Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 8, No. 3, 2023

Intelligent Identification Technology of Stratum Sub-Layer Based on Multi-Parameter Integration of Logging While Drilling

Haibo Liang, Xin Jiang, Yi Yang, Jialing Zou

School of Mechanical Engineering, Southwest Petroleum University, Chengdu 610500, China

Abstract: Stratum identification is the division of the stratum lithology of one region, which is an important part of petroleum geology research. How to effectively improve the accuracy and efficiency of stratum recognition is an important issue in oil exploration and development. During the traditional oil and gas drilling process, the logging data is commonly used as the main basis to conduct artificial stratum division. The challenges encountered are high labor intensity and excessive dependence on artificial experience for identification accuracy. By comprehensively considering the synergy of multiple parameters in oil and gas drilling, we propose an intelligent sub-layer division model based on the LightGBM algorithm. First, the data set was formed by normalizing, de-noising, and smoothing the drilling engineering parameters and combining them with the element logging parameters. Then, the LightGBM algorithm was applied to build the sub-layer division model, and the deep neural network and support vector machine was introduced for comparative analysis. Finally, the input parameters of the model were optimized by the principal component analysis method to realize the intelligent identification of the stratum sub-layer. The application results of a certain block in the central Bohai Sea oil field showed that the intelligent identification of stratum sub-layer while drilling could be realized. The use of the model and combination of the logging while drilling data with high recognition accuracy provided a crucial theoretical model for the transformation of stratum sub-layer identification technology.

Keywords: Logging While Drilling; Sub-Layer Division; LightGBM; Parameter Optimization; Intelligent Identification.

1. Introduction

During the oil and gas drilling, the bit crushes rocks in the stratum and drills through different strata until reaching the target. Different strata have different drilling characteristics, related to oil and gas capacity and drilling accidents. Therefore, precisely identifying strata provides an important guarantee for the correct selection of construction parameters, ensuring drilling safety and improving oil and gas exploration efficiency.

There are two traditional stratum identification methods in the oil drilling site. The first method is to collect and observe rock residues from rock crushing during drilling. The second method is to conduct laboratory analysis and tests on the core of the rock taken from the reservoir stratum. These methods put high requirements on the staff and are prone to low accuracy, heavy workload, and high cost of rocks core analysis. [1]

In recent years, the steady improvement of logging while drilling (LWD) technology has provided site petroleum geological experts with a comprehensive and accurate basic data source to identify sub-layers.[2] Scholars conducted research and enriched the identification methods of the stratum sub-layer while drilling. [3] ~ [5] In addition to the excellent prediction, analysis, and calculation capabilities, the use of computer and artificial intelligence in the petroleum industry makes the identification technology of the stratum sub-layer actively move towards intelligence. Pedram Masoudi [6] et al. applied a neural network to classify and identify oil layers in a carbonate environment. The results showed that the classification accuracy obtained by the neural network method was 85% higher than that obtained by the geological method. Ahmed Ali Zerrouki [7] et al. proposed a method to predict fracture porosity by fuzzy sorting and the artificial neural network. Baijie Wang [8] et al. proposed an intelligent framework integrating fuzzy sorting and a multi-layer perceptron neural network. These experiments show that the framework is capable of distinguishing the characteristics of reservoirs. Réda Samy Zazoun [9] et al. applied a conjugate gradient descent algorithm to train the neural network and predicted fractured reservoirs. Data showed that experimental predictions were in line with the actual results. Yunxin Xie [10] et al. applied five typical machine learning methods, including naive Bayes, support vector machines, artificial neural networks, random forests, and gradient tree lifting, to identify and evaluate stratum lithology. Through comparing and analyzing the classification of different models, the results show that random forest and gradient lifting tree are two prominent algorithm choices. Zhang H [11] et al. established a reasoning model for intelligent recognition of stratum lithology using the improved fuzzy clustering algorithm SVM method. The site application showed that the accuracy of the stratum recognition method can reach 90.9%. Zhou Jinhui [12] et al. designed the structure and output mode of the three-layer feedback-free feedforward artificial neural network model. The dynamic information, including drilling rate, bit pressure, and torque collected by the input sensor was collected to identify the drilled stratum, showing high performance. To solve the problem of low accuracy in identifying stratum by logging data, Xia Hongquan [13] et al. optimized the parameters of the data while drilling, combined with the logging data while drilling, and finally processed the data while drilling in real time using gray correlation analysis. This improved the capability of tracking the geological target layer in horizontal well drilling. Combining the advantages of logging information and logging data, Yang Sitong [14] et al. used BP neural network for comprehensive processing and

realized the oil and gas identification in low porosity and low permeability reservoirs. Sebtosheikh [15] et al. applied a support vector machine algorithm to successfully predict the lithology of heterogeneous carbonate reservoirs. Dong [16] et al. applied the improved Linear Discriminant Analysis (LDA) method to identify lithology and achieved good recognition results through experiments with on-site data sets. Fengqi Tan [17] et al. compared and analyzed four clustering algorithms for reservoir classification. The results show that the application of standard K-means algorithm based on division can meet the best fitting of actual reservoir geological characteristics and the maximum accuracy of reservoir classification.

Through the development of machine learning technology, Chen [18] et al. improved the Gradient Boosting Decision Tree (GBDT) and proposed an optimally distributed decision-making gradient lifting library XGBoost with high efficiency, flexibility, and portability. Dev [19] et al. and Sun [20] et al. adopted XGBoost to identify the stratum lithology and confirmed that the model was of higher potential than GBDT in stratigraphic and lithology identification by analyzing the experimental results. However, some experiments found that XGBoost was insufficient in processing high-dimensional massive data. [21] Therefore, Qi M [22] proposed another model LightGBM that overcame the problems of slow training speed and the high likelihood of over-fitting of XGBoost to a certain extent and has been widely applied in classifiers.

In conclusion, logging while drilling technology and artificial intelligence algorithm have been integrated into the practice of stratum identification engineering. The aim is to solve the low efficiency of stratum division based on artificial experience, ensuring drilling safety and efficiency, and improving the accuracy of stratum identification. Based on the logging while drilling data from a block in the central Bohai Sea oil field, we fully considered the synergy of multiple parameters in oil and gas drilling, used the LightGBM algorithm to identify the sub-layer in the study area and compared the recognition effect with traditional classification models, Support Vector Machine (SVM) and Deep Neural Networks (DNN). Finally, the principal component analysis method was applied to optimize the input parameters of the model and realized the accurate prediction of the stratum sub-layer in the study block.

The main content of the rest chapters is summarized as follows: Chapter II, the introduction to the principle of the LightGBM algorithm; Chapter III, the introduction to geological situation and process design of the study area; Chapter IV, the sub-layer division model based on LightGBM is established, including the comparison with the traditional classification model and the optimization of the model input parameters; Chapter V, the prediction accuracy and generalization capability of the model are further verified with engineering examples; the conclusion was proposed in last chapter.

2. Algorithm Principle of LightGBM

LightGBM algorithm is a gradient upgrading framework based on the decision tree algorithm. The advantages include quicker training efficiency, lower memory usage rate, and higher prediction accuracy. Also supports efficient parallel training, and can quickly process large-scale data[23].

As shown in Figure 1, the LightGBM algorithm adopts the histogram optimization algorithm, whose basic idea is to divide the continuous floating-point data into an independent intervals and construct a histogram with a width of n simultaneously.

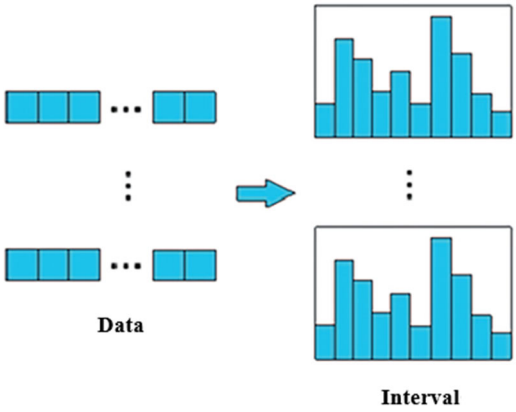


Figure 1. Histogram optimization

There are two types of information in each interval of the histogram. The sum of sample gradients and the number of samples in each interval. During data traversal, the discrete value is used as an index and the statistics are accumulated in each interval of the histogram. Then the discrete value is traversed to find the optimal segmentation point. In terms of calculation cost, the histogram optimization algorithm only generates n times calculation, reducing the storage cost and calculation cost [24].

LightGBM algorithm adopts a Leaf-wise leaf growth strategy with depth constraints, which traverse all leaf nodes before each split and select the node with the largest information gain to grow. However, it is noteworthy that the Leaf-wise strategy may generate higher trees, and can reduce over-fitting by controlling the height of the tree and the minimum number of each leaf node. Under the same splitting times, the prediction accuracy of the Leaf-wise strategy is higher and the convergence speed is faster.

In addition, the LightGBM algorithm further applies histogram subtraction technology to improve operational efficiency. The histogram of a leaf can be obtained by subtracting the histogram of its parent node and its brother node. In each node splitting, the histogram of the sub-nodes with fewer samples can be calculated. The histogram of its sibling nodes can be obtained by making a difference, with the doubled speed. LightGBM algorithm with excellent performance was applied to petroleum geology research, whose reliability and flexibility are capable of promoting the rapid development of relevant fields.

3. Geology Overview and Overall Process Design of the Study Area

3.1. Geology Overview

The study area is a block in the central Bohai Sea oil field. Based on the available geology model recognition, and combined with F-K (frequency-wave number domain filtering) to re-processing data, the sub-layers were divided into two types.

Reservoir stratum type I: Mainly located in the tectonic stress compression zone. The high part of an ancient landform, and well areas 4, 7, and 9 with low dark mineral content. The seismic facies was characterized by low-frequency strong amplitude, strong continuous reflection, F-K post-filtering

dense pointy reflection, or high-angle reflection. The drilling revealed that the reservoir stratum development thickness was 160~220m, with an average of 198m, and the average fracture development density of a single well was 2.2~6.8 fractures/m.

Reservoir stratum type II: Mainly located in the strike-slip zone of tectonic stress, the high part of ancient landform, and well areas 2Sa and 11 with low content of dark minerals. The seismic facies was characterized by medium and low-frequency medium and strong amplitude medium continuous reflection, F-K post-filtering dense pointy reflection, or high-angle reflection. The drilling revealed that the reservoir stratum development thickness was 160~190m, with an average of 175m. The average fracture development density of a single well was 2.0~4.4 fractures/m.

To study the reservoir stratum in the test area, F-K filtering data was used for analysis. The data revealed a reflection blank area in the reservoir stratum at the northeast area of the main body in the test area. There was still uncertainty in the reservoir stratum, which requires further application with production data to make a fine division of the reservoir stratum in the block.

3.2. Overall Process

Following the above description, we adopted the intelligent identification model of the stratum sub-layer based on the multi-parameter fusion of Logging While Drilling, and its model test flow chart as shown in Figure 1. The major operation steps of the experimental process were as follows: ① Pre-treating the element logging parameters, manually removing elements with less content, and conducting 3σ standard de-noising and five-point three-time smoothing for drilling engineering parameters, as well as combining with element logging parameters to form new data samples; ② Dividing the obtained data samples into training sets and test sets according to a certain proportion, and setting the parameters of each model and conducting model training; ③ Inputting the test set into each model, comparing the prediction effect of different models through evaluation indicators, and selecting the optimal prediction model; ④ Optimizing the input parameters of the optimal model, and taking a well in the study block as an example to apply and verify the model to get the result of sub-layer identification.

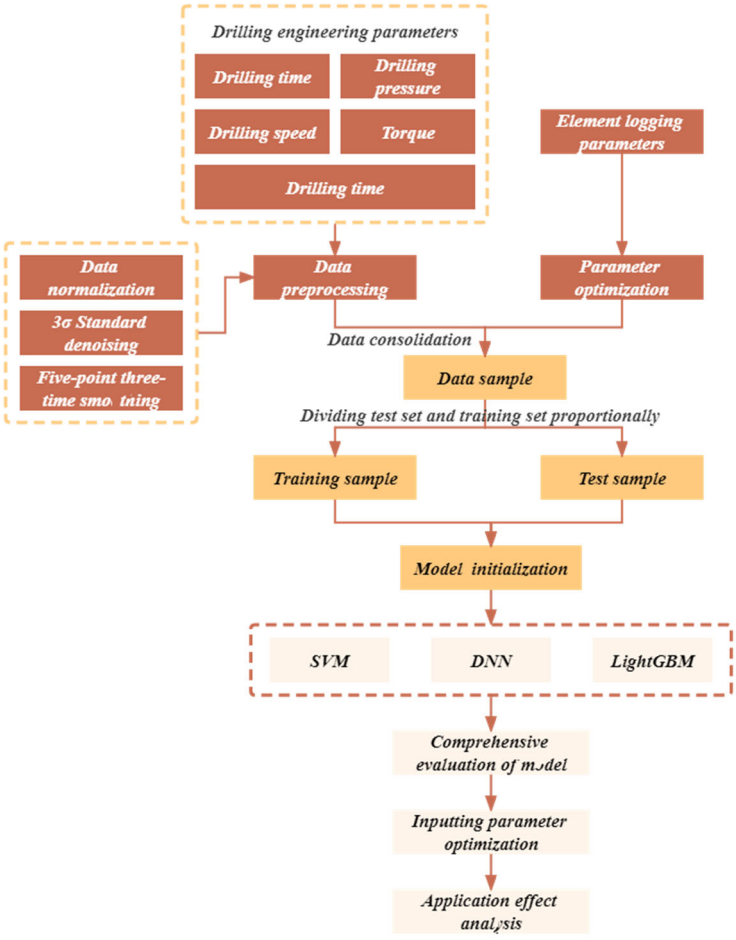


Figure 2. Flow chart of intelligent identification of sub-layer

4. Sub-Layer Division Model Based on LightGBM

4.1. Data Processing

There were 17 characteristics in element logging data, including extremely low element characteristics V, Ni, and Sr. To simplify the model input and improve the operation speed, priority was given to the element characteristics with percentage content greater than 0.01, and the selected parameters were Na, Mg, Al, Si, S, K, Ca, Ti, Mn, and Fe.

There were differences in the dimensions of the original logging while drilling parameters with varied physical meanings. The drilling time, drilling pressure, drilling speed, torque, and fracture pressure gradient, were processed and converted the original data to [0,1] intervals by linear

normalization:

$$\overline{x_i} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Where, x_i is raw data, x_{max} and x_{min} are the maximum and minimum values of data respectively.

To prevent the influence of gross errors in data sets on model training the 3σ standard was adopted. 3σ Standard requires equal precision measurement for the measured and independently obtained result$x = (x_1, x_2, \dots, x_n)$, where the mean value of x wasμ, and the standard deviation was σ. If x_i satisfied Equation (2), then x_i was regarded as an abnormal value and was eliminated[25]. After elimination, it was filled with the mean value of its front and back positions. Figure 3 shows the result that partial drilling samples passed 3σ standard treatment.

$$|x_i - \mu| \geq 3\sigma, i = 1,2, \dots, n \quad (2)$$

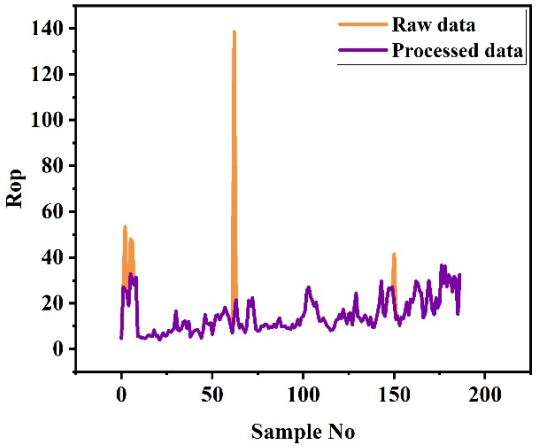


Figure 3. De-noising effect of partial drilling samples

During the logging while drilling process, the instrument itself or the measurement process may be influenced by environmental and artificial factors which may influence the quality of data at the acquisition end. To prevent the influence of noise on the sub-layer recognition model, we used the five-point three-time filtering algorithm on the original data. The principle is a processing method that applies the least squares method to smooth the discrete data with the three-time least square polynomial[26]. Assuming the calculation formula of sequence $x(n), n = 1,2, \dots, n$, a five-point three-time smoothing algorithm is shown in Formula (3).

$$\begin{cases} y(1) = \frac{1}{70}\{69x(1) + 4[x(2) + x(4)] - 6x(3) - x(5)\} \\ y(2) = \frac{1}{35}\{2[x(1) + x(5)] + 27x(2) + 12x(3) - 8x(4)\} \\ \vdots \\ y(i) = \frac{1}{35}\{-3[x(i-2) + x(i+2)] + 12[x(i-1) + x(i+1)] - 17x(i)\} \\ \vdots \\ y(n-1) = \frac{1}{35}\{2[x(n-4) + x(n)] - 8x(n-3) + 12x(n-2) - 27x(n-1)\} \\ y(n) = \frac{1}{70}\{-x(n-4) + 4[x(n-3) + x(n-1))] - 6x(n-2) + x(n)\} \end{cases} \quad (3)$$

Figure 4 shows the effect of partial drilling speed samples after five-points and three-times smoothing. The blue curve in the Figure is the original data, and the orange curve is the smoothed data.

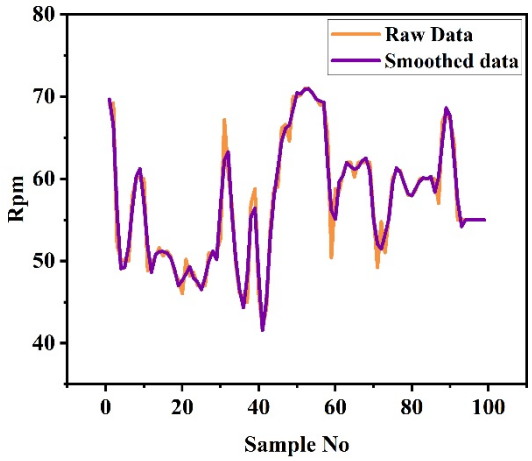


Figure 4. Smoothing effect of partial drilling speed samples

In addition, the acquisition interval depth of element logging data in the original data was 5m. The drilling engineering parameters were taken at the depth of 1m. Therefore, to ensure the consistency of data in depth, the top depth in the positioning element logging data was recorded asD_1. To find the corresponding depth of D_1 in the drilling engineering parameters, 5m from this depth was measured and marked asD_5. The average value of the depth interval $D_5 \sim D_1$ was calculated in the drilling engineering parameter at depth D_1. The method was applied to calculate the average value of the drilling parameter data set at an interval of 5m, and the final data sample can be obtained by integrating it with the element data set.

4.2. Evaluation Indicators

Considering that the number of samples in some wells is unbalanced, in order to better reflect the prediction ability of the model, this paper uses two indicators, F1-Score and AUC[27], to comprehensively evaluate the model. Among them, F1-Score has a good integration of precision and recall rate, which is more evaluative. The calculation formula of F1-Score is as follows:

$$\begin{cases} pre = TP/(TP + FP) \\ rec = TP/(TP + FN) \\ F1_{score} = 2 \times pre \times rec/(pre + rec) \end{cases} \quad (4)$$

Where $precision$ represents the precision and $recall$ represents the recall rate. The prediction results of pattern recognition can generally be classified into four categories: TP is the true case, FP is the false positive case, FN is the false negative case, and TN is the true negative case.

AUC can estimate the discrimination ability of the model without any prior information of misjudgment cost. It is independent of the threshold, which neither varies with different thresholds, nor depends on other parameters, and can effectively measure the overall performance of the classification model [28]. AUC represents the area under the $ROC(receive\ operating\ characteristic)$ curve. The higher the value, the better the robustness of the model. The ROC curve was drawn from the true positive rate TPR as the y-axis and the false positive rate TPR as the x-axis, and its value range was [0, 1]. The calculation formula of TPR and FPR was as follows:

$$\begin{cases} TPR = TP/(TP + FN) \\ FPR = FP/(FP + TN) \end{cases} \quad (5)$$

The pattern recognition result of the model is the output in probability value. At this point, a probability value as the threshold can be acquired as a set of TPR and FPR. Therefore, when all probability values are used as thresholds, multiple sets of TPR and FPR can be obtained, then ROC curve drawn, and AUC calculated.

4.3. Modeling and Analysis of Test Effect

The paper applies digital labels to calibrate the sub-layer types in the study area (Table 1).

Table 1. Sub-layer calibration table

Sub-layer	Representative No.
TypeI	0
Type II	1

We used the logging data of 9 wells in this area as the original data. After data processing, 758 data samples were obtained, including 335 samples of reservoir stratum type I and 423 samples of reservoir stratum type II. The data samples were divided into training sets and test sets according to the ratio of 8:2. A low-level prediction model based on LightGBM was built by applying the operating platform Spyder3 (Python 3.9.12) and introducing the third-party library lightgbm3.2.1. To verify the prediction effect of the proposed model, the third-party libraries skit-learn1.0.2 and torch1.12.1 were used to build a small-scale prediction model based on SVM and DNN respectively. The setting information of each model parameter is shown in Table 2. The comprehensive evaluation was conducted according to the above prediction results from evaluation indicators.

Table 2. Model parameter setting information table

Test Model	SVM	LightGBM	DNN
Initialization Parameters	Penalty coefficient C=1.0; kernel function kernel=rbf; Kernel function parameters gamma=auto; Multi-scheme classification selection decision_ function_ shape=ovo	Maximum depth of decision tree max_ depth=-1; Number of leaf nodes num_ leaves=31; Learning rate learning-rate=0.05; Learning objectives objective=binary Estimator type boosting_ type=gdbt	Number of input neurons input=14 Number of hidden layers Hidden layer=2 Number of output neurons Output=2 Activation function_ functions=tanh&Softmax loss function loss function=Cross Entropy Loss optimizer Optimizer=Adam

The test sets were input into the three trained models, and the F1-Score result of each model was calculated (Figure 5). The Figure shows that the prediction effect of the LightGBM model was better than the other two models. The F1-Score results for identifying Reservoir stratum type I and Reservoir stratum type II were 97.6% and 98.3%; the F1-Score of the SVM model was the lowest, at 74.5% and 85.6% respectively; the F1-Score score of DNN-based small-level prediction model was between the two was 85.9% and 89.8%.

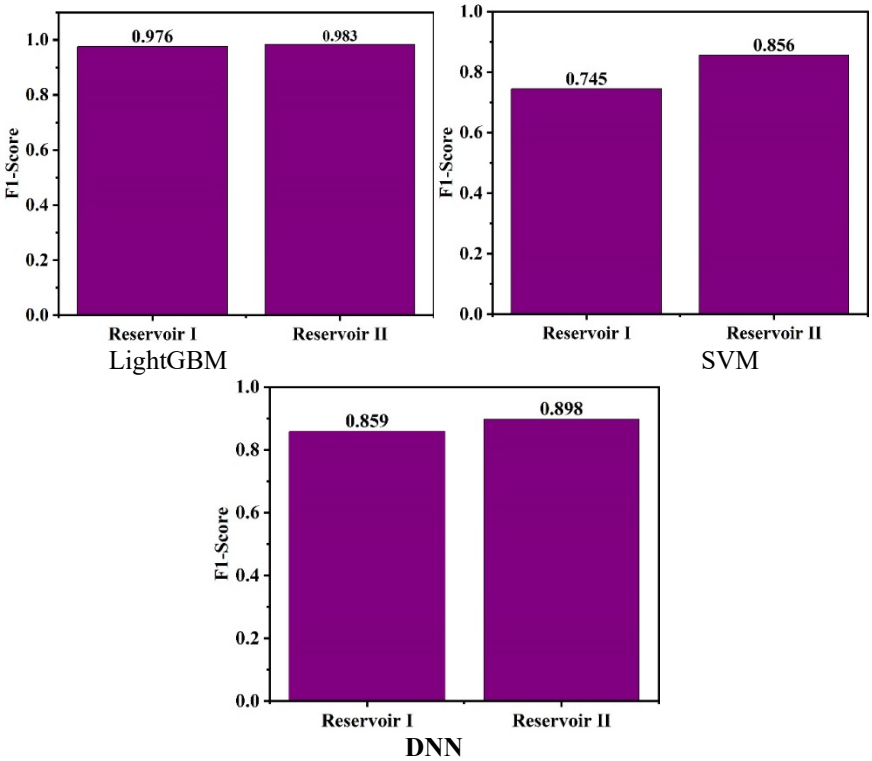


Figure 5. Statistical chart of recognition results of different models on test samples

Furthermore, the paper used the prediction results of each model on the test sample to draw the ROC curve, and the results are shown in Figure 6. The AUC value of the sub-layer prediction model based on SVM was the lowest at 0.7918. The AUC value of the LightGBM model and DNN model was greater than 0.9, and the AUC value of the LightGBM model was the highest at 0.9808, which had higher robustness.

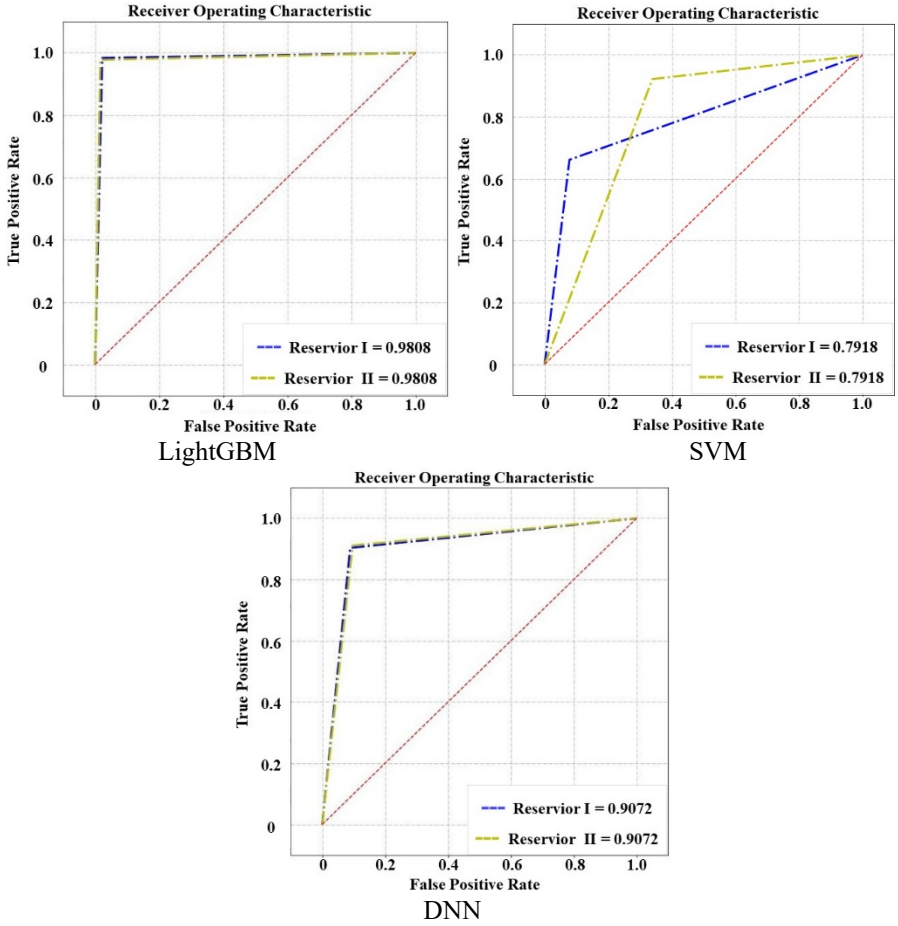


Figure 6. ROC curves of different models

In conclusion, by comparing and analyzing the $F1-Scores$ and AUC values of different models, as well as the recognition of two different layers of class I and class II, the LightGBM-based recognition model performed best. The DNN-based recognition model performance was second, while the SVM-based recognition model performance was the least. The above results showed that the distributed gradient lifting framework based on a decision tree algorithm was superior to some classical machine learning algorithms in different sub-layer recognition, with higher progressiveness and superiority.

4.4. Optimization of Input Parameter

To further study the effect of some elements with lower content on the stratum in the process of logging while drilling, the characteristics of P, Cl, Ba, V, Ni, Sr, and Zr were added based on the original element characteristics. Influenced by the high dimension of parameters, we used the PCA(principal component analysis)[29] method to extract the element features to reduce the dimension of input features and improve the operation speed of the model.

The central idea of the principal component analysis method was to calculate a group of new features in order of importance from multiple unordered original features. They were linear combinations of the original features and were not related to each other. Which was beneficial to calculate several comprehensive variables that reflected more information in the original data and were independent of each other to achieve the purpose of feature dimensionality reduction. [30]

First, we carried out the KMO(Kaiser-Meyer-Olkin) test and Bartlett spherical test[31] on 17 different element characteristics that affected the sub-layer to determine whether the data was suitable for principal component analysis. The inspection results are shown in Table 3:

Table 2. KMO and Bartlett test

KMO sampling suitability quantity		.667
Bartlett sphericity test	Approximate chi-square	7527.340
	Freedom	136
	Significance	.000

a. Correlation-based

We observed that the KMO value was 0.667, greater than 0.5. Meanwhile, Bartlett's significance level was 0, less than 0.05, indicating that there was a collinearity problem between variables, and principal component analysis could be performed. Based on the principle that the eigenvalue was greater than 1, we extracted four principal components F_1, F_2, F_3, F_4. According to the principal component extraction table (Table 4), the characteristic values of F_1, F_2, F_3, F_4 were 20.922, 2.894, 2.692, and 1.964

respectively. Their cumulative contribution rates were 68.609%, 78.099%, 86.972%, and 93.368% respectively, with less information loss. The loading matrix and eigenvector of the element feature are shown in Table (5), of which the eigenvector had the following mathematical relationship with the principal component load matrix and eigenvalues:

$$\text{Principal component eigenvector coefficient} = \frac{\text{Principal component load vector}}{\sqrt{\text{Corresponding principal component characteristic value}}} \quad (6)$$

Table 4. Principal component extraction table (Partial)

Component	Initial eigenvalue			Extracting the sum of the squares of the load		
	Total	Percent Variance	Cumulative%	Total	Percent Variance	Cumulative%
1	20.922	68.609	68.609	20.922	68.609	68.609
2	2.894	9.490	78.099	2.894	9.490	78.099
3	2.692	8.828	86.927	2.692	8.828	86.927
4	1.964	6.441	93.368	1.964	6.441	93.368

Table 5. Characteristics of the load matrix and eigenvector of element

Element	Rescaling composition				Feature vector			
	1	2	3	4	1	2	3	4
Si	-.975	.009	.220	.000	-.213	.005	.134	.000
Al	-.847	.221	-.450	.161	-.185	.130	-.274	.115
Sr	.367	-.271	.360	.124	.080	-.159	.219	.088
Cl	.328	.185	.046	-.137	.072	.109	.028	-.098
Mn	.144	.064	.016	.023	.031	.038	.010	.016
Ca	.491	.831	.230	-.056	.107	.488	.140	-.040
P	.123	.561	.096	.134	.027	.330	.059	.096
S	.273	.511	.089	.020	.060	.300	.054	.014
Zr	.284	-.372	.290	.065	.062	-.219	.177	.046
K	-.199	-.075	-.442	-.190	-.044	-.044	-.269	-.136
Ni	.029	-.047	.375	.272	.006	-.028	.229	.194
Na	-.232	-.298	.322	.137	-.051	-.175	.196	.098
Fe	.492	-.097	.181	.818	.108	-.057	.110	.584
Ti	.344	-.063	.084	.618	.075	-.037	.051	.441
Ba	.351	-.022	.125	.512	.077	-.013	.076	.365
Mg	.204	.316	.225	.485	.045	.186	.137	.346
V	.394	-.131	.396	.453	.086	-.077	.241	.323

Extraction method: Principal component analysis.
Four components were extracted.

The normalized element data for Si, Al, Sr, Cl, Mn, Ca, P, S, Zr, K, Ni, Na, Fe, Ti, Ba, Mg, and V are presented as X_ 1~X_ 17. The obtained eigenvector was multiplied by the normalized element data to obtain the principal component expression (7) ~ (10). On the basis of this expression, data dimension reduction operation is performed as shown below.

$$F_1 = -0.213X_1 - 0.185X_2 + 0.08X_3 + 0.072X_4 + 0.031X_5 + 0.107X_6 + 0.027X_7 + 0.06X_8 + 0.062X_9 - 0.044X_{10} + 0.006X_{11} - 0.051X_{12} + 0.108X_{13} + 0.075X_{14} + 0.077X_{15} + 0.045X_{16} + 0.086X_{17} \quad (7)$$

$$F_2 = 0.005X_1 + 0.13X_2 - 0.159X_3 + 0.109X_4 + 0.038X_5 + 0.488X_6 + 0.33X_7 + 0.3X_8 - 0.219X_9 - 0.044X_{10} - 0.028X_{11} - 0.175X_{12} - 0.057X_{13} - 0.037X_{14} - 0.013X_{15} + 0.186X_{16} - 0.077X_{17} \quad (8)$$

$$F_3 = 0.134X_1 - 0.274X_2 + 0.219X_3 + 0.028X_4 + 0.01X_5 + 0.14X_6 + 0.059X_7 + 0.054X_8 + 0.177X_9 - 0.269X_{10} + 0.229X_{11} + 0.196X_{12} + 0.11X_{13} + 0.051X_{14} + 0.076X_{15} + 0.137X_{16} + 0.241X_{17} \quad (9)$$

$$F_4 = 0.115X_2 + 0.088X_3 - 0.098X_4 + 0.016X_5 - 0.04X_6 + 0.096X_7 + 0.014X_8 + 0.046X_9 - 0.136X_{10} + 0.194X_{11} + 0.098X_{12} + 0.584X_{13} + 0.441X_{14} + 0.365X_{15} + 0.346X_{16} + 0.323X_{17} \quad (10)$$

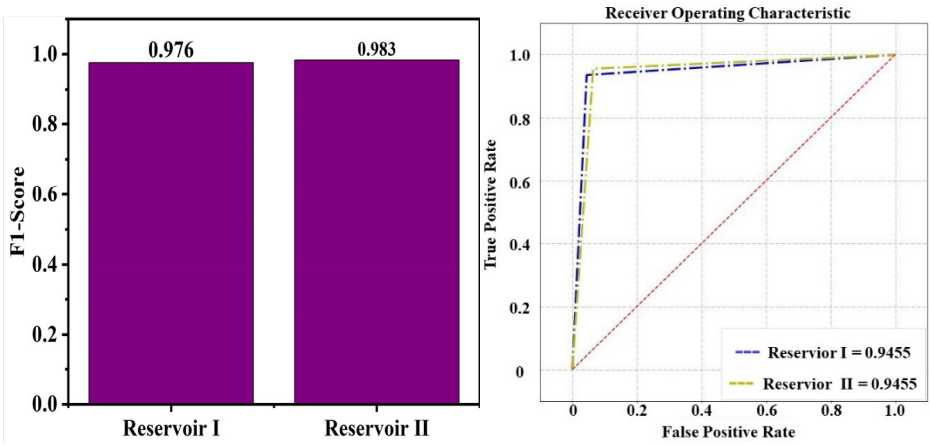
5. Application Examples and Effect Analysis

The stratum identification model established using the above method together with the logging data obtained while drilling, a stratum identification was performed while drilling for multiple wells in the study block, resulting in satisfactory outcomes. Selecting part of the logging while drilling data of a well in the research block as the test sample, a follow-up verification was performed which generates the drilling parameter data set during the collection of drilling engineering parameters, de-noised and smoothed. Subsequently, the principal component expression was used to calculate the principal component factor of element logging data to obtain the element data set. Finally, the two data sets were used to integrate and obtain the final test set (Table 7).

Table 7. Partial data of validation well test set

Well depth	Drilling time	Drilling pressure	Drilling speed	Torque	Fracture pressure gradient	F1	F2	F3	F4
4830	9.362	10.528	60	12.686	2.16	0.140	0.196	0.350	0.439
4835	9.012	7.786	58.4	9.14	2.152	0.118	0.448	0.334	0.316
4840	6.878	6.66	60	8.012	2.154	0.113	0.479	0.372	0.460
4845	9.342	7.158	60	7.756	2.156	0.126	0.292	0.315	0.280
…	…	…	…	…	…	…	…	…	…
5390	14.772	9.814	56.4	12.252	2.2	0.115	0.173	0.433	0.435
5395	18.54	10.022	55	12.248	2.204	0.114	0.155	0.378	0.422
5400	18.428	10.032	55	11.862	2.202	0.124	0.149	0.412	0.394
5405	17.514	10.33	55	12.084	2.2	0.088	0.124	0.381	0.378

As shown in the real label statistics, this data sample includes 53 samples of Reservoir stratum type I and 64 samples of reservoir stratum type II. The test set was input into the built LightGBM model to calculate the F1-Score and draw the ROC curve. The details are shown in Figure 7. The model was used to divide the verification wells into sub-layers. The F1-Score of reservoir stratum type I and reservoir stratum type II reservoirs were 93.5% and 95.6%, respectively and the AUC value was 0.9455. The prediction accuracy and generalization ability of the model meet the identification requirements.



LightGBM ROC curve

Figure 7. Testing well $F1-Score$ score and ROC curve

The model output and real values are projected into the logging curve, which is presented in Figure 8.

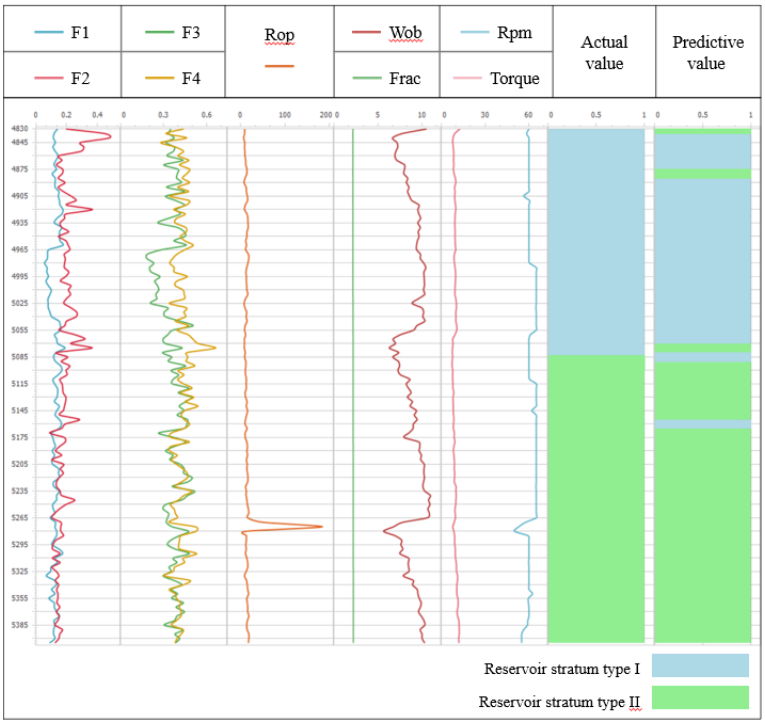


Figure 8. Comparison between conventional stratigraphic sub-layer division and predicted sub-layer division

6. Conclusion

Heavy workload and low accuracy in identifying the sub-layer of stratum by manual use of logging data is a great concern. Therefore, we comprehensively considered the synergy of multiple parameters in logging data while drilling and constructed an intelligent sub-layer division model based on the LightGBM algorithm. Then, optimized the input parameters of the model using the principal component analysis method. We realized a fine division of the sub-layer of stratum in a block of the central Bohai Sea oil field, with conclusions as follows:

Compared with the traditional classification model, deep neural network, and support vector, the LightGBM algorithm had high competence. The training speed improved, the calculation cost was lowered, and also built a small prediction model with a higher $F1-Score$ and AUC value. In addition, the LightGBM algorithm produced better accuracy and robustness, and the overall performance of the model improved.

The principal component analysis method effectively reduced the dimension of the element logging data in the study block. Only four factors attained the interpretation of 93.368% of the total 17 element characteristics, with minimal information loss. The integration with drilling parameters ensured the lower dimension of the data set, ensuring the accuracy of identification, reducing the calculation amount of the model, and improving the calculation speed.

The field application of the study block showed that the intelligent identification model of the formation sub-layer established by the multi-parameter fusion of logging while drilling made full use of the logging while drilling data. Also, it met the requirements of the identification accuracy of the formation sub-layer and compensated for the insufficiencies of artificial identification of the formation. Furthermore, the application provided a key theoretical model for the intelligent identification of the formation sub-layer of the logging while drilling in this block.

7. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by Opening Foundation of Agile and lntelligent Computing Key Laboratory of Sichuan Province, China (H23007). We gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] Shahab Mohaghegh, Reza Arefi, Sam Ameri, Khashayar Aminiand, Roy Nutter. Petroleum reservoir characterization with the aid of artificial neural networks [J]. Journal of Petroleum Science and Engineering, 1996, 16(4).

[2] Li Fengfeng, Guo Rui, Yu Yichang. Progress and prospect of sequence stratigraphic division method [J]. Geological Science and Technology Information, 2019, 38 (04): 215-224. DOI: 10.19509/j.cnki.dzkq.2019. 0422.

[3] Tang Xie, Tang Jiaqiong, Luo Yuhai, Deng Yuerui, Cui Jian. Evaluation method of horizontal well logging while drilling in thin carbonate reservoir [J]. Natural Gas Industry, 2013, 33 (09): 43-47.

[4] Han Yonggang, Feng Zhaojian, Luo Yongjun, Zhang Hui, Li Ping. Application of quantitative fluorescence logging technology in hydrocarbon reservoir classification [J]. Natural Gas Industry, 2007 (11): 24-26+131.

[5] He Ye, Zhang Hanbing, Zheng Ru, Cui Huan, Niu Wei, Chen Meijun. Shale gas horizontal well drilling analysis and reservoir evaluation parameter calculation based on element logging [J]. Natural Gas Industry, 2021, 41 (S1): 110-117.

[6] Pedram Masoudi,Bita Arbab,Hossein Mohammadrezaei. Net pay determination by artificial neural network: Case study on Iranian offshore oil fields [J]. Journal of Petroleum Science and Engineering, 2014, 123.

[7] Ahmed Ali Zerrouki,Tahar Aïfa,Kamel Baddari. Prediction of natural fracture porosity from well log data by means of fuzzy ranking and an artificial neural network in Hassi Messaoud oil field, Algeria [J]. Journal of Petroleum Science and Engineering, 2014, 115.

[8] Baijie Wang,Xin Wang,Zhangxin Chen. A hybrid framework for reservoir characterization using fuzzy ranking and an artificial neural network [J]. Computers and Geosciences, 2013, 57.

[9] Réda Samy Zazoun. Fracture density estimation from core and conventional well logs data using artificial neural networks: The Cambro-Ordovician reservoir of Mesdar oil field, Algeria [J]. Journal of African Earth Sciences, 2013, 83.

[10] Yunxin Xie,Chenyang Zhu,Wen Zhou,Zhongdong Li,Xuan Liu,Mei Tu. Evaluation of machine learning methods for stratum lithology identification: A comparison of tuning processes and model performances[J]. Journal of Petroleum Science and Engineering, 2018, 160.

[11] Zhang H , Chen Q , Ni P , et al. Study on the intelligent identification method of stratum lithology by element and gamma spectrum[J]. Neural Computing and Applications, 2021:1-9.

[12] Zhou Jinhui, Yan Taining, Tu Houze. Application of artificial neural network method to identify drilled strata [J]. Geoscience, 2000 (06): 642-646.

[13] Xia Hongquan, Chen Ping, Shi Xiaobing, Zhang Xianhui, Fan Xiangyu. Real-time identification method of stratum lithology based on drilling data [J]. Journal of Petroleum, 2004 (02): 51-54.

[14] Yang Sitong, Sun Jianmeng, Ma Jianhai, Huan Guanghui. Oil and gas identification method for logging data of low porosity and low permeability reservoirs [J]. Petroleum and Natural Gas Geology, 2007 (03): 407-412.2.

[15] M. A. Sebtosheikh,R. Motafakkerfard,M. A. Riahi,S. Moradi,N. Sabety. Support vector machine method, a new technique for lithology prediction in an Iranian heterogeneous carbonate reservoir using petrophysical well logs [J]. Carbonates and Evaporites, 2015, 30(1).

[16] Shaoqun Dong,Zhizhang Wang,Lianbo Zeng. Lithology identification using kernel Fisher discriminant analysis with well logs [J]. Journal of Petroleum Science and Engineering, 2016, 143.

[17] Fengqi Tan,Gang Luo,Duojun Wang,Yangkang Chen. Evaluation of complex petroleum reservoirs based on data mining methods[J]. Computational Geosciences,2017,21(1).

[18] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 2016: 785–794.

[19] Vikrant A. Dev,Mario R. Eden. Stratum lithology classification using scalable gradient boosted decision trees [J]. Computers and Chemical Engineering, 2019, 128.

[20] Zhixue Sun,Baosheng Jiang,Xiangling Li,Jikang Li,Kang Xiao. A Data-Driven Approach for Lithology Identification Based on Parameter-Optimized Ensemble Learning [J]. Energies, 2020, 13(15).

[21] Gu Yufeng, Zhang Daoyong, Bao Zhidong, Guo Haixiao, Zhou Liming, Ren Jihong. Using GS-LightGBM machine learning model to identify the lithology of tight sandstone stratum [J]. Geological Science and Technology Bulletin, 2021, 40 (04): 224-234. DOI: 10.19509/j.cnki.dzkq.2021.0416.

[22] Qi M. LightGBM: A Highly Efficient Gradient Boosting Decision Tree[C]// Neural Information Processing Systems. Curran Associates Inc. 2017.

[23] Wang Heng, Jiang Yanan, Zhang Xin, Zhong Hongru, Chen Qingxuan, Gao Shichen. Lithologic identification method based on gradient lifting algorithm [J]. Journal of Jilin University (Earth Science Edition), 2021, 51 (03): 940-950. DOI: 10.13278/j.cnki.jjuese.20200081.

[24] Ma Xiaojun, Sha Jinglan, Niu Xueqi. Design and application of P2P project credit rating model based on LightGBM algorithm [J]. Quantitative Economic and Technological Economic Research, 2018, 35 (05): 144-160. DOI: 10.13653/j.cnki. jqte.20180503.001.

[25] Yang Liu,Zhi-Ping Fan,Tian-Hui You,Wei-Yu Zhang. Large group decision-making (LGDM) with the participators from multiple subgroups of stakeholders: A method considering both the collective evaluation and the fairness of the alternative[J]. Computers & Industrial Engineering,2018,122.

[26] Yaguang Kong,Xuyang Tao,Zhangpin Chen. Sound field measurement and evaluation research for radiated acoustic fields in amplitude-variable sonochemical systems[J]. Measurement and Control, 2019,52(9-10).

[27] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms[J]. Pattern Recognition, 1997,30(7).

[28] Liu Saike, He Xiaoqun, Xia Liyu. Discussion on the effectiveness of model evaluation indicators under unbalanced data [J]. Statistics and Decision, 2022, 38 (19): 5-9. DOI: 10.13546/j.cnki.tjyjc.2022.19.001.

[29] Jolliffe Ian T,Cadima Jorge. Principal component analysis: a review and recent developments.[J]. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 2016, 374(2065).

[30] Yu Xiaofen, Fu Dai. Summary of Multi-index Comprehensive Evaluation Methods [J]. Statistics and Decision, 2004 (11): 119-121.

[31] Michele Biasutti,Sara Frate. A validity and reliability study of the Attitudes toward Sustainable Development scale[J]. Environmental Education Research,2016,23(2).