Advances in
Geo-Energy Research

Vol. 16, No. 1, p. 21-35, 2025

Original article

Data-driven interpretable machine learning for prediction of porosity and permeability of tight sandstone reservoir

Liu Cao[1,2], Fujie Jiang[1,2]*, Zhangxing Chen[1,3,4], Yang Gao[1,2,5], Lina Huo[1,2], Di Chen[1,2]

[1] *National Key Laboratory of Petroleum Resources and Engineering, China University of Petroleum (Beijing), Beijing 102249, P. R. China*
[2] *College of Geosciences, China University of Petroleum (Beijing), Beijing 102249, P. R. China*
[3] *Institute of Digital Twins, Eastern Institute of Technology, Ningbo 315200, P. R. China*
[4] *Chemical and Petroleum Engineering, Schulich School of Engineering, University of Calgary, Calgary T2N 1N4, Canada*
[5] *Research Institute of Petroleum Exploration and Development, PetroChina, Beijing 100089, P. R. China*

Keywords:
Data-driven modeling
interpretable machine learning
permutation importance-set
reservoir characterization

Cited as:
Cao, L., Jiang, F., Chen, Z., Gao, Y., Huo, L., Chen, D. Data-driven interpretable machine learning for prediction of porosity and permeability of tight sandstone reservoir. Advances in Geo-Energy Research, 2025, 16(1): 21-35.
https://doi.org/10.46690/ager.2025.04.04

Abstract:
Porosity and permeability are crucial indicators in the identification of high-quality reservoirs and favorable "sweet spot" zones, as well as key parameters when predicting and evaluating the development potential of fossil fuels like oil and gas. However, it is impracticable to collect enough core samples on vertical and horizontal planes for analysis due to the associated time and cost demand. Machine learning algorithms have shown remarkable capabilities in predicting the petrophysical properties by capturing non-linear relationships among logging data. In this study, to quantify the selection of logging curves and reduce the redundant logging data input, a novel and interpretable Permutation Importance-Set algorithm is proposed on the basis of logging data from the Upper Triassic Xujiahe Formation in the Sichuan Basin. The results indicate that, because of compaction, burial depth is the primary feature affecting the physical properties of tight sandstone reservoirs. Acoustic and spontaneous potential logs are critical for porosity, while density and spontaneous potential logs are pivotal for permeability, reflecting the complex diagenesis caused by the widespread sand-mud interbedding. Basin-level prediction models for porosity and permeability were developed using ten machine learning algorithms, then ablation studies confirmed the effectiveness of our feature selection and the reduced model complexity and over-fitting. This study offers a concise, interpretable prediction model with superior accuracy and interpretability for tight sandstone reservoirs.

1. Introduction

To avoid costly drilling mistakes due to the geological uncertainties and varying resource potential, the primary risk assessment indicator for oil exploration must be taken as the hydrocarbon enrichment potential of prospective reservoirs. The petrophysical properties (porosity and permeability) of the reservoir are among the crucial evaluation indicators of hydrocarbon enrichment potential (Wang et al., 2020). Within the context of continental sedimentation in China, tight sandstone reservoirs refer to a porosity $< 10\%$ and permeability < 1 mD of oil and gas reservoirs (Zou et al., 2012). Compared to conventional oil and gas reservoirs, tight oil and gas reservoirs have less favorable physical properties, stronger heterogeneity, lower reserve density ratio, and they present challenges in predicting favorable "sweet spot" zones and effective reservoirs (Zhao and Chen, 2014; Sun et al., 2019; Ampomah et al., 2017). Therefore, it is essential to derive a reliable and convenient technology that can predict reservoir porosity and permeability across an entire basin. Such tech-

*Corresponding author.
E-mail address: cl@student.cup.edu.cn (L. Cao); jfjhtb@163.com (F. Jiang); zxchen@eitech.edu.cn (Z. Chen); gaoyang1112024@163.com (Y. Gao) hln10171007@163.com (L. Huo); chendcup@cup.edu.cn (D. Chen).

Yandy Scientific Press

2207-9963 © The Author(s) 2025.
Received December 9, 2024; revised January 5, 2025; accepted February 1, 2025; available online February 5, 2025.

nology could not only identify favorable "sweet spot" zones and effective reservoirs, guiding oil and gas exploration and development efficiently and in a cost effective way, but also would be crucial to ensure interpretability and maintain credibility among exploration and development personnel (Aras and Hanifi Van, 2022).

The accurate assessment of porosity and permeability typically relies on core tests, which are both costly and time-consuming (Zhao et al., 2022). In addition, it is difficult and impracticable to collect core samples across vertical and horizontal planes for analysis. Thus, obtaining sufficiently comprehensive and precise data on porosity and permeability remains a significant challenge (Alfi et al., 2019). In contrast, logging data, which is more readily available, offers a more economic option while also providing a greater abundance of data. In the actual exploration and development process, well logging data records the formation data at every 0.125 m, which can reflect the continuous physical properties of formations, including the electrical and acoustic properties (Lu et al., 2021; Jiang et al., 2023). The first progress in this field emerged from the study by Chork et al. (1994), who pioneered sonic travel time clustering to homogenize logging data and subsequently derived porosity-permeability conversion formulas through systematic core data integration. Later studies confronted the challenge of geological nonlinearity. Zhang et al. (2020) developed a linearized rock physics inversion method that predicts accurate physical parameters but is unsuitable for nonlinear models. A limitation was similarly observed in Belhouchet et al. (2021)'s multiple regression model, which demonstrated partial success in permeability prediction but inadequately addressed complex nonlinear interdependencies. Recent research shifted focus toward specialized relative permeability modeling. Shen et al. (2020) experimentally established empirical correlations between water relative permeability and hydrate saturation, while Yang et al. (2023b) numerically simulated multiphase flow dynamics. A theoretical breakthrough was made by Chai et al. (2024), who redefined relative permeability characterization through fractal geometry principles. Despite the widespread use of logging interpretation and empirical methods for porosity and permeability prediction, those in tight sandstone reservoirs remain challenging due to their strong nonlinearity caused by heterogeneity and their complex diagenesis (Al Khalifah et al., 2020).

With the great capabilities of data-driven artificial intelligence algorithms in discerning nonlinear mapping relationships, machine learning (ML) techniques have found extensive application in porosity and permeability prediction (Zhang et al., 2023). Yu et al. (2020) combined fractal theory with deep learning to accurately determine macroscopic permeability from microscopic sandstone images. Zhao et al. (2022) used a cross-correlation matrix to analyze the relationship between logging characteristics and permeability, and developed a high-accuracy eXtreme Gradient Boosting (XGBoost) model based on core test data from three wells. Jiang et al. (2023) applied the Grey Correlation Analysis (GCA) and Back Propagation Neural Network (BPNN) methods to predict porosity and permeability using 678 datasets. Lu et al. (2021) employed linear regression to identify correlations between logging curves and porosity, selecting log features for their machine learning model based on expert judgment. Otchere et al. (2022) used Pearson and Spearman correlation coefficients along with Random Forest (RF) algorithms to determine the best method for predicting porosity and permeability, ultimately favoring RF. Zhang et al. (2021) introduced univariate and bivariate predictive patterns (UPP and BPP) to visualize how variables influence model predictions, whereas these methods are limited in uncovering variable interactions. Each commonly used feature selection algorithm has its inherent advantages and limitations: Pearson, Principal Component Analysis (PCA), and Linear Regression (LR) excel at detecting linear relationships (Jolliffe, 1986; Seber and Wild, 1989; Hauke and Kossowski, 2011), Spearman is suited for monotonic relationships (Hauke and Kossowski, 2011), and tree-based algorithms like RF have inherent biases (Breiman, 2001). In addition, due to the limitations of small datasets, there remains a paucity of discussions on the application of ML models with industrial application value (typically $R^2 \geq 80\%$) across larger geographical spaces, such as at the basin scale, hindering the development of predictive models at this extensive level (Karpatne et al., 2019).

In this study, taking logging data from the Xujiahe Formation (Xu FM) in the Sichuan Basin as a basis, a novel and highly interpretable PI-Set algorithm is proposed, which quantifies the importance of specific logging features to greatly reduce the unnecessary and redundant logging data input, alleviating model complexity and overfitting risk. Additionally, the geological reasons for selecting logging curves by the PI-Set model are discussed. Comparative experiments with ten ML algorithms are conducted to develop the best-performing basin-level porosity and permeability prediction models. Five evaluation metrics are introduced to assess the prediction performance of the ML models, and their applicability to the prediction of porosity and permeability is examined. In addition, complete feature ablation studies are conducted to illustrate the influence of each feature on the final prediction results, validating the effectiveness of the PI-Set model.

2. Geological setting

The Sichuan Basin is located at the western margin of Yangtze Block, China, with an area of approximately 180,000 km^2 (Liu et al., 2020). It was estimated to contain about 40×10^{12} cubic meters of natural gas, making it the largest reserve in China. However, its proven reserve has turned out to be merely 18.8% (Liu et al., 2020; Yang et al., 2023a). According to the structural characteristics of this basin, it can be divided into four first-order tectonic units: foreland thrust zone, foreland depression zone, foreland lope zone, and foreland uplift zone (Fig. 1(a)). Therefore, it is a classical superimposed basin with sedimentary records encompassing the Paleozoic, Mesozoic and Cenozoic eras (Deng et al., 2022; Shi et al., 2022).

The Xu FM contributes nearly 80% of the entire reserves in the Sichuan Basin, which are mainly distributed in its central and northwest parts, accounting for 65% and 29% of the pro-

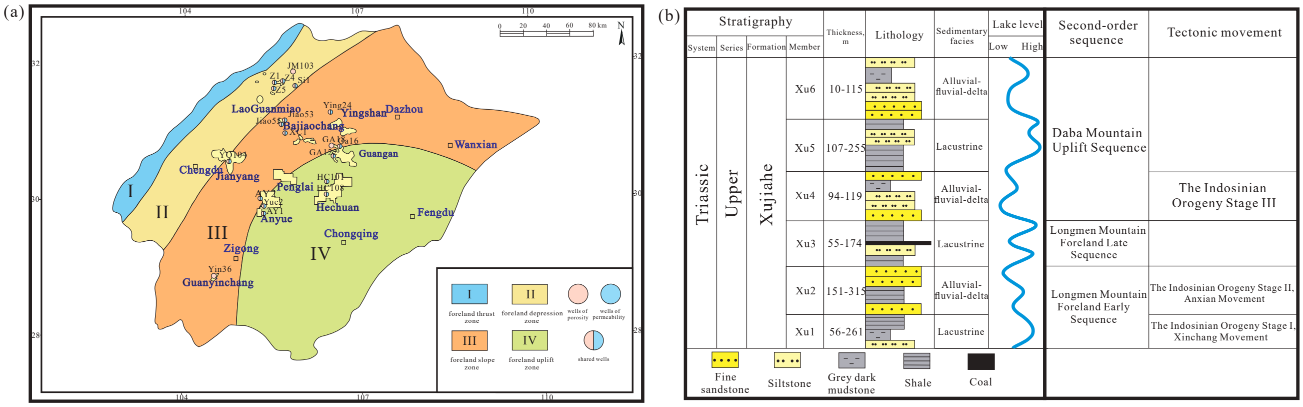


Fig. 1. Tectonic zone map and lithologic column diagram of the Sichuan Basin. (a) Tectonic zone map. (Note: the pink dots represent the well locations used by the porosity prediction model, the blue dots represent the well locations used by the permeability prediction model, and the mixed color of the two represents the well locations shared by both models.) and (b) lithologic column diagram.

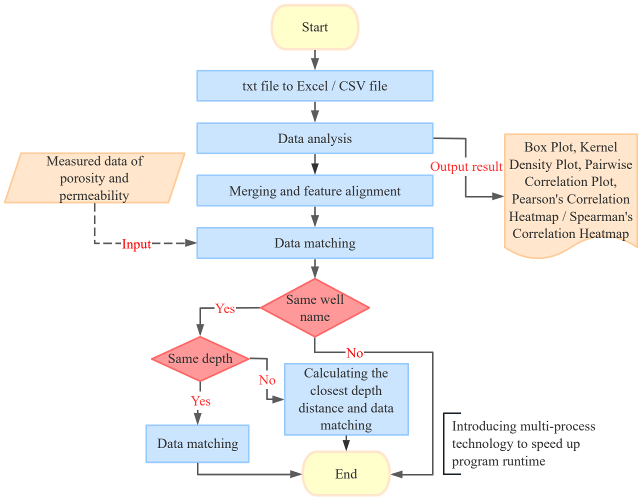

Start
txt file to Excel / CSV file
Data analysis
Output result
Box Plot, Kernel Density Plot, Pairwise Correlation Plot, Pearson's Correlation Heatmap / Spearman's Correlation Heatmap
Measured data of porosity and permeability
Merging and feature alignment
Input
Data matching
Same well name
Yes
No
Same depth
No
Yes
Calculating the closest depth distance and data matching
Data matching
End
Introducing multi-process technology to speed up program runtime

Fig. 2. Workflow of DMML.

ved reserves, respectively (Yang et al., 2023a). During the Xu FM period, the Sichuan Basin was primarily characterized by a topography that was high in the east and low in the west, forming strata gradually thickening from east to west. The Upper Triassic Xu FM constitutes a set of continental coal-bearing clastic rock strata, with a thickness ranging from 250 to 3,500 m. On the basis of sedimentary cycles and lithological characteristics, the Xu FM can be divided into 6 layers, sequentially named as Xu1 to Xu6. During the sedimentary periods of Xu1, Xu3 and Xu5, the basin witnessed periodic rises of lake water levels, leading to mudstone accumulation and the formation of main source rock beds (Liu et al., 2018; Gou et al., 2024). During the Xu2, Xu4 and Xu6 periods, lake water levels in the basin decreased, leading to the widespread development of alluvial fans, braided rivers, deltas, and littoral-shallow lake beach bars. These features represent a basin-wide sedimentary pattern characterized by a predominance of sand, which forms the main reservoir layers of the Xu FM. As such, the source rock and sandstone reservoirs of the Xu FM are closely interbedded, forming a unique "sandwich" structure (Fig. 1(b)) that is conducive to the accumulation of tight sandstone gas.

3. Data and methodology

The basic geological data include 13,640 measured porosity samples and 6,403 measured permeability samples from the Research Institute of Exploration and Development, PetroChina Southwest Oil & Gas Field Company, all conforming to the Chinese National Standard. This collection was accompanied by nine conventional well logging curves, including spontaneous potential (SP, mV), gamma ray (GR, API), caliper (CAL, cm), compensated neutron log (CNL, %), density log (DEN, g/cm^3), acoustic log (AC, µs/m), resistivity log (RT, OMM), deep resistivity log (RLLD, OMM) and shallow resistivity log (RLLS, OMM) (detailed in the Supplementary file). Due to logging issues, the RT, RLLD and RLLS curves are missing in some wells, so these three well logging curves were omitted from the feature selection and modeling process. Finally, the available data set included 2,583 measured porosity data points from 19 wells and 1,043 measured permeability data points from 16 wells (Tables 1 and 2). After relaxing input feature constraints via the PI-Set algorithm, the porosity dataset was expanded to include 6,337 measured data points from 35 wells (see Section 4.3). Owing to the limitations of the original dataset, the permeability dataset could not be expanded. All the datasets were divided into training set and test set in a ratio of 70% to 30%.

3.1 Data preprocessing and preparation

3.1.1 Data preprocessing

Matching measured data with logging data is often a cumbersome task requiring excessive manual effort by researchers. To address this challenge, we developed a software called DMML (Data Matching of Measured Data and Logging Data) (Fig. 2). This can automatically convert logging data text files within the same storage path to Excel/CSV tables and allow for customized data analysis, including Box Plot, Kernel Density (KD) Plot, Pairwise Correlation Plot, and Pearson's/Spearman's Correlation Heatmaps. It calculates the

Table 1. Statistical summary of well log data and porosity of tight sandstone.

Features	Depth (m)	SP (mV)	GR (API)	CAL (cm)	CNL (%)	DEN (g/cm^3)	AC (µs/m)	Porosity (%)
mean	2,776.35	143.55	77.57	9.25	11.01	2.52	63.35	5.784377
std	606.77	114.09	30.40	4.46	6.38	0.15	6.52	2.896617
min	1,946.31	-14.48	26.16	4.51	-0.18	1.39	46.62	0.080000
25%	2,346.54	27.52	55.71	6.62	7.20	2.47	59.23	3.750235
50%	2,757.62	162.61	68.84	7.90	8.95	2.53	62.44	5.340000
75%	3,106.13	246.71	91.40	9.52	13.42	2.61	66.50	7.230000
max	4,518.21	366.32	247.07	29.53	65.99	2.92	101.14	15.897613

Notes: Count means the total number of data, mean is the average value, std stands for the standard deviation, min is the minimum, 25% denotes the first quartile, 50% is the median, 75% represents the third quartile, and max is the maximum.

Table 2. Statistical summary of well log data and permeability of tight sandstone.

Features	Depth (m)	SP (mV)	GR (API)	CAL (cm)	CNL (%)	DEN (g/cm^3)	AC (µs/m)	Permeability (mD)
mean	2,789.75	174.18	86.66	8.46	11.01	2.50	65.22	0.533566
std	747.73	110.04	32.87	3.26	7.75	0.20	7.64	2.094230
min	2,012.64	-14.48	26.16	4.51	-0.18	1.39	46.80	0.000005
25%	2,119.33	58.73	61.41	6.45	6.52	2.46	58.82	0.004400
50%	2,570.20	183.59	75.56	7.82	8.73	2.54	63.11	0.071672
75%	3,133.68	248.99	112.12	9.50	12.60	2.62	67.46	0.241350
max	4,518.21	366.32	214.90	25.28	65.99	2.85	101.14	27.70000

minimum depth difference between each measured data and the corresponding logging data on the basis of well names and depth, and selects the logging data with the closest depth for matching.

3.1.2 Data preparation

The permeability values were logarithmically transformed from skewed distribution to normal distribution for correction. Then, a pairwise correlation plot, including porosity and permeability with depth and each logging parameter, was generated to analyze the correlations between each pair of variables in the dataset (Figs. 3 and 4).

From Figs. 3 and 4, it can be seen that SP, porosity and permeability all exhibit a horizontal distribution parallel to the X-axis by well location, making it impossible to describe their relationships with simple linear or nonlinear correlations. AC and porosity show an obvious nonlinear distribution pattern by well location, while GR with porosity and permeability presents a distinct clustered distribution, suggesting the possibility of other relationships, such as exponential correlation, to describe this distribution. From the pairwise correlation plot, it is possible to observe the data distribution between each pair of features, thus gaining an intuitive understanding of the mathematical relationships between the feature pairs, such as linear, nonlinear, and feature interaction relationships (the latter refers to a specific correlation or interaction between two or more features, i.e., the information gain resulting from the interaction of multiple features). The cross-plot illustrates the data distribution relationship between two features, while the images on the diagonal represent KD plots, which intuitively display the distribution shapes of specific feature data, including peaks, valleys and skewness.

3.2 Feature engineering

In ML problems, data and features set the upper limit of achievable results, with algorithms striving to reach this limit. To this end, feature engineering, which includes data presentation, information extraction and organization, is a pivotal process (Wood, 2023). Feature selection, a key part of feature engineering, involves choosing the most beneficial subset of features, and eliminating redundant or irrelevant ones to enhance model stability, reliability and performance. This process also mitigates overfitting and improves generalizability, rendering the model more comprehensible (Otchere et al., 2022).

Geology researchers have long used various feature selection algorithms, while these often lack interpretability and have inherent flaws. Though some algorithms excel at identifying linear or monotonic relationships (Table 3), they may miss complex interactions. For example, high AC indicates higher pore volume and lower density (DEN), revealing valuable insights into porosity prediction through their interaction.

Using common feature selection algorithms in ML models

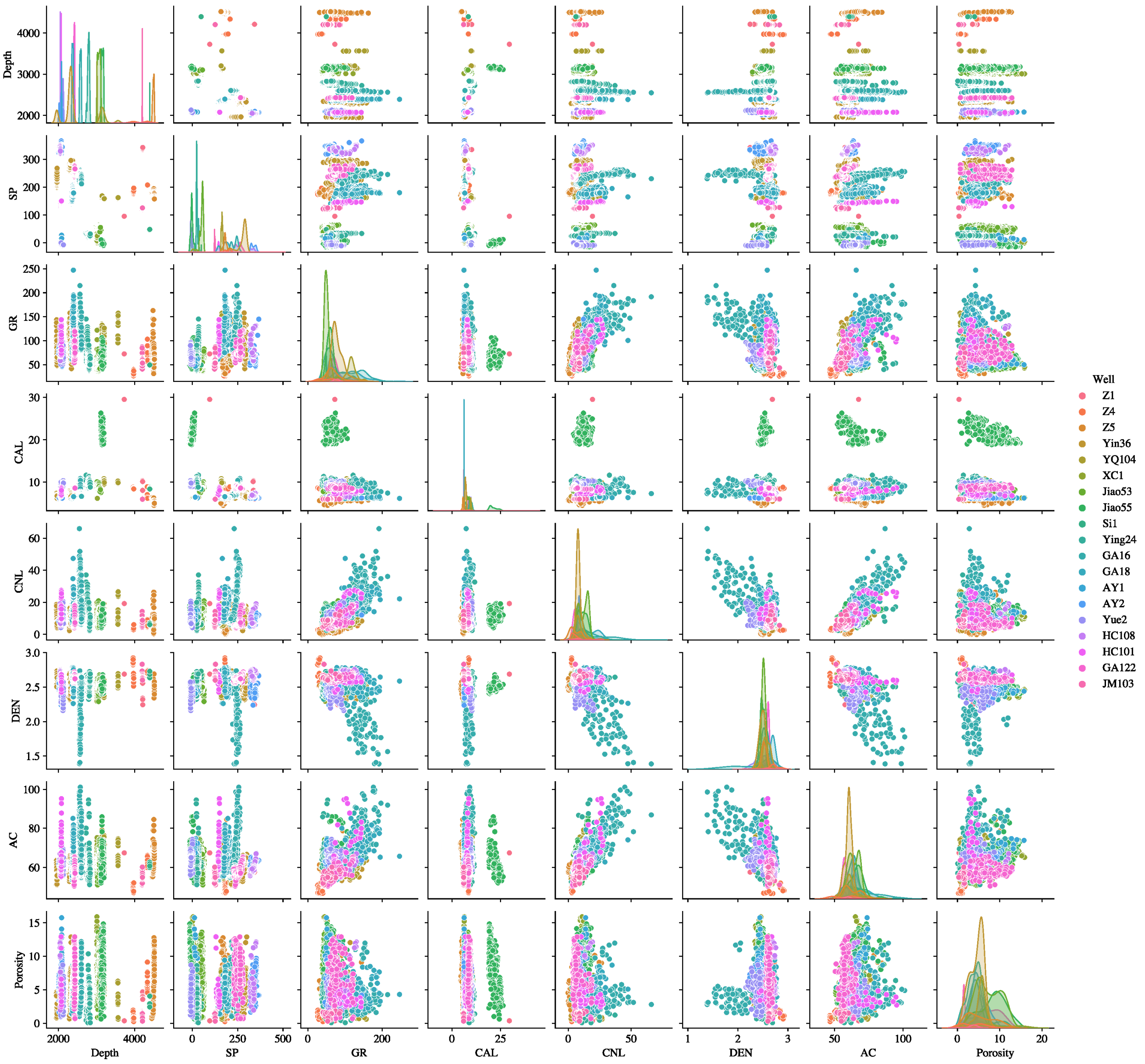


Fig. 3. Pairwise correlation plot of porosity illustrating the data distribution relationship between two features (the units are the same as in Table 1).

for porosity and permeability prediction may result in the loss of intricate information between depth and logging features, leading to reduced model accuracy and interpretability. This flaw is particularly pronounced in the Xu FM of the Sichuan Basin due to its vast area and varied well locations. To address these limitations of common feature selection algorithms, Fisher et al. (2019) proposed the model-independent Permutation Importance (PI) algorithm, which disrupts the relationship between a feature and the target variable by randomly shuffling the feature's values, assessing the impact on model performance to rank feature importance. By repeating this process with different shuffling sequences, PI accurately captures both direct and interaction information between features. This algorithm is suitable for structured data, can be applied to any model and offers high interpretability in feature importance measurement:

For each repetition k in $1 \cdots K$, column j of dataset A is randomly shuffled to generate a corrupted version of the data named $\widetilde{D}_{k,j}$, and $s_{k,j}$ is the computation score of model m on corrupted data $\widetilde{D}_{k,j}$.

However, the PI algorithm relies on the predictive performance of a single model to determine the importance of features. To ensure both high accuracy and interpretability in feature selection, inspired by Zhang (2022)'s work on medical data, this study proposes a novel PI-Set algorithm (Fig. 5). On the basis of the set theory, this algorithm selects models with the best predictive performance, using diverse mathematical methods, ensemble techniques (see Section 3.4), and Decision Tree (DT) types (Table 4). The PI-Set algorithm evaluates contributions from multiple high-performance models, reduc-

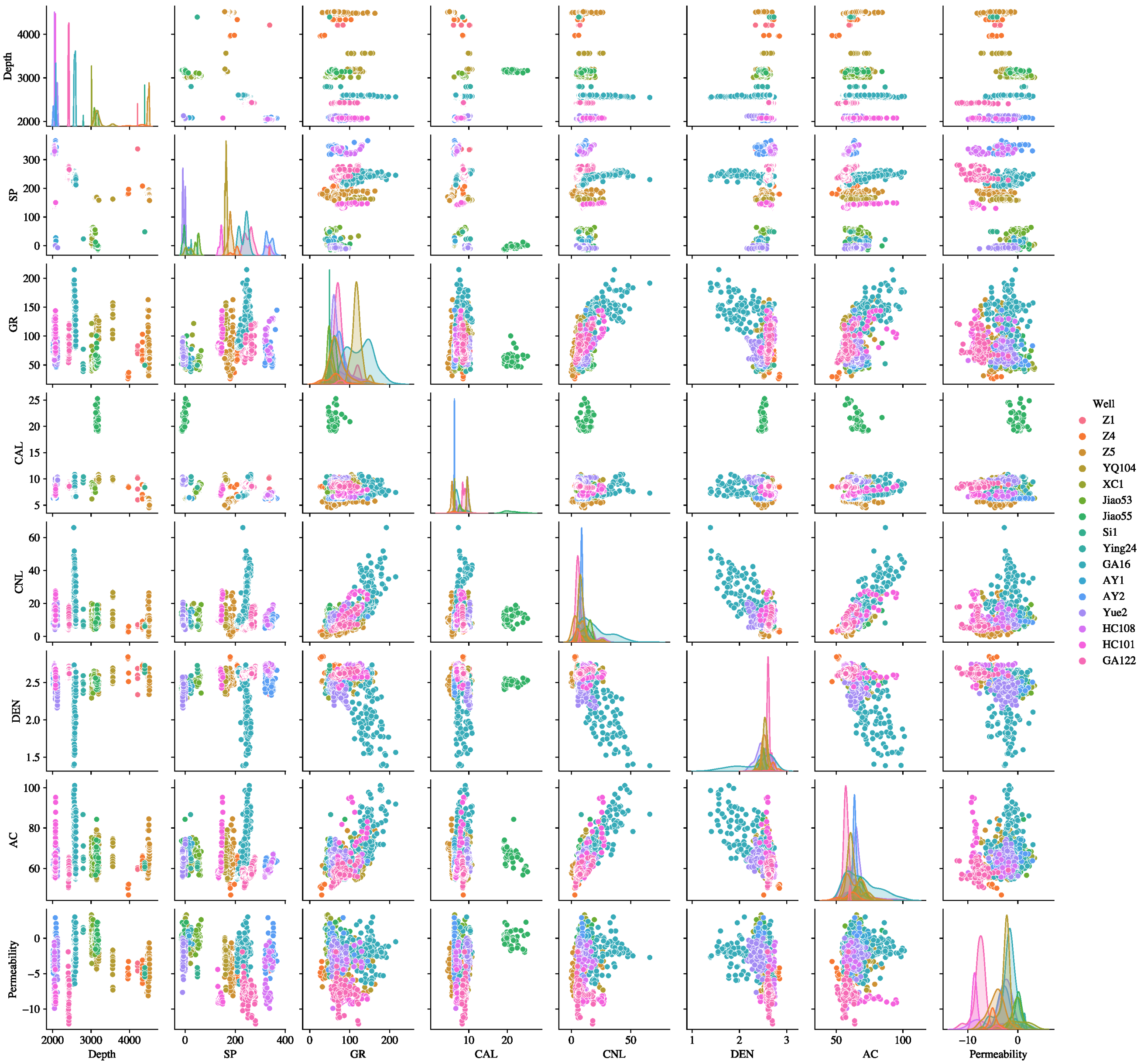


Fig. 4. Pairwise correlation plot of log permeability (the units are the same as in Table 2).

ing bias from any single model and minimizing information inaccuracies.

3.3 Various regression algorithms

Porosity and permeability prediction tasks are treated as regression problems involving numerical predictions. In this study, 10 ML algorithms (including LR (Seber and Wild, 1989), Support Vector Regression (SVR) (Drucker et al., 1996), K-Nearest Neighbors (KNN) (Cover and Hart, 1967), DT (Quinlan, 1986, 1993; Chou, 1991), BPNN (Zhao et al., 2022), RF (Breiman, 2001), XGBoost (Chen and Guestrin, 2016), Light Gradient Boosting Machine (LightGBM) (Ke et al., 2017), Category Boosting (Catboost) (Prokhorenkova et al., 2018), and Stacking (Wolpert, 1992)), encompassing traditional ML, deep learning and ensemble learning techniques, were utilized to predict porosity and permeability (detailed in the Supplementary file).

3.4 Ensemble learning-Stacking

Addressing complex data by a singular model often presents challenges, including limited resistance to noise. Therefore, the objective is to amalgamate multiple models, leveraging their respective advantages and compensating for their deficiencies, to bolster the model's overall generalization capacity, constituting the foundation of ensemble learning. Ensemble learning predominantly adopts two methodologies: a boosting framework, represented by algorithms like LightGBM, XGBoost, and CatBoost, which aims to construct a robust learner by sequentially combining base learners. This method focuses on sequentially improving the base learners by specifically addressing the errors of the previous models (Freund and Schapire, 1997). The bagging framework, as illus-

Table 3. Flaws of the common feature selection algorithms.

Algorithms	Flaws
Manual selection	Relies on individual experience
LR (Seber and Wild, 1989)	Interaction information is not considered
Pearson (Hauke and Kossowski, 2011)	Identifies only linear relationships
Spearman (Hauke and Kossowski, 2011)	Identifies only monotonic relationships
PCA (Jolliffe, 1986)	Insensitive to nonlinear relationships
GCA (Deng, 1982)	Requires a sequence among the data
AHP (Saaty, 1988)	Excessive subjectivity
Tree-based Algorithms (Breiman, 2001)	Existence of feature preferences

Notes: Pearson: Pearson's Correlation Coefficient; Spearman: Spearman's Rank Correlation Coefficient.

trated by RF, creates several independent models and merges their predictions through "voting" or "averaging" strategies to form a potent learner. This approach is known for its ability to reduce variance and improve prediction stability (Breiman, 1996).

Furthermore, Stacking represents a hybrid approach, merging the aspects of both boosting and bagging techniques. Stacking involves employing a variety of base learners to process the original dataset and then using the predictions from these base learners as inputs for a subsequent meta-learner model (Algorithm 1). The essence of Stacking is to use the predictive outputs of various base learners as new inputs for training a meta-learner. This approach capitalizes on the diversity among base learners to boost the overall prediction accuracy. One of Stacking's key benefits is its capacity to integrate the strengths of numerous base learners through a meta-learner, thereby achieving enhanced performance outcomes. Additionally, Stacking can mitigate the risk of overfitting, since the meta-learner is trained on the predictions from the base learners instead of directly on the original data features (Wolpert, 1992). Typically, it is advisable to select ML algorithms with strong performance as base learners while opting for algorithms of lower complexity (such as the LR algorithm) for the meta-learner. This approach balances the ability of Stacking to capture complex patterns with the need to maintain overall model simplicity and interpretability.

$$i_j = s - \frac{1}{K \sum_{k=1}^{K} s_{k,j}} \tag{1}$$

where j represents the numbering of feature tabular dataset A; i_j presents the computation importance for feature j; s represents the computation reference score of the validation model m on data A.

Table 4. Baseline models of the PI-Set, which utilize different mathematical computation methods, ensemble methods, and DT types.

Algorithms	Mathematical methods	Ensemble methods	DT types
RF	Gini coefficient	Bagging	CART
XGBoost	Information entropy	Boosting	GBDT
CatBoost	Information entropy	Boosting	Symmetric trees

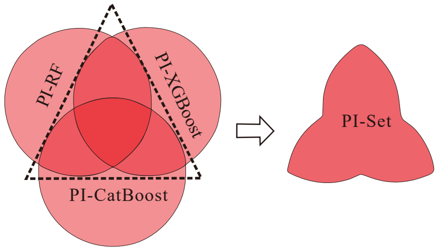


Fig. 5. Architecture of the PI-Set algorithm (a feature is deemed important and selected only if it is considered as such by two or more baseline models).

Algorithm 1: The pseudocode of Stacking.

```
Input  : Training dataset:
         $D = (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), x_n, y_n \in R$
Input  : Base learners: $\mathfrak{L}_1, \mathfrak{L}_2, ..., \mathfrak{L}_T$;
Input  : Meta-learner: $\mathfrak{L}$.
Output: $H(x) = h'(h_1(x), h_2(x), ..., h_T(x))$
for $t = 1, 2, \cdots, T$ do
    $h_t = \mathfrak{L}_t(D)$;
end
$D' = \emptyset$;
for $i = 1, 2, \cdots, m$ do
    for $t = 1, 2, \cdots, T$ do
        $z_{it} = h_t(x_i)$;
    end
    $D' = D' \cup ((zi1, zi2, ..., ziT), yi)$;
end
$h' = \mathfrak{L}(D')$;
```

Stacking can be implemented by two primary methods (Fig. 6). The first employs the K-fold cross-validation method to partition the original dataset into k subsets, it utilizes diverse ML algorithms to predict on each subset, and ultimately amalgamates the predictions. The second method involves applying different ML algorithms directly to the original dataset and integrating the outcomes of these algorithms. This strategy ensures a comprehensive utilization of the data, enhancing the ensemble's predictive performance. In this study, the second method was utilized to construct the Stacking model, integrating the outputs of various ML algorithms applied directly to the original dataset.

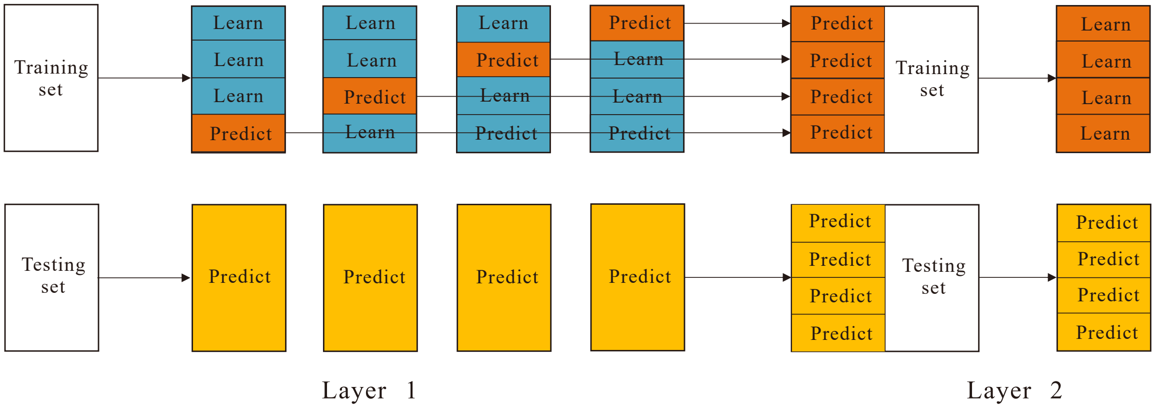


Fig. 6. Two methods of constructing the Stacking algorithms (the horizontal lines represent data sets that are not involved in modeling). The second method was utilized to construct the Stacking model.

Table 5. Mathematical expression and value ranges of common evaluation metrics for five types of regression models.

Evaluation metric	Mathematical expression	Range of values
R-squared (R^2)	$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y}_i - y_i)^2}$	$(-\infty, 1]$
Mean Squared Error (MSE)	$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$[0, +\infty)$
Root Mean Squared Error (RMSE)	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	$[0, +\infty)$
Mean Absolute Error (MAE)	$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	$[0, +\infty)$
Mean Absolute Percentage Error (MAPE)	$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left\lvert\frac{\hat{y}_i - y_i}{y_i}\right\rvert$	$[0, +\infty)$

3.5 Model evaluation metrics

In order to assess the ML model's predictive performance, it is essential to employ various evaluation metrics. Each evaluation metric offers unique benefits and limitations (detailed in the Supplementary file), underscoring the importance of selecting the most appropriate indicator based on the specific problem, the nature of the dataset, and the characteristics of the model under consideration. In practice, to gain a general view of a model's effectiveness, it is advisable to assess its performance comprehensively by integrating different metrics. This consideration is vital for accurately evaluating the predictive performance and interpretability of the ML model concerning a given problem. In this study, five evaluation metrics were utilized to thoroughly evaluate the proposed model's predictive capabilities (Table 5).

4. Results

In this section, the innovative PI-Set algorithm is employed to rank the features based on their importance, providing an in-depth explanation of the feature selection process. Then, a comprehensive ablation study is conducted to evaluate the influence of this feature selection. Following this, ten ML models, including traditional ML, ensemble learning and deep learning methods, are developed to predict porosity and permeability in tight sandstone using well logging data. The performance of these ML models is thoroughly assessed by five different regression evaluation metrics. In addition, the research includes detailed hyperparameter settings for the base learners in the Stacking model.

4.1 Feature selection

In order to address the limitations of traditional feature selection algorithms, this study proposes a novel PI-Set algorithm on the basis of the PI algorithm and the set theory. Unlike conventional approaches, this algorithm uses baseline models with distinct mathematical methods, ensemble techniques and DT types (data structures). This innovative approach allows for a comprehensive consideration of the relationships between feature variables and target variables.

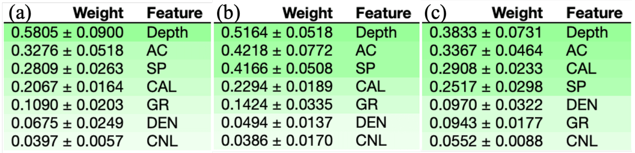
For porosity and permeability feature selection models, RF, XGBoost and CatBoost were chosen as the baseline models for the PI-Set model, aiming for optimal prediction performance (Tables 6 and 7) and significant model differentiation (Table 4). Consequently, the PI-Set feature selection models for porosity and permeability prediction were established (Figs. 7 and 8).

$$\begin{aligned}&\text{Porosity}_{PI-Set} = \\ &(\text{Depth}, \text{AC}, \text{SP}, \text{CAL}, \text{GR}, \text{DEN}, \text{CNL})_{PI-RF} \\ &\cup (\text{Depth}, \text{AC}, \text{SP}, \text{CAL}, \text{GR}, \text{DEN}, \text{CNL})_{PI-XGBoost} \\ &\cup (\text{Depth}, \text{AC}, \text{CAL}, \text{SP}, \text{DEN}, \text{GR}, \text{CNL})_{PI-LightGBM} \\ &= (\text{Depth}, \text{AC}, \text{SP}, \text{CAL}, \text{GR}, \text{DEN}, \text{CNL})\end{aligned} \quad (2)$$

On the basis of the principle that "a feature is deemed important and selected only if it is considered as such by two or more baseline models", the ranking of features by their importance to porosity, from highest to lowest, is as follows:

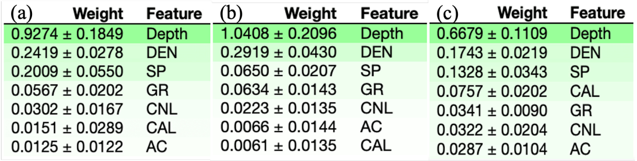
Table 6. Results of five evaluation metrics for ten porosity ML prediction models.

Category	Models	R^2 (%)	MSE	RMSE	MAE	MAPE (%)
Traditional ML	LR	22.50	6.41	2.53	1.92	54
	DT	67.92	2.65	1.63	1.07	24
	SVR	55.80	3.66	1.91	1.32	32
	KNN	69.09	2.56	1.60	1.12	27
Deep learning	BPNN	39.71	4.99	2.23	1.70	46
Ensemble learning	RF	81.66	1.52	1.23	0.86	21
	XGBoost	80.48	1.62	1.27	0.90	22
	LightGBM	78.16	1.81	1.34	0.95	24
	CatBoost	79.96	1.66	1.29	0.92	23
	Stacking	83.04	1.40	1.18	0.83	20



(a) Weight	Feature	(b) Weight	Feature	(c) Weight	Feature
0.5805 ± 0.0900	Depth	0.5164 ± 0.0518	Depth	0.3833 ± 0.0731	Depth
0.3276 ± 0.0518	AC	0.4218 ± 0.0772	AC	0.3367 ± 0.0464	AC
0.2809 ± 0.0263	SP	0.4166 ± 0.0508	SP	0.2908 ± 0.0233	CAL
0.2067 ± 0.0164	CAL	0.2294 ± 0.0189	CAL	0.2517 ± 0.0298	SP
0.1090 ± 0.0203	GR	0.1424 ± 0.0335	GR	0.0970 ± 0.0322	DEN
0.0675 ± 0.0249	DEN	0.0494 ± 0.0137	DEN	0.0943 ± 0.0177	GR
0.0397 ± 0.0057	CNL	0.0386 ± 0.0170	CNL	0.0552 ± 0.0088	CNL

Fig. 7. Results of the PI-Set feature selection model for porosity: (a) PI-RF, (b) PI-XGBoost and (c) PI-CatBoost. The first numbers in each line of the arrangement result indicate the degree of model performance attenuation. Negative values indicate that the prediction result of the disrupted feature is more accurate than the real data. The numbers after ± represent the standard deviation of multiple shuffling.



(a) Weight	Feature	(b) Weight	Feature	(c) Weight	Feature
0.9274 ± 0.1849	Depth	1.0408 ± 0.2096	Depth	0.6679 ± 0.1109	Depth
0.2419 ± 0.0278	DEN	0.2919 ± 0.0430	DEN	0.1743 ± 0.0219	DEN
0.2009 ± 0.0550	SP	0.0650 ± 0.0207	SP	0.1328 ± 0.0343	SP
0.0567 ± 0.0202	GR	0.0634 ± 0.0143	GR	0.0757 ± 0.0202	CAL
0.0302 ± 0.0167	CNL	0.0223 ± 0.0135	CNL	0.0341 ± 0.0090	GR
0.0151 ± 0.0289	CAL	0.0066 ± 0.0144	AC	0.0322 ± 0.0204	CNL
0.0125 ± 0.0122	AC	0.0061 ± 0.0135	CAL	0.0287 ± 0.0104	AC

Fig. 8. Results of the PI-Set feature selection model for permeability: (a) PI-RF, (b) PI-XGBoost and (c) PI-CatBoost.

Depth, AC, SP, CAL, GR, DEN, CNL.

$$
\begin{aligned}
&\text{Permeability}_{PI-Set} = \\
&(\text{Depth}, \text{DEN}, \text{SP}, \text{GR}, \text{CNL}, \text{CAL}, \text{AC})_{PI-RF} \\
&\cup (\text{Depth}, \text{DEN}, \text{SP}, \text{GR}, \text{CNL}, \text{AC}, \text{CAL})_{PI-XGBoost} \\
&\cup (\text{Depth}, \text{DEN}, \text{SP}, \text{CAL}, \text{GR}, \text{CNL}, \text{AC})_{PI-LightGBM} \\
&= (\text{Depth}, \text{DEN}, \text{SP}, \text{GR}, \text{CNL}, \text{CAL}, \text{AC})
\end{aligned} \tag{3}
$$

The ranking of features by their importance to permeability, from highest to lowest, is as follows: Depth, DEN, SP, GR, CNL, CAL, AC.

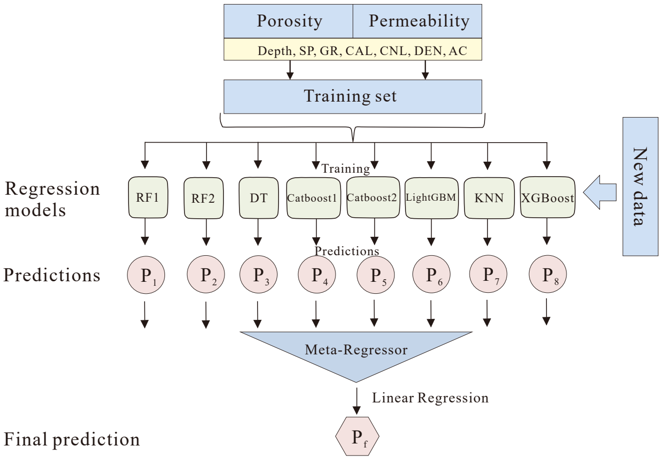


Fig. 9. The actual Stacking models for porosity and permeability prediction were constructed in this paper (to reduce the risk of overfitting, there are substantial differences between base learners).

The analysis of the PI-Set models for porosity and permeability revealed that baseline models with significant differences display remarkable consistency in the ranking of feature importance. This consistency deviates from the traditionally accepted significance of well logging features, a topic explored further in the discussion section. The next Section 4.3 presents an ablation study on input features. Both the porosity and permeability prediction models achieved their highest accuracy when using the top three features identified by the PI-Set algorithm, which validates the precision and robustness of this novel algorithm.

4.2 Prediction

Among the nine ML models evaluated for predicting porosity and permeability, DT, RF, CatBoost, LightGBM, XGBoost, and KNN achieved the best performance (Tables 6 and 7, Figs. 10 and 11). Due to their substantial differences, these six models were selected as base learners for the Stacking ensemble model, with the LR algorithm chosen as the meta-learner for its simplicity and low complexity (Fig. 9).

In order to promote model diversity and obtain comprehensive differential information without overfitting, this study used the RF and CatBoost models twice. We first used the Bayesian optimization algorithm (Pelikan et al., 2000) (detailed in the Supplementary file) to find the optimal hyperparameter configurations for RF1 and CatBoost1, then randomly set the hyperparameter values for RF2 and CatBoost2. This strategy introduced variations in the dataset and hyperparameters between RF1 and RF2, as well as CatBoost1 and CatBoost2 (detailed in the Supplementary file), with the aim to provide the Stacking model with sufficient differential information to enhance predictive accuracy and prevent overfitting.

As can be observed from Table 7, the MAPE values for the permeability prediction model display exceptionally high figures. As mentioned in Section 3.5 and Supplementary file, MAPE penalizes true values that are close to zero, leading to abnormally high MAPE values. Given that the permeability values in tight sandstone reservoirs often approach near-zero true values, MAPE is not a suitable metric for evaluating the performance of permeability prediction models. Considering

Table 7. Results of five evaluation metrics for ten permeability ML prediction models.

Category	Models	R^2 (%)	MSE	RMSE	MAE	MAPE (%)
Traditional ML	LR	22.31	7.11	2.67	2.18	3.06×10^{13}
	DT	73.35	2.44	1.56	1.05	2.46×10^{12}
	SVR	64.62	3.24	1.80	1.30	3.25×10^{11}
	KNN	80.63	1.77	1.33	0.93	1.66×10^{12}
Deep learning	BPNN	72.12	2.55	1.60	1.12	1.02×10^{13}
Ensemble Learning	RF	80.14	1.82	1.35	0.88	1.16×10^{12}
	XGBoost	80.53	1.78	1.33	0.90	1.05×10^{12}
	LightGBM	80.40	1.79	1.34	0.95	5.46×10^{11}
	CatBoost	80.84	1.75	1.32	0.92	1.67×10^{12}
	Stacking	82.84	1.57	1.25	0.86	4.64×10^{13}

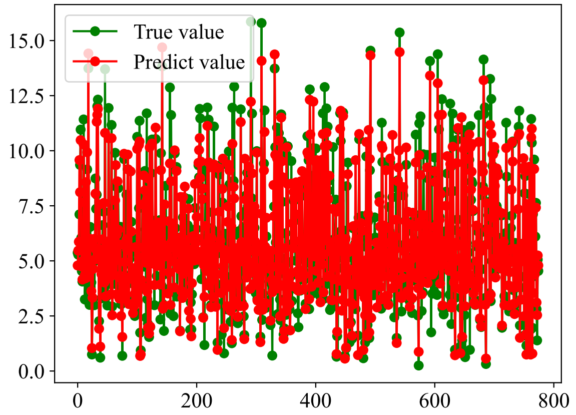


Fig. 10. The R^2 result of porosity prediction for Stacking model (The Stacking is 83.04%).

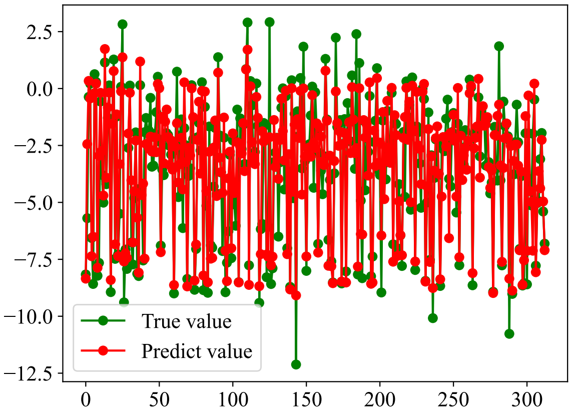


Fig. 11. The R^2 result of permeability prediction for Stacking model (The Stacking is 82.84%).

that the subject of this paper is the Xu FM across the entire Sichuan Basin, with data points distributed over several different gas fields, the vast geological block and the complex geological conditions contain a wealth of information, presenting significant challenges to model prediction. Therefore, this study posits that the most appropriate evaluation metric for a basin-level porosity and permeability prediction model is the R^2, as it measures the model's ability to capture information.

4.3 Ablation studies

Ablation studies are critical to ML research as they elucidate the impact of individual feature variables within the black box of ML models. These studies offer a tangible means to visualize the influence of specific features on the performance of a ML system. In this context, we evaluated the performance of ten ML models in predicting porosity and permeability with varying input features.

As indicated by the results in Tables 8 and 9, and Figs. 12 and 13, the porosity prediction model and the permeability prediction model both exhibited optimal performance when the PI-Set model selected the top three important features as inputs. The best-performing Stacking model attained an $R^2 = 85.15\%$ accuracy in the porosity prediction model and an $R^2 = 83.62\%$ accuracy in the permeability prediction model. These results surpassed the accuracy achieved when all seven features were input into the Stacking model, which yielded an $R^2 = 83.04\%$ for the porosity prediction and $R^2 = 82.84\%$ for the permeability prediction. It was observed that the Stacking models, which performed optimally, consistently achieved the highest prediction accuracy when the top three features identified by the PI-Set model proposed in this study were used as inputs. This outcome underscores the PI-Set model's exceptional accuracy and robustness in constructing porosity and permeability feature selection models. Specifically, this research focuses on the Xu FM across the entire Sichuan Basin, where the PI-Set model is required to extract sufficient effective information from a vast geological area and intricate logging data to identify the optimal feature combination. The high level of accuracy and consistency demonstrated by the PI-Set model in this complex setting fully attests to the effectiveness of the PI-Set algorithm.

Table 8. Results of R^2 (%) metrics of the feature ablation study for ten porosity ML prediction models.

Input feature	LR	DT	SVR	KNN	BPNN	RF	XGBoost	LightGBM	CatBoost	Stacking
Depth	0.56	49.64	23.49	72.50	3.65	65.99	65.66	61.07	64.74	72.79
Depth, AC	2.28	74.10	29.47	49.72	24.30	80.10	76.96	76.25	72.28	82.20
Depth, AC, SP	5.84	74.75	35.64	62.62	17.56	84.45	82.44	76.68	78.54	85.15
Depth, AC, SP, CAL	7.09	65.56	37.91	56.73	30.87	83.05	79.01	75.99	77.81	83.76
Depth, AC, SP, CAL, GR	6.65	68.13	41.78	64.26	32.80	81.94	80.00	75.06	77.76	83.55
Depth, AC, SP, CAL, GR, DEN	9.24	67.11	48.72	67.78	37.26	82.31	80.31	76.24	79.01	82.85
Depth, AC, SP, CAL, GR, DEN, CNL	22.50	67.92	55.80	69.09	39.71	81.66	80.48	78.16	79.96	83.04
AC, SP, CAL, GR, DEN, CNL	22.52	53.57	50.90	67.70	40.13	75.14	71.61	71.23	73.82	76.90

Table 9. Results of R^2 (%) metrics of the feature ablation study for ten permeability ML prediction models.

Input feature	LR	DT	SVR	KNN	BPNN	RF	XGBoost	LightGBM	CatBoost	Stacking
Depth	1.34	58.89	25.76	74.86	49.50	68.19	66.19	74.97	71.90	74.45
Depth, DEN	9.32	74.10	38.98	56.77	45.83	80.10	76.96	76.25	77.50	79.50
Depth, DEN, SP	19.17	76.14	56.43	77.42	75.03	82.93	82.19	81.15	81.43	83.62
Depth, DEN, SP, GR	19.41	72.94	64.64	76.71	73.14	82.85	81.27	80.86	82.89	83.07
Depth, DEN, SP, GR, CNL	19.37	66.85	63.58	78.43	69.74	81.60	81.28	81.28	82.04	82.74
Depth, DEN, SP, GR, CNL, CAL	22.34	70.74	64.89	76.67	76.99	79.81	80.41	80.86	80.21	82.31
Depth, DEN, SP, GR, CNL, CAL, AC	22.31	73.35	64.62	80.63	72.12	80.14	80.53	80.40	80.84	82.84
DEN, SP, GR, CNL, CAL, AC	21.46	58.94	49.55	72.98	61.40	75.93	77.97	77.34	77.55	78.14

4.4 Model extension

By adeptly capturing the interplay between feature variables (depth and conventional logging data) and target variables (porosity and permeability) with sufficient differentiation in the baseline model, the PI-Set model further loosens the constraint on the logging feature data. This flexibility permits the exploration of expanding the dataset utilized for building predictive models within the scope of the gathered data.

In the case of the porosity prediction model, Depth, AC, and SP-the three most crucial features identified by the PI-Set model, which also correspond to the best-performing model configuration-were chosen (Table 8 and Fig. 12). Subsequently, a porosity dataset consisting of 6,337 measured data points from 35 wells was compiled (as illustrated in Figs. S1 and S2 of the Supplementary file). This dataset covers a comprehensive range of well locations throughout the Xu FM in the entire Sichuan Basin. A split of 70% of this data was utilized as the training set for feeding into the model, while the remaining 30% served as the test set to evaluate the model's efficacy. After significantly increasing both the volume of data and the number of well locations, with only two logging curves, the optimally performing Stacking model still maintained an accuracy of $R^2 = 84.04\%$ (Table 10 and Fig. S7 of the Supplementary file).

Despite the loosening of input feature constraints by the PI-Set model, the permeability dataset could not be similarly expanded.

5. Discussion

As observed in Figs. 3 and 4, depth exhibits a well-distributed relationship with porosity and permeability on a well-by-well basis. The PI-Set model further validates that depth is an extremely important feature for the prediction of porosity and permeability. In ablation studies that only considered depth, the prediction accuracies for porosity and permeability reached 72.79% and 74.45%, respectively. Our study suggests that the compaction effect experienced during the formation of the tight sandstone reservoirs in the Xu FM

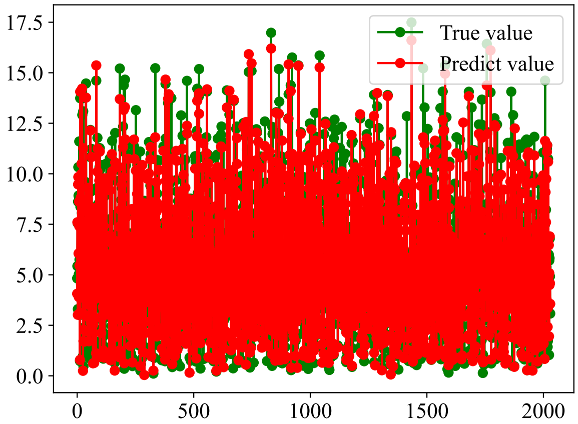


Fig. 12. R^2 results of feature ablation study for Stacking porosity prediction models (The Stacking is 85.15%).

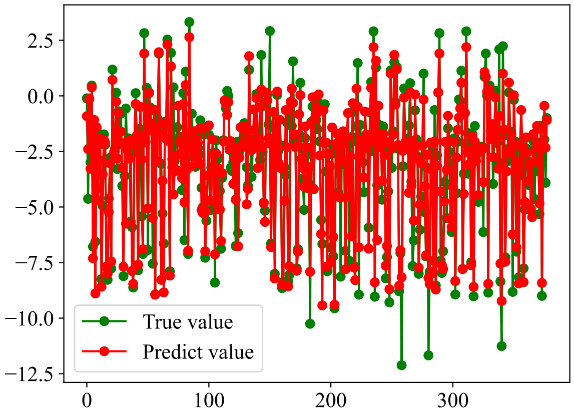


Fig. 13. R^2 results of feature ablation study for Stacking permeability prediction models (The Stacking is 83.62%).

of the Sichuan Basin has a significant impact on the reservoir's porosity and permeability. Such impacts are accurately captured and reflected in the feature importance and model accuracy by the artificial intelligence model; therefore, the depth factor should be taken seriously in the modeling process of porosity and permeability prediction.

In traditional logging interpretations, DEN, AC and CNL are often regarded as the most pertinent logs to porosity. The assessments of the PI-Set model, developed employing PI-RF, PI-XGBoost and PI-CatBoost, demonstrate remarkable consistency, unanimously identifying AC, SP and CAL as the key logging features for the porosity prediction model. Moreover, the optimal configuration for both porosity and permeability prediction models include depth plus one porosity logging curve and one lithology logging curve. Despite the PI-Set' reliance on diverse mathematical methods, DT types and ensemble techniques, this agreement persists. SP and CAL are indicators of formation lithology changes, signified through potential changes and borehole diameter variations, respectively. From our results, it can be suggested that this is due to the PI-Set model capturing unique information from the widespread sand-mud interlayering phenomena in the Xu FM of the Sichuan Basin, providing crucial insights for model predictions.

Table 10. Results of five evaluation metrics for ten porosity ML prediction models after expanding the dataset (the input features are Depth, AC and SP).

Category	Models	R^2 (%)	MSE	RMSE	MAE	MAPE (%)
Traditional ML	LR	13.85	9.04	3.01	2.37	91
	DT	70.66	3.08	1.75	1.02	31
	SVR	39.80	6.31	2.51	1.78	64
	KNN	66.13	3.55	1.89	1.30	43
Deep learning	BPNN	33.99	6.92	2.63	1.98	75
Ensemble learning	RF	81.38	1.95	1.40	0.91	31
	XGBoost	79.31	2.17	1.47	1.00	33
	LightGBM	76.10	2.51	1.58	1.12	38
	CatBoost	77.28	2.38	1.54	1.10	37
	Stacking	84.04	1.68	1.29	0.85	27

In addition, as demonstrated in Fig. 3, DEN and CNL, along with SP, exhibit collinear data distributions in opposite directions yet with relatively consistent patterns. This observation further corroborates the presence of highly correlated feature interaction information between SP, DEN and CNL, indirectly validating their interconnectedness. In other words, the SP effectively encapsulates the information pertinent to porosity prediction that is contained within DEN and CNL. The results of ablation studies show that the PI algorithm can well capture the potential nonlinear relationship between feature variables and target variables well, and effectively mine the interaction information between feature variables and target variables. Furthermore, it can effectively uncover the interaction information among feature variables. This approach maximizes the shared information across different models and ensures that the feature selection process is both interpretable and transparent ("white box") for geologists.

This study establishes base learners with substantial differences by developing diverse ML models and creating differential models among identical ML models. The application of Stacking ensemble learning method effectively leverages the disparities among base learners, which overcomes the limitations of unsatisfactory prediction accuracy inherent to single models, thereby enhancing the overall predictive performance and reducing the risk of model overfitting. Additionally, the Stacking method is trained on the predictions made by base learners rather than on the original features, which also mitigates the risk of overfitting to a certain degree. After significantly increasing both the volume of data and the number of well locations, with only two logging curves, the optimally performing Stacking model still maintained a high accuracy of $R^2 = 84.04\%$, conclusively proving the efficacy of the PI-Set model in feature selection and the robustness of the Stacking model in performing prediction across the entire basin.

Using well logging data and ML techniques to predict porosity and permeability can enhance the utilization of this data, boosting efficiency and reducing laboratory analysis costs. During the exploration phase, accurate predictions guide drilling decisions and mitigate exploration risks. In the development phase, these predictions provide support for real-time reservoir monitoring, facilitate adjustments to development plans, and increase operational efficiency. Consequently, these methods lower exploration and production costs while maximizing economic returns. Furthermore, basin-scale predictions using these interdisciplinary techniques are particularly important for fuel science: they advance data-driven research, deepen the understanding of complex reservoir properties, and enable optimized exploration and development of tight sandstone gas resources. This in turn provides critical insights and supports research advancements, contributing to more effective and sustainable energy resource management.

6. Conclusions

This study has presented a new ML model and conducted data preprocessing, feature engineering, hyperparameter tuning, model construction, extension and evaluation. Each step has been designed to be interpretable and repeatable for geological researchers. From the results, the following key points can be highlighted:

1) This study developed an automated software named "DMML" for preprocessing measured and logging data.
2) The PI-Set algorithm proposed in this study is able to quantify the relationship between feature variables and target variables, addressing the limitations of common feature selection algorithms through multi-angle exploration. The intuitive principle of this algorithm ensures the interpretability of feature selection results for researchers.
3) Five common evaluation metrics were explored in the prediction of porosity and permeability by regression models. It was found that MAPE is unsuitable for permeability predictions in tight sandstone reservoirs, as low permeability values lead to inflated MAPE scores. R^2 is recommended as the most appropriate metric for basin-level porosity and permeability prediction models, since it effectively measures the information capture ability of the model.
4) Ten ML methods, including traditional, deep learning and ensemble learning techniques, were used to predict porosity and permeability in the Xu FM of the Sichuan Basin. The Stacking algorithm from ensemble learning achieved prediction accuracies of 85.15% for porosity and 83.62% for permeability. The PI-Set model, as indicated by ablation studies, reduced the input requirements for well logging data.

In future research, we intend to investigate the effectiveness of the PI-Set algorithm in other geological problems and explore the transferability of porosity and permeability prediction models in other basins and blocks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 42302142 and 42372147) and the Collaborative Project between the PetroChina Southwest Oil & Gas Field Company and China University of Petroleum (Beijing) (No. HX20231362). We are grateful to the Research Institute of Exploration and Development, PetroChina Southwest Oil & Gas Field Company for providing the data.

Supplementary file

https://doi.org/10.46690/ager.2025.04.04

Conflict of interest

The authors declare no competing interest.

References

Alfi, M., Hosseini, S. A., Enriquez, D., et al. A new technique for permeability calculation of core samples from unconventional gas reservoirs. Fuel, 2019, 235: 301-305.

Al Khalifah, H., Glover, P. W. J., Lorinczi, P. Permeability prediction and diagenesis in tight carbonates using machine learning techniques. Marine and Petroleum Geology, 2020, 112: 104096.

Ampomah, W., Balch, R. S., Cather, M., et al. Optimum design of CO_2 storage and oil recovery under geological uncertainty. Applied Energy, 2017, 195: 80-92.

Aras, S., Hanifi Van, M. An interpretable forecasting framework for energy consumption and CO_2 emissions. Applied Energy, 2022, 328: 120163.

Belhouchet, H. E., Benzagouta, M. S., Dobbi, A., et al. A new empirical model for enhancing well log permeability prediction, using nonlinear regression method: Case study from Hassi-Berkine oil field reservoir-Algeria. Journal of King Saud University-Engineering Sciences, 2021, 33: 136-145.

Breiman, L. Bagging predictors. Machine Learning, 1996, 24: 123-140.

Breiman, L. Random forests. Machine Learning, 2001, 45: 5-32.

Chai, X., Tian, L., Wang, J., et al. A novel prediction model of oil-water relative permeability based on fractal theory in porous media. Fuel, 2024, 372: 131840.

Chen, T., Guestrin, C. XGBoost: A scalable tree boosting system. Paper Presented at the Knowledge Discovery and Data Mining, ACM, San Francisco, CA, USA, 13-17 August, 2016.

Chork, C. Y., Jian, F. X., Taggart, I. J. Porosity and permeability estimation based on segmented well log data. Journal of Petroleum Science and Engineering, 1994, 11: 227-239.

Chou, P. A. Optimal partitioning for classification and regression trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 13: 340-354.

Cover, T., Hart, P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 1967, 13: 21-27.

Deng, J. Control problems of grey systems. Systems & Control Letters, 1982, 1(5): 288-294.

Deng, J., Liu, M., Ji, Y., et al. Controlling factors of tight sandstone gas accumulation and enrichment in the slope zone of foreland basins: The Upper Triassic Xujiahe Formation in Western Sichuan Foreland Basin, China. Journal of Petroleum Science and Engineering, 2022, 214: 110474.

Drucker, H., Burges, C. J. C., Kaufman, L., et al. Support vector regression machines. Paper Presented at Neural Information Processing Systems, USA, 2-5 December, 1996.

Fisher, A., Rudin, C., Dominici, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research, 2019, 20(177): 1-81.

Freund, Y., Schapire, R. E. A Decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 1997, 55: 119-139.

Gou, M., Lu, G., Deng, B., et al. Tectonic-paleogeographic evolution of the Late Triassic in the Sichuan basin, SW China: Constraints from sedimentary facies and provenance analysis of the Xujiahe Formation. Marine and Petroleum Geology, 2024, 160: 106649.

Hauke, J., Kossowski, T. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. Quaestiones Geographicae, 2011, 30: 87-93.

Jiang, D., Chen, H., Xing, J., et al. A new method for dynamic predicting porosity and permeability of low permeability and tight reservoir under effective overburden pressure based on BP neural network. Geoenergy Science and Engineering, 2023, 226: 211721.

Jolliffe, I. T. Principal Component Analysis, Springer Series in Statistics. New York, USA, Springer New York, 1986.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., et al. Machine learning for the geosciences: Challenges and opportunities. IEEE Transactions On Knowledge and Data Engineering, 2019, 31: 1544-1554.

Ke, G., Meng, Q., Finley, T., et al. LightGBM: A highly efficient gradient boosting decision tree. Paper Presented at Neural Information Processing Systems, Long Beach, CA, USA, 4-9 December, 2017.

Liu, J., Cao, J., Hu, G., et al. Water-level and redox fluctuations in a Sichuan Basin lacustrine system coincident with the Toarcian OAE. Palaeogeography, Palaeoclimatology, Palaeoecology, 2020, 558: 109942.

Liu, Y., Hu, W., Cao, J., et al. Diagenetic constraints on the heterogeneity of tight sandstone reservoirs: A case study on the Upper Triassic Xujiahe Formation in the Sichuan Basin, southwest China. Marine and Petroleum Geology, 2018, 92: 650-669.

Lu, H., Li, Q., Yue, D., et al. Study on optimal selection of porosity logging interpretation methods for Chang 73 segment of the Yanchang Formation in the southwestern Ordos Basin, China. Journal of Petroleum Science and Engineering, 2021, 198: 108153.

Otchere, D. A., Ganat, T. O. A., Ojero, J. O., et al. Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. Journal of Petroleum Science and Engineering, 2022, 208: 109244.

Pelikan, M., Goldberg, D. E., Cantu-Paz, E. Bayesian optimization algorithm, population sizing, and time to convergence. Paper Presented at Proceedings of the Genetic and Evolutionary Computation Conference, Las Vegas, Nevada, USA, 8-12 July, 2000.

Prokhorenkova, L., Gusev, G., Vorobev, A., et al. CatBoost: Unbiased boosting with categorical features. Paper Presented at Neural Information Processing Systems, Canada, 3-8 December, 2018.

Quinlan, J. R. Induction of decision trees. Machine Learning, 1986, 1: 81-106.

Quinlan, J. R. C4.5: Programs for Machine Learning. San Mateo, California, USA, Morgan Kaufmann Publishers, 1993.

Saaty, T. L. Multicriteria Decision Making: The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation. New York, USA, McGraw-Hill, 1988.

Seber, G. A. F., Wild, C. J. Nonlinear Regression. New York, USA, Wiley, 1989.

Shen, P., Li, G., Li, B., et al. Coupling effect of porosity and hydrate saturation on the permeability of methane hydrate-bearing sediments. Fuel, 2020, 269: 117425.

Shi, Z., Zhou, T., Guo, C. Clastic sedimentary records of the Upper Triassic Sichuan Basin, China: Implications for the transition from marine to transitional environment. Geological Journal. 2022, 57: 4393-4411.

Sun, L., Zou, C., Jia, A., et al. Development characteristics and orientation of tight oil and gas in China. Petroleum Exploration and Development, 2019, 46: 1073-1087.

Wang, W., Pang, X., Chen, Z., et al. Improved methods for determining effective sandstone reservoirs and evaluating hydrocarbon enrichment in petroliferous basins. Applied Energy, 2020, 261: 114457.

Wolpert, D. H. Stacked generalization. Neural Networks, 1992, 5: 241-259.

Wood, D. A. Variable interaction empirical relationships and machine learning provide complementary insight to experimental horizontal wellbore cleaning results. Advances in Geo-Energy Research, 2023, 9(3): 172-184.

Yang, Y., Wen, L., Zhou, G., et al. New fields, new types and resource potentials of hydrocarbon exploration in Sichuan Basin. Acta Petrolei Sinica, 2023a, 44: 2045-2069. (in Chinese)

Yang, Z., Shabani, M., Solano, N., et al. Experimental determination of gas-water relative permeability for ultra-low-permeability reservoirs using crushed-rock samples: Implications for drill cuttings characterization. Fuel, 2023b, 347: 128331.

Yu, Q., Xiong, Z., Du, C., et al. Identification of rock pore structures and permeabilities using electron microscopy experiments and deep learning interpretations. Fuel, 2020, 268: 117416.

Zhang, G., Wang, Z., Mohaghegh, S., et al. Pattern visualization and understanding of machine learning models for permeability prediction in tight sandstone reservoirs. Journal of Petroleum Science and Engineering, 2021, 200: 108142.

Zhang, J., Yin, X., Zhang, G., et al. Prediction method of physical parameters based on linearized rock physics inversion. Petroleum Exploration and Development, 2020, 47: 59-67.

Zhang, L., Gao, L., Jing, B., et al. Permeability estimation of shale oil reservoir with laboratory-derived data: A case study of the chang 7 member in Ordos Basin. Applied Geophysics, 2023, 21(3): 440-455.

Zhang, S. Research on key technologies for stroke medical data mining. Zhengzhou, Zhengzhou University, 2022. (in Chinese)

Zhao, C., Chen, B. China's oil security from the supply chain perspective: A review. Applied Energy, 2014, 136: 269-279.

Zhao, X., Chen, X., Huang, Q., et al. Logging-data-driven permeability prediction in low-permeable sandstones based on machine learning with pattern visualization: A case study in Wenchang A Sag, Pearl River Mouth Basin. Journal of Petroleum Science and Engineering, 2022, 214: 110517.

Zou, C., Zhu, R., Liu, K., et al. Tight gas sandstone reservoirs in China: Characteristics and recognition criteria. Journal of Petroleum Science and Engineering, 2012, 88-89: 82-91.