# Spatial–Stratigraphic Information and Dynamic Range Attention Assist Well-Logging Lithological Interpretation

Keran Li, Jinmin Song, Shugen Liu, Zhiwu Li, Di Yang, Wei Chen, Xin Jin, Chunqiao Yan, and Shan Ren

***Abstract*— Time-series models, particularly CNN–bidirectional long short-term memory (BiLSTM) architectures, have shown advances in the lithological interpretation of well-logging data. However, CNN and attention mechanisms face challenges in training efficiency and predicting precision. To improve this deficiency, a dynamic and lightweight attention mechanism and a strategy that combines geological/spatial information have been proposed. This study introduces two novel enhancements: the spatial and stratigraphic information processing (shortened as Spatial and Strat) method and the dynamic range attention (DRA) mechanism. Spatial–stratigraphic context (SSP) integrates geological context by encoding depositional sequences as time series. DRA is a lightweight attention module that adaptively adjusts local attention ranges based on global context. Experiments on a collected dataset from the eastern Sichuan Basin (13 wells and 14 587 labeled samples) demonstrate that the proposed DRA–BiLSTM model with SSP achieves excellent performance, achieving accuracies of 0.99 on the training set, 0.97 on the validation set, and 0.92 on the testing set, with low error rates of 0.08 for Top-5 and 0.02 for Top-1. Ablation studies confirm the critical roles of SSP in capturing geological patterns and DRA in balancing computational efficiency by paying more attention to the vertical sedimentary process. These innovations significantly advance automated lithological interpretation, offering a robust framework for geophysical applications.**

***Index Terms*— Dynamic range attention (DRA), lithological interpretation, spatial and stratigraphic information, well logging.**

## I. Introduction

LITHOLOGICAL interpretation is a critical geological task that determines rock types and properties to predict hydrocarbon reservoir location, size, and production potential. Traditional methods rely on expert knowledge and physical/chemical analysis of rock samples (e.g., field observations, core analysis, and thin-section identification) [1], [2], [3]. These methods are valid in simple geological settings but are time-consuming, resource-intensive, and subjective, limiting their performance in complex environments. The logging data offers easily accessible source of rock information. Traditional logging interpretation techniques like cross-plotting and statistical methods are widely used but face limitations in complex settings, including inefficiency, reliance on subjective judgment, and poor accuracy in unconventional reservoirs [4]. To improve the gap, machine learning and deep learning techniques are emerging as promising solutions, leveraging high-dimensional data to improve prediction accuracy and optimize reservoir characterization in complex geological environments [5]. Recently, time-series models are widely used in machine learning applications such as well-logging curve completion [6], [7], [8], [9], [10], [11], underground parameter prediction [12], [13], [14], and lithological prediction [15]. These models excel at handling sequential data. Dominant architectures combine CNNs and bidirectional long short-term memory (BiLSTM) modules [16], [17]. However, the recurrent structure of LSTM limits training parallelism due to temporal dependencies [18], and its matrix operations reduce training efficiency [19]. These factors result in low-LSTM training efficiency, which worsens with CNNs and improved BiLSTM models.

Previous studies focused on logging curve completion [6], [7], [8], [9], [10], [11] but rarely explored lithological interpretation with time-series models [15]. LSTM models, due to their strong temporal dependence, struggle to directly use temporal information for classification. Adding multilayer perceptrons (MLPs) is common for classification [17], [20], [21], [22], but excessive layers can cause overfitting and reduce performance. While attention mechanisms have been integrated into well-logging models [16], [23], [24], [25], adapting Transformer-based mechanisms [26] increases computation time. Thus, improving time-series model performance in log lithology classification requires a lightweight attention module combined with CNN for feature extraction and MLP projections for classification.

To address these gaps, we introduce two key innovations in this study: the spatial and stratigraphic information processing method and the dynamic range attention (DRA) mechanism. The spatial–stratigraphic context (SSP) method processes spatial and stratigraphic information by treating depositional sequences as time series, significantly enhancing the model's ability to capture geological context. The DRA mechanism, a novel attention mechanism, adaptively adjusts the scope of local attention based on the global context of the input, enhancing the model's ability to handle long-range dependencies efficiently.

## II. Methods

### A. Spatial and Stratigraphic Information Processing

Spatial information (shortened as Spatial), represented by well coordinates, was processed as location tokens. In this study, we utilized the cosine function for transformation, defined as follows:

$$X(\alpha_i, \beta_i) = \cos(\alpha_i)\cos(\beta_i) \tag{1}$$

Received 13 February 2025; revised 10 April 2025; accepted 11 April 2025. Date of publication 18 April 2025; date of current version 15 May 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 41872150, Grant U2344209, and Grant U19B6003; and in part by the Sinopec Zhongyuan Oilfield under Grant 31300027. *(Corresponding author: Jinmin Song.)*

The authors are with the State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation, Chengdu University of Technology, Chengdu 610059, China (e-mail: keranli98@outlook.com; songjinmin@sohu.com; lsg@cdut.edu.cn; 9368099@qq.com; 185672735@qq.com; 592790004@qq.com; 269798688@qq.com; 158527608@qq.com; 1198481650@qq.com).

Digital Object Identifier 10.1109/LGRS.2025.3562350

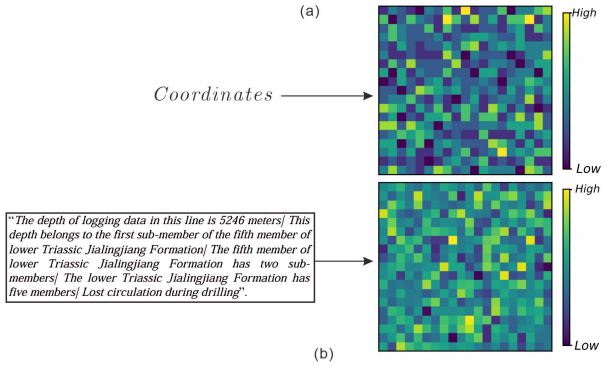


Fig. 1. Illustration of (a) generating spatial information and (b) Strat information.

where $(\alpha_i, \beta_i)$ denotes the geodetic coordinates of well $i$. After projecting by a learnable MLP, the spatial information would expand as the same shape like the well-logging data [Fig. 1(a)].

Stratigraphic information (shortened as Strat) was encoded using word embeddings from natural language processing (NLP). Specifically, we first loaded a pretrained bidirectional encoder representation from transformers (BERTs) model [27]. Subsequently, we appended a sentence to each line of logging data. The template for these sentences is

$$\text{Sentence}_{i,j} = [D, S, M, \hat{S}, \hat{D}] \tag{2}$$

where $j$ represents the depth point and $D$, $S$, $M$, $\hat{S}$, and $\hat{D}$ correspond to linguistic descriptions of depth, stratigraphy, member, submember, and drilling log, respectively. Sentences were input with a fixed format [see Fig. 1(b)].

The complete set of descriptions for each line forms a corpus $C$. Finally, the corpus $C$ is used to fine-tune the BERT model

$$\theta = \text{BERT}(C|\text{fine-tune}). \tag{3}$$

The resulting model $\theta$ generates word embeddings for the corpus $C$ and reduces the word dimensions to $k$-dimensions to keep the same number with the input logging data. In conclusion, after SSP operation, original logging data were expanded to be three channels.

### B. Dynamic Range Attention

Traditional self-attention in Transformers [26], often face challenges in balancing computational efficiency and the ability to model long-range dependencies. To address this, we propose DRA, Fig. 2, a novel attention mechanism that adaptively adjusts the scope of local attention based on the global context of the input. DRA consists of two key components: local attention and global context modeling.

Given an input sequence $X \in \mathbb{R}^{L\times C}$, where $L$ is the sequence length and $C$ is the dimension of the characteristic, local attention computes the attention weights within a sliding window $W$. For each position $i$ in the sequence, the attention weights are computed as follows:

$$\text{Attention}(Q_i, K_{[i-w/2:i+w/2]}, V_{[i-w/2:i+w/2]}) = \text{Softmax}\left(\frac{Q_i K^T_{[i-w/2:i+w/2]}}{\sqrt{C}}\right) V_{[i-w/2:i+w/2]} \tag{4}$$

where $Q_i$, $K_{[i-W/2:i+W/2]}$, and $V_{[i-W/2:i+W/2]}$ are the query, key, and value vectors calculated by [26] for position $i$ and its local neighborhood, respectively.

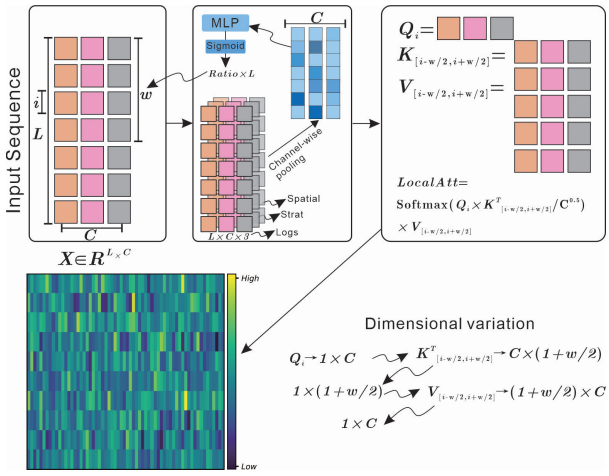


Fig. 2. Illustration of DRA mechanism.

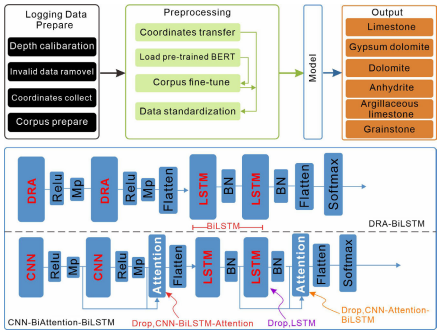


Fig. 3. Workflow of DRA–BiLSTM. The bolded red DRA can be replaced by CNN1D. The bolded white DRA can be replaced by self-attention mechanism. The bolded red BiLSTM can be replaced by LSTM.

As shown in Fig. 2, to dynamically adjust window size $W$, DRA incorporates a global context vector $g \in \mathbb{R}^C$ to fuse the SSP channelwise by

$$g = \text{GlobalPool}(X) \tag{5}$$

where GlobalPool$(\cdot)$ denotes a global pooling operation (e.g., average pooling). The window size for position $i$ is then predicted as follows:

$$W_i = \text{Sigmoid}(f(g)) \tag{6}$$

$$W = W_i \times L \tag{7}$$

where $f(\cdot)$ is a learnable MLP to gain window size probability. Multiplying with the actual length of the logging curve $L$, $W$ was dynamically calculated.

The final output of DRA is computed by combining the local attention outputs with the dynamically adjusted window sizes

$$\text{DRA}(X) = \text{LocalAttention}(X, W) + X \tag{8}$$

where the addition of $X$ represents a residual connection to facilitate gradient flow during training. This design allows DRA to efficiently model both local and global dependencies while maintaining computational efficiency.

### C. BiLSTM Networks

The BiLSTM network [28] comprises two LSTM networks. LSTM [29], excels at modeling long-term dependencies. Its architecture includes a cell state and three gating units (forget, input, and output gates). The BiLSTM uses both the past and the future context to enhance the accuracy of the prediction (see [28]).

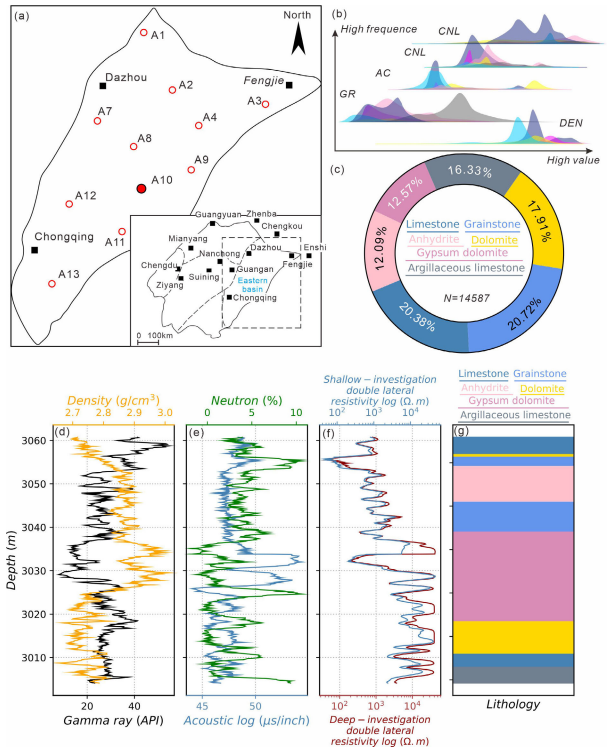


Fig. 4. (a) Well location. (b) Logging data statistic. (c) Logging data distributions. (d)–(g) Logging data visualization of well A10.

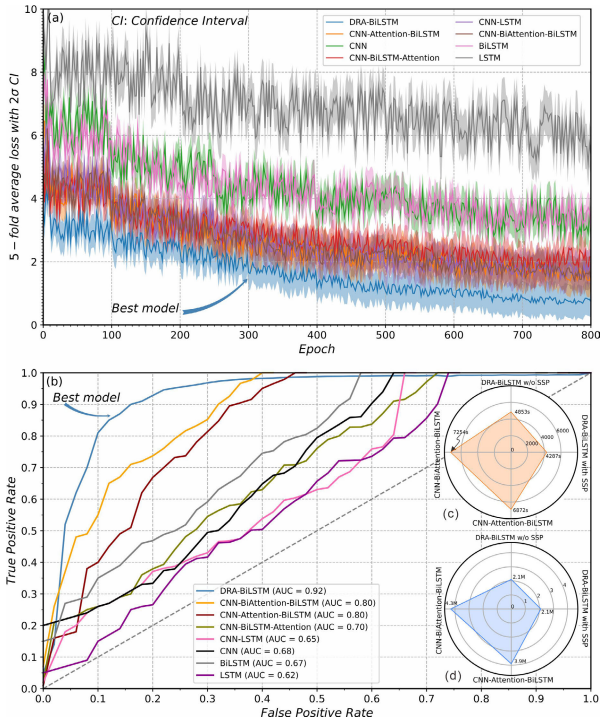


Fig. 5. Model training, ROC, and model size visualizations. (a) Fivefold average loss with $2\sigma$ confidence interval in the test dataset. (b) ROC curves of different models. (c) Radar plot of model training time. (d) Radar plot of model sizes.

### D. Workflow

The study's workflow (Fig. 3) includes data preparation (depth calibration, invalid data removal, well coordinate collection, and corpus construction) and preprocessing [well location transformation using (1), BERT fine-tuning with the corpus [(3), see Fig. 1], and data standardization]. The preprocessed data enters a DRA module (128 filters, ReLU activation, and max pooling) and a BiLSTM module (two layers: forward LSTM with 128 units and batch normalization and backward LSTM with 128 units). The backward BiLSTM outputs serve as $Q$, $K$, and $V$ for the DRA module. The model generates the output after flattening and applying the softmax function. In summary, all BiiLSTM models consist of two LSTM models, including the forward and backward layers. Even in LSTM, four LSTM layers are used to ensure fair comparison. At the same time, comparing the effects of DRA and CNN–BiAttention modules, two CNN–self-attention mechanism modules and two DRA modules were adopted. Details can be seen from Fig. 3.

## III. Experiment

### A. Data

Data from this study was collected from 13 different wells (A1–A13, the well A13 was preserved to evaluate model's generalization performance) from the eastern Sichuan basin [Fig. 4(a)]. A total of 14 587 logging data points and well-observed lithological labels were compiled by petrological experts [Fig. 4(b) and (c)]. Fig. 4(c) shows that the dataset is label-balanced. To avoid data leaching, the dataset was divided into training, validation, and testing datasets in a ratio of 7:2:1 in every single well. All data was unshuffled. The types of logging data include gamma ray (GR), density (DEN), acoustic (AC), CNL (neutron), deep-investigation double lateral resistivity log (RLLD), and shallow-investigation double lateral resistivity log (RLLS) [Fig. 4(d)–(g)].

### B. Evaluation Metrics Setting

In this study, we employed a comprehensive set of evaluation metrics to assess the performance of the models, including accuracy, precision, recall, $F1$-score, AUC, Top-1 error, and Top-5 error. The definitions and corresponding formulas followed [30], [31].

### C. Experimental Parameters

All experiments in this study were conducted on a Python-based online platform (AutoDL, https://www.autodl.com/) using an NVIDIA A100 GPU. The initial learning rate was $5 \times 10^{-5}$. The batch size was set to 128, and the number of iterations was 800 under the fivefold strategy.

## IV. Results

As shown in Fig. 5(a), all models stabilize within 800 epochs. The single LSTM model performs worst, while CNN, BiLSTM, and CNN–LSTM models show slightly better results. Combining CNN and BiLSTM improves performance by 25%. Adding an attention module between CNN and BiLSTM significantly reduces cross-entropy loss. Finally, incorporating the DRA module achieves optimal performance. In ROC curves, the DRA–BiLSTM model is the closest vertex to the top left corner [Fig. 5(b)]. Meanwhile, the DRA–BiLSTM model achieves the highest AUC of 0.92, while the LSTM model has the lowest AUC of 0.64. The curves show that models incorporating attention mechanisms and hybrid architectures generally outperform simpler models, highlighting their effectiveness in improving classification accuracy. Except for performance, Fig. 5(c) and (d) shows the comparison of models' size and running efficiency. DRA–BiLSTM with SSP and without SSP both have 2.1 M parameters but differ in runtime (4287 versus 4853 s), with the SSP version faster. Larger models, CNN–BiAttention–BiLSTM (4.3 M) and CNN–Attention–BiLSTM (3.9 M), have longer runtimes

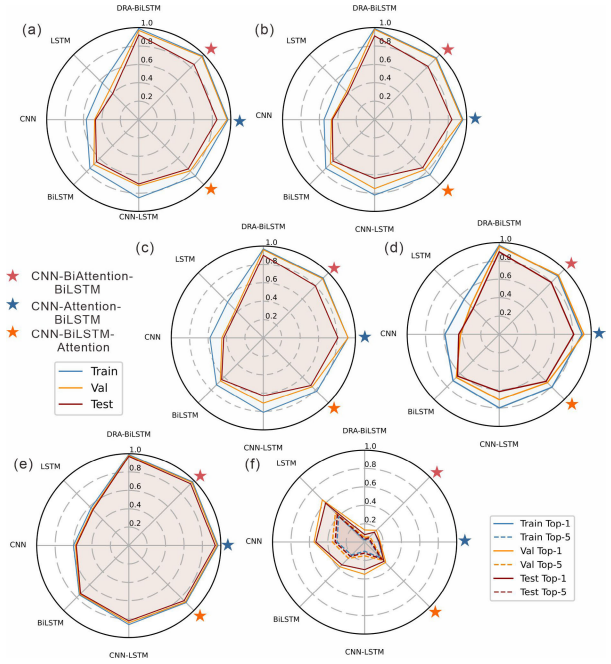


Fig. 6. Comparisons of model performance. (a) Accuracy. (b) Precision. (c) Recall. (d) $F1$-score. (e) AUC. (f) Top-1 and Top-5 error.

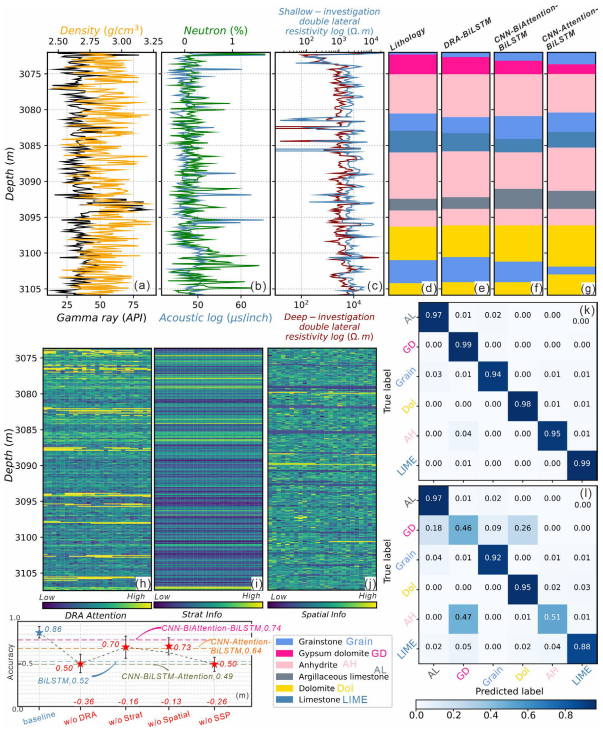


Fig. 7. Comparisons and ablation experiments results. (a)–(c) Logs of preserved well A13. (d) Actual lithological labels. (e) Lithological labels predicted by DRA–BiLSTM. (f) Lithological labels predicted by CNN–BiAttention–BiLSTM. (g) Lithological labels predicted by CNN–Attention–BiLSTM. (h) Feature maps of DRA attention in DRA–BiLSTM. (i) Feature maps of Strat information in DRA–BiLSTM. (j) Feature maps of spatial information in DRA–BiLSTM. (k) Confusion matrix of DRA–BiLSTM in test dataset. (l) Confusion matrix of DRA–BiLSTM without SSP in the testing dataset. (m) Ablation results of DRA–BiLSTM in the testing dataset.

(7254 and 6872 s), showing a tradeoff between size and efficiency. The DRA–BiLSTM with SSP outperforms traditional CNN–BiLSTM with self-attention.

Fig. 6 compares the performance metrics of various deep learning models across training, validation, and test datasets. The DRA–BiLSTM model achieves the highest accuracy (0.99, 0.97, and 0.92) and lowest error rates, outperforming other models. The CNN–BiAttention–BiLSTM model follows with accuracies of 0.98, 0.97, and 0.85, and a Top-1 error of 0.15 and Top-1 error of 0.03 on the testing set. CNN–Attention–BiLSTM and CNN–BiLSTM models show strong results (accuracies: 0.97–0.87 and Top-5/1 errors: 0.15/0.24). CNN–LSTM and BiLSTM models perform moderately (accuracies: 0.85 and 0.75 and Top-5/1 errors: 0.30/0.35). The simplest architectures, CNN and LSTM, have the lowest performance (accuracies: 0.57–0.40 and Top-5/1 errors: 0.53/0.60). The results highlight the superiority of complex models with attention mechanisms (e.g., DRA–BiLSTM and CNN–BiAttention–BiLSTM) over simpler models like CNN and LSTM, demonstrating improved accuracy and reduced errors.

## V. Discussion

Except for comparing models' performance and selecting the best closed box model, this study first utilized the preserved A13 data to test the generalization performance. Fig. 7(a)–(g) displays the original logging data and different models' predictions. Generally, no obvious lithological mis-predictions occurred. DRA–BiLSTM indicates more accurate forecasting in thin beds [the uppermost grainstone and argillaceous limestone, Fig. 7(e) and (f)]. The performance of CNN-Attention-BiLSTM model further deteriorates after dropping one self-attention module [Fig. 7(g)]. DRA module and SSP guided model are more sensitive to interlayer information perception [Fig. 7(h)–(j)]. Especially the DRA module captures fine-grained logging information. SSP further maintains the original high information entropy of logging data under the guidance of learnable parameters. To test the role of SSP, another DRA–BiLSTM model was trained on the same dataset without the SSP. Confusion matrices suggest that without SSP, the DRA–BiLSTM model could not identify the gypsum dolomite and dolomite, anhydrite and gypsum dolomite [Fig. 7(k) and (l)]. The SSP improves the lithological predicting by introducing extra data, as demonstrated by the confusion matrices. This improvement is crucial for distinguishing complex lithologies, such as gypsum dolomite versus anhydrite, and highlights the importance of SSP in enhancing model performance. However, the detailed boosts of Strat and Spatial information were still mysterious. By dropping the whole SSP operation, the accuracy of DRA–BiLSTM in test dataset declined 0.26 [Fig. 7(m)]. When only dropping the Strat and Spatial information, the accuracy only declined by 0.16 and 0.13 [Fig. 7(m)]. Both Strat and Spatial information have made almost equal contributions to SSP operation. The overall performance degradation of SSP compared to individual components can be explained by the GlobalPool of channelwise. The information fusion between the three channels will preserve more modeling information for subsequent time-series models.

The estimated 10% performance difference between CNN–BiAttention–BiLSTM and DRA–BiLSTM models indicates that the performance of CNN with multiple self-attention mechanisms is similar to that of DRA modules, but there are differences at the same time. In order to compare the differences between the two mechanisms, the feature maps of the CNN–BiAttention–BiLSTM model are shown in Fig. 8. Fig. 8(a) and (b) shows sparse attention of the two-layer 1-D convolution model on well-logging curves, and the model attention is independent of the depth direction. The receptive field of the model becomes sparser as the number of

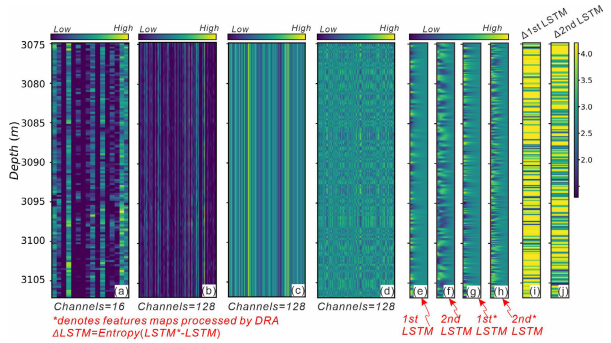


Fig. 8. Feature visualizations. (a) First layer of CNN1D module. (b) Second layer of CNN1D module. (c) First self-attention module. (d) Second self-attention module. (e) First LSTM module in BiLSTM following two CNN1D layers. (f) Second LSTM module in BiLSTM following two CNN1D layers. (g) First LSTM module in BiLSTM following DRA modules. (h) Second LSTM module in BiLSTM following DRA modules. (i) Entropy difference of the first LSTM module in BiLSTM with DRA and CNN1D. (j) Entropy difference of the second LSTM module in BiLSTM with DRA and CNN1D.

convolutional layers increases [Fig. 8(a) and (b)]. The sparse attention of convolutional layers in depth is also transmitted to subsequent self-attention mechanisms, causing the model to lose longitudinal attention between lithological formations [Fig. 8(c) and (d)]. In contrast, the significant vertical attention of DRA modules is consistent with the superposition of rocks in the vertical direction during the sedimentary process of the strata [Fig. 7(h)–(j)]. The shortcomings of vertical attention also have corresponding implications in subsequent BiLSTM modeling. The BiLSTM model processed by DRA has more complex temporal features [Fig. 8(e)–(h)] and higher information entropy than the BiLSTM model with convolution and self-attention mechanism [Fig. 8(i) and (j)].

## VI. Conclusion

This study enhances lithological interpretation by integrating SSP and DRA into a DRA–BiLSTM model. SSP encodes geological sequences as time series, while DRA improves accuracy and efficiency. Experiments show the model outperforms traditional approaches, achieving high training accuracy and strong generalization. Ablation tests confirm SSP's role in distinguishing complex lithologies and DRA's enhancement of dependency modeling. The study highlights the potential of combining domain-specific data processing with lightweight attention mechanisms for geophysical time-series models, paving the way for efficient, interpretable automated logging analysis. Meanwhile, the test results showed that DRA is insufficient in capturing thin-layer interfaces. The DRA receptive field focuses too much on the vertical subtle changes in the logging information. Balancing local and detailed information is the key to future research on intelligent interpretation of logging lithology.

## References

[1] T. Aïfa, "Neural network applications to reservoirs: Physics-based models and data models," *J. Petroleum Sci. Eng.*, vol. 123, pp. 1–6, Nov. 2014.

[2] F. Anifowose, J. Labadin, and A. Abdulraheem, "Improving the prediction of petroleum reservoir characterization with a stacked generalization ensemble model of support vector machines," *Appl. Soft Comput.*, vol. 26, pp. 483–496, Jan. 2015.

[3] J. Chang, Y. Kang, Z. Li, W. X. Zheng, W. Lv, and D.-Y. Feng, "Cross-domain lithology identification using active learning and source reweighting," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[4] J. Jeong, E. Park, I. Emelyanova, M. Pervukhina, L. Esteban, and S. Yun, "Interpreting the subsurface lithofacies at high lithological resolution by integrating information from well-log data and rock-core digital images," *J. Geophys. Res., Solid Earth*, vol. 125, no. 2, Feb. 2020, Art. no. e2019JB018284.

[5] Q. Pang, C. Chen, Y. Sun, and S. Pang, "STNet: Advancing lithology identification with a spatiotemporal deep learning framework for well logging data," *Natural Resour. Res.*, vol. 34, no. 1, pp. 327–350, Feb. 2025.

[6] D. Zhang, Y. Chen, and J. Meng, "Synthetic well logs generation via recurrent neural networks," *Petroleum Explor. Develop.*, vol. 45, no. 4, pp. 629–639, Aug. 2018.

[7] Y. Chen and D. Zhang, "Well log generation via ensemble long short-term memory (EnLSTM) network," *Geophys. Res. Lett.*, vol. 47, no. 23, Dec. 2020, Art. no. e2020GL087685.

[8] N. Pham, X. Wu, and E. Z. Naeini, "Missing well log prediction using convolutional long short-term memory network," *Geophysics*, vol. 85, no. 4, pp. WA159–WA171, Jul. 2020.

[9] J. Li and G. Gao, "Digital construction of geophysical well logging curves using the LSTM deep-learning network," *Frontiers Earth Sci.*, vol. 10, Jan. 2023, Art. no. 1041807.

[10] H. Zhang, W. Wu, and X. Song, "Well logs reconstruction based on deep learning technology," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[11] D. Haritha, N. Satyavani, and A. Ramesh, "Generation of missing well log data with deep learning: CNN-Bi-LSTM approach," *J. Appl. Geophys.*, vol. 233, Feb. 2025, Art. no. 105628.

[12] R. Huang et al., "Well performance prediction based on long short-term memory (LSTM) neural network," *J. Petroleum Sci. Eng.*, vol. 208, Jan. 2022, Art. no. 109686.

[13] W. Sun et al., "Lost circulation monitoring using bi-directional LSTM and data augmentation," *Geoenergy Sci. Eng.*, vol. 225, Jun. 2023, Art. no. 211660.

[14] M. Zhang, A. Jia, and Z. Lei, "Inter-well reservoir parameter prediction based on LSTM-attention network and sedimentary microfacies," *Geoenergy Sci. Eng.*, vol. 235, Apr. 2024, Art. no. 212723.

[15] J. Lin, H. Li, N. Liu, J. Gao, and Z. Li, "Automatic lithology identification by applying LSTM to logging data: A case study in X tight rock reservoirs," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 8, pp. 1361–1365, Aug. 2021.

[16] L. Shan, Y. Liu, M. Tang, M. Yang, and X. Bai, "CNN-BiLSTM hybrid neural networks with attention mechanism for well log prediction," *J. Petroleum Sci. Eng.*, vol. 205, Oct. 2021, Art. no. 108838.

[17] D. T. dos Santos, M. Roisenberg, and M. dos Santos Nascimento, "Deep recurrent neural networks approach to sedimentary facies classification using well logs," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[18] K. Hwang and W. Sung, "Single stream parallelization of generalized LSTM-like RNNs on a GPU," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 1047–1051.

[19] B. Li et al., "Large scale recurrent neural network on GPU," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2014, pp. 4062–4069.

[20] C. Jiang, D. Zhang, and S. Chen, "Lithology identification from well-log curves via neural networks with additional geologic constraint," *Geophysics*, vol. 86, no. 5, pp. IM85–IM100, Sep. 2021.

[21] J. Wang and J. Cao, "A lithology identification approach using well logs data and convolutional long short-term memory networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[22] J.-J. Liu and J.-C. Liu, "Integrating deep learning and logging data analytics for lithofacies classification and 3D modeling of tight sandstone reservoirs," *Geosci. Frontiers*, vol. 13, no. 1, Jan. 2022, Art. no. 101311.

[23] L. Zeng, W. Ren, and L. Shan, "Attention-based bidirectional gated recurrent unit neural networks for well logs prediction and lithology identification," *Neurocomputing*, vol. 414, pp. 153–171, Nov. 2020.

[24] M. Shi, B. Yang, R. Chen, and D. Ye, "Logging curve prediction method based on CNN-LSTM-attention," *Earth Sci. Informat.*, vol. 15, no. 4, pp. 2119–2131, Dec. 2022.

[25] I. Kumar, B. K. Tripathi, and A. Singh, "Attention-based LSTM network-assisted time series forecasting models for petroleum production," *Eng. Appl. Artif. Intell.*, vol. 123, Aug. 2023, Art. no. 106440.

[26] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[28] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[30] M. Lapin, M. Hein, and B. Schiele, "Loss functions for top-k error: Analysis and insights," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1468–1477.

[31] G. Santafe, I. Inza, and J. A. Lozano, "Dealing with the evaluation of supervised classification algorithms," *Artif. Intell. Rev.*, vol. 44, no. 4, pp. 467–508, Dec. 2015.