Surv. Geophys. 44, 463–494. https://doi.org/10.1007/s10712-022-09746-9.

Lai, J., Wang, G.W., Wang, S., Cao, J.T., Li, M., Pang, X.J., Zhou, Z.L., Fan, X.Q., Dai, Q. Q., Yang, L., He, Z.B., Qin, Z.Q., 2018a. Review of diagenetic facies in tight sandstones: diagenesis, diagenetic minerals, and prediction via well logs. Earth Sci. Rev. 185, 234–258. https://doi.org/10.1016/j.earscirev.2018.06.009.

Lai, J., Wang, G.W., Wang, Z.Y., Chen, J., Pang, X.J., Wang, S.C., Zhou, Z.L., He, Z.B., Qin, Z.Q., Fan, X.Q., 2018b. A review on pore structure characterization in tight sandstones. Earth Sci. Rev. 177, 436–457. https://doi.org/10.1016/j.earscirev.2017.12.003.

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

Li, H., Zhao, H.Y., Xu, Y.X., He, Y.B., 2021. Characteristics of debrites, turbidites, and contourites in the Upper Ordovician Pingliang formation along southwestern margin of the Ordos Basin, western China. Arab. J. Geosci. 14, 1730. https://doi.org/10.1007/s12517-021-08215-8.

Li, P.K., Tong, X.R., Wang, Y.J., Zhang, Q., 2025. Meta doubly robust: debiasing CVR prediction via meta-learning with a small amount of unbiased data. Knowl.-Based Syst. 310, 112898. https://doi.org/10.1016/j.knosys.2024.112898.

Li, S.X., Chu, M.J., Wang, T.F., Zhang, W.X., Zhou, X.P., 2017. Features of formation water and implications for hydrocarbon accumulation in Chang 6 pay zone, Jiyuan area, Ordos Basin(in Chinese). China Pet. Explor 22, 43–53. https://doi.org/10.3969/j.issn.1672-7703.2017.05.005.

Lin, M.R., Xi, K.L., Cao, Y.C., Niu, X.B., Ma, W.J., Wang, X.J., Xu, S., 2024. Palaeoenvironmental changes in the Late Triassic lacustrine facies of the Ordos Basin of Northwest China were driven by multistage volcanic activity: implications for the understanding the Carnian Pluvial Event. Palaeogeogr. Palaeoclimatol. Palaeoecol. 637, 112012. https://doi.org/10.1016/j.palaeo.2024.112012.

Liu, M.C., Wu, S.H., Yue, D.L., Xu, Z.H., Wan, X.L., Wu, H.L., Chen, Z.H., Li, Z., 2025. Palaeogeomorphological control on the depositional architecture of lacustrine gravity-flow deposits in a depression lacustrine basin: a case study of the Triassic Yanchang Formation, southern Ordos Basin, China. J. Palaeogeogr. 14, 476–500. https://doi.org/10.1016/j.jop.2025.01.003.

Lu, J.G., Liao, J.B., Liu, X.J., Li, Y., Yao, J.L., He, Q.B., Xiao, Z.L., He, X., Fu, X.Y., Li, X. M., 2022. Geochemistry of different source rocks and oil-source correlation of lacustrine sedimentary successions: a case study of the Triassic Yanchang formation in the Dingbian-Wuqi Area, Ordos Basin, Northern China. J. Asian Earth Sci., Tectonics and Sedimentology of Accretionary and Collisional Orogens 232, 105216. https://doi.org/10.1016/j.jseaes.2022.105216.

Lu, Z.Y., He, Z.L., Gluyas, J.G., Liu, G.X., Liu, T., Chen, C.F., Zou, M., 2024. Reservoir quality of the lower–middle Permian Shan 2 and He 1 members in the Ordos Basin, China: implications for depositional and diagenetic processes and the role of volcanic tuffaceous sediment in tight sandstones. J. Asian Earth Sci. 263, 106050. https://doi.org/10.1016/j.jseaes.2024.106050.

Luo, J.C., Yan, J.P., Liao, M.J., Wang, M., Geng, B., Hu, Q.H., 2024. Evaluation of pore structure characteristics of deep clastic rocks in the Huangliu formation of LD-X area, Yinggehai Basin. Mar. Pet. Geol. 167, 106969. https://doi.org/10.1016/j.marpetgeo.2024.106969.

Ma, Y.Y., Qiao, Y.H., Chen, M.X., Rui, D.N., Zhang, X.X., Liu, W.J., Ye, L., 2024. How small is big enough? Big data-driven machine learning predictions for a full-scale wastewater treatment plant. Water Res., 123041 https://doi.org/10.1016/j.watres.2024.123041.

Maldar, R., Ranjbar-Karami, R., Behdad, A., Bagherzadeh, S., 2022. Reservoir rock typing and electrofacies characterization by integrating petrophysical properties and core data in the Bangestan reservoir of the Gachsaran oilfield, the Zagros basin, Iran. J. Pet. Sci. Eng. 210, 110080. https://doi.org/10.1016/j.petrol.2021.110080.

Mao, J.Y., Chen, J., Deng, Y.J., 2025. Optimization strategy for ensemble learning models based on fusing resampling, adaptive dimensionality reduction, and Optuna in intelligent flight technology evaluation. Aerosp. Sci. Technol. 162, 110251. https://doi.org/10.1016/j.ast.2025.110251.

Mohr, F., van Rijn, J.N., 2024. Learning curves for decision making in supervised machine learning: a survey. Mach. Learn. 113, 8371–8425. https://doi.org/10.1007/s10994-024-06619-7.

Moreno-Barea, F.J., Jerez, J.M., Franco, L., 2020. Improving classification accuracy using data augmentation on small data sets. Expert Syst. Appl. 161, 113696. https://doi.org/10.1016/j.eswa.2020.113696.

Noble, W.S., 2006. What is a support vector machine? Nat. Biotechnol. 24, 1565–1567. https://doi.org/10.1038/nbt1206-1565.

Pittman, E.D., 1992. Relationship of porosity and permeability to various parameters derived from Mercury injection-capillary pressure curves for Sandstone1. AAPG Bull. 76, 191–198. https://doi.org/10.1306/BDFF87A4-1718-11D7-8645000102C1865D.

Qiao, J.C., Zeng, J.H., Jiang, S., Zhang, Y.C., Feng, S., Feng, X., Hu, H.T., 2020. Insights into the pore structure and implications for fluid flow capacity of tight gas sandstone: a case study in the upper paleozoic of the Ordos Basin. Mar. Pet. Geol. 118, 104439. https://doi.org/10.1016/j.marpetgeo.2020.104439.

Qin, Y., Liu, C.Y., Zhao, J.F., He, F.Q., Zhang, W., Yang, L.H., Du, N., Shao, D.Y., 2025. Unravelling the origin of gas in tight sandstones of the Hangjinqi gas field, Ordos Basin, China: new insights from natural gas geochemistry data. Org. Geochem. 206, 105012. https://doi.org/10.1016/j.orggeochem.2025.105012.

Rajput, D., Wang, W.J., Chen, C.C., 2023. Evaluation of a decided sample size in machine learning applications. BMC Bioinf. 24, 48. https://doi.org/10.1186/s12859-023-05156-9.

Safavi, S.J., Maldar, R., 2024. An enhancement in the petrophysical evaluation in a vuggy carbonate gas reservoir by integrating the core data and empirical methods, Zagros basin, south of Iran. J. Asian Earth Sci. X 11, 100177. https://doi.org/10.1016/j.jaesx.2024.100177.

Schmidinger, J., Schröter, I., Bönecke, E., Gebbers, R., Ruehlmann, J., Kramer, E., Mulder, V.L., Heuvelink, G.B.M., Vogel, S., 2024. Effect of training sample size, sampling design and prediction model on soil mapping with proximal sensing data for precision liming. Precis. Agric. 25, 1529–1555. https://doi.org/10.1007/s11119-024-10122-3.

Sharifi, A., Miri, R., Riazi, M., 2023. A holistic review of harsh conditions resistant surfactants for enhanced oil recovery in dense carbonate reservoir. Fuel 353, 129109. https://doi.org/10.1016/j.fuel.2023.129109.

Shehata, A.A., Ahmed, M., Kassem, A.A., Abdelrehim, R., Tsuji, T., Ismail, A., 2025. Optimizing permeability and porosity prediction with advanced machine learning: a case study unlocking the complexities of late cretaceous reservoirs, gulf of suez, Egypt. J. Afr. Earth Sci. 228, 105670. https://doi.org/10.1016/j.jafrearsci.2025.105670.

Song, S.H., Mukerji, T., Hou, J.G., Zhang, D.X., Lyu, X.R., 2022. GANSim-3D for conditional geomodeling: theory and field application. Water Resour. Res. 58, e2021WR031865. https://doi.org/10.1029/2021WR031865.

Su, Y., Lai, J., Dang, W.L., Bie, K., Zhao, Y.D., Zhao, X.J., Li, D., Zhao, F., Wang, G.W., 2024. Pore structure characterization and reservoir quality prediction in deep and ultra-deep tight sandstones by integrating image and NMR logs. J. Asian Earth Sci. 272, 106232. https://doi.org/10.1016/j.jseaes.2024.106232.

Tan, M.J., Bai, Y., Zhang, H.T., Li, G.R., Wei, X.P., Wang, A.D., 2020. Fluid typing in tight sandstone from wireline logs using classification committee machine. Fuel 271, 117601. https://doi.org/10.1016/j.fuel.2020.117601.

Tao, B.C., Zhou, H.L., Wu, W.Y., Zhang, G., Liu, B., Liu, X.Y., 2025. Porosity prediction based on improved structural modeling deep learning method guided by petrophysical information. Pet. Sci. https://doi.org/10.1016/j.petsci.2025.03.035.

Taunk, K., De, S., Verma, S., Swetapadma, A., 2019. A brief review of nearest Neighbor algorithm for learning and classification. In: 2019 International Conference on Intelligent Computing and Control Systems (ICCS). Presented at the 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pp. 1255–1260. https://doi.org/10.1109/ICCS45141.2019.9065747.

Vapnik, V.N., 2000. The Nature of Statistical Learning Theory. Springer, New York, NY. https://doi.org/10.1007/978-1-4757-3264-1.

Vera-Arroyo, A., Bedle, H., 2025. Seal and reservoir risk evaluation using hierarchical clustering analysis with seismic attributes in Northwestern Australia. J. Appl. Geophys. 232, 105556. https://doi.org/10.1016/j.jappgeo.2024.105556.

Wang, L., Lyu, Q.Q., Li, L.H., Liu, J., Luo, S.S., Sun, X.H., Zhang, L., Xu, X.S., 2024. Sedimentary characteristics of mixed source fine-grained gravity-flow and its significance for shale oil exploration in a lacustrine depression basin: a case study of the Chang 73 Sub-member of the Triassic Yanchang Formation in Ordos Basin, NW China. Sediment. Geol. 464, 106629. https://doi.org/10.1016/j.sedgeo.2024.106629.

Wang, Y., Liu, L.F., Li, S.T., Ji, H.T., Xu, Z.J., Luo, Z.H., Xu, T., Li, L.Z., 2017. The forming mechanism and process of tight oil sand reservoirs: a case study of Chang 8 oil layers of the Upper Triassic Yanchang formation in the western Jiyuan area of the Ordos Basin, China. J. Pet. Sci. Eng. 158, 29–46. https://doi.org/10.1016/j.petrol.2017.08.026.

Wood, D.A., 2020. Predicting porosity, permeability and water saturation applying an optimized nearest-neighbour, machine-learning and data-mining network of well-log data. J. Pet. Sci. Eng. 184, 106587. https://doi.org/10.1016/j.petrol.2019.106587.

Xu, X.T., Zeng, L.B., Dong, S.Q., Li, H.M., Liu, J.Z., Ji, C.Q., 2025. The characteristics and controlling factors of high-quality reservoirs of ultra-deep tight sandstone: a case study of the Dabei Gas Field, Tarim Basin, China. Pet. Sci. https://doi.org/10.1016/j.petsci.2025.03.033.

Yang, S.F., Bao, Z.D., Wang, N., Qu, X.F., Lin, Y.B., Shen, J.J., Rizwan, S.A., 2020. Diagenetic evolution and its impact on reservoir quality of tight sandstones: a case study of the Triassic Chang 6 Member, Ordos Basin, northwest China. Mar. Pet. Geol. 117, 104360. https://doi.org/10.1016/j.marpetgeo.2020.104360.

Yaniv, A., Beck, Y., 2024. Enhancing NILM classification via robust principal component analysis dimension reduction. Heliyon 10, e30607. https://doi.org/10.1016/j.heliyon.2024.e30607.

Yao, J.L., Deng, X.Q., Zhao, Y.D., Han, T.Y., Chu, M.J., Pang, J.L., 2013. Characteristics of tight oil in Triassic Yanchang Formation, Ordos Basin. Pet. Explor. Dev. 40, 161–169. https://doi.org/10.1016/S1876-3804(13)60019-1.

Zhang, Q.P., Yang, C., Gu, Y., Tian, Y., Liu, H., Xiao, W., Wang, Z.K., Mi, Z.R., 2025. Microscopic pore-throat structure and fluid mobility of tight sandstone reservoirs in multi-provenance systems, Triassic Yanchang formation, Jiyuan area, Ordos basin. Energy Geosci 6, 100407. https://doi.org/10.1016/j.engeos.2025.100407.

Zhang, Z., Zhang, H., Li, J., Cai, Z.X., 2021. Permeability and porosity prediction using logging data in a heterogeneous dolomite reservoir: an integrated approach. J. Nat. Gas Sci. Eng. 86, 103743. https://doi.org/10.1016/j.jngse.2020.103743.

Zhao, W.W., Zhang, Z.H., Liao, J.B., Zhang, J.W., Zhang, W.T., 2024. Prediction method for the porosity of tight sandstone constrained by lithofacies and logging resolution. Mar. Pet. Geol. 170, 107114. https://doi.org/10.1016/j.marpetgeo.2024.107114.

Zhao, X.B., Chen, X.J., Huang, Q., Lan, Z.J., Wang, X.G., Yao, G.Q., 2022. Logging-data-driven permeability prediction in low-permeable sandstones based on machine learning with pattern visualization: a case study in Wenchang A Sag, Pearl River Mouth Basin. J. Pet. Sci. Eng. 214, 110517. https://doi.org/10.1016/j.petrol.2022.110517.

Zou, C.N., Yang, Z., He, D.B., Wei, Y.S., Li, J., Jia, A.L., Chen, J.J., Zhao, Q., Li, Y.L., Li, Jun, Yang, S., 2018. Theory, technology and prospects of conventional and unconventional natural gas. Pet. Explor. Dev. 45, 604–618. https://doi.org/10.1016/S1876-3804(18)30066-1.