Intelligent identification method of sedimentary microfacies based on DMC-BiLSTM

Luo Renze [1*] Zhou Yang [1] Li Xingyu [1] Guo Liang [1] TuoJuanjuan [1] Kang Lixia [2]

(1. State Key Laboratory of Oil and Gas Reservoir Geology and Development Engineering, Southwest Petroleum University, School of Earth Sciences and Technology, Southwest Petroleum University, Chengdu 610500; 2. PetroChina Changqing Oilfield Company, Yinchuan, Ningxia, 750005)

Abstract: Sedimentary microfacies division is the basis of oil and gas exploration research. The traditional sedimentary microfacies division mainly depends on human experience, which is greatly influenced by human factor and is low in efficiency. Although deep learning has its advantage in solving complex nonlinear problems, there is no effective deep learning method to solve sedimentary microfacies division so far. Therefore, this paper proposes a deep learning method based on DMC-BiLSTM for intelligent division of well-logging—sedimentary microfacies. First, the original curve is reconstructed multi-dimensionally by trend decomposition and median filtering, and spatio-temporal correlation clustering features are extracted from the reconstructed matrix by Kmeans. Then, taking reconstructed features, original curve features and clustering features as input, the prediction types of sedimentary microfacies at current depth are obtained based on BiLSTM. Experimental results show that this method can effectively classify sedimentary microfacies with its recognition efficiency reaching 96.84%.

Key words: Trend decomposition; Median filtering; kmeans; BiLSTM

*Corresponding author. Email:lrzsmith@126.com.

1 Introduction

Sedimentary microfacies is one of the most important research contents in oil and gas exploration and development, and plays an especially important role in predicting the production of remaining oil enrichment area. In the traditional sedimentary microfacies division, based on previous research results and regional sedimentary background, sedimentary types in the study area are usually determined by combining paleontology, sedimentology and other theories with the analysis of key well core facies markers. The sedimentary microfacies in the study area are usually divided by comprehensively analyzing the characteristics of rock thickness and grain size and analyzing the morphological characteristics of various logging curves [1-2]. Therefore, it is difficult to find out the qualitative mapping relationship between sedimentary facies and various logging data due to the complicated and tedious cross-analysis of various data.

With the rapid development of computer technology, the main applications of machine learning in sedimentary microfacies include KNN [3], Bayesian network [4], support vector machine [5] and artificial neural network [6]. It depends on specific geological environment and sedimentary background when these methods mainly construct morphological parameters and physical parameters of GR curves. Therefore, machine learning method has limitations in constructing the identification model of sedimentary microfacies.

Although deep learning has its advantage in mining nonlinear relationship of data, there is no effective identification model of sedimentary microfacies so far. recurrent neural network can conduct sequential processing over time. In 2020, Song Hehou, et al applied LSTM network to identifying well logging curve shape [7], but it could only identify 4 kinds of logging facies of Bell shape, Funnel shape, Egg shape, Cylinder shape with only one-way sequence, and could not satisfy the demand for effectively identifying more complex log facies types in the actual sedimentary environment. Therefore, it is urgent to develop a more accurate deep learning network model with stronger generalization ability to effectively identify sedimentary microfacies.

This paper proposes a sedimentary microfacies identification model based on DMC-BiLSTM deep learning, and establishes a complete set of sedimentary microfacies recognition method of curve processing - structure characteristics - model training – model evaluation. The validity and advantage of this proposed method is proved by comparing this method with geological experts and the other two deep learning models in precision and accuracy.
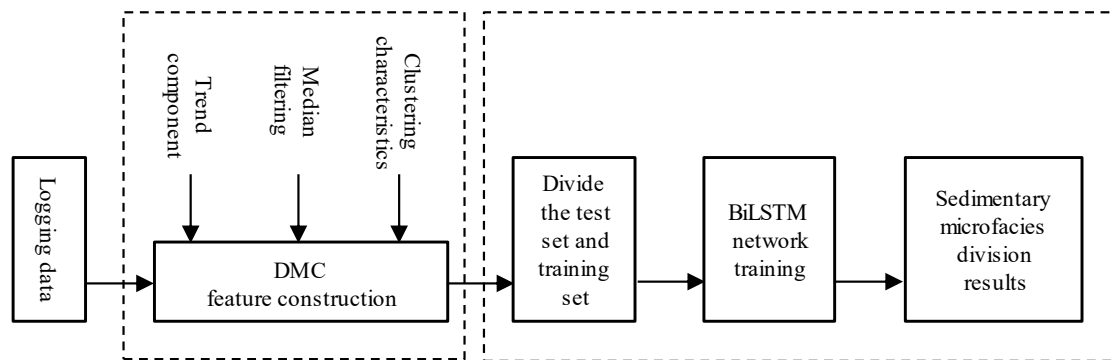
2   DMC-BiLSTM Model

Fig. 1   Overall model framework

Fig 1 is the overall flow chart of intelligent identification of sedimentary microfacies based on DMC-BiLSTM deep learning. The specific implementation process is as follows:

1)   The original GR logging curve is decomposed by STL algorithm, and the periodic component and residual term are removed from the decomposed results to obtain the trend component. In this paper GR is decomposed by different frequency windows to obtain multiple trend component features;

2)   Use the same frequency window as the trend decomposition to conduct median filtering to obtain multiple median filtering features;

3)   The obtained features and the original GR data features are used as the input features of Kmeans clustering to conduct unsupervised learning and the clustering features with spatial and temporal correlation are obtained;

4)   Establish the correspondence between logging facies and sedimentary microfacies of delta front subfacies;

5)   All feature sets generate characteristics matrix, which are input into Bidirectional Long Short-Term Memory Network (BiLSTM) for sedimentary microfacies division.

2.1 DMC Characteristics Structure

DMC feature structure: D represents trend feature, M represents median filter feature, and C represents clustering feature.In the process of logging data collection, the effective signals are mixed with various noises due to external interference or the instrument itself. The key logging facies partitioning parameters such as the shape and width of logging signal need to be kept to the maximum extent while removing noises. First, the STL method is used to decompose the curve to get its trend feature, and then the median filter is used to get the median filter feature. Then the Kmeans clustering feature is constructed. Finally, the feature set of the GR curve is obtained after normalization of all the features.

1. The trend Component

STL algorithm is a common algorithm in time series decomposition. Based on LOESS, the data at a certain moment is decomposed into trend component, periodic component and residual term [8]. According to the preliminary analysis of GR curve, the addition model can meet the decomposition requirements, namely:

$$Y_v = T_v + S_v + R_v \qquad v = 1,...,N \tag{1}$$

Where, Y_v is the data at a certain moment, T_v is the trend component, S_v is the periodic component, R_v is the residual term; v is some time.

2. Median Filtering

Basic principle of median filtering [9-10] is: by setting the window value, the output signal of a certain point can be replaced by the statistical median value of all signals within the window range of this point. One-dimensional median filtering is defined as:

$$g(x) = med\left\{ f(x-k),(k,\in W) \right\} \tag{2}$$

Where: $f(x)$ is the original signal sequence, $g(x)$ is the filtered data.

3. Clustering Characteristics

Kmeans [11] is an unsupervised learning clustering algorithm, and its steps are as follows:

1) Characteristic data of GR curve is clustered into K class (for better classification effect of the final model, K=6). First, 6 GR data points are selected as the initial center points;

2) According to the principle of minimum distance from the initial center point, all curve data is divided into the classes where each center point is located;

3) There is a large amount of data in each class, and the means of all curve sample data of 6 classes are calculated as the 6 center points of the second iteration;

4) Step 2 and 3 are repeated by using the center until convergence (the center point will not alter the table or reach the specified number of iterations), and then the clustering process is ended.

Euclidean distance is easy to understand and the most commonly used distance formula. The Euclidean distance can better reflect the similarity between GR curve sequences in the small depth section, and fully reflect the spatial correlation of GR curve. It is defined as:

$$d = \sqrt{\sum_{k=1}^{n}(x_{1k} - x_{2k})^2} \qquad (3)$$

Where: x_{1k}, x_{2k} are two n-dimensional vectors.

2.2 BiLSTM Principle

1 BiLSTM Principle

GR data can be regarded as the time series changing with the depth, and its shape classification depends on the previous output and current input. Assuming that a given GR sequence $x = \{x_1, x_2, ..., x_t, ..., x_N\}$, BiLSTM [12-14] can model each sequence forward and backward at the same time because each tag encoding contains contextual information from the past and the future, and it can better represent the long-term dependence on time series data in a richer way. Each layer of the BiLSTM network is composed of a single LSTM unit propagating forward and backward (Fig 2).
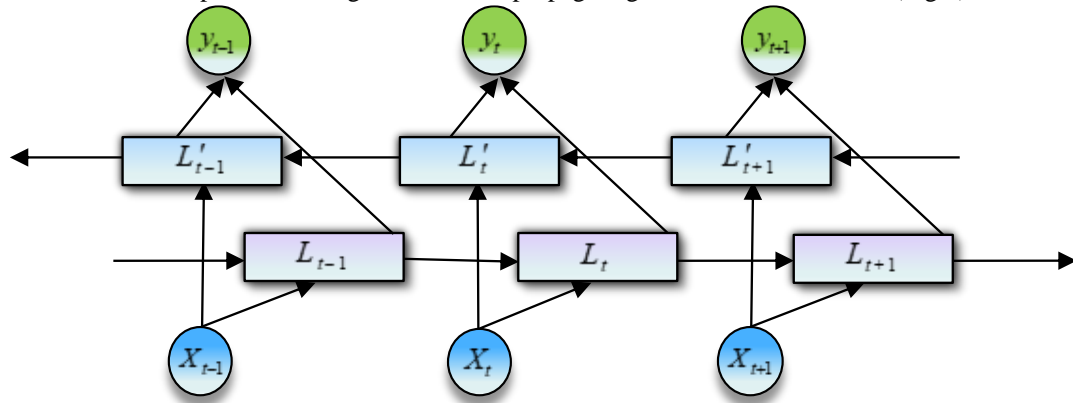

Fig. 2   BiLSTM Structure

Long Short-Term Memory Network (LSTM) can solve the problem that RNN cannot cope with its dependence on long distance [15]. The hidden layer of the original RNN has only one state, h, which is very sensitive to short-term input. Add a state c to hold long-term states, and it is called cell states. Three more gates, the forgetting gate(f_t), the input gate(i_t) and the output gate (o_t), are used to control the long-term state c. Together, the core part of the LSTM model "memory block" (Fig 3) is constituted. General formula is as follows:

$$\begin{cases} f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ c_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \\ c_t = f_t \otimes c_{t-1} + i_t \otimes \hat{c}_t \\ h_t = o_t \otimes \tanh(c_t) \end{cases} \qquad (4)$$

3

Where: σ , tanh as two activation function, c_t represents the storage unit at time t, c_t represents the candidate storage unit at time t, $W = \left\{ W_f ; W_i ; W_o ; W_c ; U_f ; U_i ; U_o ; U_C \right\}$ is the weight coefficient, updated by the back propagation, $b = \left\{ b_f ; b_i ; b_o ; b_c \right\}$ represents the bias coefficient, x_t represents the sequence value at time t, h_t represents the output of the hidden layer.
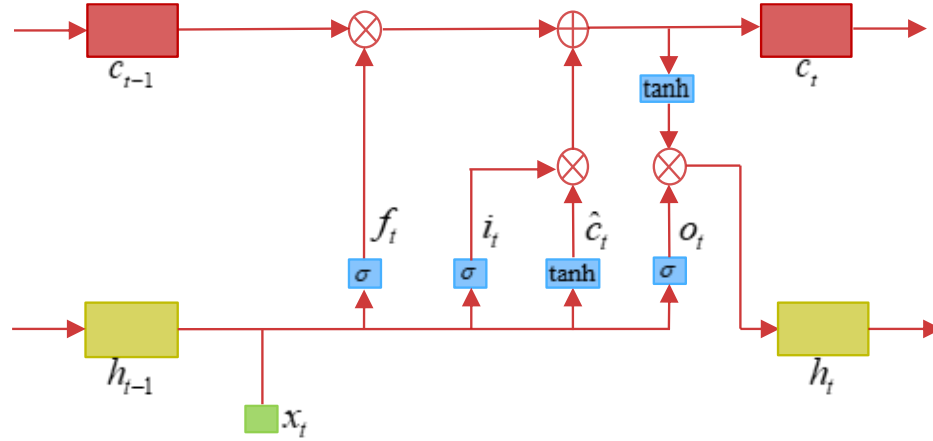

Fig. 3   LSTM Operation Flow Fig

2 Model Configuration and Training

In this paper, BiLSTM network is used to identify sedimentary microfacies, and a four-layer BiLSTM unit is built to learn the GR curve sequence. The "distributed characteristic representation" learned is mapped to the sample tag space through the fully connected layer; finally, the output vector is input to the Softmax layer, which contains 5 neural units, to perform the logging phase classification task.

In order to avoid model overfitting, the data is divided into training set and validation set. The cross-validation method is adopted. 75% of the wells are used for training and 25% for validation. A Dropout unit with a coefficient of 0.2 is added to each layer of the network, which increases the generalization ability of the model by abandoning some connections. Adam [16] optimization algorithm is used to update the gradient, and learning rate attenuation algorithm is defined to accelerate the convergence of the model.

The cross entropy loss function is used as the training loss, as follows:

$$\begin{cases} y = \text{softmax}(W_c x + b_c) \\ L = -\dfrac{1}{n} \sum_x [y \, Ina + (1-y) In(1-a)] \end{cases} \qquad (5)$$

Where, y is the expected output, a is the actual output of the neuron.

Table. 1   Parameters of BILSTM structure

Layer(type)	Output Shape	Param
L_0 (input layer)	(None, 1100, 20)	0
L_1 (Bidirection	(None, 1100, 256)	152576
L_2 (Bidirection	(None, 1100, 128)	164352
L_3 (Bidirection	(None, 1100, 128)	98816
L_4 (Bidirection	(None, 1100, 128)	98816
L_5 (Dense)	(None, 1100, 128)	16512
L_6 (Dropout)	(None, 1100, 128)	0
L_7 (Dense)	(None, 1100, 5)	645

3   Experimental Results and Analysis

3.1 Source of Log Data

The logging data in this experiment is Gamma Log Facies Type Prediction data from the artificial intelligence companies CrowdAnalytix in California's Silicon Valley. The variation of GR logging characteristics can be used as the characterization of grain size, making GR play an important role in sequence stratigraphic analysis [17].The abrupt changes in the GR log response are interpreted as sharp lithologic fractures associated with unconformity and sequence boundaries, so the basic shape of the logging curve is often used to explain the sedimentary background of the sedimentary cycle [18]. Table 2 shows the logging curve types of delta front subfacies.

Table. 2   The Logging Curve Types of Delta Front Subfacies

Curve shape	Cylindrical shape	Funnel shape	Bell shape	Symmetrical shape	Lowly Microdentate
Microfacies types	Distributary channel	Mouth bar	Channel edge	Front sheet sand	Distributary bay

3.2 Divide the Training Set and the Validation Set

The data was collected from 4,000 wells, and each well contained 1,100 data entries. The data of 3000 wells were selected as the training set, and the remaining 1000 wells as the verification set to verify the prediction ability of the classification model based on DMC-BiLSTM proposed in this paper on sedimentary microfacies. Table 3 is the specific divided training set and verification set.

Table. 3   Training Set and Verification Set

Microfacies types	Training set (number)/piece	Test set (number)/ piece
Distributary bay	1734194	579087
Front sand sheet	389971	130773
Distributary channel	395415	129961
Mouth bar	389359	132781
Channel edge	391061	127398

3.3 Data Preprocessing

All experiments in this paper were carried out on Intel (R) Core (TM) i5-8300CPU@2.3GHz and8GBRAMD devices. 4 Taking No. 9 Well as an example, the frequency window values of decomposition and filtering are both 19, and the trend component feature, median filter feature and Kmeans clustering feature are constructed for the GR curve (Fig 4). -GR represents the original GR curve data, -decompose represents the trend decomposition curve, -medfilt represents the median filtering curve, -cluster represents the clustering characteristic curve, and -label represents the true label distribution. It can be seen that such a feature set can not only remove the high-frequency noise in the curve, but also retain the geological trend characteristics and effective edge morphological characteristics, and fully reflect the spatial and temporal correlation of the curve data. In the original data, in order to make the fitting ability of the model stronger, the GR curve feature construction was carried out in all the odd filtering windows in 3-20. Finally, the 20-dimensional feature set was generated together with the original GR curve data.
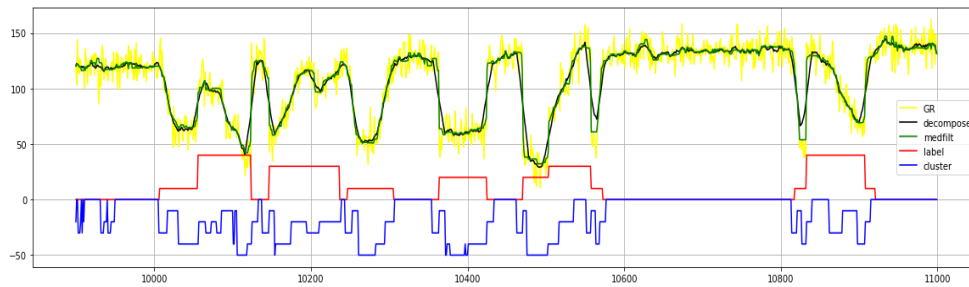
Fig. 4   GR curve feature construction of No. 9 well

3.4 Experimental Results

In the data, the types of Distributary bays are significantly more than the other four microfacies types, and there is a phenomenon of data imbalance. Therefore, it is not scientific to adopt a single evaluation index. In this paper, multiple evaluation indexes were introduced, including confusion matrix, precision, recall, F1-score, and overall accuracy [19] to comprehensively evaluate the prediction results and to comprehensively reflect the classification effect of the model.

1)   Confusion matrix

Real / Predict	Positive cases	Negative cases
Positive cases	True Positive	False Negative
Negative cases	False Positive	True Negative

2)   Accuracy: the proportion of correct results predicted by the model in the total results:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

3)   Precison: proportion of true instances in results predicted by the model as positive examples:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

4)   Recall: How many positive examples of the sample in the model prediction are correctly predicted:

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

5)   F1 value: It is the harmonic average of accuracy and recall:

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{9}$$

Where, TP is the true example, FP is the false positive example, TN is the true negative example, and FN is the false negative example.

From the confusion matrix and various evaluation indexes, it can be seen that the classification accuracy of the model in this paper reached 0.98, 0.94, 0.96, 0.95 and 0.95, respectively for Distributary bay, Front sand sheet, Distributary channel, Mouth bar and Channel edge. Among them, the classification effect for the low-amplitude microdentate Distributary bay is the best, which is due to the fact that Distributary bay is mainly composed of mudstone deposits, shows high GR value and is easy to identify. Secondly, the accuracy, recall rate and F1 value reached more than 96% in the identification of Cylinder shape Distributary channel. For the identification of other curve shapes, the values of all indexes were slightly lower, probably because the curve shapes of the three are not very different, and the only difference is between water progradation and retrogradation during sedimentation, but the accuracy of prediction of the three curve shapes were also over 94%. In conclusion, the overall prediction ability of the model on the validation set was great, with average precision, average recall rate and average F1 value all above 95%, and the accuracy rate also reached 96.84%.

Table. 4   Confusion Matrix and Evaluation Indexes

Predictive value	Real value					Total
	Distributary bay	Front sand sheet	Distributary channel	Mouth bar	Channel edge	

6

Distributary bay	572585					579087
Front sand sheet		122996				130773
Distributary channel			125722			129961
Mouth bar				123621		132781
Channel edge					120368	127398
Precision	0.98	0.94	0.96	0.95	0.95	0.956
Recall	0.99	0.94	0.97	0.93	0.94	0.954
F1-score	0.99	0.94	0.97	0.94	0.95	0.958
Accuracy: 96.84%						

The DMC-BiLSTM model proposed in this paper is compared with LSTM and (Time Convolutional Network) TCN [20]. Batch size of 100 was adopted for all the three models, 1000 rounds were trained, an early stop strategy was set (the model was within 20 rounds, if loss did not decrease, it was the final model), and model parameters such as Adam optimization algorithm were used. The difference is that the input to the DMC-BiLSTM model is the new 20-dimensional feature set. The final results are shown in Table 5. It can be seen that the DMC-BiLSTM model proposed in this paper has the best effect, with the final accuracy reaching 96.84%, while that of LSTM model is 69.76% and that of TCN model is 88.68%. The final accuracy is about 15% higher than that of LSTM model and about 8% higher than that of TCN model. At the same time, compared with the traditional LSTM model and TCN model, the prediction precision of DMC-BiLSTM model is also improved in each category. It is proved that the structural features play an important role in the identification of sedimentary microfacies, and the importance of before and after remote sequence information in the morphological analysis of GR curves is reflected.

Table. 5   The Results of Different Microfacies Identification Models

model	evaluation index	Distributary bay	Front sand sheet	Distributary channel	Mouth bar	Channel edge	average
LSTM	Precision	0.83	0.32	0.44	0.57	0.61	0.554
	Accuracy: 69.76%						
TCN	Precision	0.94	0.80	0.84	0.81	0.85	0.848
	Accuracy: 88.68%						
DMC-BiLSTM	Precision	0.98	0.94	0.96	0.95	0.95	0.956
	Accuracy: 96.84%						

Fig 5 shows the visualization results of the three models on the sedimentary microfacies division on Well 3998, where 0, 1, 2, 3 and 4 respectively corresponded to five microfacies types: Distributary bay, Front sand sheet, Distributary channel, Mouth bar and Channel edge. -dcp curve was the trend component curve of GR curve with the window value of 19, -med was the median filtering curve of GR curve with the frequency window value of 19, -cluster was the characteristic curve after clustering, and the label column was the true label value. It can be seen that DMC-BiLSTM model not only has obvious advantage in the recognition accuracy of sedimentary microfacies, consistented with the artificial partition, but also can well distinguish the boundaries between different categories.
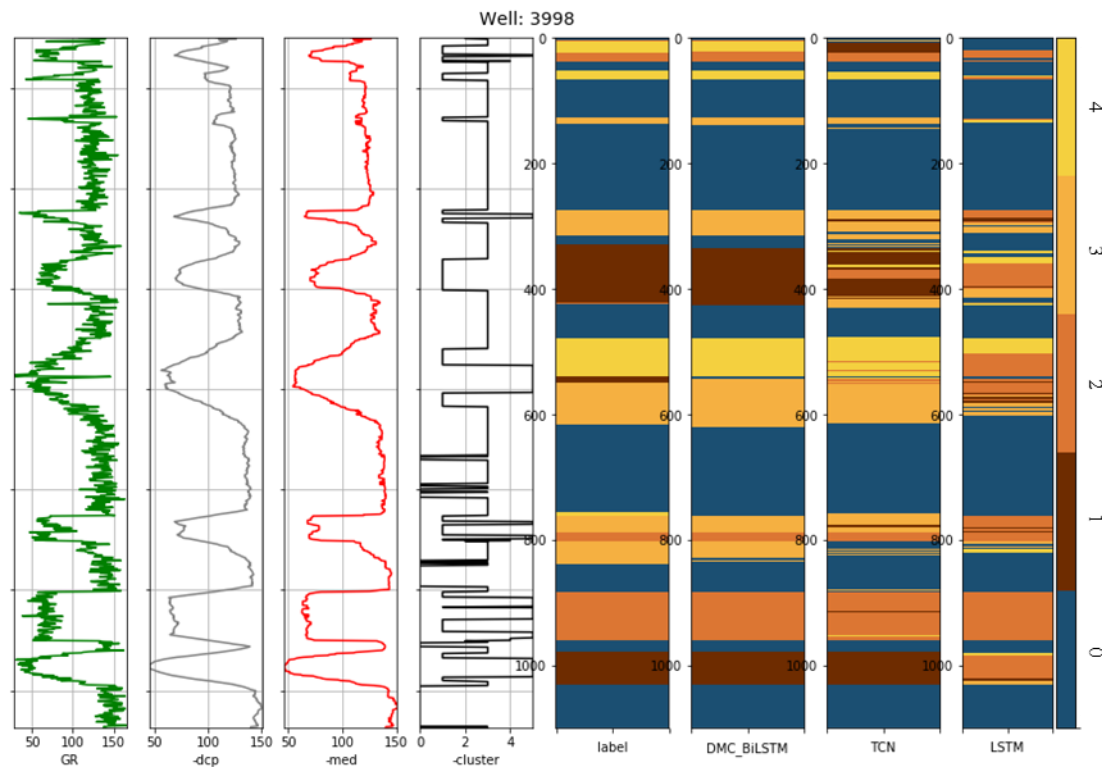
Fig. 5   The Comparison of Different Models in Predicting Sedimentary Microfacies in Well 3998

4   Conclusion

In this chapter, an intelligent identification method of sedimentary microfacies based on DMC-BILSTM deep learning is proposed.In this method, three kinds of features are constructed: geological trend feature, median filter feature and clustering feature.The median filtering feature can not only remove the high-frequency noise in the logging curve, but also retain the effective edge information. The clustering feature can better reflect the spatio-temporal correlation of the logging curve and distinguish the boundaries between different curve shapes.Comparing the DMC-BILSTM method with the TCN and LSTM models obtained from the original curve input simulation, the classification accuracy of the proposed method is 96.84%, which is higher than the 88.68% of the TCN model and 69.76% of the LSTM model.DMC-BILSTM method can be used to realize end-to-end learning of logging data, and the recognition accuracy of each kind of sedimentary microfacies is 98%, 94%, 96%, 95% and 95% for Distributary bay, Front sand sheet, Distributary channel, Mouth bar, Channel edge.Experimental results show that the DMC-BILSTM model is helpful to extract the hidden features of logging curve sequences and fully learn the differences between different shapes.It can effectively identify sedimentary microfacies, and the model has strong generalization and robustness.It provides a new method for the identification of sedimentary microfacies.

5   Acknowledgement

References

[1] Li, L., Wei, J., Yin, H., Xie, J., Ding, Y., & Shi, S. Influence of sedimentary facies on reservoir quality and distribution of diagenetic fatures in the Funing Formation, Wanglongzhuang Oilfield, Subei Basin, Eastern China. Arabian Journal of Geosciences, 2018, 11(15).

[2] Leila, M., & Moscariello, A. Seismic stratigraphy and sedimentary facies analysis of the pre- and syn- Messinain salinity crisis sequences, onshore Nile Delta, Egypt: Implications for reservoir quality prediction. Marine and Petroleum Geology, 2019, 101: 303-321.

[3] Li Yanhua, Wang Hongtao, Wang Mingchuan, Lian Peiqing, Duan Taizhong,& Ji Bingyu.Automatic identification of carbonate sedimentary facies logging based on PCA and KNN.Well Logging Technology, 2017, 41(1): 57-63.

[4] Bhattacharya, S., & Mishra, S. Applications of machine learning for facies and fracture prediction using Bayesian Network Theory

and Random Forest: Case studies from the Appalachian basin, USA. Journal of Petroleum Science and Engineering, 2018, 170: 1005-1017.

[5] Ai, Wang, & Sun. Automatic Identification of Sedimentary Facies Based on a Support Vector Machine in the Aryskum Graben, Kazakhstan. Applied Sciences, 2019, 9(21): 2076-3417

[6] Han, F., Zhang, H., Rui, J., Wei, K., Zhang, D., & Xiao, W. Multiple point geostatistical simulation with adaptive filter derived from neural network for sedimentary facies classification. Marine and Petroleum Geology, 2020, 118: 104406.

[7] Song, S., Hou, J., Dou, L., Song, Z., & Sun, S. Geologist-level wireline log shape identification with recurrent neural networks. Computers & Geosciences, 2020, 134: 104313.

[8] Aboagye-Sarfo, P., & Mai, Q. Seasonal analysis of emergency department presentations in Western Australia, 2009/10-2014/15. Journal of Applied Statistics, 2018, 45(15): 2819-2830.

[9] Chen, Y., Zu, S., Wang, Y., & Chen, X. Deblending of simultaneous source data using a structure-oriented space-varying median filter. Geophysical Journal International, 2020, 222(3).

[10] Huang, S., & Wan, S. A Total Variation Denoising Method Based on Median Filter and Phase Consistency. Sensing and Imaging, 2020 21(1): 19.

[11] Lin, P., Peng, Z., Lai, Y., Cheng, S., Chen, Z., & Wu, L. Short-term power prediction for photovoltaic power plants using a hybrid improved Kmeans-GRA-Elman model based on multivariate meteorological factors and historical power datasets. Energy Conversion & Management, 2018 177: 704-717.

[12] Li, Z., Li, Q., Zou, X., & Ren, J. Causality extraction based on self-attentive BiLSTM-CRF with transferred embeddings. Neurocomputing, 2021 423: 207-219.

[13] Deng, J., Cheng, L., & Wang, Z. Attention-based BiLSTM fused CNN with gating mechanism model for Chinese long text classification. Computer Speech & Language, 2021 68: 101182.

[14] Aslan, M. F., Unlersen, M. F., Sabanci, K., & Durdu, A. CNN-based transfer learning–BiLSTM network: A novel approach for COVID-19 infection detection. Applied Soft Computing, 2021, 98: 106912.

[15] Fan, D., Sun, H., Yao, J., Zhang, K., Yan, X., & Sun, Z. Well production forecasting based on ARIMA-LSTM model considering manual operations. Energy, 2021, 220: 119708.

[16] Kingma, D.P., Ba, J. Adam: a method for stochastic optimization. CoRR, 2014

[17] Omidpour, A., Moussavi-Harami, R., Mahboubi, A., & Rahimpour-Bonab, H. Application of stable isotopes, trace elements and spectral gamma-ray log in resolving high-frequency stratigraphic sequences of a mixed carbonate-siliciclastic reservoirs. Marine and Petroleum Geology, 2021, 125: 104854.

[18] Krassay, A.A. Outcrop and drill core gamma ray logging integrated with sequence stratigraphy: examples from Proterozoic sedimentary successions of northern Australia. AGSO Journal of Australian Geology and Geophysics, 1998, 17(4): 285-299.

[19] Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., Asadi, H. Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. AJR. American journal of roentgenology, 2019, 212(1): 38-43.

[20] Hewage, P., Behera, A., Trovati, M., Pereira, E., Ghahremani, M., Palmieri, F., & Liu, Y. Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station. Soft Computing, 2020, 24(21): 16453-16482.