

Research Article

Prediction Model of Fault Block Reservoir Measure Index Based on 1DCNN-LightGBM

Bin Wang,¹ Dawei Wu,² Kai Zhang ,³ Huaqing Zhang,⁴ and Chao Zhang¹

¹Exploration and Development Research Institute of Shengli Oil Field, SINOPEC, Dongying 257100, Shandong, China

²China University of Petroleum East China-Qingdao Campus, School of Petroleum Engineering, Qingdao 266580, Shandong, China

³School of Civil Engineering, Qingdao University of Technology, Qingdao 266520, China

⁴China University of Petroleum East China-Qingdao Campus, College of Control Science and Engineering and College of Sciences, Qingdao 266580, Shandong, China

Correspondence should be addressed to Kai Zhang; zhangkai@qut.edu.cn

Received 1 June 2022; Revised 31 July 2022; Accepted 11 October 2022; Published 3 February 2023

Academic Editor: Debo Cheng

Copyright © 2023 Bin Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In view of the shortcomings of the prediction method of future development measures and indicators of fault block reservoir in the current oilfield practical application, a prediction method of fault block reservoir measures and indicators based on the random forest method and LightGBM is proposed, which can help the oilfield make more effective decisions in the middle and later development. Firstly, using the advantages of random forest (RF) in dealing with high-dimensional data sets, the main controlling factors are selected by feature analysis. Then, the measure prediction model is established by using the 1DCNN-LightGBM algorithm. Firstly, 1DCNN processes the reservoir dynamic data and then trains the LightGBM model with the extracted time series characteristics and static data characteristics as input to predict the measure indexes of fault block reservoir. The evaluation results show that the prediction models proposed in this paper have good performance and can obtain more accurate prediction results and more stable prediction performance. It provides a basis for the future planning and optimization of the oilfield.

1. Introduction

With the continuous development of the petroleum industry, most fault block oilfields in China have entered the stage of medium and high water cut adjustment and development, and there are problems such as significant production decline and fast water cut rise. Timely evaluating the development effect of fault block oilfields and establishing a complete potential evaluation index prediction method are of great significance in achieving stable production and increasing production [1]. Compared with ordinary reservoirs, the structure of fault block reservoir is relatively more complex, faults are developed, and there is an independent oil-water system composed of multiple separate fault blocks [2]. Therefore, the relationship of this type of reservoir is more complex, and it is difficult to realize the

fine modeling of fault block reservoir structure, which leads to the significant increase in development difficulty and the deviation of the overall development effect of the reservoir. Previously, based on practical experience and theoretical derivation, oilfield workers proposed a variety of traditional methods for development potential evaluation, mainly including well pattern density method, numerical simulation method, water drive curve method, production decline method, and prediction model method [3, 4]. However, in the practical application of the abovementioned methods, due to the complex boundary conditions and oil-water distribution of the fault block reservoir, there will be a large deviation between the prediction results and the real situation.

In recent years, machine learning technology has developed rapidly and has been gradually applied to various

industrial fields, which has greatly promoted industrial production. Oilfield researchers have also found the advantages of machine learning technology in prediction and applied it to the petroleum field [5]. Machine learning methods such as artificial neural network, support vector machine, and swarm intelligence have been widely used in the field of production performance analysis [6]. It also provides a new method for index prediction of old oil fields, which is different from the previous method of studying oil reservoirs based on experience and classical reservoir engineering theory. Sun et al. used the LSTM neural network method instead of the decline curve analysis method to predict oilfield production, which improved the prediction accuracy [7]. Cai used the least squares support vector machine method based on particle swarm optimization to predict shale gas reservoir production. The results show that this method has good convergence and prediction accuracy [8]. However, at present, the relevant machine learning methods are relatively simple and do not consider the physical properties of the reservoir, which cannot meet the needs for rapid and accurate prediction of the oilfield [9]. Aiming at the problems of poor accuracy and low reliability of fault block reservoir potential evaluation methods, a new high-precision prediction method based on the 1DCNN LightGBM neural network model is proposed in this paper, which provides a scientific and accurate judgment basis for fault block reservoir planning and optimization.

2. Prediction Model of Measure Index of Fault Block Reservoir

This paper presents a fault block reservoir measure index prediction model based on the 1DCNN LightGBM neural network. Its framework process mainly includes three steps, namely, data cleaning and data standardization, feature selection, and neural network model construction.

2.1. Data Cleaning and Data Standardization. In the data directly obtained from the oil field, some defective data are often produced due to format conversion, staff operation errors, instrument failures, and other reasons. If these data are not processed, the error of prediction results will become larger and even impossible to calculate. The cleaning of the original data of index prediction of old oil fields mainly includes the following: (1) eliminating the recorded false data, such as the annual value of permeability difference and formation crude oil is 0; (2) eliminating the reservoir data with missing data; for example, there are no records of liquid production rate and oil increase in the current year in the fault block data; (3) eliminate the wrong data of logical verification, etc.

At the same time, normalizing the original data can remove the dimension of different evaluation indexes and the influence of dimension units, solve the comparability problem between data indexes, and is suitable for comprehensive comparative evaluation. The specific normalization formula is

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \quad (1)$$

where x_i and x_i^* represent the values before and after normalization, respectively; x_{\min} and x_{\max} represent the minimum and maximum values in the sample data, respectively.

As shown in Figures 1 and 2, the data of the total number of wells is partially missing, and there should be no negative value in the annual remaining recoverable reserves. Considering that the total sample size available from the oilfield is large and the proportion of such data is small, this paper chooses to delete the data containing missing values and abnormal values to avoid affecting the subsequent model training. The normalized results are shown in Figure 3. After normalization, the two indicators are in the same order of magnitude, which is suitable for subsequent comprehensive comparative evaluation.

2.2. Selection of Main Control Factors of Fault Block Reservoir with Random Forest Method. There are many factors affecting the development effect of fault block reservoirs. This paper considers the incremental oil production in the current year after implementing the measures as the index to evaluate the future development effect of the reservoir and selects the main controlling factors for the prediction of fault block reservoir from two aspects of dynamic factors and static factors. Static factors mainly refer to the geological factors related to the physical properties of the reservoir itself, which can be obtained through comprehensive logging and other data, such as the reservoir subclass and the medium depth of the reservoir, which can largely determine the development system of the reservoir and affect the final actual development effect. Dynamic factors are constantly changing factors with the development of the oilfield, which is dynamically updated with the real-time development process of the reservoir, such as oil production rate, annual remaining recoverable reserves, and recovery degree of geological reserves. These indicators can reflect the current development effect and oil reservoir production status of the oilfield from different aspects.

Considering that the dynamic and static factors of the oil field may contain a large amount of redundant information and information irrelevant to the prediction index, which are used as inputs, it will increase the sample dimension, slow down the training speed of the model, and even affect the accuracy of the model. Therefore, feature selection is of great significance to reduce the number of features [10]. Compared with other feature selection methods, the random forest method can output the relative importance of each feature [11]. Selecting the features with high correlation can obtain the preferred features for model training to reduce the complexity of the model and avoid overfitting of the model. In addition, the random forest feature selection method has the characteristics of fast training speed, high precision, and no complex parameter adjustment [12]. Therefore, this paper uses the random forest to rank the importance of dynamic and static factors of high-dimensional fault block

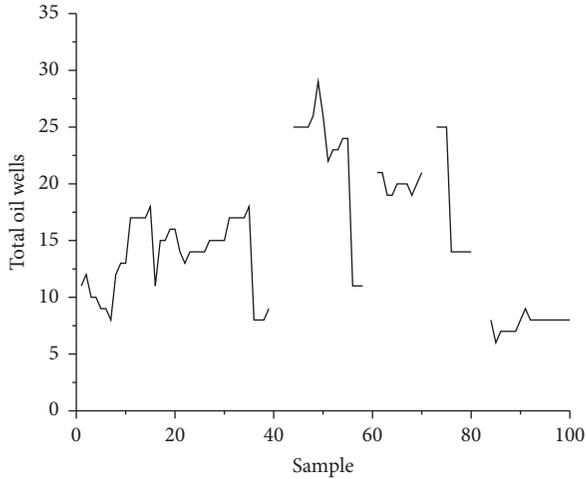


FIGURE 1: Initial data of total number of wells.

reservoirs and selects the main control factors for further machine learning model training.

2.3. 1DCNN-LightGBM Prediction Model. The research content of this paper is the prediction of measure indexes of fault block reservoirs. Due to the complex geological conditions of fault block reservoirs, traditional methods such as the water drive characteristic curve method and production attenuation method are difficult to accurately analyze the relationship between prediction indexes and various main control factors.

Firstly, the 1DCNN neural network is used to train the characteristics of dynamic data and deeply mine the time characteristics of dynamic data. Then, the integrated learning method LightGBM is used to learn the nonlinear mapping from static data and latent temporal data to incremental oil production. The prediction framework has higher prediction accuracy than the signal model.

2.3.1. 1DCNN Neural Network. The convolutional neural network imitates the construction of a biological visual perception mechanism, which has been applied in various fields. This method processes data by using the convolution layer and pooling layer rather than the full connection layer. Due to the structural characteristics such as local connection, weight sharing, and pooling, CNN is usually used for image classification, fault diagnosis, and image recognition [13].

The one-dimensional convolutional neural network (1DCNN) can be applied to a variety of one-dimensional signals. It has a strong processing ability for time series without complex feature extraction. The structure of 1DCNN can be roughly divided into two parts, namely, the convolution layer and the pooling layer, as shown in Figure 4.

As the core of DCNN, the convolution layer uses the same core to traverse the input with a fixed step size. At each traversal position, the convolution kernel and the neurons of the previous layer perform convolution operation and finally

generate the feature map. The specific convolution layer calculation formula is

$$y^{(i,j)} = \omega * x^{(i,j)} = \sum_m \sum_n x^{(i+m,j+n)} \omega^{(m,n)}, \quad (2)$$

where ω represents the weight of the convolution kernel, and $x^{(i,j)}$ represents the convolution region of the starting point (i, j) .

The pooling layer reduces the number of parameters by compressing the extracted features. Therefore, adding a pooling layer can speed up the calculation and prevent overfitting. There are generally two ways of pooling, namely, maximum pooling and mean pooling. The maximum pool is used in this study, which can output the maximum value of neurons in the perception domain. Its calculation formula is shown as follows:

$$y_i = \max_{l \times l} (x_i), \quad (3)$$

where x_i and y_i are input features and output features, respectively, and l is the width scale of the pool area.

2.3.2. LightGBM Model. LightGBM is a new integrated learning model proposed by Microsoft in 2015. The principle of LightGBM is similar to XGBoost, which is based on Taylor expansion of loss function, so the residuals can be approximated [14]. The difference between the two is reflected in the following two aspects:

- (1) By using the decision tree strategy of leaf wise (Figure 5), instead of the level-wise strategy used in XGBoost, only the node with the largest splitting gain is selected for splitting, avoiding the waste caused by the small gain of some nodes. However, the number of splits will also increase, making the final decision tree too large. Therefore, it is necessary to set a maximum depth for LightGBM.
- (2) LightGBM uses a histogram-based decision tree algorithm. The basic idea is to discretize the continuous features into k integers and construct a histogram with width k . When traversing the data, the discrete value is used as the cumulative index of the statistical data in the histogram. After traversing the data once, the histogram accumulates the required statistical information and then traverses according to the discrete value of the histogram to find the optimal segmentation point. It uses the presorting algorithm used in the accurate algorithm, which reduces the use of memory and improves the training speed of the model.

2.3.3. 1DCNN- LightGBM Hybrid Model. There are two main factors affecting the production of fault block reservoirs, namely, dynamic data and static data. The traditional method only processes them according to the same data and cannot comprehensively extract the implicit relationship between feature data and index data, which often leads to poor prediction results. This paper proposes a fault block

TABLE 1: Comparison of the accuracy of each model on the training set.

Model	Accuracy (%)
1DCNN	62.8
LightGBM	65.7
1DCNN-LightGBM	72.3

The bold value indicates the best one in all results.

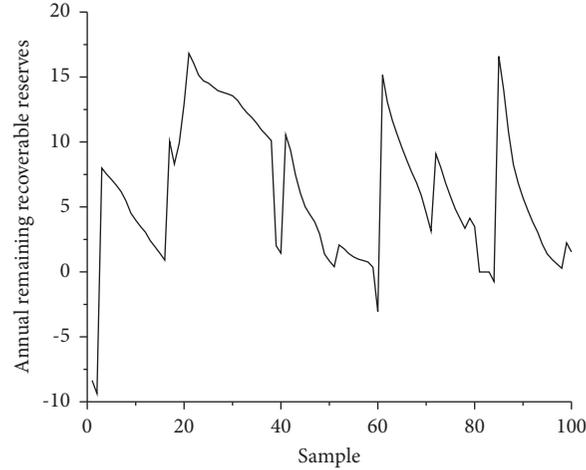


FIGURE 2: Initial data of annual remaining recoverable reserves.

index prediction method based on the 1DCNN-LightGBM hybrid model, and its model framework is shown in Figure 6.

The method mainly includes the following steps: Firstly, by preprocessing the data set of the fault block reservoir and using the random forest method to obtain the ranking of feature importance, the strong correlation characteristics of measure indicators are obtained. Secondly, the 1DCNN model is used as the feature extractor to extract the features of dynamic data, which are combined with the static features of the data and input into the LightGBM model for training. Finally, the evaluation index is calculated to evaluate the model.

Among them, 1DCNN is mainly used to adaptively extract features from input data. This layer mainly includes two convolution layers and two pooling layers. The pooling layer is the maximum pooling. The extracted feature vector is input to the LightGBM model in the form of time-series data from the last pool layer. At the same time, the static characteristics of the fault block reservoir are also input as the LightGBM model, so that the hybrid algorithm can integrate the dynamic and static information in production data and better predict the index of fault block reservoir.

3. Experimental Study

3.1. Construction of Dynamic and Static Database. In order to verify the effectiveness of the algorithm framework proposed in this paper, we use this algorithm to predict the actual data of an oilfield block. After data preprocessing and removing null values and outliers, a fault-block reservoir sample database was established, including 759 samples. There are 17 dimensions of sample feature variables in the database, and

the label that needs to be predicted is the incremental oil production in the current year after implementing measures.

3.2. Feature Selection. The features are selected based on data collected in the dynamic and static databases. By using the random forest method, the relative importance of the initial factors to the labels can be calculated, and the importance between each feature and the output is shown in Figure 7. Here, we select the top 5 factors to construct the main control factor database of fault block oilfield index prediction. The static factors in the main control factors include producing geological reserves and recoverable reserves, and the dynamic factors include the total number of oil wells, the total number of measured wells, and the number of effective wells.

3.3. Model Training and Performance Analysis. The data set after feature selection is randomly divided into training set and test set according to the ratio of 7:3. The 5 features selected are input into 1DCNN, LightGBM, and 1DCNN-LightGBM models, respectively.

After training, the prediction results of each model can be obtained, as shown in Figures 8–10. Figure 8 shows the prediction results when the 1DCNN model is used, Figure 9 shows the prediction results of the LightGBM model using the ensemble learning method, and Figure 10 shows the prediction results of the 1DCNN-LightGBM model proposed in this paper. As can be seen from the figure, although the results of the 1DCNN model can predict the general trend, there are cases where the prediction of individual samples is inaccurate; the predicting results of the LightGBM

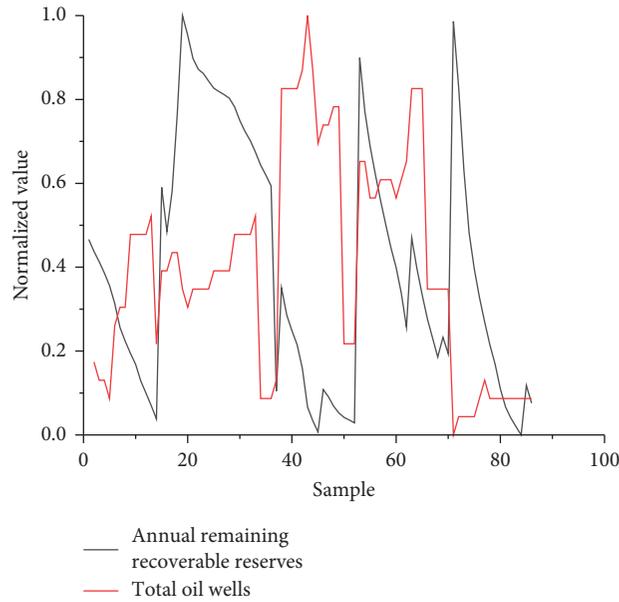


FIGURE 3: Annual remaining recoverable reserves and total number of oil wellheads after normalization.

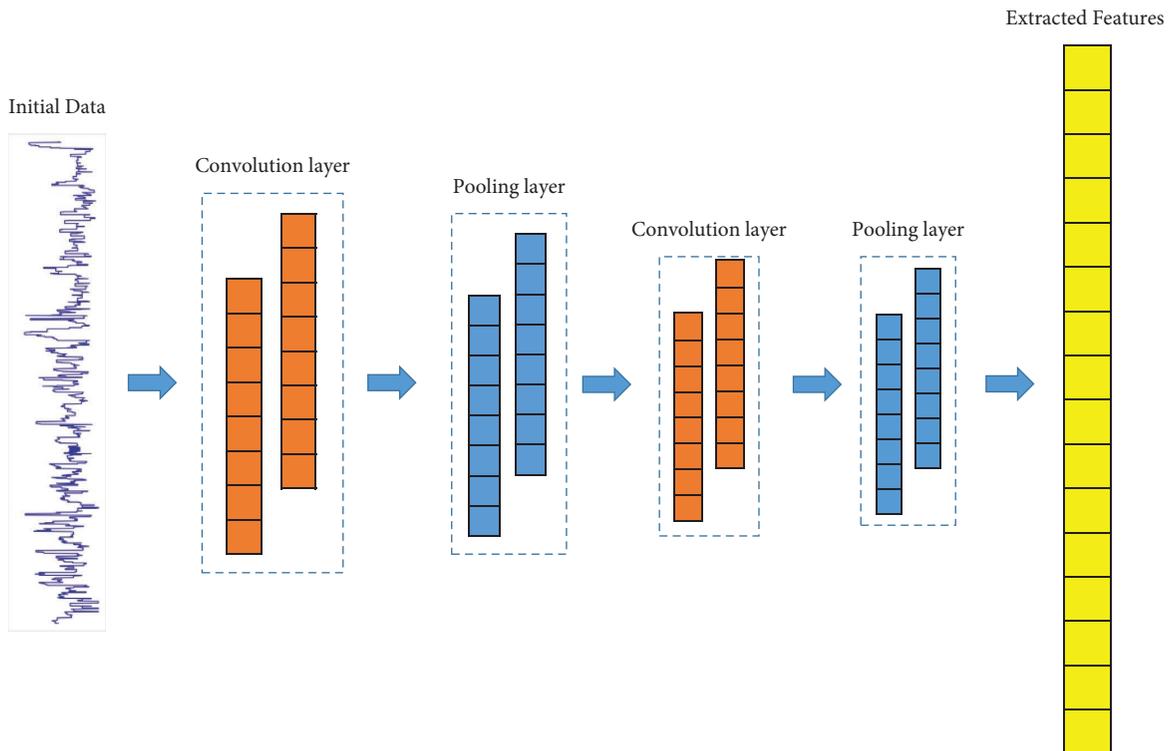


FIGURE 4: 1D convolutional neural network.

model are better than the results of the 1DCNN model, but the details are still not accurately predicted; in contrast, the prediction results of the 1DCNN-LightGBM model proposed in this paper are more in line with the actual value of the incremental oil production in the current year after implementing measures, showing stronger prediction reliability. In addition, compared with the standard 1DCNN model and the LightGBM model, the 1DCNN-LightGBM

hybrid model has higher stability and better adaptability to samples with a wider range of input and output variables. The 1DCNN-LightGBM hybrid model utilizes 1DCNN to extract high-level time series features from dynamic production data and combines it with reservoir static data as the input of ensemble learning, fully excavating the time correlation and data space nonlinear correlation contained in the data. Therefore, compared with other models in the

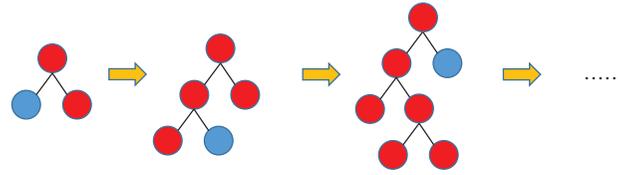


FIGURE 5: Decision tree strategy based on leaf growth of LightGBM.

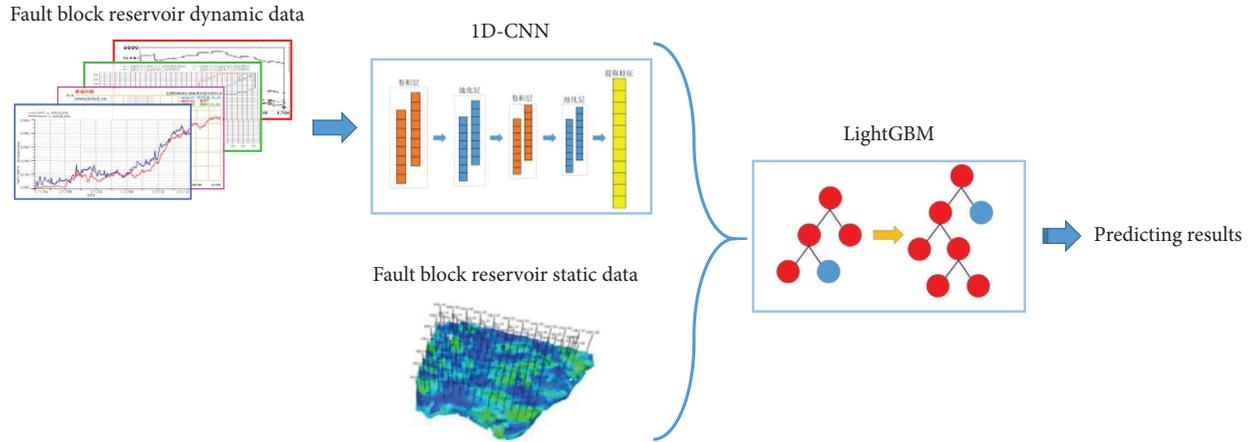


FIGURE 6: 1DCNN-LightGBM hybrid model structure.

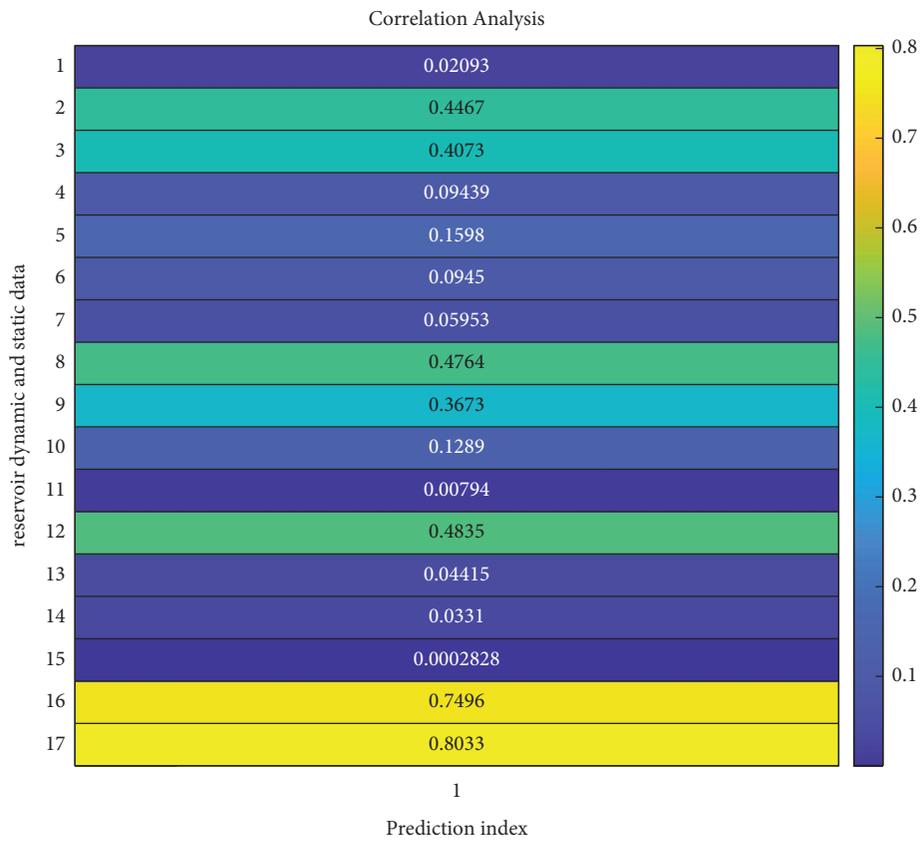


FIGURE 7: Correlation analysis between dynamic and static factors of fault block and measure index based on random forest method.

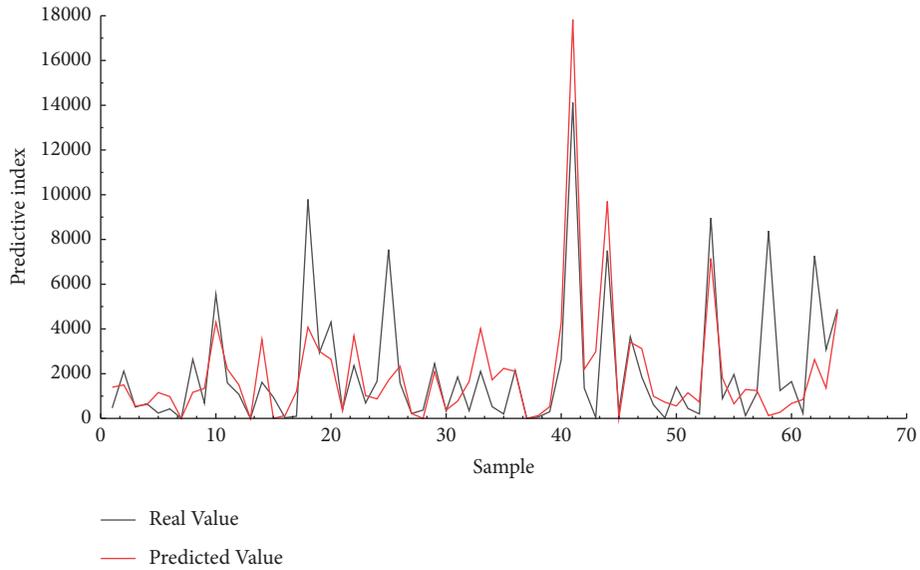


FIGURE 8: 1DCNN prediction results.

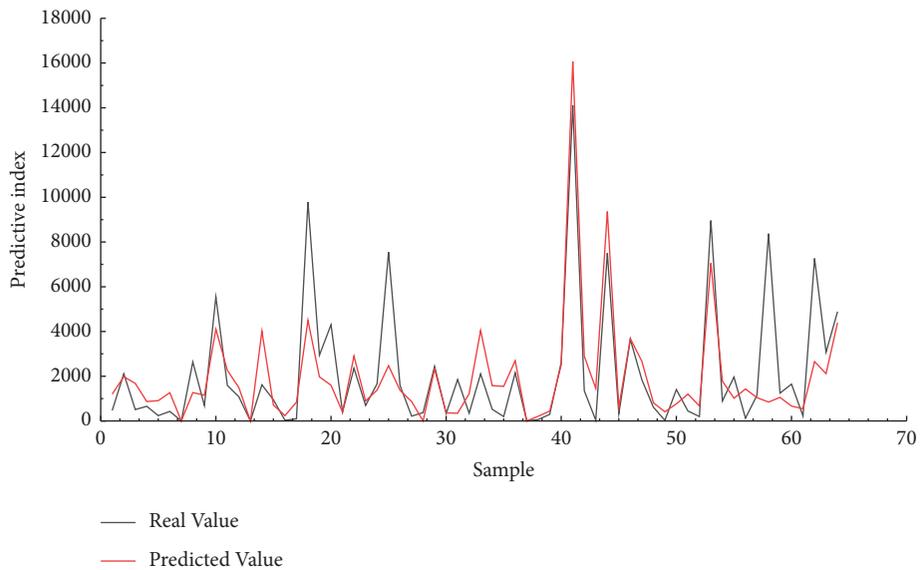


FIGURE 9: LightGBM prediction results.

combination, the prediction accuracy can be significantly improved.

In order to further verify the effectiveness of the model, each model is trained 10 times with the same data, and the average training accuracy is shown in Table 1. The accuracy rates of 1DCNN, LightGBM, and 1DCNN-LightGBM models are 62.8%, 65.7%, and 72.3%, respectively. The accuracy of 1DCNN-LightGBM is 6.6% higher than that of the

LightGBM classification method and 11.5% higher than 1DCNN. It can be seen that the accuracy of the model proposed in this paper is higher than that of the single 1DCNN and LightGBM models. The experimental results can reflect the high potential application value of the 1DCNN-LightGBM algorithm in the prediction of fault-block reservoir indicators, which can extract deep-level information from dynamic and static data.

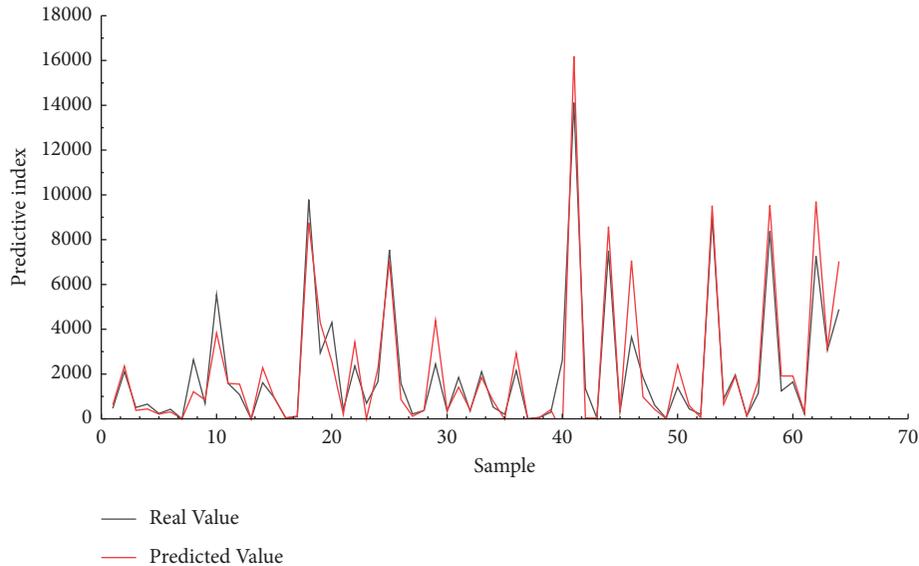


FIGURE 10: 1DCNN-LightGBM prediction results.

4. Conclusions and Discussion

Based on the characteristics of two 1DCNN and LightGBM, this paper proposes a hybrid prediction model combining LightGBM with 1DCNN. In this paper, the main controlling factor which has the highest correlation with the predictive index is selected by using the random forest method. After that, 1DCNN is adopted to extract time series features from dynamic production data which is then integrated with static data as model input to train LightGBM. The results show that the performance of the 1DCNN-LightGBM model is significantly improved compared with the 1DCNN and LightGBM models. It indicates that the mixed model has good prediction performance, provides a new way for the prediction research of reservoir measures index, and has guiding significance for the formulation of reservoir development plan.

Data Availability

Access to data is restricted as third-party rights.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Research and Application of Old Oil Field Potential Evaluation and Planning Optimization Method of Sinopec (p20070-3).

References

- [1] A. Mamghaderi and P. Pourafshary, "Water flooding performance prediction in layered reservoirs using improved

capacitance-resistive model," *Journal of Petroleum Science and Engineering*, vol. 108, pp. 107–117, 2013.

- [2] H. Zhou, X. Chang, and J. Hao, "Development technique and practice of horizontal wells for complex fault-block reservoirs in Jidong oilfield," in *Proceedings of the International Oil and Gas Conference and Exhibition in China*, Society of Petroleum Engineers, Beijing, China, December 2006.
- [3] J. Du, "A method for calculating oil field relative permeability curve by using water drive characteristic curve in high water cut stage," *Journal of Geoscience and Environment Protection*, vol. 10, no. 02, pp. 47–54, 2022.
- [4] P. Zhang, H. Chen, Q. Bai, and H. Wei, "New applications of low-frequency absorption attenuation gradient analysis for potential evaluation in oilfield development stage," *SEG International Exposition and Annual Meeting*, San Antonio, TX, USA, 2019.
- [5] K. Zhang, X. Zhao, and L. Zhang, "Current status and prospect for the research and application of big data and intelligent optimization methods in oilfield development," *Journal of China University of Petroleum (Edition of Natural Science)*, vol. 44, no. 4, pp. 28–38, 2020.
- [6] R. Zhang and H. Jia, "Production performance forecasting method based on multivariate time series and vector autoregressive machine learning model for waterflooding reservoirs," *Petroleum Exploration and Development*, vol. 48, no. 1, pp. 201–211, 2021.
- [7] J. Sun, X. Ma, and M. Kazi, "Comparison of decline curve analysis dca with recursive neural networks rnn for production forecast of multiple wells," *SPE Western Regional Meeting*, Garden Grove, CA, USA, 2018.
- [8] J. Cai, "Application of least squares support vector machines based on particle swarm optimization in production forecast of shale gas wells," *China Computer & Communication*, no. 13, pp. 35–41, 2019.
- [9] B. Lin and J. Guo, "Discussion on current application of artificial intelligence in petroleum industry," *Petroleum Science Bulletin*, vol. 4, pp. 403–413, 2019.
- [10] H. Zhang, J. Wang, Z. Sun, J. M. Zurada, and N. R. Pal, "Feature selection for neural networks using group lasso

- regularization,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 4, pp. 659–673, 2020.
- [11] X. Li, W. Chen, Q. Zhang, and L. Wu, “Building Auto-Encoder Intrusion Detection System based on random forest feature selection,” *Computers & Security*, vol. 95, Article ID 101851, 2020.
- [12] Y. Ye, Q. Wu, J. Zhexue Huang, M. K. Ng, and X. Li, “Stratified sampling for feature subspace selection in random forests for high dimensional data,” *Pattern Recognition*, vol. 46, no. 3, pp. 769–787, 2013.
- [13] L. Li, R. Situ, and J. Gao, “A hybrid model combining convolutional neural network with XGBoost for predicting social media popularity,” in *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 1912–1917, New York, NY, USA, October 2017.
- [14] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937–1967, 2020.