IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 5, 2021, accepted November 16, 2021, date of publication November 22, 2021,
date of current version December 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3129830

An eXtreme Gradient Boosting Algorithm Combining Artificial Bee Colony Parameters Optimized Technique for Single Sand Body Identification

RENZE LUO[1], LIANG GUO[1], XINGYU LI[1], JUANJUAN TUO[1], CANRU LEI[1], AND YANG ZHOU[2]
[1]State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation, School of Earth Sciences and Technology, Southwest Petroleum University, Chengdu 610500, China
[2]School of Earth Sciences and Technology, Southwest Petroleum University, Chengdu 610500, China

Corresponding author: Renze Luo (lrzsmith@126.com)

This work was supported in part by the Deep Earthquake Special Project of the National Key Research and Development Program under Grant 2016YFC0601100, and in part by the Science and Technology Project of Sichuan Province under Grant 2019CXRC0027.

ABSTRACT Due to the problems of traditional artificial single sand body identification methods such as strong subjectivity, heavy workload and low efficiency, we propose a fast and objective ABC-XGBoost. The algorithm consists of two parts: eXtreme gradient boosting (XGBoost) and artificial bee colony algorithm (ABC). XGBoost introduces a regular term, which can effectively prevent overfitting, and uses the second derivative to make the identification result more accurate. However, a large number of parameters in XGBoost need to be adjusted manually, which affects the efficiency of the algorithm. In this regard, ABC is used to optimize the parameters based on XGBoost, and then the single sand body can be identified quickly and effectively. We take the $C6_1^2$ oil-bearing layer in the second area of Dalugou, Jing'an Oilfield as the research object, and use the ABC-XGBoost to identify the single sand body in the research area. Based on the reasonable selection of physical parameter data and logging data, the partition and interlayer data should be eliminated first to avoid data redundancy. The results indicate that ABC-XGBoost is more efficient and accurate than the existing mainstream machine algorithms, such as support vector machines (SVM), random forests (RF), and XGBoost using trial and error tuning under the same logging data and computer hardware conditions. The accuracy can reach 90.6%, which has certain practical application value in the middle and late development of oil and gas fields.

INDEX TERMS Single sand body identification, artificial bee colony algorithm, machine learning, eXtreme gradient boosting, ABC-XGBoost.

I. INTRODUCTION

In the past few decades, researchers have mainly relied on traditional manual methods, combined with drilling and coring, logging, mud logging, and other data to identify single sand bodies. This method must consider comprehensive geological background factors, such as the sedimentary genesis of a single sand body, sequence stratigraphy, and spatial combination. [1]–[3]. However, the reservoir has strong heterogeneity, and the nonlinear correlation between various parameters is complicated. In areas with complex geological backgrounds, traditional methods not only require staff to analyze and compare logging data but also comprehensively consider the geological background. As a result, different staff members face the same data due to different understandings and often obtain different results. Therefore, the traditional artificial single sand body identification has limitations and uncertainties [4].

In response to the above problems, many scholars have attempted to use pure data-driven methods to quantitatively identify lithology, which effectively avoids human error and improves the efficiency of single sand body identification. Luo *et al.* used logging data as the basis, combined with rock slice identification, mineral composition analysis and

The associate editor coordinating the review of this manuscript and approving it for publication was Wenbing Zhao.

other data, and used BP neural network to identify the formation lithology with an accuracy of approximately 80% [5]. Amosu *et al.* verified the effectiveness of a data-driven support vector machine (SVM) to identify TOC enrichment areas through data from the Barnett Shale Formation in the Fort Worth Basin in North Texas and the Duvernay Shale Formation in the Western Canada Sedimentary Basin in Canada [6]. Chen *et al.*. successfully divided sand bodies by using SVM combined with Gaussian radial basis function and grid search (GS) to determine penalty factors [7]. Although the above methods for sand body identification have high accuracy, they have problems such as slow convergence speed and easy to fall into local optimum. Additionally, due to the heterogeneity of underground reservoirs, the characteristics of actual logging data are usually imbalanced, and there are certain limitations when using traditional machine learning models to solve such problems.

The ensemble learning (EL) method transforms a single weak classifier into a strong classifier to obtain higher classification accuracy and generalization ability. In recent years, methods based on decision trees (DTs) such as random forests (RF), and gradient boosting decision tree (GBDT), which have been widely used in lithology identification [8]–[13], reservoir prediction [14]–[18] and other fields. This series of methods have achieved more efficient and accurate results than traditional classification algorithms. However, whether it is SVM, RF, or eXtreme gradient boosting (XGBoost), it is often necessary to optimize a variety of empirical parameters during model training to obtain the best results [19]. The traditional parameter optimization method uses trial and error, or GS traverses different parameter combinations and determines the optimal choice by comparing the final results, which is inefficient.

Artificial bee colony algorithm (ABC) was proposed by Karaboga in 2005 to solve the optimization problem by simulating the collective behavior of bee colony honey collection [20]. Compared with the genetic algorithm (GA) and particle swarm optimization algorithm (PSO), ABC has the characteristics of simultaneous global search and local search, making it easier to jump out of the local optimum [21]. In this research, the main problem we face is that the identification results of the single sand body are unstable due to subjective reasons. Instability is mainly caused by two factors. First, each researcher has a different understanding of the geological background, resulting in different identification results. In this regard, we use data-driven XGBoost. Second, there are many parameters and the adjustment processes are complicated. Each researcher will choose different parameter adjustment methods due to different knowledge backgrounds. Identification results have difficulty achieving uniformity. Therefore, we introduce ABC into the parameter adjustment of XGBoost to make the effect of single sand body identification more objective. The best data are automatically selected, redundant data such as stratigraphic partitions and interlayers are removed in advance and combined with ABC-XGBoost, which can minimize the time-consuming single sand body identification. These measures make our research more objective and efficient.

II. ABC-XGBOOST METHOD PRINCIPLE

When the traditional DT algorithms identify the single sand body based on logging data, they need to enumerate all the possibilities of each feature, and then find the best segmentation point by calculating the mean square error, which results in the algorithm's inefficiency [22]. The XGBoost arranges the features in advance according to the percentile method and selects candidates that may become the split point. By calculating the split point to find the best split point from the candidates, the efficiency is significantly higher than that of the traditional DT. Additionally, the XGBoost pays attention to issues such as missing values and the sparseness of logging data. When the logging data are missing, the branch direction is designated for it, which greatly improves the applicability of the algorithm. In addition, XGBoost adds a regularization term to the loss function [23]. The regularization term reduces the variance in the model, makes the final model obtained by training simpler, and can effectively prevent overfitting. Finally, unlike the gradient boosting (GB) algorithm, which uses the loss function to perform the first-order derivative of the original function to calculate the pseudoresiduals, XGBoost uses the second-order derivative to make the model more accurate [24]. XGBoost also draws on RF and supports column sampling of data, which not only effectively prevents model overfitting but also reduces the number of calculations for model training [25].

Although XGBoost has great advantages in using logging data to finely identify a single sand body, it also has obvious shortcomings. The largest shortcoming is that XGBoost has many parameters, and the parameter adjustment is very complicated [26]. The quality of parameter adjustment makes the performance of the trained model very different. Therefore, we use the ABC to optimize the selection of its parameters. The ABC adopts three kinds of bee colony division mechanisms, adopts different search methods according to the different functions of each bee colony, and adjusts according to the results found at any time to reduce search complexity [27]. There is a cooperative working mechanism between the three bee colonies. When the bee colony selects the path, it refers to the information left by the previous bees, which can find the optimal solution to the problem more quickly and speed up the algorithm convergence speed [28]. Additionally, the ABC does not need to consider its information characteristics that need to be solved, but only needs to compare the pros and cons of the results, so that the ABC can be easily combined with other machine learning algorithms to improve model performance [29]. Finally, a single local optimal set in the ABC can reach the global optimal, and the model will not collapse due to individual failure [30]. The specific calculation principle and process of the combination of ABC and XGBoost applied to single sand body identification based on logging data are as follows:

XGBoost generates a new tree for the model after each iteration. The goal of the newly generated tree is to fit the residual of the previous tree to build a more accurate model. The objective function in the t round of iteration can be expressed as:

$$\text{obj}^{(t)} = \sum_{i=1}^{n} L\left[(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right] + \Omega(f_t) \tag{1}$$

where $L(y_i, \hat{y}_i)$ is the training error of the sample, and $\hat{y}_i$ and y_i denote the prediction of the sample and the actual classification label, respectively. $\Omega(f_k)$ is the regularization item, which optimizes the model structure while reducing overfitting, as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{2}$$

where T is the number of leaf nodes, ω_j is the weight of the corresponding leaf nodes, γ and λ are constants, which represent the penalty coefficient. The objective function is approximated by the Taylor formula, and the second-order Taylor expansion of Eq. (1) is:

$$\text{obj}^{(t)} \cong \sum_{i=1}^{n} \left[L(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \tag{3}$$

where g_i and h_i are the first and second derivatives of a pair, respectively, so Eq. (3) can be simplified to:

$$\text{obj}^{(t)} \simeq \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i\right) \omega_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right) \omega_j^2\right] + \gamma T \tag{4}$$

Take the partial derivative of the objective function and set it equal to 0 to obtain the optimal weight that minimizes the objective function:

$$\omega_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{5}$$

Substitute Eq. (5) into Eq. (4) to obtain the optimal value of the objective function:

$$\text{obj}^* = -\frac{1}{2} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{6}$$

XGBoost needs to set many empirical parameters in modeling, such as the number of CART leaf nodes, learning rate, and regularization coefficients. A large number of parameters make the tuning process extremely complicated. The steps to optimize XGBoost using the ABC algorithm are as follows:

(1) The dimensionality is eliminated while retaining the log shape by standardizing the original log dataset. The formula is as follows:

$$y_i = \frac{x_i - \bar{x}}{s} \tag{7}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{8}$$

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{9}$$

where x_i is the log data of a certain sampling point, y_i is the result of normalization of x_i, and $\bar{x}$ and s are the mean and variance, respectively, of the log data.

(2) The standardized dataset is divided into a training set and a test set. Select the feature vector and initialize the ABC parameters. For example, we need to set the number of parameters to be optimized in XGBoost as D and the number of honey sources as S. then, in D-dimensional space, randomly generate S initial solutions ($i = 1,2,\ldots,S$). We also need to set the total number of bees N, the maximum number of individual updates L, and the value range of the parameters to be optimized.

The objective function f_i of the XGBoost is set to cross-entropy loss, and the K-fold cross-validation experiment is used to calculate the fitness value fit_i corresponding to each nectar source:

$$fit_i = \frac{1}{f_i + 1} \tag{10}$$

$$f_i = -\frac{1}{n} \sum_{i=1}^{n} y_i \cdot \log\left(p(y_i^*)\right) \tag{11}$$

where fit_i is the fitness of the $i-th$ solution, n is the number of samples, y_i is the measured value, and y_i^* is the predicted value.

(3) The initial solution of the hired bee is to search for the nectar source to generate a new nectar source:

$$x'_{id} = x_{id} + \varphi_{id}\left(x_{id} - x_{jd}\right) \tag{12}$$

where φ_{id} is a random number between [−1,1], $i \in [1, N]$, $d \in [1, D]$, x_{jd} is a randomly selected candidate nectar, and $i \neq j$. Calculate the fitness value according to formula (10). If the new result is higher than before the search, assign the value x'_{id} to x_{id}.

(4) The observing bee recalculates the fitness value of x_{id} based on the nectar source location shared by the hired bees and selects the nectar source according to the probability related to the fitness. The probability of each nectar source being selected is:

$$P_i = \frac{\text{fit}_i}{\sum_{i=1}^{S} \text{fit}_i} \tag{13}$$

The better the fitness is, the greater the probability of selecting the nectar source.

(5) If the nectar source X_i cannot be improved after a preset number of updates L, the solution corresponding to this position is abandoned. The hired bees in that location also become scouts that randomly search for new nectar sources:

$$x_{id} = \min_d x_{id} + \text{rand}(0, 1) \times \left(\max_d x_{id} - \min_d x_{id}\right) \tag{14}$$

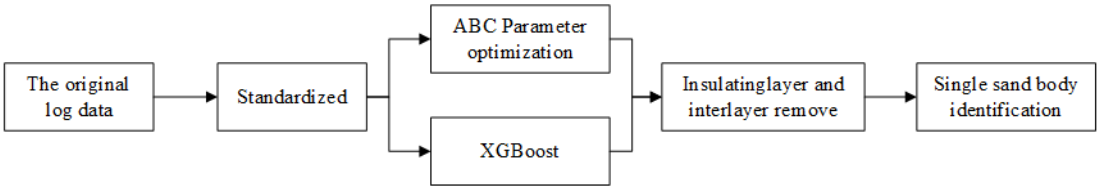

FIGURE 1. The application of the ABC-XGBoost model in single sand body identification.

(6) Judge whether the stop condition has been met. If so, the optimal feature subset and XGBoost parameters are obtained. Otherwise, repeat the previous steps and start a new round of calculations.

(7) The optimal feature subset and parameters are substituted into XGBoost to train the model and output the classification results, and then the correctness of the feature subset and parameters is verified. Finally, the model is applied on the unlabeled test samples and the identification effect of the model is checked. The specific process of the XGBoost model based on the ABC is shown in Figure 1.

III. GEOLOGICAL BACKGROUND

The study area is Jing'an Oilfield Dalugou II, located in the north-central Shanbei Slope Basin, and the overall monoclinic structure west inclined angle is less than 1°. The structure in the area is relatively simple, with only a few developed faults and folds. The research target layer is $C6_1^2$ oil-bearing small interval. According to the actual study area situation, combined with the analysis of the existing geological research results, it is found that the current single sand body identification is not precise enough and the distribution is not clear. As a result, the number of low-yield and low-efficiency wells is large, the basis for stable production is weak, and the remaining oil is enriched but it is difficult to tap the potential. Therefore, the fine identification of single sand bodies is carried out in the study area to provide a strong geological basis for the subsequent targeted improvement of reservoir water flooding efficiency and the formulation of long-term stable production plans.

Based on the detailed comparative study of the strata in the study area, the analysis of the sedimentary microfacies characteristics, and the regional geological background, we formulated the basic principles suitable for the identification of single sand bodies in the study area. First, based on subdivided layer groups and sedimentary microfacies, we used horizon, cycle control, "elevation" comparison methods combined with core observations and field outcrops, to determine two types of barriers and interlayers: argillaceous interlayer and calcareous interlayer.

We removed the partition and interlayer data, and only keep the data segment of a single sand body. On the one hand, it can reduce redundant data and speed up the identification rate of single sand bodies. On the other hand, it can eliminate the negative impact of barrier and interlayer data and can maximize the data characteristics between individual sand bodies.

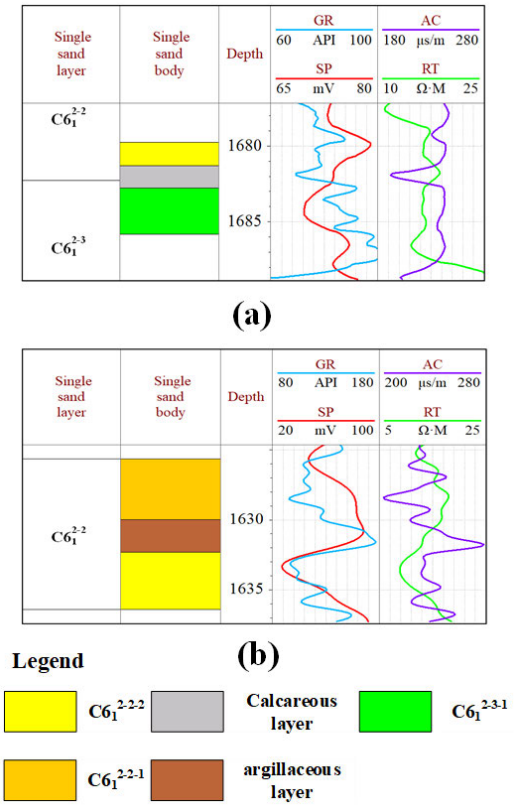

FIGURE 2. Electrical characteristics of the interlayer. (a) is a calcareous interlayer, (b) is a muddy interlayer. Calcareous interlayers mostly appear between estuary dams and underwater distributary channel microfacies. Typical electrical characteristics are as follows: acoustic time difference and natural gamma value decrease, resistivity value rises obviously, and the curve is sharp (Figure 2(a)). The argillaceous interlayer is mainly manifested by a sudden increase in the natural gamma value, and the resistivity curve shows a low value (Figure 2(b)).

Therefore, the experiment was divided into two stages: the elimination of the insulating layer, the interlayer stage and the single sand body identification stage.

IV. EXPERIMENTAL DATA SOURCE AND SELECTION

A. EXPERIMENTAL DATA SOURCE

We selected 5,161 sampling points from 20 wells in the study area where the logging data were complete and the single sand body identification had been completed by experts using traditional methods as the data set. Each well contained natural gamma (GR), spontaneous potential (SP), formation true resistivity (RT), acoustic wave (AC) logging response

TABLE 1. Part of the input data and labels.

Depth (*m*)	GR (*API*)	SP (*mV*)	RT (*Ω·m*)	AC (*μs/m*)	POR (*%*)	PERM (*mD*)	SW (*%*)	Microfacies	Category
1,639.6	71.5	43.3	25.2	236.3	12.7	4.7	45.8	Mouth bar	$C6_1^{2\text{-}1\text{-}1}$
1,639.8	88.7	40.6	24.8	237.8	12.9	5.1	45.5	Mouth bar	$C6_1^{2\text{-}1\text{-}1}$
	…		…		…		…	…	…
1,648.9	94.3	75.4	21.4	218.9	9.3	1.1	59.3	Distributary channel flank	Calcareous interlayer
1,649.0	103.7	71.0	23.8	204.7	9.1	0.8	75.3	Distributary channel flank	Calcareous interlayer
	…		…		…		…	…	…
1,659.8	110.4	71.2	19.7	239.2	-	-	-	Interdistributary bay	Mudstone
1,659.9	107.6	69.8	20.4	237.3	-	-	-	Interdistributary bay	Mudstone
	…		…		…		…	…	…
1,670.6	83.8	67.7	22.2	254.3	15.2	12.3	41.9	Distributary channel flank	$C6_1^{2\text{-}4\text{-}1}$
1,670.8	90.2	69.7	22.2	258.9	15.8	15.7	40.5	Distributary channel flank	$C6_1^{2\text{-}4\text{-}1}$

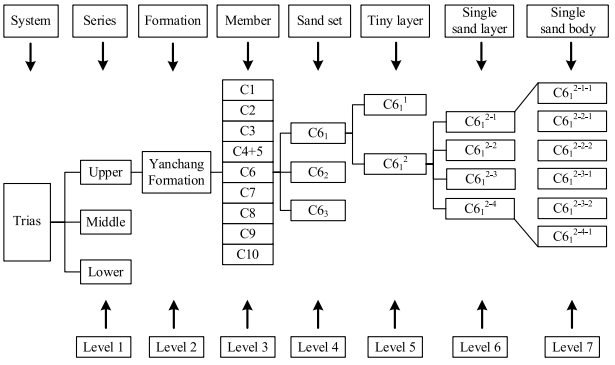

FIGURE 3. Types of small interlayer development in C6, Dalugou second district. Level 1-4 is the stratum division, Level 5-6 is the original single sand body division in the study area, and Level 7 is the goal of this research. Comprehensive logging data and physical property parameters redivide the study area $C6_1^2$ based on the original level 6 and 4 single sand layers into level 7 which consists of six single sand body units:$C6_1^{2\text{-}1\text{-}1}$, $C6_1^{2\text{-}2\text{-}1}$, $C6_1^{2\text{-}2\text{-}2}$, $C6_1^{2\text{-}3\text{-}1}$, $C6_1^{2\text{-}3\text{-}2}$, $C6_1^{2\text{-}4\text{-}1}$.

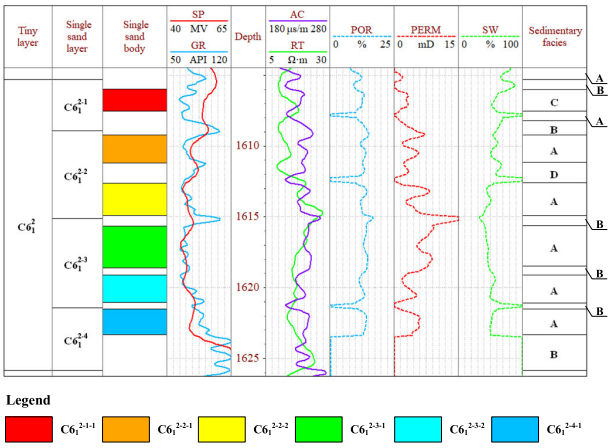

FIGURE 4. The single well phase analysis diagram of *L31-34* well. Tiny layer $C6_1^2$ contains A, B, C, and D sedimentary facies. A is the distributary, B is the interdistributary, C is the mouth bar and D is the distributary channel flank.

parameters and porosity (POR), permeability (PERM), water saturation (SW) reservoir physical parameters and depth data and sedimentary microfacies category. Some logging data and identified results are shown in Table 1.

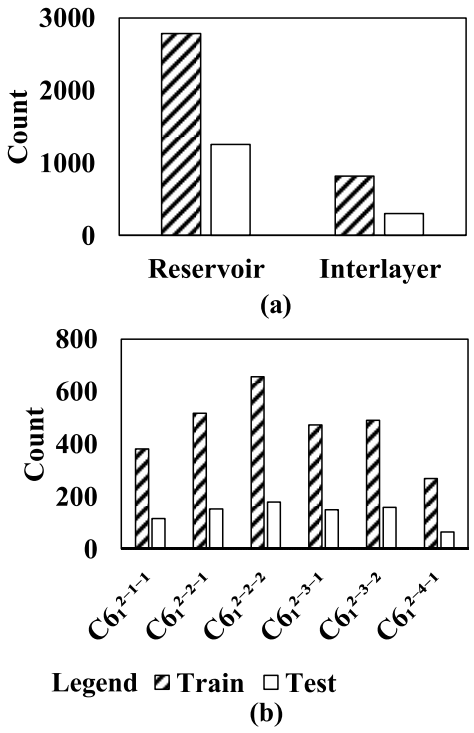

FIGURE 5. Histogram of the distribution training data sample. (a) is the distribution of experimental data for interlayer removal, and (b) is the distribution of experimental data for single sand body identification.

We selected 16 wells as the training dataset for model training, and the remaining 4 wells were used to evaluate the identification effect of the model. In the experiment, the distribution of experimental datasets that correspond to the two stages of the insulating layer and interlayer removal and single sand body identification are shown in Figure 5.

B. DATA SELECTION OF THE INSULATING LAYER AND THE INTERLAYER REMOVAL EXPERIMENT

Based on the results of manual logging interpretation data, we selected GR and SP curves that were more sensitive to changes in shale content, and RT and AC curves that were

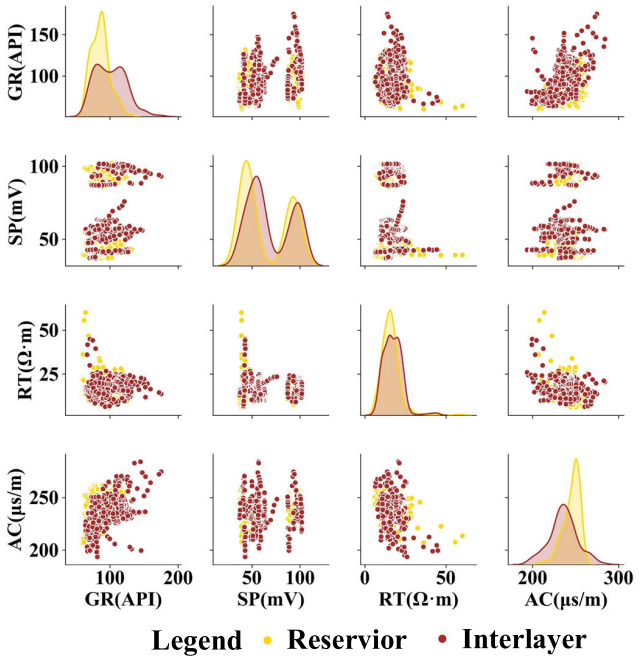

FIGURE 6. Intersection diagram of logging curve response characteristics and logging interpretation.

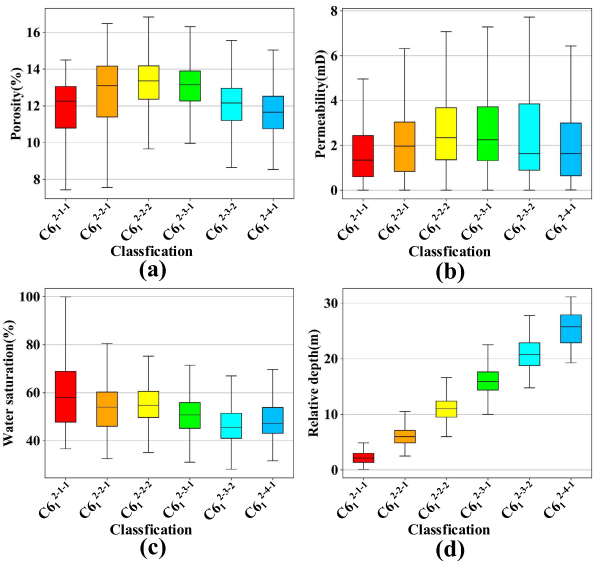

FIGURE 7. Relationship between single sand body classification and petrophysical property and relative depth. (a), (b) and (c) are the relationship between the physical property parameters of a single sand body and the single sand body category, and (d) is the relationship between the relative depth and the single sand body category.

closely related to lithological changes. The intersection analysis of these 4 types of logging parameters for this block is shown in Figure 6. It shows the relationship between the four types of logging parameters corresponding to reservoirs, barriers and interlayers. It can be seen in Figure 6 that the four types of parameters have a strong correlation with the identification of reservoirs. Therefore, GR was selected. The four logging curve data of, SP, RT and AC were used as the basis for the removal of clamps and barriers.

C. DATA SELECTION OF THE SINGLE SAND BODY IDENTIFICATION EXPERIMENT

In the process of workers identifying a single sand body, when the logging curve does not change significantly, the elevation and thickness of the single sand body are relatively stable within a certain range, and the single sand body can be divided by the ''equal elevation method''. In the single sand body identification stage, the physical parameters of the reservoir can be added as a feature vector, and a new single sand body identification model can be established and trained with the aid of the elevation of each single sand body. After completing the division of the single sand body from the barrier and the interlayer, the data section that had nothing to do with the single sand body was excluded, and the top of $C6_1^2$ was used as the base level to obtain the relative elevations of all sampling points within the range of the well. Figure 7 shows the relationship between single sand body classification and reservoir physical parameters and relative elevation.

The type of single sand body has a strong correlation with the distribution of physical parameters and relative elevation. Therefore, the difference in physical properties between single sand bodies is the main basis, and the thickness and elevation changes are the secondary basis. Based on the four types of logging curve parameters GR, SP, RT, and AC, we incorporated POR, PERM, SW and relative depth into the characteristic variables to establish and train a new single sand body identification model. The output of the model was six types of single sand bodies: C612-1-1, C612-2-1, C612-2-2, C612-3-1, C612-3-2, C612-4-1.

V. MODEL TRAINING AND ANALYSIS

A. PARAMETER SETTINGS

To verify the identification effect of ABC-XGBoost, we inputted the same data in the experiment and compared the identification effect with the three models of SVM, RF, one-dimensional convolutional neural network (1D-CNN) and XGBoost. The parameter adjustments of the SVM, RF, and XGBoost were optimized by ABC. We repeated experiments according to reference [31] and set the maximum number of iterations of ABC to 400, the total number of bees to 50, and the maximum number of individual updates to 30. The traditional XGBoost used GS to obtain an optimized parameter combination. SVM, RF, XGBoost initial parameter setting and optimization range are shown in Table 2.

B. EVALUATION INDEX

In this experiment, we used a confusion matrix to evaluate the performance of the model. By revealing the relationship between the actual category and the predicted category of the sample data, it obtained four sets of data including true positive examples (*TP*), false positive examples (*FP*), true negative examples (*TN*) and false negative examples (*FN*) [32].

TABLE 2. Initial value setting and optimization range.

Model	Parameters	Initial value	Search range
SVM	Penalty coefficient(c)	1	0.001~1,000
	Kernel function smoothing factor(γ)	1	0.001~1,000
RF	Number of classification trees (k)	200	100~1,000
	Learning rate(η)	0.3	0.001~1
	max_depth	6	0.001~1
	min_sample_split	1	3~10
	min_sample_leaf	1	1~10
XGBoost	Number of classification trees (k)	200	100~1,000
	Learning rate(η)	0.3	0.001~1
	Regular coefficient(λ)	0.1	0~100
	max_depth	6	3~10
	min_child_weight	1	0~100
	gamma	0	0~1
	subsample	0.8	0~1
	colcample_bytree	0.8	0~1

TABLE 3. Confusion matrix of classification results.

Real result	Predictive result	
	Positive example	Negative example
Positive example	*TP*	*FN*
Negative example	*FP*	*TN*

The distribution of correct and incorrect in each category is shown in Table 3.

In the evaluation of the binary classification problem of the insulating layer and interlayer elimination, the confusion matrix selects the receiver operating characteristic curve (ROC) that simply and intuitively reflects the classification performance. The method is based on the results of the classification model prediction. The ordinate of the curve is the true positive rate (TPR) and the abscissa is the false positive rate (FPR). The two calculation formulas are Eq. (15) and Eq. (16) [32].

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{15}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}} \tag{16}$$

Therefore, the closer the ROC curve is to the upper left corner, the better the model effect. The area under the curve (AUC) can intuitively and accurately evaluate the performance of the classifier. The larger the value is, the better the model effect. When evaluating the multiclassification problem of single sand body identification, the precision, recall, and *F1* values are calculated to evaluate the accuracy of the model based on introducing the confusion matrix. The calculation formulas for the three evaluation indicators are shown in Eq. (17), Eq. (18), and Eq. (19) [32].

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{17}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{18}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{19}$$

C. MODEL TRAINING

The experimental research environment was a CentOS7, an Intel Core i5-8400 processor, and an NVIDIA RTX2080Ti GPU. The programming language was Python 3.6.6, and various machine learning models were constructed using sklearn library. Table 4 shows the optimized parameters of each model in the insulating layer and interlayer elimination stages, which are quite different from the initial values. This proves the necessity of parameter tuning.

Figure 8 shows the optimization process of ABC. In the model optimized by the ABC, the fitness value significantly increased. The ABC optimized XGBoost has the maximum fitness value in both experiments. This shows that the model has the strongest identification ability and stability in different scales of the insulating layer, interbed removal and single sand body identification and classification tasks based on logging data.

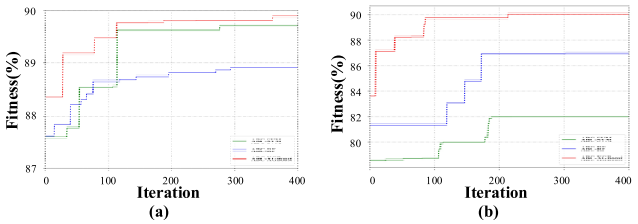

FIGURE 8. The optimization process of the ABC algorithm. (a) is the fitness of the insulating layer and interlayer removal experiment, (b) is the fitness of the single sand body identification experiment.

D. EXPERIMENTAL RESULTS AND ANALYSIS

Figure 9 shows the ROC curves and AUC values of the five models on the test set for insulating layer and interlayer elimination experiments. ABC-XGBoost has the best results in the elimination and interlayer experiments, followed by traditional GS-XGBoost and SVM. RF has relatively poor elimination effects.

The second stage of the single sand body identification experiment was carried out after the insulating layer and interlayer removal was completed, and the confusion matrix of the identification result is shown in Figure 10. It can be seen in Figure 10 that the SVM performs relatively poorly in high-precision identification tasks. The RF and traditional XGBoost identification effects are similar and better than the SVM. ABC-XGBoost has the best effect. Misidentification of various models is mainly concentrated on the single sand bodies of $C6_1^{2\text{-}2\text{-}2}$, $C6_1^{2\text{-}3\text{-}1}$, and $C6_1^{2\text{-}3\text{-}2}$ in the middle and lower parts of the $C6_1^2$ oil-bearing layer. The reason is that the above single sand bodies mostly have superimposed contact in the vertical direction, and the later formed single sand bodies have no obvious erosion in the early stage. Therefore, there is no obvious boundary mark and sedimentary discontinuity in the strata. $C6_1^{2\text{-}4\text{-}1}$ has the lowest degree of sand body development and the smallest proportion of data. The identification rate of $C6_1^{2\text{-}4\text{-}1}$ by SVM and RF is lower than that of the XGBoost model. This shows that SVM and SF are

TABLE 4. Optimization results of experimental parameters at different stages.

EXPERIMENT	MODEL	PARAMETERS AFTER OPTIMIZATION
INSULATING LAYER AND INTERLAYER REMOVAL EXPERIMENT	ABC-SVM	C=0.1, Γ=0.15
	ABC-RF	K=168, MAX_DEPTH=9, MIN_SAMPLE_SPLIT=5, MIN_SAMPLE_LEAF=3
	GS-XGBOOST	K=200, H=0.05, Λ=0.1, MAX_DEPTH=3, GAMMA=0.5, SUBSAMPLE =0.80, COLCAMPLE_BYTREE =0.70
	ABC-XGBOOST	K=159, H=0.01, Λ=0, MAX_DEPTH=5, GAMMA=1.9, SUBSAMPLE =0.73, COLCAMPLE_BYTREE =0.78
SINGLE SAND BODY IDENTIFICATION EXPERIMENT	ABC-SVM	C=10,Γ=0.3
	ABC-RF	K=414, MAX_DEPTH=7, MIN_SAMPLE_SPLIT=9, MIN_SAMPLE_LEAF=4
	GS-XGBOOST	K=400, H=0.005, Λ=0.1, MAX_DEPTH=6, GAMMA=0.4, SUBSAMPLE =0.75, COLCAMPLE_BYTREE =0.85
	ABC-XGBOOST	K=350, H=0.03, Λ=0.1, MAX_DEPTH=7, GAMMA=0.1, SUBSAMPLE =0.90, COLCAMPLE_BYTREE =0.56

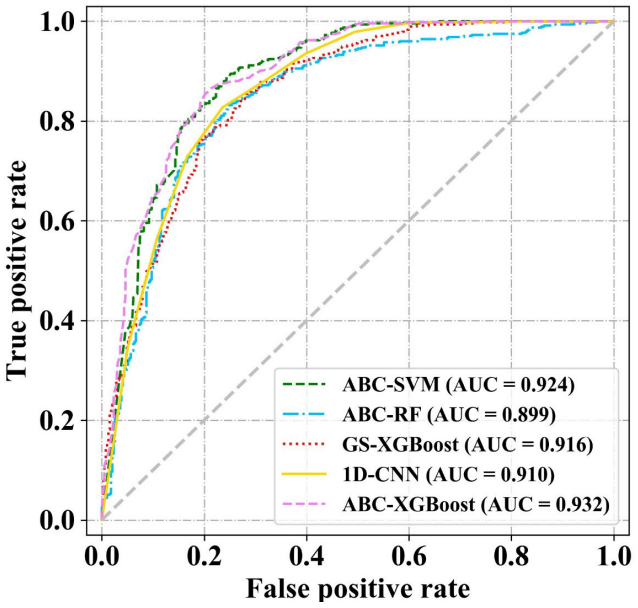

FIGURE 9. ROC curves of five models in the reservoir division. Each column in the confusion matrix represents the predicted category of the sample type, and each row represents the true attribution of the sample. The data on the diagonal is the number of correct classifications for each reservoir type.

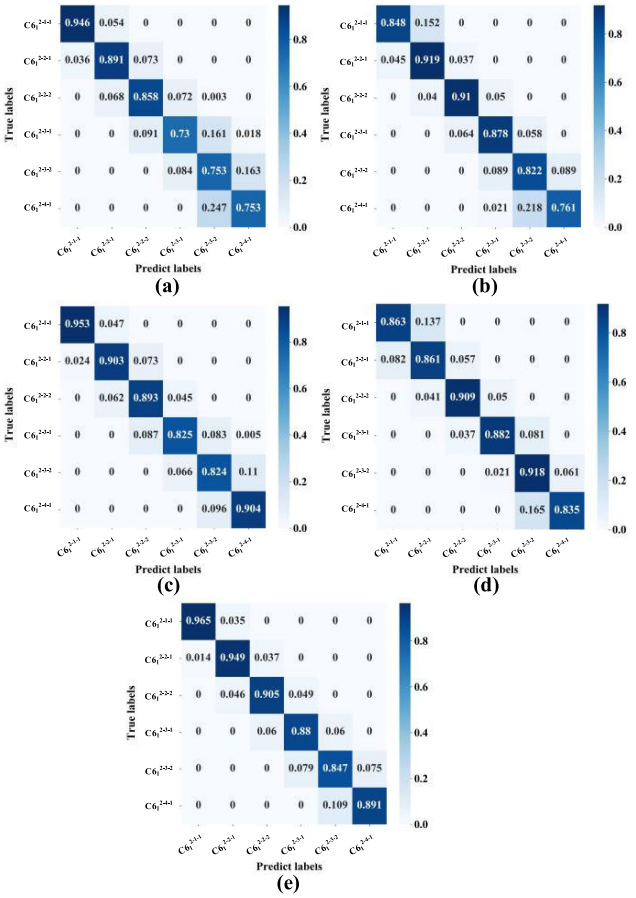

FIGURE 10. Confusion matrix of five models in single sand body identification. (a), (b), (c), (d), and (e) are ABC-SVM, ABC-RF, GS-XGBoost, 1D-CNN, ABC-XGBoost.

greatly affected by the number of samples, and it also reflects that XGBoost has higher stability in the case of unbalanced samples.

Table 5 shows the accuracy rate, recall rate, *F1* value and comprehensive time-consuming of each model. ABC-XGBoost has the highest accuracy rate of 0.907 and an *F1* value of 0.906, which is 3% higher in accuracy and generalization than the XGBoost model that uses the grid search method to adjust parameters. Generally, RF and XGBoost using EL have more advantages than SVM in dealing with multiclassification problems. The single sand body identification effect of the ABC-XGBoost model is better than the other three models. ABC-XGBoost and RF require less time to participate in model training, and SVM takes slightly longer. Compared with XGBoost that uses grid search to adjust parameters, ABC-XGBoost is significantly more efficient.

VI. PRACTICAL APPLICATION OF THE MODEL

To verify the actual application effect of the ABC-XGBoost model, which was applied to wells of *LJ 36-361* and

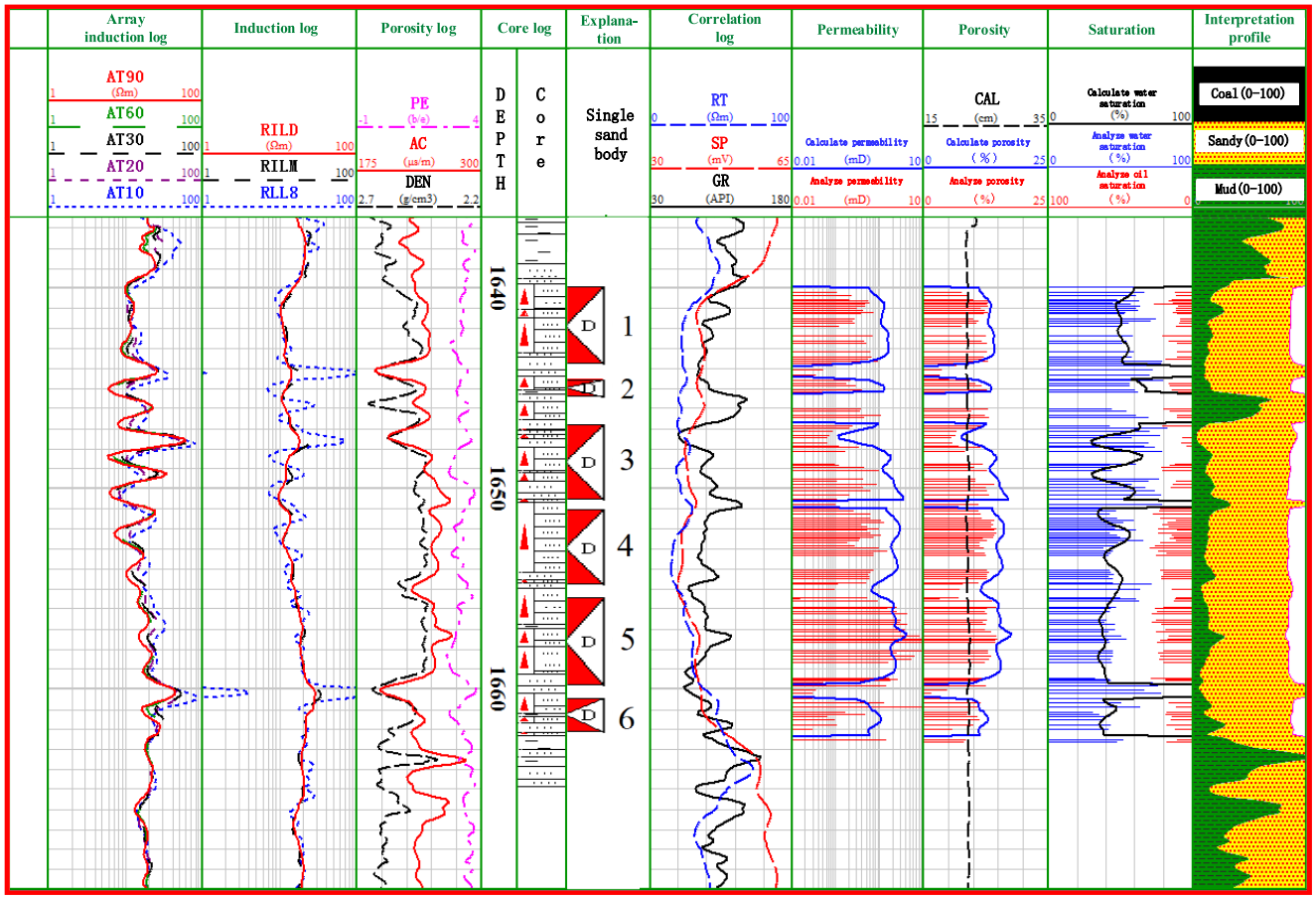

FIGURE 11. Artificial identification result of LJ 36-361 single sand body.

TABLE 5. Results of various single sand body identification methods.

Algorithm	Precision	Recall	F1	Time
ABC-SVM	0.835	0.823	0.826	6m1s
ABC-RF	0.879	0.874	0.875	4m39s
GS-XGBoost	0.878	0.878	0.878	18m52s
1D-CNN	0.880	0.880	0.880	12m47s
ABC-XGBoost	0.907	0.906	0.906	4m21s

LJ 34-365 in the study area, we used the other five models to compare and analyze the identification results of ABC-XGBoost. Taking *LJ 36-361* as an example, this well not only does conventional logging work but also has complete coring logging and other data, as shown in Fig. 11. We used core wells as the reference standard for model identification results.

A. APPLICATION OF PARTITION AND INTERLAYER REJECTION MODEL

The experimental results of the insulating layer and interlayer removal of the five models are shown in Figure 12 and Table 6. We selected two core wells *LJ 36-361* and *LJ 38-365* as application objects. Based on the core data, we divided wells *LJ 36-361* and *LJ 38-365* into five interlayers: I, II, III, IV and V. The location of the top and bottom interface of the interlayer is shown in Figure 12 where the red line is located. In Table 6, we counted the top and bottom interface and thickness errors of the five models, where "\" denotes that the model did not identify the interlayer. The data show that the maximum identification error of ABC-XGBoost is less than 1 *m*, and the performance is better than the other four models. The error in some layers of ABC-XGBoost is larger than that of the other four models. For example, in *LJ 38-385* Interlayer I, the thickness error is 0.625 m. The errors of other models are less than or equal to 0.625 *m*. However, the rest of the models caused misjudgments on the stratum above the top interface. A similar situation also appeared in the LJ 38-385 interlayer III. We found that the reason is the similar lithology above and below the interface. For example, LJ 38-385 Interlayer I is mudstone with a small amount of sandstone, and the top interface is siltstone and argillaceous siltstone. The similar lithology of the upper and lower interface leads to small changes in the logging curve. In addition to errors, each model has unidentified situations. For example, the thickness of LJ 38-385 interlayer V (Figure 2(b)) is less

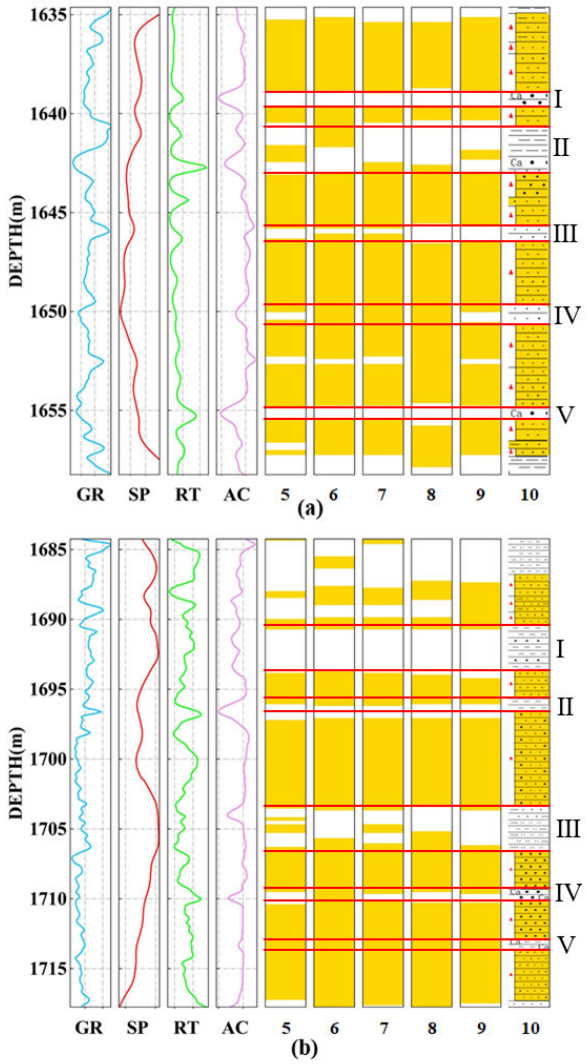

FIGURE 12. Experimental results of removing insulating layers and interlayers. (a) is the application result of *LJ* 36-361, (b) is the application result of *LJ* 38-365. The 4 traces on the left are the logging response values of the curve selected in the experiment that are eliminated with depth. Tracks 5 and 6 are the identification results by SVM and RF respectively. Track 7 is the classification result of the GS-XGBoost model using the grid search method to adjust the parameters. Track 8 is the classification result of the 1D-CNN. Track 9 is the reservoir conclusion of ABC-XGBoost classification proposed in this paper. Track 10 is the identification conclusion obtained by the experts based on the mud logging core. The depth section filled with color in the figure is the reservoir, and the unfilled depth section is the interlayer and interlayer.

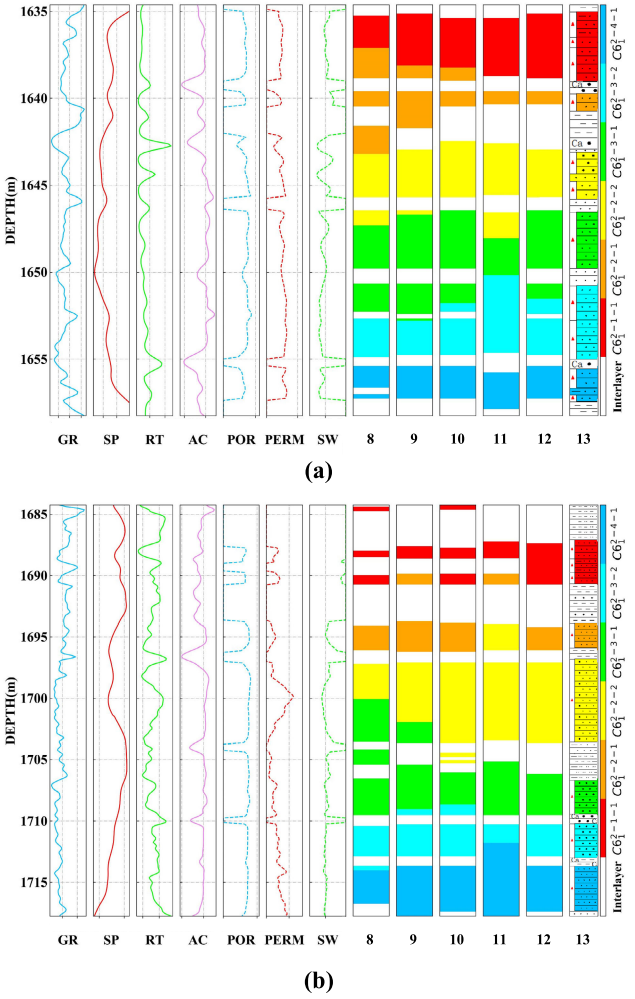

FIGURE 13. Single sand body identification results in 2 core wells. (a) is the application result of *LJ* 36-361, (b) is the application result of *LJ 38-365*. Tracks 1-7 are for 4 types of logging response parameters and 3 types of reservoir physical parameter values. Tracks 8-12 are the results of intelligently identifying a single sand body based on the logging data of two core wells for five models.

than 1 m from the core data. It is a very thin layer, which makes it difficult to identify the model. Therefore, the small change in lithology of the interface and the thin layer will make it difficult for the model to identify the interlayer. The performance of the ABC-XGBoost model is stable, and even in the extreme situations mentioned above, there are very few misidentifications (Figure 12).

In general, we used quantitative identification to better identify most of the reservoirs, barriers and interlayers in unknown wells, which is more sensitive to logging data changes. It can identify some extremely thin interlayers. However, there are limitations in the resolution of some layers that are easy to confuse and have less obvious characteristics. Comparing the five intelligent models, the identification effect of XGBoost is slightly better than that of SVM and RF, and it has certain advantages in the use of logging data to identify reservoirs.

B. APPLICATION OF SINGLE SAND BODY IDENTIFICATION MODEL

After completing the first phase of the test, we eliminated redundant data, leaving the reservoir data consisting of a single sand body. We used them as input data to carry out the second stage of single sand body identification experiment. Figure 13 and Table 7 shows the single sand body identification results of the five models in this subsection.

TABLE 6. Insulating layer and interlayer experimental error.

Identification	Error category	ABC-SVM(*m*)		ABC-RF(*m*)		GS-XGBoost(*m*)		1D-CNN(*m*)		ABC-XGBoost(*m*)	
		LJ 36-361	LJ 38-365	LJ 36-361	LJ 38-365	LJ 36-361	LJ 38-365	LJ 36-361	LJ 38-365	LJ 36-361	LJ 38-365
InterlayerI	Top error	0.125	0.250	0.125	0.250	0.000	0.250	0.250	0.250	0.125	0.250
	Bottom error	0.000	0.250	0.000	0.125	0.000	0.250	0.000	0.375	0.000	0.375
	Thickness error	0.125	0.500	0.125	0.375	0.000	0.500	0.250	0.625	0.125	0.625
InterlayerII	Top error	0.125	0.375	1.125	0.500	0.125	0.500	0.250	0.250	0.250	0.250
	Bottom error	1.500	0.375	0.125	0.375	1.625	0.375	0.500	0.375	0.500	0.375
	Thickness error	1.675	0.750	1.250	0.875	1.500	0.825	0.250	0.625	0.750	0.625
InterlayerIII	Top error	0.250	0.250	0.125	0.375	0.125	0.375	0.125	0.000	0.000	0.375
	Bottom error	0.125	0.375	0.625	0.875	0.500	0.500	0.125	1.375	0.000	0.375
	Thickness error	0.375	0.625	0.750	1.250	0.625	0.875	0.250	1.375	0.000	0.750
InterlayerIV	Top error	0.375	0.250	\	\	\	0.375	\	0.125	0.375	0.375
	Bottom error	0.250	0.375	\	\	\	0.000	\	0.250	0.000	0.250
	Thickness error	0.625	0.625	\	\	\	0.375	\	0.375	0.375	0.625
InterlayerV	Top error	0.000	\	0.250	\	0.250	\	0.250	\	0.125	\
	Bottom error	0.000	\	0.000	\	0.000	\	0.000\	\	0.000	\
	Thickness error	0.000	\	0.250	\	0.250	\	0.250	\	0.125	\

TABLE 7. Single sand body identification experiment error.

Identification	Error category	ABC-SVM(m)		ABC-RF(m)		GS-XGBoost(m)		1D-CNN(m)		ABC-XGBoost(m)	
		LJ 36-361	LJ 38-365	LJ 36-361	LJ 38-365	LJ 36-361	LJ 38-365	LJ 36-361	LJ 38-365	LJ 36-361	LJ 38-365
$C6_1^{2\text{-}1\text{-}1}$	Top error	0.250	0.750	0.125	0	0.375	0.000	0.375	0	0.125	0.125
	Bottom error	1.875	0.000	0.850	2.125	0.125	0.000	0.250	2.125	0.125	0.000
	Thickness error	2.150	0.750	1.000	2.125	1.125	0.000	0.625	2.125	0.25	0.125
$C6_1^{2\text{-}2\text{-}1}$	Top error	0.000	0.000	0.000	0.375	0.000	0.250	0.000	\	0.000	0.125
	Bottom error	0.125	0.000	0.125	0.125	0.125	0.000	0.125	\	0.125	0.000
	Thickness error	0.125	0.000	0.125	0.250	0.125	0.250	0.125	\	0.125	0.125
$C6_1^{2\text{-}2\text{-}2}$	Top error	0.125	0.125	0.125	0.000	1.000	0.000	0.500	0.000	0.125	0.000
	Bottom error	0.000	3.500	0.000	1.625	0.000	1.750	0.125	0.125	0.000	0.125
	Thickness error	0.125	3.375	0.125	1.625	1.000	1.750	0.375	0.125	0.125	0.125
$C6_1^{2\text{-}3\text{-}1}$	Top error	0.850	0.000	0.250	1.125	0.000	0.500	1.625	1.375	0.000	0.375
	Bottom error	0.000	0.000	0.000	0.500	0.000	0.875	0.000	0.000	0.000	0.000
	Thickness error	0.850	0.000	0.250	0.625	0.000	0.375	1.625	1.375	0.000	0.375
$C6_1^{2\text{-}3\text{-}2}$	Top error	1.875	0.000	2.125	0.000	1.125	0.000	1.000	0.000	0.875	0.000
	Bottom error	0.000	0.000	0.000	0.000	0.000	0.000	0.250	1.125	0.000	0.000
	Thickness error	1.875	0.000	2.125	0.000	1.125	0.000	0.750	1.125	0.875	0.000
$C6_1^{2\text{-}4\text{-}1}$	Top error	0.000	0.250	0.000	0.000	0.000	0.000	0.375	1.875	0.000	0.000
	Bottom error	0.000	0.625	0.000	0.375	0.000	0.000	0.625	0.375	0.000	0.000
	Thickness error	0.000	0.375	0.000	0.375	0.000	0.000	0.250	1.500	0.000	0.000

The identification effect of the five models is significantly higher than that of the interlayer experiment. Combined with the analysis of the results of the interlayer removal experiment, there are three reasons. First, in the first phase of the experiment, we eliminated the interlayer data, which reduced the data redundancy and simultaneously reduced the interference of the interlayer. Second, the thickness of a single sand body is generally higher than that of the interlayer, and the model can better identify it. For example, the total thickness of a single sand body of *LJ 36-361* is 17 *m*, the total thickness of the interlayer is 6.625 *m*, and the single sand body accounts for 71.96%. Finally, the lithology difference between the single sand bodies is obvious, the shape difference of the corresponding logging curves is prominent, and the data characteristics change obviously. This can also improve the identification effect of the model.

For a single model, the ABC-SVM identification effect is poor. For example, the $C6_1^{2\text{-}2\text{-}1}$ in *LJ 36-361* has a thickness error of 2.15 *m*. This has exceeded the tolerance for single sand body identification. At the same time, there are many misidentifications, such as $C6_1^{2\text{-}2\text{-}1}$ in *LJ 36-361*, $C6_1^{2\text{-}1\text{-}1}$ in *LJ 38-365*, and $C6_1^{2\text{-}4\text{-}1}$ in *LJ 38-365*, et *al.*. The reason is that ABC-SVM is good at dealing with dichotomous classification problems, and single sand body recognition belongs to multiclassifications, which leads to a decrease in the performance of the ABC-SVM model. In the single sand body identification experiment based on the ABC-XGBoost model, the maximum error of thickness is 0.75 *m* (*LJ 36-361*, $C6_1^{2\text{-}3\text{-}2}$). The recognition performance is better than the other four models. At the same time, ABC-XGBoost is rarely misidentified.

In general, the ABC-XGBoost algorithm proposed in this paper has higher accuracy in the boundary division of a single sand body and can better identify the interlayer in the composite sand body, and has the highest agreement with the actual classification. The absolute error is less than 1 *m*, which meets the actual application requirements in the later development and production.

VII. CONCLUSION

This paper proposed an ABC-XGBoost model for single sand body identification that consists of parameter optimization and intelligent identification. When faced with artificial single sand body identification, due to different geological backgrounds and personal understandings, the identification results were different. We adopted the XGBoost model driven by pure data. In the face of the fact that XGBoost had many parameters in single sand body identification and it was difficult to adjust the parameters, we used ABC to automatically adjust the XGBoost parameters. This avoided the problem of different identification results caused by a large number of adjustment parameters and low efficiency, which greatly improved the work efficiency and identification effect. Finally, we divided the experiment into two phases: insulating layer, interlayer removal and single sand body identification. This not only eliminated the interference of barriers and interlayers but also reduced data redundancy, and maximized the efficiency of single sand body identification. To further evaluate the effectiveness of the model, we compared the current mainstream five models and verified them with two core wells. The verification results show that ABC-XGBoost has the highest identification efficiency and accuracy, and the identification effect is better than the traditional single-model SVM and RF. The absolute error of the top and bottom interface of a single sand body is within 1 *m*, which can be used in the later production of oil fields.

REFERENCES

[1] Y. Zhang, L.-W. Qiu, B.-L. Yang, J. Li, and Y.-L. Wang, "Effects of water level fluctuation on sedimentary characteristics and reservoir architecture of a lake, river dominated delta," *J. Central South Univ.*, vol. 23, no. 11, pp. 2958–2971, Nov. 2016, doi: 10.1007/s11771-016-3360-1.

[2] X. F. Shang, T. Z. Duan, J. G. Hou, and Y. Li, "Spatial configuration of sand and mud in the lacustrine nearshore sand bar deposits and its geological implications," *Petroleum Explor. Develop.*, vol. 46, no. 5, pp. 954–968, Oct. 2019, doi: 10.1016/S1876-3804(19)60252-1.

[3] J. Sheng, J. C. Liu, Q. S. Qi, X. Li, W. L. Tan, X. Chai, B. Guo, X. J. Yang, D. B. Yi, S. S. Lu, Y. A. Li, and C. Y. Zhang, "Fine characterization of sand body in the front of the fluvial delta: Taking the VII oil group of N_2^1 reservoir in Gasikule oilfield as an example," *Geofluids*, vol. 2021, no. 4, pp. 1–14, Feb. 2021, doi: 10.1155/2021/6671855.

[4] Z. W. Si, F. W. Lin, D. F. Liu, X. S. Kong, Q. L. Yin, and D. Z. Zhuang, "Research on well logging evaluation method of igneous reservoir in Nanpu No. 5 structure," *Energy Sources, A, Recovery, Utilization, Environ. Effects*, vol. 2020, pp. 1–13, Jul. 2020, doi: 10.1080/15567036.2020.1798565.

[5] L. Han, L. Fuqiang, D. Zheng, and X. Weixu, "A lithology identification method for continental shale oil reservoir based on BP neural network," *J. Geophys. Eng.*, vol. 15, no. 3, pp. 895–908, Jun. 2018, doi: 10.1088/1742-2140/aaa4db.

[6] A. Amosu, M. Imsalem, and Y. F. Sun, "Effective machine learning identification of TOC-rich zones in the Eagle Ford shale," *J. Appl. Geophys.*, vol. 188, no. 9, pp. 1–9, May 2021, doi: 10.1016/j.jappgeo.2021.104311.

[7] X. Chen, S. Y. Xu, S. M. Li, H. He, Y. M. Han, and X. Y. Qu, "Identification of architectural elements based on SVM with PCA: A case study of sandy braided river reservoir in the Lamadian Oilfield, Songliao Basin, NE China," *J. Petroleum Sci. Eng.*, vol. 198, pp. 1–11, May 2021, doi: 10.1016/j.petrol.2020.108247.

[8] Y. L. Ao, L. P. Zhu, S. Guo, and Z. G. Yang, "Probabilistic logging lithology characterization with random forest probability estimation," *Comput. Geosci.*, vol. 144, pp. 1–10, Nov. 2020, doi: 10.1016/j.cageo.2020.104556.

[9] K. Zhou, J. Zhang, Y. Ren, Z. Huang, and L. Zhao, "A gradient boosting decision tree algorithm combining synthetic minority oversampling technique for lithology identification," *Geophysics*, vol. 85, no. 4, pp. WA147–WA158, Jul. 2020, doi: 10.1190/GEO2019-0429.1.

[10] L. X. Zhao, J. S. Liu, Y. X. Yao, K. Zhong, J. Q. Ma, C. F. Zou, Y. Y. Chen, X. W. Fu, X. J. Zhu, W. L. Zhu, and J. H. Geng, "Quantitative seismic characterization of source rocks in lacustrine depositional setting using the random forest method: An example from the Changjiang sag in East China sea basin," *Chin. J. Geophys.-Chin. Ed.*, vol. 64, no. 2, pp. 700–715, Feb. 2021, doi: 10.6038/cjg2021O0123.

[11] J. J. Liu and J. C. Liu, "An intelligent approach for reservoir quality evaluation in tight sandstone reservoir using gradient boosting decision tree algorithm—A case study of the Yanchang formation, mid-eastern Ordos basin, China," *Mar. Petroleum Geol.*, vol. 126, pp. 1–15, Apr. 2021, doi: 10.1016/j.marpetgeo.2021.104939.

[12] Y. Xie, C. Zhu, R. Hu, and Z. Zhu, "A coarse-to-fine approach for intelligent logging lithology identification with extremely randomized trees," *Math. Geosci.*, vol. 53, no. 5, pp. 859–876, Aug. 2020, doi: 10.1007/s11004-020-09885-y.

[13] J. Sun, R. Zhang, M. Chen, Q. Li, Y. Sun, L. Ren, and W. Zhang, "Real-time updating method of local geological model based on logging while drilling process," *Arabian J. Geosci.*, vol. 14, no. 9, pp. 1–17, May 2021, doi: 10.1007/s12517-021-07034-1.

[14] H. Y. Han, Z. Z. Wang, and L. C. Wang, "The application of decision tree method to the identification of formation types in tight gas reservoirs," *Appl. Mech. Mater.*, vols. 170–173, pp. 969–974, May 2012, doi: 10.4028/www.scientific.net/AMM.170-173.969.

[15] T. L. Insua, L. Hamel, K. Moran, L. M. Anderson, and J. M. Webster, "Advanced classification of carbonate sediments based on physical properties," *Sedimentology*, vol. 62, no. 2, pp. 590–606, Feb. 2015, doi: 10.1111/sed.12168.

[16] S. A. Naghibi, H. R. Pourghasemi, and B. Dixon, "GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran," *Environ. Monitor. Assessment*, vol. 188, no. 1, pp. 1–27, Jan. 2016, doi: 10.1007/s10661-015-5049-6.

[17] S. Tewari and U. D. Dwivedi, "Ensemble-based big data analytics of lithofacies for automatic development of petroleum reservoirs," *Comput. Ind. Eng.*, vol. 128, pp. 937–947, Feb. 2019, doi: 10.1016/j.cie.2018.08.018.

[18] P. Wang, X. Chen, B. Wang, J. Li, and H. Dai, "An improved method for lithology identification based on a hidden Markov model and random forests," *Geophysics*, vol. 85, no. 6, pp. IM27–IM36, Nov. 2020, doi: 10.1190/GEO2020-0108.1.

[19] J. Sun, R. Zhang, M. Chen, B. Chen, X. Wang, Q. Li, and L. Ren, "Identification of porosity and permeability while drilling based on machine learning," *Arabian J. Sci. Eng.*, vol. 46, no. 7, pp. 7031–7045, Jul. 2021, doi: 10.1007/s13369-021-05432-x.

[20] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," ISE Lab., Blacksburg, VT, USA, Tech. Rep. TR-2009-14, Oct. 2009.

[21] D. Karaboga and B. Basturk, "On the performance of artificial bee colony (ABC) algorithm," *Appl. Soft Comput.*, vol. 8, no. 1, pp. 687–697, Jan. 2008, doi: 10.1016/j.asoc.2007.05.007.

[22] Z. H. Wang, X. M. Cui, D. B. Yuan, H. Liu, and J. F. Wang, "The research on forest resources change detection based on C5.0 algorithm and neighborhood correlation image analysis," *Key Eng. Mater.*, vol. 500, pp. 701–708, Jan. 2012, doi: 10.4028/www.scientific.net/KEM.500.701.

[23] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. Mohammadian, "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis," *Accident Anal. Prevention*, vol. 136, Mar. 2020, Art. no. 105405, doi: 10.1016/j.aap.2019.105405.

[24] A. Samat, E. Li, W. Wang, S. Liu, C. Lin, and J. Abuduwaili, "Meta-XGBoost for hyperspectral image classification using extended MSER-guided morphological profiles," *Remote Sens.*, vol. 12, no. 12, p. 1973, Jun. 2020, doi: 10.3390/rs12121973.

[25] H. Li, Y. Li, and F. Porikli, "DeepTrack: Learning discriminative feature representations online for robust visual tracking," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1834–1848, Apr. 2015.

[26] W. XingFen, Y. Xiangbin, and M. Yangchun, "Research on user consumption behavior prediction based on improved XGBoost algorithm," in *Proc. IEEE Int. Conf. Big Data, Big Data*, Dec. 2018, pp. 4169–4175, doi: 10.1109/BigData.2018.8622235.

[27] W.-L. Xiang, X.-L. Meng, Y.-Z. Li, R.-C. He, and M.-Q. An, "An improved artificial bee colony algorithm based on the gravity model," *Inf. Sci.*, vol. 429, pp. 49–71, Mar. 2018, doi: 10.1016/j.ins.2017.11.007.

[28] D. Karaboga and B. Gorkemli, "A quick artificial bee colony (qABC) algorithm and its performance on optimization problems," *Appl. Soft Comput.*, vol. 23, no. 1, pp. 227–238, Oct. 2014, doi: 10.1016/j.asoc.2014.06.035.

[29] O. Ercin and R. Coban, "Identification of linear dynamic systems using the artificial bee colony algorithm," *Turkish J. Elect. Eng. Comput. Sci.*, vol. 20, pp. 1175–1188, Oct. 2012, doi: 10.3906/elk-1012-956.

[30] M. S. Kıran and O. Fındık, "A directed artificial bee colony algorithm," *Appl. Soft Comput.*, vol. 26, pp. 454–462, Jan. 2015, doi: 10.1016/j.asoc.2014.10.020.

[31] N. Veček, S.-H. Liu, M. Črepinšek, and M. Mernik, "On the importance of the artificial bee colony control parameter 'limit,'" *Inf. Technol. Control*, vol. 46, no. 4, pp. 566–604, Dec. 2017, doi: 10.5755/j01.itc.46.4.18215.

[32] Z. Wang, S. Lou, S. Liang, and X. Sheng, "Multi-class disturbance events recognition based on EMD and XGBoost in ϕ-OTDR," *IEEE Access*, vol. 8, pp. 63551–63558, 2020, doi: 10.1109/ACCESS.2020.2984022.