Petroleum Science 21 (2024) 1135–1148

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

A real-time intelligent lithology identification method based on a dynamic felling strategy weighted random forest algorithm

Tie Yan [a, b], Rui Xu [a, *], Shi-Hui Sun [a], Zhao-Kai Hou [a], Jin-Yu Feng [a]

[a] School of Petroleum Engineering, Northeast Petroleum University, Daqing, 163318, Heilongjiang, China
[b] Sanya Offshore Oil & Gas Research Institute, Northeast Petroleum University, Sanya, 572025, Hainan, China

ARTICLE INFO

Article history:
Received 23 November 2022
Received in revised form
19 April 2023
Accepted 12 September 2023
Available online 16 September 2023

Edited by Jia-Jia Fei

Keywords:
Intelligent drilling
Closed-loop drilling
Lithology identification
Random forest algorithm
Feature extraction

ABSTRACT

Real-time intelligent lithology identification while drilling is vital to realizing downhole closed-loop drilling. The complex and changeable geological environment in the drilling makes lithology identification face many challenges. This paper studies the problems of difficult feature information extraction, low precision of thin-layer identification and limited applicability of the model in intelligent lithologic identification. The author tries to improve the comprehensive performance of the lithology identification model from three aspects: data feature extraction, class balance, and model design. A new real-time intelligent lithology identification model of dynamic felling strategy weighted random forest algorithm (DFW-RF) is proposed. According to the feature selection results, gamma ray and 2 MHz phase resistivity are the logging while drilling (LWD) parameters that significantly influence lithology identification. The comprehensive performance of the DFW-RF lithology identification model has been verified in the application of 3 wells in different areas. By comparing the prediction results of five typical lithology identification algorithms, the DFW-RF model has a higher lithology identification accuracy rate and *F*1 score. This model improves the identification accuracy of thin-layer lithology and is effective and feasible in different geological environments. The DFW-RF model plays a truly efficient role in the real-time intelligent identification of lithologic information in closed-loop drilling and has greater applicability, which is worthy of being widely used in logging interpretation.

1. Introduction

Real-time intelligent lithology identification is the essential early work of downhole closed-loop intelligent steering drilling technology. It is used to sense the drilling formation environment to provide the basis for intelligent decision-making in the drilling operation. The lithology identification effect directly affects the intelligent development process of downhole closed-loop drilling technology (Zhao et al., 2021; Wu et al., 2022; Xie et al., 2022).

Traditional lithology identification methods mainly include the cross-plot method (Zhou et al., 2016; Baisakhi and Rima, 2018), the probability statistics method (Phillip et al., 2017), and the cluster analysis method (Wang et al., 2018; Amjad and Chen, 2020). Although these lithology identification methods are simple in principle and easy to operate, the judgment results have intense subjectivity and time lag, which is challenging to meet the demand for real-time intelligent identification of closed-loop drilling. In recent years, the rapid development of artificial intelligence technology has provided a new technical approach to solving the problem of real-time intelligent lithology identification. Many scholars have applied various intelligent information processing technologies to lithology identification, including but not limited to Support Vector Machines (Al-Mudhafar, 2017), Naïve Bayes (Rosid et al., 2019), Random Forests (Zou et al., 2021) and Neural Networks (Xu M.H. et al., 2022). These methods are now becoming the primary method to obtain logging lithology information quickly and accurately. Jorge et al. (2018) used a Support Vector Machine to identify lithology types automatically. Karimzadeh and Tangestani (2021) combined the Principal Component Analysis and the Support Vector Machine to realize the effective identification of lithology. Liang et al. (2022) established a Support Vector Machine model based on simulated annealing optimization to realize rapid and intelligent lithology identification. Dong et al. (2022) established a logging discrimination model for carbonate based on the

* Corresponding author.
E-mail address: sygcxytyb@163.com (R. Xu).

https://doi.org/10.1016/j.petsci.2023.09.011
1995-8226/

Fisher discriminant method. Ren et al. (2022) combined the K-means algorithm, Fuzzy theory and Decision Tree algorithm and proposed a Fuzzy Decision Tree model for lithology identification. Abdelhakim et al. (2020) used a Random Forest algorithm to classify lithology automatically and analyzed the importance of different variables in lithology classification. Stephen et al. (2019) proposed a complex lithology identification method based on a rough set-Random Forest algorithm, which has higher discriminative stability than the traditional Random Forest method. Wu et al. (2021) constructed a lithology identification method based on a Long Short-Term Memory (LSTM) recurrent neural network that can extract and learn the characteristics of lithologic depositional sequences. Miao et al. (2021) considered the spatial interdependence between sediments and the spatial coupling between logging data and proposed a spatial deep recurrent neural network lithology classifier. Zeng et al. (2020) analyzed the correlation between different logging series and the actual depth accumulation effect. They established a quantitative lithology identification method based on attention-based bidirectional gated recurrent unit neural networks. Li et al. (2021) used the Extreme Learning Machine (ELM) of semi-supervised learning mode to classify lithology. Alzubaidi et al. (2021), Polat et al. (2021) and Becerra et al. (2022), used the core image as input, and the Convolution Neural Network (CNN) was used to extract the image features of rock slices for lithology classification. Merembayev et al. (2021), Sun et al. (2021), Zhang et al. (2022) and Kumar et al. (2022) compared the performance of machine learning methods such as Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting Decision Tree and Artificial Neural Network in formation lithology identification. Reviewing the previous research results shows that the research on applying artificial intelligence technology to lithology identification has been profound and extensive. However, there are still some areas for improvement in the current research that is difficult to overcome. For example, the Support Vector Machine method's classification effect is too dependent on kernel function and penalty parameters selection. The Random Forest method has a multi-value bias. The neural network method is slow in convergence and has poor visualization and interpretability.

From the current research, although many artificial intelligence algorithms can be applied to lithology identification, there is still much room for improvement in the applicability and comprehensive performance of the model. The adaptability between datasets and algorithms is critical of lithology intelligent identification technology. Further research needs to use the massive logging data more effectively and construct a lithology identification method with a high recognition rate and computational efficiency that can be applied to various complex geological environments. This paper examines the problems of difficulty extracting feature information, the low recognition accuracy of thin-layer, and the limited applicability of the model in lithology identification. A new lithology identification model of dynamic felling strategy weighted random forest algorithm is established to realize real-time intelligent and accurate lithology identification while drilling.

2. Data preprocessing method

2.1. Feature selection method

Before performing the lithology identification task, using feature selection technology to select LWD parameters sensitive to lithology information can reduce the number of features, decrease the model's learning difficulty and computational cost, and improve the model's generalization ability. This paper uses the minimum Redundancy Maximum Relevance algorithm (mRMR) (Peng et al., 2005) to mine the correlation between LWD information and lithology categories. Taking mutual information (Hjelm et al., 2018) as the calculation criterion, the LWD parameter set with the most significant correlation with the lithology category and the least redundancy among LWD parameters is selected as the lithology identification feature. The principle of the mRMR algorithm is as follows:

$$\max_{f_r \in F-F_{m-1}} \left[I(f_r, c) - \frac{1}{m-1} \sum_{f_o \in F_{m-1}} I(f_r, f_o) \right] \tag{1}$$

where F represents the original LWD parameter set, m denotes the number of LWD parameters in the LWD parameter set, I (f_r, f_o) indicates the mutual information between the LWD parameter f_r and the LWD parameter f_o, c is the lithology category, and I (f_r, c) refers to the mutual information between the LWD parameter f_r and the lithology category c.

Equation (2) is the mutual information between LWD parameters. Eq. (3) is the mutual information between LWD parameters and lithology categories.

$$I(f_r, f_o) = \iint p(f_r, f_o) \log \frac{p(f_r, f_o)}{p(f_r)p(f_o)} \mathrm{d}f_r \mathrm{d}f_o \tag{2}$$

$$I(f_r, c) = \iint p(f_r, c) \log \frac{p(f_r, c)}{p(f_r)p(c)} \mathrm{d}f_r \mathrm{d}c \tag{3}$$

where p (f_r) represents the probability density of the LWD parameter f_r, p (f_o) denotes the probability density of the LWD parameter f_o, p (f_r, f_o) indicates the combined probability density of the LWD parameter f_r and the LWD parameter f_o, $p(c)$ is the probability density of the lithology c, and p (f_r, c) refers to the combined probability density of the LWD parameter f_r and the lithology c.

2.2. Thin-layer sample data processing method

In the process of lithology identification, the geological environment is complex and changeable, and many thin-layer are difficult to identify. Because the number of samples in the thin-layer is far less than that in other formations, the characteristic lithology information in the thin-layer is scarce. The lithology identification model does not learn enough about the thin-layer, which leads to the accuracy of model identification always being biased towards other rock formations with sufficient samples, and the identification effect of the thin-layer is unsatisfactory. The Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) is used to process thin-layer sample data to solve the problem of scarcity of thin-layer feature information. According to the similarity of thin-layer sample data in the feature space, the number of thin-layer samples is expanded by linear interpolation. The calculation steps of the SMOTE oversampling method are as follows:

Assume that the original sample dataset is T, in which the non-thin-layer sample dataset is T_{majority}, and the number of samples is M; the thin-layer sample dataset is T_{minority}, and the number of samples is N; the sampling ratio is P (P is a positive integer not less than 1).

Step 1. For each thin-layer sample x_i, calculate the Euclidean distance from x_i to all the samples in the thin-layer sample dataset, and obtain the k near neighbor samples of sample x_i ($k > P$), which are denoted as y_j (j = 1,2, …,k).

Step 2. Select P nearest neighbor samples from the k near neighbor samples of x_i, and then calculate the newly generated thin-layer sample x_{new} according to Eq. (4), where rand(0,1) represents a random number between 0 and 1.

$$x_{\text{new}} = x_i + \text{rand}(0,1) \times \left(y_j - x_i\right) \ (j = 1, \cdots, P) \tag{4}$$

Step 3. Repeat step 2, and then add all the newly generated thin-layer samples x_{new} to the T_{minority} to obtain the thin-layer sample dataset after SMOTE oversampling.

3. Lithology identification algorithm

3.1. The dynamic felling strategy weighted random forest algorithm

This paper aims to develop a lithology identification model for downhole closed-loop intelligent drilling. The model's input data is real-time LWD data, and the output result is lithologic type. In addition to high-quality data preprocessing techniques, the selection of identification algorithms is crucial to the accuracy and speed of lithology identification. Random Forest (RF) (Breiman, 2001) is an ensemble classification algorithm based on the decision tree. The algorithm has good tolerance to noise and outliers, has high prediction accuracy, and is suitable for lithology identification.

The generalization error of the RF algorithm is expressed as Eq. (5). The correlation and classification intensity of the decision tree in the RF algorithm are the main factors that affect the lithology identification effect of the model. The upper bound of the generalization error of the RF algorithm is positively correlated with the correlation of any two decision trees in the forest and negatively correlated with the classification intensity of each decision tree. The stronger the classification strength and the smaller the correlation between decision trees, the smaller the upper limit of the generalization error of the model and the lower the error rate of the RF model.

$$\text{Mar}(Q, V_T) = \text{ave}(V_T) - E^* \le \frac{\overline{\rho}(1 - s^2)}{s^2} \tag{5}$$

where E^* denotes the generalization error of the RF algorithm, $\overline{\rho}$ refers to the average correlation of decision trees, and s indicates the overall classification strength of the decision tree.

To effectively reduce the upper limit of model generalization error and improve the accuracy and efficiency of model identification, this paper designs a dynamic felling strategy weighted random forest algorithm from the perspectives of reducing the correlation between decision trees and improving the influence of decision trees with good classification effect.

The dynamic felling strategy is adopted to reduce the correlation between decision trees. Calculate the correlation of the confusion matrix between two decision trees from Eq. (6). The two decision trees will be retained if the correlation is less than 70%. If the correlation is greater than 70%, the decision trees need to be cut down. According to the *AUC* values of the two decision trees, the decision tree with the lower *AUC* value should be cut down. The other decision tree is retained for the next iteration. Only the decision trees with weak correlation remain in the forest through the dynamic felling strategy. It can be seen from Eq. (5) that reducing the correlation between decision trees can improve the accuracy of the random forest model.

$$Sim(a, b) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|\boldsymbol{a}\| \times \|\boldsymbol{b}\|} \tag{6}$$

where Sim(a, b) represents the confusion matrix similarity between the decision tree a and the decision tree b, $\boldsymbol{a} \cdot \boldsymbol{b}$ indicates the dot product of the confusion matrix of the decision tree a and the confusion matrix of the decision tree b, $\|\boldsymbol{a}\|$ is the length of the confusion matrix of the decision tree a, $\|\boldsymbol{b}\|$ is the length of the confusion matrix of the decision tree b.

The *AUC* is an index to measure the performance of the classification learner, with values ranging from 0.5 to 1.0. A model with better overall performance has an *AUC* value close to 1. The *AUC* is the area under the receiver operating characteristics (ROC) curve that represents the trade-off between *TPR* and *FPR*, given by Eqs. (7) and (8). The lithology identification results can be divided into four cases: true positive (*TP*), false positive (*FP*), true negative (*TN*), and false negative (*FN*).

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

$$FPR = \frac{FP}{TN + FP} \tag{8}$$

The result of lithology identification by RF algorithm is equal weight voting for all decision tree identification results, which will lead to the decision tree with good lithology identification effect being unable to play a better role. In contrast, the decision tree with a poor lithology identification effect will hurt the identification results. Therefore, this paper evaluates the identification effect of each decision tree based on the out-of-bag data, *OOB*. According to the identification accuracy of *OOB*, the decision tree in the random forest is weighted to vote. By setting the weights, the voting power of the decision tree with a good identification effect is improved, and the influence of the decision tree with a poor impact on the lithology identification results is reduced. The weight calculation method is shown in Eq. (9):

$$\text{Mar}(Q, V_T) = \text{ave}(V_T) - w_i = \frac{OOB_{\text{correct}}(i)}{count(OOB)} \tag{9}$$

where w_i is the weight of decision tree i, $OOB_{\text{correct}}(i)$ indicates the correct number of samples predicted by the i decision tree in the *OOB*, and *count* (*OOB*) represents the overall sample size of the *OOB*.

3.2. Construction of lithologic identification model

According to the principle of the dynamic felling strategy weighted random forest algorithm, a real-time intelligent identification model of lithology based on closed-loop drilling (DFW-RF) is constructed, as shown in Fig. 1.

The implementation process of the DFW-RF lithology identification model is as follows.

Step 1: Preprocess the original LWD dataset. The mRMR algorithm performs feature selection, and the thin-layer data is processed by SMOTE sampling technology. After preprocessing, 70% of the data are randomly selected to form a training dataset, and the remaining data constitute a test dataset. The training dataset is used to learn the relationship between independent and dependent variables. The testing dataset evaluates the model's ability to predict unknown data points. The stratified random sampling method ensures that different lithology types are uniformly distributed in the training and testing datasets.
Step 2: The Bootstrap sampling technique extracts n training subset from the training dataset. Each training subset has the same capacity as the training dataset. Every time the data that is not extracted is recorded as out-of-bag data (*OOB*).
Step 3: n training subsets are generated into n decision trees, and the initial random forest is formed. Calculate each decision

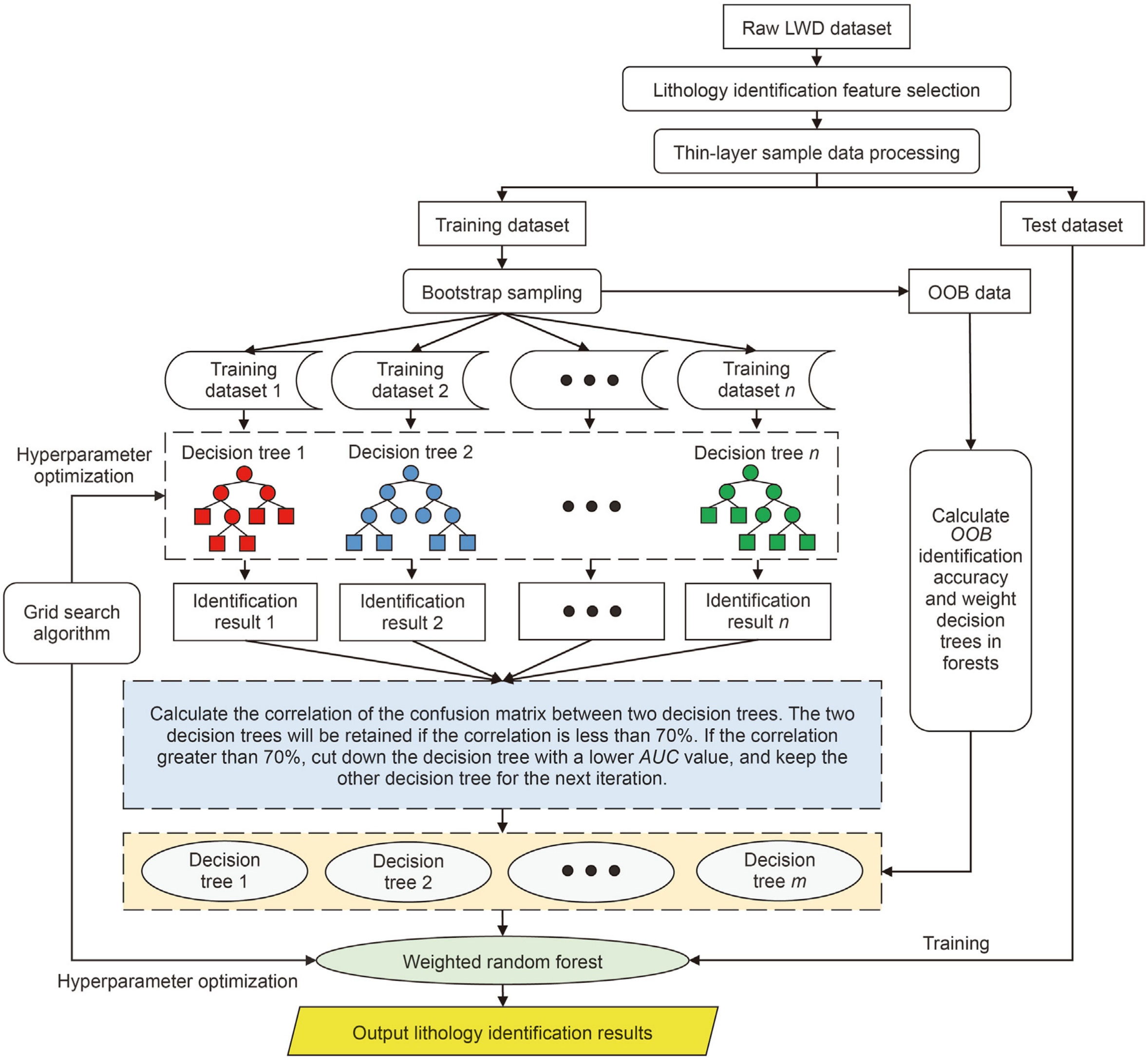

Raw LWD dataset
Lithology identification feature selection
Thin-layer sample data processing
Training dataset
Test dataset
Bootstrap sampling
OOB data
Training dataset 1
Training dataset 2
Training dataset n
Hyperparameter optimization
Decision tree 1
Decision tree 2
Decision tree n
Calculate OOB identification accuracy and weight decision trees in forests
Grid search algorithm
Identification result 1
Identification result 2
Identification result n
Calculate the correlation of the confusion matrix between two decision trees. The two decision trees will be retained if the correlation is less than 70%. If the correlation greater than 70%, cut down the decision tree with a lower AUC value, and keep the other decision tree for the next iteration.
Decision tree 1
Decision tree 2
Decision tree m
Training
Weighted random forest
Hyperparameter optimization
Output lithology identification results

Fig. 1. DFW-RF lithology identification model.

tree's confusion matrix and *AUC* value in the initial random forest.

Step 4: Calculate the correlation of the confusion matrix between two decision trees in the initial random forest. The two decision trees will be retained if the correlation is less than 70%. If the correlation is greater than 70%, cut down the decision tree with a lower *AUC* value, and keep the other decision tree for the next iteration. After the calculation, the remaining decision trees will form a new random forest.

Step 5: The weight of each decision tree in the new random forest is calculated according to the test accuracy of the out-of-bag data. Input the test dataset, and vote the classification result of the decision tree by weight to obtain the lithology classification result.

The performance of the model is evaluated using the accuracy rate. The accuracy rate, defined by Eq. (10), measures the percentage of correctly identified samples.

$$\text{Accuracy rate} = \frac{TP + TN}{TP + FP + FN + TN} \tag{10}$$

4. Model training

4.1. Training data analysis

This paper selected 6 wells in area A with typical structural characteristics and sedimentary environment in X oilfield as the model training research objects. Each well has complete LWD data and named lithology information. LWD parameters obtained from the field include gamma ray (GR), caliper (CAL), 2 MHz phase resistivity, 400 kHz attenuation resistivity, compensated neutron log (CNL), density (DEN), and acoustic (AC). The depth range of the logging section in the experiment is 4204.43~4320.81 m. According to the coring and logging data, the primary lithologies drilled in these 6 wells within the depth range of the selected logging section are sandstone, dacite, limestone, and mudstone. The upper part is interbedded with sandstone and mudstone of equal thickness; the middle is mainly composed of dacite with thin limestone layers; the lower part is interbedded with unequal thickness of mudstone and sandstone. The LWD data of 6 wells were sampled at equal depth intervals of 0.125 m in the selected logging section 931 sample points were obtained from each well, with a total of 5586 sample points. The samples of sandstone, dacite, limestone, and mudstone accounted for 31.59%, 31.08%, 4.00%, and 33.33%, respectively, in the total sample points. Some LWD data are shown in Table 1.

Boonen et al. (2005), Sun et al. (2019) and Xu T. et al. (2022) found that the variation laws of formation petrophysical parameters obtained by logging while drilling and wireline logging are the same. This information can be effectively used for geological evaluation. The wireline logging is measured after the casing is running. Due to the time problem, the logging value measured by wireline logging will be affected by mud invasion. Compared with wireline logging, LWD measures simultaneously during the drilling and records the geological information when the bit is drilled through the formation. At this time, the borehole has not collapsed obviously, and the invasion of mud into the formation is shallow or even negligible, so the LWD data can better reflect the information of the original formation (Li et al., 2023; Han et al., 2023; Sui et al., 2022). In addition to real-time and accuracy, LWD technology can save much drilling time and reduce drilling costs. Reasonable use of LWD data helps realize real-time intelligent identification of lithology information with high efficiency and low cost.

Based on the LWD data of the 6 wells in Block A of X oilfield, box plot analysis was performed on sandstone, dacite, limestone, and mudstone samples. As shown in Fig. 2, due to the different physical and chemical properties of various lithologies, there are specific differences in the numerical response laws of different LWD parameters. In GR, the higher the intensity of gamma rays emitted by the decay process of radionuclides in the rock, the greater the GR value. The average GR value of dacite is the highest, followed by mudstone, sandstone, and limestone, with the lowest average GR. Observing the 2 MHz phase resistivity and the 400 kHz attenuation resistivity, it can be seen that the electromagnetic wave resistivity while drilling from limestone-dacite-sandstone-mudstone shows a changing law from large to small. In CNL and DEN, mudstone has the highest CNL but low DEN and limestone have the lowest CNL but high DEN. In AC, mudstone shows a high AC value, and limestone shows a low AC value. The differences in the LWD data statistics of various lithologies indicate that different lithologies are separable based on LWD data. However, only rough qualitative lithology identification results can be obtained only through statistical analysis of data, and lithology cannot be accurately and quantitatively identified. This kind of method cannot real-time intelligently identify lithology during drilling, so it is vital to study further the lithology identification method that can be used while drilling.

4.2. Determine lithologic identification characteristic parameters

This paper uses the mRMR feature selection algorithm to select LWD parameters for lithology identification of 6 wells in Block A of X oilfield. According to Eqs. (2) and (3), the mutual information between LWD parameters and lithology categories are calculated. Then bring the calculated mutual information into Eq. (1) to get the mRMR score of each LWD parameter, as shown in Fig. 3.

As displayed in Fig. 3, GR, 2 MHz phase resistivity, 400 kHz attenuation resistivity, and CNL are more sensitive to lithology of block A of X oilfield. GR reflects rock mineral skeleton characteristics and sedimentary information, with an mRMR score of 0.3949, contributing the most to lithology identification. The 2 MHz phase resistivity and 400 kHz attenuation resistivity reflect the comprehensive electrical characteristics of rocks and fluids at different detection distances during drilling, and the mRMR scores are 0.2234 and 0.1093, respectively. Attenuation resistivity and phase resistivity can meet the measurement accuracy when the resistivity is low. Still, when the formation resistivity is greater than 20 Ω•m, the measurement accuracy of attenuation resistivity is extremely low. At this time, the 2 MHz phase resistivity measurement is more accurate. The CNL reflects the ability of the formation to store fluids, and the mRMR score is 0.1378. Compared with the above LWD parameters, the CAL, DEN, and AC have lower contribution rates to lithology identification.

The mRMR algorithm determines the correlation between LWD parameters and lithology categories, but the number of characteristic parameters is not determined. Too few characteristic parameters cannot fully obtain lithologic characteristics, and too many characteristic parameters will increase the burden of model training. Therefore, the number of characteristic parameters needs to be determined through experiments. Firstly, according to the mRMR score, the LWD parameters are sorted, and the LWD parameters sensitive to lithology are preferentially selected. Then, lithology identification is carried out through different numbers of characteristic parameters. Finally, the optimal number of characteristic parameters is determined according to the lithology identification results of the DFW-RF model under the different number of characteristic parameters. The lithology identification result of the DFW-RF model with the different number of characteristic parameters is displayed in Fig. 4.

According to the experimental results in Fig. 4, it can be seen that feature selection has a great influence on the identification effect of the lithologic identification model. With the increase in the number of characteristic parameters, the accuracy of lithology identification of the DFW-RF model is constantly improved, but it also takes more training time. With the increase of characteristics, the information on lithologic characteristics is more abundant, which is beneficial for distinguishing different lithology types. However, with the increase in the number of characteristics, the calculation cost of the model increases exponentially, and the training time is significantly prolonged. Comparing the lithology identification results with the different number of characteristic parameters, when the number of characteristic parameters is less than four, the lithology identification accuracy rate of the training dataset and the testing dataset is lower than 76%. When the number of characteristic parameters is greater than four, the accuracy rate of lithology identification in the training dataset is 98.02%~98.13%, and that in the testing dataset is 96.12%~96.39%. However, the training time of the model is longer. The training time

Table 1
Partial LWD data.

Sample point number	GR, API	CAL, cm	2 MHz phase resistivity, Ω·m	400 kHz attenuation resistivity, Ω·m	CNL, %	DEN, g/cm^3	AC, μs/m	Lithology category
1	48.49	25.49	7.66	6.92	21.50	2.31	201.03	Sandstone
2	50.52	25.10	7.04	6.50	19.79	2.28	210.20	Sandstone
3	49.32	25.14	7.61	6.67	20.55	2.25	210.19	Sandstone
4	49.38	25.27	8.01	6.93	21.01	2.25	211.30	Sandstone
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
5583	88.53	27.41	1.54	1.52	27.46	2.11	276.82	Mudstone
5584	88.67	28.22	1.43	1.36	27.81	2.04	260.04	Mudstone
5585	86.58	28.27	1.60	1.50	26.40	2.17	249.45	Mudstone
5586	87.44	26.18	1.04	1.19	27.71	2.17	347.77	Mudstone

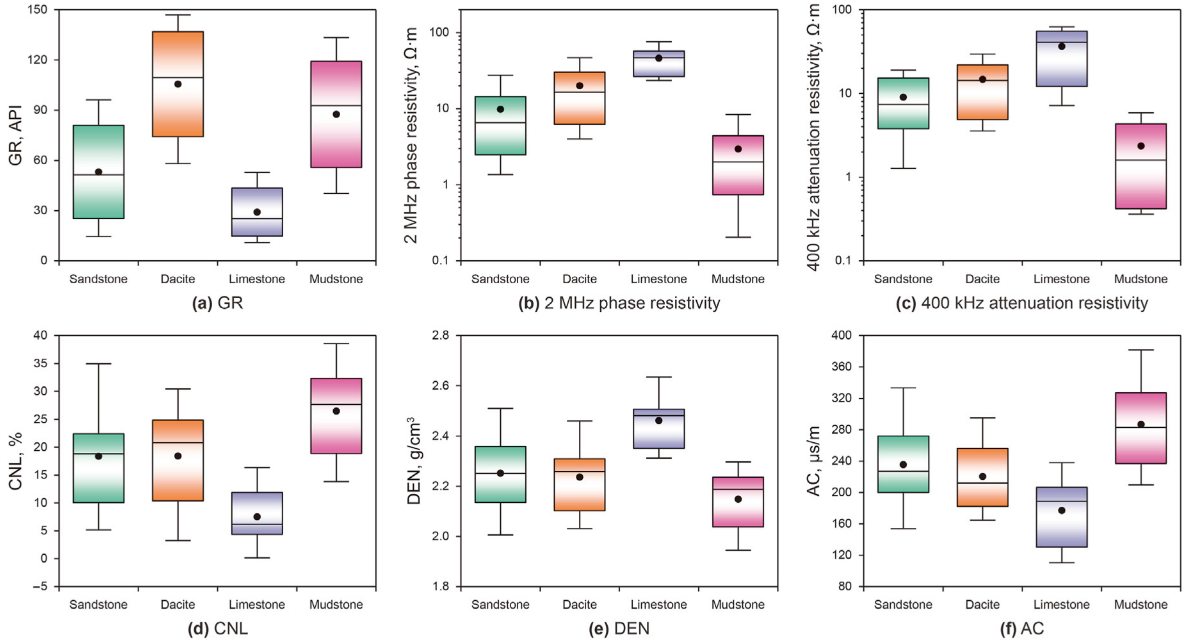


Fig. 2. Box plot of LWD parameters.

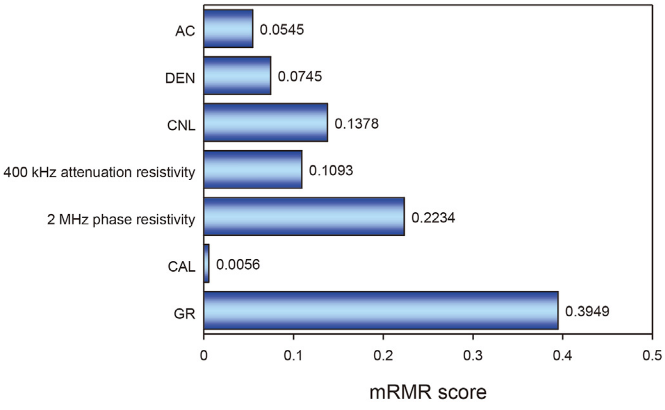
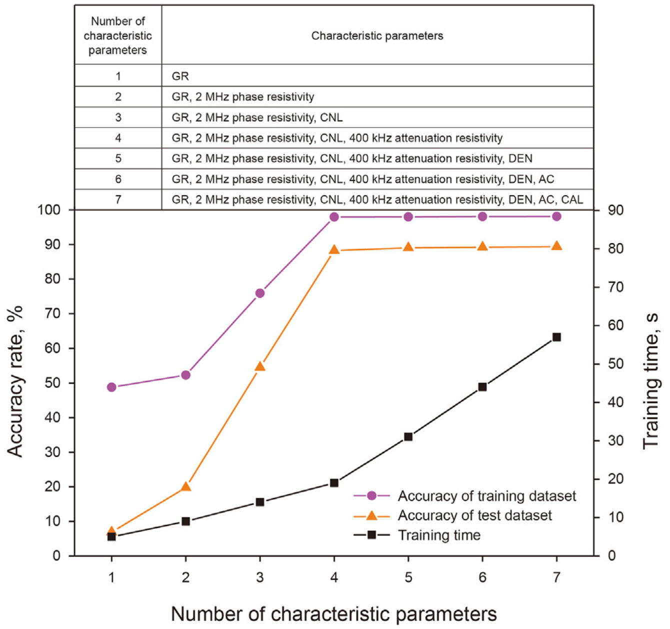


Fig. 3. LWD parameter mRMR score.



Number of characteristic parameters	Characteristic parameters
1	GR
2	GR, 2 MHz phase resistivity
3	GR, 2 MHz phase resistivity, CNL
4	GR, 2 MHz phase resistivity, CNL, 400 kHz attenuation resistivity
5	GR, 2 MHz phase resistivity, CNL, 400 kHz attenuation resistivity, DEN
6	GR, 2 MHz phase resistivity, CNL, 400 kHz attenuation resistivity, DEN, AC
7	GR, 2 MHz phase resistivity, CNL, 400 kHz attenuation resistivity, DEN, AC, CAL

Fig. 4. The lithology identification result of DFW-RF model with different number of characteristic parameters.

of the model with five characteristic parameters is 12 s longer than that with four characteristic parameters. When the number of parameters is four, the comprehensive performance of the DFW-RF lithology identification model is the best. The accuracy rate of lithology identification in the training dataset is 97.99% that in the testing dataset is 95.54%, and the training time is 19 s. Considering the model's identification accuracy and training time, four LWD parameters, GR, 2 MHz phase resistivity, 400 kHz attenuation resistivity and CNL, are selected as the characteristic parameters for real-time intelligent identification lithology of closed-loop drilling.

4.3. A sampling of thin-layer sample data

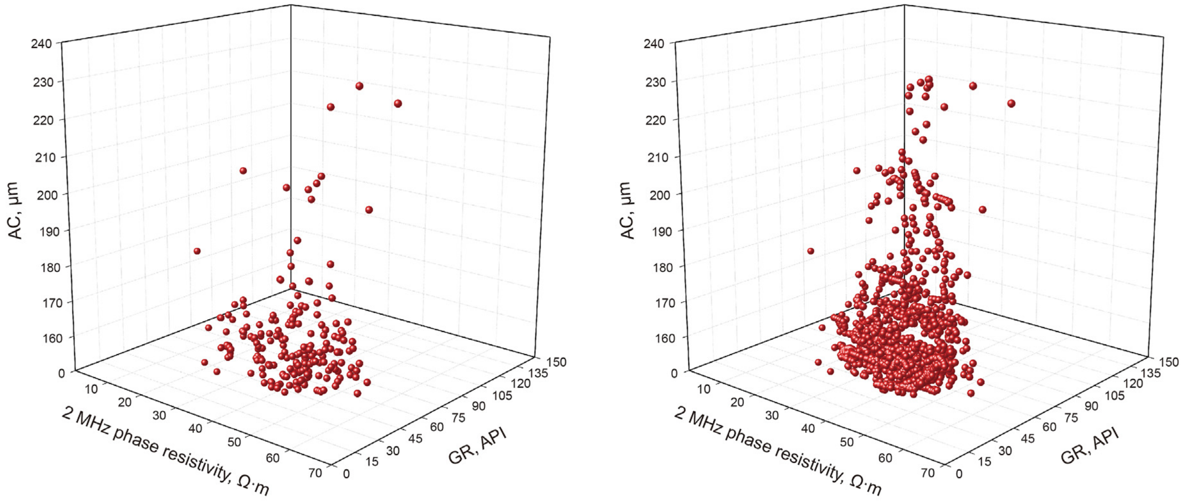
The single layer thickness of the sandstone layer, dacite layer and mudstone layer in the formation drilled in 6 wells are all greater than 4.00 m, and the single layer thickness of the limestone layer is only 0.30~0.72 m. The number of limestone samples only accounts for 4.00% of the total samples. Therefore, it is necessary to process the sample data of the limestone thin-layer by SOMTE sampling technology. Table 2 shows the sample capacities of different lithologies before and after SMOTE sampling.

After sampling, the number of four types of lithologic samples is balanced. The number of sandstone, dacite, limestone, and mudstone samples accounted for 24.92%, 24.51%, 24.29%, and 26.28% of the total sample points, respectively. The limestone thin-layer sample data was expanded from 223 to 1721 after sampling, and the sample size was boosted by 6.72 times compared with the original data. Fig. 5 presents the limestone sample capacity before and after sampling.

SMOTE sampling increases the sample number of the limestone thin-layer and balances the number of samples of different types of lithology in the dataset. However, whether SMOTE sampling is

Table 2
The sample capacity of different lithologies before and after SMOTE sampling.

Lithology category	Number of samples before sampling	Number of samples after sampling
Sandstone	1765	1765
Dacite	1736	1736
Limestone	223	1721
Mudstone	1862	1862
Total	5586	7084



(a) Limestone sample before SMOTE sampling **(b)** Limestone sample after SMOTE sampling

Fig. 5. Limestone sample capacity before and after sampling.

beneficial to lithology identification needs experimental demonstration. Fig. 6 shows the lithology identification results of the DFW-RF model before and after SMOTE sampling.

As shown in Fig. 6, the identification accuracy rate of limestone before SMOTE sampling is 58.32%. After SMOTE sampling, the identification accuracy rate of the limestone is 96.53%, which is increased by 38.21%. Before and after SMOTE sampling, the identification accuracy rate of sandstone, dacite and mudstone did not change significantly. The DFW-RF model lithology identification results show that the identification accuracy rate of the training and test datasets has improved by 8.87% and 11.18%, respectively, after SMOTE sampling. According to the identification results of four lithologies, the DFW-RF model performance in this comparative experiment is mainly due to the improvement of limestone thin-layer identification accuracy. With the increase in the number of samples after SMOTE sampling, the training time of the DFW-RF model increased by 3 s. The experiment proves that SMOTE sampling technology can effectively solve the problem that the model learns insufficiently about thin-layers due to the scarcity of thin-layer samples. After sampling, the identification accuracy of thin-layers is significantly improved, and the overall lithology identification effect of the DFW-RF model is also better.

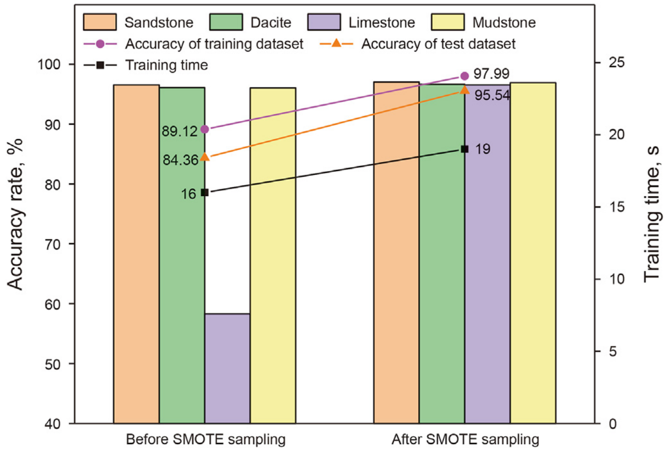


Fig. 6. Lithologic identification results of DFW-RF model before and after SMOTE sampling.

4.4. Analysis of model training results

Classification and Regression Tree (CART) is adopted for all decision trees during DFW-RF model training. The classification criterion is to minimize the Gini index. The grid search algorithm determines the optimal model parameters (Ghawi and Pfeffer, 2019; Yao et al., 2021). Cross entropy is the loss function to evaluate the model's performance. Training dataset training and test dataset testing are conducted simultaneously. Table 3 shows the setting values of relevant parameters of the DFW-RF model.

Table 3
Parameter setting values of DFW-RF model.

Model	Model parameter

DFW-RF	max_features = 4
	n_estimators = 58
	max_depth = 100
	min_samples_split = 2
	min_samples_leaf = 1
	max_leaf_nodes = None
	min_impurity_split = 0

The training results of DFW-RF model are shown in Fig. 7. The changing trend of loss value and accuracy rate of the training and testing datasets are synchronized. After 350 iterations, the loss value and accuracy rate of the DFW-RF lithology identification model on the training dataset and test dataset tend to converge. After training, the final lithology identification accuracy of the training and test datasets is higher than 95%, which takes 19 s. Compared with the test dataset, the training dataset contains more samples, and after 200 times of training, the loss value of the training dataset is less than 0.1. The final lithology identification accuracy of DFW-RF model training dataset is 97.99%, 2.45% higher than that of test dataset, and the final lithology identification accuracy of test dataset is 95.54%.

In order to verify that the Dynamic Felling Strategy Weighted Random Forest algorithm can effectively improve the performance of the lithologic identification model, this paper compares the lithologic identification results of the Random Forest model (RF) and the Dynamic Felling Strategy Weighted Random Forest model (DFW-RF). In the contrast experiment, the selection of characteristic parameters, the processing of sample data and the setting of model parameters of the RF model are consistent with those of the DFW-RF model.

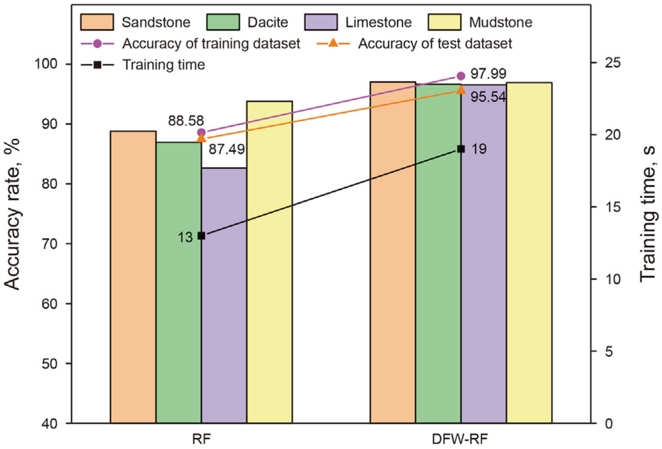


Fig. 8. Lithology identification results of the RF model and the DFW-RF model.

Fig. 8 indicates the lithology identification results of the RF and DFW-RF models. The accuracy rate of lithology identification of the DFW-RF model is higher than that of the RF model in both training and testing datasets. The DFW-RF model not only improves the identification accuracy rate of each lithology but also makes all kinds of lithology identification effects more stable. The identification accuracy rate of the four lithologies is more than 96%, and the average identification accuracy rate is 96.77%. In comparison, the identification effect of the RF model on the four lithologies is quite different. The identification effect of mudstone is outstanding, and the accuracy rate is 93.77%. The identification effect of limestone is poor, and the accuracy rate is only 82.64%. The average identification accuracy rate of the four lithologies under the RF model is 88.04%. The experimental results certificate that the Dynamic Felling Strategy Weighted Random Forest algorithm successfully improves the accuracy of lithology identification by reducing the correlation between decision trees and improving the influence of decision trees with good classification effects. Although cutting down trees and weighting resulted in the training time of the DFW-RF model being 6 s longer than that of the RF model, the sacrifice of a few seconds of training time has exchanged for a significant improvement in the accuracy and stability of lithology identification. In general, the lithology identification performance of the DWF-RF model is better than that of the RF model.

Compared with the RF model, the DFW-RF model proposed in this paper improves the RF model through dynamic felling and weighted voting. In order to verify that dynamic felling and weighted voting positively improve the accuracy of model lithology identification, the author designed a comparative experiment. The feasibility and effectiveness of the dynamic felling strategy weighted random forest algorithm are verified by comparing the lithology identification results of the RF model after dynamic felling and weighted voting. The lithology identification results of the RF model after dynamic felling and the RF model after weighted voting are shown in Fig. 9.

Comparing the experimental results of Figs. 8 and 9, it can be seen that compared with the RF model, the lithology identification accuracy rate of the RF model after dynamic felling or weighted voting is significantly improved. The accuracy rate of lithology identification in training set and testing set is greater than 91.00%. The average identification accuracy rate of four kinds of lithology in the RF model after dynamic felling is 4.64% higher than that in the RF model, and the average identification accuracy rate of four kinds of lithology in the RF model after weighted voting is 4.10% higher than that in the RF model. The results show that both dynamic felling and weighted voting have a positive effect on improving the lithology identification effect of the RF model. By comparing the

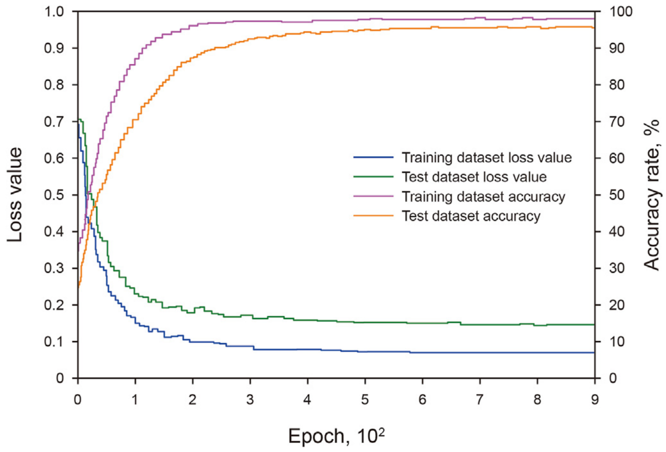


Fig. 7. DFW-RF lithology identification model training results.

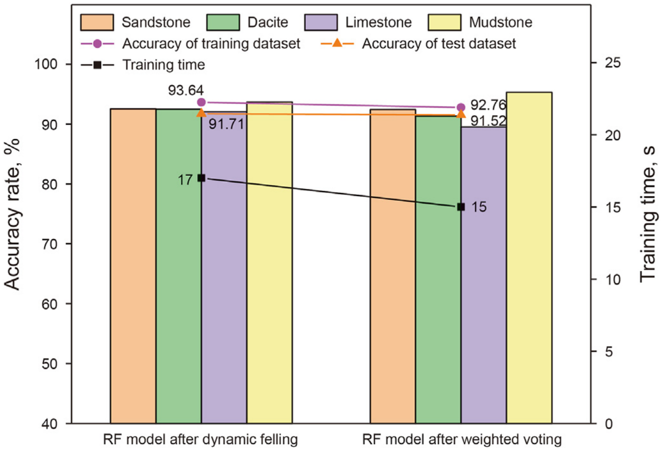


Fig. 9. Lithology identification results of the RF model after dynamic felling and the RF model after weighted voting.

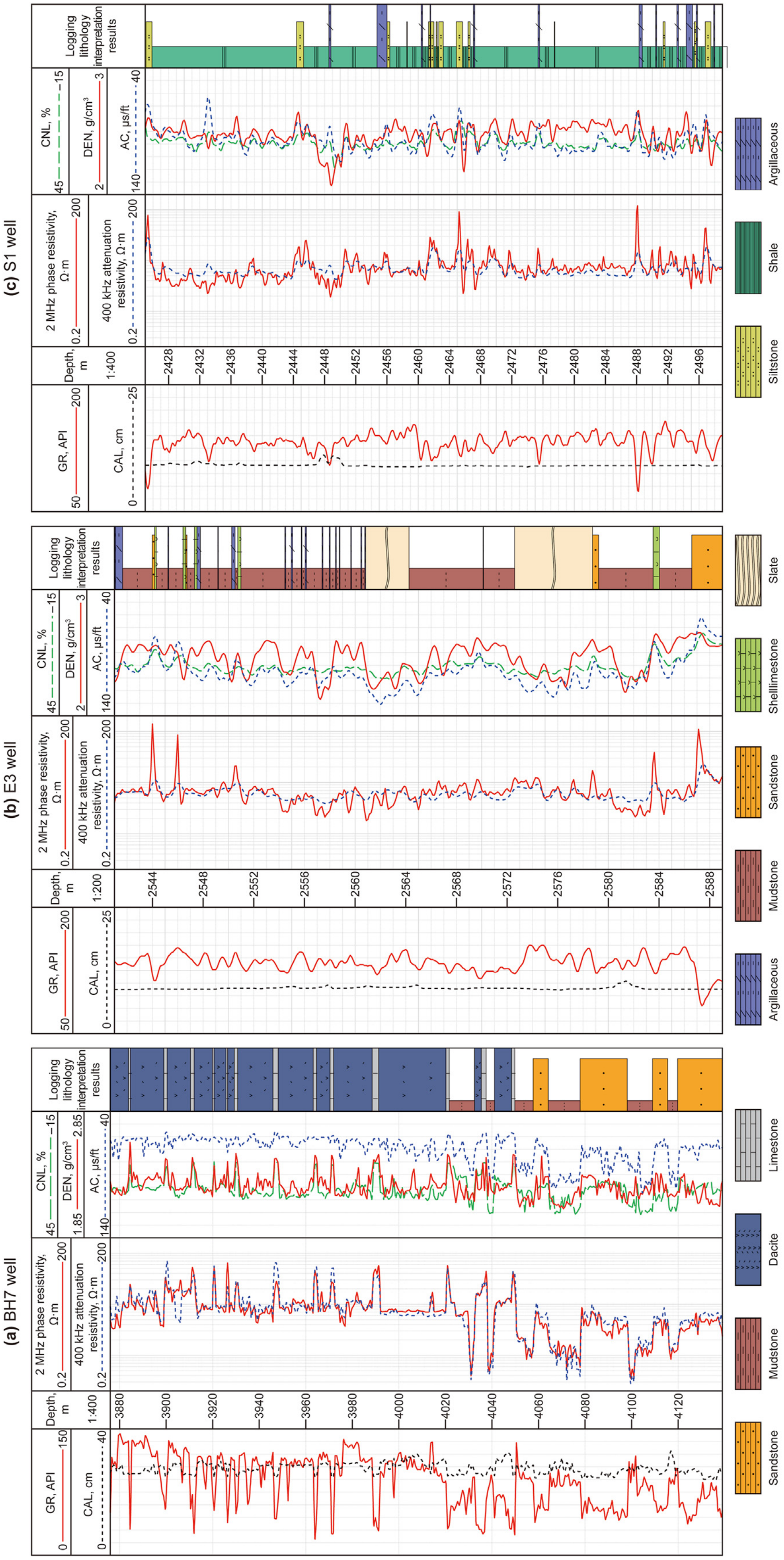


Fig. 10. The LWD data and logging lithology interpretation results of Well BH7, Well E3 and Well S1.

Table 4
The characteristic parameter selection and data sampling results of BH7 well, E3 well and S1 well.

Well	Lithology identification characteristic parameters	Lithology	Number of samples before SMOTE sampling	Number of samples after SMOTE sampling
BH7	GR	Limestone	156	378
	2 MHz phase resistivity	Mudstone	410	410
	DEN	Dacite	1119	1119
	400 kHz attenuation resistivity	Sandstone	427	427
E3	GR	Slate	77	77
	2 MHz phase resistivity	Mudstone	255	255
	AC	Sandstone	27	76
	CNL	Shell limestone	10	63
		Argillaceous	15	70
S1	GR	Siltstone	45	495
	2 MHz phase resistivity	Shale	520	520
	DEN	Argillaceous	31	372
	AC			

Table 5
Model parameter setting.

Model	Model parameter
CART	splitter = best
	max_features = 4
	max_depth = 100
	min_samples_split = 2
	min_samples_leaf = 1
	max_leaf_nodes = None
	min_impurity_split = 0
RF	max_features = 4
	n_estimators = 58
	max_depth = 100
	min_samples_split = 2
	min_samples_leaf = 1
	max_leaf_nodes = None
	min_impurity_split = 0
SVM	kernel = RBF
	gamma = 0.125
	C = 87
BPNN	training method = traingd
	net.trainParam.epochs = 500
	net.trainParam.goal = 1e-5
	net.trainParam.lr = 0.05
	net.trainParam.time = inf
	net.trainParam.min_grad = 1e-10
DFW-RF	max_features = 4
	n_estimators = 58
	max_depth = 100
	min_samples_split = 2
	min_samples_leaf = 1
	max_leaf_nodes = None
	min_impurity_split = 0

lithology identification results of the RF model after dynamic felling, the RF model after weighted voting and the DFW-RF model, it can be seen that the DFW-RF model combined with dynamic felling and weighted voting is better than the RF model only with dynamic felling or weighted voting. The average identification accuracy rate of the four lithology of DFW-RF model is 8.73% higher than that of the RF model, which also shows that the combination of dynamic felling and weighted voting improves the classification performance of the RF model more effectively.

The author further analyzes the influence of SMOTE sampling technology and the Dynamic Felling Strategy Weighted Random Forest algorithm on lithology identification. Compared with Figs. 6 and 8, it can be seen that the primary function of SMOTE sampling is to improve the lithology identification accuracy rate of the thin-layer. However, the identification accuracy rate of other lithologies with sufficient samples has not been greatly improved. In contrast, the Dynamic Felling Strategy Weighted Random Forest algorithm positively impacts the overall lithology identification performance of the model. It has significantly improved the identification accuracy and stability of each type of lithology. For lithology identification, we should not only consider the problem that thin-layer lithology is challenging to identify but also consider the accuracy and stability requirements of lithology identification. Therefore, combining SMOTE sampling technology with the Dynamic Felling Strategy Weighted Random Forest algorithm is necessary.

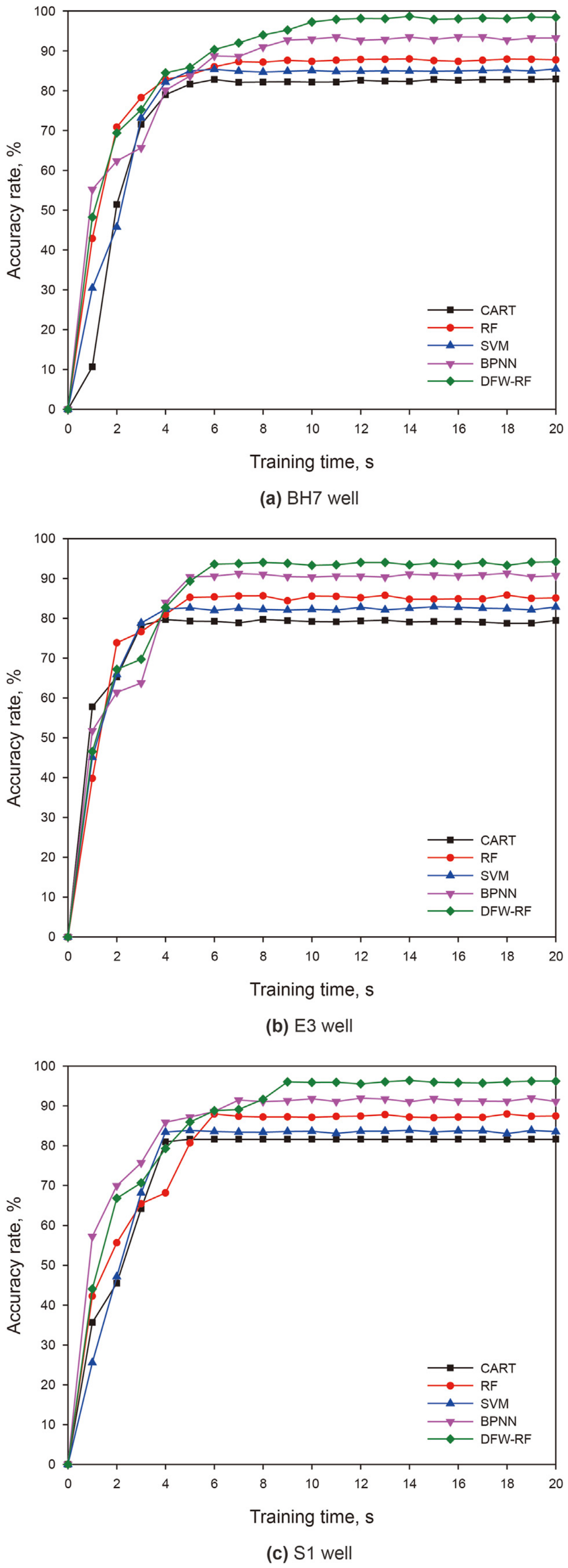
5. Field application

To verify the application effect of the DFW-RF model, the DFW-RF model is used to identify the lithology of the BH7 well in block B of X oilfield, E3 well in the east area of H oilfield and S1 well in block II of S oilfield. By comparing with the actual logging lithology interpretation results, the lithology identification results of the DFW-RF model are verified. The experiment was ducted at the High-Efficiency Drilling and Rock Breaking Technology Laboratory of Northeast Petroleum University, Daqing. The computing server adopts the operating system of Red Hat Enterprise Linux Server Release 6.2, and the processor is Intel Xeon E5-2650. The total storage capacity is 1.5 PB, with 120 computing nodes and 1920 processor cores. The algorithm is programmed in a Python language environment.

Fig. 10 shows the LWD data and logging lithology interpretation results of BH7 well, E3 well and S1 well. All 3 wells were sampled at equal depth intervals of 0.125 m in the selected logging section. The depth of the selected logging section in Well BH7 ranges from 3875.89~4139.99 m, with a total of 2112 samples. According to the logging lithology interpretation results, there are 156 limestone sample points, 410 mudstone sample points, 1119 dacite sample points and 427 sandstone sample points. The depth of the selected logging section of Well E3 ranges from 2541.07~2589.16 m, with a total of 384 samples. According to the logging lithology interpretation results, there are 77 slate sample points, 255 mudstone sample points, 27 sandstone sample points, 10 shell limestone sample points and 15 argillaceous sample points. The depth of the selected logging section of Well S1 ranges from 2425.10~2499.61 m, with a total of 596 samples. According to the logging lithology interpretation results, there are 45 siltstone sample points, 520 shale sample points and 31 argillaceous sample points.

According to the lithology identification process of the DFW-RF model (Fig. 2), the lithology identification characteristic parameters of BH7, E3, and S1 wells are selected through the mRMR algorithm. The thin-layer lithology samples of BH7, E3, and S1 wells are processed using SMOTE sampling technology. The characteristic parameter selection and data sampling results are shown in Table 4.

It is concluded from Table 4 that the lithologic identification characteristic parameters of the BH7, E3, and S1 wells are different.

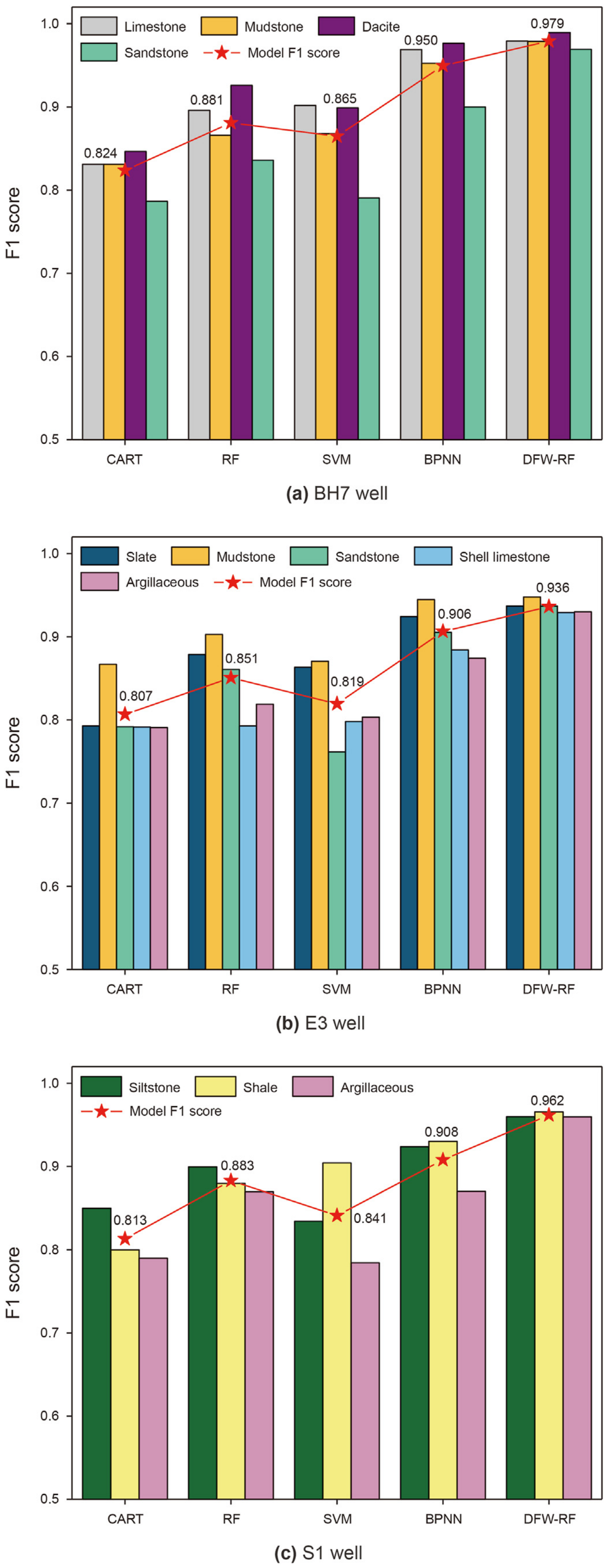


(a) BH7 well

(b) E3 well

(c) S1 well

Fig. 11. Lithology identification accuracy rate and training time of different models.



(a) BH7 well

(b) E3 well

(c) S1 well

Fig. 12. The *F*1 scores of each model and the identification accuracy rate of different lithologies.

That is because each well is in a different geographical environment, and the drilled section's sedimentary environment and structural characteristics differ. Among the seven LWD parameters, GR and 2 MHz phase resistivity are sensitive to the lithology of the three wells. Through SMOTE sampling, the thin-layer sample data of three wells have been expanded to some extent. The single-layer thickness of the mudstone layer, dacite layer, and sandstone layer in the BH7 well is more than 3.00 m, and the single-layer thickness of the limestone layer is 0.72~1.44 m. The number of limestone samples only accounts for 7.39% of the total samples, so processing the limestone sample data is necessary. After SMOTE sampling, the number of limestone samples in the BH7 well has been expanded from 156 to 378. The single-layer thickness of slate in the E3 well is more than 3.45 m, and that of the mudstone layer is more than 1.55 m. The single-layer thickness of the sandstone, shell limestone, and argillaceous layer is less than 0.50 m. Sandstone, shell limestone, and argillaceous samples account for less than 7.0% of the total samples. It is necessary to process the sample data of sandstone, shell limestone, and argillaceous. After SMOTE sampling, the number of sandstone, shell limestone, and argillaceous samples in the E3 well is 76, 63, and 70, respectively. The total number of samples in the E3 well increased by 157. The single-layer thickness of the shale layer in the S1 well is more than 1.75 m, the average thickness of the siltstone layer is 0.40 m, and the average thickness of the argillaceous layer is 0.33 m. The samples of siltstone and argillaceous account for less than 7.5% of the total samples, respectively. The sample data of siltstone and argillaceous need to be processed. After SMOTE sampling, the number of samples of siltstone and argillaceous in the S1 well is 495 and 372, respectively. The total number of samples in the S1 well increased by 791.

Input the processed sample data into the DFW-RF model for lithology identification. At the same time, the lithology identification results of the Classification and Regression Tree model (CART), Random Forest model (RF), Support Vector Machine model (SVM) and Back Propagation Neural Network model (BPNN) are compared and analyzed in the experiment. The lithology identification characteristic parameters and sample data input by the above model are the same as those of the DFW-RF model. The grid search algorithm determines the optimal value of super parameters. Table 5 shows the parameter settings of each model. The lithology identification accuracy rate and training time of each model are shown in.

By comparing the lithology identification accuracy rate and training time of different models in Fig. 11, it can be seen that the DFW-RF model has the highest lithology identification accuracy rate, followed by the BPNN model. The RF model composed of multiple decision trees is more accurate than the CART model of a single decision tree in lithology identification. The lithology identification effect of the SVM model is worse than that of the RF model but better than that of the CART model. The model with a simpler algorithm structure has a faster training speed in terms of training time. The training time of SVM and CART models is short, followed by RF models, and the training time of BPNN and DFW-RF models is relatively long.

The training speed of the RF model, SVM model and CART model is fast. However, the lithology identification accuracy rate of BH7, E3 and S1 wells is lower than 90.0% in the above model. Under the BPNN model, the lithologic identification accuracy rate of BH7, E3 and S1 wells is 95.25%, 90.41% and 91.47%, respectively, and the training time of the model is 9, 5 and 7 s, respectively. The average identification accuracy rate of three wells under the BPNN model is 92.38%, and the average training time is 7.0 s. Under the DFW-RF model, the lithology identification accuracy rate of BH7, E3 and S1 wells is 97.92%, 93.61% and 96.04%, respectively, and the model training time is 11, 6 and 9 s, respectively. The average identification accuracy rate of the three wells under the DFW-RF model is 95.86%, and the average training time is 8.7 s. The average training time of the DFW-RF model is 1.7 s longer than that of the BPNN model, but the average lithology identification accuracy rate is increased by 3.48%. The accuracy of lithology identification is more critical for underground geological guidance. The calculation time of the DFW-RF model is seconds, which can meet the real-time requirements of intelligent steering drilling technology, so the comprehensive performance of the DFW-RF model is better.

In order to evaluate the comprehensive performance of different models, the $F1$ score is introduced. The $F1$ score is an index to measure the model's comprehensive performance, which considers both the model's precision rate and recall rate. The value range of the $F1$ score is [0, 1], and the higher the $F1$ score, the higher the model quality. The calculation method of the $F1$ score is shown in Eqs. (11)–(13). The $F1$ scores of each model and the identification accuracy rate of different lithologies are shown in Fig. 12.

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{11}$$

$$\text{precision} = \frac{TP}{TP + FP} \tag{12}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{13}$$

where precision indicates the percentage of real positive samples that are classified as positive, recall indicates the proportion of actual positive samples within the samples that are predicted to be positive.

According to the $F1$ score of different models in Fig. 12, the DFW-RF model has the highest $F1$ score, followed by the BPNN, RF, SVM, and CART models. The $F1$ score of the RF, SVM and CART models in BH7, E3, and S1 wells are all less than 0.900. The $F1$ score of the BPNN model in BH7, E3 and S1 wells are 0.950, 0.906 and 0.908, respectively, and the average $F1$ score is 0.921. The $F1$ score of the DFW-RF model in BH7, E3 and S1 wells are 0.979, 0.936 and 0.962, respectively, and the average $F1$ score is 0.959. The average $F1$ score of the DFW-RF model is 3.8% higher than that of the BPNN model, which indicates that the comprehensive performance of the DFW-RF model is better than the BPNN model.

By analyzing the $F1$ score of different lithology under five models, it is concluded that the DFW-RF model has improved the identification effects of various lithologies to varying degrees, and the lithology identification effect of this model is more stable. In the BH7 well, the $F1$ score of limestone, mudstone, dacite and sandstone under the DFW-RF model has been improved, and the $F1$ score of four lithology types is more than 0.965. DFW-RF model not only improves the identification effects of limestone thin-layer in BH7 well but also effectively solves the problem of low identification effects of sandstone by the other four methods. In E3 well, the five models show good identification effects for mudstones with sufficient sample data, but the identification effects for other lithologies are unstable. The BPNN model improves the identification effect of slate and sandstone in the E3 well, but the $F1$ score of shell limestone and argillaceous is still lower than 0.875. Through the DFW-RF model, the $F1$ score of shell limestone and argillaceous in the E3 well has been significantly improved. Compared with the BPNN model, the shell limestone $F1$ score under the DFW-RF model increased from 0.884 to 0.929. The $F1$ score of argillaceous is increased from 0.874 to 0.930. In the S1 well, the $F1$ score of the CART model and RF model for siltstone, shale and argillaceous is lower than 0.900. The SVM model has a better identification effect on shale but a poor identification effect on thin layers of siltstone and argillaceous. The $F1$ score of siltstone and shale in the BPNN model is more than 0.920, but the $F1$ score of argillaceous is only 0.870. The BPNN model is not effective in identifying argillaceous. Under the DFW-RF model, the $F1$ score of siltstone, shale and argillaceous in the S1 well is more than 0.960. The DFW-RF model can more effectively identify the lithology of siltstone and argillaceous. The lithology identification results of the DFW-RF model in BH7, E3 and S1 wells show that the DFW-RF model can accurately and stably identify different lithology types in different geological environments. The DFW-RF model has good applicability and popularization, providing technical support for real-time intelligent lithology identification in closed-loop drilling.

6. Conclusion

A real-time intelligent lithology identification model (DFW-RF) based on a dynamic felling strategy weighted random forest algorithm is developed using common LWD data in oil and gas fields. Aiming at the problems of LWD information selection and thin-layer lithology scarcity in closed-loop drilling, the DFW-RF model adopts the minimum Redundancy Maximum Relevance algorithm to extract lithology-sensitive LWD parameters and introduces SMOTE sampling technology to expand the thin-layer lithology information capacity. At the same time, it also strengthens the comprehensive performance of the model by cutting down the decision tree with strong correlation and voting weighted according to the classification effect.

Based on LWD data from three wells in different areas, the field application proves that the DFW-RF lithology identification model has higher lithology identification accuracy and efficiency than the Decision Tree model, Random Forest model, Support Vector Machine model and Back Propagation Neural Network model. The DFW-RF model shows high accuracy and stability in dealing with thin-layer lithology identification problems. This model's calculation time is in the second, which can meet the real-time requirements of intelligent steering drilling technology. Integrating the DFW-RF lithology identification model into an intelligent drilling system can solve the problems of low intelligence, unstable identification effect and low accuracy of thin-layer identification in closed-loop drilling to some extent. As the applicability and popularization of the DFW-RF model proposed in this paper have been well proved, this new lithology identification method is also effective and feasible for other lithology identification, so this method has high engineering value and application prospects.

Conflict of interest

No conflict of interest exits in the submission of this manuscript, and manuscript is approved by all authors for publication.

Acknowledgements

This work is financially supported by the National Natural Science Foundation of China (No. 52174001); the National Natural Science Foundation of China (No. 52004064); the Hainan Province Science and Technology Special Fund “Research on Real-time Intelligent Sensing Technology for Closed-loop Drilling of Oil and Gas Reservoirs in Deepwater Drilling” (ZDYF2023GXJS012), and Heilongjiang Provincial Government and Daqing Oilfield's first batch of the scientific and technological key project “Research on the Construction Technology of Gulong Shale Oil Big Data Analysis System” (DQYT-2022-JS-750).

Abbreviation

AUC	The index to measure the performance of the classification learner
$\boldsymbol{a}$	The confusion matrix vector of the decision tree a
$\boldsymbol{a}\cdot\boldsymbol{b}$	The dot product of the confusion matrix of a decision tree a and the confusion matrix of the decision tree b
$\lVert\mathbf{a}\rVert$	The length of the confusion matrix of the decision tree a
$\boldsymbol{b}$	The confusion matrix vector of decision tree b
$\lVert\mathbf{b}\rVert$	The length of the confusion matrix of the decision tree b
c	The lithology category
$count(OOB)$	The overall sample size of the OOB
$E*$	The generalization error of the RF algorithm
f_r	LWD parameter
f_o	LWD parameter
F	The original LWD parameter set
$I(f_r, f_o)$	The mutual information between the LWD parameter f_r and the LWD parameter f_o
$I(f_r, c)$	The mutual information between the LWD parameter f_r and the lithology category c
k	Number of adjacent samples
m	Number of LWD parameters in the LWD parameter set
M	Number of samples in the non-thin-layer sample set
N	Number of samples in the thin-layer sample set
OOB	Out-of-bag data
$OOB_{\text{correct}}(i)$	The correct number of samples predicted by the i decision tree in the OOB
$p(f_r)$	The probability density of the LWD parameter f_r
$p(f_o)$	The probability density of the LWD parameter f_o
$p(c)$	The probability density of the lithology c
$p(f_r, f_o)$	The combined probability density of the LWD parameter f_r and the LWD parameter f_o
$p(f_r, c)$	The combined probability density of the LWD parameter f_r and the lithology c
precision	The accuracy rate of the lithology identification model
P	The sampling ratio
recall	The recall rate of the lithology identification model
s	The overall classification strength of the decision tree
$\text{Sim}(a, b)$	The confusion matrix similarity between decision tree a and the decision tree b
T	The original sample data set
T_{majority}	The non-thin-layer sample data set
T_{minority}	The thin-layer sample data set
w_i	The weight of the decision tree i
x_i	The thin-layer sample
x_{new}	The newly generated thin-layer sample
y_j	The adjacent sample of the thin-layer sample
$\bar{\rho}$	The average correlation of decision trees

References

Abdelhakim, L., Abdellah, E.H., Karima, B., et al., 2020. Mapping specific groundwater vulnerability to nitrate using random forest: case of Sais basin, Morocco. Modeling Earth Systems and Environment 6 (3), 1451–1466. https://doi.org/10.1007/s40808-020-00761-6.

Al-Mudhafar, J.W., 2017. Integrating kernel support vector machines for efficient rock facies classification in the main pay of Zubair formation in South Rumaila oil field, Iraq. Modeling Earth Systems and Environment 3 (1), 1–8. https://doi.org/10.1007/s40808-017-0277-0.

Alzubaidi, F., Mostaghimi, P., Swietojanski, P., et al., 2021. Automated lithology classification from drill core images using convolutional neural networks. J. Petrol. Sci. Eng. 197, 107933. https://doi.org/10.1016/j.petrol.2020.107933.

Amjad, A., Chen, S.C., 2020. Characterization of well logs using K-mean cluster analysis. J. Pet. Explor. Prod. Technol. 10 (6), 2245–2256. https://doi.org/10.1007/s13202-020-00895-4.

Baisakhi, D., Rima, C., 2018. Well log data analysis for lithology and fluid identification in Krishna-Godavari basin, India. Arabian J. Geosci. 11, 1–12. https://doi.org/10.1007/s12517-018-3587-2.

Becerra, D., Pires, D.L.R., Galvis-Portilla, H., et al., 2022. Generating a labeled data set to train machine learning algorithms for lithologic classification of drill cuttings. Interpretation 10 (3), SE85–SE100. https://doi.org/10.1190/INT-2021-0194.1.

Boonen, P., Valant-Spaight, B., de-Andre, C.A., et al., 2005. A comparison of logging-while-drilling and wireline nuclear porosity logs in shales from wells in Brazil. Petrophysics 46 (4), 295–301. https://doi.org/10.1144/1354-079304-636.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Chawla, N.V., Bowyer, K.W., Hall, L.O., et al., 2002. Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953.

Dong, S.Q., Zeng, L.B., Du, X.Y., et al., 2022. Lithofacies identification in carbonate reservoirs by multiple kernel Fisher discriminant analysis using conventional well logs: a case study in A oilfield, Zagros Basin, Iraq. J. Petrol. Sci. Eng. 210, 110081. https://doi.org/10.1016/j.petrol.2021.110081.

Ghawi, R., Pfeffer, J., 2019. Efficient hyperparameter tuning with grid search for text categorization using kNN approach with BM25 similarity. Open Computer Science 9 (1), 160–180. https://doi.org/10.1515/comp-2019-0011.

Han, X., Feng, F.P., Zhang, X.C., et al., 2023. An unequal fracturing stage spacing optimization model for hydraulic fracturing that considers cementing interface integrity. Petrol. Sci. 20 (4), 2165–2186. https://doi.org/10.1016/j.petsci.2023.05.010.

Hjelm, R.D., Alex, F., Samuel, L.M., et al., 2018. Learning deep representations by mutual information estimation and maximization. Statistics 2, 1–24. http://arxiv.org/abs/1808.06670.

Jorge, A.L., Luis, H.O., Carmen, C.C., 2018. Automatic identification of calcareous lithologies using support vector machines, borehole logs and fractal dimension of borehole electrical imaging. Earth Sci. Res. J. 22 (2), 75–82. https://doi.org/10.15446/esrj.v22n2.68320.

Karimzadeh, S., Tangestani, H.M., 2021. Evaluating the VNIR-SWIR datasets of WorldView-3 for lithological mapping of a metamorphic-igneous terrain using support vector machine algorithm; a case study of Central Iran. Adv. Space Res. 68 (6), 2421–2440. https://doi.org/10.1016/j.asr.2021.05.002.

Kumar, T., Seelam, N.K., Rao, G.S., 2022. Lithology prediction from well log data using machine learning techniques: a case study from Talcher coalfield, Eastern India. J. Appl. Geophys. 199, 104605. https://doi.org/10.1016/j.jappgeo.2022.104605.

Li, S.Q., Chen, Z., Li, W., et al., 2023. An FE simulation of the fracture characteristics of blunt rock indenter under static and harmonic dynamic loadings using cohesive elements. Rock Mech. Rock Eng. 56 (4), 2935–2947. https://doi.org/10.1007/s00603-022-03214-x.

Li, Z., Wu, Y., Kang, Y., et al., 2021. Feature-depth smoothness based semi-supervised weighted extreme learning machine for lithology identification. J. Nat. Gas Sci. Eng. 96, 104306. https://doi.org/10.1016/j.jngse.2021.104306.

Liang, H., Chen, H., Guo, J., et al., 2022. Research on lithology identification method based on mechanical specific energy principle and machine learning theory. Expert Syst. Appl. 189, 116142. https://doi.org/10.1016/j.eswa.2021.116142.

Merembayev, T., Kurmangaliyev, D., Bekbauov, B., et al., 2021. A comparison of machine learning algorithms in predicting lithofacies: case studies from Norway and Kazakhstan. Energies 14 (7), 1896. https://doi.org/10.3390/en14071896.

Miao, T., Henning, O., Xu, H.M., 2021. Inversion of well logs into lithology classes accounting for spatial dependencies by using hidden Markov models and recurrent neural networks. J. Petrol. Sci. Eng. 196, 107598. https://doi.org/10.1016/j.petrol.2020.107598.

Peng, H.C., Long, F.H., Ding, C., 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27 (8), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159.

Phillip, J.H., Trudy, D., John, A.W., et al., 2017. Elemental differences: geochemical identification of aboriginal silcrete sources in the Arcadia Valley, eastern Australia. J. Archaeol. Sci.: Report 15 (15), 570–577. https://doi.org/10.1016/j.jasrep.2016.11.032.

Polat, Ö., Polat, A., Ekici, T., 2021. Automatic classification of volcanic rocks from thin section images using transfer learning networks. Neural Comput. Appl. 33 (18), 11531–11540. https://doi.org/10.1007/s00521-021-05849-3.

Ren, Q., Zhang, D., Zhao, X., et al., 2022. A novel hybrid method of lithology identification based on k-means++ algorithm and fuzzy decision tree. J. Petrol. Sci. Eng. 208, 109681. https://doi.org/10.1016/j.petrol.2021.109681.

Rosid, M.S., Haikel, S., Haidar, M.W., 2019. Carbonate reservoir rock type classification using comparison of Naïve Bayes and Random Forest method in field “S” East Java. AIP Conf. Proc. 2168 (1), 1–9. https://doi.org/10.1063/1.5132446.

Stephen, K., Matthew, J.C., Anya, M.R., 2019. Lithological mapping in the central African Copper Belt using random forests and clustering: strategies for optimized results. Ore Geol. Rev. 112, 103015. https://doi.org/10.1016/j.oregeorev.2019.103015.

Sui, Y., Cao, G.S., Guo, T.Y., et al., 2022. Development of gelled acid system in high-temperature carbonate reservoirs. J. Petrol. Sci. Eng. 216, 110836. https://doi.org/10.1016/j.petrol.2022.110836.

Sun, J., Chen, M., Li, Q., et al., 2021. A new method for predicting formation lithology while drilling at horizontal well bit. J. Petrol. Sci. Eng. 196, 107955. https://doi.org/10.1016/j.petrol.2020.107955.

Sun, J., Li, Q., Chen, M.Q., et al., 2019. Optimization of models for a rapid identification of lithology while drilling-a win-win strategy based on machine learning. J. Petrol. Sci. Eng. 176, 321–341. https://doi.org/10.1016/j.petrol.2019.01.006.

Wang, X.D., Yang, S.C., Zhao, Y.F., et al., 2018. Lithology identification using an optimized KNN clustering method based on entropy-weighed cosine distance in Mesozoic strata of Gaoqing field, Jiyang depression. J. Petrol. Sci. Eng. 166, 157–174. https://doi.org/10.1016/j.petrol.2018.03.034.

Wu, Z.B., Wang, J., Xi, K.K., et al., 2022. Research on control system of small intelligent drilling rig based on lithology identification. J. Phys. Conf. 2181 (1), 012039. https://doi.org/10.1088/1742-6596/2181/1/012039.

Wu, Z.Y., Zhang, X., Zhang, C.L., et al., 2021. Lithology identification based on LSTM recurrent neural network. Lithologic Reservoirs 33 (3), 120–128. https://doi.org/10.12108/yxyqc.20210312 (in Chinese).

Xie, H.M., Zhou, J., Zhang, P.F., 2022. Simulation research on vibration parameters model of drill string. Fuel 315, 122351. https://doi.org/10.1016/j.fuel.2021.122351.

Xu, M.H., Zhu, X.Y., Gao, S.L., et al., 2022. Joint use of multi-seismic information for lithofacies prediction via supervised convolutional neural networks. Geophysics 87 (5), 151–162. https://doi.org/10.1190/GEO2021-0554.1.

Xu, T., Zhang, W.T., Li, J., et al., 2022. Domain generalization using contrastive domain discrepancy optimization for interpretation-while-drilling. J. Nat. Gas Sci. Eng. 105, 104685. https://doi.org/10.1016/j.jngse.2022.104685.

Yao, L., Fang, Z., Xiao, Y., et al., 2021. An intelligent fault diagnosis method for lithium battery systems based on grid search support vector machine. Energy 214, 118866. https://doi.org/10.1016/j.energy.2020.118866.

Zeng, L.L., Ren, W.J., Shan, L.Q., 2020. Attention-based bidirectional gated recurrent unit neural networks for well logs prediction and lithology identification. Neurocomputing 414, 153–171. https://doi.org/10.1016/j.neucom.2020.07.026.

Zhang, J., He, Y., Zhang, Y., et al., 2022. Well logging based lithology classification using machine learning methods for high quality reservoir identification: a case study of Baikouquan formation in Mahu area of Junggar basin, NW China. Energies 15 (10), 3675. https://doi.org/10.3390/en15103675.

Zhao, L.X., Zou, C.F., Chen, Y.Y., et al., 2021. Fluid and lithofacies prediction based on integration of well-log data and seismic inversion: a machine-learning approach. Geophysics 86 (4), 151–165. https://doi.org/10.1190/GEO2020-0521.1.

Zhou, Z.L., Wang, G.W., Ran, Y., et al., 2016. A logging identification method of tight oil reservoir lithology and lithofacies: a case from Chang 7 member of Triassic Yanchang Formation in Heshui area, Ordos Basin, NW China. Petroleum Exploration and Development Online 43 (1), 65–73. https://doi.org/10.1016/S1876-3804(16)30007-6.

Zou, C.F., Zhao, L.X., Xu, M.H., et al., 2021. Porosity prediction with uncertainty quantification from multiple seismic attributes using Random Forest. J. Geophys. Res. Solid Earth 126 (7), e2021JB021826. https://doi.org/10.1029/2021JB021826.