

An Intelligent Well Log Based Lithology Identification Method Integrating Unsupervised Clustering and BiLSTM-CRF

Zihao Mu

15701219426@163.com

Northeast Petroleum University <https://orcid.org/0009-0001-4798-3230>

Liwei Song

Northeast Petroleum University <https://orcid.org/0000-0003-1039-8131>

Jiacheng Huang

Northeast Petroleum University

Haiwei Mu

Hainan Vocational University of Science and Technology

Xiaochang Lv

Northeast Petroleum University

Guohua Zhang

Northeast Petroleum University

Guangjie Fu

Shenyang Urban Construction University

Yan Shi

Northeast Petroleum University

Xinyu Liu

Northeast Petroleum University

Yuhang Wang

Northeast Petroleum University

Research Article

Keywords: Lithology identification, Well logs, Gaussian mixture model, Deep learning, Reservoir characterization

Posted Date: June 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9667239/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

An Intelligent Well Log Based Lithology Identification Method Integrating Unsupervised Clustering and BiLSTM-CRF

Zihao Mu^{1,2,3}, Liwei Song^{1,*}, Jiacheng Huang³, Haiwei Mu⁴, Xiaochang Lv², Guohua Zhang²,
Guangjie Fu⁵, Yan Shi¹, Xinyu Liu², Yuhang Wang²

¹Sanya Offshore Oil & Gas Research Institute, Northeast Petroleum University, Sanya, China.

²Qinhuangdao Campus, Northeast Petroleum University, Qinhuangdao, China.

³School of Earth Sciences, Northeast Petroleum University, Daqing, China.

⁴Hainan Vocational University of Science and Technology, Haikou, China.

⁵School of Mechanical Engineering, Shenyang Urban Construction University, Shenyang, China.

*Author to whom correspondence should be addressed.

*Corresponding author at: Sanya Offshore Oil & Gas Research Institute, Northeast Petroleum
University, Sanya 572025, China; Email address: zhida090@163.com

Highlights

1. A GMM-based unsupervised labelling strategy is developed to reduce reliance on manually interpreted lithology labels.
2. Well-log reconstruction and feature screening are incorporated to improve model robustness under incomplete logging conditions.
3. A BiLSTM-CRF sequence model is used to jointly capture bidirectional depth-dependent features and lithological transition constraints.

Abstract

Accurate lithology identification from well logs is essential for reservoir characterization, stratigraphic interpretation, and hydrocarbon exploration. However, lithology prediction in heterogeneous reservoirs remains challenging because well log responses are nonlinear, lithological classes are strongly imbalanced, and logging curves are often incomplete. To address these issues, this study proposes an intelligent lithology identification method that integrates Gaussian mixture model based unsupervised clustering with a bidirectional long short-term memory conditional random field network. First, well log curves are standardized, reconstructed, and screened to reduce dimensional inconsistency and improve data continuity. Second, GMM is used to model the statistical distribution of multi-parameter logging responses and generate geologically constrained pseudo-labels. Finally, BiLSTM extracts bidirectional sequential features along the depth domain, while CRF imposes global constraints on adjacent lithological transitions. The proposed workflow is evaluated

using classification metrics, error analysis, and multi-well prediction experiments from the Viking Graben well log dataset. Results show that the proposed GMM-BiLSTM-CRF model achieves an overall accuracy of 93.18% and a single-well average accuracy of 90.1%, while maintaining stable performance under incomplete logging-curve conditions. The method provides a data-driven and geologically consistent solution for automatic lithology identification and can support reservoir evaluation and geological interpretation in complex subsurface settings.

Keywords: Lithology identification; Well logs; Gaussian mixture model; Deep learning; Reservoir characterization

1.Introduction

Lithology identification is a fundamental component of well log interpretation and formation characterization, and its reliability has a direct influence on reservoir classification, geological modeling, and development decision making (Chen et al., 2025). Compared with core observations, well log data provide continuous depth coverage and can be acquired at relatively low cost, making them an essential data source for lithology prediction and stratigraphic correlation (Jin et al., 2025). However, the mapping relationship between well log responses and lithological categories is commonly nonlinear, ambiguous, and affected by multiple geological factors. In heterogeneous intervals and lithological transition zones, different lithologies may exhibit similar or overlapping responses in multidimensional log feature space, thereby increasing the uncertainty of lithology discrimination. Consequently, accurate lithology identification from well logs remains a challenging classification problem that requires robust feature representation and effective nonlinear decision modeling (He et al., 2023; Mousavi et al., 2024).

Traditional lithology identification methods have mainly relied on geological expertise and statistical discrimination strategies, including single curve threshold analysis and crossplot based interpretation using multiple logging curves (Mu et al., 2024, Mehama et al., 2025). Although these methods are intuitive and geologically interpretable, their performance is strongly dependent on expert experience and subjective judgement. With the increasing scale, dimensionality, and complexity of well log data, conventional interpretation approaches have shown limitations in terms of identification accuracy, prediction stability, and practical applicability. These limitations have promoted the development of intelligent lithology identification methods (Chang et al., 2022).

With the continuous advancement of artificial intelligence techniques, clustering

analysis has been increasingly applied to lithology identification (Bhattacharya et al., 2016; Prabowo et al., 2025). Compared with supervised learning algorithms that require a large amount of labeled data, unsupervised clustering algorithms can identify intrinsic data structures without prior label information (Kiselev et al., 2019; Mu et al., 2025). Peng et al. (Peng et al., 2023) integrated Fcola Loss with extreme gradient boosting, namely XGBoost, and proposed a semi supervised clustering method, which demonstrated favorable generalization ability under certain conditions. Zerui Li et al. (Li et al., 2021) developed a semi supervised clustering method based on a feature deep smoothing strategy, showing good robustness under complex sample conditions. Dunham et al. (Dunham et al., 2019) applied semi supervised clustering to well log data and obtained relatively stable identification results. Mohana et al. (Mehana et al., 2021) used unsupervised K means clustering for oil well performance evaluation and further employed the evaluation results for reservoir classification in new wells. Yan Guo et al. (Guo et al., 2023) constructed a supervised lithology identification method based on Yolov5 and achieved high accuracy on the validation set, although the collaborative fusion of multimodal information remained insufficient. Xialin Zhang et al. Xialin Zhang et al. (Zhang et al., 2023) introduced an attention mechanism into a supervised lithology identification model, but further improvement is still required for thin bed identification.

In recent years, bidirectional long short term memory networks, namely BiLSTM, have shown strong performance in geological applications because they can capture bidirectional dependencies along the depth dimension of well log sequences (Bian, n.d.; Chen et al., 2017). Qifu Fan et al. (Fu, et al., 2025) proposed a BiLSTM method integrated with an attention mechanism and achieved favorable classification performance in highly heterogeneous shale reservoirs. Wenjie Cui et al. (Cui, et al., 2023) developed a GWO BiLSTM method, in which GWO refers to the Grey Wolf Optimizer. However, this optimization strategy requires a large number of iterative evaluations, and its performance may be constrained under imbalanced data distributions. Yingqiang Qi et al. (Qi et al., 2025) combined BiLSTM with CNN and Transformer architectures for lithology classification and obtained satisfactory prediction results.

Overall, current intelligent lithology identification methods still face several challenges. First, labels generated by unsupervised algorithms often have limited accuracy, which may introduce noise into subsequent model training. Second, the imbalanced distribution of lithology samples restricts the ability of models to identify

minority lithology classes. Third, missing well log data are common in practical oilfield applications, which may reduce model stability and prediction reliability. To address these issues, this study proposes an intelligent well log lithology identification method that integrates unsupervised clustering with GMM-BiLSTM-CRF, following the principle of complementary advantages (Yang et al., 2017). In this framework, BiLSTM is used to extract bidirectional sequential features from well log data, while CRF imposes global consistency constraints during sequence labeling. The collaborative coupling of BiLSTM and CRF improves the robustness and generalization ability of the proposed model under complex reservoir conditions, thereby providing technical support for intelligent lithology identification.

2.Data description and analysis

2.1Study Area

The dataset used in this study was derived from the Viking Graben region and includes 118 wells (Badley et al., 1988). The spatial coverage extends across the South Viking Graben and the North Viking Graben, as shown in Fig. 1. The study area has a large lateral extent, and marked differences in structural configuration and sedimentary filling characteristics occur among different subbasins. These geological variations provide diverse training samples for well logging based lithology identification.

In terms of stratigraphic development, the lithological assemblages in the Norwegian offshore area are generally dominated by shale and shale rich sediments, followed by sandstone, limestone, marl, and tuff. In addition, lithology proportions vary among different data subsets. Evaporite lithologies, including Halite and Anhydrite, are relatively abundant in some subsets, which further increases the imbalance of the data distribution.

From the perspective of regional geological evolution, the formation and filling of the Viking Graben were controlled by multiple stages of rifting and subsidence (Brekke et al., 2001). From the Triassic to the Cretaceous, the study area experienced successive evolutionary stages, including extensional rifting, uplift and erosion, and renewed rifting. These processes resulted in fluvial and lacustrine deposits, deltaic deposits, and widely developed organic rich shale deposits. The combined effects of multiphase tectonic activity and sedimentary filling led to diverse lithology types, frequent vertical and lateral lithological variations, and complex well log response characteristics. Consequently, higher requirements are imposed on the stability and adaptability of lithology identification models.

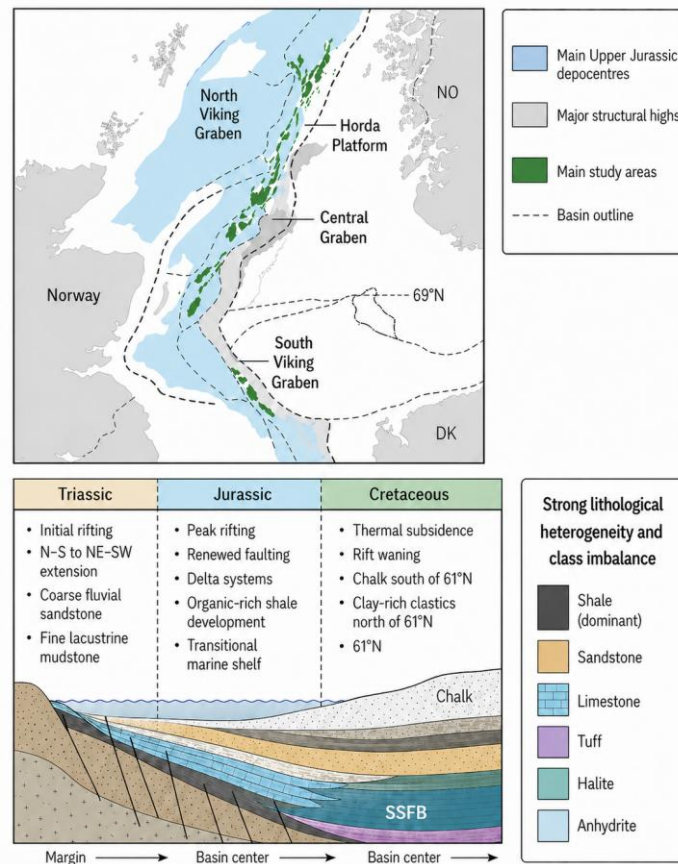


Fig.1.Overview of the study area in the South and North Viking graben

2.2Lithological Characteristics

The sedimentary filling process in the study area was jointly controlled by rifting activity, transgression and regression cycles, and variations in sedimentary dynamic conditions. As a result, the lithological assemblages within the well intervals show pronounced vertical sequence differentiation and lateral distribution heterogeneity. Clastic rocks and mud rich clastic sediments account for a relatively high proportion, whereas carbonate rocks and evaporites are intermittently developed. Tuff and coal seams occur locally, and crystalline basement is encountered in several basal well intervals (Ramm, 1992). Such multi lithology, multi interface, and strongly heterogeneous lithological assemblages cause evident superposition and overlap effects in well log responses across different well intervals, thereby increasing the uncertainty of lithology identification (Redfield et al., 2016).

Based on regional stratigraphic data, well log response characteristics, and integrated interpretation results, the lithologies in the study area were classified into 12 categories: Sandstone, Sandstone and Shale Interbeds, Shale, Marl, Dolomite, Limestone, Chalk, Halite, Anhydrite, Tuff, Coal, and Crystalline Basement. The lithology types and their geological characteristics are presented in Table 1.

Table 1. Lithology Types and Characteristics of the Viking Graben

Lithology type	Abbreviation	Main features	Depositional setting
Sandstone	SS	Quartz-rich, well sorted	Fluvial / shallow-marine sandstone body
Sandstone/Shale	SSS	Sand–shale interbeds, heterogeneous	Delta front / tidal flat / mixed depositional setting
Shale	SH	Mud-rich, laminated	Deep-water shelf / offshore mud deposition
Marl	MR	Carbonate – clay mixed rock	Transitional marine shelf / restricted basin
Dolomite	DL	Crystalline, locally porous	Restricted platform / evaporitic platform
Limestone	LS	Calcite-rich, compact	Shallow-marine carbonate platform
Chalk	CK	Micritic, soft, porous	Open marine carbonate deposition
Halite	HL	Evaporite, plastic	Evaporitic basin / restricted lagoon
Anhydrite	ANH	Dense, brittle evaporite	Sabkha / evaporitic platform
Tuff	TF	Volcaniclastic, variable	Volcanic ash deposition / reworked sedimentary setting
Coal	CO	Organic-rich, low density	Swamp / delta plain / coastal marsh
Crystalline Basement	CB	Dense, hard basement rock	Basement setting

As shown in Fig. 2, the well log curves from different well intervals exhibit significant differences in vertical distribution patterns and response amplitudes. In the well shown in Fig. 2(a), GR values are relatively low and show limited fluctuation, while the RHOB and NPHI curves vary more smoothly. DTC and RDEP display certain differentiation characteristics in some intervals, indicating that the lithological assemblage of this well is relatively stable and is dominated by clastic rocks with low to moderate shale content. In contrast, the well shown in Fig. 2(b) exhibits more pronounced oscillatory characteristics. The GR and SGR curves show markedly enhanced fluctuations, RHOB and NPHI present a strong inverse correlation, and the response range of DTC is enlarged. These features indicate frequent lithological variations and more complex development of interlayer interfaces within this well interval.

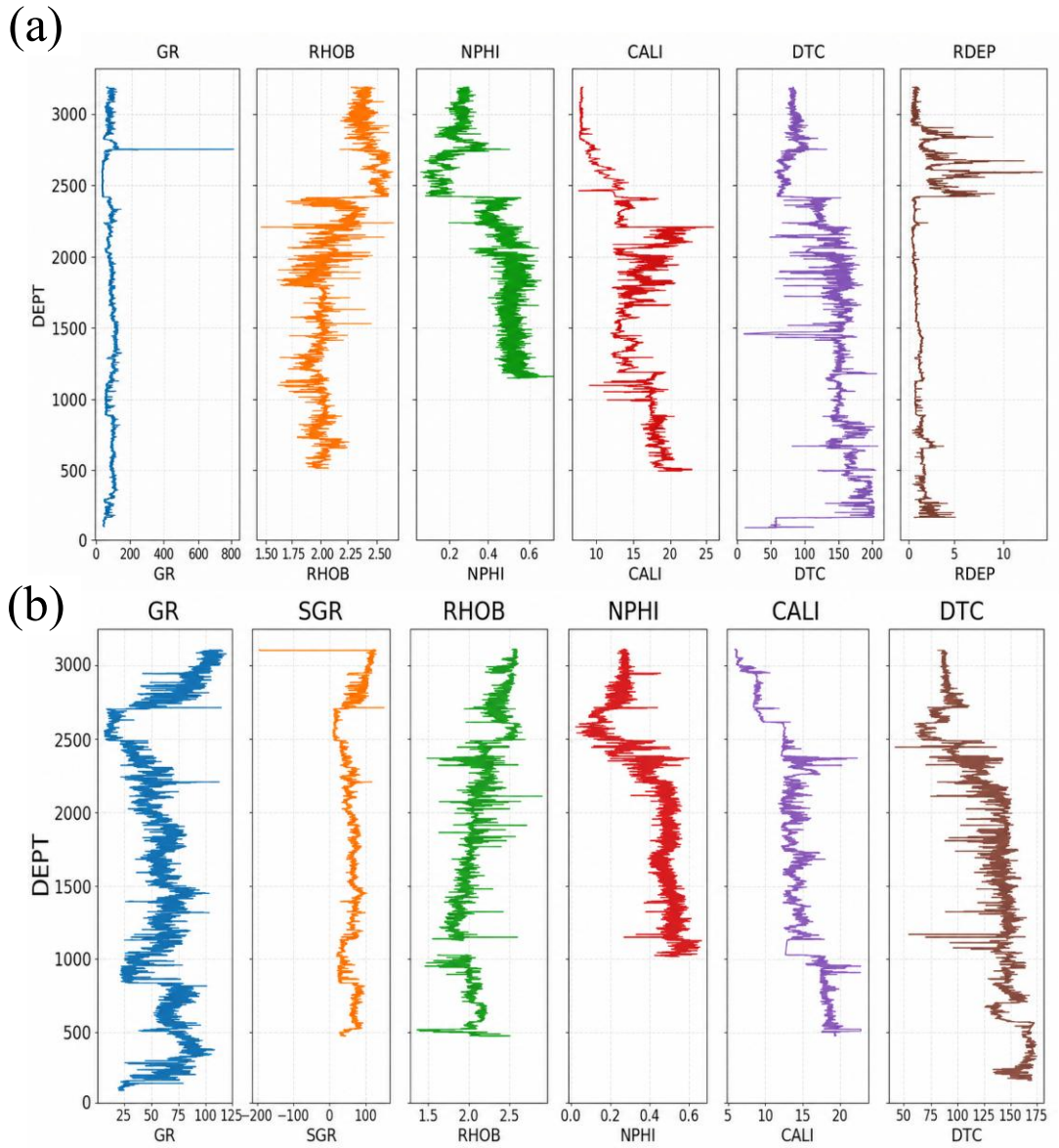


Fig.2. Integrated Well Log Responses of Multiple Well
 (a) Well log response characteristics of Well 15_9_15
 (b) Well log response characteristics of Well 15_2_17

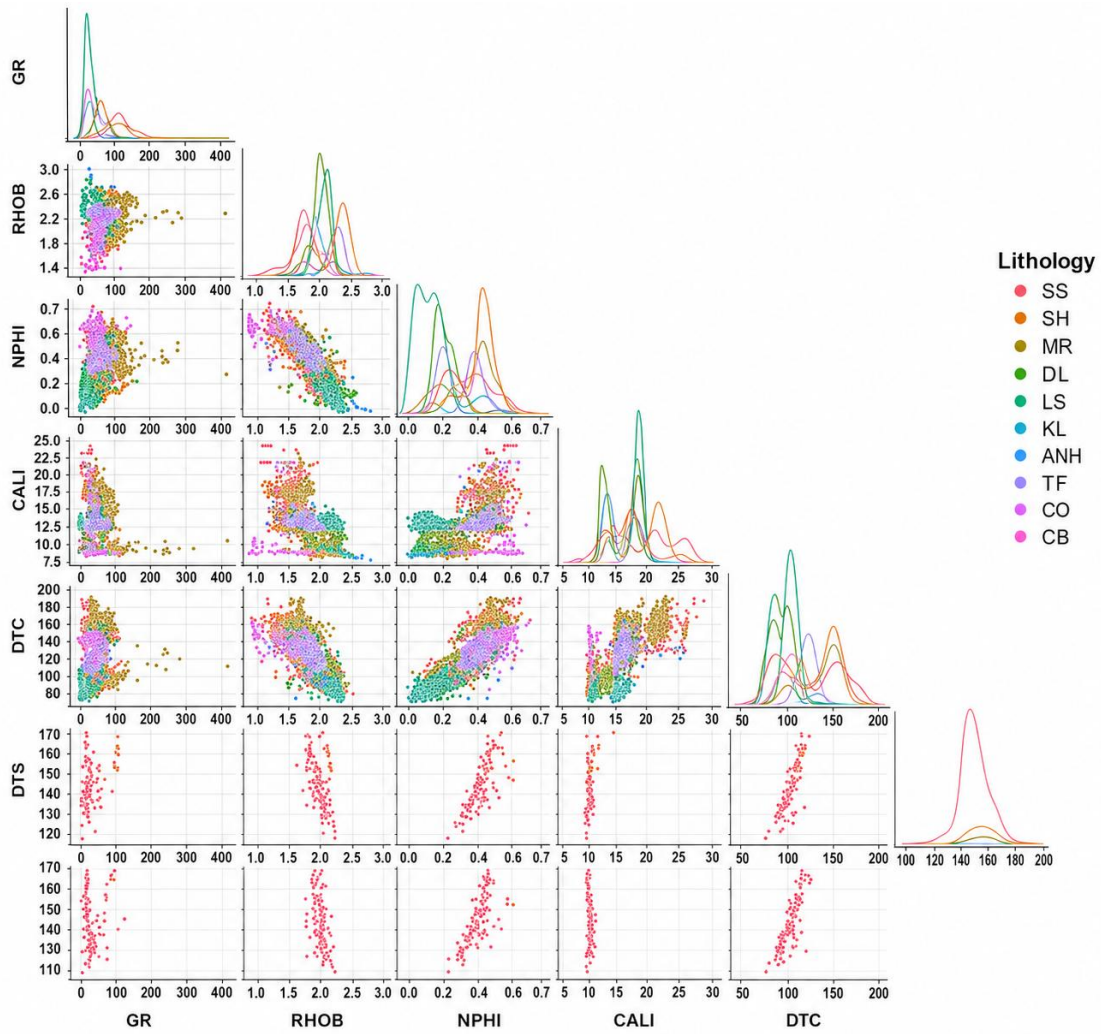


Fig. 3 Cross plot of well log responses and lithology

As shown in Fig. 3, the cross plots of well log lithology parameters indicate that different lithologies exhibit distinct distribution patterns in the multilog parameter space. Sandstone and transitional sandstone shale lithologies generally show relatively low GR values, whereas fine grained lithologies such as shale and marl commonly correspond to higher GR values. In the cross plot of bulk density, namely RHOB, and neutron porosity, namely NPHI, different lithology types are distributed along trend bands, reflecting the controlling effects of lithological composition and pore structure differences on well log responses. The distribution ranges of acoustic slowness parameters, including DTC and DTS, and the caliper parameter, namely CAL, also differ markedly among lithologies. In particular, carbonate and evaporite lithologies show relatively concentrated distribution characteristics in acoustic responses.

2.3 Well-Log Response Characteristics

Based on the statistical characteristics of well log response boxplots for different lithologies in the Force 2020 dataset (SnapBormann et al., 2020), distinct differences can be observed among lithology types in the multiparameter logging responses, as

shown in Fig. 4. The GR curve shows good discriminatory capability between sandstone related lithologies, including SS, SSS, and SH, and argillaceous lithologies, including MR and DL. Sandstone related lithologies generally exhibit moderate to low gamma ray responses, whereas argillaceous and mixed lithologies commonly correspond to higher GR values, indicating the dominant control of shale content on gamma ray responses.

RHOB and NPHI are relatively sensitive to variations in rock framework minerals and pore structure. Carbonate and evaporite lithologies generally present a combined response pattern characterized by higher density and lower neutron porosity, whereas some clastic lithologies mostly show moderate density and relatively higher porosity. Acoustic slowness parameters show relatively stable response interval differences among different lithologies, especially with clear zonation between tight lithologies and high porosity lithologies, effectively reflecting variations in rock mechanical properties and compaction degree. The CAL curve exhibits significantly increased dispersion in lithologies with high shale content or poor borehole wall stability, suggesting that caliper responses are controlled not only by lithology but also by additional factors such as borehole collapse (Fetni et al., 2025)

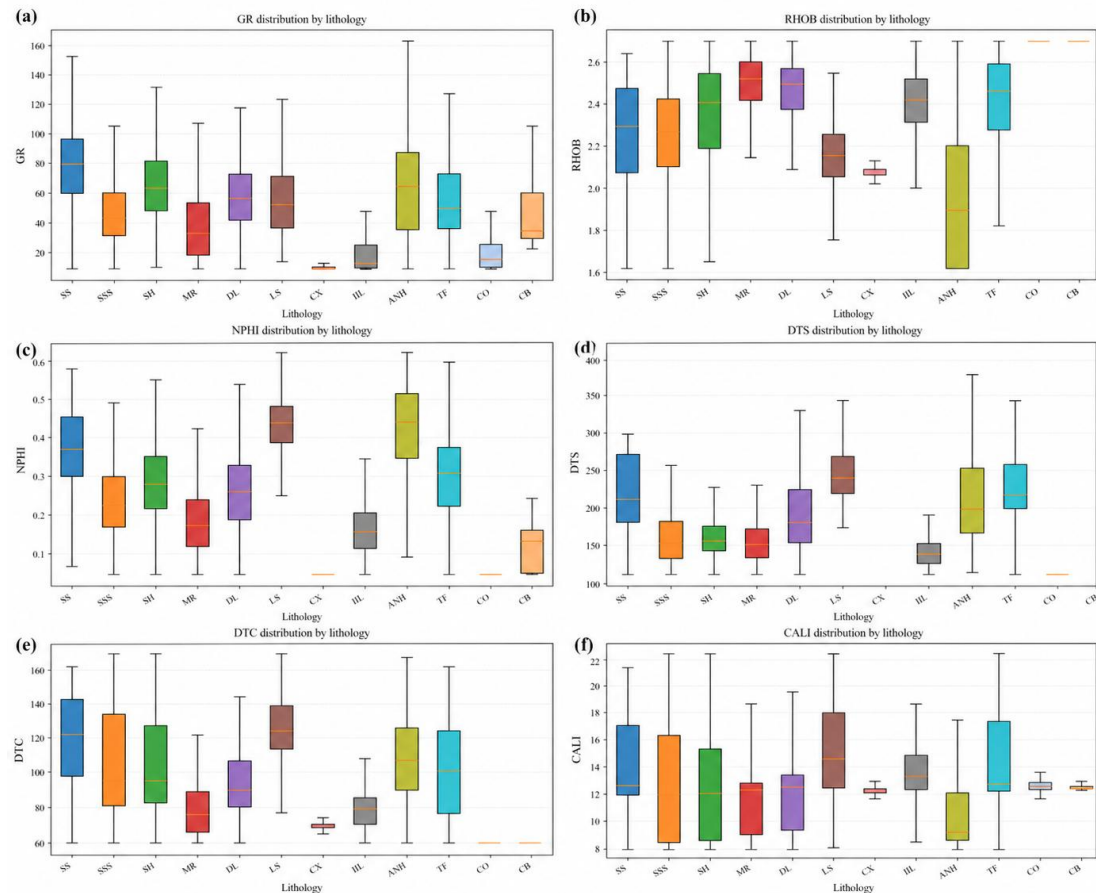


Fig 4. Boxplots of Well Log Response Characteristics for Different Lithologies

3. Methodology

3.1 Workflow of the Proposed Method

The framework of the intelligent well log lithology identification algorithm consists of three modules, as shown in Fig. 5. The data construction module is responsible for data balancing and feature parameter selection of the original well log curves, thereby constructing a dataset suitable for sequence modeling. The unsupervised lithology labeling module uses GMM to probabilistically model the distribution characteristics of well log data and generate initial sample labels automatically, providing reliable training samples for subsequent supervised learning. Finally, the lithology identification module integrates bidirectional long short-term memory networks with conditional random fields, which effectively improves lithological boundary recognition and overall classification accuracy, enabling high precision intelligent lithology identification from well logs under complex geological conditions.

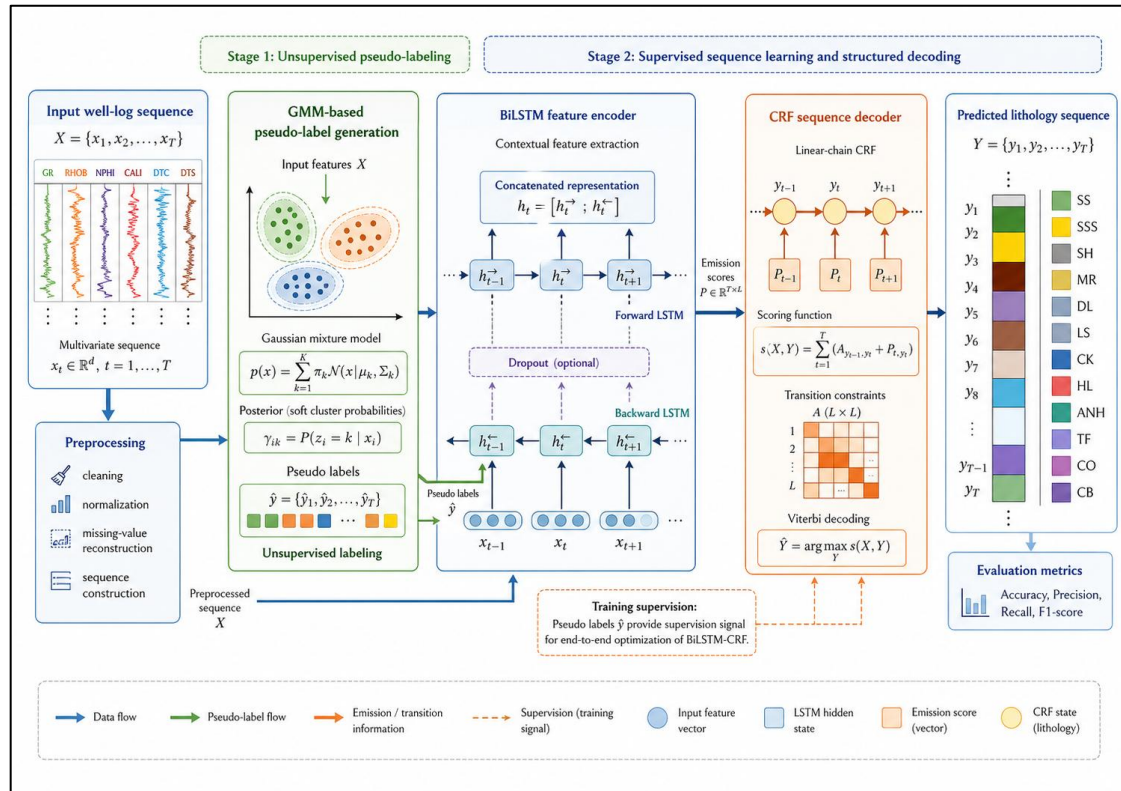


Fig.5 Framework diagram of the GMM-BiLSTM-CRF-based intelligent lithology identification method using well logs

3.2 Data Preprocessing

(1) Data balancing

The statistical distribution of lithology samples is presented in Table 2, showing pronounced differences in sample size among lithology categories. The SS class contains 1,754,086 samples, representing the largest category in the dataset. The SSS and SH classes also account for relatively high proportions, with 415,408 and 361,640 samples, respectively. In contrast, lithologies such as ANH, TF, CO, and CB contain considerably fewer samples, with the CB class including only 282 samples. This distribution exhibits a typical long tail pattern. If such class imbalance is not properly addressed, the model may become biased toward majority classes during training, thereby reducing overall classification accuracy and, more importantly, weakening its ability to identify minority lithology classes. (Martin-Fernandez et al., 2018)

Tab.2 Category statistics of lithology samples in the dataset

Lithology	Count	Lithology	Count
SS	1754086	CK	29424
SSS	415408	HL	28086
SH	361640	ANH	9508
MR	138996	TF	4782
DL	82076	CO	3616
LS	34862	CB	282

To alleviate the above imbalance problem, a combined strategy of random undersampling (Wang et al., 2020) and oversampling using SMOTE (Camacho et al., 2022) was adopted to balance the lithology samples. SMOTE controls the generation and distribution of synthetic samples based on minority class samples, thereby improving the balance of the dataset. Specifically, a sample is first randomly selected from the minority class sample set, and one of its neighbors is then randomly selected from its K nearest neighbors. A random number within the interval (Reference 0; Chen et al., 2025) is subsequently multiplied by the difference between the corresponding feature vectors to generate a synthetic sample, as expressed in Eq. (1).

$$c_{new} = c_i + (\hat{c}_i - c_i) \times \delta \quad (1.0)$$

where c_i denotes a minority class sample, $\hat{c}_i$ denotes one of the K nearest neighbor samples of c_i , and δ is a random number within the interval (Reference 0; Chen et al., 2025). After the above balancing procedure, the sample distribution of different lithology categories becomes more reasonable, which helps improve the ability of the model to learn representative features of minority lithology classes. (Chen, n.d.)

(2) Well log data reconstruction

For the well log data from well TF2 13, a curve reconstruction method was introduced to improve the completeness of the well log sequences (Chen, n.d.). Six curves, including GR, RHOB, NPHI, DTS, DTC, and CAL, were first processed for missing value identification and outlier removal. Extreme values caused by non-

geological factors were set as null values, and interpolation along the depth direction was then performed to obtain continuous initial curves. On this basis, a multivariate collaborative reconstruction strategy was introduced (Velásquez Hernández CA, 2020). The statistical correlations among multiple well log curves were used to iteratively estimate and complete the missing intervals. A light smoothing procedure was further applied to the reconstructed results to suppress local high frequency noise while preserving the overall curve trends.

As shown in Fig. 6, before reconstruction, the curves contain discontinuities, spikes, and local abrupt changes, which reduce the usability of well log curves along the depth dimension. After reconstruction, all six curves are continuously reconstructed over the entire depth range, the missing intervals are completed, and abnormal fluctuations are constrained. These reconstructed curves provide stable and reliable input data for the subsequent lithology identification model.

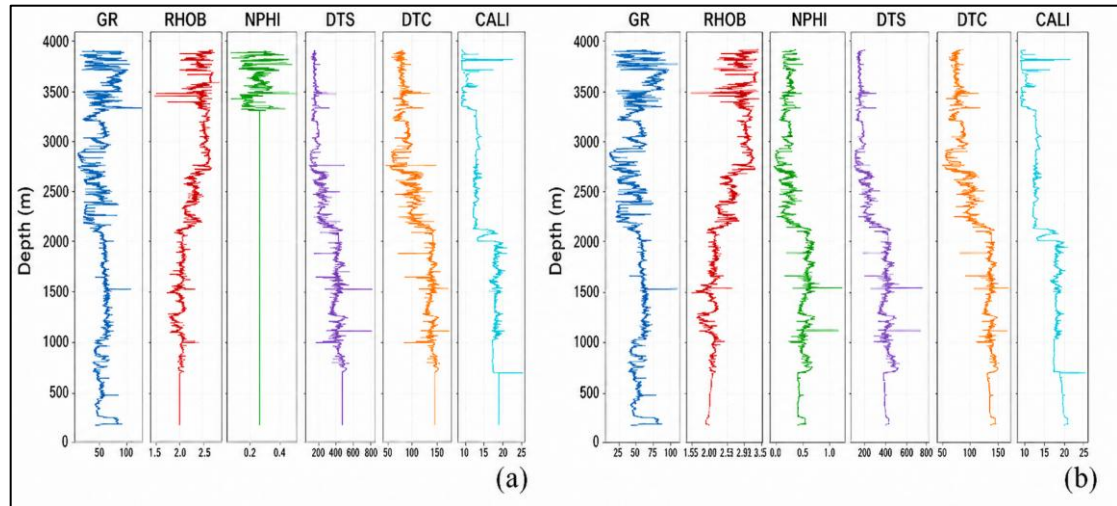


Fig.6 Comparison of logging data before and after reconstruction
(a) Logging data before reconstruction (b) Logging data after reconstruction

3.3 Feature Selection

The original well log features include six curves, namely gamma ray, bulk density, neutron porosity, compressional wave slowness, shear wave slowness, and caliper. To reduce redundant information and computational burden caused by high dimensional input, effective screening of the well log curves is required. The analysis results indicate that GR, RHOB, NPHI, and acoustic slowness curves exhibit clearer distributional differences among different lithologies. On the one hand, the medians and interquartile ranges corresponding to different lithologies show obvious offsets, providing strong interclass discrimination information. On the other hand, most lithologies show relatively low dispersion, indicating that these curves have good stability in vertical

responses and have the potential to serve as dominant features for lithology identification.

In contrast, CAL is a caliper curve, and its variations are more susceptible to non-geological factors such as borehole collapse, borehole enlargement, mud cake development, and drilling conditions. Therefore, CAL may introduce additional noise in some intervals, which is unfavorable for the model to extract the true lithological response patterns. Based on the comprehensive analysis results presented in Table 3, GR, RHOB, NPHI, DTC, and DTS were selected as the input features for model identification.

Tab.3 Contribution rate and cumulative contribution rate of principal component analysis for well logging curves

Well Log Curve	Contribution Rate(%)	Cumulative Contribution Rate(%)
GR	28.7	28.7
RHOB	22.5	51.2
NPHI	19.1	70.3
DTC	15.6	85.9
DTS	10.2	96.1
CAL	3.9	100

3.4 GMM-BiLSTM-CRF Lithology Identification Method

To address the dependence of lithology sample labeling on manual expertise and the limited accuracy of clustering based labels in lithology identification, this study develops an intelligent lithology identification method based on well log sequence data. The proposed method integrates the complementary strengths of Gaussian mixture model, bidirectional long short term memory network, namely BiLSTM, and conditional random field, thereby improving the reliability of label generation, sequential feature representation, and lithological boundary discrimination.

3.4.1 GMM-Based Unsupervised Labelling Module

In lithology identification using well logging data, reliable labels usually rely on core interpretation and comprehensive expert judgment. However, due to the high cost of coring, insufficient coverage of cored intervals, and uncertainty in interwell correlation, manually interpreted labels are often limited in both quantity and continuity. In addition, obvious subjectivity may exist among different geologists. If manual interpretation is solely used to complete missing labels, its efficiency can hardly meet the requirements of large-scale modeling. More importantly, differences in interpretation scale and empirical bias may introduce systematic noise, thereby

affecting the transferability and generalization stability of the model. Therefore, an unsupervised labeling module is introduced during model training. Without relying on manual labels, this module generates pseudo label sequences with relatively high consistency based on the probabilistic statistical structure of multiple logging parameters, providing learnable supervision signals and an entry point for sequence prior constraints for the subsequent BiLSTM-CRF model.

As a probabilistic generative model, the Gaussian mixture model assumes that the observed samples are generated from a linear combination of multiple Gaussian distributions, with each Gaussian component corresponding to a potential lithology type.

(1) Modeling of the logging feature space and probabilistic assumptions

Let the standardized feature vector of multiple logging parameters be defined as:

$$x_i = (x_1, x_2, \dots, x_i) \in R^d \quad (1)$$

where i denotes the dimension index of the selected logging curve parameter. The GMM assumes that the overall sample distribution can be represented as a weighted superposition of K Gaussian components, and its probability density function is expressed as:

$$p(x_i) = \sum_{k=1}^k \pi_k N(x_i | \mu_k, \Sigma_k) \quad (2)$$

where π_k is the mixture coefficient, and μ_k and Σ_k denote the mean and covariance of the k th component, respectively.

(2) EM estimation and soft clustering output

The model parameters are estimated iteratively using the expectation maximization algorithm. In the E step, the posterior probability that a sample belongs to each component is calculated as:

$$vik = \frac{\pi_k N(x_i | \mu_k, \Sigma_k)}{\sum_{j=1}^k \pi_k N(x_i | \mu_j, \Sigma_j)} \quad (3)$$

where π_k represents the mixture coefficient of the k th Gaussian component, satisfying $\sum_{k=1}^K \pi_k = 1$ and $\pi_k \geq 0$. μ_k and Σ_k are the mean vector and covariance matrix of this component, respectively. The EM iteration continues until the increase in log likelihood is lower than a predefined threshold or the maximum number of iterations is reached, thereby ensuring stable convergence of parameter estimation.

(3) Generation of pseudo labels

After GMM convergence, point level pseudo labels are generated according to the maximum posterior probability principle:

$$\hat{y}_i = \operatorname{argmax}(vik) \quad (4)$$

Based on the statistical characteristics of samples within each cluster under geological constraint indicators, and combined with lithology prior knowledge, the clusters are semantically named to form three macro labels, namely C1, C2, and C3. The strategy of first performing unsupervised grouping and then conducting geological semantic merging can reduce the noise caused by fine scale lithology imbalance, while maintaining the consistency of logging responses within each cluster. This provides favorable conditions for the subsequent BiLSTM-CRF model to learn stable transition patterns at the sequence level.

$$ci = \max_k(vik) \quad (5)$$

As shown in Fig. 7, the unsupervised clustering results based on the Gaussian mixture model present relatively clear lithology distribution characteristics in the principal component space. After dimensionality reduction by principal component analysis, different lithology samples form three major cluster centers in the feature space, corresponding to three lithology assemblages, with distinct spatial positions and distribution patterns. The cluster on the left is mainly dominated by clastic lithologies. Its sample distribution is relatively concentrated, but a certain degree of overlap exists in the boundary area, reflecting the continuous variation of logging responses within clastic rocks. The cluster in the middle mainly covers carbonate lithologies, with a larger number of samples and a wider distribution range, indicating the complexity and heterogeneity of this lithology type in terms of multiple logging parameter responses. The clustering visualization results indicate that the GMM can effectively identify the statistical distribution differences among different lithologies in the feature space under unsupervised conditions, while retaining reasonable probability overlap characteristics in lithology transition zones. This provides initial sample labels for the subsequent lithology identification model with temporal constraints.

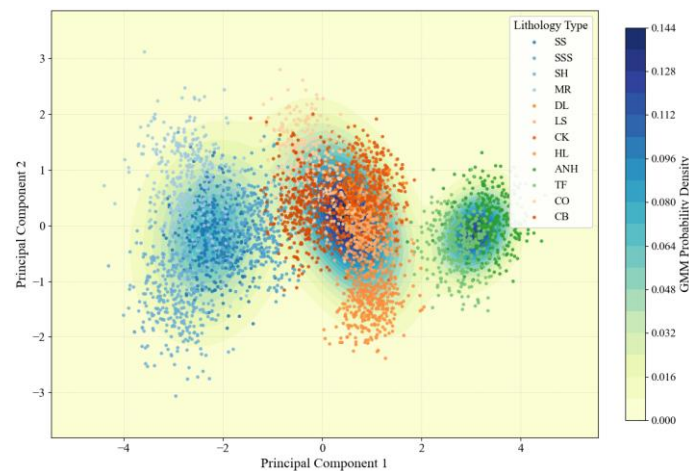


Fig.7 Lithology clustering distribution based on GMM analysis

3.4.2 BiLSTM-CRF-Based Lithology Identification Module

After obtaining the weighted feature vectors, the GMM is used to generate point level lithology pseudo labels under unsupervised conditions, thereby providing supervision signals for model training. However, the GMM is essentially a point wise discrimination method and does not incorporate the continuity of lithology along the depth direction or the constraints of stratigraphic sequence structure. In actual formations, lithology usually exhibits interval based continuous distribution, and the label sequence has significant contextual dependence. If only point level classification is used, discrete misclassification with a salt and pepper pattern may easily occur in thin interbeds, noise disturbed intervals, or transition zones. Therefore, a BiLSTM-CRF identification module is constructed in this section. The BiLSTM is used to learn bidirectional temporal features from logging sequences, while the CRF performs global decoding of label transitions.

(1) BiLSTM temporal encoding and gating mechanism

In the LSTM unit, the forget gate controls which historical information from the logging data should be retained or discarded. The forget gate reads the output h_{t-1} from the previous time step and the current input x_t , generates the forget vector f_t , and performs element wise multiplication with the previous cell state C_{t-1} , thereby filtering historical information, as expressed in Eq. (6).

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (6)$$

where W_f is the weight coefficient matrix of the forget gate corresponding to the logging curve information, b_f is the bias of the forget gate, $\sigma(\cdot)$ denotes the sigmoid function, and f_t is the decision vector of the forget gate. The input gate determines the amount of new information written at the current time step. In the input gate, the tanh layer creates a new candidate state vector $\tilde{C}_t$, which is added to the new state to replace the old information that needs to be discarded, thereby determining the updated information, as expressed in Eq. (7).

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (7)$$

where W_C is the weight coefficient matrix corresponding to the input gate, and b_C is the bias matrix of the input gate. The cell state of the logging parameter sequence is then updated from C_{t-1} to C_t . Specifically, the old state is multiplied by f_t to discard irrelevant information, and the new candidate information $i_t \cdot \tilde{C}_t$ is added according to the update degree determined in this step, as expressed in Eq. (8).

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (8)$$

Finally, the output gate determines the output at the current time step. First, the activation of the output gate is calculated. Then, the updated cell state is processed by

the tanh function and multiplied by the output of the sigmoid gate to obtain the hidden output, as expressed in Eqs. (9) and (10).

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (9)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (10)$$

where W_o is the weight coefficient matrix corresponding to the input information x_t in the output gate, and b_o is the bias of the output gate.

(2) CRF sequence constraint and global decoding

Although BiLSTM can effectively extract temporal features from logging curves, directly applying softmax for point wise classification after the BiLSTM output still implies an assumption of label independence. This makes it difficult to explicitly constrain the transition patterns of lithology along the depth direction. To introduce global constraints into the lithology sequence, a conditional random field model, namely CRF, is added after the BiLSTM output layer to jointly model the lithology label sequence. The CRF regards the lithology prediction of the entire depth sequence as a whole. By simultaneously considering the classification score of each point and the transition relationship between adjacent labels, it globally optimizes the prediction results.

Under the linear chain CRF framework, given the feature sequence H , the conditional probability of the label sequence Y is defined as:

$$P(Y|H) = \frac{1}{Z(H)} \exp(-E(Y, H)) \quad (11)$$

where Y denotes the set of lithology classes, and H represents the output features of the BiLSTM for the sequence. $E(Y, H)$ is the energy function of the sequence, and $Z(H)$ is the normalization factor.

The model parameters are learned by maximizing the log conditional likelihood of the true lithology sequence:

$$T(\theta) = \sum_{n=1}^N \log P(Y^{(n)} | H^{(n)}) \quad (12)$$

where N denotes the number of training samples, and θ represents the set of CRF parameters, including the emission parameters and label transition parameters. Through this optimization process, the model can adaptively learn the transition patterns among different lithology classes, enabling the final prediction results to preserve local recognition capability while satisfying the global structural constraints of the sequence.

3.4.3 GMM-BiLSTM-CRF Model Training

This section describes the complete training procedure of the GMM-BiLSTM-

CRF model. Overall, a two stage collaborative strategy is adopted. First, GMM generates lithology pseudo labels and their uncertainty measures in an unsupervised manner. Then, the pseudo labels are used as supervisory signals to train the BiLSTM-CRF model, enabling temporal feature learning and sequence level global decoding of well log sequences. This workflow reduces the dependence on manually refined annotations and, with the structural constraints of CRF, suppresses the propagation of pseudo label noise near lithological boundaries.

To improve training stability and parameter optimization efficiency, this study employs the gradient based adaptive optimization algorithm Adam (Adaptive Moment Estimation) to update the model parameters (Le-Duc T et al., 2024). Adam integrates first order and second order moment information to dynamically adjust the learning step size of different parameters, making it suitable for optimization problems involving significant scale differences and large gradient fluctuations in well log sequences. The parameter update process is described as follows.

Taking the minimization of the loss function as the training objective, at the t -th iteration, the objective function is first calculated based on the current mini batch samples, and the trainable parameters are updated as follows:

(1) Gradient calculation

The moment vectors are first initialized. Before the model converges, iterative calculation is repeatedly performed, and a new gradient value is obtained in each iteration, as shown in Eq. (13).

$$g_t = \nabla_{\theta} f_t(\theta_{t-1}) \quad (13)$$

where g_t denotes the gradient information of the mini batch samples, $f_t(\theta)$ denotes the objective function, and (x_t, y_t) denotes the features and labels of a mini batch sample. In each iteration, the gradient $g_t = \nabla_{\theta} f_t(\theta_{t-1})$ is calculated, where $f_t(\theta)$ represents the current loss function and θ_{t-1} is the parameter value from the previous iteration.

(2) Momentum and weighted averaging

After the gradient is obtained, Adam performs exponential weighted averaging on the first order and second order moments of the gradient, respectively, to characterize the historical trend and fluctuation magnitude of the gradient, as shown in Eqs. (14-15).

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t \quad (14)$$

$$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2 \quad (15)$$

where m_t denotes the first moment estimate vector, v_t denotes the second moment estimate vector, and β_1 and β_2 denote the decay rates, which are set to 0.9 and 0.999, respectively. They control the response strength of the momentum term and the smoothing term to the current gradient.

(3) Bias correction

Since the momentum and weighted averages may be biased at the initial stage, Adam adjusts them through bias correction. The bias corrected first moment estimate and bias corrected second moment estimate are calculated as shown in Eqs. (16-17).

$$\hat{m}_t = m_t / (1 - \beta_1^t) \quad (16)$$

$$\hat{v}_t = v_t / (1 - \beta_2^t) \quad (17)$$

where $\hat{m}_t$ and $\hat{v}_t$ denote the bias corrected first moment estimate vector and second moment estimate vector, respectively. α denotes the learning rate. ϵ denotes the smoothing term, which is usually set to 10^{-8} to prevent division by zero.

(4) Weight parameter update

Finally, the corrected first order and second order moments are used together to determine the parameter update, and the model parameters are iteratively optimized with the learning rate α . The specific calculation is given in Eq. (18).

$$\theta_t = \theta_{t-1} - \alpha \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon) \quad (18)$$

Figure 8 shows the loss convergence process of the GMM-BiLSTM-CRF model during training. It can be observed that both the training loss and validation loss decrease continuously with the increase in training epochs, showing a clear overall convergence trend. In the first 10 epochs, the loss decreases rapidly, indicating that the model can quickly learn the main pattern features of the samples at the early training stage. After 10 epochs, the decrease in loss becomes gradual, suggesting that the model enters a stable optimization stage. After approximately the 25th epoch, both the training loss and validation loss tend to stabilize, with only slight fluctuations, indicating that no significant overfitting occurs during training and that the model has good generalization stability. Overall, the constructed GMM-BiLSTM-CRF model can effectively learn the vertical sequence features and lithology transition patterns from well log curves, demonstrating strong training stability.

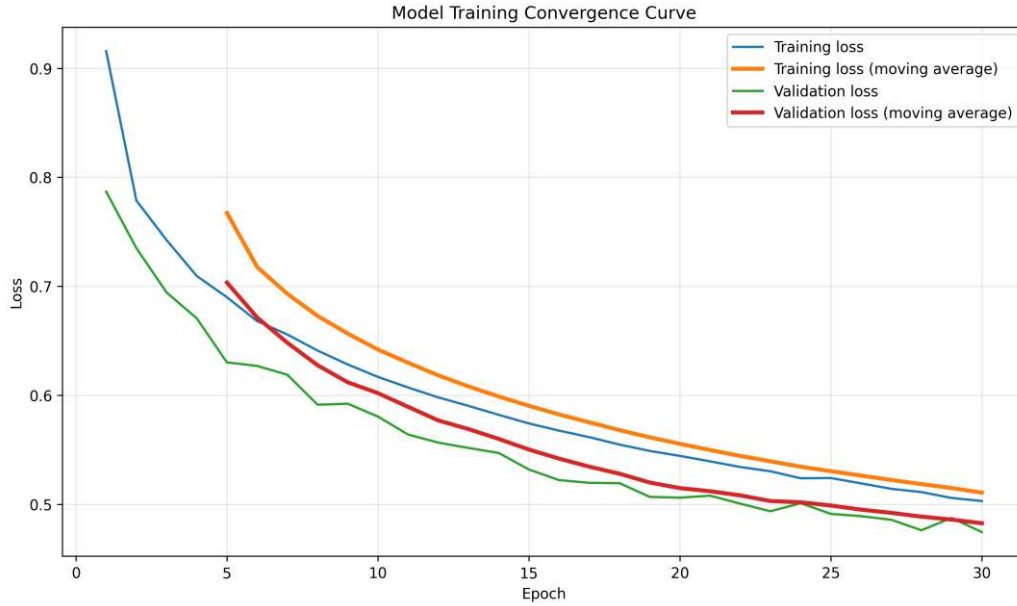


Fig.8 Training iteration convergence curve

4.Results

The experimental samples were derived from real well log data from oilfields in the Viking Graben area of the Norwegian Sea. To comprehensively verify the effectiveness of the proposed method in this chapter, model metric evaluation experiments, model error experiments, and multi well lithology prediction experiments were designed and conducted. The experiments systematically evaluated the stability of the constructed model from multiple perspectives, including overall performance, labeling accuracy and efficiency, error distribution, missing well log data, and multi well applicability. The specific configuration of the experimental equipment was as follows: the CPU was an Intel Xeon Silver 4210R, the memory was 64 GB, and the GPU was an RTX 6000/8000. The operating system was Ubuntu 20.04.3, and the experimental framework was PyTorch 1.7.1.

4.1 Classification-Metric Evaluation

To comprehensively evaluate the classification performance of the proposed method in the well log lithology identification task, LSTM, BiLSTM, BiLSTM-CRF, and GMM-BiLSTM were selected as comparison models. Under the same training and test set partitioning conditions, Precision, Recall, Accuracy, and F1-score were used to evaluate the overall classification performance of the models. In addition, the confusion matrix was used to analyze the identification results among different lithology classes.

As shown in Figures 9 and 10, the results of the training and test sets are generally consistent. The proposed GMM-BiLSTM-CRF model achieves the highest values in

Precision, Recall, Accuracy, and F1-score, indicating that the model has good comprehensive performance in lithology identification. For the test set, the Precision, Recall, Accuracy, and F1-score of GMM-BiLSTM-CRF are 91.36%, 90.82%, 93.18%, and 90.97%, respectively, which are higher than those of LSTM, BiLSTM, GMM-BiLSTM, and BiLSTM-CRF. Compared with the models that introduce GMM or CRF alone, the fused model still shows further improvement in all metrics, indicating that the lithology pseudo labels generated by GMM and the sequence constraints of CRF can play complementary roles. Among them, the results of GMM-BiLSTM and BiLSTM-CRF are relatively close, and the two models show different advantages in different metrics, suggesting that unsupervised labeling and sequence constraints can both improve model performance but contribute in different ways. In the training set, the Precision, Recall, Accuracy, and F1-score of GMM-BiLSTM-CRF are 92.17%, 91.54%, 94.27%, and 91.85%, respectively. The differences between the training and test results are small, indicating that the model does not simply fit the training samples but maintains stable identification capability on the test data. Overall, GMM-BiLSTM-CRF performs better in well log sequence feature extraction and lithology class discrimination, providing reliable technical support for intelligent lithology identification under complex well log conditions.

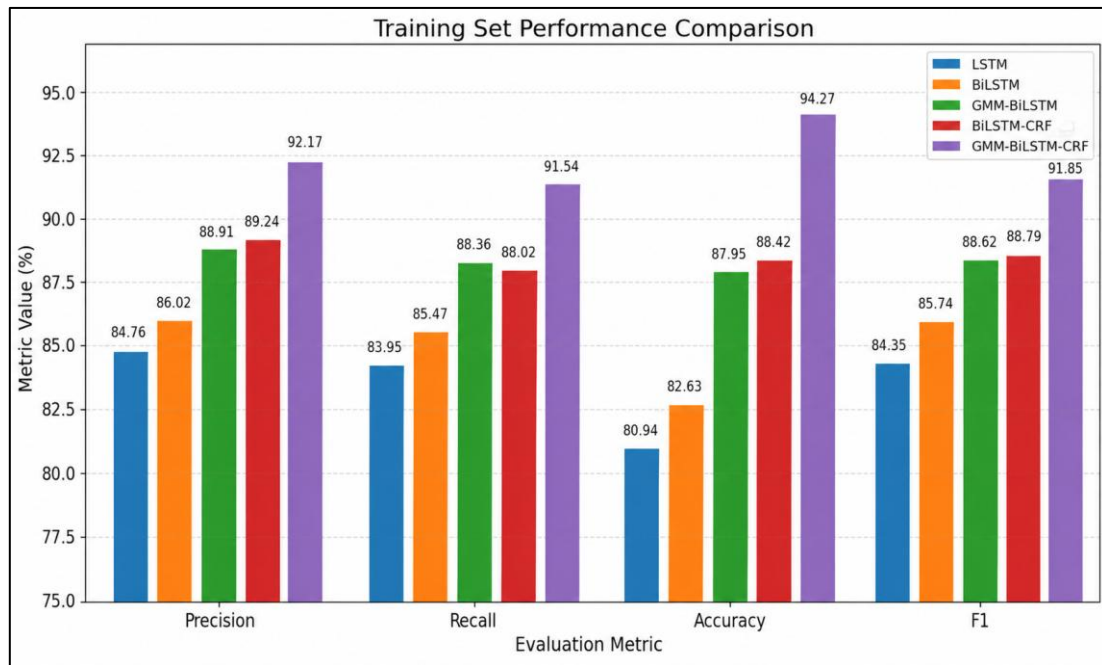


Fig. 9. Identification results of the training set

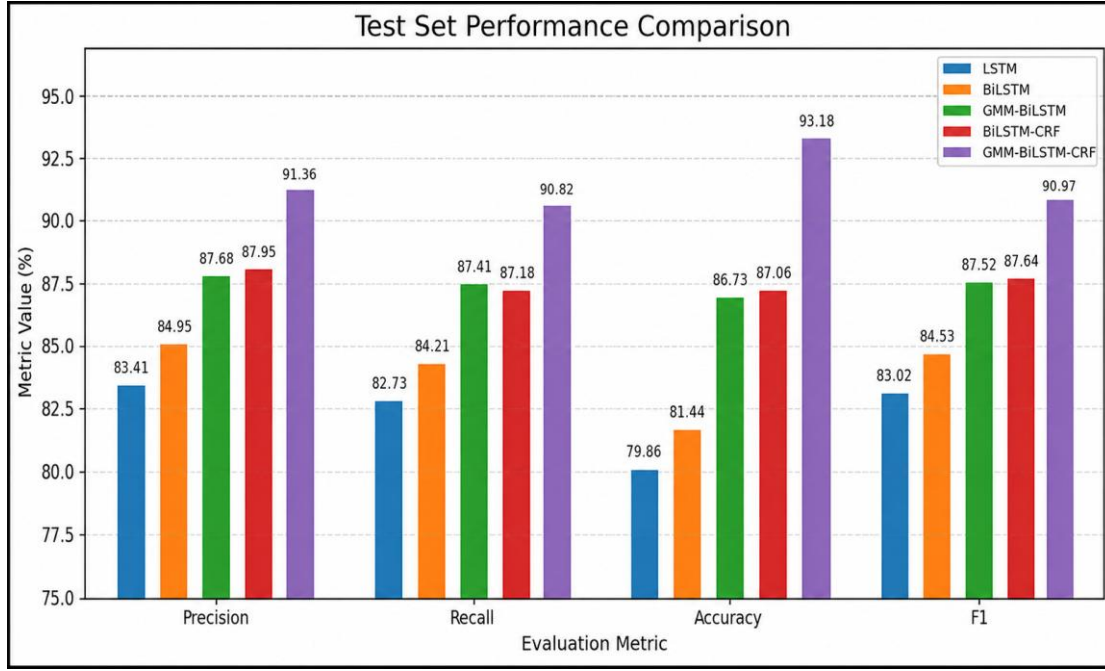


Fig. 10. Recognition performance on the test set

Based on the confusion matrices shown in Figure 11, different models exhibit certain differences in their identification performance for various lithology classes. LSTM and BiLSTM show relatively low scores for some lithology classes and well log parameter combinations, indicating that a single sequence model still has limited capability in distinguishing complex lithological responses. After introducing GMM, the overall identification performance of GMM-BiLSTM is improved, suggesting that the pseudo labels generated by unsupervised clustering can enhance the model's ability to learn the distribution characteristics of lithology classes. In contrast, the proposed GMM-BiLSTM-CRF method maintains higher and more stable identification scores for most lithology classes, indicating that it has a more balanced discrimination capability for different lithology types.

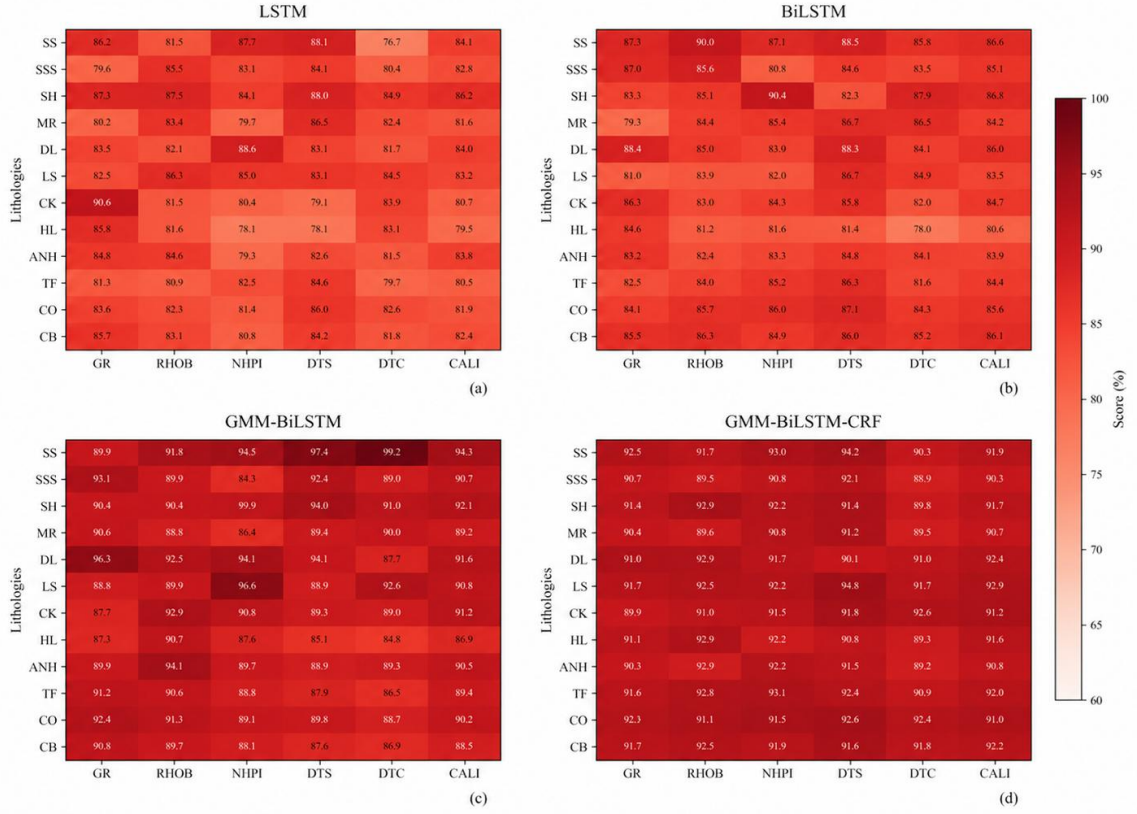


Fig.11. Convergence of the Confusion Matrix

4.2 Error Analysis Evaluation

As shown in Table 4, the four models exhibit relatively consistent variation patterns in the error metrics on both the training and test sets. Overall, with the continuous optimization of the model structure, RMSE, MSE, and MAE gradually decrease, while R^2 continues to increase, indicating that the model's ability to characterize lithological features is progressively enhanced and that the identification results become more stable.

On the test set, BiLSTM improves all metrics compared with LSTM, indicating that the bidirectional temporal structure can more fully utilize the forward and backward contextual information of well log curves along the depth direction. GMM-BiLSTM further reduces the errors on this basis, suggesting that after introducing clustering results, the model can more adequately represent the internal structural characteristics of lithology samples. In contrast, GMM-BiLSTM-CRF achieves the best performance across all metrics on the training set, indicating that after adding sequence labeling constraints to bidirectional feature extraction, the model can learn lithological variation characteristics more effectively.

567

Table 4. Comparison of Overall Model Error Results

Lithofacies type	Error	LSTM	BiLSTM	GMM-BiLSTM	GMM-BiLSTM-CRF
Training Set	RMSE	0.286	0.253	0.218	0.181
	MSE	0.082	0.064	0.048	0.033
	MAE	0.207	0.174	0.139	0.103
	R ²	0.872	0.895	0.927	0.954
Test Set	RMSE	0.296	0.262	0.226	0.190
	MSE	0.088	0.070	0.052	0.037
	MAE	0.214	0.182	0.147	0.111
	R ²	0.861	0.887	0.918	0.946

568

569

570

571

572

573

Based on the single well error results shown in Figure 12, the error variation trends of different models in the four randomly selected wells are generally consistent. LSTM has the highest error, followed by BiLSTM, while GMM-BiLSTM further reduces the error. GMM-BiLSTM-CRF achieves the lowest MAE, MSE, and RMSE values. This indicates that the proposed method has stable identification performance across different wells and does not achieve good results only in a specific well or local samples.

574

575

576

577

578

579

580

Specifically, in random well 1, the MAE, MSE, and RMSE of GMM-BiLSTM-CRF are 0.116, 0.040, and 0.200, respectively, all of which are lower than those of the other models. In random well 2, the three error metrics further decrease to 0.109, 0.036, and 0.190. In random well 3 and random well 4, the model still maintains the lowest errors, with RMSE values of 0.187 and 0.184, respectively. In contrast, the RMSE of LSTM remains around 0.302 in the four wells, indicating that its adaptability to inter well differences in lithological responses is relatively limited.

581

582

583

Overall, GMM-BiLSTM-CRF shows lower error levels in all four randomly selected wells, with small fluctuations among different wells, demonstrating good cross well stability.

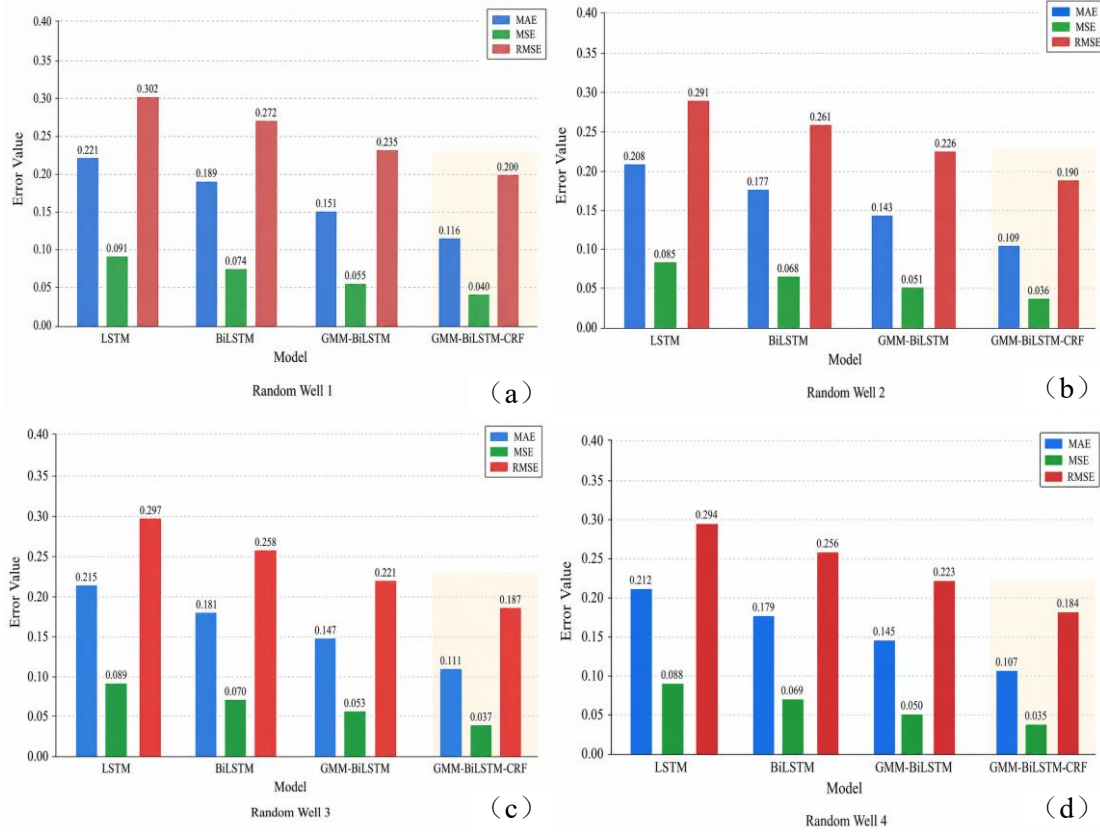


Fig.12. Comparison of Error Metrics of Different Models in four Random Wells

4.3 Well Log Data Missingness Evaluation

To verify the adaptability of the model under incomplete well log data conditions, a robustness experiment with missing well log curves was further conducted. In this experiment, missing ratios of 0%, 5%, 10%, 15%, 20%, 25%, and 30% were set to simulate common problems in actual well log data, such as curve absence, local missing intervals, and data quality degradation. The changes in identification accuracy of GMM, GMM-BiLSTM, BiLSTM-CRF, and GMM-BiLSTM-CRF under different missing ratios were then compared.

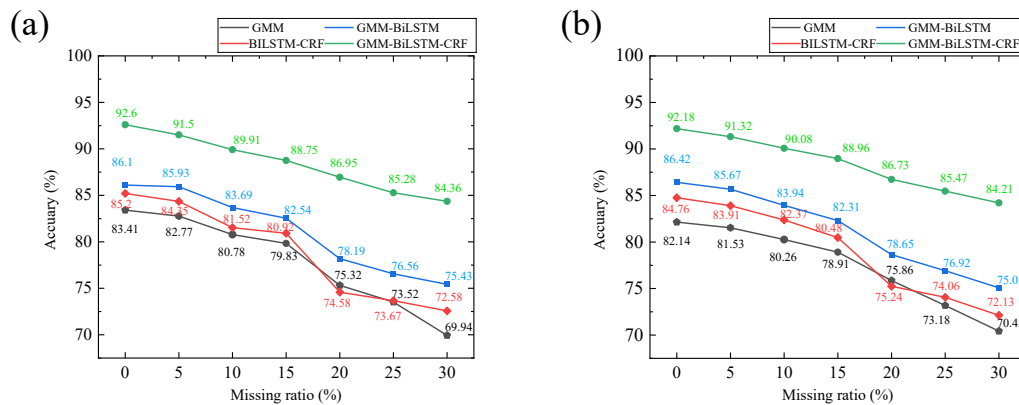


Fig.13. Comparison of identification accuracy under different missing ratios of well log curves in the test set.

As shown in Figure 13, the experimental results indicate that as the missing ratio of well log curves gradually increases from 0% to 30%, the identification accuracy of all models on both the training and test sets decreases to varying degrees. This suggests that missing well log information weakens the models' ability to effectively represent and discriminate lithological well log response characteristics. However, the degree of accuracy degradation differs significantly among different models.

The test set results show that under complete well log curve conditions, the Accuracy of the GMM-BiLSTM-CRF model is 92.18%. When the missing ratio increases to 30%, its Accuracy remains at 84.21%, with a decrease of only 7.97 percentage points. In contrast, the Accuracy values of GMM, BiLSTM-CRF, and GMM-BiLSTM decrease from 82.14%, 84.76%, and 86.42% to 70.42%, 72.13%, and 75.08%, respectively, corresponding to decreases of 11.72, 12.63, and 11.34 percentage points. This indicates that these comparison models are more sensitive to missing well log curves. The training set results show a similar pattern. The Accuracy of GMM-BiLSTM-CRF decreases from 93.74% to 85.63%, with a decrease of 8.11 percentage points, whereas the Accuracy values of the other comparison models all fall below 80% under the 30% missing condition. This further demonstrates that the proposed model has more stable identification performance under missing well log information conditions.

As shown in Figure 14, further comparison between the training and test set results shows that the GMM-BiLSTM-CRF model achieves the highest identification accuracy under all missing ratios, and the performance difference between the training and test sets is small. This indicates that the model does not exhibit obvious overfitting and has good generalization stability. Especially under high missing ratio conditions, the accuracy of the traditional GMM method and deep learning models without sufficient sequence constraint mechanisms decreases more significantly, whereas the GMM-BiLSTM-CRF model still maintains an Accuracy above 84%.

In summary, the proposed GMM-BiLSTM-CRF model demonstrates strong robustness under missing well log curve conditions and can provide methodological support for intelligent lithology identification when actual well log data quality is uneven.

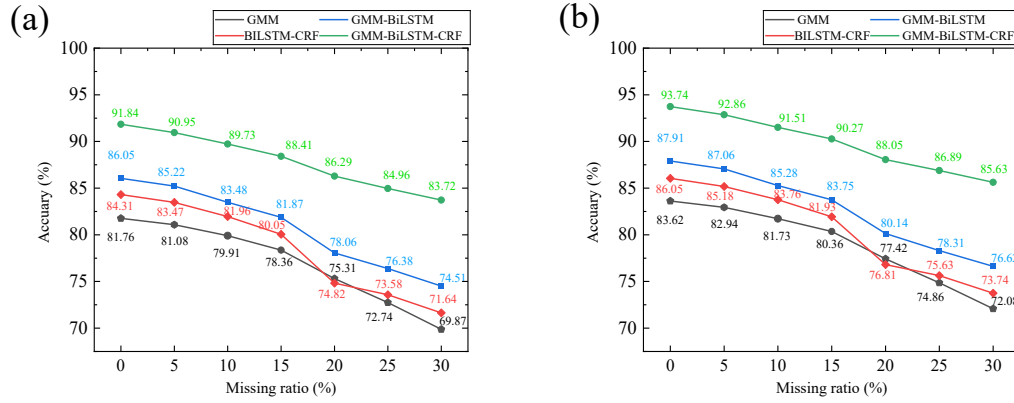


Fig.14. Comparison of identification accuracy under different missing ratios of well log curves in the training set.

4.4 Multi well Lithology Prediction Experiment Evaluation

To further verify the applicability and stability of the proposed method under different inter well conditions, multi well lithology identification experiments were conducted in this section. Statistical analyses of the identification results for 12 lithology classes were performed on two groups of well samples. The first group includes S1, S3, E1, and E3, while the second group includes TW1, TW3, TN1, and TN3. The distribution of identification accuracy for different lithology classes in each well is also presented. The lithology types include SS, SSS, SH, MR, DL, LS, CK, HL, ANH, TF, CO, and CB.

Figure 15 shows the reservoir isopach map and favorable area distribution map. From the performance of different wells, E1 and TN1 show relatively better overall identification results, achieving high accuracy for most lithology classes. In particular, the identification results for major lithology types such as SS, SH, and DL are more prominent. In contrast, S3 and TW3 show slightly lower overall accuracy, but the identification results for various lithology classes still remain at a stable level, without obvious distortion or sharp decline. This indicates that under conditions where inter well differences exist in well log responses, the model can still maintain good identification consistency and has a certain degree of generalization capability.

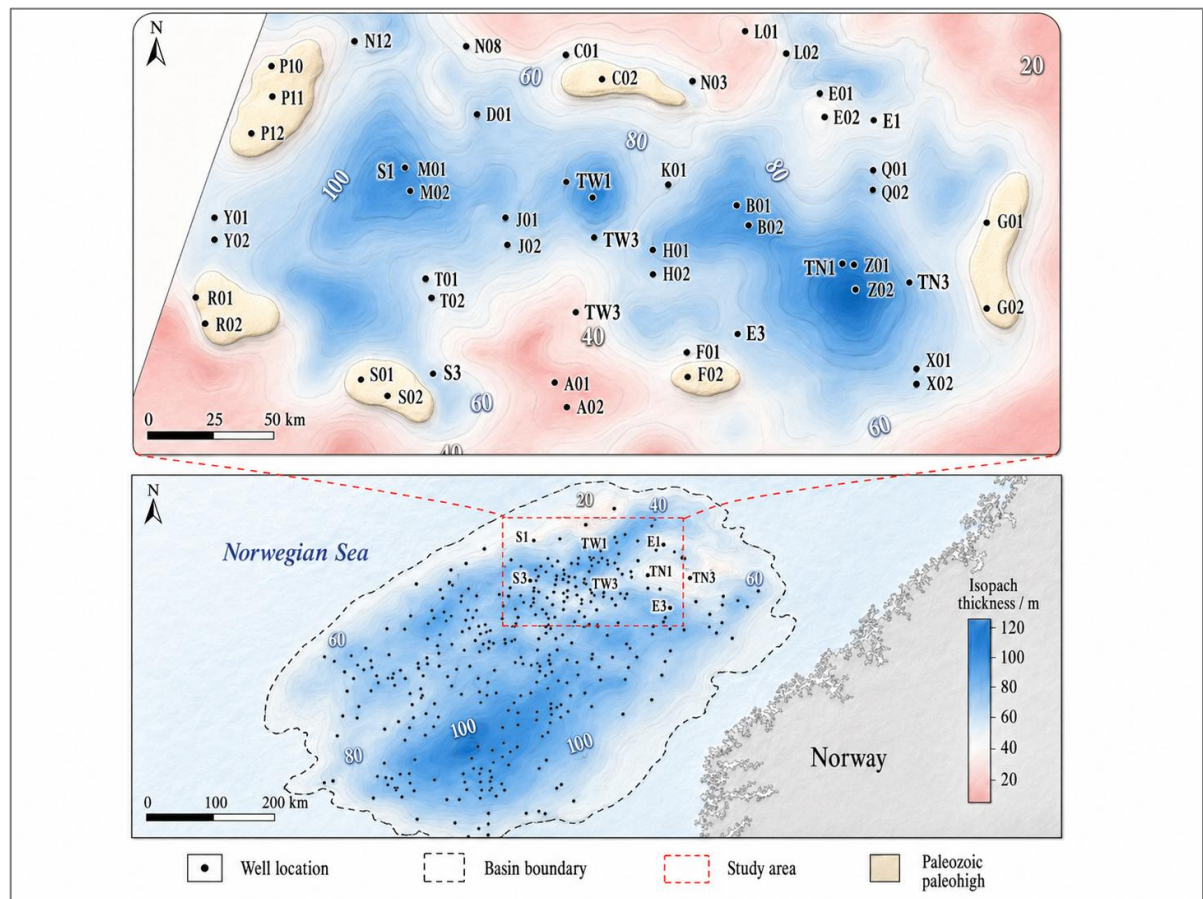


Fig.15. Reservoir isopach map and favorable area distribution map in the study area.

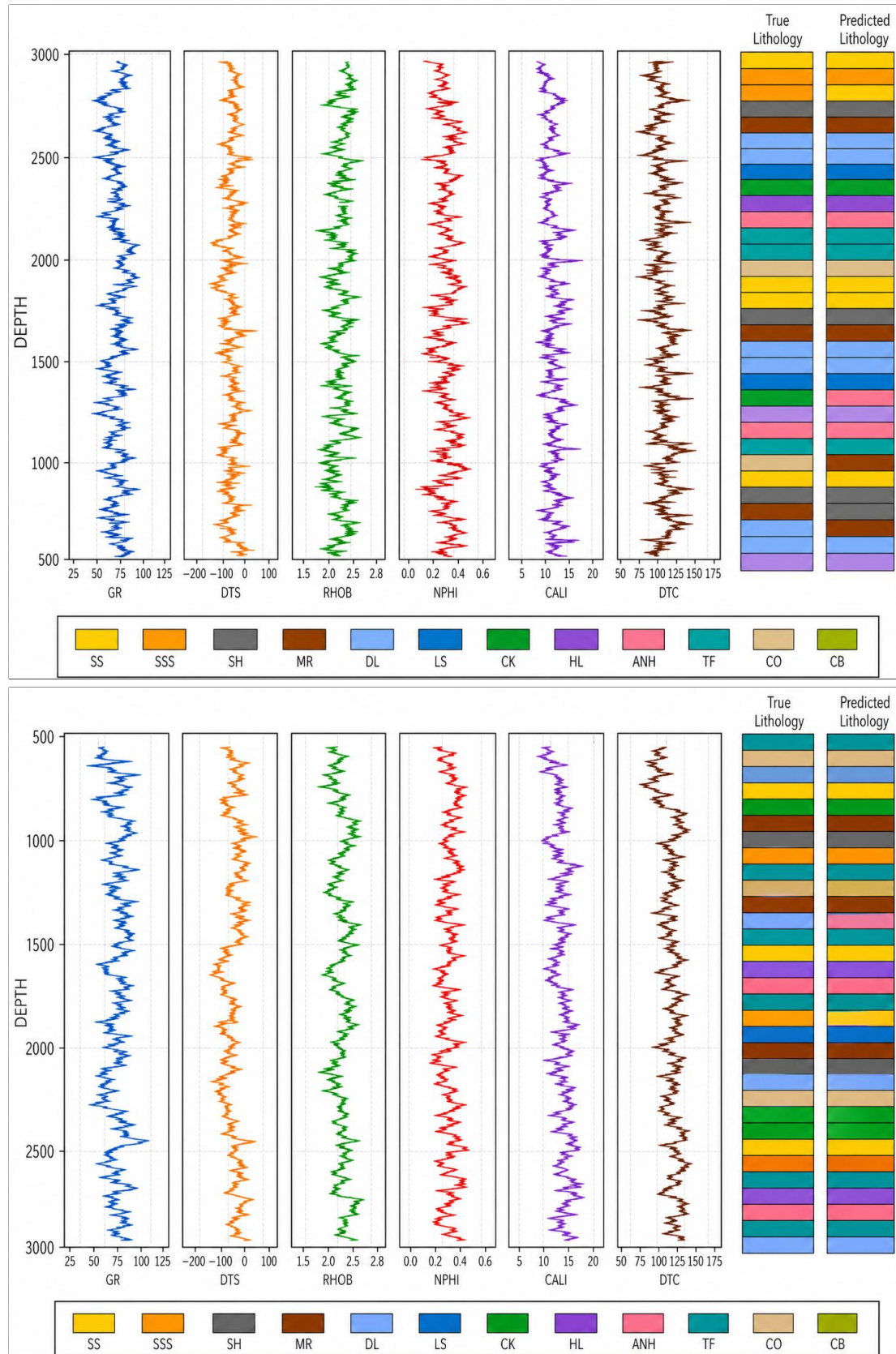


Fig.16. Comparison between original lithology and predicted lithology identification results for two randomly selected single wells

To examine the applicability and stability of the model under different well log

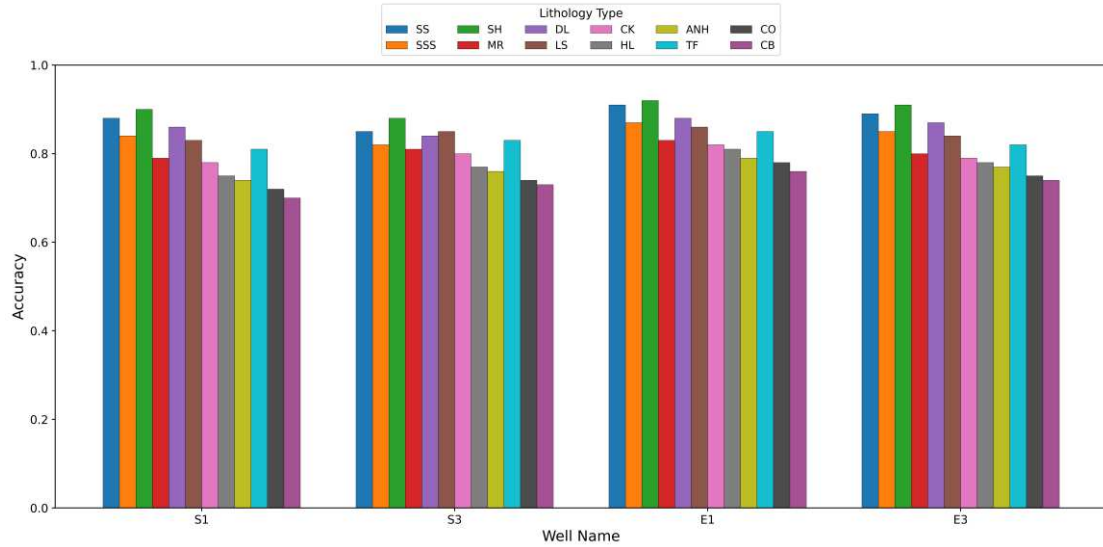
response conditions, Figure 16 presents a comparison of lithology identification results for two wells. For both wells, six well log curves, including GR, DTS, RHOB, NPHI, CAL, and DTC, were used as inputs. The predicted results show high consistency with the original lithology interpretation results in terms of vertical distribution, with an average identification accuracy of 90.1%. Overall, the model can effectively identify the continuous variation of different lithologic units along the depth direction. The predicted lithology is generally consistent with the true lithology in terms of color combinations, boundary positions, and vertical superposition relationships in the main intervals. This indicates that the proposed method can effectively extract nonlinear response features from multiple well log parameters while maintaining good continuity along the depth sequence.

From the perspective of local details, a small number of lithology misclassifications still occur in the two wells, mainly concentrated in thin interbedded intervals where lithology changes frequently or where differences in well log responses are not obvious. Such intervals usually exhibit strong heterogeneity, and the well log responses of different lithologies may overlap to some extent, causing the model to produce local shifts or short interval misidentifications near adjacent lithologic boundaries. Nevertheless, these identification errors do not significantly alter the overall lithologic assemblage characteristics or the recognition results of the main reservoir intervals. The model still maintains good vertical continuity overall and avoids the high frequency fluctuations commonly seen in traditional point by point classification methods, indicating that it has strong sequence constraint capability and good geological consistency in complex well sections.

Figures 17 and 18 show the results of multi well lithology identification. From the perspective of different lithology classes, the identification accuracies of SS, SH, and DL are generally high, and their variations among different wells are relatively small, indicating that the well log response characteristics of these lithologies are relatively clear and that the model has strong discrimination ability for them. The identification results of LS and TF are also relatively stable, with the accuracy in most wells remaining within a relatively high range, indicating that the model also performs well in distinguishing carbonate related lithologies. In contrast, the accuracies of CO, CB, HL, and ANH are relatively lower, and the fluctuations among wells are slightly larger. This suggests that these lithologies may be relatively weaker in terms of sample size, differences in well log responses, or the clarity of class boundaries, making them more difficult to identify. In particular, minority lithologies such as CO and CB are more easily affected by class imbalance and local response fluctuations, resulting in a certain

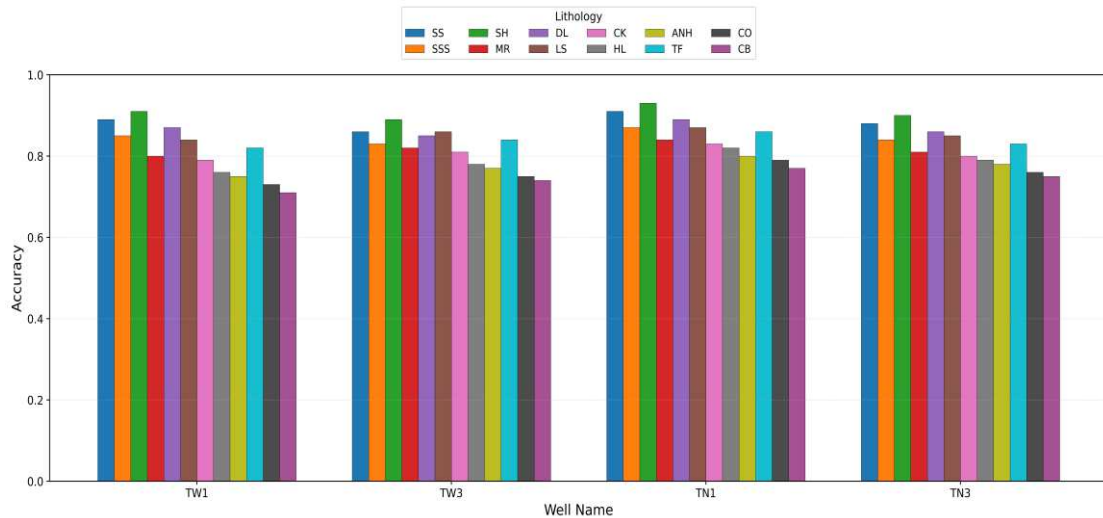
decline in identification accuracy.

Fig.17. Lithology identification results for the first group of wells



Overall, the multi well lithology experiment results indicate that the proposed method achieves good identification performance across different wells and lithology classes. For dominant lithologies and classes with relatively distinct response characteristics, the model can achieve high accuracy identification. For minority lithologies or lithologies with complex boundary transitions, although the accuracy decreases slightly, the overall results remain within an acceptable range. This experiment further demonstrates that the proposed method has good stability and generalization capability in cross well identification tasks, providing a reliable lithological basis for subsequent reservoir distribution prediction and favorable interval identification.

Fig.18. Lithology identification results for the second group of wells



5. Discussion

The experimental results demonstrate that the intelligent well log lithology identification method integrating GMM, BiLSTM, and CRF achieves good identification accuracy, error control capability, and cross well applicability under complex geological conditions. Compared with LSTM, BiLSTM, BiLSTM-CRF, and GMM-BiLSTM, the proposed method obtains better overall performance on both the training and test sets. On the test set, the Precision, Recall, Accuracy, and F1-score of the GMM-BiLSTM-CRF model reach 91.36%, 90.82%, 93.18%, and 90.97%, respectively, indicating that the model can more effectively characterize the nonlinear relationships between multiple well log parameters and lithology classes while maintaining high classification stability under complex lithological assemblages.

The error experiments further verify the stability and generalization capability of the proposed method. The test set results show that the GMM-BiLSTM-CRF model achieves the lowest RMSE, MSE, and MAE, with RMSE reduced to 0.190 and R^2 reaching 0.946. This indicates that the proposed method has clear advantages in reducing prediction errors and improving fitting consistency. Compared with the LSTM model, the proposed model performs better not only in classification metrics but also in error metrics, showing stronger stability. The small differences between the training and test set metrics suggest that the model does not simply fit the training samples but still maintains good identification capability on unseen data. The error comparison results of four randomly selected wells also show that GMM-BiLSTM-CRF maintains low MAE, MSE, and RMSE values across different wells, indicating that its advantage is not limited to a single well section or local samples but has certain cross well applicability.

It should be noted that the identification performance still differs among lithology classes. For dominant lithologies such as SS and SH, the model usually achieves higher identification accuracy because of relatively sufficient samples and clearer well log response characteristics. For minority lithologies such as CO and CB, the accuracy is relatively lower and the inter well fluctuations are more obvious. Although a sample balancing strategy is adopted in this study, class imbalance and insufficient minority class samples remain important factors affecting identification performance. In addition, the pseudo labels generated by GMM are essentially approximate labels in a statistical sense, and their quality may be affected by clustering structure, class boundaries, and differences in sample distribution. This is also a key issue that needs to be addressed for further improvement of model performance.

From the multi well lithology prediction results, the proposed model can maintain

good consistency between the predicted lithology columns and the original lithology interpretation results. In the identification experiments of two single wells, the predicted lithology is generally consistent with the original lithology in terms of major lithological assemblages, boundary positions, and vertical superposition relationships, with an average identification accuracy of 90.1%. This indicates that the model can maintain stable identification performance under different well log curve patterns and lithological assemblage conditions. For practical reservoir evaluation, this stability is important because well log responses often vary significantly among different wells due to differences in sedimentary environment, diagenesis, borehole conditions, and logging data quality. The model still achieves good lithology identification results under multi well conditions, suggesting that it has certain potential for engineering application.

6.Conclusions

This study proposes an intelligent well log lithology identification method integrating GMM-BiLSTM-CRF, aiming to address problems such as strong dependence on labels and imbalanced sample distribution in lithology identification under complex geological conditions. The experimental results show that the proposed method outperforms the comparison models in overall classification performance, error control, and multi well adaptability, demonstrating its ability to more effectively characterize the vertical variation patterns and lithological transition features in well log data.

Future Prospects

Although minority classes and complex transitional lithologies still present certain challenges, the overall results maintain good stability and acceptable generalization boundaries. The multi well lithology prediction experiments further confirm that the method has good robustness for cross well application and potential for engineering promotion.

In summary, this study integrates unsupervised feature clustering, deep temporal modeling, and sequence constraint optimization, providing an effective technical pathway for improving well log lithology identification under complex reservoir conditions. This method not only provides a new technical approach for automatic well log lithology identification but also offers a reliable basis for subsequent reservoir evaluation and geological interpretation.

Author Contributions

Author Contributions: Conceptualization, Z.M. and L.S.; methodology, Z.M. and L.S.; software, Z.M.; validation, Z.M., X.Lv. and G.Z.; formal analysis, Z.M. and H.M.; investigation, X.Lv., G.Z. and G.F.; resources, L.S. and G.F.; data curation, Y.S., X.Liu. and Y.W.; writing—original draft preparation, Z.M.; writing—review and editing, L.S., H.M. and G.Z.; visualization, Z.M. and X.Liu.; supervision, L.S.; project administration, L.S. and G.F.; funding acquisition, L.S. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability

Data will be made available on request.

Conflicts of Interest

The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Name
Adam	Adaptive Moment Estimation
BiLSTM	Bidirectional Long Short-Term Memory
CAL	Caliper
CRF	Conditional Random Field
DTC	Compressional-wave slowness
DTS	Shear Wave Slowness
GMM	Gaussian Mixture Model
GMM-BiLSTM	Gaussian Mixture Model with Bidirectional Long Short-Term Memory
GMM-BiLSTM-CRF	Gaussian Mixture Model with Bidirectional Long Short-Term Memory and Conditional Random Field
GR	Gamma Ray log
NPHI	Neutron Porosity
RHOB	Bulk Density

7.Reference

- Badley, M.E., Price, J.D., Dahl, C.R., Agdestein, T.: The structural evolution of the northern Viking Graben and its bearing upon extensional modes of basin formation. *Journal of the Geological Society* 145(3), 455-472 (1988). <https://doi.org/10.1144/gsjgs.145.3.0455>
- Bhattacharya, S., Carr, T.R., Pal, M.: Comparison of supervised and unsupervised approaches for mudstone lithofacies classification: case studies from the Bakken and Mahantango-Marcellus Shale, USA. *Journal of Natural Gas Science and Engineering* 33, 1120-1133 (2016). <https://doi.org/10.1016/j.jngse.2016.04.055>
- Bian, H.: Coal structure identification method based on Bayesian-optimized CNN-BiLSTM-Attention.
- Brekke, H., Sjulstad, H.I., Magnus, C., Williams, R.W.: Sedimentary environments offshore Norway - an overview. In: Martinsen, O.J., Dreyer, T. (eds.) *Sedimentary Environments Offshore Norway - Palaeozoic to Recent*. Norwegian Petroleum Society Special Publications, vol. 10, pp. 7-37. Elsevier, Amsterdam (2001). [https://doi.org/10.1016/S0928-8937\(01\)80006-0](https://doi.org/10.1016/S0928-8937(01)80006-0)
- Camacho, L., Douzas, G., Bacao, F.: Geometric SMOTE for regression. *Expert Systems with Applications* 193, 116387 (2022). <https://doi.org/10.1016/j.eswa.2021.116387>
- Chang, J., Kang, Y., Li, Z., Zheng, W.X., Lv, W., Feng, D.-Y.: Cross-domain lithology identification using active learning and source reweighting. *IEEE Geoscience and Remote Sensing Letters* 19, 1-5 (2022). <https://doi.org/10.1109/LGRS.2020.3041960>
- Chen, H.: Research progress in fine characterization of complex lithologic reservoirs.
- Chen, J., Gui, Z., Rui, Y., Zhao, X., Pan, X., Wang, Q., Pu, Y., Zheng, L., Liu, M.: A dual attention-based deep learning model for lithology identification while drilling. *Journal of Rock Mechanics and Geotechnical Engineering* 18(2) (2025). <https://doi.org/10.1016/j.jrmge.2025.03.051>
- Chen, T., Xu, R., He, Y., & Wang, X. (2017). Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications*, 72, 221-230.
- Cui, W.: Research and application of a GWO-BiLSTM-based lithology identification method.
- Dunham, M.W., Welford, J.K.: Improved well log classification using semi-supervised algorithms. *SEG Technical Program Expanded Abstracts*, 2398-2402 (2019). <https://doi.org/10.1190/segam2019-3208817.1>
- Fetni, Y., Ameer-Zaimeche, O., Wood, D. A., Heddarn, S., Al-Mudhafar, W. J., & Kechiched, R. (2025). Explainable machine-learning workflow to improve real-time lithofacies identification using mudlogging data from a heterogeneous clastic reservoir sequence. *Computational Geosciences*, 29, 56. <https://doi.org/10.1007/s10596-025-10394-2>

827 Fu, Q.: Lithology log classification of shale reservoirs based on the BiLSTM-Attention model.

828 Guo, Y., Li, Z., Lin, W., Zhou, J., Feng, S., Zhang, L., Liu, F.: Automatic lithology identification
829 method based on efficient deep convolutional network. *Earth Science Informatics* 16(2), 1359-
830 1372 (2023). <https://doi.org/10.1007/s12145-023-00962-4>.

831 He, Y., Li, W., Dong, Z., Zhang, T., Shi, Q., Wang, L., Wu, L., Qian, S., Wang, Z., Liu, Z., Lei, G.:
832 Lithologic identification of complex reservoir based on PSO-LSTM-FCN algorithm. *Energies*
833 16(5), 2135 (2023). <https://doi.org/10.3390/en16052135>

834 Jin, H., Ci, Y., Xuan, L.: Research status and prospects of intelligent logging lithology identification.
835 *Measurement Science and Technology* 36(1), 012010 (2025). [https://doi.org/10.1088/1361-](https://doi.org/10.1088/1361-6501/ad9cad)
836 [6501/ad9cad](https://doi.org/10.1088/1361-6501/ad9cad)

837 Kiselev, V.Y., Andrews, T.S., Hemberg, M.: Challenges in unsupervised clustering of single-cell
838 RNA-seq data. *Nature Reviews Genetics* 20, 273-282 (2019). [https://doi.org/10.1038/s41576-018-](https://doi.org/10.1038/s41576-018-0088-9)
839 [0088-9](https://doi.org/10.1038/s41576-018-0088-9)

840 Le-Duc T, Nguyen-Xuan H, Lee J (2024) Sequential motion optimization with short-term adaptive
841 moment estimation for deep learning problems. *Engineering Applications of Artificial Intelligence*
842 129:107593. <https://doi.org/10.1016/j.engappai.2023.107593>

843 Li, Z., Wu, Y., Kang, Y., Lv, W.: Feature-depth smoothness based semi-supervised weighted extreme
844 learning machine for lithology identification. *Journal of Natural Gas Science and Engineering* 96,
845 104306 (2021). <https://doi.org/10.1016/j.jngse.2021.104306>

846 Logging lithology identification of conglomeratic reservoirs based on the XGBoost algorithm.

847 Martin-Fernandez, J.A., Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R.: Advances in
848 principal balances for compositional data. *Mathematical Geosciences* 50(3), 273-298 (2018).
849 <https://doi.org/10.1007/s11004-017-9712-z>

850 Mehana M, Guiltinan E, Vesselinov V, et al. Machine-learning predictions of the shale wells'
851 performance[J/OL]. *Journal of Natural Gas Science and Engineering*, 2021, 88: 103819.
852 DOI:10.1016/j.jngse.2021.103819.

853 Mou, D.: Comparison of three classical machine learning algorithms for volcanic-rock logging
854 lithology identification.

855 Mousavi, S.H.R., Hosseini-Nasab, S.M.: A novel approach to classify lithology of reservoir
856 formations using GrowNet and Deep-Insight with physic-based feature augmentation. *Energy*
857 *Science and Engineering* 12(10), 4453-4477 (2024). <https://doi.org/10.1002/ese3.1895>

858 Mu, Z., Li, C., Liu, Z., Liu, T., Zhang, K., Wang, H., Mu, H., Yang, Y.: An automatic lithofacies
859 sample labeling method by integrating deep learning and Gaussian mixture model. *Earth Science*
860 *Informatics* 18(3), 434 (2025). <https://doi.org/10.1007/s12145-025-01937-3>

- Peng, Y.: Sand-mudstone identification method based on the FL-XGBoost algorithm: a case study of the Niuzhuang area, Shengli Oilfield.
- Prabowo, U.N., Sudarmaji, Sismanto: Multi-parameter post-stack seismic inversion based on the cycle loop-semi-supervised learning. *Computational Geosciences* 29(1), 7 (2025).
- Qi, Y., Luo, S., Tang, S., Ruan, J., Gao, D., Liu, Q., Li, S.: Prediction of carbonate reservoir porosity based on CNN-BiLSTM-Transformer. *Applied Sciences* 15(7), 3443 (2025). <https://doi.org/10.3390/app15073443>
- Ramm - 1992 - Porosity-depth trends in reservoir sandstones Theoretical models related to jurassic sandstones off.pdf[Z/OL]. [2026-02-17]. <https://sci.bban.top/pdf/10.1016/0264-8172%252892%252990066-n.pdf>.
- Redfield, T.F., Hermanns, R.L.: Gravitational slope deformation, not neotectonics: revisiting the Nordmannvikdalen feature of northern Norway. *Norwegian Journal of Geology* 96(4), 251-272 (2016).
- Research progress on artificial-intelligence methods for seismic data processing and interpretation.
- SnapBormann, P., Aursand, P., Dilib, F., Manral, S., Dischington, P.: FORCE 2020 well well log and lithofacies dataset for machine learning competition. *Zenodo Dataset* (2020). <https://doi.org/10.5281/zenodo.4351156>
- Velásquez Hernández CA, Prieto Ortiz FA (2020) A real-time map merging strategy for robust collaborative reconstruction of unknown environments. *Expert Systems with Applications* 145:113109. <https://doi.org/10.1016/j.eswa.2019.113109>
- Wang, Z., Cao, C., Zhu, Y., et al.: Entropy and confidence-based undersampling boosting random forests for imbalanced problems. *IEEE Transactions on Neural Networks and Learning Systems* 31(12), 5178-5191 (2020). <https://doi.org/10.1109/TNNLS.2020.2964585>
- Yang, C., Huang, Q., Li, Z., Liu, K., Hu, F.: Big Data and cloud computing: innovation opportunities and challenges. *International Journal of Digital Earth* 10(1), 13-53 (2017). <https://doi.org/10.1080/17538947.2016.1239771>
- Zhang, X., Wen, J., Sun, Q., Wang, Z., Zhang, L., Peng, L.: Lithology identification technology of logging data based on deep learning model. *Earth Science Informatics* 16(3), 2545-2557 (2023). <https://doi.org/10.1007/s12145-023-01051-2>