

Original Paper

Enhancing reservoir characterization via lithology-constrained machine learning methods: The deep Jurassic Sangonghe tight sandstones in the Turpan-Hami Basin, China

Hong-Bin Li ^{a,b}, Yin-Hong Tian ^{a,b}, Gui-Wen Wang ^{a,b,*}, Zhi-Bin He ^{a,b}, Lin-Bo Shao ^{a,b}, Jin Lai ^{a,b}

^a State Key Laboratory of Petroleum Resources and Engineering, China University of Petroleum (Beijing), Beijing, 102249, China

^b College of Geosciences, China University of Petroleum (Beijing), Beijing, 102249, China

ARTICLE INFO

Article history:

Received 24 September 2025

Received in revised form

28 January 2026

Accepted 15 May 2026

Available online 20 May 2026

Keywords:

Deep tight sandstone reservoir

Parameter prediction

Machine learning

SHAP-based interpretability

Sangonghe Formation

Turpan-Hami Basin

ABSTRACT

Reliable porosity prediction in deep tight sandstones remains a persistent challenge due to complex lithological heterogeneity and weak correlations between conventional logs and reservoir parameters. This study systematically evaluates the performance of four predictive models, namely SVM, CatBoost, 2-D CNN and LSTM, in predicting NMR-derived effective porosity using nine conventional log curves and lithology as inputs. A total of 190 test samples from well J7-2-3H and blind well validation from the Jurassic Sangonghe Formation in the Turpan-Hami Basin were employed for performance benchmarking. The lithologies are mainly siltstone, fine-grained sandstone, medium-grained sandstone, coarse-grained sandstone, and sandy conglomerate. Results show that CatBoost consistently outperforms both traditional method and the other two deep learning models, achieving the highest accuracy ($R^2 = 0.94$, MAE = 0.24%) and demonstrating superior robustness. SHAP analysis was used to interpret model behavior and confirmed that CatBoost provides physically consistent predictions, assigning geologically reasonable weights to key features such as gamma ray, density, and lithology. The model–data compatibility and the structured treatment of discrete geological information are essential for enhancing predictive performance. Furthermore, lithology serves not only as an auxiliary input but also as a soft petrophysical constraint that enables improved model generalization across heterogeneous reservoirs. Despite the greater architectural complexity of deep learning models, their limited capacity to handle categorical features under data-scarce conditions results in suboptimal performance compared to CatBoost. This study presents a comparative modeling framework and interpretability strategy for evaluating reservoir parameter prediction models, emphasizing the value of incorporating geological priors in data-driven workflows. Medium- and coarse-grained sandstones exhibit relatively high porosity, with the latter showing a slightly superior porosity range. The proposed approach offers practical guidance for algorithm selection and feature engineering in intelligent reservoir characterization. This methodology enables precise calculation of reservoir parameters for deep tight sandstone reservoirs in the absence of NMR logging tools, thereby providing both theoretical frameworks and technical support for reservoir evaluation and reserve assessment of deep tight sandstones.

© 2026 Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

The depletion of conventional oil and gas resources and the growing energy demand have heightened the importance of unconventional oil and gas resources (Desbois et al., 2011; McGlade et al., 2013; Bera and Shah, 2021; Wang et al., 2024). Deep tight sandstone gas is an important type of unconventional hydrocarbon resource with rich reserves and substantial development potential

* Corresponding author.

E-mail address: wanggw@cup.edu.cn (G.-W. Wang).

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

(Jiang et al., 2015; Lai et al., 2023; Dai et al., 2024; Su et al., 2026). Porosity is one of the critical petrophysical parameters for characterizing deep tight sandstone gas reservoirs (Bagheri and Rezaei, 2019; Moosavi et al., 2023; Tariq et al., 2023). The tight sandstone reservoirs commonly exhibit strong heterogeneity due to complex diagenetic modifications and tectonic stress (Huang et al., 2026; Lai et al., 2026), leading to weak and ambiguous well logging responses, thereby challenging traditional petrophysics-based parameter evaluation methods (Ahmadi, 2015; Li et al., 2025). Nuclear magnetic resonance (NMR) logging represents an advanced geophysical technology capable of overcoming the intrinsic limitations of empirical petrophysical models and thereby improving the relatively accurate estimates of reservoir parameters (Xiao et al., 2018; Jin et al., 2019; Guo et al., 2023; Lai et al., 2024). Nevertheless, its prohibitive acquisition cost renders large-scale deployment in conventional wells economically unviable, as it is acquired only in a limited suite of key wells, restricting its broader spatial coverage and data availability (da Rocha et al., 2019; Liao et al., 2021). For the majority of conventional exploration wells, reservoir characterization still relies on low-cost, widely acquired conventional logging data (Dong et al., 2023; Tian et al., 2024). Therefore, determining how to effectively utilize the conventional logging data to achieve the most accurate possible inversion or prediction of reservoir parameters in unconventional reservoirs remains a challenging work (Handhal et al., 2020; Wu et al., 2025).

In recent years, machine learning (ML) techniques have emerged as a powerful toolkit for predicting petrophysical properties in unconventional reservoirs (Bergen et al., 2019; Lai et al., 2022; Shi et al., 2024), capable of uncovering complex, nonlinear patterns within multi-dimensional logging data (Raeesi et al., 2012; Otchere et al., 2021; Bione et al., 2024). Established ML algorithms, such as support vector machines (SVM) and random forests (RF), have proven effective in estimating key parameters like porosity and permeability (Chithra Chakra et al., 2013; Handhal et al., 2020; Zhao et al., 2025). For instance, Zhang et al. (2021) combined the flow zone indicator with a PSO-SVM algorithm to improve permeability and porosity predictions in heterogeneous dolomite reservoirs. Moreover, the application of CatBoost in synthetic NMR log generation further exemplifies the versatility of ML in reservoir evaluation (Tamoto et al., 2023).

Unlike conventional ML algorithms, deep learning (DL) models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) and hybrid frameworks, excel at capturing intricate nonlinear relationships and interactions within complex geoscientific data (Chen and Zhang, 2020; Yang et al., 2023), including conventional logging data, seismic attributes, and core measurements (Saporetti et al., 2022; Zhang and Yang, 2025). For instance, Wang et al. (2025) constructed a One-Dimensional Convolutional Neural Network-Bidirectional Long Short-Term Memory (1DCNN-BLSTM) model that leverages spatial and temporal features for reservoir classification. Similarly, Hu et al. (2024) introduced a GRU-based framework for water saturation estimation, trained on extensive conventional petrophysical logs. The method demonstrated robust temporal modeling and generalization across diverse lithologies, underscoring the efficacy of sequence-based architectures in analyzing depth-dependent logging data (Hu et al., 2024; Wang et al., 2025).

Both machine learning (ML) and deep learning (DL) have shown strong predictive performance in unconventional reservoir evaluation (Bagheri and Riahi, 2015). However, they differ fundamentally in their algorithmic structures and applicability across geological settings (Male et al., 2020; Latrach et al., 2024). ML models offer higher computational efficiency and perform well on

simpler datasets with clear correlations, while DL excels in handling complex, high-dimensional, and nonlinear data through advanced feature extraction (LeCun et al., 2015; Sagi and Rokach, 2018). Integration of ML with conventional and NMR logging data improves the accuracy and efficiency of reservoir parameter quantification (Tamoto et al., 2023). However, a key challenge for both approaches remains “black-box” nature, which reduces interpretability and limits geological insight into predictions (Gupta et al., 2023; Hassija et al., 2024).

Motivated by these considerations, this study focuses on the deep tight sandstone reservoirs of the Jurassic Sangonghe Formation in the Taibei Sag of the Turpan-Hami Basin. Several representative ML and DL models, validated in previous studies, were analyzed to evaluate their capabilities in porosity prediction under complex logging responses. By systematically comparing their predictive accuracies and further dissecting the results from the perspective of algorithmic principles, data characteristics, and model–data coupling relationships, the underlying reasons for the performance discrepancies are analyzed among different models under identical inputs. The effective porosity derived from NMR measurements was treated as the label for supervised learning, given its superior accuracy and consistency in deep tight sandstone reservoirs. Four representative artificial intelligence methods were employed for predictive modeling, including the conventional machine learning algorithms (SVM and CatBoost), alongside the deep learning architectures (2-D CNN and LSTM). The porosity prediction performance of each model was evaluated and compared using multiple regression assessment metrics. Then, through a comparative analysis of how models with varying performance levels extract and utilize input features, we systematically investigate the fundamental causes behind discrepancies in porosity prediction accuracy, emphasizing the interplay between model architecture and data utilization strategies. Finally, to further elucidate the model behavior, the SHAP method was introduced to visualize the internal decision-making processes of both the best-performing and underperforming models. Overall, this study presents a comparative investigation into the performance of machine learning and deep learning algorithms for reservoir parameter prediction, emphasizing the critical importance of algorithm–data compatibility in complex unconventional reservoirs. These insights can support more informed model selection and feature engineering in future applications of AI-assisted reservoir characterization in tight sandstones.

2. Geological settings

The Turpan-Hami Basin, positioned in northwestern China, marks the convergence of the Tarim, Kazakhstan, and Siberian plates (Fig. 1(a)) (Wang et al., 2018; Li et al., 2020). This narrow intermontane basin extends east–west, encompassing approximately 53,000 km², and is flanked by the Jueluotage Mountains to the south, the Bogeda Mountains to the north, the Balikun Mountains to the east, and the Kalawucheng Mountains to the west (Ni et al., 2019; Wang et al., 2023). The structural framework is divided into four major tectonic units: the Turpan Depression, the Aydingkol Lake Slope, the Liaodun Uplift, and the Hami Depression, with the Turpan Depression being the primary feature (Fig. 1(a)) (Zhi et al., 2024).

In the Early Jurassic, the basin developed north-dipping slopes under N–S compression, forming listric thrust faults along the southern margin of the central uplift’s western flank and initiating a faulted nose structure (Gong et al., 2016). During the Middle to Late Jurassic, enhanced detachment folding at the thrust front reshaped the basin, forming a depression near the northern foothills resembling a foreland basin (Zhang et al., 2020). By the

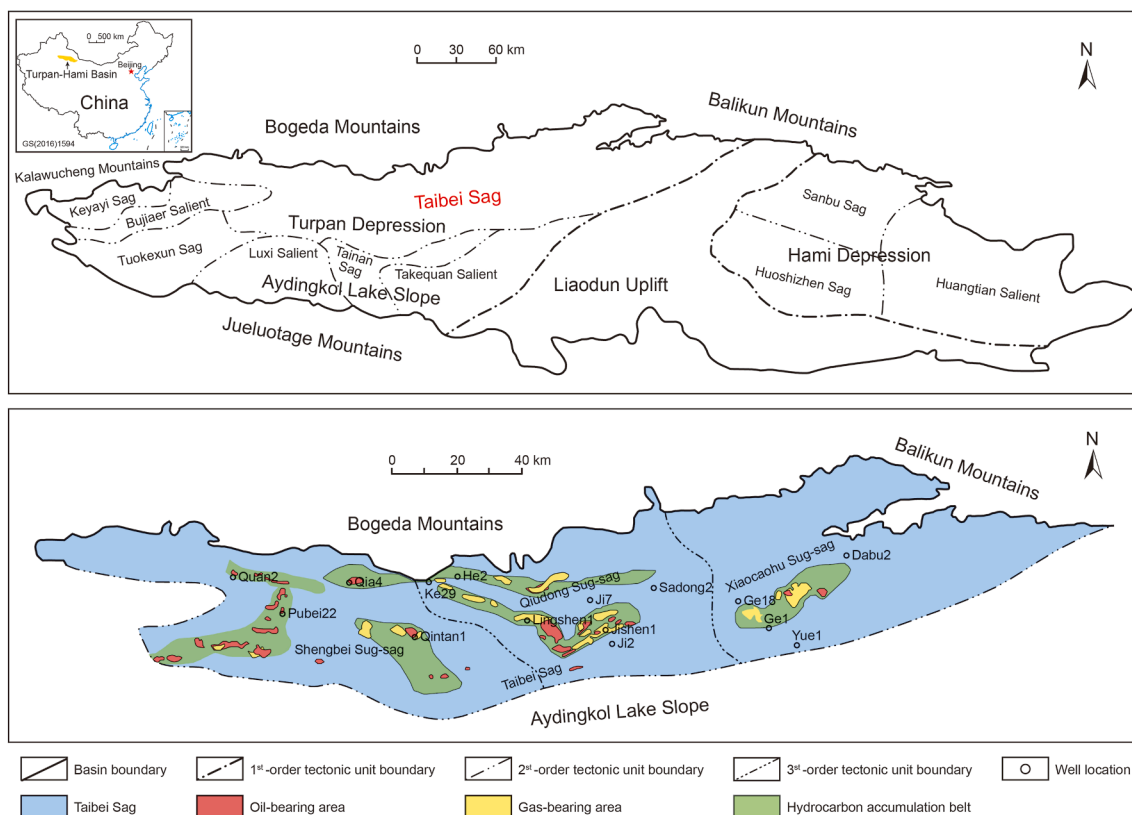


Fig. 1. (a) Location and tectonic units of the Turpan-Hami Basin, NW China (Zhi et al., 2024); (b) distribution of principal oil and gas fields in the Taibei Sag, Turpan Depression (Chen et al., 2024).

Cretaceous, southern basin subsidence occurred alongside northward migration of paleo-highs (Wang et al., 2018; Zhen et al., 2024). The Himalayan orogeny later induced tectonic inversion, uplifting anticlines and generating near N–S strike-slip faults such as the Shengbei faulted anticline (Liu et al., 2023).

The Taibei Sag, a secondary tectonic unit within the Turpan Depression, covers an area of about 10,000 km² (Li et al., 2020). It comprises three smaller sub-sags: the Shengbei, Qiudong, and Xiaocaochu (Fig. 1(b)) (Wang et al., 2023). The Jurassic, which reaches a thickness of up to 4600 m, is predominantly found within the Taibei Sag (Ni et al., 2019). Notably, the Middle–Lower Jurassic Shuixigou Group consists of the Xishanyao, Sangonghe, and Badaowan formations, deposited in varied environments including warm, humid shore-shallow lakes, swamps, deltas, rivers, and braided river deltas (Fig. 2). The Xishanyao and Sangonghe formations are particularly notable for the development of deep tight sandstone gas resources in the basin (Liu et al., 2023).

3. Data and methodology

This study employs core experimental data and logging data to predict porosity in tight sandstone reservoirs. Core experimental data provide key petrophysical parameters, while nuclear magnetic resonance (NMR) logging reflects properties in intervals lacking core data. Nine conventional logging curves are utilized for porosity prediction in wells devoid of both core and NMR data. Subsequently, lithology and conventional logging data serve as inputs for porosity prediction using Support Vector Machine (SVM), Categorical Boosting (CatBoost), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) models, with evaluation of their predictive performance. The SHapley

Additive exPlanations (SHAP) method is introduced to quantify feature contributions and interpret model predictions.

3.1. Data

A total of 1140 samples from four wells (J7, J701H, J702H and J7-2-3H) from the Jurassic Sangonghe Formation in Turpan-Hami Basin were collected to predict the reservoir porosity. Core observations from eleven wells identified that the study area predominantly comprises siltstone, fine-grained sandstone, medium-grained sandstone, coarse-grained sandstone, and sandy conglomerate. Core samples of eight wells were acquired. A total of 101 core samples from eight cored wells were analyzed for helium porosity and air permeability at confining pressures of 800 psi (5.5 MPa) net effective stress.

Standard modern wire-line logs comprise gamma ray (GR), photoelectric factor (Pe), caliper (CAL), spontaneous potential (SP), deep resistivity (RD), shallow resistivity (RS), compensated neutron (CNL), acoustic slowness (DT), and bulk density (DEN). The core-to-log depth calibration correlates the GR log with core descriptions.

3.2. Methods

3.2.1. Support vector machine (SVM)

The Support Vector Machine (SVM), pioneered by Cortes and Vapnik (1995), represents a formidable machine learning technique anchored in statistical learning theory. The quintessential objective of SVM is to either classify or regress by constructing a hyperplane capable of distinguishing between data sets (Huang and Wang, 2006; Deng et al., 2019). By leveraging kernel

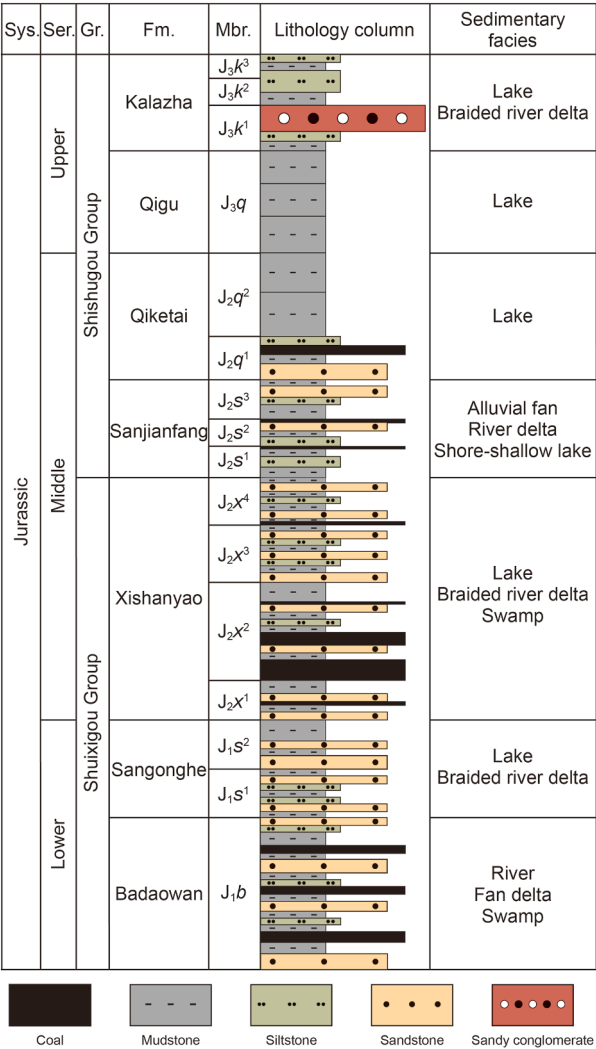


Fig. 2. The lithology and stratification characteristics of the Jurassic strata in the Taipei Sag, Turpan-Hami Basin (modified from Chen et al., 2024).

functions (Cervantes et al., 2020), SVM projects the input features into a higher-dimensional space, transforming a nonlinear problem within the original space into a linear one in this new feature space (Fig. 3) (Dong et al., 2020).

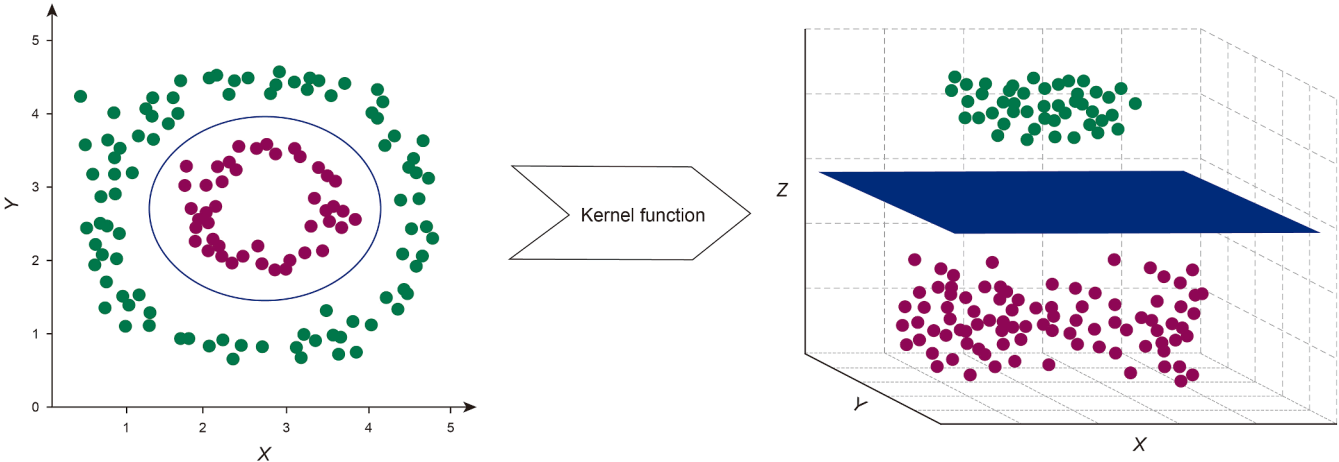


Fig. 3. Visualization of data mapping from input space to feature space via SVM kernel function.

SVM operates under the principle of structural risk minimization, which is part of statistical learning theory (Cervantes et al., 2020). It utilizes an optimization model to derive the decision function expressed as:

$$\min \left[\frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^m (\xi_i + \xi_i^*) \right] \tag{1}$$

where ω represents the weight vector, C is the penalty factor indicative of the misclassification cost, ξ_i and ξ_i^* are the slack variables denoting the error margins. SVM can accommodate various kernel functions, including linear, polynomial, sigmoid, and radial basis function (RBF) kernels (Otchere et al., 2021). Due to the high dimensionality involved in feature extraction, this study employs the RBF kernel function, defined as:

$$K(\mathbf{x}, \mathbf{x}_i) = \exp \left(- \frac{\| \mathbf{x} - \mathbf{x}_i \|^2}{2\sigma^2} \right) \tag{2}$$

where $\mathbf{x}$ and $\mathbf{x}_i$ are sample vectors, and σ is the kernel function's parameter controlling the influence range of a training example.

Characterized by a limited number of hyperparameters, predominantly the penalty factor C and the kernel function parameter σ , the SVM model allows some degree of misclassification to achieve a more generalized decision boundary (Ali et al., 2024). The parameter σ crucially defines the extent of influence of a training sample on the decision boundary, affecting the model's complexity and generalization capability. By strategically selecting the RBF kernel and tuning these parameters, the SVM algorithm adeptly manages the high-dimensional feature space (Fig. 3) pertinent to predicting oil and gas reservoir parameters, thus ensuring robust and precise modeling outcomes (Ao et al., 2019).

3.2.2. Categorical Boosting (CatBoost)

Categorical Boosting (CatBoost) is a gradient boosting decision tree (GBDT) framework that employs symmetric oblivious trees as base learners (Prokhorenkova et al., 2018; Dev and Eden, 2019). It excels in handling categorical variables by automatically converting them into numerical representations, thereby reducing the need for extensive hyperparameter tuning (Wen et al., 2019; Bentéjac et al., 2021). A key innovation of CatBoost is its ordered boosting technique, which mitigates prediction shift and gradient bias caused by training noise, thereby reducing overfitting and enhancing model accuracy and generalization (Tanha et al., 2020).

The dataset for CatBoost is defined as follows:

$$D = (I_x, J_x) \quad (3)$$

where $x = 1, 2, \dots, t$ is the number of sample groups, $I_x = (i_x^1, i_x^2, \dots, i_x^t)$ denotes the sample groups, and i_x^i represents the feature vector of the i -th sample group, and J_x denotes the label value. CatBoost's methodology includes several critical steps (Hancock and Khoshgoftaar, 2020; Mienye and Sun, 2022):

- (1) Random generation of multiple rankings: Multiple random rankings of the dataset are generated for learning.
- (2) Calculation of categorical feature transformation values: For each feature, the algorithm identifies samples within the same class and calculates their categorical feature transformation values:

$$\tilde{i}_x^z = \frac{\sum_{y=1}^t \varphi(i_y^z = i_x^z) J_y + \alpha u}{\sum_{y=1}^t \varphi(i_y^z = i_x^z) + \alpha} \quad (4)$$

where $\tilde{i}_x^z$ is the categorical feature transformation value, φ is the indicator function, which equals 1 when $\{i_y^z = i_x^z\}$ is satisfied and 0 otherwise. J_y represents the label value, u is the prior value, and α is the prior weight.

- (3) Greedy construction of feature combinations: CatBoost employs a greedy strategy to construct feature combinations, improving its capacity to capture complex dependencies. For the corresponding sample group I_x , only one model within the training set incorporates it, while models from other training sets exclude I_x . Using this greedy approach, the classification feature combination is built and the tree structure is chosen. The gradient I_x was computed via the ordered boosting algorithm and subsequently applied to train the weak learner.
- (4) Ordered boosting algorithm: The ordered boosting algorithm calculates gradients and uses them to train the weak learners iteratively.
- (5) Weighted development of the final model: The final model is developed by aggregating the weak learners with appropriate weights.

CatBoost's robust framework suits applications demanding precise handling of categorical data with complex interactions. Its ordered boosting algorithm, combined with automated feature transformation and combination, reduces overfitting and improves model performance (Tang et al., 2021). As a result, CatBoost proves highly effective in predicting oil and gas reservoir parameters, effectively addressing specific challenges in the petroleum industry (Lu et al., 2022).

3.2.3. Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is a class of deep learning models designed to efficiently learn hierarchical spatial representations from grid-structured data (Krizhevsky et al., 2017; Sakhavi et al., 2018). As illustrated in Fig. 4, a CNN architecture typically comprises a sequence of convolutional and pooling layers, followed optionally by fully connected layers for task-specific mapping. In the convolutional layers, learnable kernels (filters) of size $k \times k$ traverse the input matrix with a given stride s , performing element-wise multiplications between kernel weights w and the corresponding input region x ,

summing the results, and adding a bias term b before applying a nonlinear activation function $f(\cdot)$. The general form of the 2-D convolution operation for an output feature map y can be expressed as:

$$y_{ij} = f\left(\sum_{p=1}^P \sum_{q=1}^Q w_{p,q} \cdot x_{i+p-1, j+q-1} + b\right) \quad (5)$$

where p and q are the kernel height P and width Q , i and j are spatial coordinates in the output map. The kernel size directly determines the receptive field, while the stride controls the density of sampling. For an input of size $n \times n$, the feature map dimensions are given by $((n-k)/s+1, (n-k)/s+1)$ (Fig. 4).

Weight sharing and local connectivity, which are the two hallmark design principles of CNNs, significantly reduce the number of trainable parameters compared with fully connected layers, thereby improving computational efficiency and mitigating overfitting (Song et al., 2022; Tian et al., 2024). Early convolutional layers typically capture low-level descriptors such as edges and texture gradients (Fig. 4(a) and (b)), whereas deeper layers progressively abstract high-level structural patterns relevant to the task (Fig. 4(c)). The pooling layers, most commonly implemented as max pooling or average pooling (Fig. 4(d)), subsample the feature maps to remove redundant information, reduce the risk of overfitting, and enhance translation invariance, while preserving the semantic integrity of the learned representations.

In this study, although the original logging data are acquired as one-dimensional sequences along depth, we implemented a structured transformation to enable Two-Dimensional Convolutional Neural Network (2-D CNN) processing. Specifically, during data preprocessing, we defined a fixed depth window of 9 consecutive depth samples (with a sampling interval of 0.125 m in the study area). For each depth window, the logging responses from nine conventional curves were extracted and stacked across the feature axis to form a 9×9 matrix data. This matrix data was used as a single input sample, with the NMR-derived effective porosity value at the final (deepest) depth point in the depth window serving as the corresponding label.

3.2.4. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks extend the conventional recurrent neural network (RNN) architecture by incorporating a gated memory cell that facilitates selective forgetting, conditional updating, and controlled exposure of information at each time step, thereby effectively mitigating the vanishing and exploding gradient problems encountered in long-sequence modeling (Ordóñez and Roggen, 2016; Liu and Guo, 2019; Yang et al., 2023). As shown in Fig. 5, the LSTM cell regulates information flow through three gating mechanisms: the forget gate f_t , the input gate i_t , and the output gate o_t . At each time step, the cell receives the current input x_t , the previous hidden state h_{t-1} , and the previous cell state c_{t-1} . In matrix form, with concatenated weights and biases:

$$\mathbf{W} = \begin{bmatrix} W_f \\ W_i \\ W_c \\ W_o \end{bmatrix}, \mathbf{U} = \begin{bmatrix} U_f \\ U_i \\ U_c \\ U_o \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_f \\ b_i \\ b_c \\ b_o \end{bmatrix} \quad (6)$$

the gate and candidate computations are expressed as follows:

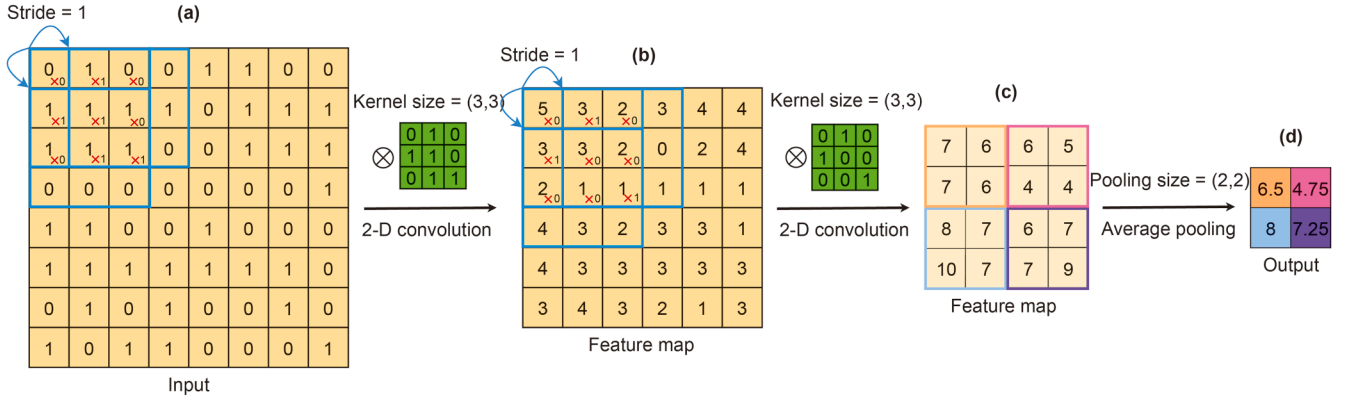


Fig. 4. Illustration of the forward computation in a 2-D CNN with two successive 3×3 convolutional layers (stride = 1, padding = 0): (a) the 8×8 input matrix; (b) first convolution yields a 6×6 feature map; (c) second convolution yields a 4×4 feature map; (d) average pooling with a 2×2 window produces a 2×2 output (modified from Tian et al., 2024).

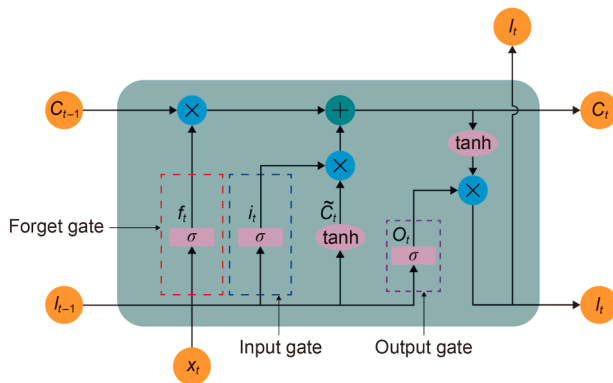


Fig. 5. Schematic representation of the Long Short-Term Memory (LSTM) cell architecture, illustrating the flow of information through the forget, input, and output gates, as well as the additive cell state update and gated hidden state output.

$$\begin{bmatrix} f_t \\ i_t \\ \tilde{c}_t \\ O_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \tanh \\ \sigma \end{bmatrix} (\mathbf{W} \cdot \mathbf{x}_t + \mathbf{U} \cdot \mathbf{l}_{t-1} + \mathbf{b}) \quad (7)$$

where σ is the sigmoid function and $\tanh$ is the hyperbolic tangent function. By adopting an additive update, the cell state maintains sustained gradient propagation across time steps:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (8)$$

and the hidden state exposed to the next layer or to the output is:

$$l_t = O_t \odot \tanh(c_t) \quad (9)$$

Beyond its structural design, LSTM's principal strength lies in its ability to capture both short-term and long-term dependencies in sequential data (Zhang et al., 2024). Its gating mechanisms act as a dynamic filtering system, allowing the network to retain slowly evolving patterns while rapidly incorporating relevant short-term variations (Lan et al., 2025). This capability is especially valuable in domains where temporal context is heterogeneous, such as geophysical logging and reservoir monitoring (Haritha et al., 2025; Zhou et al., 2025). Furthermore, the cell's additive state update mitigates catastrophic forgetting, enabling effective learning on sequences spanning hundreds or thousands of steps without significant degradation in predictive accuracy. By selectively gating

the flow of information, LSTMs can adapt to varying signal-to-noise ratios, focus on salient features, and disregard irrelevant fluctuations, making them particularly well-suited for modeling the temporal data encountered in scientific and engineering applications (Kocoglu et al., 2024; Chen et al., 2025).

3.2.5. SHapley additive exPlanations (SHAP)

SHAP (SHapley Additive exPlanations) is a method derived from cooperative game theory that explains machine learning model outputs by quantifying each feature's contribution to predictions (Mi et al., 2020; Belle and Papantonis, 2021). Based on Shapley values, it offers instance-specific feature impact measurements, enabling detailed interpretation of model decisions (Antwarg et al., 2021). The approach constructs an additive explanation model in which all features contribute to the final prediction, with SHAP values computed as follows (Aas et al., 2021):

$$f(\mathbf{x}_i) = \varphi_0 + \sum_{j=1}^n \varphi_{i,j} \mathbf{z}'_{i,j} \quad (10)$$

Where $x_{i,j}$ is the j -th feature of the i -th sample, n is the total number of samples, $f(\mathbf{x}_i)$ is the predicted outcome for $\mathbf{x}_i$, φ_0 is the baseline value assigned to the model, $\varphi_{i,j}$ is the SHAP value of $x_{i,j}$, and $\mathbf{z}'_{i,j}$ belongs to a feature combination vector in the $\mathbf{z}'_{i,j} \in \{0, 1\}$, which 0 indicates that the vector does not exist and 1 indicates that the vector exists. The calculation of the SHAP value for a feature $x_{i,j}$ is defined as:

$$\varphi_{i,j} = \sum_{\mathbf{S} \subseteq \mathbf{N} \setminus \{x_{i,j}\}} \frac{|\mathbf{S}|!(n - |\mathbf{S}| - 1)!}{n!} [f_{\mathbf{x}_i}(\mathbf{S} \cup \{x_{i,j}\}) - f_{\mathbf{x}_i}(\mathbf{S})], j \geq 1 \quad (11)$$

where $\mathbf{N}$ is the set of all features in the dataset, $\mathbf{S}$ is a subset of $\mathbf{N}$, and its spatial size is denoted by $|\mathbf{S}|$, $f_{\mathbf{x}_i}(\mathbf{S})$ is the value predicted using the samples of $\mathbf{S}$ in the dataset, and when $\mathbf{S}$ is an empty set, $\varphi_0 = f_{\mathbf{x}_i}(\mathbf{S})$ is the base value. This calculation thoroughly accounts for the influence of all possible subsets of features and their contributions to the prediction, ensuring a comprehensive measure of each feature's impact.

Incorporating SHAP into model evaluation enables detailed analysis of feature importance and interactions (Ekanayake et al., 2022). This approach is especially valuable in oil and gas reservoir parameter prediction, as it clarifies the contribution of

geological and geophysical features, thereby improving result interpretability and reliability (Li, 2022).

4. Results

This section outlines the construction of the porosity prediction dataset and the associated modeling workflow. NMR-derived effective porosity and conventional logging data from four wells are preprocessed using lithology-based filtering and normalization. Lithology is encoded as a categorical input to reflect its control on porosity. Four machine learning models are developed and optimized using the Optuna framework. Model performance is evaluated using multiple regression metrics, and generalization is assessed through blind well testing. Finally, SHAP analysis is applied to quantify feature importance and interpret the influence of lithology and logging parameters on porosity prediction.

4.1. Lithology, pore systems and reservoir quality

4.1.1. Lithological characteristics and pore systems

Core observations reveal that the Jurassic Sangonghe Formation in the Taibei Sag exhibits complex and diverse lithologies with rapid vertical variations, pronounced heterogeneity, and the presence of various sedimentary structures. The lithologies consist of mudstones, siltstone, fine-grained sandstone, medium-grained sandstone, coarse-grained sandstone, and sandy conglomerate (Fig. 6). Siltstone contains minor argillaceous components, displaying visible parallel lamination (Fig. 6(a)). Fine-grained, medium-grained, and coarse-grained sandstones predominantly

exhibit massive structures (Fig. 6(b)–(d)), with certain cores showing parallel lamination, cross-bedding, or wedge-shaped cross-bedding. Sporadic gravels are dispersed within some medium-grained and coarse-grained sandstones (Fig. 6(e) and (f)). Sandy conglomerate, characterized by the coarsest grain size (Fig. 6(g)), is primarily developed at the base of channel deposits, indicating strong hydrodynamic energy. Upward, it is typically overlain by coarse-grained sandstone and medium-grained sandstone, forming an overall fining-upward sequence reflecting diminishing hydrodynamic energy. Variations in hydrodynamic intensity facilitate the development of multiple fining-upward sequences, grading from coarse-grained to fine-grained sediments from base to top.

Detailed microscopic analyses (Fig. 7), including both thin sections and scanning electron microscopy (SEM), reveal that the pore system in the Jurassic Sangonghe Formation primarily consists of intergranular and intragranular dissolution pores, with micropores and microfractures also playing crucial roles (Fig. 7(a)–(f)). Intergranular pores, often formed along the peripheries of clastic grains and linked via grain-edge joints, are frequently accompanied by sporadic primary intergranular pores (Fig. 7(a) and (b)). Intragranular dissolution pores are characterized by the selective dissolution of soluble components within feldspar or rock fragments, primarily developing inside the particles (Fig. 7(c)–(e)). Furthermore, micro-fractures, often narrower than 0.1 mm and not detectable in core samples, are visible under microscopic (Fig. 7(f)). SEM analysis has categorized micropores into two main types: those associated with feldspar dissolution and those dominated by clay minerals (Fig. 7(g) and (h)). SEM



Fig. 6. Lithological characteristics of the Sangonghe Formation in the Taibei Sag, Turpan-Hami Basin. (a) Siltstone, 5313 m, J₁s, well J702H; (b) medium-grained sandstone, 5304.4 m, J₁s, well J702H; (c) coarse-grained sandstone, 5326.74 m, J₁s, well J702H; (d) fine-grained sandstone, 5327.18 m, J₁s, well J702H; (e) coarse-grained sandstone with a small amount of gravel, 5249.77 m, J₁s, well J702H; (f) medium-grained sandstone with a small amount of gravel, 5316.14 m, J₁s, well J702H; (g) conglomeratic sandstone, 5256.5 m, J₁s, well J702H.

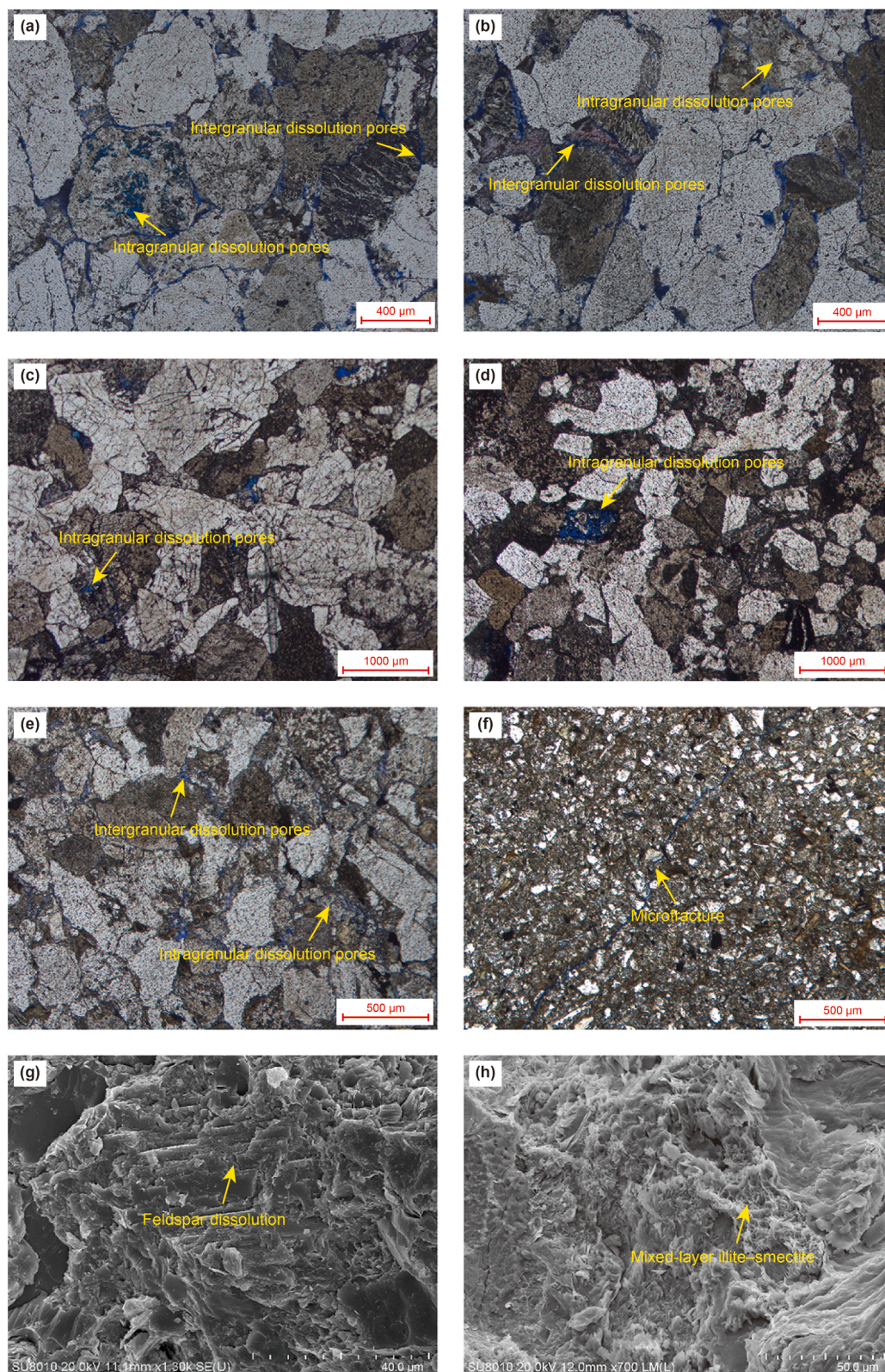


Fig. 7. Thin section and SEM images illustrating the lithologies and pore systems of the Jurassic Sangonghe Formation tight sandstones in the Taibei Sag, Turpan-Hami Basin. **(a)** Intergranular and intragranular dissolution pores, 5242.30 m, J₁s, well J7-2-3H (–); **(b)** intergranular and intragranular dissolution pores, 5242.30 m, J₁s, well J7-2-3H (–); **(c)** intragranular dissolution pores, 5141.97 m, J₁s, well J701H (–); **(d)** intragranular dissolution pores, 5135.39 m, J₁s, well J701H (–); **(e)** intergranular and intragranular dissolution pores, 5309.49 m, J₁s, well J7-2-3H (–); **(f)** microfracture, 5287.90 m, J₁s, well J7-2-3H (–); **(g)** feldspar dissolution, 5305.49 m, J₁s, well J702H, SEM; **(h)** micropores in mixed-layer illite-smectite, 5315.14 m, J₁s, well J702H, SEM.

analysis reveals a plethora of intergranular pores originating from the dissolution of feldspars and rock fragments (Fig. 7(g)), as well as significant microporosity within authigenic clay minerals like illite and mixed-layer illite-smectite, indicating a complex, albeit compact, pore network. Thin-section and SEM observations show that medium- to coarse-grained sandstones commonly contain abundant intergranular and intragranular dissolution pores, lower matrix contents, and better pore-throat connectivity, thereby forming high-quality reservoirs. By contrast, siltstone, fine-grained sandstone, and sandy conglomerate typically have higher clay contents and poorer pore development, resulting in poorer reservoir quality (Fig. 7).

4.1.2. Reservoir quality

The Jurassic Sangonghe Formation tight sandstones exhibit notably constrained porosity, ranging from 1.8% to 5.9% with an average of 4.03%. These formations also demonstrate exceptionally low permeability, from 0.002 mD to 0.7 mD, averaging around 0.085 mD (Fig. 8). These data highlight the deep tight sandstone reservoir of the Jurassic Sangonghe Formation as having poor reservoir quality due to its extreme heterogeneity. Notably, there is a significant variance in permeability even at constant porosity levels, where minor changes in porosity can result in permeability adjustments spanning one to three orders of magnitude (Fig. 8). This variability underscores that permeability is influenced by factors beyond porosity, including the intricate microscopic pore structures and the presence of microfractures (Fig. 7). The weak correlation between permeability and porosity (correlation coefficient: $R^2 = 0.52$) further attests to the dominant influence of pore structure over simple porosity measurements (Fig. 8). In summary, the deep tight sandstone in the Jurassic Sangonghe Formation is characterized by low porosity, very low permeability and pronounced heterogeneity, with lithofacies exerting a dominant control on reservoir quality parameters such as porosity. Medium- and coarse-grained sandstones exhibit the most favorable porosity and permeability among the five dominant lithologies.

4.2. Petrophysical properties data features

This study assembles a four-well dataset from the Taibei Sag, with NMR-derived effective porosity as the supervised label and

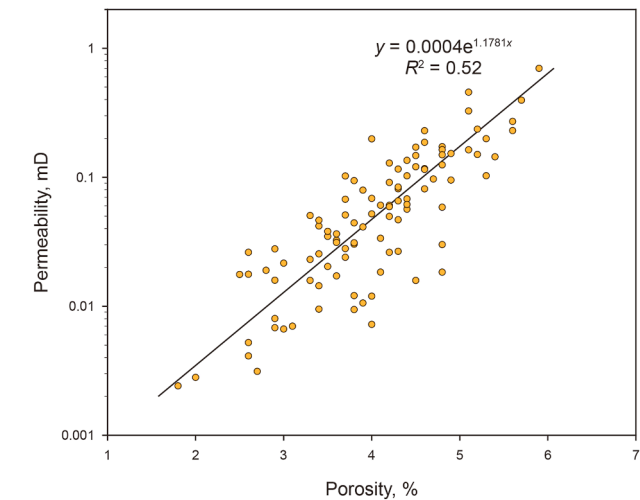


Fig. 8. Cross-plot of core porosity versus core permeability in Jurassic Sangonghe Formation tight sandstones, demonstrating that minor variations in porosity correspond to significant permeability changes across several orders of magnitude in the Taibei Sag, Turpan-Hami Basin.

nine conventional logging data as input data. To ensure the reliability of the NMR-derived effective porosity, core-calibrated cutoff values of the T_2 distribution were established by matching the NMR logging with core experimental data. Consequently, the NMR-derived effective porosities used as labels in this study exhibit a strong correspondence with core porosity, with minimal error, validating their suitability as training targets for machine learning models. In addition, intervals dominated by mudstone were excluded to maintain geological relevance, and caliper-based quality control was applied to remove intervals affected by significant borehole enlargement. The resulting dataset retains only reservoir intervals suitable for modeling (Table 1).

Fig. 9 illustrates the probability density distributions of nine conventional logging curves, along with effective porosity, across four wells: J7, J701H, J702H, and J7-2-3H. These distributions offer critical insights into the petrophysical properties of the wells, underscoring the variability and prevalent trends within the data. Notably, the natural gamma logging values predominantly range between 40 and 80 API, indicative of coarse-grained sandstone in the analyzed statistical intervals. Concurrently, the photoelectric factor (Pe) values are primarily concentrated between 4 and 8 b/e, characteristic of sandstone reservoirs. This tight clustering across the wells suggests consistent lithological characteristics. Moreover, the density logs are closely grouped around 2.55 to 2.65 g/cm³, typical for quartz-rich sandstone. The slight variations in density across the wells offer valuable insights into porosity variations and fluid content within the reservoir. Additionally, the sonic transit time log displays a broad distribution, particularly in well J7, pointing to significant lithological variations. Faster transit times typically imply denser, less porous lithology, whereas slower times might suggest higher porosity. Resistivity logs exhibit a wide range of values, with lower resistivity generally associated with more conductive (water-saturated) formations and higher resistivity indicative of the presence of hydrocarbons or more resistive lithologic types. The effective porosity distributions mostly show peaks at lower porosity values, especially in wells J701H and J702H, with a broader distribution in J7—this pattern suggests more pronounced variation in pore space connectivity and size. Furthermore, the caliper log demonstrates significant variation across the wells, signaling differences in borehole size and potentially conditions such as borehole enlargement due to drilling-induced stress or formation weakness. These distributions are skewed towards higher caliper readings, particularly in wells J701H and J702H, hinting at borehole instability in these areas. To preserve the integrity of the logging data, sections displaying significant borehole diameter variations, likely due to phenomena such as caving or enlargement, have been meticulously excluded from the analysis.

In addition to conventional logging data, the lithology was also incorporated as an input feature to enhance the predictive capacity of the models. Unlike continuous well-log measurements, lithology varies as a categorical variable along the vertical profile of each well, reflecting discrete depositional units. To enable model ingestion, we encoded the five lithology classes (siltstone, fine-

Table 1
Summary of wells, depth intervals, sample counts, and data partition for the effective porosity dataset of the Jurassic Sangonghe Formation in the Taibei Sag, Turpan-Hami Basin.

Well	Depth, m	Samples	Data type
J7	5435–5547.25	405	Training dataset
J701H	5372.375–5458.75	236	Training dataset
J702H	5126–5228.375	309	Training dataset
J7-2-3H	5210–5233.625	190	Test dataset

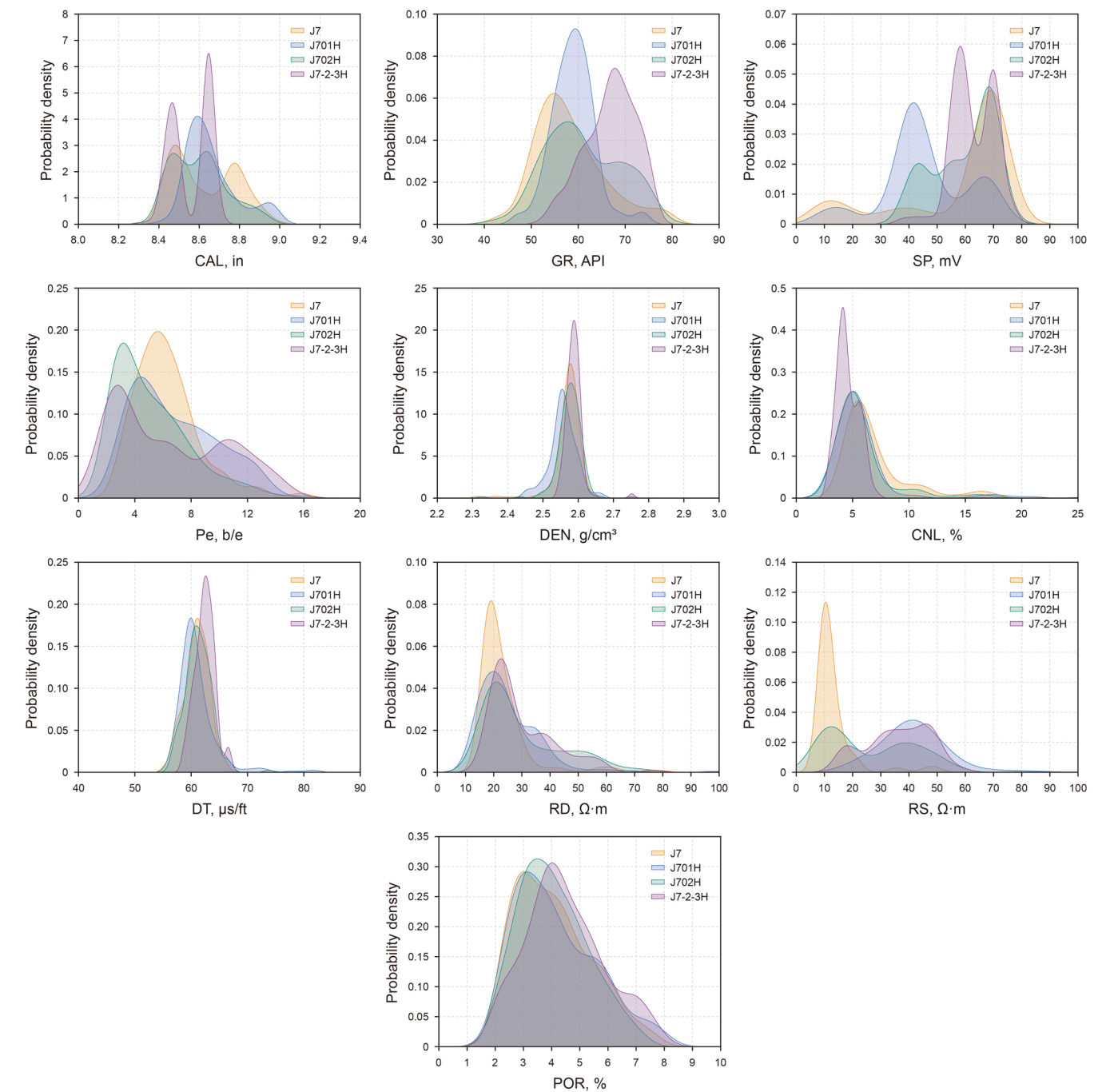


Fig. 9. Probability density distribution of nine conventional log curves and effective porosity of the Jurassic Sangonghe Formation in the Taibei Sag, Turpan-Hami Basin.

grained sandstone, medium-grained sandstone, coarse-grained sandstone, and sandy conglomerate) into integers ranging from 0 to 4, as summarized in Table 2. This lithology encoding was

Table 2
Lithology encoding and distribution in the model input data of the Jurassic Sangonghe Formation in the Taibei Sag, Turpan-Hami Basin.

Lithology	Code	Samples	Wells covered
Siltstone	0	90	J7, J7-2-3H
Fine-grained sandstone	1	193	J7, J701H, J702H, J7-2-3H
Medium-grained sandstone	2	324	J7, J701H, J702H, J7-2-3H
Coarse-grained sandstone	3	387	J7, J701H, J702H, J7-2-3H
Sandy conglomerate	4	146	J701H, J702H, J7-2-3H

concatenated with the numerical logging features at each depth point to form the complete input vector for model training. The inclusion of lithology not only introduces valuable geological context but also allows tree-based and deep learning models to learn class-conditioned responses that are essential for characterizing tight, heterogeneous reservoirs.

Fig. 10 depicts the Pearson correlation matrix heatmap, which illustrates the relationships between various well logs and effective porosity within the study area. As shown in the figure, the correlation coefficients between conventional logs and NMR-derived effective porosity are uniformly low, all falling below 0.2. This weak correlation reflects the subdued logging responses of the Jurassic Sangonghe Formation, likely caused by strong

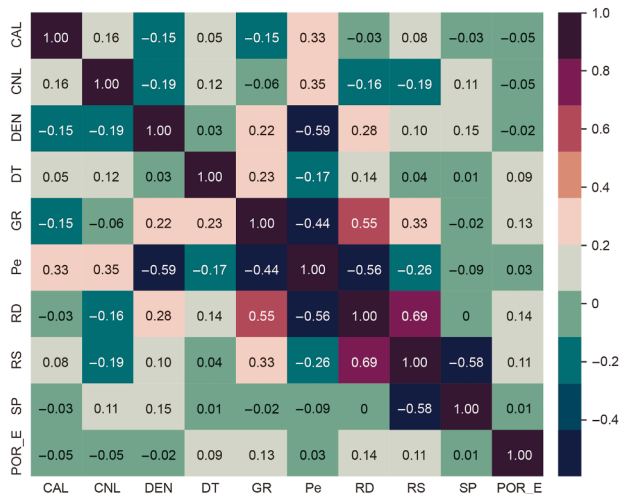


Fig. 10. Correlation coefficient matrix heat map displaying the correlation between nine logging data, namely CAL, CNL, DEN, DT, GR, Pe, RD, RS, SP, and effective porosity data, POR_E.

compaction. Consequently, conventional regression models that rely on one or a few logging curves are unlikely to provide accurate porosity predictions in this setting, thereby underscoring the need for more advanced, nonlinear modeling approaches.

4.3. Training process and evaluation metrics

Appropriate model-architecture design and systematic hyperparameter tuning are essential for achieving accurate predictions. To establish a nonlinear mapping between conventional well logs and effective porosity for accurate porosity prediction, four machine learning algorithms, namely SVM, CatBoost, 2-D CNN and LSTM, were implemented. Moreover, to obtain optimal hyperparameters and maximize regression performance, we employed Optuna, a Python-based open-source hyperparameter optimization framework that utilizes efficient search algorithms (e.g., TPE, CMA-ES) within predefined search spaces to automatically identify the best configurations. To eliminate inconsistencies in magnitude and range and improve model convergence, the logging data were normalized using the Min-Max scaling method. Each logging curve was rescaled to the range of [0, 1] based on its minimum and maximum values observed within the dataset.

For the deep learning models, the mean squared error (MSE) loss function was employed to quantify the discrepancy between predicted and true porosity values. To mitigate the risk of gradient vanishing during training, we incorporated early stopping and an adaptive learning rate strategy. Specifically, the Adam optimizer, known for its robustness in handling sparse gradients, was adopted with an initial learning rate of 0.01. If the validation error did not decrease for 50 consecutive epochs, the learning rate was reduced by 2% of the initial learning rate, and training continued with the updated learning rate. The training process was terminated when no further improvement was observed over 300 epochs, at which point the model weights corresponding to the lowest validation error were preserved. All experiments were implemented in PyTorch and trained on an NVIDIA RTX 4060Ti GPU with 16 GB of memory. After systematic hyperparameter tuning, the finalized hyperparameters for all four models are reported in Table 3. The SVM model is relatively simple, involving only two hyperparameters ($C = 25$ and $\gamma = 0.06$), which highlights its low model complexity and limited flexibility in handling nonlinear relationships. In contrast, as a tree-based

gradient boosting model, CatBoost has the most complex hyperparameter set, including a high tree depth (14), 5000 boosting rounds, a learning rate of 0.18, and additional regularization and sampling parameters, reflecting its ability to robustly handle categorical features such as lithology and to model nonlinear interactions among logging data. For the two deep learning models, the final network architectures and hyperparameters were determined through extensive ablation experiments combined with automated hyperparameter tuning. The 2-D CNN model consists of three convolutional layers with 32, 64, and 64 filters, respectively, using kernel sizes of 3×3 , 2×2 , and 2×2 . The LSTM employs a four-layer recurrent stack with decreasing hidden sizes (32, 32, 16, 16) to capture depth-wise sequential dependencies while progressively limiting capacity to mitigate overfitting, and is followed by two fully connected layers (with 32 and 8 units, respectively). In both models, a dropout rate of 0.2 is applied between layers as a regularization measure. Notably, both architectures conclude with a single-unit fully connected output layer with a linear activation function that performs the final regression mapping to effective porosity.

The predictive performance of the four models (SVM, CatBoost, 2-D CNN, LSTM) for NMR-derived effective porosity is evaluated using mean absolute error (MAE), mean absolute percentage error (MAPE), mean squared error (MSE) and the coefficient of determination (R^2). These regression metrics are defined as follows:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i| \quad (12)$$

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^N \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (13)$$

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2 \quad (14)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{\sum_{i=1}^N (y_i - \text{aver}(y_i))^2} \quad (15)$$

where y_i and $\hat{y}_i$ are the true and predicted porosity values, respectively, $\text{aver}(y_i)$ is the mean of the true porosity values, and N is the number of samples. Given the pronounced heterogeneity of deep tight sandstones, relying on a single metric can be misleading. MAE reflects typical absolute error, MAPE expresses relative error with respect to scale, MSE emphasizes large deviations, and R^2 quantifies the fraction of variance explained. A higher R^2 together with lower MAE, MAPE and MSE indicates superior predictive accuracy.

The complete modeling workflow for predicting reservoir parameters in the deep tight sandstones of the Jurassic Sangonghe Formation is illustrated in Fig. 11. The process begins with data preprocessing, where non-reservoir intervals are excluded based on lithology classification and borehole quality control. A total of 1140 depth-sampled data points from four wells (J7, J701H, J702H, and J7-2-3H) are retained for analysis. Basic analysis, correlation assessment, and feature normalization are conducted to prepare the dataset for modeling. Specifically, wells J7, J701H, and J702H are assigned as the training set, while J7-2-3H serves as the blind test well for validation. Then four representative machine learning models, namely SVM, CatBoost, 2-D CNN, and LSTM are constructed to capture the complex nonlinear relationships between conventional well logs and NMR-derived effective porosity. Each model undergoes automated hyperparameter tuning using the Optuna framework, which utilizes efficient search algorithms such as TPE and CMA-ES to identify optimal configurations within the

Table 3
Optimal hyperparameters of four machine learning models.

Models	Hyperparameters	Optimization range	Optimal hyperparameters
SVM	C	[0.01, 100]	25
	Gamma	[0.01, 100]	0.06
CatBoost	Learning_rate	[0.001, 10]	0.18
	Iterations	[500, 5000]	5000
	l2_leaf_reg	[0, 1]	0.8
	Colsample_bylevel	[0, 1]	0.1
	Max_depth	[5, 16]	14
	Bootstrap_type	[Bayesian, Bernoulli, MVS]	Bayesian
	Min_data_in_leaf	[1, 20]	17
	Bagging_temperature	[0, 10]	4
2-D CNN	Batch_size	[16, 64]	32
	Conv1_filters	[16, 128]	32
	Conv1_kernel_size	[2, 5]	3×3
	Stride1	[2, 3]	2
	Conv2_filters	[16, 128]	64
	Conv2_kernel_size	[2, 3]	2×2
	Stride2	–	1
	Conv3_filters	[16, 128]	64
	Conv3_kernel_size	–	2×2
	Stride3	–	1
	Fc1_units	[32, 512]	128
	Fc2_units	[16, 256]	32
	Fc3_units	–	1
	Batch_size	[16, 128]	32
LSTM	Lstm1_units	[8, 128]	32
	Lstm2_units	[8, 128]	32
	Lstm3_units	[8, 128]	16
	Lstm4_units	[8, 128]	16
	Fc1_units	[8, 64]	32
	Fc2_units	[8, 64]	8
	Fc3_units	–	1

defined parameter spaces. During training, model performance is quantitatively assessed using regression metrics, primarily mean squared error (MSE). If the predefined accuracy threshold is not met, the optimization loop continues with adjusted search ranges or re-initialized parameters. Once the optimal performance is achieved on the test dataset, all models are applied to an additional blind well, whose data are excluded entirely from both the training and test sets, to evaluate the four models' generalization capability. Additionally, to investigate the internal decision-making mechanisms of the models, we apply SHAP algorithm to the best performing machine learning model to quantify feature importance and enhance interpretability, with particular emphasis on the influence of lithology and key well logs on porosity predictions.

5. Discussion

5.1. Comparative evaluation of porosity prediction models

In this study, a total of 190 samples from Well J7-2-3H were designated as the test dataset, while data from the remaining three wells (J7, J701H, and J702H) were used for model training. The selection of features, lithology encoding, model architecture design, training strategies, and evaluation metrics were comprehensively described in section 4. Building on this foundation, the current section conducts both qualitative and quantitative comparisons of the four models, SVM, CatBoost, 2-D CNN, and LSTM to assess their predictive performance in estimating effective porosity. Through visual inspection of prediction trends and statistical evaluation of fitting accuracy, we aim to elucidate each model's robustness, generalization capacity, and applicability in the context of deep tight sandstone reservoir characterization.

Fig. 12 presents point-by-point comparisons between predicted porosity (in teal) and NMR-derived effective porosity (in red) for the test samples. Each panel shows the same 190 samples, allowing a direct visual assessment of local errors. The vertical gap between paired markers reflects the residual for each sample. As shown in Fig. 12, CatBoost yields the closest alignment overall. Predicted points track the measured sequence with small and relatively uniform gaps, including around local peaks and troughs, which indicates good stability across the full porosity range. LSTM follows the measured trend well and maintains modest errors for most samples, although slight underestimation is visible at high-porosity excursions and occasional over-smoothing occurs at sharp transitions. The 2-D CNN shows larger departures than LSTM, with a noticeable tendency to smooth local extremes; several peaks and lows are not fully captured, producing wider residuals near turning points. In contrast, SVM performs the worst, with frequent large gaps and clear phase lags relative to the measured sequence, indicating a limited capacity to model the nonlinear, lithology-conditioned logging–porosity relationship.

The scatter plots of the predicted versus effective porosity further illustrate that the CatBoost model delivers the most accurate performance on the test dataset (Fig. 13). Its predictions exhibit the lowest error magnitude, with the corresponding data points densely clustered around the 1:1 reference line, yielding a coefficient of determination (R^2) of 0.94. Moreover, the majority of CatBoost-derived estimates fall within the $\pm 0.5\%$ error band delimited by the two light-green dashed lines. Although both the LSTM and 2-D CNN models display commendable predictive capability, their prediction errors are comparatively larger, particularly within the porosity range of 2%–4%.

To further quantify and compare predictive performance, four statistical metrics were computed for all models, as summarized in Table 4. CatBoost exhibits superior accuracy and consistently outperforms the other methods across all criteria. Direct visual comparisons in Fig. 14 illustrate these findings: CatBoost yields an MAE of only 0.24%, significantly lower than LSTM (0.50%), 2-D CNN (0.87%), and SVM (1.45%) (Fig. 14(a)). Likewise, its MAPE is markedly reduced to 6.87%, versus 8.81%, 11.85%, and 17.48% for LSTM, 2-D CNN, and SVM, respectively (Fig. 14(b)). With respect to MSE, SVM exhibits the largest error, whereas the deep learning models, particularly LSTM, achieve competitive results; nevertheless, CatBoost still delivers the lowest MSE (Fig. 14(c)). Notably, the tree-based CatBoost attains an R^2 of 0.94, substantially exceeding that of the advanced deep learning baseline model, i.e., 2-D CNN ($R^2 = 0.83$) (Fig. 14(d)). Although LSTM ranks best among the baseline models, CatBoost further improves performance by reducing MAE, MAPE, and MSE by 52.00%, 22.02%, and 55.17%, respectively, and elevating R^2 by 8.05%. These results unequivocally demonstrate that CatBoost provides superior accuracy, reliability, and robustness for porosity prediction in deep tight sandstones of the Jurassic Sangonghe Formation.

5.2. A comparative analysis of porosity prediction models in core well profile

In this section, the previously established porosity prediction models were applied to a field well in the Turpan-Hami Basin to verify their applicability in deep tight sandstone reservoirs. As illustrated in Fig. 15, the CatBoost, LSTM, and 2-D CNN models were employed to predict NMR-derived effective porosity in a blind well. The first five tracks contain nine conventional logging curves that serve as model input data, including gamma ray (GR), photoelectric factor (Pe), caliper (CAL), spontaneous potential (SP), deep resistivity (RD), shallow resistivity (RS), compensated neutron (CNL), acoustic slowness (DT), and bulk density (DEN).

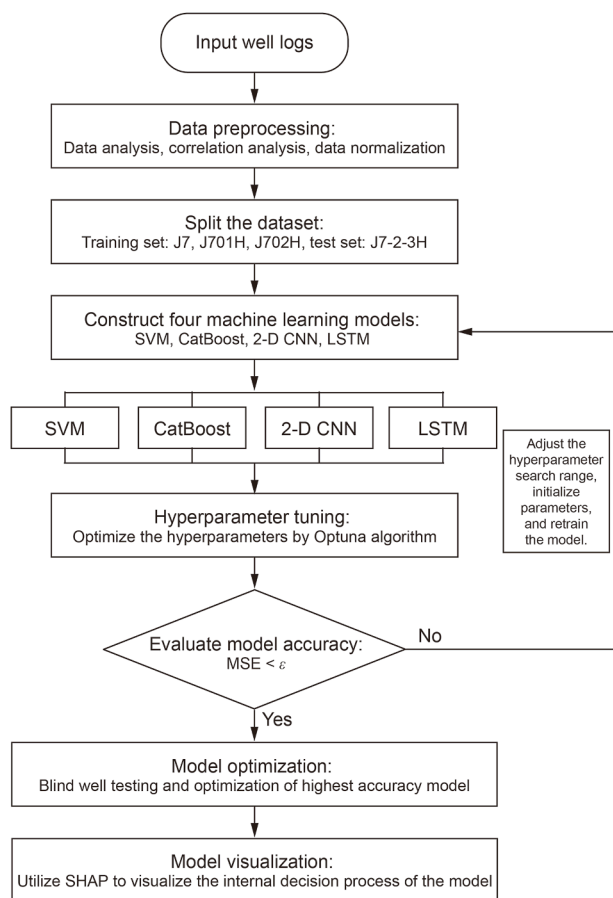


Fig. 11. Comprehensive workflow for porosity prediction and interpretability: data preprocessing, model development with Optuna tuning, blind well validation, and SHAP analysis.

Tracks 6 and 7 present the FMI static and dynamic images, while Track 8 displays the NMR T_2 spectrum from which effective porosity is computed. Tracks 9 to 11 illustrate the porosity prediction results from CatBoost, LSTM, and 2-D CNN models, respectively, with reference to core measurements and NMR-derived effective porosity. The final track shows the lithofacies profile, which is also part of the input data. A comparative assessment reveals that the porosity predicted by the CatBoost model exhibits the highest consistency with the NMR-derived effective porosity, demonstrating a closer match across the full depth interval. The LSTM model shows moderate prediction accuracy, with notable fluctuations along the wellbore, while the 2-D CNN model produces the largest deviation from the effective porosity.

Fig. 16 and Table 5 provide a detailed performance comparison of the three models in the blind well. The scatter plot of CatBoost predictions shows a tight clustering of points around the 1:1 reference line (Fig. 16(a)), achieving a coefficient of determination (R^2) of 0.901, confirming its superior accuracy in predicting porosity in deep tight sandstone reservoirs. The LSTM model yielded an R^2 of 0.845, with underestimation at higher porosity ranges and overestimation at lower values (Fig. 16(b)). In contrast, the 2-D CNN model demonstrated the poorest performance, with an R^2 of only 0.654, and a wide scattering of prediction points indicating limited reliability in capturing porosity characteristics in complex reservoir settings (Fig. 16(c)). In terms of prediction errors, the CatBoost model achieved the lowest values across all

three metrics: MAE (0.354%), MAPE (7.993%), and MSE (0.582%). The LSTM model followed with moderately increased errors (MAE = 0.784%, MAPE = 9.841%, MSE = 1.415%), while the 2-D CNN model showed the highest errors (MAE = 1.674%, MAPE = 19.35%, MSE = 4.560%). These scatter plots and evaluation metrics collectively affirm the superior prediction accuracy, stability, and generalizability of the CatBoost model, confirming its practical applicability for porosity prediction under the complex geological conditions of the study area.

5.3. Impact of discrete lithology data on model predictions

In unconventional reservoirs such as deep tight sandstones, relying solely on advanced model architectures and learning algorithms is insufficient for accurate reservoir parameter prediction. Embedding appropriate geological prior constraints with petrophysical consistency, such as lithologic classifications, into the input features often promotes significant improvements in prediction accuracy and model robustness (Feng et al., 2018; Huang et al., 2025). In the study area, lithologic heterogeneity is pronounced, with interbedded sequences of mudstone, siltstone, fine-grained sandstone, medium-grained sandstone, coarse-grained sandstone, and conglomerate contributing to complex petrophysical behaviors. These lithologic units exhibit significant variability in grain size distribution, compaction, cementation, and pore-throat structures (Wang et al., 2023), all of which critically impact the logging response and subsequently the model's ability to generalize across different lithology types. In our experiments, despite exhaustive preparation that included careful architecture refinement and comprehensive hyperparameter tuning to give each algorithm its best attainable performance, the comparative results show that CatBoost still surpasses the more sophisticated deep learning models (LSTM and 2-D CNN). This outcome is noteworthy because prior studies have repeatedly demonstrated strong performance of LSTM and 2-D CNN for reservoir parameter prediction. Nevertheless, under identical inputs, CatBoost attains higher accuracy, suggesting that the effective exploitation of input features is more critical than architectural complexity within the study area.

During the blind well test presented in Fig. 15, although LSTM performed worse than CatBoost overall, it produced closer agreement with effective porosity in specific depth intervals and thereby demonstrates considerable generalization capability. For instance, in the 5315–5322 m interval, the LSTM predictions are more closely aligned with the effective porosity, whereas CatBoost exhibits larger deviations. A similar pattern is also observed in the 5342–5346.5 m interval. Upon integrating geological interpretations and conducting a detailed analysis of the logging curve characteristics, lithological variations were identified as a key explanatory factor for this discrepancy. The deep tight sandstone reservoirs in the study area are characterized by strong vertical heterogeneity and frequent lithological transitions, which pose significant challenges for model prediction. Under such conditions, the lithology type embedded as an input feature plays a critical role, and the extent to which each model utilizes this information becomes a decisive factor in prediction accuracy. Consequently, in zones with rapid lithologic variation, such as the 5307–5315 m and 5348–5361 m intervals, CatBoost clearly outperforms LSTM. Within these lithologically heterogeneous intervals, the LSTM model fails to establish a high-fidelity nonlinear mapping between input logging data and porosity, whereas the CatBoost model effectively captures the relationships between input features and porosity under varying lithologic constraints, yielding more accurate porosity predictions.

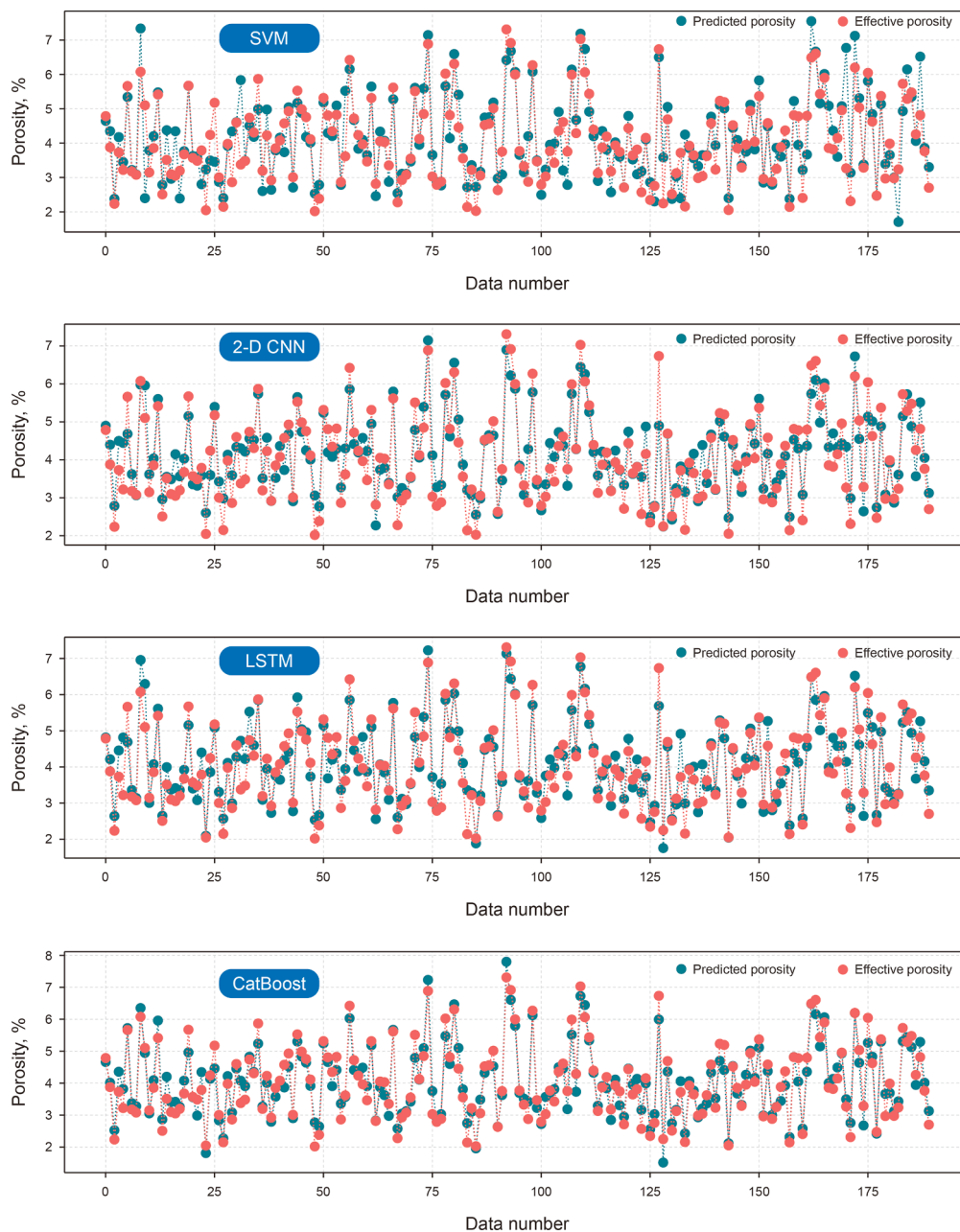


Fig. 12. Comparison of predicted porosity and effective porosity for the four models on the test dataset of the Jurassic Sangonghe Formation in the Taibei Sag, Turpan-Hami Basin.

According to lithology and petrophysical analysis (Fig. 17), a clear correspondence is observed between lithology types and porosity distribution in the deep tight sandstone reservoirs of the Jurassic Sangonghe Formation. Medium-grained and coarse-grained sandstones exhibit relatively high porosity, with the latter showing a slightly superior porosity range. While medium-grained sandstones present a broad porosity distribution, including both low and high values (2%–8%), coarse-grained sandstones tend to be concentrated in the higher porosity domain (5%–9%). Fine-grained sandstones exhibit noticeably lower porosity than the medium-grained and coarse-grained sandstones, with values primarily concentrated around the median range (3%–6%). In comparison, siltstones and conglomerates both demonstrate lower porosity levels overall (<4%). Each lithology displays a specific porosity range

(Fig. 17(a) and (b)). Although partial overlap exists, mainly because different pore structures develop within each lithology, as suggested by the geometric mean of T_2 (T_{2gm}) in Fig. 17(a). Nonetheless, dominant lithologies such as medium-grained and coarse-grained sandstones still exhibit comparatively distinct porosity ranges (Fig. 17(b)). Therefore, incorporating lithology data that exhibit such distribution characteristics as input features can be regarded as implicitly embedding a form of soft petrophysics constraints into the input (Sacchi et al., 2016). This petrophysical embedding significantly enhances the model's predictive performance by guiding it toward geologically plausible solutions.

To further assess the performance of each model under varying lithological conditions, we calculated the average prediction errors (MSE) of the four models within each lithological

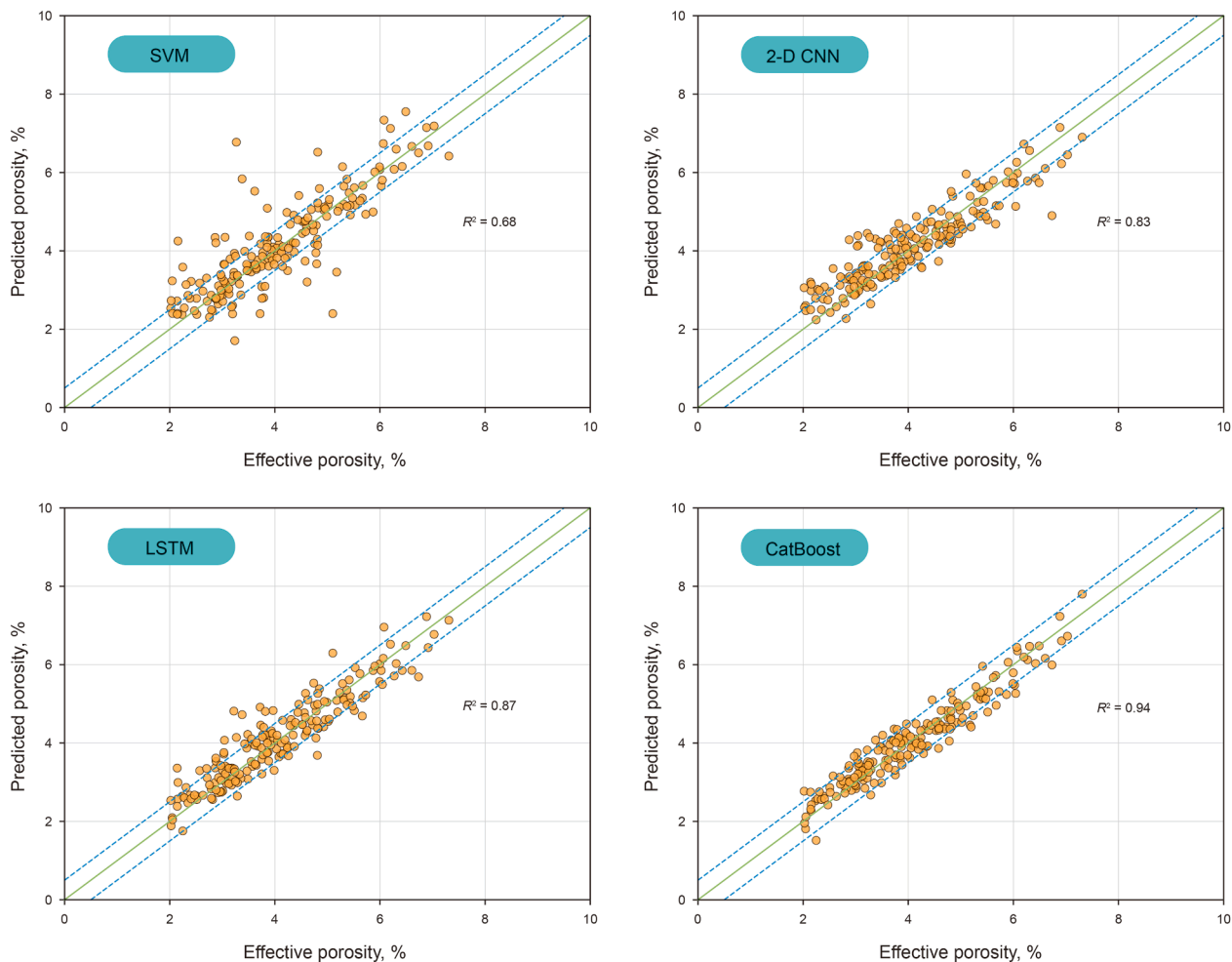


Fig. 13. Scatter plots of predicted versus NMR effective porosity for the four models on the test dataset of the Jurassic Sangonghe Formation in the Taibei Sag, Turpan-Hami Basin.

Table 4
Quantitative comparison of four statistical metrics across the evaluated models on the test dataset.

Models	MAE, %	MAPE, %	MSE, %	R^2
SVM	1.45	17.48	3.96	0.68
2-D CNN	0.87	11.85	1.55	0.83
LSTM	0.50	8.81	0.87	0.87
CatBoost	0.24	6.87	0.39	0.94

class in the test dataset. As shown in Fig. 18, CatBoost consistently achieved the lowest prediction errors across all lithological types, particularly in medium-grained and coarse-grained

sandstones, underscoring its robustness in capturing complex reservoir heterogeneity. This superior performance highlights the advantage of incorporating lithological information as a soft constraint into the predictive modeling process. The two deep learning models, LSTM and 2D CNN, exhibited intermediate performance. Notably, the LSTM model exhibited prediction errors in fine-grained sandstones that were comparable to those of CatBoost. This observation aligns with the results shown in Fig. 15, where LSTM demonstrated strong predictive capability in relatively homogeneous lithological intervals and even outperformed CatBoost in specific depth intervals. This lithology-based performance breakdown not only strengthens the conclusion that geological priors improve prediction

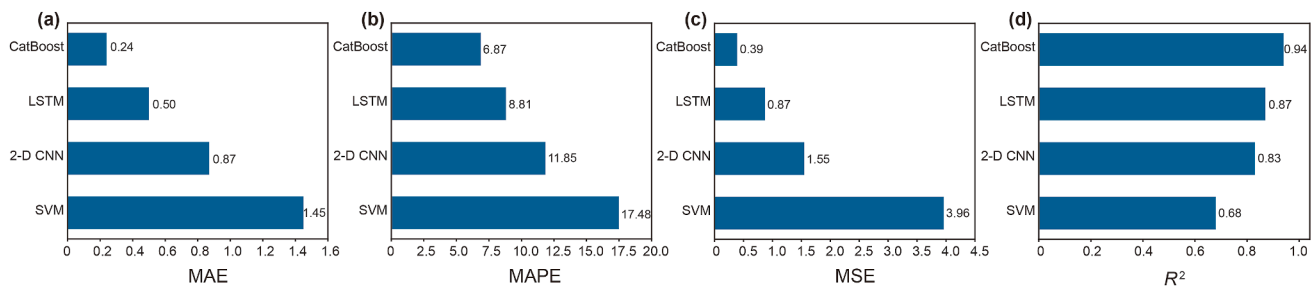


Fig. 14. Comparison of prediction metrics among evaluated models of the Jurassic Sangonghe Formation in the Taibei Sag, Turpan-Hami Basin.

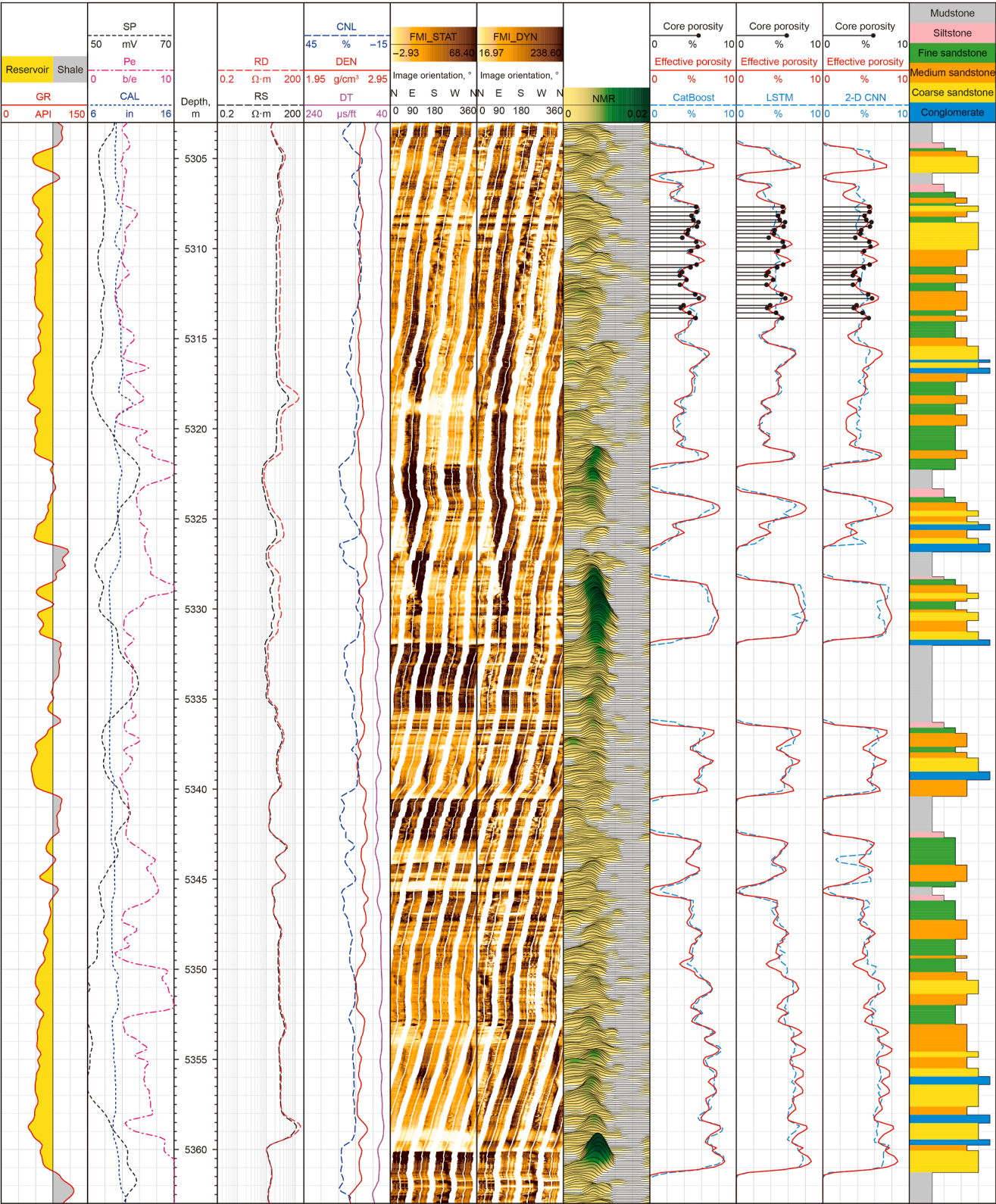


Fig. 15. Comparative well log interpretation and porosity predictions in the blind well of the Jurassic Sangonghe Formation. Tracks 1-5 illustrate logging curves including SP, GR, CAL, Pe, RD, RS, DT, CNL, DEN. Tracks 9 to 11 present the porosity predictions from the CatBoost, LSTM, and 2-D CNN models compared with the NMR-derived effective porosity, with core porosity included for reference. CatBoost exhibits the closest agreement with effective porosity and core data.

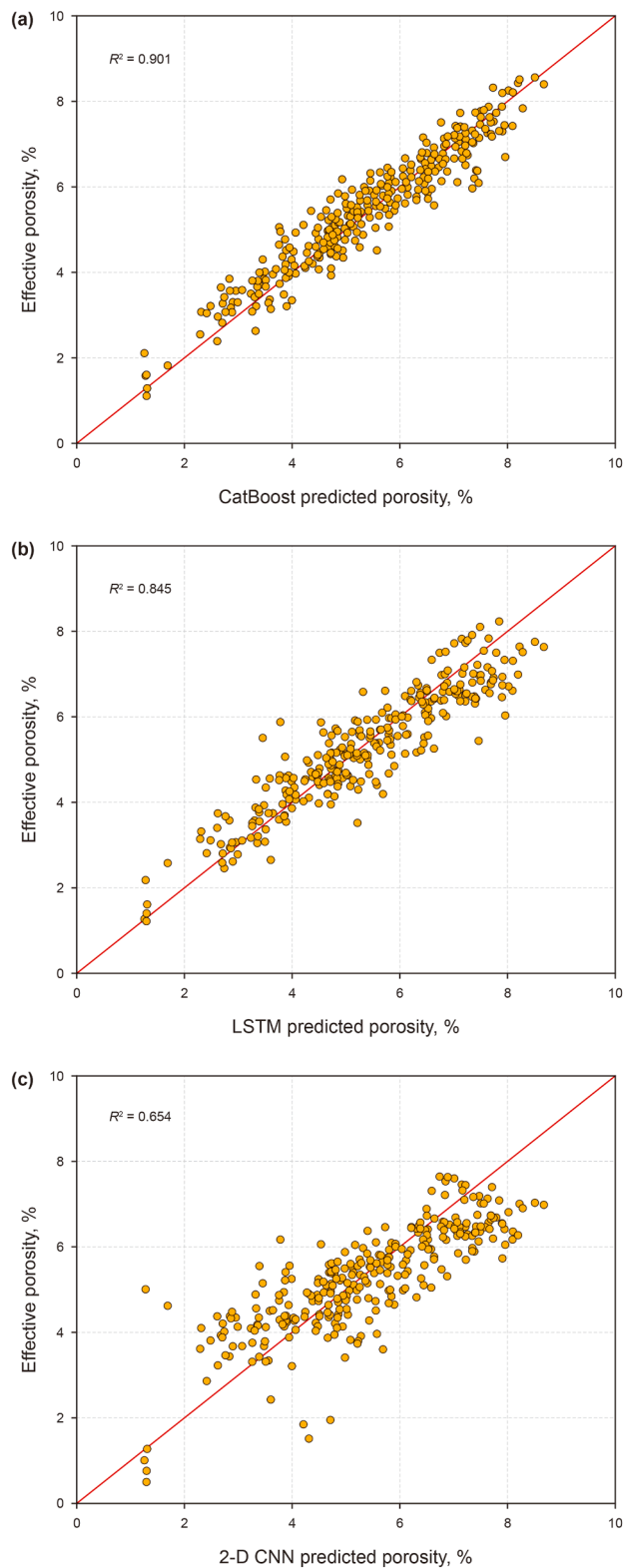


Fig. 16. Cross-plots of predicted porosity versus effective porosity in the blind well of the Jurassic Sangonghe Formation in the Taibei Sag, Turpan-Hami Basin. (a) CatBoost; (b) LSTM; (c) 2-D CNN.

generalization but also highlights the necessity of integrating domain knowledge into machine learning workflows for reservoir characterization in complex stratigraphic settings.

Table 5
Quantitative comparison of prediction errors among three models for the blind well of the Jurassic Sangonghe Formation in the Taibei Sag, Turpan-Hami Basin.

Models	MAE, %	MAPE, %	MSE, %	R^2
2-D CNN	1.674	19.35	4.560	0.654
LSTM	0.784	9.841	1.415	0.845
CatBoost	0.354	7.993	0.582	0.901

5.4. Adaptability of CatBoost in complex heterogeneous lithologies

As previously discussed, the deep tight sandstone reservoirs of the Jurassic Sangonghe Formation exhibit strong vertical heterogeneity in lithologic composition. Simultaneously, the logging responses in these intervals are often subtle or ambiguous, which poses substantial challenges for predictive modeling. For instance, even advanced deep learning architectures such as 2-D CNN and LSTM have struggled to deliver satisfactory performance under such conditions. However, the lithology-specific porosity distribution ranges provided a geologically meaningful constraint that allowed the CatBoost model, a traditional machine learning approach, to utilize categorical lithology and emerge as the best performer. This highlights the critical importance of integrating geological priors, especially under conditions of strong reservoir heterogeneity and weak logging responses.

Regarding the employed models, a noteworthy aspect lies in the intrinsic algorithmic advantage of the CatBoost model in handling discrete categorical input features. Specifically, CatBoost possesses a native capability to process discrete variables such as lithology by leveraging ordered target statistics combined with Bayesian priors (Hancock and Khoshgoftaar, 2020). This encoding approach effectively mitigates target leakage, stabilizes rare category estimation, and allows for the efficient capture of interactions between lithological classes and continuous well logs within its symmetric (oblivious) tree structure (Tanha et al., 2020; Bentéjac et al., 2021; Mienye and Sun, 2022). As a result, CatBoost naturally incorporates discrete lithologic features into its tree-based learning framework. Even in intervals with rapid lithologic transitions or ambiguous log responses, the model maintains geologically consistent and highly accurate porosity predictions. This inherent strength explains its superior predictive performance in blind well tests, particularly in zones characterized by frequent lithological changes. By contrast, deep learning networks such as LSTM and 2-D CNN are effective at learning complex nonlinear relations from continuous inputs but lack explicit mechanisms to represent the structured semantics of discrete features unless provided with substantial and well-balanced data (Yang et al., 2025). The limited number of training samples in this study may also constrain the predictive capacity of deep learning models, as these architectures generally require large-scale data to achieve stable generalization. Consequently, under the relatively small dataset conditions in this study, even with lower model complexity, CatBoost can outperform these more sophisticated architectures. This also indicates that embedding geological constraints such as lithology into the modeling workflow can partially reduce the model's excessive dependence on training sample sizes, thereby enhancing the applicability of such strategies in data-scarce regions with sparse well control. The result underscores that model effectiveness depends not only on algorithmic depth or network size but also on judicious feature engineering and the choice of algorithms that are well matched to the data.

Overall, treating lithology as an input feature effectively embeds lithology-dependent porosity ranges as a petrophysics constraint that guides the learning process. This can be viewed as an implementation of a petrophysics-informed learning strategy,

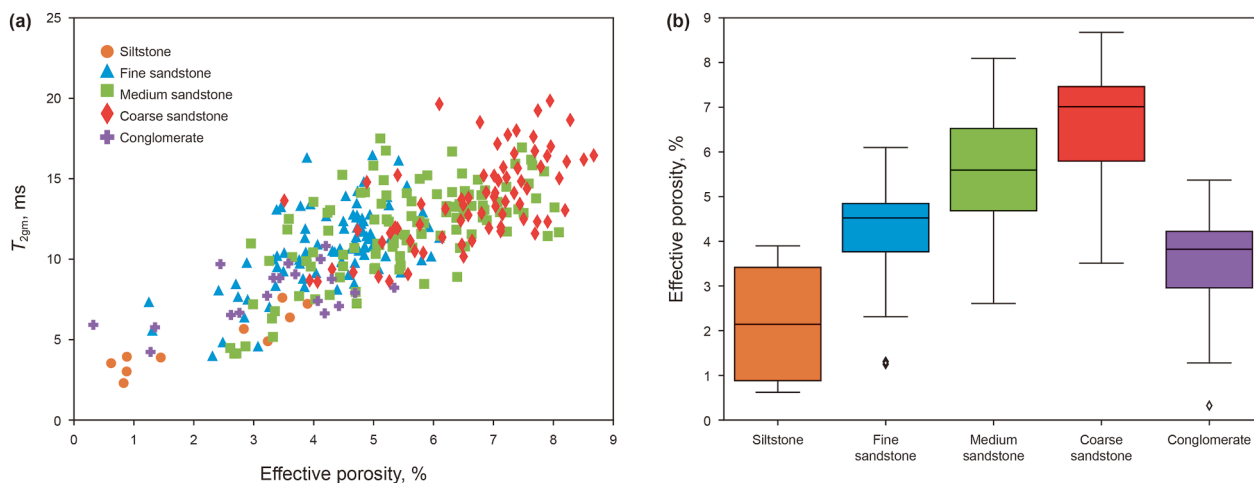


Fig. 17. Lithology-controlled porosity characteristics of the Jurassic Sangonghe Formation in the Taibei Sag, Turpan-Hami Basin. (a) Cross-plot of the relationship between NMR T_{2gm} and effective porosity across various lithologies; (b) boxplots of effective porosity distributions for each lithology.

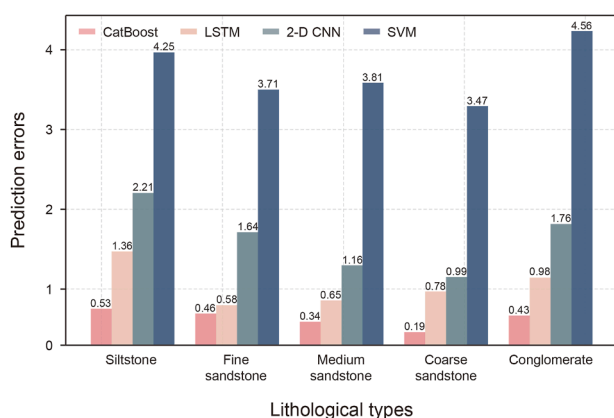


Fig. 18. Comparison of average prediction errors (MSE) of four models across different lithology types in the Jurassic Sangonghe Formation of the Taibei Sag, Turpan-Hami Basin.

analogous to an alternative form of the PINN framework. Simultaneously, it exemplifies how refined, geologically meaningful feature design can enhance model performance, emphasizing the principle that “high-quality data is the foundation of high-performance predictive models”.

5.5. Model explanation and visualization analysis

In the preceding sections, we thoroughly examined why the CatBoost model achieved superior performance in predicting effective porosity within the deep tight sandstone reservoirs of the Jurassic Sangonghe Formation. This assessment included discussions from the perspectives of data preprocessing strategies, algorithmic characteristics, and lithologic heterogeneity sensitivity. While these analyses provided a solid rationale for CatBoost’s strong predictive capabilities, the internal decision-making logic of the model remains largely opaque, underscoring the “black-box” nature. To gain insights into how different input features influence the model’s predictions, this section introduces the SHAP method to interpret the CatBoost model. By applying SHAP to the porosity prediction task, we aim to visualize and understand the internal decision-making process of the CatBoost model,

particularly how various logging curves affect porosity predictions at both the global and local levels.

5.5.1. SHAP values of different well logs

The contribution of different logging curves to the porosity prediction varies distinctly. For each sample, the model’s predicted value emerges as a cumulative outcome of the SHAP values attributed to all variables in that sample. Fig. 19 presents a SHAP summary plot for the nine input logging data used in CatBoost. Points to the right indicate positive contributions to the porosity prediction, points to the left indicate negative contributions, and the color encodes the logging data magnitude from low (blue) to high (red). A consistent pattern emerges for lithology-sensitive logging curves such as GR and DEN (Fig. 19). Samples with higher GR or higher DEN are associated with negative SHAP values, whereas lower GR or lower DEN tend to yield positive SHAP values. This monotonic directionality is geologically plausible. Increasing GR generally indicates finer grain size and elevated clay content, which is typical of a lithology shift from coarse-grained or medium-grained sandstones toward siltstone, and is usually accompanied by reduced pore-throat connectivity and lower effective porosity. Negative SHAP contributions in the high GR regime therefore decrease the model’s porosity prediction. Likewise, higher DEN reflects stronger compaction and cementation, often corresponding to poorly developed or absent secondary porosity, resulting in predominantly negative SHAP values.

The SHAP values for each data point across the nine logging curves are depicted in Fig. 20. Notably, when the GR values range between 40 and 70 API, the SHAP values are relatively high, diminishing rapidly as GR exceeds 70 API. This pattern correlates with the GR characteristics of different lithologies within the study area, where GR values of less than 70 API typically correspond to medium-grained and coarse-grained sandstones, which possess the highest porosity. The CatBoost model astutely captures this relationship, assigning higher SHAP values to the GR variable when indicative of higher porosity. Conversely, as GR values increase beyond this range, the model assigns lower SHAP values, effectively reducing the predicted porosity value. Moreover, it is observed that SHAP values associated with the CAL exhibit little variation when CAL ranges from 8.4 in to 8.8 in but rise sharply once the CAL exceeds 8.8 in (Fig. 20). This behavior reflects the sensitivity of NMR logging to borehole enlargement. Despite the removal of most enlarged-borehole intervals during

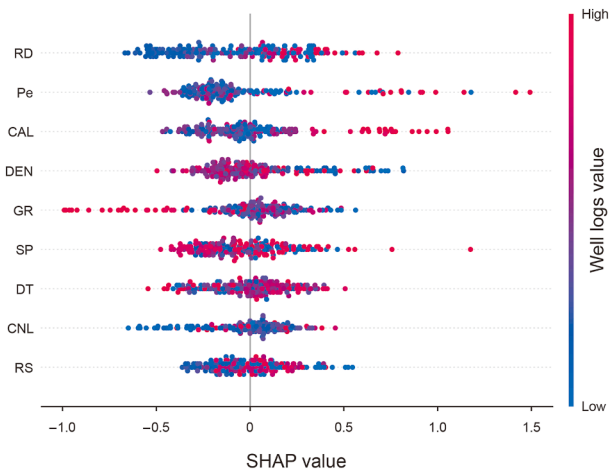


Fig. 19. SHAP value analysis of the nine logging curves in the CatBoost model for the Jurassic Sangonghe Formation in the Taibei Sag, Turpan-Hami Basin.

preprocessing, residual borehole enlargement data can still exert a significant influence on porosity prediction. These insights not only elucidate the predictive behavior of the CatBoost model, but also contribute to a deeper understanding of how specific logging measurements affect reservoir parameter estimation when applying machine learning in complex reservoirs.

5.5.2. SHAP values of different variables

Quantitative representation of the contributions of input feature variables to predicted porosity through SHAP values effectively elucidates and interprets the model's predictive results. Figs. 19 and 20 illustrate the SHAP values for 190 data points derived from nine logging curves in the test dataset. Next, we focus on analyzing the internal decision-making process of the models at a single sample point, which comprises nine input features (i.e., nine logging data at a specific depth). For a comprehensive comparison, we present the prediction outcomes of both the CatBoost and SVM for the same sample point. This facilitates the investigation of the internal decision-making processes of different

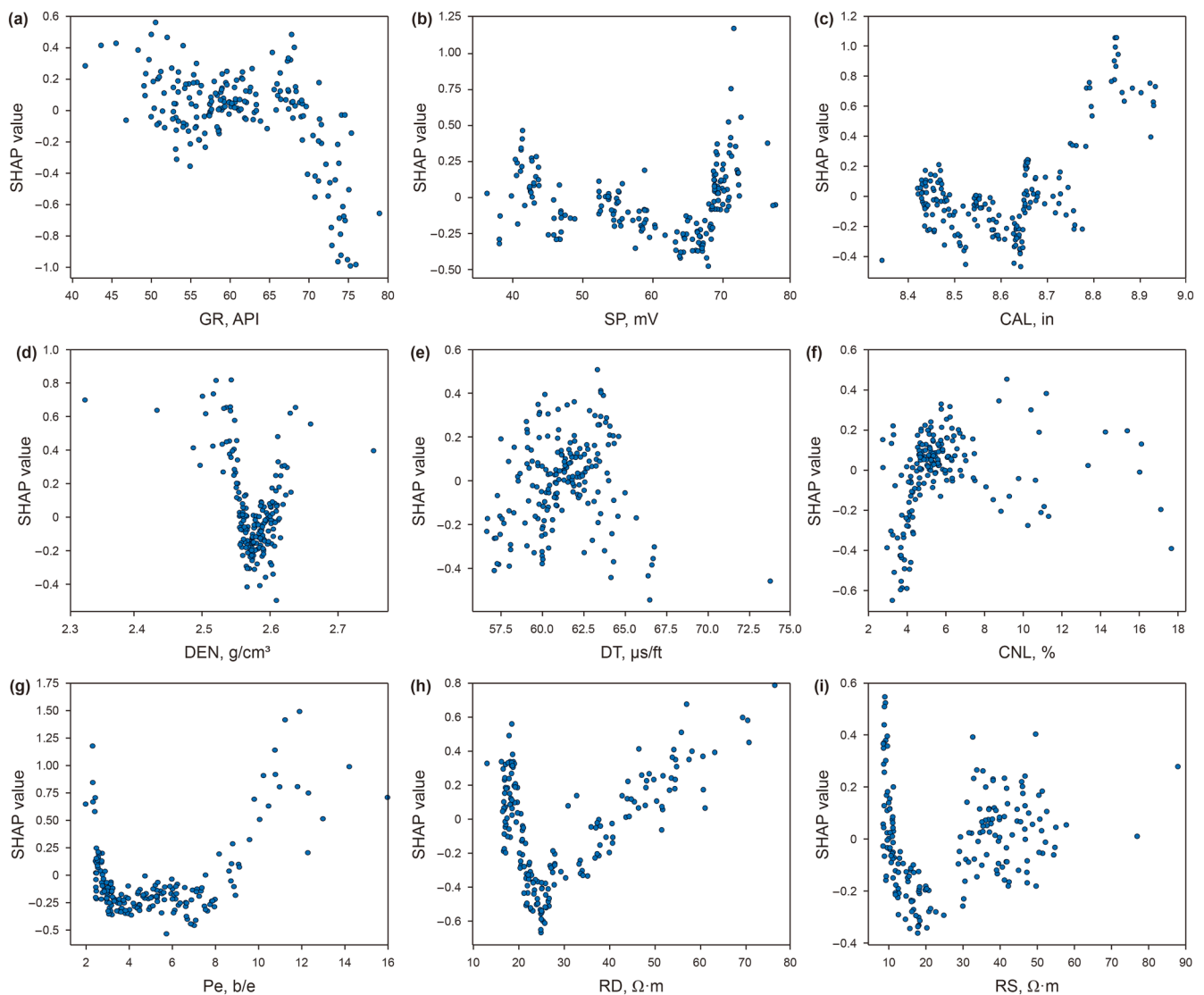


Fig. 20. Distribution of SHAP values for each data point of the nine logging curves in the CatBoost model for the Jurassic Sangonghe Formation in the Taibei Sag, Turpan-Hami Basin. (a)–(i) represent SHAP value distributions for GR, SP, CAL, DEN, DT, CNL, Pe, RD, and RS, respectively.

models when handling identical input data, thereby revealing the underlying causes for their differing predictive accuracies.

Fig. 21 showcases the prediction results for a specific sample point (depth of 5231.175 m, sample number 171 (Table 1) in the test set using both the SVM and CatBoost models. The force plots offer a visual explanation of the internal decision-making processes of the CatBoost and SVM models, respectively (Fig. 21(a) and (b)). Here, blue and red bars represent positive and negative SHAP values that reflect the physical relevance of each feature (e.g., GR, DEN, RD) highlights its robustness and interpretability in the context of heterogeneous deep tight sandstone reservoirs. This interpretability not only enhances trust in the model's predictions but also bridges the gap between data-driven modeling and geological reasoning.

performance by accurately analyzing the impact of each feature variable, progressively approximating the true value.

In summary, the CatBoost consistently provides more geologically reasonable and quantitatively accurate porosity predictions due to its superior handling of complex feature interactions and structured learning from both continuous and categorical variables. The model's ability to assign directionally consistent SHAP values that reflect the physical relevance of each feature (e.g., GR, DEN, RD) highlights its robustness and interpretability in the context of heterogeneous deep tight sandstone reservoirs. This interpretability not only enhances trust in the model's predictions but also bridges the gap between data-driven modeling and geological reasoning.

6. Conclusions

This study systematically evaluated and compared the performance of four predictive models, namely SVM, CatBoost, 2-D CNN, and LSTM, for effective porosity estimation in deep tight sandstone reservoirs of the Jurassic Sangonghe Formation. Utilizing nine conventional logs and lithology as input features, we benchmarked model accuracy using both standard test sets and blind-well validation, supported by NMR-derived porosity as the label.

The results clearly demonstrate that CatBoost outperforms both classical machine learning and advanced deep learning models in terms of accuracy, robustness, and generalization capability.

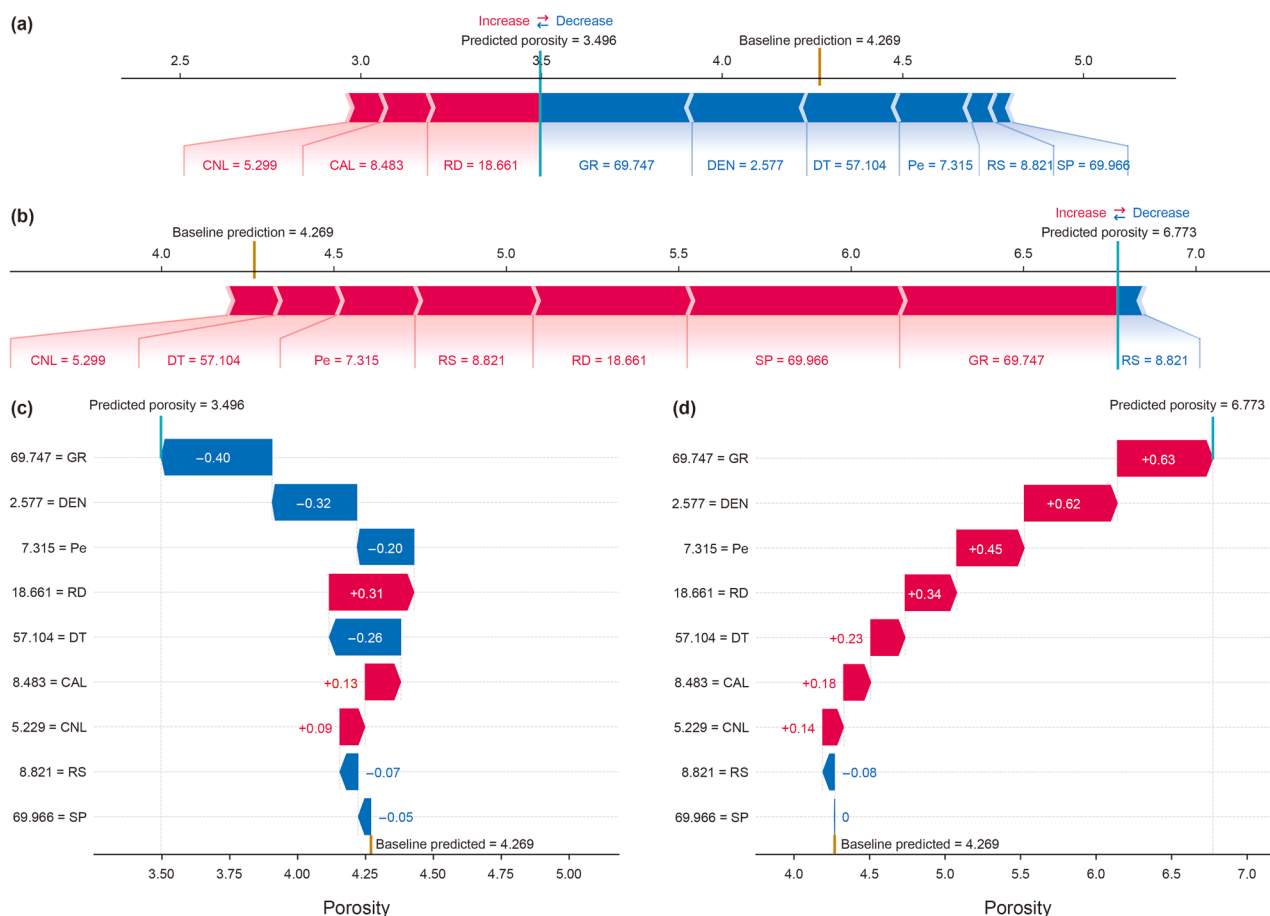


Fig. 21. Comparative analysis of prediction results by SVM and CatBoost models for identical sample points in the test set of the Jurassic Sangonghe Formation in the Taibei Sag, Turpan-Hami Basin. (a) Force diagram visualization of the decision-making process in the CatBoost model; (b) force diagram visualization of the decision-making process in the SVM model; (c) waterfall diagram for the CatBoost model; (d) waterfall diagram for the SVM model.

CatBoost not only captured global porosity trends but also maintained predictive fidelity in intervals characterized by abrupt lithologic transitions. In blind-well scenarios, CatBoost consistently produced the closest agreement with core porosity and NMR measurements, confirming its practical applicability for field deployment. One of the central findings of this work is the critical role of lithologic information in enhancing model performance. Unlike deep learning architectures, which require extensive data and struggle with discrete features, CatBoost natively integrates categorical lithology inputs using an ordered target statistics approach with Bayesian regularization. This capability enables the model to effectively learn lithology-dependent porosity behaviors, even in data-sparse environments with strong heterogeneity.

SHAP-based interpretability analysis revealed that the CatBoost model assigns physically meaningful contributions to features, including gamma ray, density, and lithology. This interpretability not only enhances user trust in model outputs but also bridges the gap between black-box AI and geological understanding. For instance, GR or density values were consistently associated with negative contributions to porosity prediction, aligning with petrophysical expectations for fine-grained or cemented lithologies.

Collectively, these results emphasize that algorithm–data compatibility, geologically informed feature engineering, and model interpretability are more impactful for predictive success than algorithmic complexity alone. Treating lithology as a soft petrophysical constraint improves the physical consistency and stability of AI models and represents a promising direction for hybrid reservoir characterization strategies. In summary, the proposed workflow offers a robust and interpretable framework for porosity prediction in complex, unconventional reservoirs. The insights gained here provide practical guidance for selecting appropriate algorithms and designing effective feature sets, with broader implications for the integration of geological priors into data-driven reservoir evaluation.

CRedit authorship contribution statement

Hong-Bin Li: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Yin-Hong Tian:** Writing – review & editing, Validation. **Gui-Wen Wang:** Writing – review & editing, Supervision, Funding acquisition. **Zhi-Bin He:** Writing – review & editing, Validation. **Lin-Bo Shao:** Writing – review & editing, Validation. **Jin Lai:** Writing – review & editing.

Funding

This work was jointly supported by the CNPC Key S&T Special Project (Grant No. 2022DJ2107), and the National Science and Technology Major Project of China (Grant No. 2025ZD1401104).

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.

References

Aas, K., Jullum, M., Løland, A., 2021. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artif. Intell.* 298, 103502. <https://doi.org/10.1016/j.artint.2021.103502>.
Ahmadi, M.A., 2015. Connectionist approach estimates gas-oil relative permeability in petroleum reservoirs: Application to reservoir simulation. *Fuel* 140, 429–439. <https://doi.org/10.1016/j.fuel.2014.09.058>.

Ali, M., Zhu, P.M., Jiang, R., et al., 2024. Data-driven lithofacies prediction in complex tight sandstone reservoirs: A supervised workflow integrating clustering and classification models. *Geomech. Geophys. Geo Energy Geo Resour.* 10 (1), 70. <https://doi.org/10.1007/s40948-024-00787-5>.
Antwarg, L., Miller, R.M., Shapira, B., et al., 2021. Explaining anomalies detected by autoencoders using Shapley Additive Explanations. *Expert Syst. Appl.* 186, 115736. <https://doi.org/10.1016/j.eswa.2021.115736>.
Ao, Y., Li, H., Zhu, L., et al., 2019. Identifying channel sand-body from multiple seismic attributes with an improved random forest algorithm. *J. Pet. Sci. Eng.* 173, 781–792. <https://doi.org/10.1016/j.petrol.2018.10.048>.
Bagheri, M., Rezaei, H., 2019. Reservoir rock permeability prediction using SVR based on radial basis function kernel. *Carbonates Evaporites* 34 (3), 699–707. <https://doi.org/10.1007/s13146-019-00493-4>.
Bagheri, M., Riahi, M.A., 2015. Seismic facies analysis from well logs based on supervised classification scheme with different machine learning techniques. *Arab. J. Geosci.* 8 (9), 7153–7161. <https://doi.org/10.1007/s12517-014-1691-5>.
Belle, V., Papantonis, I., 2021. Principles and practice of explainable machine learning. *Front. Big Data* 4, 688969. <https://doi.org/10.3389/fdata.2021.688969>.
Bentéjac, C., Csörgő, A., Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* 54 (3), 1937–1967. <https://doi.org/10.1007/s10462-020-09896-5>.
Bera, A., Shah, S., 2021. A review on modern imaging techniques for characterization of nanoporous unconventional reservoirs: Challenges and prospects. *Mar. Petrol. Geol.* 133, 105287. <https://doi.org/10.1016/j.marpetgeo.2021.105287>.
Bergen, K.J., Johnson, P.A., de Hoop, M.V., et al., 2019. Machine learning for data-driven discovery in solid Earth geoscience. *Science* 363 (6433). <https://doi.org/10.1126/science.aau0323>.
Bione, F.R.A., Venancio, I.M., Santos, T.P., et al., 2024. Estimating total organic carbon of potential source rocks in the Espírito Santo Basin, SE Brazil, using XGBoost. *Mar. Petrol. Geol.* 162, 106765. <https://doi.org/10.1016/j.marpetgeo.2024.106765>.
Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., et al., 2020. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing* 408, 189–215. <https://doi.org/10.1016/j.neucom.2019.10.118>.
Chen, X., Jia, X.L., Wang, B., et al., 2024. Main controlling factors for the enrichment of oil and gas in deep tight sandstone in the Qudong sub-sag of the Turpan-Hami Basin. *China Offshore Oil Gas* 36 (6), 26–38. <https://doi.org/10.11935/j.issn.1673-1506.2024.06.003> (in Chinese).
Chen, X.Y., Weng, C.K., Tao, L., et al., 2025. A novel method for predicting formation pore pressure ahead of the drill bit by embedding petrophysical theory into machine learning based on seismic and logging-while-drilling data. *Pet. Sci.* 22 (7), 2868–2883. <https://doi.org/10.1016/j.petsci.2025.04.012>.
Chen, Y., Zhang, D., 2020. Physics-constrained deep learning of geomechanical logs. *IEEE Trans. Geosci. Remote Sens.* 58, 5932–5943. <https://doi.org/10.1109/TGRS.2020.2973171>.
Chithra Chakra, N., Song, K.Y., Gupta, M.M., et al., 2013. An innovative neural forecast of cumulative oil production from a petroleum reservoir employing higher-order neural networks (HONNs). *J. Petrol. Sci. Eng.* 106, 18–33. <https://doi.org/10.1016/j.petrol.2013.03.004>.
Cortes, C., Vapnik, V., 1995. Support-vector networks. *Mach. Learn.* 20 (3), 273–297. <https://doi.org/10.1007/BF00994018>.
da Rocha, H.O., da Costa, J.L.S., Carrasquilla, A.A.G., et al., 2019. Petrophysical characterization using well log resistivity and rock grain specific surface area in a fractured carbonate pre-salt reservoir in the Santos Basin, Brazil. *J. Petrol. Sci. Eng.* 183, 106372. <https://doi.org/10.1016/j.petrol.2019.106372>.
Dai, J.X., Dong, D.Z., Ni, Y.Y., et al., 2024. Distribution patterns of tight sandstone gas and shale gas. *Petrol. Explor. Dev.* 51 (4), 767–779. [https://doi.org/10.1016/S1876-3804\(24\)60505-7](https://doi.org/10.1016/S1876-3804(24)60505-7).
Deng, W., Yao, R., Zhao, H.M., et al., 2019. A novel intelligent diagnosis method using optimal LS-SVM with improved PSO algorithm. *Soft Comput.* 23 (7), 2445–2462. <https://doi.org/10.1007/s00500-017-2940-9>.
Desbois, G., Urai, J.L., Kukla, P.A., et al., 2011. High-resolution 3D fabric and porosity model in a tight gas sandstone reservoir: A new approach to investigate microstructures from mm- to nm-scale combining argon beam cross-sectioning and SEM imaging. *J. Petrol. Sci. Eng.* 78 (2), 243–257. <https://doi.org/10.1016/j.petrol.2011.06.004>.
Dev, V.A., Eden, M.R., 2019. Formation lithology classification using scalable gradient boosted decision trees. *Comput. Chem. Eng.* 128, 392–404. <https://doi.org/10.1016/j.compchemeng.2019.06.001>.
Dong, S.Q., Wang, L.T., Zeng, L.B., et al., 2023. Fracture identification in reservoirs using well log data by window sliding recurrent neural network. *Geoenergy Sci. Eng.* 230, 212165. <https://doi.org/10.1016/j.geoen.2023.212165>.
Dong, S.Q., Zeng, L.B., Lyu, W.Y., et al., 2020. Fracture identification by semi-supervised learning using conventional logs in tight sandstones of Ordos Basin, China. *J. Nat. Gas Sci. Eng.* 76, 103131. <https://doi.org/10.1016/j.jngse.2019.103131>.
Ekanayake, I.U., Meddage, D.P.P., Rathnayake, U., 2022. A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using shapley additive explanations (SHAP). *Case Stud. Constr. Mater.* 16, e01059. <https://doi.org/10.1016/j.cscm.2022.e01059>.
Feng, R., Luthi, S.M., Gislöf, D., et al., 2018. Reservoir lithology classification based on seismic inversion results by hidden markov models: Applying prior

- geological information. *Mar. Petrol. Geol.* 93, 218–229. <https://doi.org/10.1016/j.marpetgeo.2018.03.004>.
- Gong, D., Cao, Z., Ni, Y., et al., 2016. Origins of Jurassic oil reserves in the Turpan-Hami Basin, northwest China: Evidence of admixture from source and thermal maturity. *J. Petrol. Sci. Eng.* 146, 788–802. <https://doi.org/10.1016/j.petrol.2016.07.025>.
- Guo, J., Xie, R., Wang, Y., et al., 2023. Variational mode decomposition for NMR echo data denoising. *IEEE Trans. Geosci. Remote Sens.* 61, 5902014. <https://doi.org/10.1109/TGRS.2023.3237925>.
- Gupta, L.K., Koundal, D., Mongia, S., 2023. Explainable methods for image-based deep learning: A review. *Arch. Comput. Meth. Eng.* 30 (4), 2651–2666. <https://doi.org/10.1007/s11831-023-09881-5>.
- Hancock, J.T., Khoshgoftaar, T.M., 2020. CatBoost for big data: An interdisciplinary review. *J. Big Data* 7 (1), 94. <https://doi.org/10.1186/s40537-020-00369-8>.
- Handhal, A.M., Al-Abadi, A.M., Chafeet, H.E., et al., 2020. Prediction of total organic carbon at Rumaila oil field, southern Iraq using conventional well logs and machine learning algorithms. *Mar. Petrol. Geol.* 116, 104347. <https://doi.org/10.1016/j.marpetgeo.2020.104347>.
- Haritha, D., Satyavani, N., Ramesh, A., 2025. Generation of missing well log data with deep learning: CNN-Bi-LSTM approach. *J. Appl. Geophys.* 233, 105628. <https://doi.org/10.1016/j.jappgeo.2025.105628>.
- Hassija, V., Chamola, V., Mahapatra, A., et al., 2024. Interpreting black-box models: A review on explainable artificial intelligence. *Cogn. Comput.* 16 (1), 45–74. <https://doi.org/10.1007/s12559-023-10179-8>.
- Hu, X., Meng, Q., Guo, F., et al., 2024. Deep learning algorithm-enabled sediment characterization techniques to determination of water saturation for tight gas carbonate reservoirs in Bohai Bay Basin, China. *Sci. Rep.* 14, 12179. <https://doi.org/10.1038/s41598-024-63168-8>.
- Huang, P., Liu, M.-J., Cao, B., et al., 2026. Diagenesis and implication for the reservoir quality of the tight sandstone reservoir in the depression zone of foreland basin, Upper Triassic Xujiahe Formation, western Sichuan foreland basin, China: Evidence from petrography and geochemistry. *Pet. Sci.* 23, 1650–1674. <https://doi.org/10.1016/j.petsci.2025.12.016>.
- Huang, C.L., Wang, C.J., 2006. A GA-based feature selection and parameters optimization for support vector machines. *Expert Syst. Appl.* 31 (2), 231–240. <https://doi.org/10.1016/j.eswa.2005.09.024>.
- Huang, W., Wang, Y., Wang, P., et al., 2025. Dual-driven gas reservoir productivity prediction with small sample data: Integrating physical constraints and GAN-based data augmentation. *Geoenergy Sci. Eng.* 247, 213711. <https://doi.org/10.1016/j.geoen.2025.213711>.
- Jiang, Z.X., Li, Z., Li, F., et al., 2015. Tight sandstone gas accumulation mechanism and development models. *Pet. Sci.* 12 (4), 587–605. <https://doi.org/10.1007/s12182-015-0061-6>.
- Jin, G., Xie, R., Liu, M., et al., 2019. Petrophysical parameter calculation based on NMR echo data in tight sandstone. *IEEE Trans. Geosci. Remote Sens.* 57 (8), 5618–5625. <https://doi.org/10.1109/TGRS.2019.2901119>.
- Kocoglu, Y., Gorell, S.B., Emadi, H., et al., 2024. Improving the accuracy of short-term multiphase production forecasts in unconventional tight oil reservoirs using contextual bi-directional long short-term memory. *Geoenergy Sci. Eng.* 235, 212688. <https://doi.org/10.1016/j.geoen.2024.212688>.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60 (6), 84–90. <https://doi.org/10.1145/3065386>.
- Lai, J., Li, D., Bai, T., et al., 2023. Reservoir quality evaluation and prediction in ultra-deep tight sandstones in the kuqa Depression, China. *J. Struct. Geol.* 170, 104850. <https://doi.org/10.1016/j.jsg.2023.104850>.
- Lai, J., Su, Y., Xiao, L., et al., 2024. Application of geophysical well logs in solving geologic issues: Past, present and future prospect. *Geosci. Front.* 15 (3), 101779. <https://doi.org/10.1016/j.gsf.2024.101779>.
- Lai, J., Wang, G., Fan, Q., et al., 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., Xiao, L., Song, X., et al., 2026. Mechanism of reservoir property variations in deep and ultra-deep sandstones in the kuqa depression, china. *J. Asian Earth Sci.* 309, 107150. <https://doi.org/10.1016/j.jseas.2026.107150>.
- Lan, W., Tao, Y., Liang, B., et al., 2025. Shear wave velocity prediction based on Bayesian-optimized multi-head attention mechanism and CNN-BiLSTM. *Comput. Geosci.* 195, 105787. <https://doi.org/10.1016/j.cageo.2024.105787>.
- Latrach, A., Malki, M.L., Morales, M., et al., 2024. A critical review of physics-informed machine learning applications in subsurface energy systems. *Geoenergy Sci. Eng.* 239, 212938. <https://doi.org/10.1016/j.geoen.2024.212938>.
- LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature* 521 (7553), 436–444. <https://doi.org/10.1038/nature14539>.
- Li, D., Wang, G.W., Bie, K., et al., 2025. Formation mechanism and reservoir quality evaluation in tight sandstones under a compressional tectonic setting: The Jurassic Ahe Formation in Kuqa Depression, Tarim Basin, China. *Petrol. Sci.* 22 (3), 998–1020. <https://doi.org/10.1016/j.petsci.2024.12.026>.
- Li, T., Huang, Z., Feng, Y., et al., 2020. Reservoir characteristics and evaluation of fluid mobility in organic-rich mixed siliciclastic-carbonate sediments: A case study of the lacustrine Qiketai Formation in Shengbei Sag, Turpan-Hami Basin, Northwest China. *J. Petrol. Sci. Eng.* 185, 106667. <https://doi.org/10.1016/j.petrol.2019.106667>.
- Li, Z., 2022. Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Comput. Environ. Urban Syst.* 96, 101845. <https://doi.org/10.1016/j.compenurbysys.2022.101845>.
- Liao, G.Z., Chen, W.L., Zong, F.R., et al., 2021. NMR fluid analyzer applying to petroleum industry. *Pet. Sci.* 18 (1), 54–91. <https://doi.org/10.1007/s12182-020-00529-8>.
- Liu, G., Guo, J., 2019. Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing* 337, 325–338. <https://doi.org/10.1016/j.neucom.2019.01.078>.
- Liu, P., Wang, X., Liu, C., et al., 2023. Kinetic isotope effect and reaction pathways of thermogenic propane generation at early maturation stage: Insights from position-specific carbon isotope distribution. *Mar. Petrol. Geol.* 153, 106252. <https://doi.org/10.1016/j.marpetgeo.2023.106252>.
- Lu, C., Zhang, S., Xue, D., et al., 2022. Improved estimation of coalbed methane content using the revised estimate of depth and CatBoost algorithm: A case study from southern Sichuan Basin, China. *Comput. Geosci.* 158, 104973. <https://doi.org/10.1016/j.cageo.2021.104973>.
- Male, F., Jensen, J.L., Lake, L.W., 2020. Comparison of permeability predictions on cemented sandstones with physics-based and machine learning approaches. *J. Nat. Gas Sci. Eng.* 77, 103244. <https://doi.org/10.1016/j.jngse.2020.103244>.
- McGlade, C., Speirs, J., Sorrell, S., 2013. Unconventional gas – A review of regional and global resource estimates. *Energy* 55, 571–584. <https://doi.org/10.1016/j.energy.2013.01.048>.
- Mi, J.X., Li, A.D., Zhou, L.F., 2020. Review study of interpretation methods for future interpretable machine learning. *IEEE Access* 8, 191969–191985. <https://doi.org/10.1109/ACCESS.2020.3032756>.
- Mienye, I.D., Sun, Y., 2022. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access* 10, 99129–99149. <https://doi.org/10.1109/ACCESS.2022.3207287>.
- Moosavi, N., Bagheri, M., Nabi-Bidhendi, M., et al., 2023. Porosity prediction using Fuzzy SVR and FCM SVR from well logs of an oil field in south of Iran. *Acta Geophys.* 71 (2), 769–782. <https://doi.org/10.1007/s11600-022-00944-y>.
- Ni, Y., Liao, F., Gong, D., et al., 2019. Stable carbon and hydrogen isotopic characteristics of natural gas from Taibei sag, Turpan-Hami Basin, NW China. *Petrol. Explor. Dev.* 46 (3), 531–542. [https://doi.org/10.1016/S1876-3804\(19\)60033-9](https://doi.org/10.1016/S1876-3804(19)60033-9).
- Ordóñez, F., Roggen, D., 2016. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16 (1), 115. <https://doi.org/10.3390/s16010115>.
- Otchere, D.A., Arbi Ganat, T.O., Gholami, R., et al., 2021. Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models. *J. Petrol. Sci. Eng.* 200, 108182. <https://doi.org/10.1016/j.petrol.2020.108182>.
- Prokhorenkova, L., Gusev, G., Vorobev, A., et al., 2018. CatBoost: Unbiased boosting with categorical features. In: Bengio, S., Wallach, H., Larochelle, H. (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc.
- Raeesi, M., Moradzadeh, A., Doulati Ardejani, F., et al., 2012. Classification and identification of hydrocarbon reservoir lithofacies and their heterogeneity using seismic attributes, logs data and artificial neural networks. *J. Petrol. Sci. Eng.* 82–83, 151–165. <https://doi.org/10.1016/j.petrol.2012.01.012>.
- Sacchi, Q., Salina Borello, E., Weltje, G.J., et al., 2016. Increasing the predictive power of geostatistical reservoir models by integration of geological constraints from stratigraphic forward modeling. *Mar. Petrol. Geol.* 69, 112–126. <https://doi.org/10.1016/j.marpetgeo.2015.10.018>.
- Sagi, O., Rokach, L., 2018. Ensemble learning: A survey. *Wires Data Min. Knowl. Discov.* 8 (4), e1249. <https://doi.org/10.1002/widm.1249>.
- Sakhavi, S., Guan, C., Yan, S., 2018. Learning temporal information for brain-computer interface using convolutional neural networks. *IEEE Transact. Neural Networks Learn. Syst.* 29 (11), 5619–5629. <https://doi.org/10.1109/TNNLS.2018.2789927>.
- Saporetta, C.M., Fonseca, D.L., Oliveira, L.C., et al., 2022. Hybrid machine learning models for estimating total organic carbon from mineral constituents in core samples of shale gas fields. *Mar. Petrol. Geol.* 143, 105783. <https://doi.org/10.1016/j.marpetgeo.2022.105783>.
- Shi, J.X., Zhao, X.Y., Zeng, L.B., et al., 2024. Identification of reservoir types in deep carbonates based on mixed-kernel machine learning using geophysical logging data. *Pet. Sci.* 21 (3), 1632–1648. <https://doi.org/10.1016/j.petsci.2023.12.016>.
- Song, C., Lu, W., Wang, Y., et al., 2022. Reservoir prediction based on closed-loop CNN and virtual well-logging labels. *IEEE Trans. Geosci. Remote Sens.* 60, 5919912. <https://doi.org/10.1109/TGRS.2022.3205301>.
- Su, Y., Lai, J., Dang, W., et al., 2026. Geological factors and fracture distribution in deep and ultra-deep sandstones in Kuqa Depression, Tarim Basin, China. *Solid Earth* 17, 643–664. <https://doi.org/10.5194/se-17-643-2026>.
- Tamoto, H., dos Santos Gioria, R., de Carvalho Carneiro, C., 2023. Prediction of nuclear magnetic resonance porosity well-logs in a carbonate reservoir using supervised machine learning models. *J. Petrol. Sci. Eng.* 220, 111169. <https://doi.org/10.1016/j.petrol.2022.111169>.
- 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>.
- Tanha, J., Abdi, Y., Samadi, N., et al., 2020. Boosting methods for multi-class imbalanced data classification: An experimental review. *J. Big Data* 7 (1), 70. <https://doi.org/10.1186/s40537-020-00349-y>.
- Tariq, Z., Gudala, M., Yan, B., et al., 2023. A fast method to infer Nuclear Magnetic Resonance based effective porosity in carbonate rocks using machine learning techniques. *Geoenergy Sci. Eng.* 222, 211333. <https://doi.org/10.1016/j.geoen.2022.211333>.

- Tian, Y.H., Wang, G.W., Li, H.B., et al., 2024. A novel deep learning method based on 2-D CNNs and GRUs for permeability prediction of tight sandstone. *Geoenergy Sci. Eng.* 238, 212851. <https://doi.org/10.1016/j.geoen.2024.212851>.
- Wang, W., Li, T., Xiao, D., et al., 2023. Geological conditions and controls on accumulation of tight sandstone gas, deep part of the Shengbei sub-sag, Turpan-Hami Basin, NW China. *Mar. Petrol. Geol.* 158, 106513. <https://doi.org/10.1016/j.marpetgeo.2023.106513>.
- Wang, Y., Cui, M., Xie, B., et al., 2025. Tight sandstone reservoir classification based on 1DCNN-BLSTM with conventional logging data. *Acta Geophys.* 73 (3), 2373–2389. <https://doi.org/10.1007/s11600-024-01506-0>.
- Wang, Y.P., Zou, Y.R., Zhan, Z.W., et al., 2018. Origin of natural gas in the Turpan-Hami Basin, NW China: Evidence from pyrolytic simulation experiment. *Int. J. Coal Geol.* 195, 238–249. <https://doi.org/10.1016/j.coal.2018.06.007>.
- Wang, Z.Y., Lu, S.F., Zhou, N.W., et al., 2024. Complementary testing and machine learning techniques for the characterization and prediction of Middle Permian tight gas sandstone reservoir quality in the northeastern Ordos Basin, China. *Pet. Sci.* 21 (5), 2946–2968. <https://doi.org/10.1016/j.petsci.2024.08.008>.
- Wen, Z.Y., Shi, J.S., He, B.S., et al., 2019. Exploiting GPUs for efficient Gradient Boosting Decision Tree training. *IEEE Trans. Parallel Distr. Syst.* 30 (12), 2706–2717. <https://doi.org/10.1109/TPDS.2019.2920131>.
- Wu, J., Luo, R.Z., Luo, L., et al., 2025. LLMs-guided parameters prediction of tight sandstone reservoirs. *Pet. Sci.* 22 (12), 5005–5019. <https://doi.org/10.1016/j.petsci.2025.08.029>.
- Xiao, L., Li, J., Mao, Z., et al., 2018. A method to determine nuclear magnetic resonance (NMR) T_2 cutoff based on normal distribution simulation in tight sandstone reservoirs. *Fuel* 225, 472–482. <https://doi.org/10.1016/j.fuel.2018.03.178>.
- Yang, L., Wang, S., Chen, X., et al., 2023. High-fidelity permeability and porosity prediction using deep learning with the self-attention mechanism. *IEEE Transact. Neural Networks Learn. Syst.* 34 (7), 3429–3443. <https://doi.org/10.1109/TNNLS.2022.3157765>.
- Yang, W., Qin, Y., Wang, K., et al., 2025. A discrete convolutional network for entity relation extraction. *Neural Netw.* 184, 107117. <https://doi.org/10.1016/j.neunet.2024.107117>.
- Zhang, M.Z., Jia, A.L., Lei, Z.X., 2024. Inter-well reservoir parameter prediction based on LSTM-Attention network and sedimentary microfacies. *Geoenergy Sci. Eng.* 235, 212723. <https://doi.org/10.1016/j.geoen.2024.212723>.
- Zhang, X., Nie, F., Su, X., et al., 2020. Relationships between Meso-Cenozoic denudation in the Eastern Tian Shan and uranium mineralization in the Turpan-Hami Basin, NW China: Constraints from apatite fission track study. *Ore Geol. Rev.* 127, 103820. <https://doi.org/10.1016/j.oregeorev.2020.103820>.
- Zhang, Y., Yang, H., 2025. Direct inversion of 3D seismic reservoir parameters based on dual learning networks. *Pet. Sci.* 22 (10), 4037–4051. <https://doi.org/10.1016/j.petsci.2025.05.027>.
- Zhang, Z., Zhang, H., Li, J., et al., 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, F., Lai, J., Xia, Z.L., et al., 2025. Coaly source rock evaluation using well logs: The Jurassic Kezilenuer Formation in Kuqa Depression, Tarim Basin, China. *Pet. Sci.* 22 (9), 3599–3612. <https://doi.org/10.1016/j.petsci.2025.05.029>.
- Zhen, Y., He, D., Li, D., et al., 2024. Episodic evolution of intracontinental orogenic deformation: Insight from the southern Bogda fold-and-thrust belt of the Tianshan Mountains, NW China. *Tectonophysics* 876, 230266. <https://doi.org/10.1016/j.tecto.2024.230266>.
- Zhi, D., Li, J., Yang, F., et al., 2024. Whole petroleum system in Jurassic coal measures of Taibei Sag in Tuha Basin, NW China. *Petrol. Explor. Dev.* 51 (3), 519–534. [https://doi.org/10.1016/S1876-3804\(24\)60485-4](https://doi.org/10.1016/S1876-3804(24)60485-4).
- Zhou, W., Zhao, H., Li, X., et al., 2025. Missing well logs reconstruction based on cascaded bidirectional long short-term memory network. *Expert Syst. Appl.* 259, 125270. <https://doi.org/10.1016/j.eswa.2024.125270>.