Petroleum Science 23 (2026) 5359–5383

Contents lists available at ScienceDirect

Petroleum Science

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

Original Paper

# A novel abnormal pore pressure prediction framework based on machine learning and drilling–rock mechanics parameter data

Hua-Yang Li [a], Quan-You Liu [a,*], Jia-Ao Chen [b], Yan-Chao Pang [c], Fu-Zhi Chen [d], Dan-Tong Liu [e], Ze-Hui Shi [a]

[a] Institute of Energy, School of Earth and Space Sciences, Peking University, Beijing, 100871, China
[b] School of Pipeline and Civil Engineering, China University of Petroleum (East China), Qingdao, 266580, Shandong, China
[c] Key Laboratory of Ministry of Education of Orogenic Belts and Crustal Evolution, School of Earth and Space Sciences, Peking University, Beijing, 100871, China
[d] Petroleum Engineering School, Southwest Petroleum University, Chengdu, 610500, Sichuan, China
[e] College of Future Technology, Peking University, Beijing, 100871, China

ARTICLE INFO

*Article history:*
Received 29 November 2025
Received in revised form
26 April 2026
Accepted 1 June 2026
Available online 6 June 2026

Edited by Xiu-Fang Hu

*Keywords:*
Xihu sag
Pore pressure
Machine learning
Pressure mechanism
Abnormal high pressure

ABSTRACT

The Xihu Sag in the East China Sea Shelf Basin exhibits complex geology (e.g., deep faulting and fragmented pressure systems) and anomalous overpressure with pressure reversal in target formations, posing severe challenges to conventional pore pressure prediction methods. This study proposes an integrated framework for pore pressure prediction that fuses nine drilling parameters with three rock mechanical parameters. The framework overcomes the common limitation of previous studies relying on a single data source (well log or seismic data). Model development and validation are demonstrated using eight vertical wells across three blocks in the Xihu Sag. Five machine learning (ML) algorithms (Back Propagation, K-Nearest Neighbor, Support Vector Regression, CatBoost, LightGBM) are optimized, with a detailed investigation of dataset partitioning strategies (direct vs. randomized division) to eliminate high-pressure prediction bias. Comprehensive evaluation via six metrics (mean squared error (MSE), mean absolute error (MAE), mean absolute percentage error (MAPE), etc., including training time) demonstrates that the LightGBM model outperforms the others: it achieves near-perfect fitting ($R^2 = 0.999$), minimal prediction errors (MSE $= 0.001$, MAPE $= 0.239\%$), and short training time (0.45 s), with a narrow relative error range ($-1.83\%$ to $+1.61\%$) for both normal and overpressure zones. Practical validation on an adjacent well and four regional wells confirms its robust generalization (average accuracy >97%), with its prediction results consistent with the "normal pressure–overpressure–pressure reversal" distribution pattern of the target block. Importantly, the ML model is successfully applied to two adjacent structural units and, without any parameter adjustment, maintains prediction accuracy above 92%. This work provides a high-precision, field-applicable pore-pressure prediction tool that mitigates drilling risks and offers a reproducible reference framework for pore pressure prediction in other similar complex overpressured basins.


## 1. Introduction

Pore pressure refers to the pressure exerted by fluids (oil, gas, and water) within the rock pores. In petroleum geology, pore pressure is a critical parameter for assessing subsurface conditions. In oil and gas drilling, accurate prediction of pore pressure is essential for safe drilling, minimizing non-productive time, and reducing operational costs.

As a crucial geotechnical parameter closely related to petroleum geology, drilling engineering, and production, the quantitative study of pore pressure began in 1965 and has nearly 60 years of research history. Based on the sources of data, traditional pore pressure calculation methods can be broadly categorized into two types. The first type is the drilling data analysis method, which

* Corresponding author.
*E-mail address:* liuqy@pku.edu.cn (Q.-Y. Liu).
Peer review under the responsibility of China University of Petroleum (Beijing).

https://doi.org/10.1016/j.petsci.2026.06.002
1995-8226/

**Nomenclatures**

| | |
|---|---|
| Depth | Well depth, m |
| Wob | Weight on bit, psi |
| Rop | Rate of penetration, m/h |
| Torque | Torque, kn·m |
| Spp | Slurry pump pressure, MPa |
| Flow | Drilling fluid flow, $m^3/s$ |
| Mwi | Drilling fluid density, $g/cm^3$ |
| Resist | Resistivity, s/m |
| Dc | Dc index, dimensionless |
| Ecd | Equivalent circulating density, $g/cm^3$ |
| UCS | Unconfined compression strength, MPa |
| $E_d$ | Dynamic Young's modulus of rock, GPa |
| $\mu_d$ | Dynamic Poisson's ratio of rock, dimensionless |
| $E$ | Static Young's modulus of rock, GPa |
| $\mu$ | Static Poisson's ratio of rock, dimensionless |
| $P_p$ | Pore pressure, $g/cm^3$ & MPa |
| BP | Back-Propagation Neural Network |
| KNN | K-Nearest Neighbors |
| SVR | Support Vector Regression |
| CatBoost | Categorical Boosting |
| LightGBM | Light Gradient Boosting Machine |
| LL | Liulang Formation |
| YQ | Yuquan Formation |
| LJ | Longjing Formation |
| HG | Huagang Formation |
| HG1 | Upper member of Huagang Formation |
| HG2 | \|Lower member of Huagang Formation |
| PH | Pinghu Formation |
| BS | Baoshi Formation |
| ML | Machine learning |
| MSE | Mean squared error |
| RMSE | Root mean squared error |
| MAE | Mean absolute error |
| MAPE | Mean absolute percentage error |
| $R^2$ | Coefficient of determination |
| SHAP | SHapley Additive exPlanations |

relies on the inductive understanding of pore pressure behavior to develop corresponding calculation formulas for predicting pore pressure using logging data. The second type is the geophysical exploration method, which establishes empirical models linking geophysical parameters to various rock mechanics parameters, predominantly using seismic data for pore pressure prediction (Li et al., 2023a). Predictions made using logging data from the oilfield tend to be accurate and reliable; however, they occur later in the process, classifying them as retrospective predictions. Conversely, the main drawback of using geophysical data, such as seismic information, for pore pressure predictions is the difficulty in accounting for other geological factors and the influence of normal pressure, resulting in lower prediction accuracy.

The pore pressure calculation formulas widely used in academia and industry today include the Eaton method, Bowers method, Fillippone method, and Foster method (also known as the equivalent depth method). However, traditional pore pressure prediction methods have many limitations in predicting different geological structures, pressure mechanisms, and rock compositions. These methods are based on empirical or semi-empirical physical models that rely on certain assumptions, establishing empirical relationships between pore pressure and specific geophysical parameters. This can lead to discrepancies with actual field conditions, limiting their applicability and effectiveness. Moreover, the shift in oil and gas exploration from conventional to complex unconventional resources, moving from shallow and shallow-water areas to deep, ultra-deep, and challenging deep-water regions, introduces intricate pressure mechanisms. Traditional physical models and calculation formulas struggle to adequately describe the complex variations in formation pore pressure, rendering their results less useful for field guidance. Traditional pore pressure calculation formulas often include one or more empirical coefficients, such as the Eaton index ($n$) in the Eaton formula and loading/unloading coefficients in the Bowers formula. The selection of these coefficients heavily depends on preliminary statistical analysis, making them susceptible to human bias. The reliance on numerous empirical parameters can lead to cumulative errors, complicate the calculation process, and limit precision, which ultimately undermines drilling safety. In newly developed areas, a lack of relevant data and available reference methods significantly increases operational costs. In response to this, various improved methods have been proposed by Liu et al. (2023), Lockhart et al. (2023), Wang et al. (2024), Zhao et al. (2024a), Lyu et al. (2026), among others, building upon traditional pore pressure prediction methods.

In recent years, the continuous development and widespread adoption of AI and big data technologies have led scholars to increasingly apply ML and deep learning techniques to address complex challenges in the field of petroleum engineering (Li et al., 2023b, 2025a; Chen et al., 2024; Gao et al., 2026a). These techniques have found extensive applications in various areas, including lithology identification (Khoshouei et al., 2022; Bai et al., 2023), rock mechanics parameter prediction (Tariq et al., 2017; Mahmoodzadeh et al., 2022; Alakbari et al., 2024; Wei et al., 2025), drag and torque prediction (Hegde et al., 2015, 2019), production prediction (Noshi et al., 2019; Song et al., 2020), seismic inversion (Ogbamikhumi and Ebeniro, 2021; Yan et al., 2021), and logging interpretation (Caja et al., 2019). Given their high self-learning, self-adaptive capabilities, and strong generalization ability, ML algorithms are well-suited for addressing complex nonlinear problems such as pore pressure prediction. Intelligent pore pressure prediction methods based on ML and deep learning technologies offer significant advantages over traditional calculation formulas, including broader applicability, higher computational efficiency, and improved accuracy. Recent years have seen substantial progress in using ML methods for predicting pore pressure (Bruna Teixeira Silveira et al., 2023; Mgimba et al., 2023; Deng et al., 2024; Zhang et al., 2025; Ehsan et al., 2025; Liang et al., 2025; Gao et al., 2026b), marking an essential step towards the digital transformation of the global petroleum industry. However, overall, intelligent pore pressure prediction methods are still in development, and the related technologies have not yet matured, presenting several notable drawbacks. For instance, existing studies that use ML for pore pressure prediction primarily rely on well logging data as input variables. The main differences lie in the selection of parameter combinations or different ML algorithms, leading to a narrow focus and limited data sources. Furthermore, most current research focuses on normally pressured areas with relatively simple geological conditions. The model construction and training processes are straightforward, typically using a limited range of ML algorithms and training data mainly derived from logging information. While these models can accurately predict pore pressure in normal pressure areas, their applicability in other regions or under abnormal high-pressure conditions remains uncertain.

This study focuses on the Xihu Sag in the East China Sea Shelf Basin. The Xihu Sag is dominated by tight hydrocarbon reservoirs and exhibits complex internal structures. Extensive deep faulting has fragmented the regional pressure systems, and deep strata display both anomalous overpressure and pressure reversal. Because the mechanisms that generate these pressures are highly complex, conventional pore-pressure prediction methods often fail to achieve high accuracy. Moreover, prior studies have not applied ML techniques to predict pore pressure specifically in such anomalous pressure blocks. To address these gaps, a novel data-driven framework for comprehensive pore pressure prediction in abnormally high-pressure blocks was developed using five ML algorithms, including LightGBM. ML models were constructed, validated, and their practical performance evaluated. The main innovations are threefold. (1) Multi-source data fusion. Contrary to most previous studies that rely solely on well-log or seismic inputs, this work combines nine field drilling parameters with three rock mechanical parameters as model inputs. Drilling parameters record the formation's dynamic response to the drill bit and therefore provide near-real-time, wide-coverage information on formation pressure dynamics. Rock mechanics parameters capture the formation's intrinsic mechanical behavior. Integrating both data types increases field relevance and enables the model to learn pore pressure distribution patterns directly from the interaction between drilling operations and the geological formation. (2) Dataset partitioning methodology. A quantitative comparison between direct depth-based splitting and random splitting was conducted for pore pressure prediction in abnormally overpressured intervals. A random-splitting strategy is recommended for datasets with pronounced pressure-state heterogeneity to reduce evaluation bias. (3) Engineering validation and generalizability. Feature importance analysis, cross-block generalization tests across three study areas, and prediction error analysis using measured pore pressure data from eight wells show that the model reliably captures pore-pressure patterns across the normal pressure zone, the overpressure zone, and the pressure reversal zone. Within-block prediction accuracy exceeds 97%, while cross-block accuracy remains above 92%. These results indicate that the proposed framework is robust, scalable, and sufficiently accurate for pore pressure prediction and drilling-risk mitigation in other analogous complex overpressured basins. The technical roadmap of this study is illustrated in Fig. 1.

## 2. Machine learning algorithm theory

As shown in Table 1, this study compares and applies five types of ML regression algorithms: BP (Back-Propagation Neural Network), KNN (K-Nearest Neighbors), SVR (Support Vector Regression), CatBoost (Categorical Boosting), and LightGBM (Light Gradient Boosting Machine). These algorithms were selected to cover representative model paradigms (neural networks, neighborhood models, kernel methods, and gradient-boosted trees), with the aim of evaluating the performance differences of different models on complex pore pressure profiles characterized by abnormal high pressure and pressure reversal. This table summarizes the core characteristics of each algorithm and its adaptation logic in the context of this study, enabling a quick understanding of the selection rationale for different algorithms and the underlying causes of their performance differences.

## 3. Overview of research block and establishment of model dataset

The completeness, accuracy, and quality of data are crucial for calculating the pore pressure of abnormally pressured zones and for developing ML models. However, the field data provided by oilfields (such as logging data) often contain missing values and significant outliers, making them unsuitable for direct use. Therefore, the primary focus of this paper is to create a high-quality dataset.

### 3.1. Geological setting

Among all the marginal basins in China's coastal waters, the East China Sea Shelf Basin is the largest (Li et al., 2021a), rich in oil and gas resources (Yang et al., 2023) and of significant strategic importance (Quan et al., 2022), attracting considerable attention from domestic and international oil professionals over the years (Zhou et al., 2020). The Xihu Sag, located southeast of Shanghai, is a key area for oil and gas exploration in the East China Sea Shelf Basin. Tectonically, it is a secondary structural unit within the basin and is positioned in its northeastern part (Cui et al., 2019). The Xihu Sag extends in a NNE direction, flanked by the Diaoyu Islands Folded-Uplift Belt to the east and the Haijiao Uplift to the west (Xu et al., 2024). The total area of the Xihu Sag is approximately 51,800 $km^2$ (Zhang et al., 2014). The sag can be divided from west to east into the western slope belt, western sub-sag, central inversion belt, eastern sub-sag, and eastern fault terrace zone (He et al., 2023), as shown in the structural map in Fig. 2 (Zhou et al., 2020; Shen et al., 2022).

The stratigraphic system of the Xihu Sag, from oldest to youngest, primarily consists of the Eocene Bajiaoting, Baoshi, and Pinghu formations, the Oligocene Huagang Formation, the Miocene Longjing, Yuquan, and Liulang formations (Liang et al., 2024), and the Pliocene Santan Formation (Fig. 2(c)). Among these, the Huagang Formation is the most developed reservoir unit (Su et al., 2013), while the Pinghu Formation serves as the main source rock within the Xihu Sag (Kang et al., 2020; Liu et al., 2020; Su et al., 2020).

The Xihu Sag not only contains conventional oil and gas reservoirs but also develops low-porosity and low-permeability reservoirs. These two types of reservoirs differ significantly in accumulation characteristics and controlling factors. The primary target formations, the Pinghu and Huagang formations, contain both low-porosity, low-permeability reservoirs and medium-to-high permeability reservoirs. Conventional pore pressure calculation methods have become inadequate for ensuring safe drilling in this area. Additionally, the area has undergone multiple periods of inversion structure and rifting (Zhu et al., 2019), resulting in a complex structural style and geological architecture in the Xihu Sag. The development of deep faults has segmented reservoirs into multiple isolated blocks, severely impacting the migration and preservation of oil and gas (McClay et al., 2002; Lin et al., 2024). Following faulting, the isolated blocks are influenced by local structural effects, leading to unpredictable pressure points and uneven lateral distribution of pore pressure. Additionally, the deep formations (specifically the Huagang and Pinghu formations) generally exhibit abnormally high pressures (Fan and Wang, 2021; Li et al., 2021a). Statistically, the highest pore pressure coefficient in drilled evaluation and exploration wells in the Xihu Sag has reached 1.8, with nearly 50% of drilled wells encountering abnormal high pressure (Yang et al., 2013). The pressure generation mechanism is complex, making traditional pore pressure prediction methods less applicable in the Xihu Sag, and accurate prediction and calibration of pore pressure is challenging. This has led to numerous challenges and complications during the early stages of drilling projects, such as frequent borehole collapses, pipe sticking, gas invasions, and lost-circulation issues. These issues severely impact drilling safety and efficiency, increasing the duration and cost of drilling and hindering oil and gas

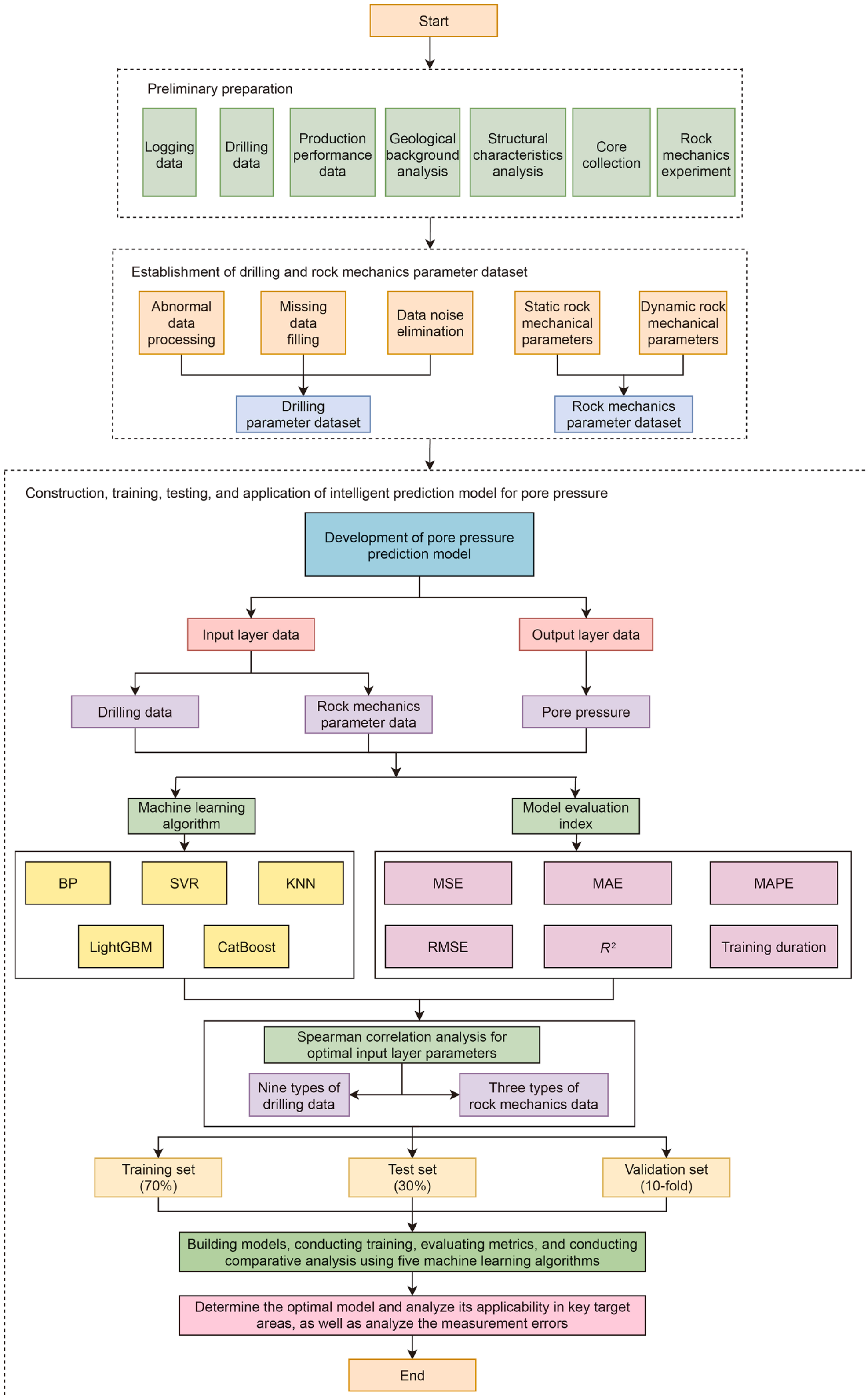


**Fig. 1.** Technology roadmap.

development in the region. Therefore, there is an urgent need for accurate pore pressure predictions to prevent complex incidents.

This study focuses on two vertical wells in the western slope belt of the Xihu Sag. Well S-1 serves as the training well, providing sample data for model training, while the nearby S-2 well is used to validate the model's predictive performance. Table 2 presents the basic information for these two wells.

**Table 1**
Overview of employed ML algorithms.

| Algorithm | Paradigm (Category) | Key characteristics | Role in this study/Rationale for inclusion | Main strengths/ Weaknesses | References |
|---|---|---|---|---|---|
| BP | Neural network | General-purpose nonlinear function approximator using layered neurons and backpropagation training | Serves as a baseline deep-learning model to evaluate a black-box approach for complex nonlinear relationships | Strengths: flexible representation capacity. Weaknesses: requires careful hyperparameter tuning, longer training time, limited interpretability. | Rumelhart et al. (1986) |
| KNN | Instance-based/ local regression | Lazy learner performing prediction by averaging outcomes of nearest training instances in feature space | Tests local similarity-based prediction behavior and robustness to spatial clustering of pressure states | Strengths: simple, intuitive. Weaknesses: sensitive to noise, poor extrapolation ability—performed poorly for deep/high-pressure extrapolation in this study. | Cover and Hart (1967); Chen et al. (2021); Guo and Wang (2024) |
| SVR | Kernel method | Kernel-based regression that fits a margin-tolerant function, suitable for small-sample settings | Provides a theoretically grounded small-sample baseline to assess stability under limited labeled pressure points | Strengths: robust theoretical properties. Weaknesses: sensitive to kernel and hyperparameters; shows high variability for abnormal-pressure segments. | Vapnik and Lerner (1963); Cortes and Vapnik (1995); Salcedo-Sanz et al. (2016) |
| CatBoost | Gradient boosting decision trees (GBDT, CatBoost implementation) | GBDT variant optimized for categorical features and ordered boosting to reduce overfitting | Represents modern boosting-tree approaches for interpretable tree-based modeling of pressure drivers | Strengths: relatively interpretable feature importances. Weaknesses: longer training time in our experiments; under-prediction of deep overpressure in some cases. | Prokhorenkova et al. (2017) |
| LightGBM | Gradient boosting decision trees (GBDT, LightGBM implementation) | Efficient GBDT implementation with GOSS, EFB, and leaf-wise growth for fast training and high accuracy | Final recommended model: balances prediction accuracy and computational efficiency for integrated drilling + rock-mechanics inputs | Strengths: fast training, low memory use, balanced feature importance, stable fit to abnormal-pressure points. Weaknesses: requires hyperparameter tuning for optimal performance. | Ke et al. (2017); Korstanje (2021); Wen et al. (2021) |

The construction and preprocessing of the training dataset are crucial for the predictive accuracy of ML models. In this study, the training dataset (i.e., the learning samples) is divided into input and output data. The input data for the model consist of drilling data and rock mechanics parameters from the two wells, while the output data represents the equivalent density of formation pore pressure.

### 3.2. Establishment of drilling parameter dataset

Drilling parameters are monitored and recorded by drilling parameter instruments during the drilling process. This study collects ten types of drilling parameters based on actual field data provided by the oilfield: depth (Depth), weight on bit (Wob), rate of penetration (Rop), torque (Torque), slurry pump pressure (Spp), flow rate (Flow), density (Mwi), resistivity (Resist), Dc index (Dc), and equivalent circulating density (Ecd). The sampling frequency for these data is one data point per meter of depth. For example, in Well S-1, drilling parameters were recorded from a depth of 134 m to a depth of 3840 m, resulting in a total of 3707 data entries.

The quality of data directly affects the stability of the model structure and the magnitude of prediction errors during the construction of ML models. However, the raw drilling data contain certain outliers due to the influence of the equipment's working environment, making them unsuitable for direct use. Therefore, it is necessary to preprocess the raw drilling data.

The preprocessing of raw drilling data primarily involves detecting and removing outliers, as well as addressing missing values. As illustrated in Fig. 3, using the drilling data from Well S-1, specifically weight on bit, rate of penetration, torque, and flow rate, the red circles highlight visually obvious outliers; however, there are additional anomalies that are not easily identifiable. These outliers can be addressed using the 3σ method or other techniques, such as box plots, each with its own advantages. The advantage of the box plot method lies in its ability to represent data visually, providing a more intuitive understanding while maintaining high accuracy in detection. Therefore, this study employs the box plot method to identify and address outliers in the raw drilling data.

A box plot consists of six components: extreme outliers, maximum, upper quartile (also known as the third quartile), median (or second quartile), lower quartile (also known as the first quartile), and minimum, as shown in Fig. 4. The box plot provides a clear visual representation of outliers in the raw drilling parameter data, allowing for the removal of extreme values while retaining normal data points. To illustrate the effectiveness of the preprocessing procedure, Fig. 5 presents a side-by-side comparison of the original and preprocessed drilling parameters. As shown in Fig. 5(a), the raw data contain a number of outliers that deviate significantly from the general trend, which are likely attributable to equipment operating conditions or measurement anomalies. After applying the box plot-based outlier detection and removal method, the preprocessed data (Fig. 5(b)) exhibit smooth, continuous curves that retain the primary variation patterns while

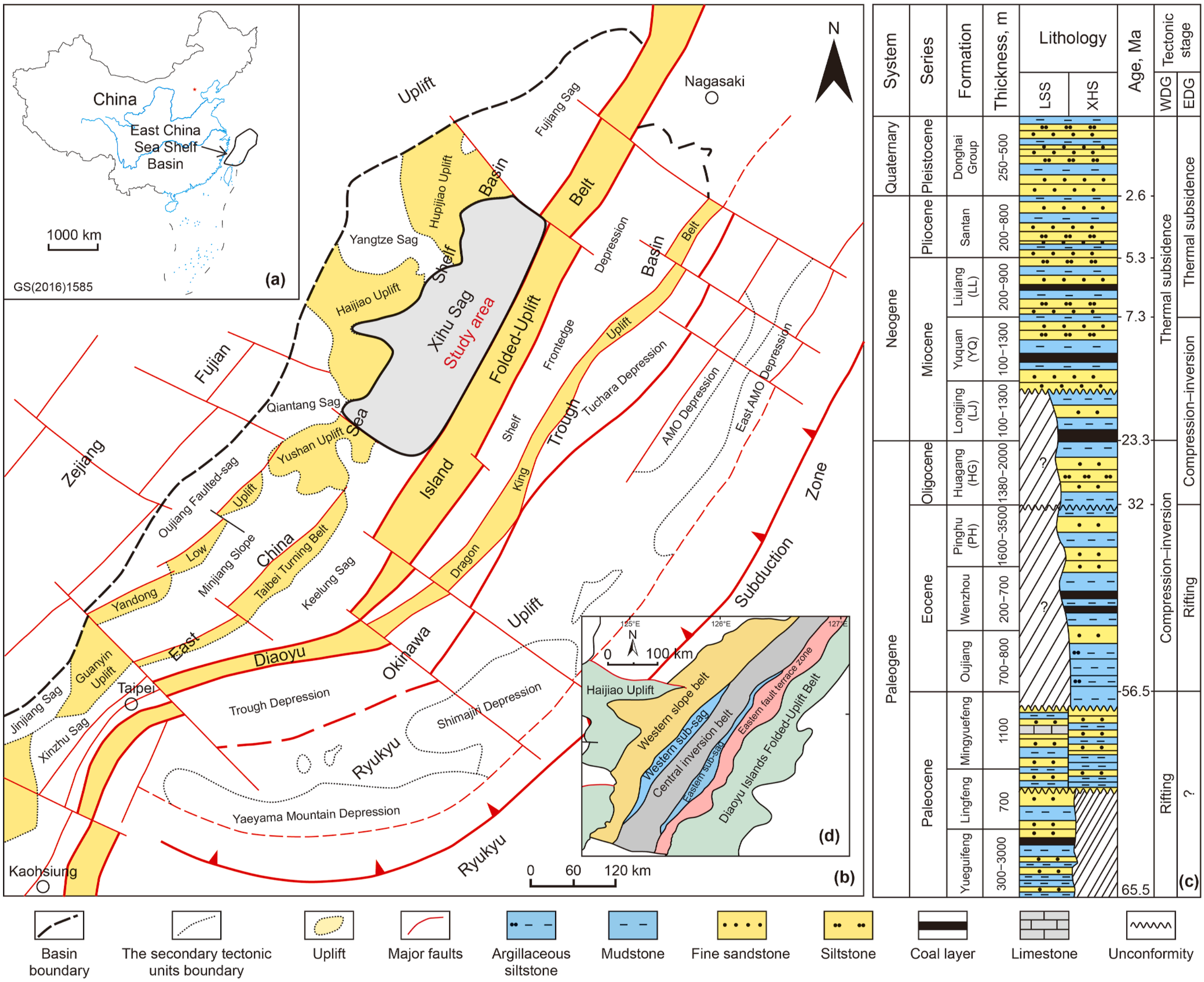


**Fig. 2.** **(a)** Location of the East China Sea Shelf Basin (modified after Li et al. (2025a; 2025b; 2025c)), **(b)** the tectonic framework of the East China Sea Shelf Basin (modified after Yang et al. (2020)), **(c)** Cenozoic stratigraphic column of the East China Sea Shelf Basin (modified after Zhang et al. (2016), Liang and Wang (2019), Li et al. (2021b), and Liu et al. (2022)), **(d)** the tectonic framework of the Xihu Sag (modified after Li et al. (2025a; 2025b; 2025c)). LSS: Lishui Sag; XHS: Xihu Sag; WDG: Western Depression Group; EDG: Eastern Depression Group.

**Table 2**
Basic information table for wells S-1 and S-2 in the western slope belt of Xihu Sag.

| Location | Well name | Well pattern | Well depth, m | Completion interval |
|---|---|---|---|---|
| Western slope belt | S-1 | Vertical well | 4670 | Lower member of Pinghu Formation |
| | S-2 | Vertical well | 4415 | Lower member of Pinghu Formation |

eliminating extreme values. Importantly, the preprocessing does not introduce artificial distortions or alter the overall depth-dependent trends, ensuring that the cleaned dataset faithfully represents the formation response during drilling.

### 3.3. Establishment of rock mechanics parameter dataset

Rock mechanics parameters are intrinsic properties of formation rocks and serve as fundamental data for drilling and fracturing operations, as well as important parameters for oil and gas exploration and development (Bowers, 1995). Commonly used rock mechanics parameters include unconfined compressive strength, Poisson's ratio, internal friction angle, and cohesion. These parameters can be obtained through two main methods: conducting laboratory rock mechanics experiments or calculating them from logging data (Ward et al., 1994). Parameters obtained from laboratory experiments are referred to as static rock mechanics parameters, while parameters calculated from logging data are known as dynamic rock mechanics parameters. Typically, a conversion between the two is required to evaluate the formation.

#### 3.3.1. Basic information of core

This study collected a core sample from Well S-1 at a depth of 4185.25 m (belonging to the Pinghu Formation), which is a primary

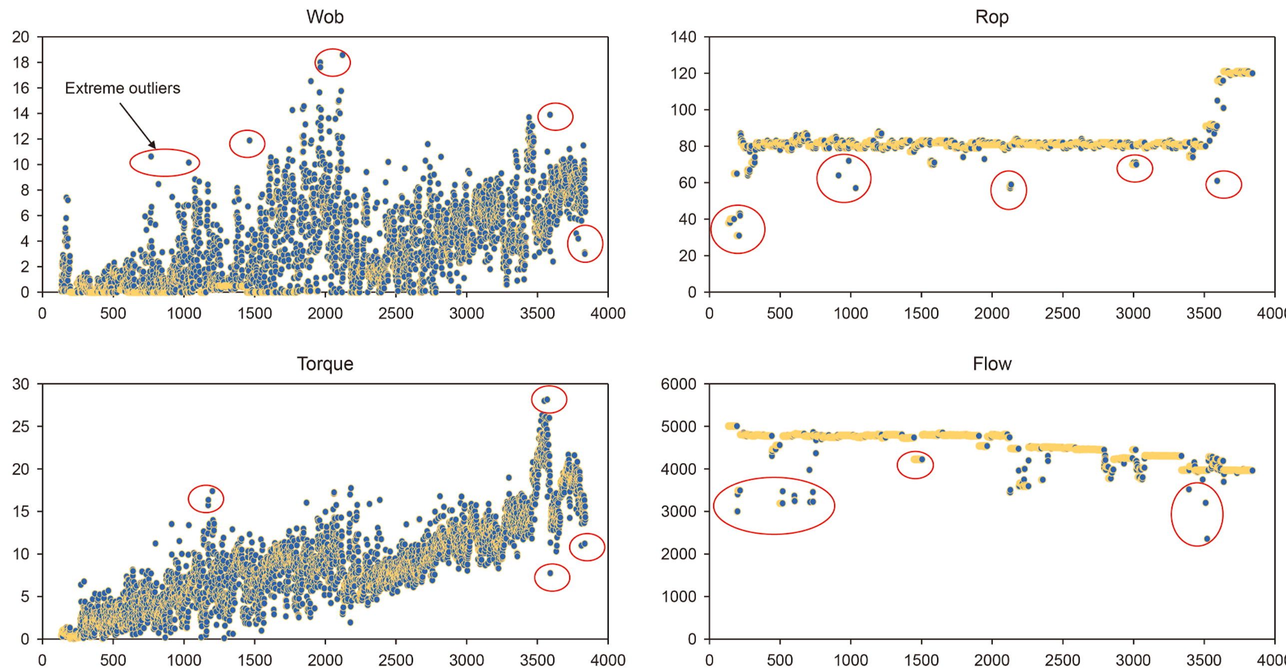


**Fig. 3.** Partial original drilling data of Well S-1.

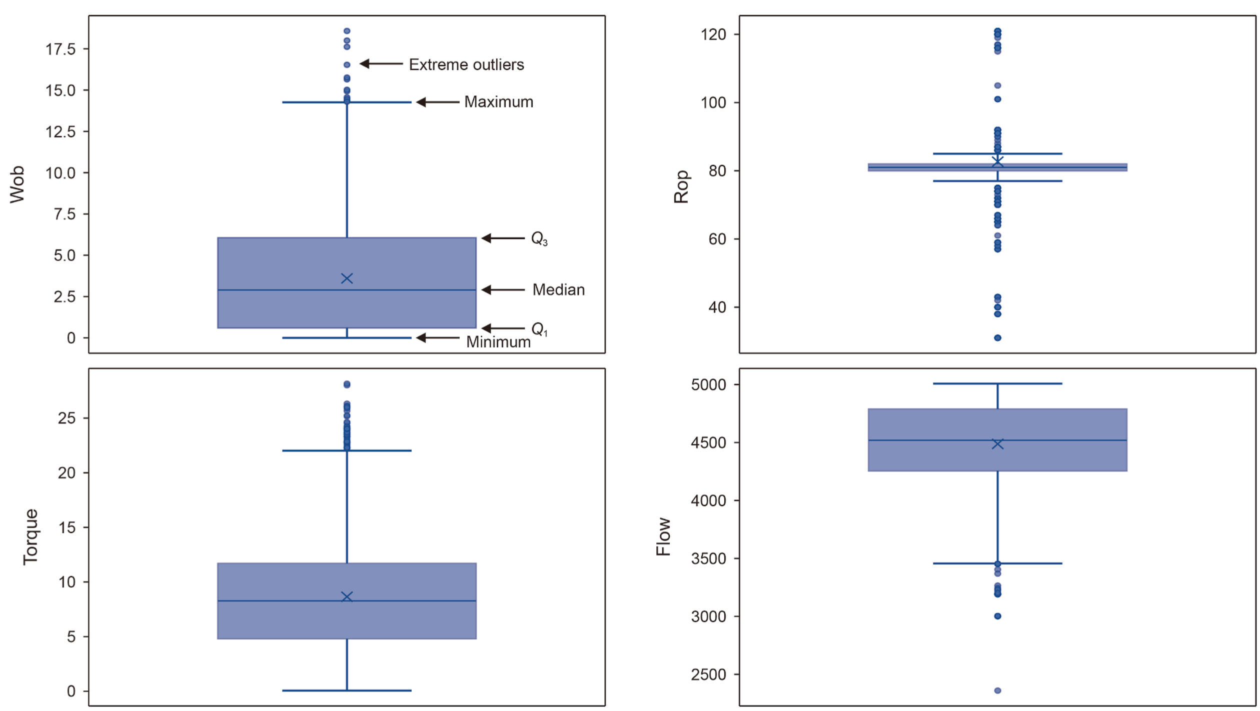


**Fig. 4.** Box plot of partial original drilling data for Well S-1.

target for development in the Xihu Sag; therefore, the core sample was selected to ensure representativeness. The basic information of the core sample is presented in Table 3, and an image of the core is shown in Fig. 6.

### 3.3.2. Rock mechanics experiment

The rock mechanics parameters to be measured in this experiment include three key values: unconfined compressive strength (UCS, MPa), Young's modulus ($E$, GPa), and Poisson's ratio ($\mu$, dimensionless). The experimental equipment used in this study is the TAW-1000 Deepwater Pore Pressure Servo Test System, as shown in Fig. 7. The software associated with the equipment was used to plot the stress–strain curve for the rock sample, with the experimental results presented in Fig. 8. The calculation formulas for Young's modulus and Poisson's ratio are given by Eqs. (1) and (2), respectively. The final calculated rock mechanics parameters for the S-1-1 sample are shown in Table 4.

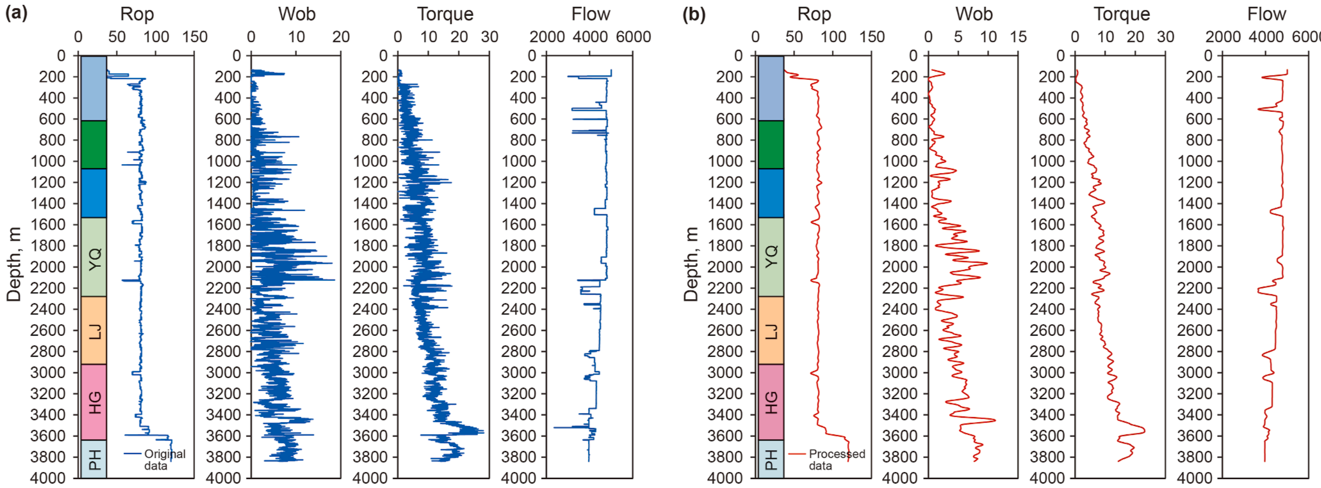


**Fig. 5.** Comparison of original and preprocessed drilling parameters for Well S-1.

**Table 3**
Basic information of the rock core S-1-1.

| Sample number | Depth, m | Lithology | Length, mm | Diameter, mm | Mass, g |
|---|---|---|---|---|---|
| S-1-1 | 4185.25 | Silty mudstone | 45.21 | 26.02 | 64.9 |


**Fig. 6.** Core image of the sample S-1-1 from the Well S-1.

**Fig. 7.** Device diagram of the TAW-1000 Deepwater Pore Pressure Servo Test System.

$$E = \frac{\Delta \sigma_a}{\Delta \varepsilon_a} \tag{1}$$

$$\mu = \frac{\Delta \varepsilon_r}{\Delta \varepsilon_a} \tag{2}$$

where $E$ is the static Young's modulus, GPa; $\mu$ is the static Poisson's ratio, dimensionless; $\Delta\varepsilon_a$ is the axial strain increment; $\Delta\sigma_a$ and $\Delta\varepsilon_r$ are the axial stress increment and radial strain increment, respectively.

Due to the insufficient amount of measured rock mechanics data, which does not meet the minimum requirements for model training, the primary task is to obtain continuous and adequate rock mechanics parameter data. The drawback of obtaining rock mechanics data through laboratory core tests is that it is complex and expensive. For highly fractured brittle rocks, it can be challenging to obtain core samples, and the resulting data are discrete rather than continuous. Consequently, calculating rock mechanics parameters using logging data has become widely adopted. Calculating dynamic rock mechanics parameters from logging data and calibrating them with measured data has the advantage of relatively easy acquisition of logging data, which effectively provides a continuous representation of the formation, ensuring data

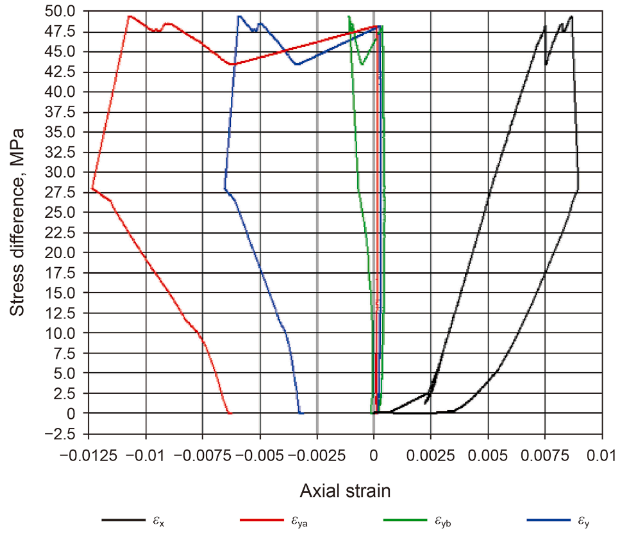


**Fig. 8.** Stress–strain curve of sample S-1-1. Note: $\varepsilon$ = strain; subscripts x/y refer to axial/circumferential directions; a/b denote two perpendicular circumferential extensometer measurement points.

**Table 4**
Rock mechanical parameters of the core sample S-1-1.

| Sample number | UCS, MPa | Young's modulus, GPa | Poisson's ratio |
|---|---|---|---|
| S-1-1 | 49.45 | 13.044 | 0.203 |

accuracy, continuity, and completeness. The calculation formulas are presented in Eqs. (3)–(5) below.

The formula for calculating the dynamic Young's modulus of rock is:

$$E_d = \frac{\rho {V_s}^2 \left(3{V_p}^2 - 4{V_s}^2\right)}{{V_p}^2 - {V_s}^2} \tag{3}$$

The formula for calculating the dynamic Poisson's ratio of rock is:

$$\mu_d = \frac{0.5{V_p}^2 - {V_s}^2}{{V_p}^2 - {V_s}^2} \tag{4}$$

The formula for calculating the dynamic unconfined compressive strength of rock is:

$$\mathrm{UCS} = A(1-\mu_d)\left(\frac{1-2\mu_d}{1+\mu_d}\right)^2 \rho^2 V_p^4 (1+0.87M_d) \tag{5}$$

where $E_d$ is the dynamic Young's modulus of the rock, GPa; $\mu_d$ is the dynamic Poisson's ratio of the rock, dimensionless; UCS is the dynamic unconfined compressive strength of the rock, MPa; $\rho$ is the density of the rock, g/cm$^3$; $V_p$ and $V_s$ are the rock's compressional and shear wave velocities, respectively, m/s; $M_d$ is the shale content, dimensionless; $A$ is a coefficient.

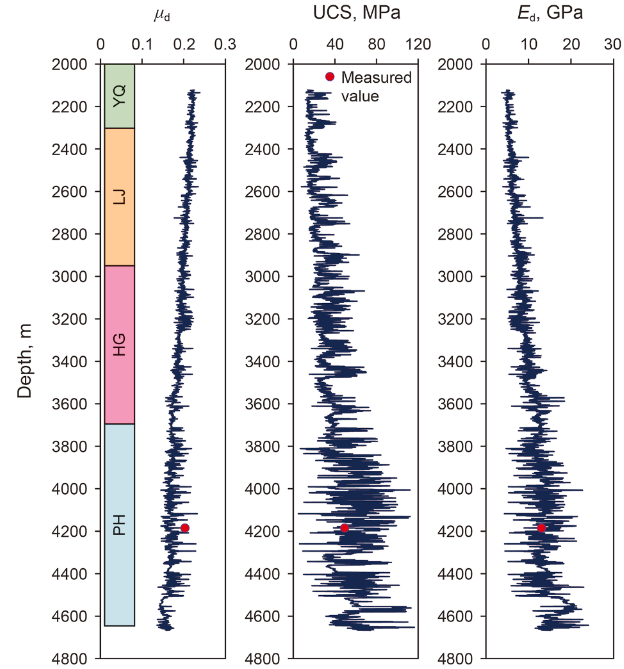


**Fig. 9.** Rock mechanics parameter profile of Well S-1.

By integrating the background of the block and expert experience, we calculated the uniaxial compressive strength, Young's modulus, and Poisson's ratio for Well S-1 using Eqs. (3)–(5), and calibrated the rock mechanics parameter profile with measured values to establish the actual values. The results indicate that the prediction errors for the UCS, Young's modulus, and Poisson's ratio of Well S-1 were 1.84%, −4.615%, and −3.448%, respectively, with absolute prediction errors not exceeding 5%. Ultimately, a sample dataset of rock mechanics parameters suitable for model training was established, as detailed in Table 5 and Fig. 9. It should be noted that due to the high cost of coring in offshore drilling, only a limited number of core samples were obtained in this study. However, these cores were collected from the Pinghu Formation, the main hydrocarbon-bearing interval, and are therefore highly representative. In future research, if conditions permit, acquiring more core samples from different stratigraphic horizons for multi-point calibration of measured data will help further improve the accuracy of input variable datasets.

### 3.4. Establishment of pore pressure dataset

The output layer of the model represents the desired variable, which in this paper is the formation pore pressure. Because the number of measured pore pressure data points is insufficient to support model development and training, most related studies use pore pressure calculated from formulas as the raw data, which is then adjusted using expert experience in conjunction with

**Table 5**
Analysis table of prediction error of rock mechanics parameters in Well S-1.

| Sample number | Type | UCS, MPa | Young's modulus, GPa | Poisson's ratio |
|---|---|---|---|---|
| S-1-1 | Actual value | 49.45 | 13.044 | 0.203 |
| | Predicted value | 50.36 | 12.442 | 0.196 |
| | Absolute error | 0.91 | −0.602 | −0.007 |
| | Relative error, % | 1.84 | −4.615 | −3.448 |

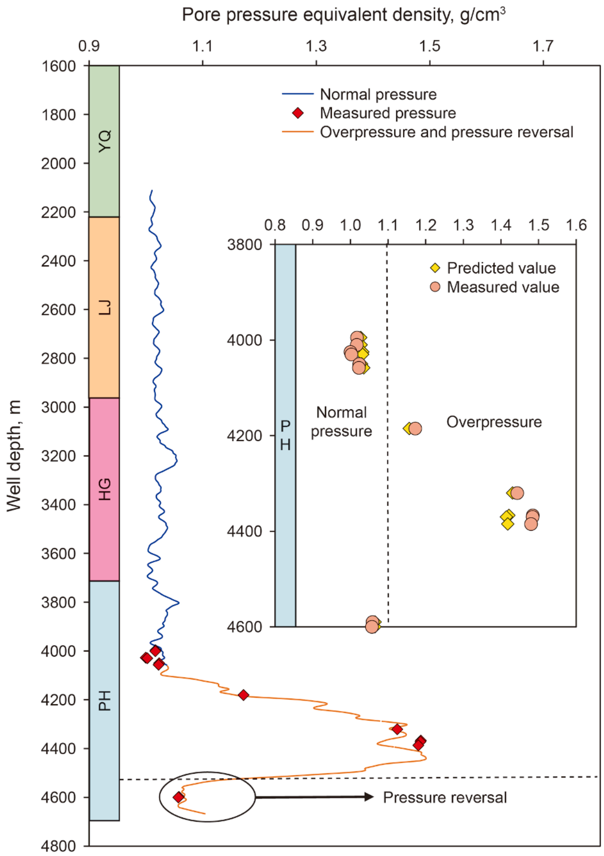


Fig. 10. Pore pressure profile of Well S-1.

measured pore pressure data. This paper utilizes this approach to construct the dataset for the model's output layer variable.

This study employs a combined method of the Eaton (1972, 1975) and Bowers (1995, 2001) techniques (see Eqs. (6)–(9)) to determine the formation pore pressure for Well S-1, which is then calibrated with the measured pore pressure data under expert guidance. The final results are presented in Fig. 10. Based on the pore pressure state classification table proposed by Fan (2016; see Table 6), formations with a pore pressure coefficient greater than 1.1 are classified as high-pressure formations.

(1) Eaton method:

$$P_{\mathrm{p}H} = P_{\mathrm{v}H} - (P_{\mathrm{v}H} - P_H) \times (\Delta t_z / \Delta t)^n \tag{6}$$

where $P_{\mathrm{p}H}$ is the pore pressure at depth $H$, MPa; $P_{\mathrm{v}H}$ is the overburden pressure at depth $H$, MPa; $P_H$ is the hydrostatic column pressure at depth $H$, MPa; $\Delta t_z$ is the acoustic time for normally compacted mudstone, μs/ft; $\Delta t$ is the measured acoustic time for mudstone at a certain depth, μs/ft; $n$ is the Eaton index, derived from measured formation pressure in the target block.

(2) Bowers method:

When $H_{\max} > H$:

$$P_{\mathrm{p}} = G - \frac{\left(\frac{\frac{10^6}{\Delta t} - \frac{10^6}{\Delta t_{\max}}}{a}\right)^{\frac{1}{b}}}{H} \tag{7}$$

When $H_{\max} \le H$:

$$P_{\mathrm{p}} = G - \frac{\sigma_{\max}^{(1-U)} \left(\frac{\frac{10^6}{\Delta t} - \frac{10^6}{\Delta t_{\max}}}{a}\right)^{\frac{U}{b}}}{H} \tag{8}$$

where,

$$\sigma_{\max} = \left(\frac{\frac{10^6}{\Delta t} - \frac{10^6}{\Delta t_{\max}}}{a}\right)^{\frac{1}{b}} \tag{9}$$

where $V_{\max}$ is the maximum acoustic velocity during unloading, m/s; $H_{\max}$ is the depth at which $V_{\max}$ is measured, m; $H$ is the total vertical depth, m; $P_{\mathrm{p}}$ is the pore pressure, MPa; $G$ is the overburden pressure, MPa; $\Delta t$ is the acoustic time, s/m; $\Delta t_{\max}$ is the acoustic time corresponding to $V_{\max}$, s/m; $\sigma_{\max}$ is the maximum effective stress in the vertical direction at the onset of unloading, MPa; $a$ and $b$ are parameters obtained through data fitting, dimensionless; $U$ is the unloading parameter, which measures the plastic deformation of the rock, dimensionless.

The study found that Well S-1 is situated in normal pressure formations in the Yuquan, Longjing, Huagang, and upper member of Pinghu formations, with abnormal high pressure starting in the middle member of the Pinghu Formation. The increase in pore pressure from the middle to the lower members of the Pinghu Formation shows a distinct stepped trend. It was also observed that the pore pressure in the lower member of the Pinghu Formation experiences a significant drop, exhibiting pressure reversal characteristics; therefore, it is referred to as the pressure reversal segment.

Pore pressures in undercompacted formations and formations governed by unloading mechanisms were calculated using the Eaton and Bowers methods, respectively, to establish the training label dataset for the model. Empirical parameters for both methods were determined by integrating the geological characteristics and measured data of the Xihu Sag, and the selected values are consistent with regional engineering practices. Specifically, the Eaton exponent $n$ is set to 3.5, and the unloading-related parameters in the Bowers method ($a$, $b$, $U$, and $H_{\max}$) are 1.06, 1.01, 2.6, and 4100 m, respectively. It should be noted that the Eaton exponent ($n$) and Bowers coefficients ($a$, $b$, $U$) used in this study were derived from drilling data of the western slope belt in the Xihu Sag and are suitable for the geological conditions of this area. If this framework is applied to other blocks, recalibration of these empirical parameters is recommended to match the local compaction trends and pressure-generation characteristics. The prediction accuracy of pore pressure for Well S-1 exceeds 95%. This specific trend in the predicted pore pressure aligns well with the measured values and corresponds to the field understanding of the block, allowing us to consider the calculated pore pressure as the true formation pore pressure for this well.

**Table 6**
Classification table of pore pressure status.

| Type | Ultra-low pressure | Underpressure | Normal pressure | Overpressure | Ultra-high pressure |
|---|---|---|---|---|---|
| Pore pressure coefficient | <0.75 | 0.75–0.9 | 0.9–1.1 | 1.1–1.5 | >1.5 |

The reliability of training data constitutes the fundamental prerequisite for both model accuracy and research credibility. It should be noted that while using measured data to construct a learning sample dataset is the optimal approach, the high cost of acquiring such data often leads to extreme scarcity, forcing researchers to adopt alternative dataset-building methods. This common and practical limitation is prevalent in research across numerous fields (He et al., 2021; Li et al., 2022; Zhang et al., 2025). In this study, the number of directly measured pore pressure points obtained via Drill Stem Testing (DST) is extremely limited and discontinuously distributed. These points are entirely insufficient to support the training of supervised ML models, which require continuous data labels. This is the fundamental reason why this study aims to construct a continuous "ground-truth" pore pressure profile. Compared with the direct use of scarce measured data, which is in fact quantitatively inadequate, the method employed in this study integrates logging data calculation, measured data verification, and expert experience adjustment to generate continuous, high-quality labels. This approach ensures the reliability of the learning sample data and significantly enhances the generalization ability of the ML model.

## 4. Establishment of intelligent prediction model for pore pressure

Abnormal high pressure significantly influences the generation, migration, and accumulation of oil and gas resources (Hunt, 1990). The Xihu Sag, a region rich in oil and gas resources within the East China Sea Shelf Basin (Zhao et al., 2024b), holds substantial exploration and development potential. However, the area has undergone a series of tectonic movements characterized by initial stretching followed by compression, resulting in complex internal structural styles. These variations in pore pressure distribution between different structures have led to abnormal high pressure in the vertical profile of the Xihu Sag, and horizontal variations in pore pressure initiation points are inconsistent, posing significant challenges for accurately calculating pore pressure in key target areas of the Xihu Sag.

In traditional methods for predicting pore pressure, two key factors are essential for accuracy: selecting an appropriate calculation model and carefully choosing the parameters used in the formulas. Typically, the process of predicting pore pressure in exploration wells involves using parameters derived from nearby drilled wells, which assumes that the pressure distribution in the target block is relatively uniform, allowing for similarities between different wells within the same block. This ensures a significant degree of reference in the pore pressure calculations. However, this approach fails in complex geological structures like the Xihu Sag, where there are considerable variations in the vertical and horizontal distribution of pore pressure. Therefore, this study constructs an intelligent pore pressure prediction model based on drilling and rock mechanics data, using two vertical wells from the western slope zone of the Xihu Sag as a case study, and evaluates its application.

### *4.1. Input variable selection*

Building on the previous discussion, this study ultimately selects ten drilling parameters and three rock mechanics parameters as preliminary candidate input variables for the ML model. However, not all of these parameters are suitable for model training, as using parameters unrelated to pore pressure can reduce model performance. Therefore, it is necessary to filter out redundant parameters. Given that these parameters typically exhibit a non-normal distribution, this study employs Spearman correlation analysis to quantitatively assess the relationship between different parameters and formation pore pressure. The formula for calculating the correlation coefficient is shown in Eq. (10) (Ma et al., 2022).

$$R_z = 1 - \frac{6\sum_{i=1}^{N} C_i^2}{N\left(N^2-1\right)} \tag{10}$$

where $R_z$ is the correlation coefficient, ranging from $-1$ to $+1$; $N$ is the observed number of data points; $i$ represents the serial number of each individual sample, from 1 to $N$; $C_i$ stands for the rank difference of the $i$-th paired data.

Fig. 11 displays a heatmap of the correlation coefficients between the input variables and formation pore pressure. In the figure, darker colors and higher absolute values of the correlation coefficients indicate stronger associations, while lighter colors suggest weaker associations. Positive coefficients indicate a positive correlation, while negative coefficients indicate a negative correlation. Additionally, Table 7 outlines the criteria for assessing the strength of these correlations (Zhao et al., 2022). From Fig. 11, it is evident that Depth, Wob, Rop, Torque, Spp, Mwi, Resist, Ecd, UCS, and $E_d$ show a positive correlation with pore pressure. In contrast, Flow and $\mu_d$ demonstrate a negative correlation with pore pressure. Considering the absolute values of the correlation coefficients, Depth (0.72), Ecd (0.66), and Mwi (0.64) exhibit the strongest correlations with pore pressure, while the Dc (0.094) has the weakest correlation with pore pressure. Following the criteria outlined in Table 7 and considering the weak correlation coefficient (0.094) as well as the fact that the information contained in the Dc index is already covered by its constituent parameters (Rop, Wob, etc.) in the input feature set, the irrelevant Dc index was excluded from the final model inputs to avoid potential redundancy, while the remaining parameters were retained. This decision aims to avoid noise introduced by redundant features during model training, while enabling the model to learn more complex pressure response patterns through nonlinear interactions among the original parameters.

### *4.2. Preliminary establishment of machine learning model structure*

#### *4.2.1. Model input variables and structural design*

The input layer data for the ML model includes nine drilling parameters: Depth, Wob, Rop, Torque, Spp, Flow, Mwi, Resist, and Ecd. Additionally, it includes three rock mechanics parameters: UCS, $E_d$, and $\mu_d$. The output layer data of the ML model represents the formation pore pressure, and the bridge for analyzing the complex nonlinear relationship between the two is formed by five ML algorithms. The process of model development is illustrated in Fig. 12.

This study samples data at 1-m intervals in depth, with the research intervals and data volume for the two wells presented in Table 8. The prediction of pore pressure in this study is categorized as a regression problem, requiring the establishment of appropriate evaluation metrics to assess the model's generalization capability. Different evaluation metrics typically yield varying assessment results, Therefore, this study employs multiple metrics to evaluate the model's generalization performance. The evaluation metrics include the mean squared error (MSE), root mean

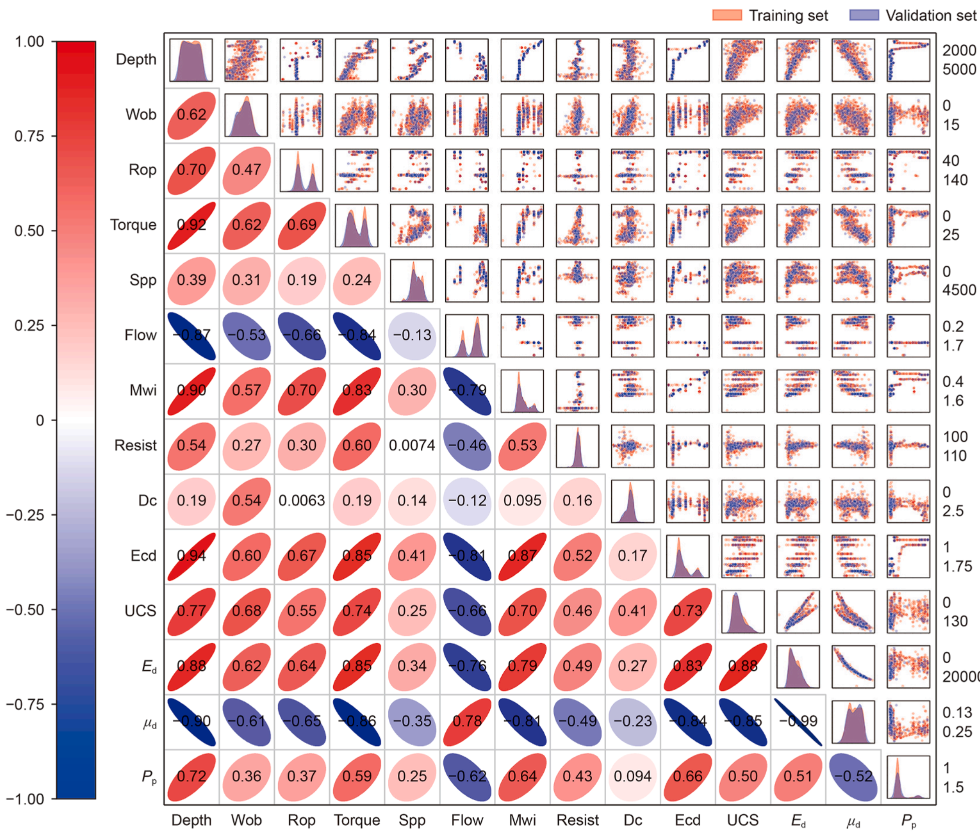


**Fig. 11.** Correlation analysis between drilling parameters, rock mechanics parameters, and pore pressure.

**Table 7**
Criteria for determining the strength of correlation between input-layer and output-layer variable data (Zhao et al., 2022).

| No. | Correlation coefficient value | Correlation strength |
|---|---|---|
| 1 | $0.8 \le \lvert R_z \rvert \le 1.00$ | Very strong |
| 2 | $0.6 \le \lvert R_z \rvert \le 0.79$ | Strong |
| 3 | $0.4 \le \lvert R_z \rvert \le 0.59$ | Moderate |
| 4 | $0.2 \le \lvert R_z \rvert \le 0.39$ | Weak |
| 5 | $0.0 \le \lvert R_z \rvert \le 0.19$ | Very weak |

square error (RMSE), coefficient of determination ($R^2$), mean absolute error (MAE), mean absolute percentage error (MAPE), and training duration. Smaller values of MSE, MAE, RMSE, and MAPE, along with a larger $R^2$, indicate higher model accuracy. The corresponding calculation formulas can be found in Eqs. (11)–(15).

$$\mathrm{MSE} = \frac{1}{S}\sum_{i=1}^{S}\left(x_i - x_i^{\mathrm{pre}}\right)^2 \tag{11}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{S}\sum_{i=1}^{S}\left(x_i - x_i^{\mathrm{pre}}\right)^2} \tag{12}$$

$$\mathrm{MAE} = \frac{1}{S}\sum_{i=1}^{S}\left|x_i - x_i^{\mathrm{pre}}\right| \tag{13}$$

$$\mathrm{MAPE} = \frac{1}{S}\sum_{i=1}^{S}\left|\frac{x_i - x_i^{\mathrm{pre}}}{x_i}\right| \times 100\% \tag{14}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{S}\left(x_i - \overline{x_i}\right)\left(x_i^{\mathrm{pre}} - \overline{x_i^{\mathrm{pre}}}\right)}{\sqrt{\sum_{i=1}^{S}\left(x_i - \overline{x_i}\right)^2\left(x_i^{\mathrm{pre}} - \overline{x_i^{\mathrm{pre}}}\right)^2}} \tag{15}$$

where $S$ is the sample size; $x_i$ is the actual formation pore pressure equivalent density, g/cm$^3$; $x_i^{\mathrm{pre}}$ is the predicted pore pressure equivalent density, g/cm$^3$; $\overline{x_i}$ is the average of the actual pore pressure equivalent densities, g/cm$^3$; $\overline{x_i^{\mathrm{pre}}}$ is the average of the predicted pore pressure equivalent densities, g/cm$^3$.

Given that there are 12 different types of input variables and significant dimensional differences among them, it is necessary to eliminate the effects of these dimensional differences during the model training process, to ensure that all data values remain within the same range for easier comprehensive analysis. This study employs the max-min normalization method, with the formula presented in Eq. (16). Table 9 shows the results of the normalization for certain input layer parameters from Well S-1.

$$x' = \frac{x - \min}{\max - \min} \tag{16}$$

where $x$ is the original value of a data point before normalization; $x'$ is the value of a data point after normalization; min is the minimum value of the sample containing that data point; max is the maximum value of the sample containing that data point.

The BP neural network model is considered one of the most classic and commonly used artificial intelligence models. In this study, the BP model has 12 nodes in the input layer and 1 node in

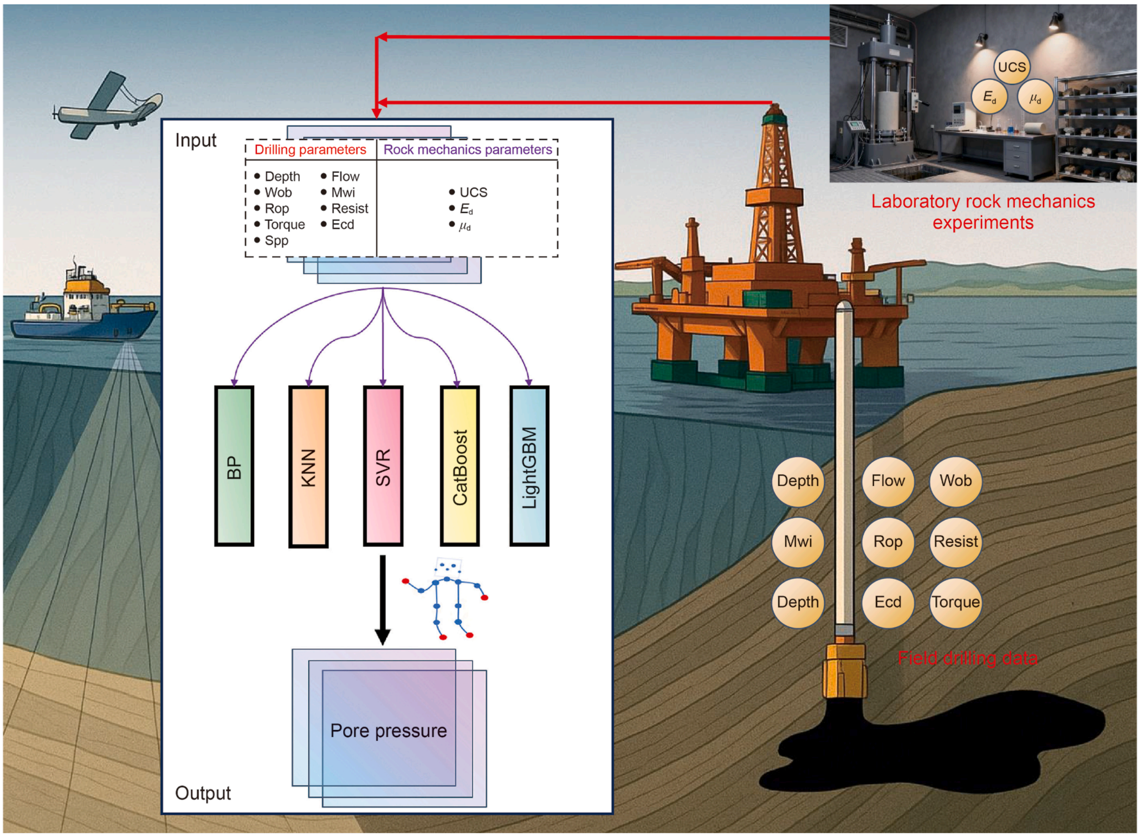


**Fig. 12.** Schematic diagram of model construction process.

**Table 8**
Statistical table of data volume after preprocessing of wells S-1 and S-2.

| Well name | Type | Top depth, m | Bottom depth, m | Data size, line |
|---|---|---|---|---|
| S-1 | Training well | 2123 | 4667 | 2545 |
| S-2 | Testing well | 1600 | 4400 | 2801 |

**Table 9**
Table of normalized input variable.

| No. | Input variable name | | | |
|---|---|---|---|---|
| | Wob | Torque | Spp | Resist |
| 1 | 0.096551724 | 0.033910035 | 0.308113489 | 0.880945301 |
| 2 | 0.075862069 | 0.042906574 | 0.297162768 | 0.880954047 |
| 3 | 0.186206897 | 0.213148789 | 0.308611249 | 0.880980286 |
| 4 | 0.137931034 | 0.2 | 0.289696366 | 0.881225182 |
| 5 | 0.131034483 | 0.248096886 | 0.283225485 | 0.880901569 |

the output layer, with the crucial task being to determine the number of nodes in the hidden layer, as this directly affects the predictive accuracy of the model. Therefore, this study first employs an empirical formula (Eq. (17)) to determine the minimum number of hidden layer nodes, followed by using a grid search algorithm to optimize the number of hidden layer nodes, using the mean squared error (MSE) of the training set as the criterion, where a lower MSE indicates stronger model fitting capability. The study found that the MSE is minimized at 0.018064 when the number of hidden layer nodes is set to 5. The results are illustrated in Fig. 13.

$$N = \sqrt{n + m} + a \tag{17}$$

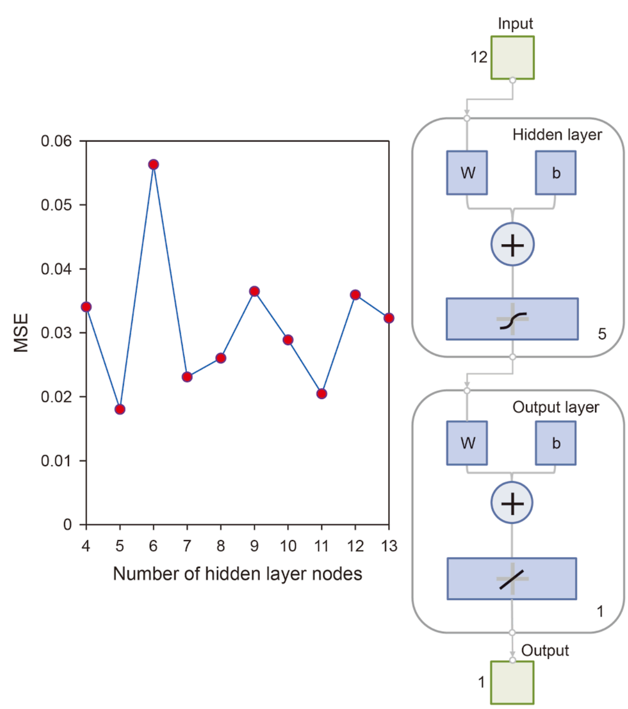


**Fig. 13.** Training set MSE under different numbers of hidden layer nodes.

where $N$ is the number of nodes in the hidden layer; $n$ is the number of nodes in the input layer; $m$ is the number of nodes in the output layer; $a$ is a constant between 1 and 10.

#### 4.2.2. Dataset partitioning strategy: implications for abnormal pressure prediction

It should be clearly pointed out that for wells with complex vertical pore pressure systems, dataset partitioning is not a simple, neutral procedure but rather one of the key factors influencing the generalization ability of a model.

In terms of data partitioning, this study divides the training and test sets in a 70:30 ratio, and also implements a ten-fold cross-validation set, as shown in Fig. 14. It is important to note that there are currently two methods for dividing the dataset: one involves a direct partitioning of the dataset as is, using 70% of the upper formation data as the training set and 30% of the lower formation data as the validation set. The other method involves randomly shuffling the dataset before partitioning. Analyzing the vertical distribution of pore pressure in Well S-1, the deep members belong to an abnormal high-pressure formation, with equivalent pore pressure densities ranging from 1.2 g/cm$^3$ to 1.49 g/cm$^3$, while the middle and upper members are at normal pressure, where the majority of formations have equivalent pore pressure densities around 1.0 g/cm$^3$. There is a significant difference between these two pressure states. Using a model trained on the upper normal pressure formation data to predict the pore pressure of the lower high-pressure formations will likely lead to large errors in the predictions, as the model has not been trained on high-pressure formation data. Therefore, it is preliminarily concluded that a simple direct partitioning of the dataset is not advisable.

This study examines how the method of dividing a quantitative research dataset affects the model. We used a direct approach, training the model with data from the upper well member and validating it with data from the lower well member. Fig. 15 presents the training results. As shown in Fig. 15, the model's predictions for the high-pressure well section remained around a standard equivalent density of 1.0 g/cm$^3$, significantly deviating from the actual high-pressure conditions. The maximum absolute prediction error reached 32.104%. This discrepancy arises because the model did not learn from the data of the high-pressure section, leading to inaccurate predictions of pore pressure under those conditions. Clearly, the model's predictions failed.

In summary, the examination of dataset partitioning strategies constitutes one of the innovations of this study. By comparing two approaches, namely direct depth-based partitioning (using shallow data as the training set and deep data as the test set) and random partitioning (optionally stratified by pore pressure conditions), this paper systematically quantifies such impacts in abnormally high-pressure zones and pressure reversal zones, an issue seldom quantitatively investigated in prior studies. For this study and similar geological data with non-stationary, segmental characteristics (vertical normal-pressure to overpressure segmentation), the sample partitioning strategy for training and test

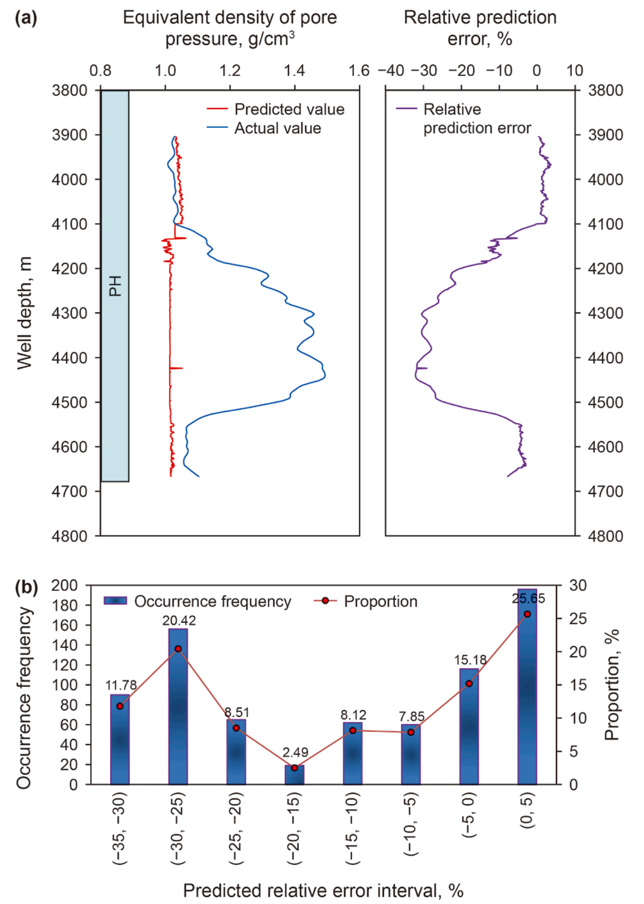


**Fig. 15.** Performance of the BP model on the test set without randomly partitioning the dataset.

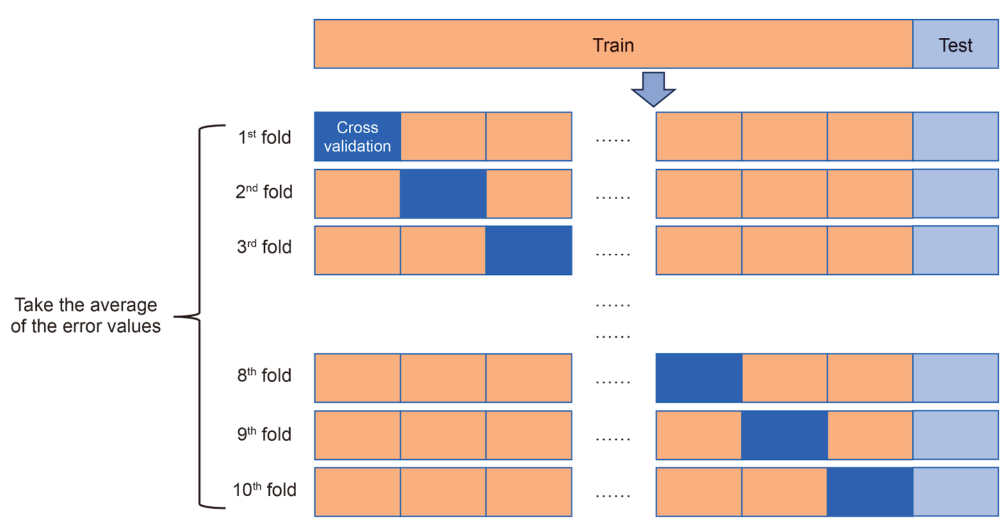


**Fig. 14.** Schematic diagram of 10 fold cross validation algorithm principle.

sets exerts a decisive influence on the generalization performance of ML models. The root cause of this issue lies in distribution shift and insufficient expressive space of training samples: for wells with significant differences in pore pressure conditions, direct depth-based dataset partitioning leads to non-overlapping distributions of the target variable (pore pressure) between the training and test sets. This restricts the model to learning only statistical features of the normal-pressure interval, while lacking representative samples from the high-pressure interval, thereby resulting in severe generalization errors. Random partitioning (or the more rigorous stratified random partitioning by pressure conditions) retains samples from different pressure states in the training set, ensuring the model learns corresponding features under all pressure conditions. This fundamentally mitigates bias risks and reduces drilling risk decision-making errors caused by sample imbalance.

Thus, the methodological innovation of proposing the “regional geological characteristics-data partitioning strategy” adaptation criterion in this paper not only ensures the model performance of this study but also provides a replicable methodological reference for related research in similar regions (such as deep-water areas, ultra-deep formations, and multi-tectonic activity blocks).

### *4.3. Training, evaluation, and application of an intelligent model for pore pressure prediction*

This study develops an intelligent pore pressure prediction model using five ML algorithms: BP, KNN, SVR, CatBoost, and LightGBM. Fig. 16 and Table 10 display six evaluation metrics for the five ML models across the training, testing, and cross-validation datasets. These quantitative metrics allow us to assess the predictive performance of the different algorithms. In terms of training time, the CatBoost model took the longest at 17.214 s, significantly exceeding the other four models. The BP model followed with a training time of 1.849 s. The training times for the LightGBM, KNN, and SVR models were 0.45 s, 0.313 s, and 0.267 s, respectively. While training time is an important aspect of model evaluation, it should not be the sole criterion for determining the effectiveness of ML models.

The purpose of comparing different ML models is to identify the optimal one. This requires the model to perform well not only on the training dataset but also on new data, specifically the test dataset. This paper focuses on the performance of five models across five evaluation metrics on the test set. In terms of MSE, the KNN, CatBoost, and LightGBM models showed the best performance, all achieving a value of 0.001. For the other evaluation metrics, LightGBM consistently outperformed the other models, achieving values of 0.004 for RMSE, 0.003 for MAE, 0.239 for MAPE, and 0.999 for $R^2$. The BP, SVR, KNN, and CatBoost models showed significant gaps compared to LightGBM across all five evaluation metrics. For instance, the SVR model increased its RMSE, MAE, and MAPE values by 1400%, 1433.333%, and 1644.770%, respectively, compared to LightGBM, indicating substantial differences. Regarding $R^2$, the BP, SVR, KNN, and CatBoost models showed reductions of 24.625%, 22.422%, 4.304%, and 1.101% compared to LightGBM, respectively. Therefore, although LightGBM ranked third among the five models in terms of training duration, it exhibited the best overall performance. This indicates that LightGBM model effectively learns the complex nonlinear relationships among drilling parameters, rock mechanics parameters, and pore pressure, demonstrating its high generalizability and stability.

Tree-based ML algorithms, such as CatBoost and LightGBM, offer more interpretability compared to black-box models like BP, KNN, and SVR, as they allow for the direct calculation of feature importance for each input variable after training. As shown in Fig. 17, both CatBoost and LightGBM identify Depth as the most important feature, indicating its significant impact on the model’s accuracy and construction. However, the two models differ in several ways. In the CatBoost model, the feature importance of Depth reaches 56.3%, accounting for more than half of the total importance. Furthermore, the feature importance among the input variables in the CatBoost model varies greatly. The importance of Ecd is 16.40%, while the importance of other variables is quite low, with Torque and $E_{\rm d}$ at only 0.90% and 0.80%, respectively. In contrast, the LightGBM model shows a more balanced distribution of feature importance across input variables, suggesting that it takes all input factors into account more comprehensively.

The feature importance rankings presented in Fig. 17 not only illustrate the dependence of the two models on each input variable, but also carry substantial geological and petrophysical implications, aligning closely with the deep geological evolution and pore pressure generation mechanisms in the Xihu Sag. Burial depth emerges as the most critical feature, as it dominates stratal compaction, fluid trapping, and tectonic compression, acting as the fundamental geological control on the formation and evolution of abnormal overpressure in the study area. In the CatBoost model, burial depth accounts for 56.3% of feature importance, indicating strong sensitivity to and excessive reliance on this dominant macroscopic factor. By contrast, burial depth contributes only 32.0% to feature importance in the LightGBM model, which retains key macroscopic controls while better integrating multi-scale geological and engineering data, demonstrating superior feature capture capability. Differences in the importance of other parameter types reflect the distinct preferences of the two models for capturing dynamic response signals of pore pressure. Drilling parameters directly reflect pressure equilibrium between the wellbore and formation, as well as overpressured fluid migration, consistent with the geological characteristics of well-developed overpressure zones and active fluid flow in the Xihu Sag. Rock mechanics parameters indirectly indicate abnormal pore pressure by characterizing the effective stress state of rock formations, matching the regional geological setting of intense tectonic compression and strong spatial heterogeneity in effective stress. These two categories of parameters provide complementary insights from the perspectives of dynamic drilling responses and static mechanical properties, respectively. The LightGBM model responds reasonably to both drilling and rock mechanics parameters, enabling it to fully characterize the complex evolutionary mechanisms of pore pressure rather than relying solely on burial depth. This is the core mechanistic reason for its higher pore pressure prediction accuracy compared with the CatBoost model. It better reflects the actual geological processes of multi-genetic overpressure superposition in the Xihu Sag, further verifying the geological rationality of the model’s learning outcomes.

The balanced distribution of feature importance in the LightGBM model stems from its unique optimization mechanisms. These mechanisms allow the model to allocate attention more efficiently when processing high-dimensional, non-uniform datasets, preventing over-reliance on dominant features (e.g., Depth) and resulting in a relatively balanced distribution of feature importance. Unlike CatBoost’s ordered boosting and random tree permutation, LightGBM adopts three key strategies: first, the Gradient-based One-Side Sampling (GOSS) prioritizes sampling data points with high gradients while retaining a portion of low-gradient samples to maintain diversity; second, Exclusive Feature Bundling (EFB) reduces dimensionality by bundling mutually exclusive features; third, the leaf-wise growth strategy captures complex interaction structures more thoroughly while preserving computational efficiency. Collectively, these

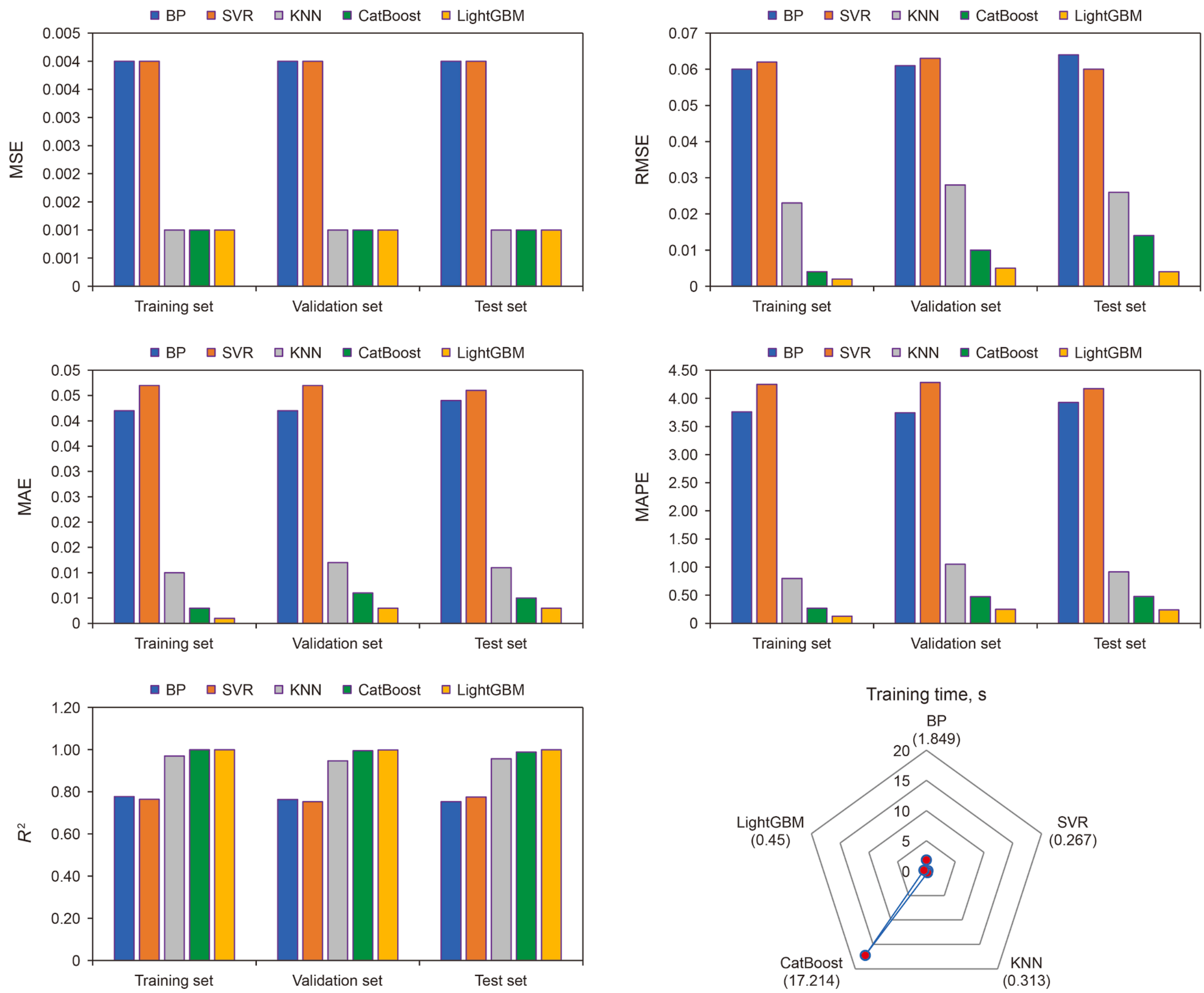


**Fig. 16.** Evaluation results of the five algorithm models.

**Table 10**
Differences in evaluation metrics between the BP, SVR, KNN, and CatBoost models, compared to the LightGBM model.

| ML models | Evaluation index | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MSE | Difference, % | RMSE | Difference, % | MAE | Difference, % | MAPE | Difference, % | $R^2$ | Difference, % |
| BP | 0.004 | 300 | 0.064 | 1500 | 0.044 | 1366.667 | 3.927 | 1543.096 | 0.753 | −24.625 |
| SVR | 0.004 | 300 | 0.06 | 1400 | 0.046 | 1433.333 | 4.17 | 1644.770 | 0.775 | −22.422 |
| KNN | 0.001 | 0 | 0.026 | 550 | 0.011 | 266.667 | 0.915 | 282.845 | 0.956 | −4.304 |
| CatBoost | 0.001 | 0 | 0.014 | 250 | 0.005 | 66.667 | 0.477 | 99.582 | 0.988 | −1.101 |
| LightGBM | 0.001 | | 0.004 | | 0.003 | | 0.239 | | 0.999 | |

mechanisms enhance the model's ability to learn "weak signals from multiple features" and complex interactions between features. Particularly in scenarios with strong vertical data heterogeneity, such as the Xihu Sag, LightGBM can better capture the interactions among multiple parameters. In contrast, CatBoost's category handling and residual fitting may, on specific datasets, concentrate importance on a small number of highly correlated features (e.g., Depth) to achieve rapid fitting. This results in one or a few features accounting for a large proportion in feature importance evaluations. From an engineering application perspective, the balanced weight distribution of LightGBM holds practical value. It indicates that the model does not over-rely on a single variable but instead establishes broader coupling relationships between various drilling parameters and rock mechanics parameters. This generally improves the model's generalization ability for pore pressure prediction across different wells and its robustness in cases involving abnormal or rare samples. In comparison, although CatBoost's high reliance on Depth in this case facilitates a good fit with the training well data, it may amplify generalization errors if the relationship between well depth and pore pressure in the study area is disrupted or locally biased. A typical example is the decoupling of local pore pressure systems caused by the development of deep faults.

Figs. 18 and 19 illustrate the comparison between predicted and actual values for five algorithm models on the test set. Analyzing these figures together provides a clearer picture of the models'

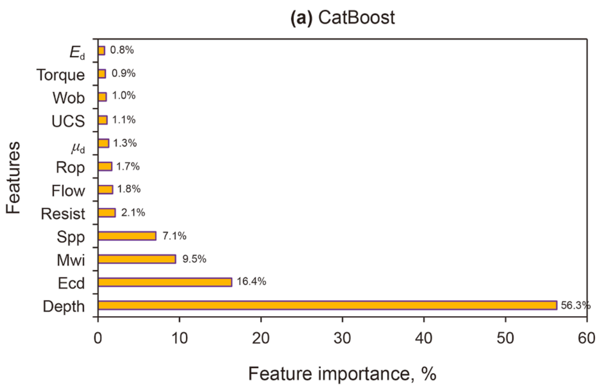


**Fig. 17.** The feature importance in the CatBoost model and LightGBM model.

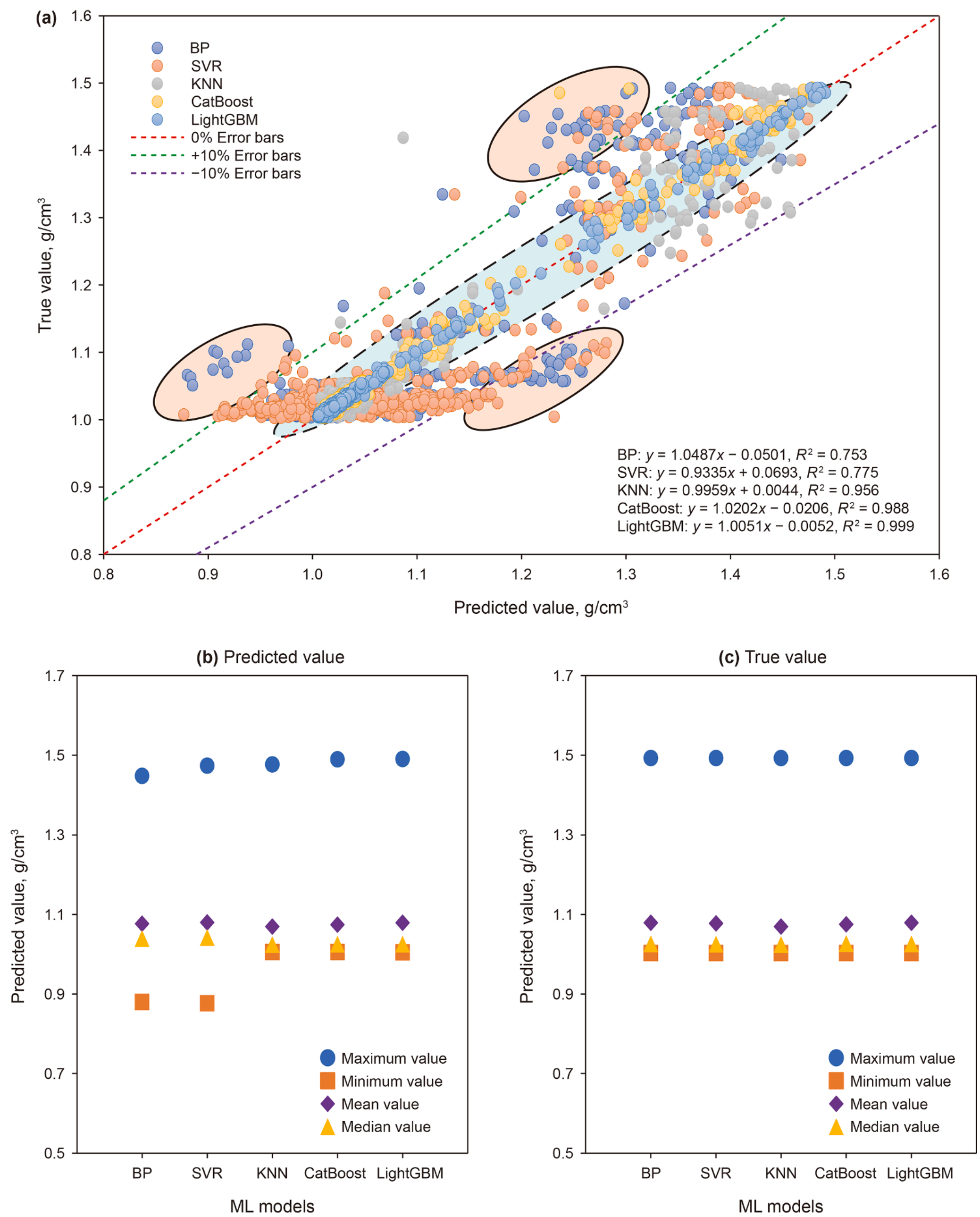


**Fig. 18.** Performance of the five algorithm models on the test data.

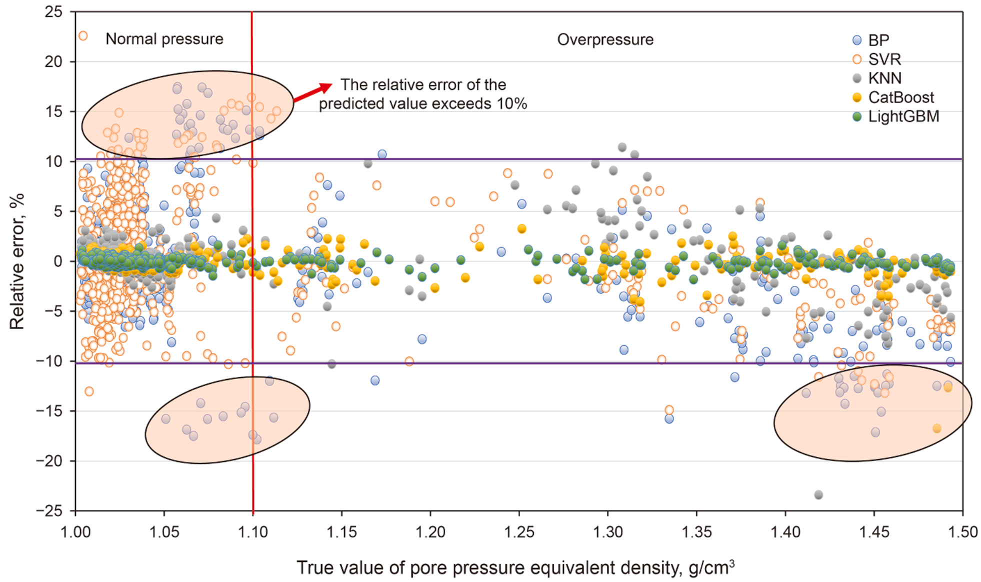


**Fig. 19.** Performance of the five algorithm models on the test data.

performance. From these figures, it is evident that among all test data, the LightGBM model exhibits the smallest prediction error, with a relative error range of (−1.830% to +1.613%). In contrast, the BP, SVR, KNN, and CatBoost models have relative error ranges of (−17.852% to +17.453%), (−14.888% to +22.592%), (−23.401% to +11.428%), and (−16.753% to +3.254%), respectively, all significantly higher than that of LightGBM model. It is easy to understand that ML models can predict normal pressure data points accurately, but struggle with abnormal high-pressure points. However, whether the data points are at normal or high pressure, the LightGBM model consistently aligns well with actual values, maintaining low prediction errors. The other four models exhibit a problematic pattern, where prediction errors are small for normal pressure points but significantly larger for high-pressure points. Notably, the BP and SVR models have prediction errors exceeding 10%, and in some cases, even 15% for a substantial number of normal pressure data points. This analysis confirms that the LightGBM model is the optimal choice among the five algorithms, achieving over 98% prediction accuracy for both normal and high-pressure data points.

To further validate the generalization ability of the five ML models and assess their final performance in real-world applications, the previously trained BP, SVR, KNN, CatBoost, and LightGBM models were used to predict the pore pressure of the adjacent Well S-2 relative to Well S-1. The prediction results from different models were compared and analyzed, with detailed equivalent densities for pore pressure shown in Figs. 20 and 21.

Figs. 20 and 21 reveal that, overall, the pore pressure predictions from the CatBoost and LightGBM models exhibit low volatility, while the predictions from the BP and SVR models show significant fluctuations. The predictions from the KNN model are clearly erroneous and do not align with existing research on pore pressure, indicating that the KNN model has failed in practical applications; therefore, it will be excluded from further consideration. Furthermore, a significant portion of the pore pressure predictions from the BP and SVR models do not conform to the geological understanding of the block. For instance, between depths of 3200 m and 3400 m, the BP and SVR models predict a

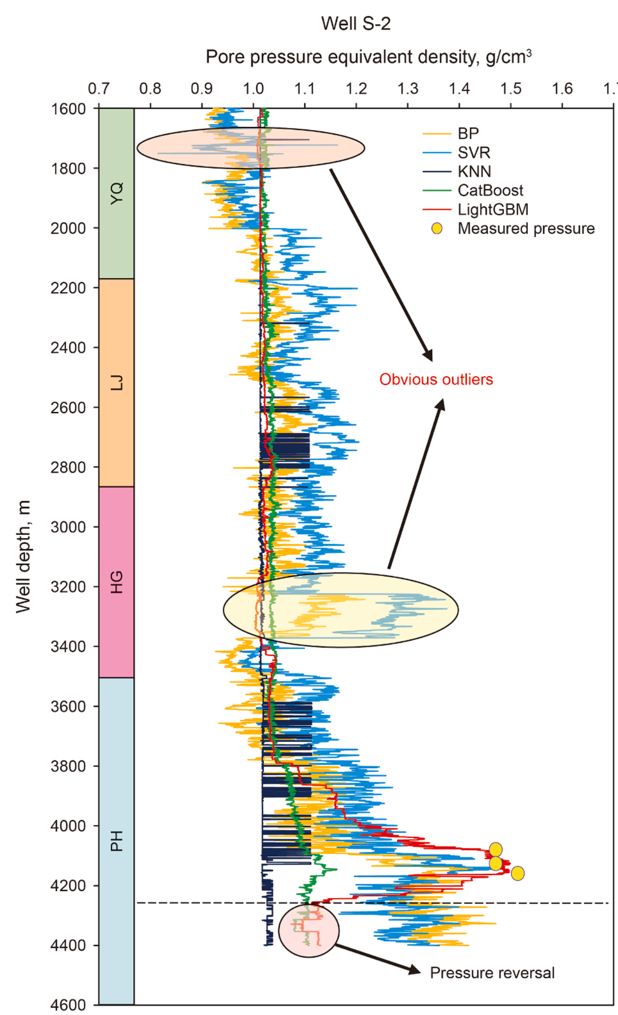


**Fig. 20.** Predictions of pore pressure for Well S-2.

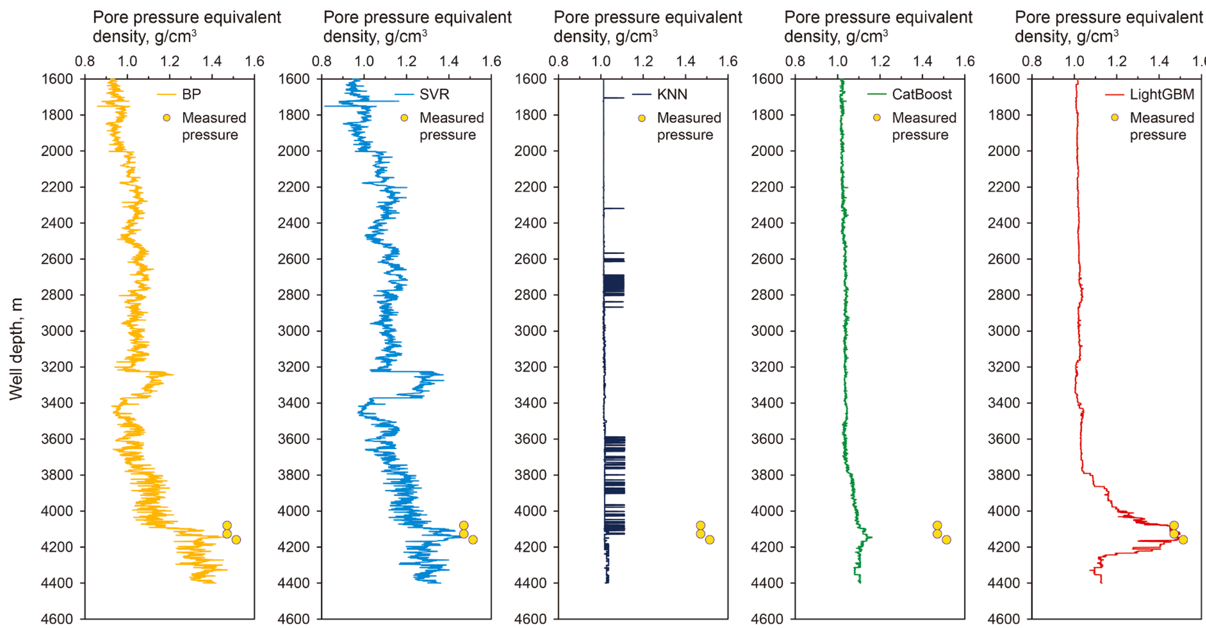


**Fig. 21.** Predictions of pore pressure for Well S-2.

sudden increase in equivalent pore pressure density to around 1.2 g/cm$^3$, followed by a return to normal pressure, which is clearly an error. Additionally, based on prior pore pressure analyses, the formations in the Yuquan, Longjing, and Huagang groups should remain at normal pressure. However, the SVR model shows drastic fluctuations in pore pressure predictions between 1600 m and 1800 m, with values ranging from a maximum of 1.164 g/cm$^3$ to a minimum of 0.815 g/cm$^3$, which are inconsistent with actual engineering conditions. Moreover, the pore pressure predictions from the BP and SVR models are significantly higher than those from the CatBoost and LightGBM models, with a difference of about 0.2 g/cm$^3$. Since these formations should be at normal pressure, the expected pore pressure density should be around 1.0–1.1 g/cm$^3$, indicating that the predictions from BP and SVR are clearly incorrect. Therefore, the BP and SVR models will also be excluded. In contrast, the predictions from the CatBoost and LightGBM models for the upper and middle members of the Well S-2 remain around the normal pressure of 1.0 g/cm$^3$, aligning with geological background and pore pressure distribution expectations.

Analyzing the trend of increasing pore pressure, the BP, SVR, CatBoost, and LightGBM models all indicate a certain upward pressure trend in the deeper members of the Pinghu Formation, specifically between 4000 m and 4200 m. Among these, the CatBoost model shows the smallest increase, while the LightGBM model exhibits the largest. Previous drilling data indicate that abnormal high pressure begins to develop in the central part of the Pinghu Formation in this block. Thus, all four ML models are capable of predicting the trend of increasing pore pressure. However, the pore pressure predictions from the CatBoost model fall significantly short of the actual measured values. As shown in Fig. 22, the CatBoost model has prediction errors of −24.983%, −23.897%, and −25.014% at three measured pore pressure data points, indicating large prediction errors; therefore, the CatBoost model will be excluded. Although the BP and SVR models align somewhat better with the measured values at the pore pressure data points compared to the CatBoost model, the BP model has prediction errors of −21.530%, −13.481%, and −10.779% at the three measured pore pressure data points. The SVR model shows prediction errors of −14.939%, −8.399%, and −8.147% at the same points. However, as noted in the previous analysis, the pore pressure predictions from the BP and SVR models do not align with geological background and pore pressure distribution patterns, leading to their exclusion.

In contrast, the pore pressure predictions from the LightGBM model closely match the actual measured values, and its overall vertical distribution aligns with previous studies on pore pressure distribution in the block. Fig. 22 shows that the LightGBM model achieves prediction accuracies of 97.594%, 98.115%, and 98.038% at the three measured pore pressure data points, indicating high accuracy that meets field engineering requirements.

This analysis demonstrates that the LightGBM-based intelligent prediction model for pore pressure has high accuracy and robust generalization capability, proving effective in practical applications and providing more accurate predictions and decision support for oilfield drilling operations.

### *4.4. Analysis of measured pore pressure data and distribution patterns in the target block*

This study focuses on two vertical wells, S-1 and S-2, located within the western slope belt of the Xihu Sag. To validate that the pore pressure predictions from the LightGBM model for Well S-2 align with the pore pressure distribution patterns in the western slope belt, this study also includes four neighboring wells S-3, S-4, S-5, and S-6 within the same block as S-1 and S-2 (the relative positions of these six wells are shown in Fig. 23). The measured pore pressure data from the six wells were collected and analyzed for their vertical distribution patterns, as detailed in Fig. 24. To facilitate an intuitive understanding of pore pressure conditions, this study adopts the classification table proposed by Fan (2016) (see Table 6), designating formations with a pore pressure coefficient greater than 1.1 as overpressure formations.

Due to the varying depths of stratigraphic layers across different wells, a distribution map of measured pore pressure data at various formations was created to provide a clear assessment of the equivalent density of pore pressure for the six wells, as shown in Fig. 24. The measured pore pressure coefficients for the six wells

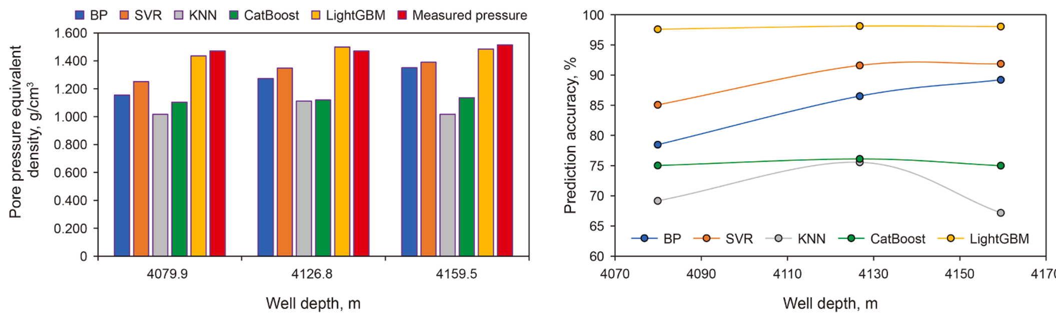


**Fig. 22.** Prediction error of measured pore pressure in Well S-2.

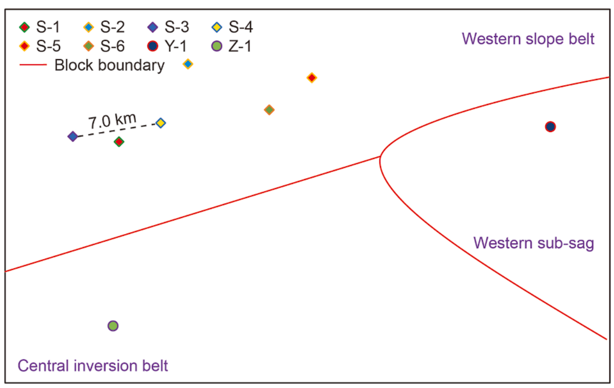


**Fig. 23.** The relative positional relationships of the eight research wells within the Xihu Sag, including wells S-1 to S-6 in the western slope belt, Well Y-1 in the western sub-sag, and Well Z-1 in the central inversion belt.

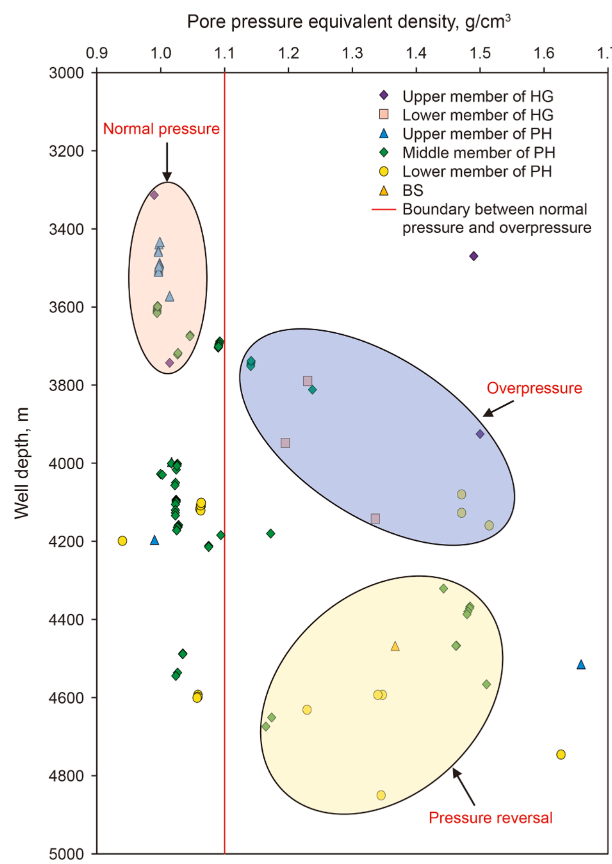


**Fig. 24.** Vertical distribution map of measured pore pressure data from the different formations of the six wells in the western slope belt of the Xihu Sag.

range from a minimum of 0.940 (at a depth of 4198.71 m in the lower member of the Pinghu Formation) to a maximum of 1.658 (at a depth of 4514.5 m in the upper member of the Pinghu Formation). Vertically, the pore pressure initiation point is around 3700 m, where overpressure is generally developed in deeper reservoirs. As burial depth increases, pore pressure rises in a “step-like” manner. Additionally, it was observed that a significant number of measured pore pressure data points show a decrease in pressure coefficient with increasing depth, indicating a reversal in pressure trends. Therefore, the strata can be categorized vertically into normal pressure zones, pressure transition–overpressure zones, and pressure reversal zones.

The research found that, apart from the upper member of the Pinghu Formation, which generally exhibits normal pore pressure, other formations show a mix of normal and abnormally high pore pressure data points. This indicates that the Xihu Sag has undergone multi-period tectonic movement and has developed deep faults, resulting in a fragmented pore pressure system underground. Different wells exhibit varying initiation pressures, and there are significant differences in the top interface depth for abnormal pore pressure. Overall, as the depth of the strata increases, the pore pressure gradually rises, with the layers first entering a high-pressure zone before transitioning into a pressure reversal zone.

Figs. 20 and 21 illustrate that the LightGBM model predicts the pore pressure in Well S-2 to follow the distribution pattern of “normal pressure zone–overpressure zone–pressure reversal zone”. Additionally, the pore pressure increases in a “stair-step” manner with depth. The predicted pore pressure values from the LightGBM model align closely with the measured values, achieving an average prediction accuracy of over 97%, indicating a high level of predictive reliability. This analysis further confirms that the pore pressure predictions made by the LightGBM model are consistent with the distribution patterns observed in the target area. Therefore, the use of the LightGBM model for intelligent predictions of pore pressure is recommended.

### 4.5. Model validation across tectonic units: an applied analysis of the western sub-sag and the central inversion belt

As shown in Fig. 2(d), the Xihu Sag is divided into several tectonic units from west to east, including the western slope belt, the western sub-sag, and the central inversion belt. These units show distinct differences in tectonic evolution, fault development intensity and the characteristics of pressure systems. To further verify the generalizability and robustness of the LightGBM model developed in this study across different blocks, Well Y-1 (in the western sub sag) and Well Z-1 (in the central inversion belt) were selected as cross-tectonic unit validation wells. This selection addresses the limitation of previous research, where all study wells were concentrated in a single block. The predicted pore pressure results for the two wells are presented in Fig. 25, and a quantitative error analysis was performed by combining these predictions with measured pressure data (see Fig. 26).

As shown in Fig. 25(a), the predicted pore pressure curve of Well Y-1 is highly consistent with the measured values overall. In the normal pressure zone (Yuquan, Longjing, and the upper member of the Huagang formations, at a depth of 1800–5100 m), the predicted pore pressure equivalent density stabilizes at approximately 0.95–1.10 g/cm$^3$. Pore pressure begins to rise at the boundary zone between the Huagang and Pinghu formations at a well depth of about 5100 m, and peaks at a maximum value of 1.17 g/cm$^3$ at 5285 m (within the Pinghu Formation). The predicted pore pressure profile accurately captures the vertical distribution pattern of the well, which transitions from the normal pressure zone (Yuquan, Longjing, and upper member of Huagang formations) to the overpressure transition zone (lower member of the Huagang Formation) and then to the overpressure zone (lower member of the Huagang Formation and Pinghu Formation), with pore pressure showing a stepwise increasing trend. Quantitative analysis of 8 measured pore pressure points, all located in the upper member of the Huagang Formation (see Fig. 26(a)), reveals a mean predicted relative error of 2.44%, a maximum predicted relative error of 7.58%, and a minimum predicted relative error of 0.30%.

The predicted pore pressure profile of Well Z-1 is similar to that of Well Y-1 with minor local discrepancies. First, Well Z-1 was drilled to completion within the lower member of the Huagang Formation and did not penetrate the Pinghu Formation, with a total depth much shallower than that of Well Y-1. In addition, due to the influence of local tectonic activities, the burial depths of the top and bottom surfaces of different strata in Well Z-1 differ significantly from those in Well Y-1, whereas its stratigraphic burial characteristics are highly consistent with those in Well S-1. Overall, owing to the relatively shallow total depth of Well Z-1, the predicted pore pressure values across the entire well section fall within the normal pressure range, with the equivalent density ranging from 0.99 g/cm$^3$ to 1.03 g/cm$^3$, and no overpressure or pressure inversion is observed. As shown in Fig. 26(b), the predicted relative errors for the two measured pore pressure points are −0.55% and −0.65%, respectively, indicating a high level of prediction accuracy.

In summary, the LightGBM model exhibits strong predictive performance even in relatively unstudied tectonic blocks. Predictions for both validation wells exhibited no notable fluctuations or systematic bias, producing an overall accuracy above 92%, which

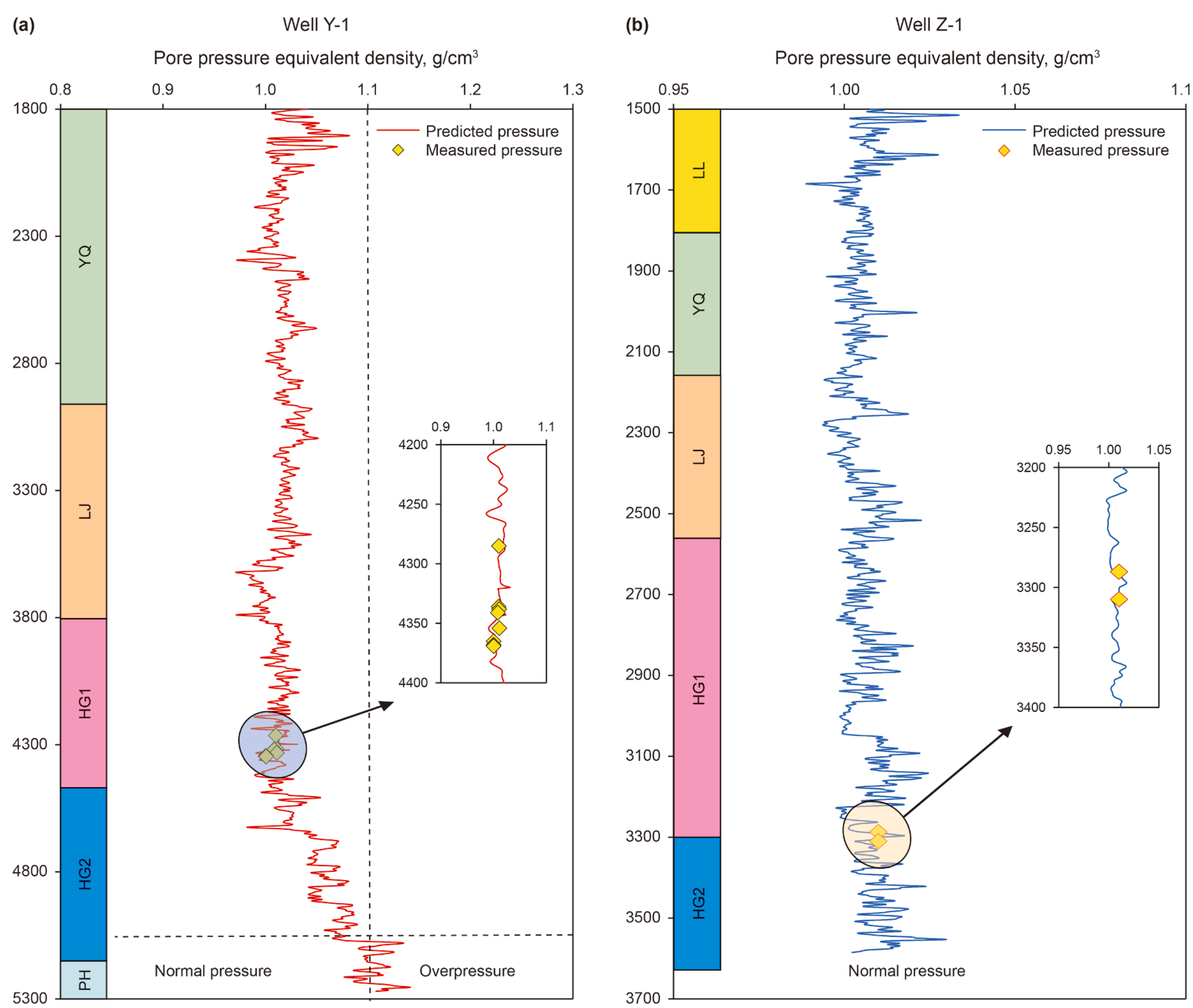


**Fig. 25.** Predicted pore pressure of wells Y-1 and Z-1.

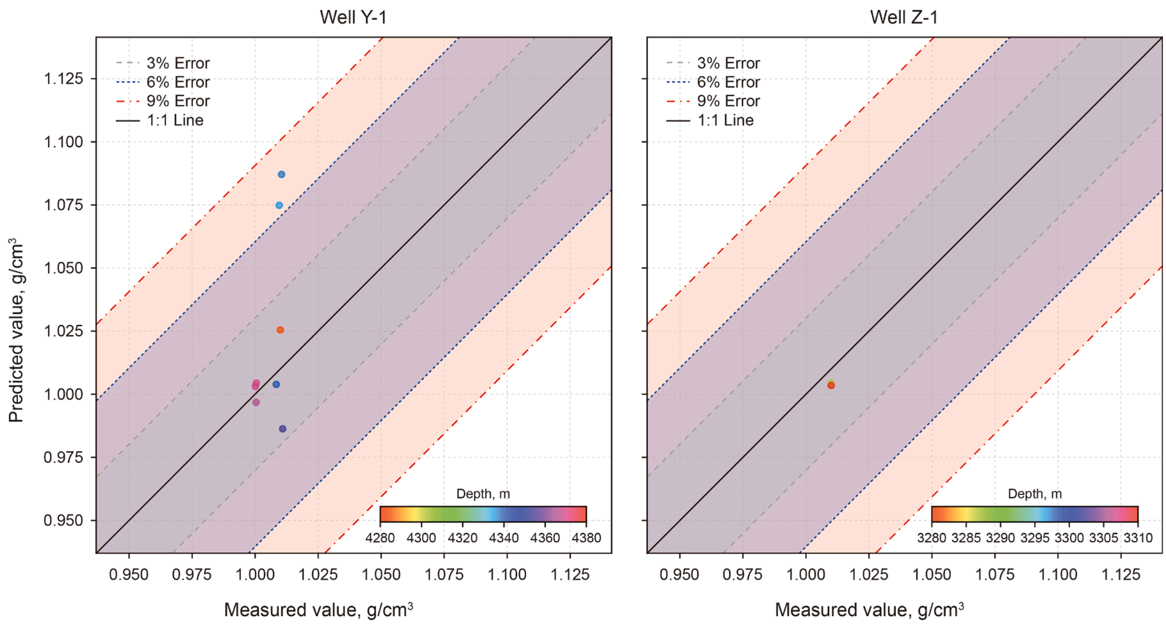


**Fig. 26.** Prediction errors of the measured pore pressure data for the wells Y-1 and Z-1.

is comparable to the Western Slope Belt validation (mean accuracy >97%). This demonstrates the model's stable performance, with no notable decline in accuracy amid variations in geological conditions, and confirms its high robustness. Whether in the Western Sub Sag or the Central Inversion Structural Belt, the LightGBM model proposed in this study can accurately reproduce the pore pressure distribution pattern without the need for retraining or parameter adjustment. This capability stems from three key factors: (1) the diversity of input variables (drilling parameters reflecting real-time dynamic responses and rock mechanics parameters capturing static properties), (2) the optimization mechanisms of LightGBM (GOSS and EFB ensuring the balanced utilization of input features), and (3) the partitioning strategy adopted for the training sample dataset. These findings verify that the technical framework established in this study has promising potential for cross-scenario application and can serve as a reference for pore pressure prediction in other analogous complex petroliferous basins.

### *4.6. Limitations and future work*

This research uses the Xihu Sag as a case study to introduce an innovative approach for predicting pressure in abnormally high-pressure formations. While this method significantly enhances the accuracy, usability, and efficiency of such predictions, the study has certain limitations that merit attention in future studies.

(1) Data requirements and representativeness: One of this study's principal innovations is an ML model that integrates drilling and rock-mechanics parameters. However, the effective implementation of this approach is highly dependent on multi-source, continuous, and high-quality datasets, which becomes the primary constraint on its widespread application. In field operations, sudden events such as wellbore collapse, mud loss, or equipment failure often produce large numbers of anomalous or missing drilling records. Furthermore, the model's training and testing drew solely from well data in the Xihu Sag. Although this data effectively represents local conditions, other areas, such as the Bohai Bay Basin, may differ in rock composition, pressure patterns, and structural features. Future research should broaden testing to include diverse settings and refine model settings based on unique local traits to improve its broader application.

(2) Model interpretability: The LightGBM model developed here attains high accuracy and efficiency, and global interpretability was provided via feature importance analysis, which identified the key input variables (for example, well depth and density). However, this global analysis characterizes aggregate model behaviour and offers limited insight into predictions for individual cases. In the future, integrating local interpretability methods (e.g., SHAP) could provide more granular explanations, facilitating a more comprehensive understanding of the model's decision-making process and thereby enhancing its credibility and practicality in engineering decision-making.

(3) Real-time deployment and operational applicability: The proposed model achieves high predictive accuracy with short training times, demonstrating real-time prediction capability in laboratory tests. Translating this capability into field deployment and online decision support, however, presents several practical challenges. The primary limitation of the model lies in its data acquisition approach: its three input rock mechanics parameters are derived from log-based calculations calibrated with field measurements. In standard drilling operations, these parameters are typically obtained via wireline logging only after well completion. Real-time parameter input while drilling is therefore impractical unless costly LWD acoustic tools are deployed continuously. Accordingly, under conventional operational conditions, the model cannot function as a true real-time risk early-warning system while drilling. Instead, it is better suited for pre-drilling pressure system evaluation of adjacent wells, near-real-time decision support when LWD tools are available, and post-drilling pore pressure back-analysis. In addition, laboratory training and validation use preprocessed, clean data; field measurements, by contrast,

are subject to tool vibration and electromagnetic interference that distort signals. Without an online preprocessing pipeline (for example, real-time outlier detection and signal smoothing), feeding raw field data into the model can cause a sharp drop in prediction accuracy.

It is important to note that these limitations stem mainly from data specifics, geological variations, and practical usage requirements, yet they do not diminish the core value or real-world relevance of this study. We intend to address these challenges in upcoming work, by advancing prediction techniques to offer stronger support for efficient offshore resource development.

## 5. Conclusions

The Xihu Sag of the East China Sea Shelf Basin is a petroliferous area of global interest, but its multi-stage tectonic inversion and rifting result in complex geology (e.g., deep faults, discontinuous pressure systems) and abnormally high pressures in target formations with strong pore pressure heterogeneities. This poses significant challenges to safe development. With a focus on two vertical wells in the sag's western slope belt, an intelligent pore pressure prediction method has been developed using nine drilling and three rock mechanics parameters, combined with five ML algorithms. Key conclusions are as follows:

(1) Unlike conventional logging-based approaches, drilling parameters (field-derived) and rock mechanics parameters (laboratory-measured, log-calibrated) have been innovatively adopted as model inputs. Redundancies were eliminated via Spearman correlation analysis, with 12 parameters finalized to construct a high-quality dataset for pore pressure prediction.
(2) The BP, SVR, KNN, CatBoost, and LightGBM models—with optimized hyperparameters and data partitioning—have been established and evaluated via six metrics (MSE, MAE, MAPE, RMSE, $R^2$, and training time). CatBoost exhibited the longest training time (17.214 s) and SVR the shortest (0.267 s), though SVR performed poorly in the other five metrics. The LightGBM model achieved the best overall performance, featuring a short training time (0.45 s), minimal errors (MSE = 0.001, RMSE = 0.004, MAE = 0.003, MAPE = 0.239), and near-perfect fitting ($R^2$ = 0.999), significantly outperforming the other four models.
(3) Feature importance analysis for CatBoost and LightGBM identified Depth as the most critical input. While CatBoost showed extreme variability in feature weights, LightGBM exhibited a balanced importance distribution—enabling it to capture complex parameter-pressure relationships, a key driver of its superiority.
(4) On regression test data, LightGBM had the narrowest error range (−1.830% to +1.613%), compared to ~20% absolute errors for the other four models. It accurately predicted both normal and overpressure points, achieving >98% test-set accuracy.
(5) Practical validation on adjacent Well S-2 (relative to Well S-1) and four regional wells confirmed the robustness of the LightGBM model: its predictions aligned with the western slope's pore pressure distribution, and it achieved an average accuracy exceeding 97% against the three measured values of Well S-2. In contrast, KNN predictions deviated significantly from regional patterns; BP and SVR models, while relatively accurate in deep high-pressure zones, exhibited two critical issues: maximum prediction errors exceeding 20% against Well S-2's measured values, and substantial fluctuations in upper-middle layer predictions that contradicted the expected normal pressure conditions. For the CatBoost model, though its upper-middle layer predictions indicated normal pressure and deep formations showed a slight upward pressure trend, the pressure increase was insufficient and failed to meet expectations, with prediction errors of approximately 25% against measured values, indicating low accuracy. The LightGBM model is recommended for pore pressure prediction in the Xihu Sag's western slope belt.

The integrated data-driven framework and the insights gained, particularly regarding data preparation for heterogeneous pressure systems, provide a transferable paradigm that can enhance drilling safety and efficiency in other complex basins worldwide facing similar challenges of abnormal pore pressure prediction.

## CRediT authorship contribution statement

**Hua-Yang Li:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Quan-You Liu:** Writing – review & editing, Resources, Project administration, Funding acquisition. **Jia-Ao Chen:** Writing – review & editing, Software. **Yan-Chao Pang:** Software. **Fu-Zhi Chen:** Writing – original draft. **Dan-Tong Liu:** Software. **Ze-Hui Shi:** Software.

## Data availability

The datasets used or analysed during the current study are available from the corresponding author on reasonable request.

## Funding

This study was supported by the National Natural Science Foundation of China (No. 42488101).

## Declaration of competing interest

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

## Acknowledgements

We extend our appreciation to the editors and reviewers for their valuable contributions towards improving this paper.

## References

Alakbari, F.S., Mohyaldinn, M.E., Ayoub, M.A., et al., 2024. A gated recurrent unit model to predict Poisson's ratio using deep learning. J. Rock Mech. Geotech. Eng. 16 (1), 123–135. https://doi.org/10.1016/j.jrmge.2023.04.012.
Bai, J., Wang, S., Xu, Q., et al., 2023. Intelligent real-time predicting method for rock characterization based on multi-source information integration while drilling. Bull. Eng. Geol. Environ. 82 (4), 150. https://doi.org/10.1007/s10064-023-03182-8.
Bowers, G.L., 1995. Pore pressure estimation from velocity data: Accounting for overpressure mechanisms besides undercompaction. SPE Drill. Complet. 10 (2), 89–95. https://doi.org/10.2118/27488-pa.
Bowers, G.L., 2001. Determining an appropriate pore-pressure estimation strategy. In: Offshore Technology Conference. https://doi.org/10.4043/13042-ms.
Bruna Teixeira Silveira, B.S., Claudio Rabe, C.R., Cristian Mejia, C.M., et al., 2023. Neural network model to predict pore pressure and wellbore stability from LWD log and drilling parameter. In: 57th U.S. Rock Mechanics/Geomechanics Symposium. ARMA. https://doi.org/10.56952/arma-2023-0916.
Caja, M.Á., Peña, A.C., Campos, J.R., et al., 2019. Image processing and machine learning applied to lithology identification, classification and quantification of

thin section cutting samples. In: SPE Annual Technical Conference and Exhibition. https://doi.org/10.2118/196117-ms.
Chen, F., Gao, J., Feng, Y., et al., 2024. Optimizing the wellbore trajectory of directional wells considering wellbore stability subjected to the non-independence and uncertainty of geomechanical parameters. Geoenergy Sci. Eng. 241, 213085. https://doi.org/10.1016/j.geoen.2024.213085.
Chen, Y., Zhou, L., Pei, S., et al., 2021. KNN-BLOCK DBSCAN: Fast clustering for large-scale data. IEEE Trans. Syst. Man Cybern. Syst. 51 (6), 3939–3953. https://doi.org/10.1109/tsmc.2019.2956527.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297. https://doi.org/10.1007/BF00994018.
Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theor. 13 (1), 21–27. https://doi.org/10.1109/TIT.1967.1053964.
Cui, X., Dai, L., Li, S., et al., 2019. Control of strike-slip and pull-apart processes to tectonic transition of the southern East China Sea Shelf Basin. Geol. J. 54 (2), 850–861. https://doi.org/10.1002/gj.3363.
Deng, S., Pan, H.Y., Wang, H.G., et al., 2024. A hybrid machine learning optimization algorithm for multivariable pore pressure prediction. Pet. Sci. 21 (1), 535–550. https://doi.org/10.1016/j.petsci.2023.09.001.
Eaton, B.A., 1972. The effect of overburden stress on geopressure prediction from well logs. J. Petrol. Technol. 24 (8), 929–934. https://doi.org/10.2118/3719-pa.
Eaton, B.A., 1975. The equation for geopressure prediction from well logs. In: Fall Meeting of the Society of Petroleum Engineers of AIME, Richardson. Society of Petroleum Engineers. https://doi.org/10.2118/5544-ms.
Ehsan, M., Manzoor, U., Chen, R., et al., 2025. Pore pressure prediction based on conventional well logs and seismic data using an advanced machine learning approach. J. Rock Mech. Geotech. Eng. 17 (5), 2727–2740. https://doi.org/10.1016/j.jrmge.2024.09.049.
Fan, C., Wang, G., 2021. The significance of a piecemeal geometric model of mudstone compaction: Pinghu Slope, Xihu Depression, Eastern China. Mar. Petrol. Geol. 131, 105138. https://doi.org/10.1016/j.marpetgeo.2021.105138.
Fan, H., 2016. Analysis Methods and Applications of Abnormal Formation Pressure. Science Press, Beijing.
Gao, J., Chen, F., Zhang, P., et al., 2026a. Probabilistic inversion of horizontal *in situ* stresses combining Bayesian theory and borehole failure data to quantify the uncertainty of wellbore instability risk. Rock Mech. Rock Eng. 59 (2), 1685–1714. https://doi.org/10.1007/s00603-025-04894-x.
Gao, J., Yang, W., Chen, F., et al., 2026b. Pore pressure prediction using machine learning methods and logging data considering Gaussian mixture clustering model. Geoenergy Sci. Eng. 257, 214188. https://doi.org/10.1016/j.geoen.2025.214188.
Guo, Y.S., Wang, W., 2024. A method for identifying lithology based on a feature-weighted KNN model. Geophys. Geochem. Explor. 48 (2), 428–436. https://doi.org/10.11720/wtyht.2024.1260 (in Chinese).
He, J., Wang, T., Tang, Y., et al., 2023. Geochemical characteristics and gas-washing accumulation mechanism of hydrocarbon from the Xihu Sag, East China Sea Basin. Geoenergy Sci. Eng. 224, 211639. https://doi.org/10.1016/j.geoen.2023.211639.
He, M., Zhang, Z., Li, N., 2021. Deep convolutional neural network-based method for strength parameter prediction of jointed rock mass using drilling logging data. Int. J. GeoMech. 21 (7), 04021111. https://doi.org/10.1061/(ASCE)GM.1943-5622.0002074.
Hegde, C., Millwater, H., Gray, K., 2019. Classification of drilling stick slip severity using machine learning. J. Petrol. Sci. Eng. 179, 1023–1036. https://doi.org/10.1016/j.petrol.2019.05.021.
Hegde, C., Wallace, S., Gray, K., 2015. Real time prediction and classification of torque and drag during drilling using statistical learning methods. In: SPE Eastern Regional Meeting, Richardson. Society of Petroleum Engineers. https://doi.org/10.2118/177313-ms.
Hunt, J.M., 1990. Generation and migration of petroleum from abnormally pressured fluid compartments. AAPG Bull. 74 (1), 1–12. https://doi.org/10.1306/0c9b21eb-1710-11d7-8645000102c1865d.
Kang, S., Shao, L., Qin, L., et al., 2020. Hydrocarbon generation potential and depositional setting of Eocene oil-prone coaly source rocks in the Xihu Sag, East China Sea shelf basin. ACS Omega 5 (50), 32267–32285. https://doi.org/10.1021/acsomega.0c04109.
Ke, G., Meng, Q., Finley, T., et al., 2017. LightGBM: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30, 3146–3154. https://doi.org/10.5555/3294996.3295074.
Khoshouei, M., Bagherpour, R., Jalalian, M.H., 2022. Rock type identification using analysisof the acoustic signal frequency contents propagated while drilling operation. Geotech. Geol. Eng. 40 (3), 1237–1250. https://doi.org/10.1007/s10706-021-01957-y.
Korstanje, J., 2021. Gradient boosting with XGBoost and LightGBM. In: Advanced Forecasting with Python: With State-of-the-Art-Models Including LSTMs, Facebook's Prophet, and Amazon's Deepar. Springer, Berkeley, CA, pp. 193–205. https://doi.org/10.1007/978-1-4842-7150-6_15.
Li, H., Liu, Q., Chen, J., et al., 2025a. A comprehensive prediction framework for offshore downhole collapse pressure based on machine learning and multi-attribute decision analysis: insights from the East China Sea. Eng. Anal. Bound. Elem. 180, 106510. https://doi.org/10.1016/j.enganabound.2025.106510.
Li, H., Liu, Q., Chen, J., et al., 2025b. Predicting offshore downhole rock drillability in complex geological conditions using explainable machine learning and multi-source data. Earth Sci. Inf. 18 (4), 510. https://doi.org/10.1007/s12145-025-02002-9.
Li, H., Liu, Q., Zhu, S., et al., 2025c. Predicting downhole rock friction angles in complex geological settings: machine learning approaches and application to the Xihu sag. Eng. Anal. Bound. Elem. 177, 106279. https://doi.org/10.1016/j.enganabound.2025.106279.
Li, H., Tan, Q., Deng, J., et al., 2023a. A comprehensive prediction method for pore pressure in abnormally high-pressure blocks based on machine learning. Processes 11 (9), 2603. https://doi.org/10.3390/pr11092603.
Li, H., Tan, Q., Li, B., et al., 2023b. Physically-data driven approach for predicting formation leakage pressure: a dual-drive method. Appl. Sci. 13 (18), 10147. https://doi.org/10.3390/app131810147.
Li, J., Zhao, J., Hou, Z., et al., 2021a. Origins of overpressure in the central Xihu depression of the East China Sea shelf basin. AAPG Bull. 105 (8), 1627–1659. https://doi.org/10.1306/02262118112.
Li, Z., Li, G., Yu, H., et al., 2022. Fracability evaluation based on the three-dimensional geological numerical simulation of in situ stress: Case study of the Longmaxi Formation in the Weirong shale gas field, southwestern China. Math. Geosci. 54 (6), 1069–1096. https://doi.org/10.1007/s11004-022-10001-5.
Li, Z.Q., Chen, J.L., Zou, H., et al., 2021b. Mesozoic–Cenozoic tectonic evolution and dynamics of the Songliao Basin, NE Asia: implications for the closure of the Paleo-Asian Ocean and Mongol-Okhotsk Ocean and subduction of the Paleo-Pacific Ocean. Earth Sci. Rev. 218, 103471. https://doi.org/10.1016/j.earscirev.2020.103471.
Liang, J., Luo, M., Li, W., et al., 2025. Machine learning enhanced formation pressure prediction using integrated well logging and mud logging. Sci. Rep. 15, 23357. https://doi.org/10.1038/s41598-025-07048-9.
Liang, J., Wang, H., 2019. Cenozoic tectonic evolution of the East China Sea Shelf Basin and its coupling relationships with the Pacific Plate subduction. J. Asian Earth Sci. 171, 376–387. https://doi.org/10.1016/j.jseaes.2018.08.030.
Liang, X., Chen, S., Ma, B., et al., 2024. Changes of tectonic regime of the East China Sea Shelf Basin since Mesozoic: Insights from the Tiantai slope belt, East China Sea. J. Asian Earth Sci. 259. https://doi.org/10.1016/j.jseaes.2023.105900.
Lin, J., Dong, C., Lin, C., et al., 2024. The impact of tectonic inversion on diagenesis and reservoir quality of the Oligocene fluvial sandstones: The upper Huagang Formation, Xihu Depression, East China Sea Shelf Basin. Mar. Petrol. Geol. 165, 106860. https://doi.org/10.1016/j.marpetgeo.2024.106860.
Liu, J., Kang, S., Shen, W., et al., 2020. Petrology and hydrocarbon significance of the coaly source rocks from the Pinghu Formation in the Xihu Sag, East China Sea Shelf Basin. Energy Explor. Exploit. 38 (5), 1295–1319. https://doi.org/10.1177/0144598720912344.
Liu, K., Ma, T., Zhou, X., et al., 2023. Probabilistic prediction method of formation pore pressure. In: 57$^{th}$ U.S. Rock Mechanics/Geomechanics Symposium. ARMA. https://doi.org/10.56952/arma-2023-0120.
Liu, S., Zhang, C., Zhu, R., et al., 2022. Thickness variation characteristics of tidal rhythmites: An example from the Pinghu Formation, Xihu Sag, East China Sea shelf basin. Front. Earth Sci. 10, 698061. https://doi.org/10.3389/feart.2022.698061.
Lockhart, L.P., Flemings, P.B., Nikolinakou, M., et al., 2023. Velocity-based pore pressure prediction in a basin with late-stage erosion: Delaware Basin. U.S. Mar. Petrol. Geol. 150, 106159. https://doi.org/10.1016/j.marpetgeo.2023.106159.
Lyu, C., Cai, W., Zhang, X., et al., 2026. CT 3D reconstruction and analysis of mesoscale damage and permeability evolution in salt rock under thermo-mechanical coupling. Nondestr. Test. Eval. 41 (6), 3272–3299. https://doi.org/10.1080/10589759.2025.2524605.
Ma, T., Zhang, Y., Qiu, Y., et al., 2022. Effect of parameter correlation on risk analysis of wellbore instability in deep igneous formations. J. Petrol. Sci. Eng. 208, 109521. https://doi.org/10.1016/j.petrol.2021.109521.
Mahmoodzadeh, A., Mohammadi, M., Ghafoor Salim, S., et al., 2022. Machine learning techniques to predict rock strength parameters. Rock Mech. Rock Eng. 55 (3), 1721–1741. https://doi.org/10.1007/s00603-021-02747-x.
McClay, K.R., Dooley, T., Whitehouse, P., et al., 2002. 4-D evolution of rift systems: Insights from scaled physical models. AAPG Bull. 86 (6), 935–959. https://doi.org/10.1306/61eedbf2-173e-11d7-8645000102c1865d.
Mgimba, M.M., Jiang, S., Nyakilla, E.E., et al., 2023. Application of GMDH to predict pore pressure from well logs data: A case study from southeast Sichuan Basin, China. Nat. Resour. Res. 32 (4), 1711–1731. https://doi.org/10.1007/s11053-023-10207-2.
Noshi, C.I., Eissa, M.R., Abdalla, R.M., 2019. An intelligent data driven approach for production prediction. In: Offshore Technology Conference. https://doi.org/10.4043/29243-ms.
Ogbamikhumi, A., Ebeniro, J.O., 2021. Reservoir properties estimation from 3D seismic data in the alose field using artificial intelligence. J. Petrol. Explor. Prod. 11 (3), 1275–1287. https://doi.org/10.1007/s13202-021-01105-5.
Prokhorenkova, L., Gusev, G., Vorobev, A., et al., 2017. CatBoost: Unbiased boosting with categorical features. arXiv:1706.09516 [cs.LG]. https://doi.org/10.48550/arXiv.1706.09516.
Quan, Y., Chen, Z., Jiang, Y., et al., 2022. Hydrocarbon generation potential, geochemical characteristics, and accumulation contribution of coal-bearing source rocks in the Xihu Sag, East China Sea Shelf Basin. Mar. Petrol. Geol. 136, 105465. https://doi.org/10.1016/j.marpetgeo.2021.105465.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536. https://doi.org/10.1038/323533a0.
Salcedo-Sanz, S., Deo, R.C., Carro-Calvo, L., et al., 2016. Monthly prediction of air temperature in Australia and New Zealand with machine learning algorithms.

Theor. Appl. Climatol. 125 (1), 13–25. https://doi.org/10.1007/s00704-015-1480-4.
Shen, W.C., Shao, L.Y., Zhou, Q.Y., et al., 2022. Geochemistry of argillaceous rocks of the Eocene Pinghu Formation from Xihu Depression in East China Sea basin and its geological significance. Acta Geol. Sin. 96 (6), 2078–2093. https://doi.org/10.19762/j.cnki.dizhixuebao.2022287 (in Chinese).
Song, X.Y., Liu, Y.T., Xue, L., et al., 2020. Time-series well performance prediction based on Long short-term memory (LSTM) neural network model. J. Petrol. Sci. Eng. 186, 106682. https://doi.org/10.1016/j.petrol.2019.106682.
Su, A., Chen, H., Wang, C., et al., 2013. Genesis and maturity identification of oil and gas in the Xihu Sag, East China Sea Basin. Petrol. Explor. Dev. 40 (5), 558–565. https://doi.org/10.1016/S1876-3804(13)60073-7.
Su, A., Chen, H., Zhao, J.X., et al., 2020. Natural gas washing induces condensate formation from coal measures in the Pinghu slope belt of the Xihu Depression, East China Sea Basin: Insights from fluid inclusion, geochemistry, and rock gold-tube pyrolysis. Mar. Petrol. Geol. 118, 104450. https://doi.org/10.1016/j.marpetgeo.2020.104450.
Tariq, Z., Elkatatny, S., Mahmoud, M., et al., 2017. A new approach to predict failure parameters of carbonate rocks using artificial intelligence tools. In: SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition. https://doi.org/10.2118/187974-ms.
Vapnik, V.N., Lerner, A.Y., 1963. Recognition of patterns with help of generalized portraits. Avtom. Telemekhanika 24 (6), 774–780.
Wang, H., Ma, J., Li, L., 2024. Pore pressure prediction based on rock physics theory and its application in seismic inversion. J. Appl. Geophys. 229, 105494. https://doi.org/10.1016/j.jappgeo.2024.105494.
Ward, C.D., Coghill, K., Broussard, M.D., 1994. The application of petrophysical data to improve pore and fracture pressure determination in north sea central graben HPHT wells. In: SPE Annual Technical Conference and Exhibition. https://doi.org/10.2118/28297-ms.
Wei, M., Qiu, D., Yue, T., et al., 2025. New method prediction of rock mechanics parameters in drilling process based on machine learning. Petrol. Sci. Technol. 43 (4), 360–383. https://doi.org/10.1080/10916466.2023.2298998.
Wen, X., Xie, Y., Wu, L., et al., 2021. Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with LightGBM and SHAP. Accid. Anal. Prev. 159, 106261. https://doi.org/10.1016/j.aap.2021.106261.
Xu, B., Wu, Z., Cheng, Y., et al., 2024. Cenozoic structural and tectonic evolution in the Western Xihu Basin, East China Sea Shelf Basin. Int. J. Earth Sci. 113 (3), 713–732. https://doi.org/10.1007/s00531-024-02396-9.
Yan, X., Zhang, M., Wu, Q., 2021. Big-data-driven pre-stack seismic intelligent inversion. Inf. Sci. 549, 34–52. https://doi.org/10.1016/j.ins.2020.11.012.
Yang, C., Han, B., Yang, C., et al., 2020. Mesozoic basin evolution of the East China Sea Shelf and tectonic system transition in Southeast China. Geol. J. 55 (1), 239–252. https://doi.org/10.1002/gj.3409.
Yang, C.H., Sun, P., Tian, C., et al., 2013. Distribution and formation mechanism of overpressure in Pinghu Formation, Xihu Sag, East China Sea. Offshore Oil 33 (3), 8–12. https://doi.org/10.3969/j.issn.1008-2336.2013.03.008 (in Chinese).
Yang, Y., Xie, X., Li, Y., et al., 2023. Formation and distribution of coal measure source rocks in the Eocene Pinghu Formation in the Pinghu Slope of the Xihu Depression, East China Sea Shelf Basin. Acta Oceanol. Sin. 42 (3), 254–269. https://doi.org/10.1007/s13131-023-2176-8.
Zhang, D., Ma, T., Liu, Y., et al., 2025. Heterogeneous multi-task learning approach for rock strength prediction in real-time during drilling based on mud log data. Geoenergy Sci. Eng. 246, 213608. https://doi.org/10.1016/j.geoen.2024.213608.
Zhang, J., Li, S., Suo, Y., 2016. Formation, tectonic evolution and dynamics of the East China Sea Shelf Basin. Geol. J. 51 (S1), 162–175. https://doi.org/10.1002/gj.2808.
Zhang, S.L., Zhang, J.P., Tang, X.J., et al., 2014. Geometry characteristic of the fault system in Xihu sag in East China Sea and its formation mechanism. Mar. Geol. Quat. Geol. 34 (1), 87–94. https://doi.org/10.3724/SP.J.1140.2014.01087 (in Chinese).
Zhao, G., Ding, W., Tian, J., et al., 2022. Spearman rank correlations analysis of the elemental, mineral concentrations, and mechanical parameters of the Lower Cambrian Niutitang shale: A case study in the Fenggang block, Northeast Guizhou Province, South China. J. Petrol. Sci. Eng. 208, 109550. https://doi.org/10.1016/j.petrol.2021.109550.
Zhao, H., Zhang, Z.W., Zhao, R.R., et al., 2024a. Three-pressure prediction method of jointing well-seismic data in JT1 well area of Sichuan Basin in China. Sci. Rep. 14 (1), 16044. https://doi.org/10.1038/s41598-024-67085-8.
Zhao, Z., Zhao, J., Zhao, H., et al., 2024b. Re-evaluation of the genetic mechanism of overpressure: western slope zone of the Xihu Depression, East China Sea Shelf Basin. Mar. Petrol. Geol. 160, 106599. https://doi.org/10.1016/j.marpetgeo.2023.106599.
Zhou, X., Xu, G., Cui, H., et al., 2020. Fracture development and hydrocarbon accumulation in tight sandstone reservoirs of the Paleogene Huagang Formation in the central reversal tectonic belt of the Xihu Sag, East China Sea. Petrol. Explor. Dev. 47 (3), 499–512. https://doi.org/10.1016/S1876-3804(20)60068-4.
Zhu, W., Zhong, K., Fu, X., et al., 2019. The formation and evolution of the East China Sea Shelf Basin: A new view. Earth Sci. Rev. 190, 89–111. https://doi.org/10.1016/j.earscirev.2018.12.009.