GeoStorage

Vol. 2, No. 2, p. 156-171, 2026

Research article

Fluid Identification in Tight Sandstone Reservoirs Using a Bayesian-Optimized CNN–BiLSTM Model

Zhiyang Zhang, Sinan Fang*, Shaman Li, Hao Ma

School of Geophysics and Petroleum Resources, Yangtze University, Wuhan 430100, China

Keywords:
Tight sandstone
fluid identification
CNN–BiLSTM
Bayesian optimization
gas reservoir

Cited as:
Zhang ZY, Fang SN, Li SM, et al. 2026. Fluid Identification in Tight Sandstone Reservoirs Using a Bayesian-Optimized CNN–BiLSTM Model. *GeoStorage*, 2(2), 156-171.
https://doi.org/10.46690/gs.2026.02.04

Abstract:
Fluid identification is a crucial component in the development of tight sandstone gas reservoirs, and its accuracy directly determines the reliability of reservoir evaluation and development planning. Due to the complex lithology of the reservoir, the highly overlapping logging response and the uneven distribution of sample categories, the traditional identification methods based on conventional cross-plot methods and conventional machine learning have limited performance in discriminating complex categories such as gas-water coexistence layers. Consequently, these methods struggle to meet the demands of precise fluid identification and efficient reservoir development. Therefore, a CNN-BiLSTM deep learning model based on Bayesian optimization is proposed for fluid type identification in tight sandstone reservoirs. Firstly, the composite parameters are constructed based on the original LAS logging data, the input feature dimension is expanded, and a sliding window sequence is formed to express the temporal change trend of fluid response. Secondly, convolutional neural network (CNN) and bidirectional long short-term memory network (BiLSTM) are combined to model local spatial and temporal features. At the same time, Focal Loss is introduced to improve the discrimination ability of the model for minority classes. Finally, Bayesian hyperparameter optimization is carried out by using the Optuna framework to obtain the optimal model structure and learning rate configuration. The results show that the optimized CNN-BiLSTM model has an accuracy of 93.5%, which has good identification ability and application prospect.

1 Introduction

With the continuous reduction of recoverable conventional oil and gas resources, the development of traditional oil and gas fields is becoming increasingly difficult, and the complexity of exploration and stimulation is increasing. The focus of oil and gas exploration and development is rapidly shifting towards more challenging unconventional oil and gas reservoirs, including low-permeability-ultra-low permeability reservoirs, high water cut reservoirs, low-resistivity oil and gas reservoirs, especially tight gas reservoirs. As an important part of unconventional oil and gas, tight sandstone gas reservoirs have a rich resource base, huge recoverable reserves and broad development potential. Studies have estimated that the global tight gas resources that can be recovered by technology reach about 5.4 $\times 10^4$Tcf (about 1.53 $\times 10^{15}$ m^3) (Dong et al., 2015). Therefore, tight gas has become an important growth point and a strategic hydrocarbon energy resource in the global natural gas industry.

Despite the huge potential of tight gas resources, their development is also difficult. Taking the L block of Ordos Basin as an example, the reservoir in this area is a typical tight sandstone gas reservoir with low porosity (<10%), ultra-low permeability (<0.1 mD) and strong heterogeneity. The complex diagenesis results in diverse lithological assemblages of reservoirs, complex pore throat structures, diverse conductive mechanisms, and extremely irregular distribution of gas-water interfaces (Li, 2025). In this context, the water layer has a significant impact on the production capacity of the gas layer: water breakthrough can cause water-block effects, increase the water cut, and ultimately reduce gas production, which seriously restricts the develop-

Yandy Scientific Press

* Corresponding author.
E-mail address: 2023720502@yangtzeu.edu.cn (Zhang ZY); fangsinan@163.com (Fang SN); a17372820906@163.com (Li SM); 3165657704@qq.com (Ma H)

Received Mar. 10, 2026; revised Apr. 1, 2026; accepted Apr. 24, 2026; available online May. 3, 2026.

ment effect and economic benefits of tight gas reservoirs.

Therefore, accurate identification of fluid properties in reservoirs has become a key prerequisite for the exploration and development of tight gas reservoirs. For tight sandstone reservoirs, fluid identification mainly relies on logging data, and the distinction between gas, water and dry layers is achieved through multi-parameter curve response characteristic analysis. This process is not only related to the identification accuracy of effective reservoirs, but also directly affects the well location deployment, fracturing renovation design and capacity prediction results, and is a key link to guide the formulation of oil and gas field development plans.

However, under complex geological conditions, the difference between the response of the gas layer and the water layer in the conventional logging curve is often significantly weakened. It is difficult for traditional empirical methods and single-parameter discrimination techniques to effectively differentiate between various fluid types, which not only increases the uncertainty of reservoir interpretation, but also reduces the scientific validity of development decisions. How to achieve high-precision fluid identification in the context of multi-cause and multi-scale heterogeneity has become one of the core scientific problems restricting the efficient development of tight gases.

Fluid identification runs through the whole process of exploration, evaluation and production of tight gas reservoirs, and is an important link between geological understanding and engineering practice. It directly determines the reliability of reservoir gas content determination, capacity prediction and fracturing transformation evaluation, which is of decisive significance for optimizing the development plan, reducing exploration risks and improving economic benefits. Therefore, accurate fluid identification is not only fundamental to understanding reservoir petrophysics and fluid distribution, but also serves as a critical technical prerequisite for the efficient development of tight gas.

At present, the methods for fluid identification in tight sandstone reservoirs can be divided into three categories:

(1) Empirical discrimination methods based on conventional logging data. This type of method is used to distinguish between conventional logging data by means of conventional cross-plots and overlapping curves (Bai et al., 2022), and the principle is intuitive and widely applied. However, in tight sandstone reservoirs with low porosity and low permeability and high overlap in logging response, the difference in curve characteristics is weak and the identification effect is limited.

(2) Basic shallow machine learning method. With the development of artificial intelligence, algorithms such as support vector machines, random forests, and gradient lifting trees have been introduced into fluid recognition research (Fang et al., 2020; Lei et al., 2025; Luo et al., 2022; Yan et al., 2012; Yu et al., 2005; Zhang et al., 2022; Zhao et al., 2025, 2018). They can partially establish nonlinear relationships and improve accuracy, but they still rely on artificial feature construction, which is difficult to capture the temporal and spatial correlation of high-dimensional logging data, and the generalization ability is limited.

(3) In recent years, the development of deep learning technology has provided a new way for fluid identification in complex reservoirs. Typical models include convolutional neural networks (CNNs), recurrent neural networks (RNNs), autoencoders, and transformers (Gu et al., 2018; Lim et al., 2021; Lindemann et al., 2021). These models can automatically extract high-dimensional features and establish nonlinear mapping relationships, which significantly improves the recognition accuracy compared with traditional methods. However, tight sandstone reservoirs generally have the characteristics of small pore throat structure, strong heterogeneity, weak difference in logging response and serious overlap. Although CNN can effectively extract local spatial features, its convolutional kernel is dominated by local slippage, and its sensitivity to deep spatial continuity and multi-scale response is insufficient, making it difficult to characterize lithological abrupt changes or interlayer fluid transition characteristics. Although RNN and its variants have the ability of temporal modeling, they are prone to gradient attenuation in deep networks, making it difficult to capture the long-range dependence of logging curves across layers. Although structures such as autoencoders and transformers have stronger global expression capabilities, they have high requirements for sample size and data equilibrium, making it difficult to stably converge in tight gas data with uneven sample distribution and scarce labels.

In view of this, this paper adopts a hybrid structure combining CNN and bidirectional long short-term memory network (BiLSTM) to give full play to the advantages of CNN in local spatial feature extraction, and at the same time uses BiLSTM to model the bidirectional temporal dependence of logging curves to achieve cooperative characterization of spatial and temporal features. Furthermore, in order to overcome the subjectivity and randomness of network hyperparameter settings, Bayesian optimization algorithm is introduced to adaptively adjust the model hyperparameters, so as to improve the stability and generalization performance of the model, and enhance its ability to identify complex reservoir fluid distributions.

In conclusion, due to its complex lithology, delicate pore throat structure, and serious overlap in logging responses, the traditional empirical method and shallow machine learning method have problems such as insufficient feature extraction, insufficient characterization of timing relationship, and limited model generalization ability in fluid identification. In order to solve this series of challenges, based on the systematic investigation of existing methods, this paper proposes a deep learning model combining convolutional neural network and bidirectional long short-term memory network to realize the co-modeling of spatial and temporal features of logging curves. At the same time, Bayesian optimization algorithm is introduced to adaptively adjust the network hyperparameters, so as to improve the stability and prediction accuracy of the model. This method effectively alleviates the problems of weak fluid response difference and class imbalance in tight sandstone reservoirs, and provides a feasible new idea for high-precision fluid identification in complex reservoirs.

2 Geological overview

The Ordos Basin is located at the junction of the stable massif in eastern China and the active tectonic belt in the west, with a

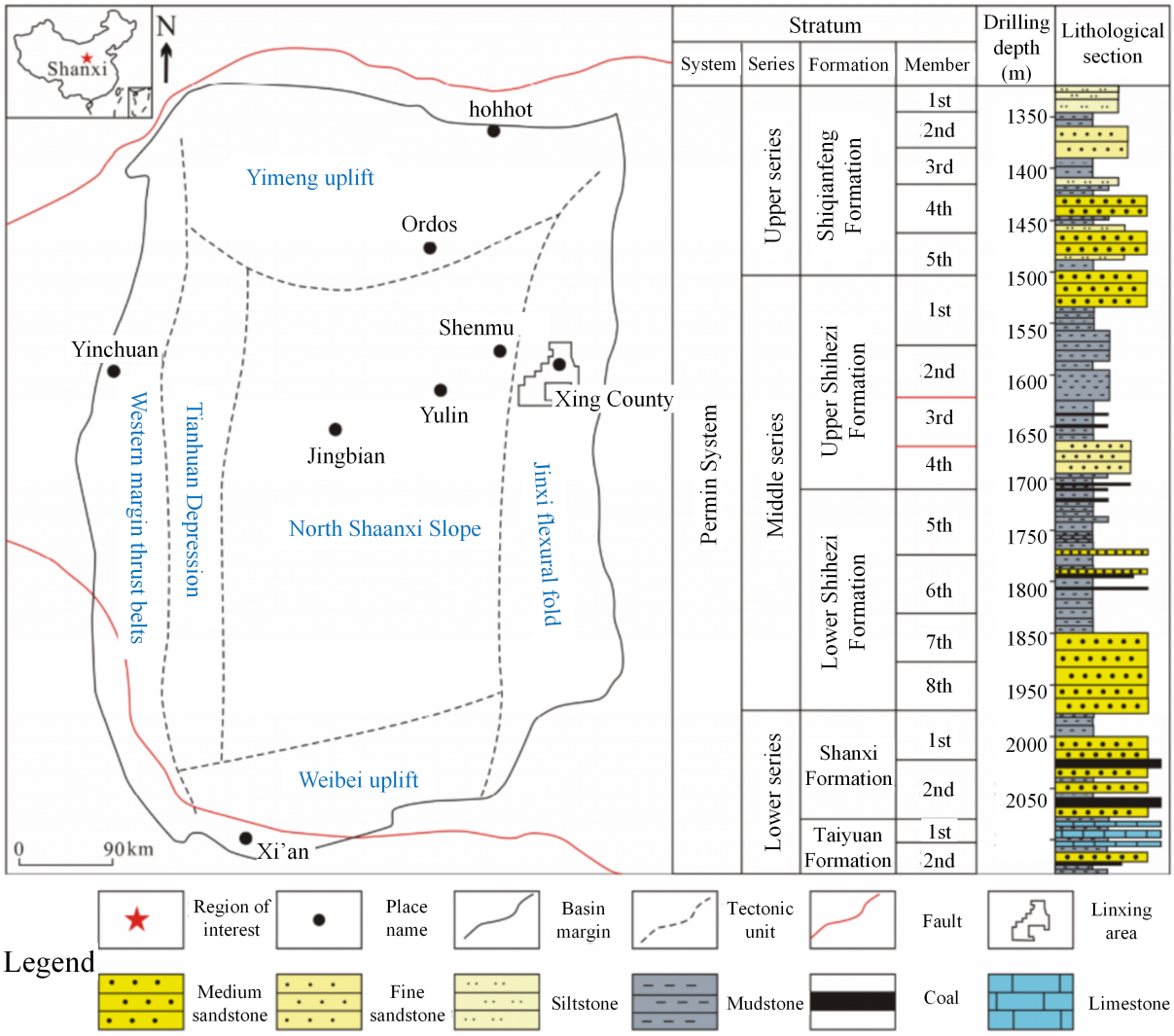

Fig. 1 Structural location of the study area within the Ordos Basin

rectangular syncline of north-south trending and steep east and west, surrounded by orogenic belts. The basin is mainly divided into six tectonic units, including the Yimeng uplift, the Weibei uplift, the western margin thrust zone, the Tianhuan depression zone, the Yishan slope belt and the western Shanxi deflection fold zone.

The L block of the study area is located in the transition zone between the western Shanxi deflection zone and the Yishan slope on the eastern margin of the basin, and the administrative division belongs to X County and L County in the western part of Shanxi Province. From bottom to top, the Carboniferous Benxi Formation, Permian Taiyuan Formation, Shanxi Formation, Xiashihe Formation, Shangshihe Formation and Shiqianfeng Formation are mainly developed, and high-quality source rocks such as coal-based and mudstone are large, abundant, gas-generating intensity and long hydrocarbon generation period, which provides favorable conditions for sufficient gas supply in the reservoir.

The lithological assemblage of the reservoir is diverse, mainly feldspar lithic sandstone, lithic feldspar sandstone, lithic sandstone, local development of lithoclastic quartz sandstone and quartz sandstone, the content of volcanic rock and pyroclastic material in the clastic components is high, the interstitial material is mainly clay, sericite and calcite, and the conductive mineral content is low. The rock structure is dense, the sortability is medium, the particle support is relatively developed, the secondary enlargement phenomenon of quartz is common, the degree of rock weathering alteration is deep, the particle size is mainly sub-rounded, and the reservoir space is dominated by residual intergranular pores, dissolving pores and local fracture pores.

Among them, the sand body of the channel and the tributary channel is the main high-quality reservoir, while the physical properties of the argillaceous interlayer and fine-grained sedimentation between the channels are poor, forming obvious physical property mutations and reservoir segmentation. This multiphase sedimentary superposition structure causes frequent changes in the pore throat structure inside the reservoir, complex conductivity mechanism, and uneven gas and water distribution, resulting in weak differences and serious overlap in logging response characteristics, which significantly increases the difficulty of fluid identification and reservoir evaluation (Mi

et al., 2022; Zhu et al., 2022).

Due to the coupling of tectonic activity, sedimentary system evolution and diagenesis, the reservoirs in the study area show strong heterogeneity in both lateral and longitudinal directions: the Shangshi Box Formation develops composite sedimentary systems such as meandering river-braided river delta and lagoon-tidal flat-barrier coast, resulting in complex sand body distribution, large physical property fluctuations, wide transition bands, and significantly weakened logging response differences. The porosity of typical tight sandstone reservoirs is mostly about 5%–10%, and the permeability is often less than 0.5 mD, and only a few high-quality sand bodies can reach 1 mD.

At the methodological level, recent studies have also shown that using only a single network (pure CNN or pure RNN) is prone to the problem of insufficient feature expression or long-term dependent modeling instability in dense sandstone scenarios with "response superposition, class imbalance, and low signal-to-noise ratio". In contrast, the progress of deep learning in multiple disciplines (computer vision, speech, and natural language) demonstrates that CNNs excel at automatically extracting multi-scale local features, while BiLSTMs are highly effective at capturing bidirectional sequential dependencies.

To solve these problems, the deep learning model based on the fusion of convolutional neural network (CNN) and bidirectional long short-term memory network (BiLSTM) proposed in this paper can learn the local spatial pattern and longitudinal timing law of the logging curve without over-relying on manual features, and alleviate the recognition bias caused by response superposition and category imbalance. Combined with Bayesian optimization, the network hyperparameters are further adaptively searched to reduce the uncertainty of manual parameter adjustment, improve generalization and stability, so as to achieve higher accuracy fluid identification of complex tight sandstone reservoirs.

3 Bayesian-optimized CNN–BiLSTM model

On the basis of extracting multi-dimensional physical property features, the convolutional neural network and bidirectional long-short-term memory network are used to fuse the structure for fluid identification, and the Bayesian optimization algorithm is combined to achieve hyperparameter adaptive optimization, so as to achieve efficient and accurate identification of fluid types in tight sandstone reservoirs.

3.1 Workflow overview

Tight sandstone fluid identification requires on-site logging personnel to collect data, but there are often problems such as missing and inconsistent units, which will lead to various errors in the data, so it is necessary to preprocess LAS files. After the data processing is completed, the optimal feature combination is screened out through analysis and experiments, and the deep learning model is combined with the deep learning model to achieve efficient identification of tight sandstone fluid types (the flow is shown in Fig. 2).

Firstly, a variety of physical property parameters are extracted based on the LAS logging file, including density, acoustic time difference (DTC, DTS), neutron porosity, resistivity, natural gamma, muddy content, porosity, permeability, saturation, Young's modulus, etc. When examining the original data, it was found that the curves of different wells were often missing or local sections were missing, and the units were not uniform. If the errors caused by these engineering factors are not reasonably handled, it will affect the reliability of model training and prediction. The LAS file used in this paper has a sampling interval of 0.05 m, while traditional fluid identification methods are usually based on the average of a certain layer, such as the average of about 40 0.05 m interval points for a reservoir with a thickness of 2 m, which may lead to local information loss. Therefore, this paper uses a sliding window to structure the logging sequence, retains the longitudinal trend and local continuity, and generates a time series sample suitable for the deep learning model, so as to retain the information of the original data as much as possible.

After data processing, the combined structure of convolutional neural network and bidirectional long short-term memory network is explored: CNN is used to extract local spatial features of logging curves, and BiLSTM is used to capture the contextual dependencies of logging sequences.

At the same time, in order to further improve the modeling efficiency and performance stability, Bayesian optimization algorithm is used to automatically optimize key hyperparameters (such as network layer number, convolutional kernel size, learning rate, etc.) based on the Optuna framework, avoiding the blindness and inefficiency of traditional manual parameter adjustment.

3.2 Model architecture

Building upon the CNN, BiLSTM, and Bayesian optimization strategies introduced earlier, this study constructs a hybrid deep neural network that integrates Convolutional Neural Networks with Bidirectional Long Short-Term Memory units. In addition, a Focal Loss function is incorporated to enhance the model's sensitivity to minority classes. The overall architecture is illustrated in Fig. 3.

Based on the CNN, BiLSTM and Bayesian optimization methods introduced above, a deep neural network model integrating CNN and BiLSTM is constructed, and Focal Loss is introduced to enhance the recognition ability of minority samples. The overall structure of the model is shown in Fig. 3. The input data is constructed into a logging sequence sample through a sliding window, and the features include conventional logging parameters and some combined features, and all features are standardized. The CNN module extracts local spatial features, and BiLSTM captures the bidirectional timing dependence to realize the joint modeling of local and global information. To prevent overfitting, a Dropout layer is added after the BiLSTM output and the final classification is completed through the fully connected layer. The model consists of 1 convolutional layer, 1 bidirectional LSTM layer, and 2 fully connected layers, with a Dropout in the middle to reduce the risk of overfitting. The convolutional layer is responsible for extracting the local spatial features of the logging curve, and the BiLSTM layer is used to capture the temporal dependence of the depth direction,

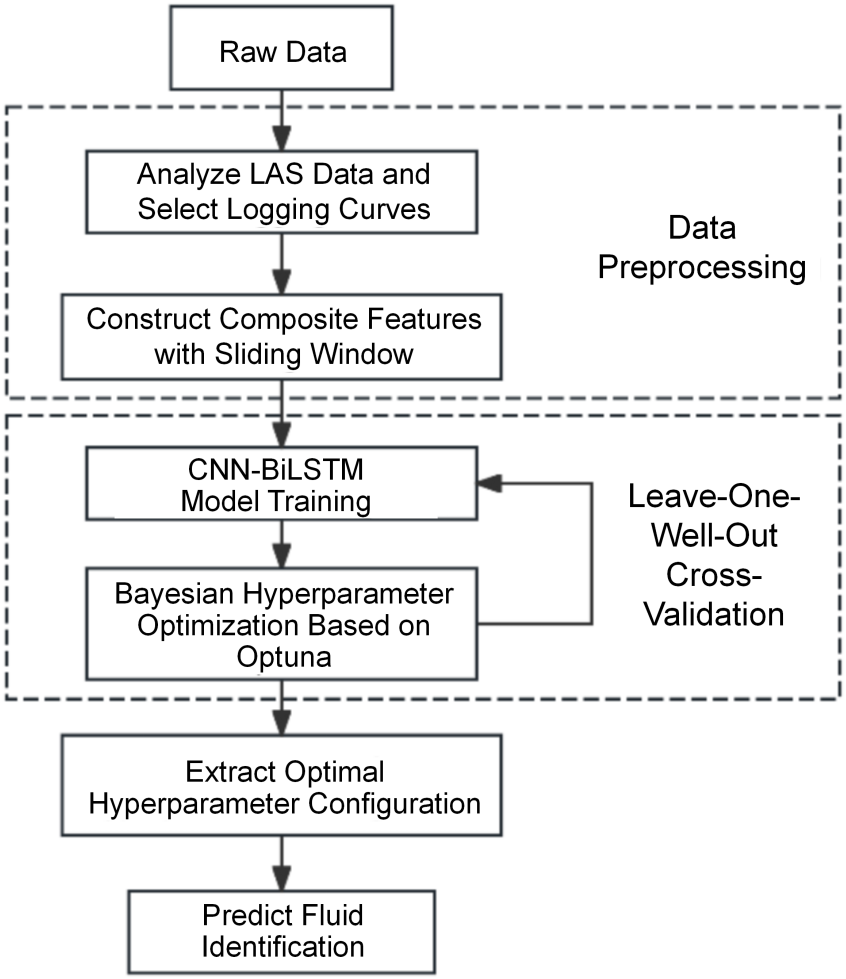

Fig. 2 Overall workflow of the proposed methodology

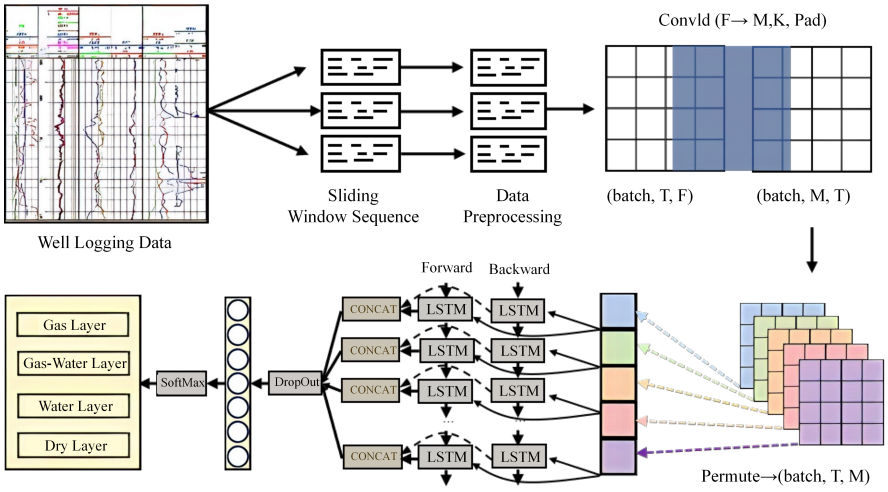

Fig. 3 Overall architecture of the proposed CNN–BiLSTM model

and finally the prediction probability of the four types of fluid types is output through the fully connected layer. Focal Loss was used in training to mitigate category imbalance, and Adam was selected as the optimizer. To further improve the model performance, Bayesian optimization is implemented using the Optuna framework to automatically search and tune key hyperparameters.

3.3 Convolutional neural network (CNN)

After completing the overall modeling process design, it

is necessary to efficiently extract the local spatial features of the logging curve. Convolutional neural networks are a deep learning model that excels in extracting local spatial features, which were initially applied to the field of image processing and have also shown good performance in geophysical data mining in recent years (Liao et al., 2020). Since logging curves often show obvious geological response patterns in small-scale depth intervals (such as high GR and low porosity in dry layers, low GR and high porosity in gas layers), CNN can effectively identify these local features through local sensing mechanisms and weight-sharing structures.

In this study, CNN was used for spatial dimensional feature extraction of logging sliding window sequences. Firstly, the input data is organized into tensors according to the window size, and then the local coupling relationship and trend change between each channel are extracted through the one-dimensional convolutional layer.

3.4 Bidirectional Long Short-Term Memory Network (BiLSTM)

However, relying solely on CNN to extract spatial features is not enough to characterize the dynamic law of logging curves with depth. In order to capture the temporal dependencies of curve data and maintain longitudinal continuity, a bidirectional long-term short-term memory network model is introduced. BiLSTM can simultaneously model sequence features in both front and rear directions, which is suitable for characterizing fluid response correlations between different depth points, thereby further improving the identification accuracy of complex reservoirs. In recent years, many researchers in the field of logging have used deep learning as a new method of machine learning, which has strong feature extraction capabilities, especially when processing timing data, and the use of recurrent neural networks is more effective. In the early days, RNN and LSTM were used, but RNN had problems such as long-term dependence, gradient disappearance or gradient explosion in processing time series, and the researchers proposed LSTM to alleviate these problems to a certain extent, but its structure can only use the front and bottom information, and there are still modeling limitations.

In order to more effectively model the sequence characteristics in logging data, Bidirectional Long Short-term Memory Network (BiLSTM) is used as the main classification model. As shown in Figure 3, traditional forward-facing LSTM models can only use contextual information from the past to the present (Chen et al., 2023), while BiLSTM can obtain the contextual information of any time step in the sequence at the same time by training two independent LSTM networks in two directions, namely forward and backward, to learn the dependencies between temporal features more comprehensively.

In this paper, the BiLSTM network takes a sliding window sequence of logging data as input, with multiple logging curve features in each time step. The model first encodes the input sequence through a bidirectional LSTM layer, and outputs the stitching results of the forward and reverse hidden states.

Let the input sequence be $X = \{x_1, x_2, \ldots, x_T\}$, where $x_t \in \mathbb{R}^d$ denotes the d-dimensional feature vector at depth step t. The BiLSTM generates its output by combining the hidden states from both the forward and backward temporal directions. At each depth step t, the forward LSTM computes the hidden state $\overrightarrow{h}_t$ and the backward LSTM computes the hidden state $\overleftarrow{h}_t$.

$$\overrightarrow{h}_t = \mathrm{LSTM}_{\mathrm{forward}}(x_t, \overrightarrow{h}_{t-1}) \tag{1}$$

$$\overleftarrow{h}_t = \mathrm{LSTM}_{\mathrm{backward}}(x_t, \overleftarrow{h}_{t+1}) \tag{2}$$

$$h_t = [\overrightarrow{h}_t \oplus \overleftarrow{h}_t] \tag{3}$$

where $\bigoplus$ denotes the concatenation operation, and h_t is the combined output hidden state at step t that integrates both past and future contextual information. The basic $LSTM$ unit controls the information flow via an input gate i_t a forget gate f_t and an output gate o_t effectively mitigating the vanishing gradient problem.

Although the LAS data are sampled by depth, the depth series is highly similar to the time series: there is a strong correlation between adjacent sampling points, and the continuity of stratigraphic sedimentation makes logging curves usually show a significant longitudinal trend. This feature enables BiLSTM to give full play to its advantages in modeling: its bidirectional structure can integrate global information in the temporal dimension while highlighting the feature expression of key positions, considering both the response characteristics of local layers and the correlation between upper and lower layers. This method is especially suitable for modeling nonlinear and nonstationary features in geological sequences, and has good adaptability in reservoir fluid identification tasks in complex geological backgrounds.

3.5 Bayesian optimization and the Optuna framework

The performance of deep learning models depends not only on the network structure design but also on the hyperparameter settings. Traditional manual parameter adjustment or grid search methods are often time-consuming and easy to fall into local optimum. In order to improve the efficiency and stability of the model, Bayesian optimization algorithm is introduced under the Optuna framework to realize hyperparameter adaptive global search and automatic tuning. Bayesian optimization is an efficient method for optimizing black-box functions (Cui & Yang, 2018), especially in scenarios where functions cannot be resolved, underivable, or expensive to evaluate. In the field of machine learning, it is widely used for hyperparameter tuning, such as the learning rate of neural networks, the number of hidden units, and the batch size. The core idea is to guide the selection of new sampling points by establishing the posterior probability distribution of the objective function in the case of high sampling cost, so as to find the optimal solution with fewer evaluations.Let the objective function be $f(\theta)$ where θ denotes

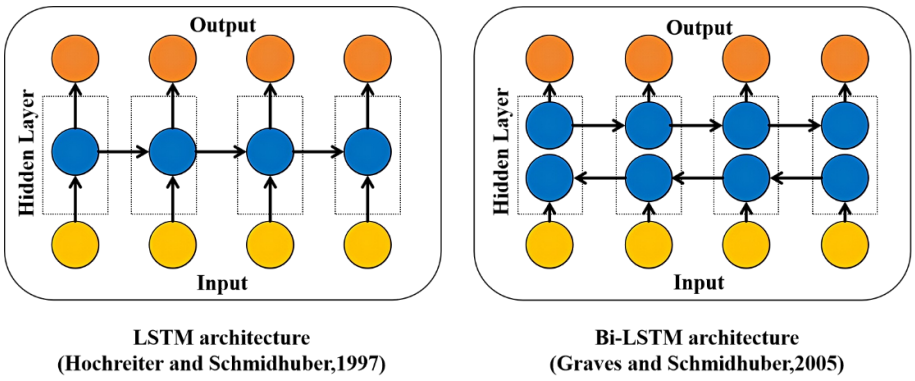

Fig. 4 Structural comparison of the LSTM and BiLSTM models

a candidate point in the hyperparameter space Θ. The goal of Bayesian optimization is to find

$$\theta^* = \text{argmax}_{\theta \in \Theta} f(\theta) \tag{4}$$

A surrogate model—commonly a Gaussian Process (GP)—is used to approximate the behavior of $f(\theta)$, expressed as $f(\theta) \sim \mathcal{GP}(\mu(\theta), k(\theta, \theta'))$ where $\mu(\theta)$ is the mean function and $k(\theta, \theta')$ is the covariance kernel. Based on the surrogate model, Bayesian optimization estimates both the predicted mean and uncertainty at any point in the hyperparameter space. These estimates are then used to construct an acquisition function $\alpha(\theta)$, which balances exploration (evaluating uncertain regions) and exploitation (refining promising regions). The next sampling point is selected by

$$\theta_{next} = \text{argmax}_{\theta}\, a(\theta|D) \tag{5}$$

where D denotes the set of previously evaluated samples. Common acquisition functions include Expected Improvement (EI), Probability of Improvement (PI), and Upper Confidence Bound (UCB).

In this study, Bayesian optimization is implemented through the Optuna framework. Optuna is a flexible and efficient open-source hyperparameter optimization tool that uses Tree-structured Parzen Estimator (TPE) as a proxy model and has built-in dynamic sampling and intelligent pruning mechanisms, which can terminate poorly performing parameter combinations in time during the search process and allocate more computing resources to regions with higher potential, thereby accelerating the convergence process and improving model performance and cross-well generalization capabilities.

4 Logging dataset analysis and preprocessing

4.1 Data preprocessing

In this study, the measured logging data of Block X of tight sandstone gas reservoir in the Ordos Basin were selected as experimental data, and a total of 134 wells were counted in the area, covering different tectonic sites and reservoir types. For each well, the conventional logging curve data is collected and integrated, and finally 15 standardized logging characteristics are selected as model input variables, covering key logging response parameters such as resistivity, density, porosity, elastic modulus and acoustic time difference, so as to fully reflect the electrical, permeable and lithological characteristics of the reservoir. The list of characteristic parameters and their physical meanings is shown in Table 1.

The target labels for model training were determined based on well-testing data, expert interpretation, and integrated geological evidence. Using drilling reports, flowback curves, well-testing summaries, and expert-reviewed interpretation tables, reservoir intervals were categorized into four fluid types: gas layer, gas–water coexistence layer, water layer, and dry layer.

In this paper, the logging curves are serialized by sliding window method, and each sliding window sample is composed of multi-dimensional logging features at several consecutive depth points. The window length and step length are optimized by considering the distribution characteristics of logging data and the experimental verification results, so as to achieve a balance between retaining longitudinal information and controlling sample redundancy. In view of the problems of defects, noise and unit inconsistencies in the logging curves of each well, the following treatments were made:

1. Missing value filling: Forward filling (ffill) and interpolation method are used to ensure the continuity of the input sequence.

2. Standardization processing: Z-score standardization of all features uniformly eliminates dimensional differences; For example, the porosity (POR) range is 0.02 ~ 0.30, and the density (ZDEN) range is 1.8 ~ 3.0. Without normalization, the model will focus more on features with large values (such as density) and ignore features with small but equally important values (such as porosity). After Z-score normalization, both curves are converted to a scale with a mean of 0 and a standard deviation of 1, making features more comparable.

3. Label screening: Eliminate sliding window samples with unclear labels or incomplete logging curves to improve the effectiveness and reliability of training samples.

4. Unified units: Standardize common logging curve units: acoustic time difference (DT) is unified as μs/ft, density (ZDEN) is unified as g/cm³, porosity (POR) is unified as decimal form (e.g., 0.2 instead of 20%), and mechanical parameters (e.g., Young's modulus, shear modulus) are unified as GPa. Through the unification of units, the data comparability of different wells and different logging conditions is ensured.

Tab. 1 Logging features and their physical significance

No.	Feature	Meaning	Physical Interpretation / Application
1	M2R1	Radial Resistivity Channel 1	Investigation closest to borehole; invasion & flushed zone
2	ZDEN	Formation Density	Reflects lithology and porosity variations
3	CNCF	Compensated Neutron Porosity	Indicates hydrogen content and fluid characteristics in pore space
4	GR	Gamma Ray	Identifies shale content and depositional environment
5	VSH	Shale Volume	Quantifies clay/shale content
6	POR	Porosity	Measures reservoir storage capacity
7	PERM	Permeability	Indicates fluid mobility within the reservoir
8	RQI	Reservoir Quality Index	Evaluates pore–permeability compatibility and reservoir quality
9	Swi	Irreducible Water Saturation	Represents the proportion of immobile water
10	Sw	Water Saturation	Reflects fluid type and distribution
11	CMPR	Compressibility Coefficient	Indicates rock compressibility and sensitivity to fluids
12	YMOD	Young's Modulus	Reflects rock stiffness and fracability
13	POIS	Poisson's Ratio	Describes rock elastic properties and brittleness
14	DTS	Shear Slowness	Used to distinguish lithology and fluid variations
15	DTC	Compressional Slowness	Used to compute elastic properties and infer pore characteristics

These preprocessing steps ensure that the sequential input data are continuous, standardized, and physically consistent, enabling more robust learning of spatial–temporal features within the CNN–BiLSTM framework.

4.2 Sample distribution and analysis of logging-response characteristics

As shown in Fig. 5, a total of 5,428 valid reservoir-interval samples were extracted from the study area. The dataset exhibits a strongly imbalanced class distribution. Among the four fluid categories, dry layers constitute the largest proportion with 2,284 samples (42.1%), followed by water layers with 1,488 samples (27.4%). Gas–water coexistence intervals account for 1,031 samples (19.0%), while gas layers represent the smallest class with only 625 samples (11.5%). This imbalance may cause the model to overfit the majority classes—particularly dry layers—during training, thereby reducing recognition accuracy for minority classes such as gas layers. To mitigate this issue, the Focal Loss mechanism is incorporated into the loss function to strengthen the model's ability to learn from underrepresented categories.

To further explore the differences in the logging response characteristics of different fluid types, Figure 6 shows the box plot distribution of the four types of fluids under the nine main logging parameters. The results show that there is a certain degree of distinction between different fluid types in some parameters, but in general, it still shows strong numerical overlap and transition characteristics, which further indicates that it is difficult to achieve high-precision fluid identification with traditional single-parameter discrimination.

In view of the differences in the sensitivity of different logging parameters to fluid response characteristics, this paper calculates the average difference to standard deviation ratio of each parameter under different fluid types, which is defined as follows:

$$S_f = \frac{|\mu_{\max} - \mu_{\min}|}{\sigma_{\text{all}}} \tag{6}$$

Where, S_f represents the separability index of feature f. The terms μ_{max} and μ_{min} denote the maximum and minimum mean values of that feature across the four fluid categories, respectively, while σ_{all} is the global standard deviation computed from all samples.

Given that different logging parameters exhibit varying sensitivity to fluid properties, a quantitative metric was introduced to evaluate feature separability across fluid categories. For each parameter, a separability score S_f was computed based on the ratio of inter-class mean difference to the overall standard deviation:

This analysis provides a quantitative foundation for feature selection and supports the use of the CNN–BiLSTM architecture to capture complex multivariate relationships that cannot be resolved through conventional methods.

The results are shown in Fig. 7. It can be observed that parameters such as formation density (ZDEN), porosity (POR), and water saturation indicators (S_w and S_{wi}) exhibit comparatively high sensitivity and stronger separability among fluid types. In contrast, acoustic slowness parameters (DTC, DTS) and resistivity-related attributes (CNCF, M2RX) display weaker inter-class differences. These results indicate that fluid responses in tight sandstone reservoirs are governed by multifactor interactions, and no single logging curve can adequately characterize gas–water distribution. Under the "low-porosity,

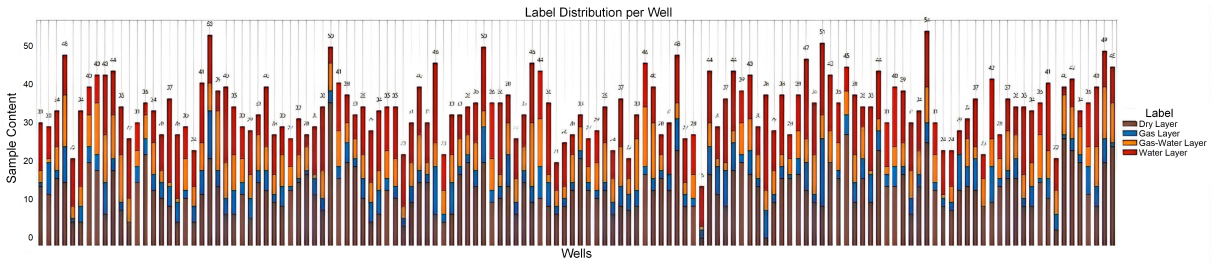

Fig. 5 Distribution of fluid-type labels for each well

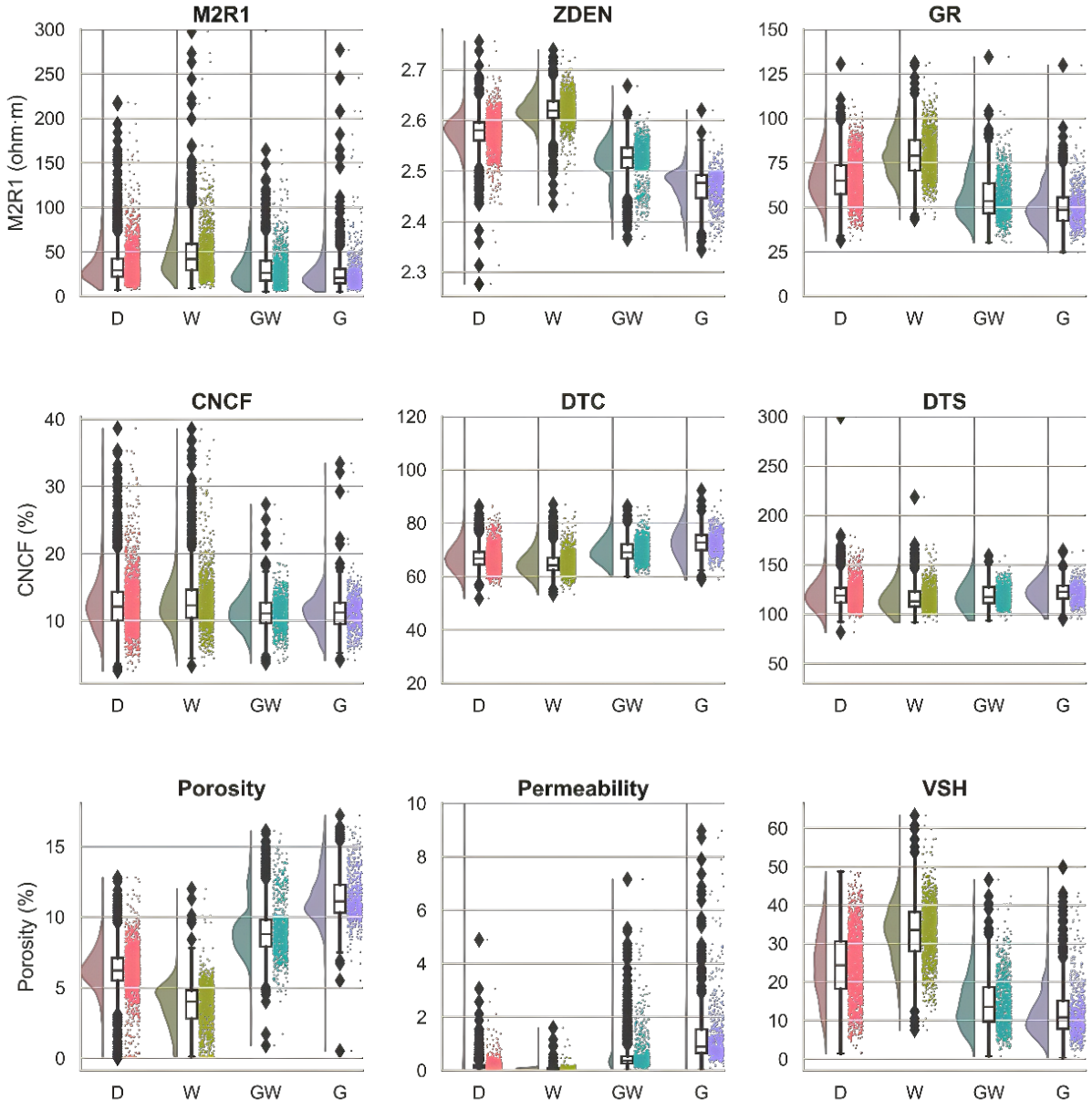

Fig. 6 Multi-parameter raincloud plots for different fluid types (Abbreviations: D – Dry layer; W – Water layer; GW – Gas–water layer; G – Gas layer)

low-permeability, strongly heterogeneous" conditions typical of tight sandstones, the interactions among multiple logging parameters become complex, further restricting the effectiveness of conventional single-curve interpretation methods.

Although the S_f indicator provides a quantitative measure of the discriminative power of individual logging parameters and

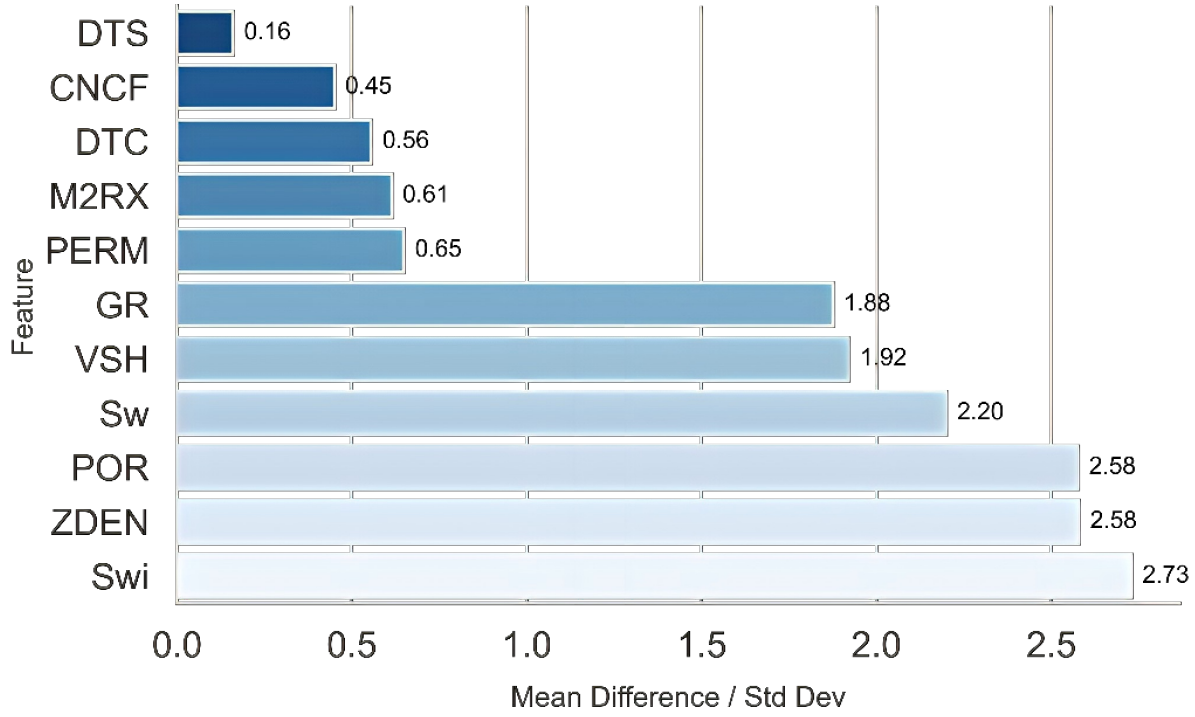

Fig. 7 Ratio of inter-class mean difference to overall standard deviation for different fluid types

helps identify key features contributing to fluid identification, evaluating feature separability from a single dimension cannot capture the joint distribution patterns of multiple features.

For a more intuitive analysis of the joint distribution of multiple features, Figure 8 shows the cross-plots of typical logging parameters. From the analysis results, it can be seen that although some fluid types have certain divisibility in a single feature or a combination of two, the overall logging response shows the characteristics of high overlap, significant transition zone and staggered sample distribution. This makes it difficult to accurately identify all fluid types, especially in complex reservoirs such as "gas-water co-layer".

In order to enhance the feature expression ability, several composite parameters are constructed from the physical mechanism:

Volume compression coefficient (volume modulus): This parameter is calculated by the combination of longitudinal wave time difference (DTC), shear wave time difference (DTS) and density (ZDEN), which can reflect the overall elastic response of rock skeleton and pore fluid. Studies have shown that the volumetric compression coefficient can effectively characterize the elastic difference between the gas layer and the water layer, reduce the misjudgment of the low resistance layer as the water layer, and improve the reliability of fluid identification (Zhao, 2024).

RQI (Reservoir Quality Index): Introducing physically constrained composite parameters such as RQI into the input features helps the model to learn the intrinsic coupling relationship between logging curves more accurately, improve its sensitivity and generalization ability to fluid distribution changes in complex reservoirs, and provide a more robust feature foundation for deep learning models.

5 Experimental design and results

5.1 Experimental design

5.1.1 Objective

The effectiveness and generalization ability of the BOA-CNN-BiLSTM model in the identification of dense sandstone fluids are verified, and compared with the traditional machine learning model to evaluate its performance stability and recognition accuracy under multi-well conditions.

5.1.2 Data sources and preprocessing

The experimental data were collected from a tight gas reservoir block in northern Shaanxi in the Ordos Basin, and logging data from 134 wells were collected for model training and prediction, covering 15 typical curve features such as resistivity, density, neutrons, acoustic waves, gamma, porosity, and permeability. In order to ensure the data quality, the original LAS file is systematically preprocessed: outliers are eliminated and the units are unified, the missing samples are repaired by median interpolation method, all features are standardized by Z-score, and finally a multi-classification sample set containing dry layer, gas layer, gas-water coexistence layer is established in combination with the interpretation result table.

5.1.3 Validation strategy

Model training was completed in a GPU environment, based on PyTorch 2.0 + CUDA 11.8 platform, with a batch size of 128 and a maximum training round (epoch) of 30. In order to comprehensively evaluate the generalization ability of the model, the leave-one-well-out (LOO) cross-validation method is adopted. The specific method is to select one well as the test

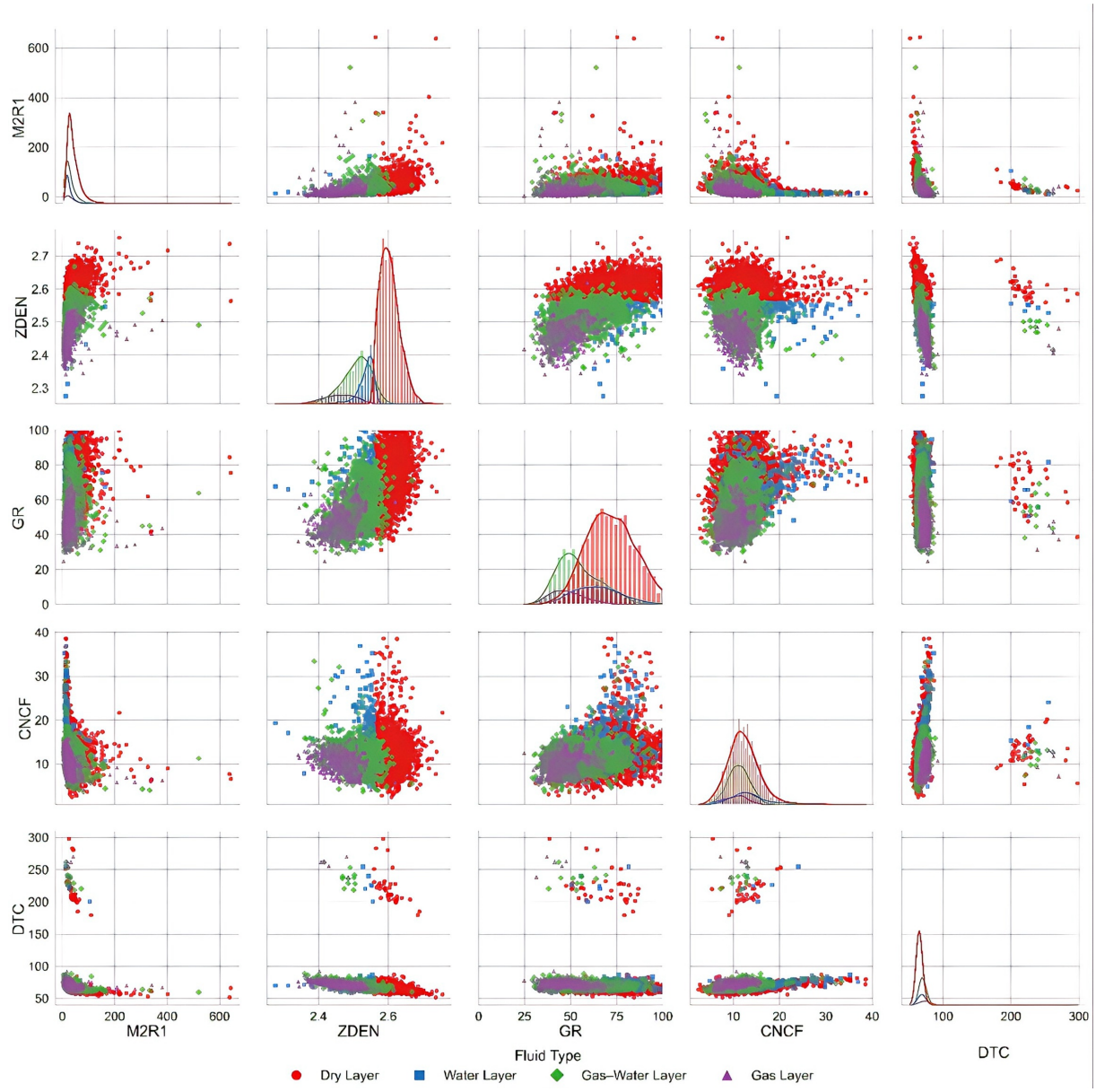

Fig. 8 Cross-plots of representative logging parameters

set and the rest of the wells as the training set, and cycle through all 134 wells, so that each well participates in an independent verification. This strategy can effectively test the stability and migration ability of the model under different geological differences between wells. In each round of verification, the accuracy, recall, F1 value and confusion matrix are calculated separately, and the average value of each round of results is taken to comprehensively evaluate the overall identification performance of the model under multi-well conditions.

5.2 Experimental results and analysis

5.2.1 Overall model performance

Across all 134 LOO validation rounds, the proposed CNN–BiLSTM model achieved an average accuracy of 93.5%. Notably, the recall rate for the gas–water coexistence category improved significantly compared with conventional approaches, indicating enhanced robustness and discriminative capability in identifying complex transitional intervals.

The results demonstrate that the BOA–CNN–BiLSTM model maintains strong and stable performance across multiple wells, with particularly notable improvement in the minority fluid categories. This confirms the model's suitability for fluid identification in heterogeneous tight sandstone reservoirs.

5.2.2 Confusion matrix analysis

The results of the confusion matrix show that there is a certain degree of confusion between the gas layer and the gas-water coexistence layer, mainly due to the strong overlap between the logging curve characteristics of the two types of samples, especially the acoustic time difference, density and neutron curve. In contrast, the CNN-BiLSTM model can better capture the continuous changes in the depth direction, making the boundary layer recognition smoother and the classification results more

Tab. 2 Performance of fluid-type classification

Class	Precision	Recall	F1-score
Dry Layer	0.994	0.985	0.990
Gas Layer	0.808	0.827	0.817
Gas–Water Coexistence	0.845	0.839	0.842
Water Layer	0.965	0.972	0.969
Overall Accuracy		0.935	

Tab. 3 Comparison of model performance and feature-learning capabilities

Model	Validation Method	Average Accuracy	Advantages
XGBoost	LOO by well	0.874	Simple structure, strong interpretability
CNN-BiLSTM	LOO by well	0.935	Strong generalization, effective spatial–temporal modeling
BiLSTM	LOO by well	0.927	Lacks local spatial feature extraction
CNN	LOO by well	0.914	Captures only spatial features, limited temporal awareness

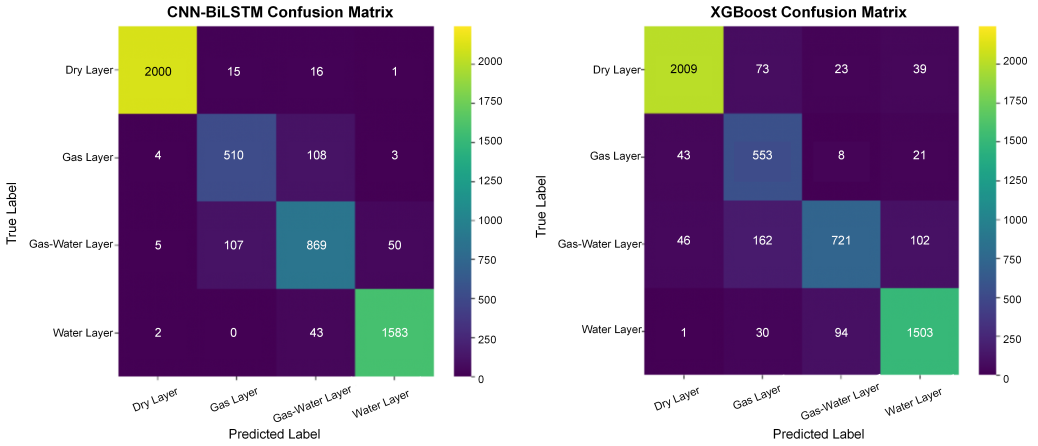

Fig. 9 Comparison of confusion matrices for fluid-type classification

stable.

5.2.3 Bayesian hyperparameter optimization and performance analysis

In order to strike a balance between recognition accuracy and computational efficiency, Bayesian optimization algorithm is used to adaptively optimize the key hyperparameters of the CNN-BiLSTM model.

In the optimization process, considering the complexity of the model in spatial-temporal feature extraction and the dependence of multi-dimensional parameters, key parameters such as sliding window length, number of convolutional channels, convolutional kernel size, number of BiLSTM hidden elements, dropout rate, learning rate, weight decay coefficient and batch size are selected as optimization variables.

In terms of experimental results, the benchmark model trained with the default hyperparameters (convolutional kernel size 3, convolutional channel number 64, hidden unit count 128, learning rate 1×10^{-3}, dropout rate 0.3) as a control showed an average accuracy of 0.902 and an overall F1 value of 0.885 under multi-well LOO cross-validation. After Bayesian optimization, the average accuracy is improved to 0.935, and the F1 value is increased to 0.930, and the overall performance is significantly improved compared with the benchmark model.

The optimized model has a particularly significant improvement in the recognition effect of a few class samples (such as gas layers and gas-water coexistence layers), which is about 3.7% and 4.4% higher than before optimization, indicating that Bayesian optimization effectively alleviates the recognition bias under the condition of class imbalance. The Precision, Recall and F1-score of each category are shown in Table 3, which further verifies the stability and robustness of the optimization strategy in the complex environment of multiple wells. Figure 9 shows the changes in the accuracy of the model before and after optimization for different fluid types, indicating that Bayesian optimization significantly improves the overall recognition performance, especially in the gas layer recognition task, showing higher sensitivity and generalization ability.

To further elucidate the mechanism of hyperparameter tuning to improve performance, Figure 11 shows the distribution of key parameters in the optimization process. The experimental

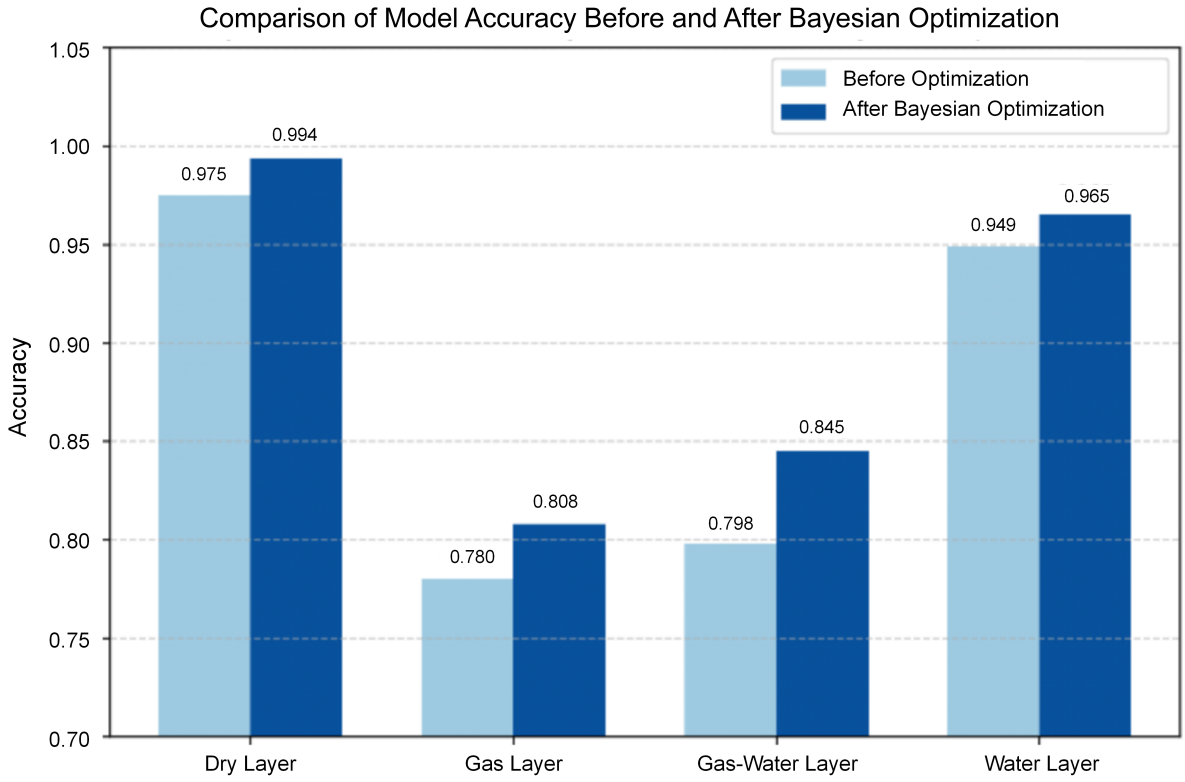

Fig. 10 Accuracy comparison of fluid-type predictions before and after Bayesian optimization

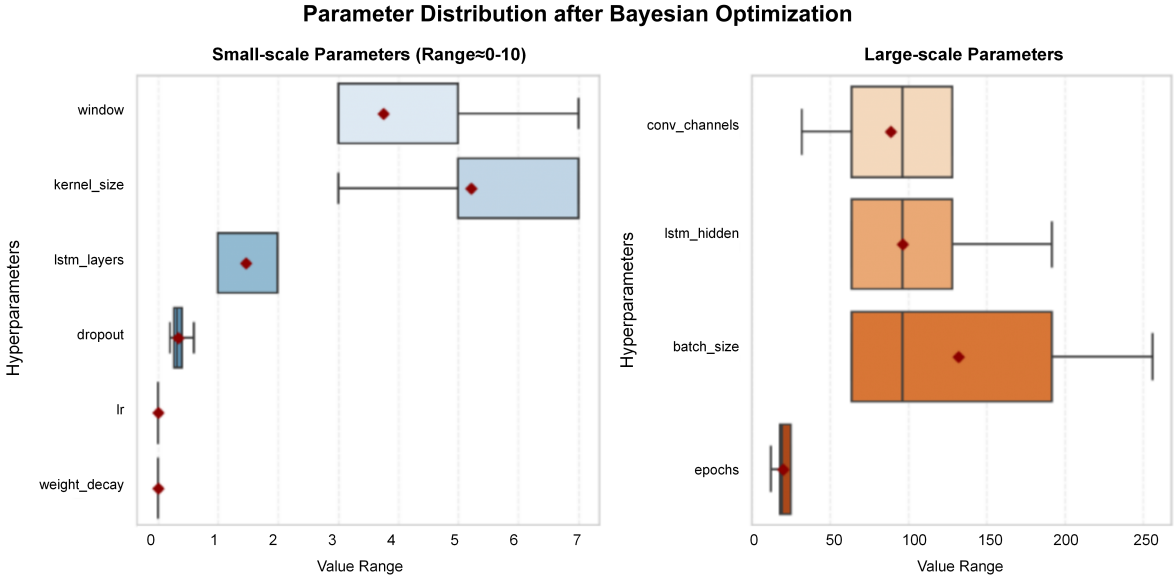

Fig. 11 Parameter distributions during Bayesian optimization

results show that after multiple iterations, all parameters tend to be stable in the numerical interval, which indicates that the optimization process has good convergence and the results are repeatable.

In terms of small-scale parameters, the sliding window length is mainly concentrated in 3–7, the convolutional kernel size is concentrated in 5–7, and the dropout rate is about 0.3, indicating that the model tends to extract local features in shorter logging segments to balance detail capture with overfitting risk. In terms of large-scale parameters, the number of convolutional channels is mainly distributed in the range of 64–96, the number of hidden units in BiLSTM is about 96, the learning rate is concentrated in the order of 1×10^{-3}, and the weight attenuation coefficient is about 1×10^{-4}. The combination of parameters

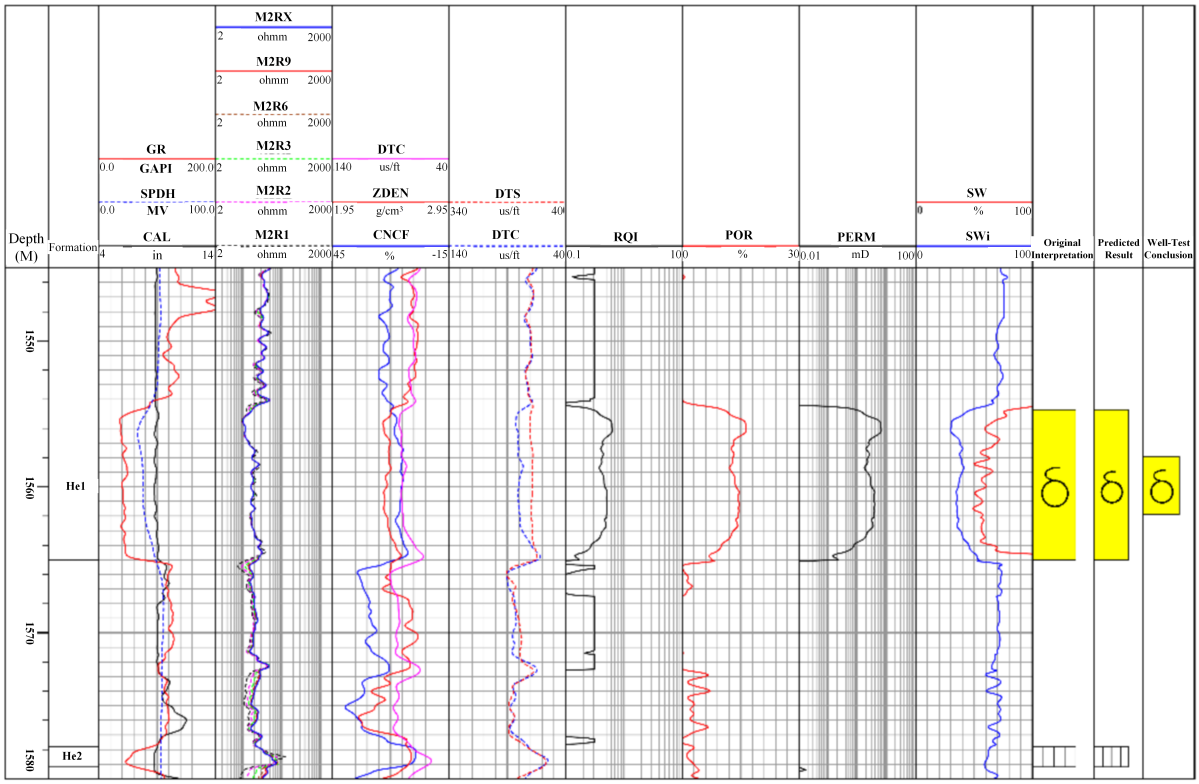

Fig. 12 Field validation example from Well X-151

reflects the optimal compromise between structural complexity and training stability: the medium-scale network structure can ensure the feature expression ability while avoiding overfitting; A reasonable learning rate and regularization intensity ensure a smooth convergence of the training process.

Overall, Bayesian optimization enables the CNN–BiLSTM model to achieve efficient hyperparameter configuration and performance improvement under complex multi-well conditions. The optimized parameter distribution has significant regularity and engineering rationality: the short window enhances the local sensitivity of the model to the logging response, the medium network scale ensures the complete expression of timing features, and the moderate learning rate and weight attenuation improve the overall stability and generalization ability. It can be seen that the constructed BOA-CNN-BiLSTM model has excellent performance in recognition accuracy, robustness and generalization ability, which can provide reliable support for high-precision intelligent identification of fluid types in tight sandstone reservoirs.

5.2.4 Comparison with traditional models and field validation

A comparative evaluation was conducted to benchmark the proposed BOA–CNN–BiLSTM model against several commonly used machine-learning and deep-learning approaches.

To further validate the practical applicability of the proposed approach, Well X-151 was selected as a representative case. The He1 interval (1555–1565 m) is shown in Fig. 12.

The average GR value of this section is about 43.86 API, which is significantly lower than that of the adjacent mudstone background, which clearly indicates that this section is a sandstone reservoir. The SP curve as a whole has a low amplitude response (average of about 39.09 mV), which reflects certain permeability characteristics and is consistent with the judgment of sandy lithology. The density curve (ZDEN) fluctuates in the range of 2.38–2.50 g/cm^3, and the neutron porosity (CNCF) and acoustic time difference (DTC) show relatively high values, which together reveal that this layer has a high effective porosity (about 13%) and a certain storage capacity.

It is worth noting that the neutron-density curve has a clear intersection in this section, showing a typical gas layer "excavation effect". Combined with the comprehensive analysis of lithology, electricity and porosity, it can be considered that this section has the characteristics of gas charging. The longitudinal contrast resistivity curve shows that although the porosity is about 13%, the reservoir is still a dense phase, which makes the resistivity amplitude not sensitive enough in fluid type discrimination. At the same time, there are no obvious flushed zones or intrusion characteristics in the curves, which is consistent with the electrical response characteristics of "low permeability-weak intrusion" in tight sandstone gas reservoirs.

6 Conclusions

In this study, a CNN-BiLSTM hybrid model based on Bayesian optimization (BOA-CNN-BiLSTM) was constructed and verified around the three core problems of "high overlap of curve response, complex lithological change, and sample class imbalance" in fluid identification of tight sandstone reservoirs. Through the systematic data standardization, composite feature construction and hyperparameter automatic optimization

process, the performance bottleneck of traditional methods is effectively broken. The main conclusions are as follows:

(1) The tight sandstone logging data generally has problems such as unit confusion, missing curves, and large noise, and the establishment of a unified data cleaning and standardization process is crucial to ensure the quality of model input.

(2) Due to the significant overlap between traditional logging curves, it is difficult to effectively distinguish different fluid types with a single feature. The composite characteristics of structural properties, brittleness, and pore structure can significantly enhance the gas-water identification ability.

(3) Bayesian optimization can efficiently complete the global search of key hyperparameters of the model, significantly reduce the trial and error cost compared with manual parameter adjustment, and improve the training efficiency and model stability.

(4) In the Leave-One-Well-Out (LOO) verification of 134 wells, the overall identification accuracy of the BOA-CNN-BiLSTM model reached 93.5%. In complex reservoirs that are most easily misjudged by traditional models, such as gas-water transition layers and low-resistivity gas layers, the F1 value is increased by about 4%-5% compared with the benchmark model, showing stronger robustness and generalization ability, especially in distinguishing between gas layers and water layers.

In general, the Bayesian optimization deep learning framework proposed in this study can effectively reduce the false judgment rate of gas-water coexistence layers and improve the reliability of reservoir gas content evaluation and fracturing design. This method provides a replicable and generalizable technical path for fluid identification in tight sandstone reservoirs, and has important engineering application value and theoretical significance for the intelligent development of unconventional natural gas.

Acknowledgements

This work was developed by the Chinese National Natural Science Foundation Youth Project under grant 42204127, the Hubei Provincial Department of Education Science and Technology Research Program for Young Talents under grant Q20221304, the China Postdoctoral Science Foundation General Program under grant 2017M622382, and the Open Fund Project of the Key Laboratory of Oil and Gas Resources and Exploration Technology, Ministry of Education under grant K2018-16.

Conflict of interest

The authors declare no competing interest.

References

Bai Z, Tan MJ, Shi YJ, et al. 2022. Genesis of low-resistivity oil pays and a fluid identification method for tight sandstone reservoirs: A case study of the Chang 8 Formation in the Longdong West area, Ordos Basin. *Geophysical Prospecting for Petroleum*, **61**(4): 750–760. doi:10.3969/j.issn.1000-1441.2022.04.018.

Chen G, Zhang Y, Wang J, et al. 2023. Application of bidirectional LSTM networks to lithology identification in beach-bar sandstone reservoirs. *Well Logging Technology*, **47**(3): 319–325. doi:10.16489/j.issn.1004-1338.2023.03.010.

Cui J, Yang B. 2018. A review of Bayesian optimization methods and applications. *Journal of Software*, **29**(10): 3068–3090. doi:10.13328/j.cnki.jos.005607.

Fang SN, Zhang ZS, Wang Z, et al. 2020. Principal Slip Zone determination in the Wenchuan earthquake Fault Scientific Drilling project—Hole 1: Considering the Bayesian discriminant function. *Acta Geophysica*, **68**(6): 1–13. doi:10.1007/s11600-020-00496-z.

Gu J, Wang Z, Kuen J, et al. 2018. Recent advances in convolutional neural networks. *Pattern Recognition*, **77**: 354–377.

Lei CJ, Luo RZ, Wu J, et al. 2025. Cost-sensitive gradient boosting tree for identifying water-rich tight sandstone fluids: A case study of the X gas field in the Ordos Basin. *Progress in Geophysics*, Online First. doi:10.1002/2008JB006013.

Li B. 2025. Efficient development practice and understanding of deep coalbed methane in the eastern margin of the Ordos Basin. *Drilling & Production Technology*, **48**(3): 119–126. doi:10.3969/J.ISSN.1006-768X.2025.03.14

Liao GZ, Li YZ, Xiao LZ, et al. 2020. Prediction of microscopic pore structure of tight reservoirs using convolutional neural network model. *Petroleum Science Bulletin*, **5**(1): 26–38. doi:10.3969/j.issn.2096-1693.2020.01.003

Lim B, Arik SO, Loeff N, et al. 2021. Temporal fusion transformers for interpretable multi-horizon time-series forecasting. *International Journal of Forecasting*, **37**(4): 1748–1764. doi:10.1016/j.ijforecast.2021.03.012

Lindemann B, Müller T, Vietz H, et al. 2021. A survey on long short-term memory networks for time series prediction. *Procedia CIRP*, **99**: 650–655. doi:10.1016/j.procir.2021.03.088

Luo G, Xiao LZ, Shi YQ, et al. 2022. Machine learning for reservoir fluid identification with logs. *Petroleum Science Bulletin*, **7**(1): 24–33. doi:10.3969/j.issn.2096-1693.2022.01.003

Mao KY. 2016. Logs Fluid Typing Methods and Adaptive Analysis of Tight Sandstone Reservoir of Yingcheng Formation in Lishu Fault. *Advances in Earth Science*, **31**(10): 1056–1066. doi:10.11867/j.issn.1001-8166.2016.10.1056

Mi HG, Zhang B, Zhu GH, et al. 2022. Geological characteristics and development potential of the Linxing tight sandstone gas reservoir. *Special Oil & Gas Reservoirs*, **29**(6): 65–72. doi:10.3969/j.issn.1006-6535.2022.06.008.

Yan XF, Cao H, Yao FC, et al. 2012. Bayesian lithology discrimination and pore-fluid detection in tight sandstone reservoirs. *Oil Geophysical Prospecting*, **47**(6): 945–950.

Yu DG, Sun JM, Zhang ZC, et al. 2005. Fluid property identification of reservoirs using support vector machines. *Xinjiang Petroleum Geology*, **26**(6): 675–677.

Zhang HT, Yang XM, Chen Z, et al. 2022. Logging data reconstruction using enhanced bidirectional long short-term memory networks. *Progress in Geophysics*, **37**(3): 1214–1222. doi:10.6038/pg2022FF0194

Dong ZZ, Holditch SA, Ayers QB. 2015. Probabilistic evaluation of global technically recoverable tight gas resources. *SPE Economics & Management*, **7**(3): 112–119. doi:10.2118/165704-PA.

Zhao B, Chen X, Gao CQ, et al. 2025. Fluid property identification of carbonate reservoirs based on Bayesian-optimized BiGRU. *Journal of Yangtze University (Natu-*

ral Science Edition), Online First. doi:10.16772/j.cnki.1673-1409.20240418.002.

Zhao Q. 2024. Fluid property identification of tight sandstone reservoirs based on well-logging data. Master's thesis, Yangtze University. doi:10.26981/d.cnki.gjhsc.2024.000497.

Zhao Q, Yang B, Li X, et al. 2018. Application of a chart-plate decision tree in fluid identification. *Well Logging Technology*, **42**(6): 641–646. doi:10.16489/j.issn.1004-1338.2018.06.006.

Zhu GH, Li BL, Li ZC, et al. 2022. Unconventional natural gas exploration practices and development direction in the eastern margin of the Ordos Basin: A case study of the Linxing–Shenfu gas field. *China Offshore Oil and Gas*, **34**(4): 16–29. doi:10.11935/j.issn.1673-1506.2022.04.002.