Journal of Innovation and Development
ISSN: 2958-4752 | Vol. 5, No. 3, 2023

Research on Intelligent Recognition Technology in Lithology Based on Multi-parameter Fusion of Logging While Drilling

Haibo Liang, Jiaguo Xiong

School of Mechatronic Engineering, Southwest Petroleum University, Chengdu Sichuan 610500, China

Abstract: In oil and gas drilling, timely and accurate identification of formation lithology is an important guarantee of drilling safety. Aiming at the problems of inaccurate identification of lithology in drilling by traditional methods, and low efficiency due to the fact that even modern instruments cannot respond to lithology in real time. the Categorical Boost (CatBoost) model was applied to lithology identification in this study. However, since CatBoost uses more hyperparameters in its modeling, it is difficult to optimize model prediction by manually tuning the parameters. Therefore, the introduction of Kernel Principal Component Analysis (KPCA) extracts fewer and more important features from the original data, eliminates the redundant information contained therein, and combines with Bayesian Optimization (BO) algorithm to optimize the hyperparameters during the training process, thus improving the prediction performance of CatBoost. Two experiments were designed to verify the recognition ability of the model, and the final test results of the model showed that the KPCA-BO-CatBoost model proposed in this paper had the best overall performance, and the lithology recognition accuracy reached more than 90%. The model was effective in identifying the formation lithology, realized real-time lithology identification by combining the parameters of logging while drilling, improved the efficiency and accuracy of lithology identification, and was of great significance in guiding the subsequent drilling work.

Keywords: Lithology Identification; Catboost; Bayesian Optimization; Kernel Principal Component Analysis.

1. Introduction

Lithology identification is a very important component in the drilling process and is directly related to the selection of drilling construction parameters and drilling efficiency [1,2]. Conventional lithology identification was generally based on cores or observations from returned rock chips, and comparative judgments were made based on manual experience, which obviously biases the lithology results. Later, researchers used the crossplot method to identify lithologies, which typically consisted of two or three representative logs and, when combined with logging data, could characterize different types of simple lithologies and establish different lithological criteria depending on the distribution of logging points [3]. Chongwain [4] used multiple sets of logging parameters and cross plots to distinguish between petrography and fluids. However, as more unconventional oil and gas fields were developed, the lithology types were becoming more complex, which can lead to the overlap of complex lithology measurement points within the crossplot, and therefore the method was not conducive to assessing complex lithology. To overcome this drawback, with the development of science and technology, many scholars have devoted themselves to the research of combining lithology identification techniques with machine learning [5,6]. Thus, complex lithology prediction based on machine learning has gradually become a hot spot of research in recent years.

Integrated learning algorithms are widely recognized as excellent predictors, formed by combining multiple weak learners through policies to form strongly supervised models, and have shown great advantages when dealing with large data [7,8], and thus have received widespread attention. The classical gradient boosted decision tree (GBDT) uses a linear combination of basic learners to continuously reduce the residuals generated during the training process through multiple iterations [9], thus achieving the classification purpose. Yu [10] accurately distinguished the lithologic interface between breccia, tuff, and rhyolite using the GBDT model. However, GBDT is difficult to perform effective feature space partitioning when faced with high-dimensional sparse features. XGBoost (extreme gradient boosting tree) improves the objective function based on GBDT by using Taylor's formula for its second-order expansion, while adding regularization to control the complexity of the model and using a special gradient descent method for training, which makes it perform well when dealing with high-dimensional sparse data [11]. Gu [12] introduced a continuous Boltzmann machine and particle swarm optimization XGBoost to achieve good results for the lithology of dense sandstone reservoirs. Kuchin [13] used XGBoost to automatically determine the rock composition along the borehole and ensure the smooth mining of uranium ore. In general, XGBoost showed satisfactory results in lithology identification. CatBoost is further improved for prediction by using a weighted cross-entropy loss function in training, employing the greedy strategy that effectively improves prediction accuracy, applying ordered boosting, optimizing the gradient bias, and using oblivious trees as the base predictor to reduce the risk of overfitting [14].

Many scientific studies have been published on CatBoost, but few studies have been conducted on lithology recognition. In addition, it is worth noting that CatBoost permutes features when training data, and a large number of irrelevant features consume a lot of time, while unoptimized hyperparameters make it difficult for the model to achieve optimal results. Therefore, this study introduces KPCA to extract fewer and more important features [15], KPCA is an extension of PCA

(Principal Component Analysis), PCA mines the most dominant feature components of high-dimensional data through linear transformation, covariance matrix and matrix diagonalization, but PCA is a linear dimensionality reduction method, which is not applicable to nonlinear data analysis, KPCA can overcome the shortcomings of PCA by mapping the original space of samples to high-dimensional feature space mapping, using kernel functions for feature identification in the high-dimensional space, and then using PCA to achieve nonlinear dimensionality reduction of the data, which in turn effectively avoids dimensional disasters and simplifies the data [16]. The predictive performance of the model is further improved by Bayesian optimization of hyperparameters, which is also an adaptive hyperparameter search method like grid search and random search [17,18], but grid search performs slightly worse with more hyperparameters and is very computationally expensive, while random search has randomness by randomly sampling the search area and the results cannot be guaranteed, Bayesian optimization updates the posterior distribution of the objective function by continuously adding sample points on a given objective function, which in short takes into account the information of the previous parameters to better adjust the current parameters. Based on this, a new lithology identification prediction model, KPCA-BO-CatBoost, is proposed in this study, and the model is introduced and validated step by step in the subsequent work.

2. Methodology

2.1. CatBoost

CatBoost is a new machine learning algorithm framework based on GBDT, which constructs a learner along the direction of the steepest gradient at each iteration and optimizes it using gradient descent. The model is defined as follows:

$$F(x,w)=\sum_{t=0}^{E}\alpha_t h_t(x,w_t)=\sum_{t=0}^{E}f_t(x,w_t) \tag{1}$$

Where: $F(\cdot)$ is the output of the entire decision tree; x is the input sample; w is the parameter of the entire decision tree; α_t is the weight of tree t. E is the number of trees; $h_t(\cdot)$ is the output of the t-th decision tree; w_t is the parameter of the t-th decision tree; $f_t(\cdot)$ is the output of the t-th decision tree after weighting.

The parameters of the optimal model are obtained by minimizing the los s function as:

$$(\alpha_i,w_i)=\arg\min\sum_{i=0}^{N}L\big(y_i,F(x_i,w)\big) \tag{2}$$

Where: $L(\cdot)$ is the loss function; y_i is the actual output of sample i; x_i is the input of sample i; N is the number of samples.

CatBoost solves the overfitting problem of GBDT by using the Ordered Boosting method to obtain an unbiased gradient estimation to mitigate the effect of gradient estimation bias. In GBDT, the label average is used as the criterion for node splitting, expressed as follows:

$$\hat{x}_k^i=\frac{\sum_{i=1}^{N}I_{(x_j=x_i)}y_j}{\sum_{j=1}^{N}I_{(x_j=x_i)}} \tag{3}$$

Where: x_k^i is the i-th category feature of the k-th training sample; $\hat{x}_k^i$ is its average value; y_j is the label of the j-th sample; I is the indicator function, that is the two quantities in brackets are equal to 1, otherwise 0.

The disadvantage of this approach is that the features contain more information than the labels, and when the average of the labels is used to represent the features, the problem of conditional bias arises when the structure and distribution of the training and test data are different. CatBoost incorporates prior terms and weighting factors that can reduce the effect of noise and low-frequency category data on the data distribution:

$$\hat{x}_k^i=\frac{\sum_{j=1}^{N}I_{(x_j=x_i)}y_j+ap}{\sum_{j=1}^{N}I_{(x_j=x_j)}+a} \tag{4}$$

Where: p is the added a priori item; a is the weighting factor.

In addition, CatBoost uses an oblivious tree as the base predictor and applies the same partitioning law to the entire layer of the tree in each iteration, with the left and right subtrees being perfectly symmetric and balanced. This not only helps to avoid overfitting, but also speeds up the model operation by encoding the index of each leaf node in the symmetric tree as a binary vector, and the length of the vector is equal to the depth of the tree. catBoost uses the above method to perform a binary conversion operation on the acquired features, and then completes the evaluation of the model by calculating the feature values.

2.2. KPCA

Since CatBoost will arrange and combine features in the training data, in order to eliminate the problems caused by redundant information in the original sample and reduce the training time, KPCA is used to map the original sample space to the high-dimensional feature space, and polynomial, Gaussian kernel, neural network and other kernel functions are used for feature recognition in the high-dimensional space, but the polynomial kernel function needs to calculate the inner product, which may cause overflow and other computational problems, The selection of neural network kernel function to construct the kernel transformation matrix is more demanding on the parameter selection, which is easy to cause pathological kernel matrix, linear kernel function is only a special case of Gaussian kernel function, and Gaussian kernel function has better positivity and has significant advantages in dealing with nonlinear data, so Gaussian kernel function is selected for processing, and finally PCA is used to realize nonlinear dimensionality reduction of data, and then effectively avoid dimensional disasters and extract important features.

First, m sets of high-dimensional data $z_j=\{z_{ij}\}$ are constructed using a single set of lithology samples of number n. Let $\boldsymbol{\tau}(z_j)$ be the mapping matrix of z_j in the feature space, its covariance is:

$$\boldsymbol{\psi}=\frac{1}{m}\sum_{j=1}^{m}\boldsymbol{\tau}(z_j)\boldsymbol{\tau}(z_j)^{\mathrm{T}} \tag{5}$$

Where: i is the serial number of the sample； j is the serial number of the sample group.

To achieve dimensionality reduction, $\boldsymbol{\psi}$ must be diagonalized so that the values on the non-diagonal are minimized. As described above, a Gaussian kernel k is

introduced for nonlinear dimensionality reduction. The Gaussian kernel matrix $\boldsymbol{K}(z_j)$ associated with z_j is constructed using k, the relation between its eigenvalues λ and eigenvectors $\boldsymbol{V}$ is:

$$\lambda_j \boldsymbol{V}_j = \boldsymbol{K}(z_j)\boldsymbol{V}_j \tag{6}$$

Diagonalize $\boldsymbol{K}(z_j)$, solve for λ and $\boldsymbol{V}$, and reorder them from largest to smallest if the sum of the first b eigenvalues and the sum of all eigenvalues satisfy the following conditions:

$$\ell = \left(\sum_{a=1}^{b} \lambda_a / \sum_{j=1}^{m} \lambda_j\right) \geqslant 80\% \tag{7}$$

Then the kernel principal component $\boldsymbol{Z}$ of z_j is:

$$\boldsymbol{Z} = \sum_{j=1}^{m} V_j \boldsymbol{K}(z_j) \tag{8}$$

Finally, different ℓ thresholds can be set according to practical needs so that $\boldsymbol{Z}$ retains enough z_j features.

2.3. Bayesian Optimization.

The CatBoost model has more hyperparameters, and during the training process, different hyperparameters will have different effects on the lithology classification results, and it will take a lot of time to find the optimal one for all parameters. The parameters that have a deeper impact on the model are selected for optimization, so that a balance can be achieved between the efficiency and accuracy of the model.

As described in the introduction section, Bayesian optimization is chosen in this study, which mainly includes two core components: the agent model generally has Gaussian Parzen (GP) and Tree Parzen estimator (TPE), and the collection function generally has maximum Probability Increment (PI) and maximum Expectation Increment (EI). The TPE proxy model, compared to GP, converts the hyperparameter search space from a graph structure to a tree structure and uses nonparametric estimation instead of parametric estimation to achieve better gains in efficiency and accuracy. The EI collection function not only integrates the probability of lifting compared to PI, but also reflects different amounts of lifting, balancing the relationship between depth and width. For this reason, the TPE agent model and the EI collection function are selected to construct the Bayesian optimization algorithm.

First, based on the prior probability distribution of the hyperparameters$p(x \mid y)$, the corresponding distribution of the value-at-risk of the objective function $p(y)$ is estimated using TPE, where $\{x^{(1)}, x^{(2)}, x^{(3)}, \cdots, x^{(k)}\}$ is the hyperparameter and y is the value at risk. The next hyperparameter is selected according to the EI, and the above process is repeated to continuously select the hyperparameters using the posterior distribution of the agent model until the optimal solution is obtained, and the probability distribution of the TPE model is defined as shown in equation (9).

$$p(x \mid y) = \begin{cases} e(x) & \text{if} \quad y < y^* \\ g(x) & \text{if} \quad y \geq y^* \end{cases} \tag{9}$$

Where: $e(x)$ denotes the density formed by the observation $\{x^{(i)}\}$, whose corresponding risk loss $y = f(x^{(i)}) < y^*$; $g(x)$ denotes the density formed using the remaining observations except for $\{x^{(i)}\}$.

The TPE algorithm chooses y^* as some quantile γ of the current observed risk value y, satisfying $p(y < y^*) = \gamma$,without the specific model of $p(y)$, and divides the set of hyperparameters into less risky and riskier parts by $e(x)$ and $g(x)$ of the TPE algorithm. Next, it is further optimized by maximum expectation boosting, and the maximum expectation boosting EI is defined as shown in equation (10).

$$\mathrm{EI}_{y^*}(x) = \int_{-\infty}^{y^*} (y^* - y) p(y \mid x) \mathrm{d}y \tag{10}$$

According to Bayes' theorem, the above equation can be rewritten as：

$$\mathrm{EI}_{y^*}(x) = \int_{-\infty}^{y} (y^* - y) \frac{p(x \mid y) p(y)}{p(x)} \mathrm{d}y \tag{11}$$

According to $p(y < y^*) = \gamma$, and $p(x) = \int_R p(x \mid y) p(y) \mathrm{d}y = \gamma e(x) + (1-\gamma) g(x)$, the above equation can be further obtained as：

$$\mathrm{EI}_{y^*}(x) = \frac{\gamma y^* e(x) - e(x) \int_{-\infty}^{y^*} p(y) \mathrm{d}y}{\gamma e(x) + (1-\gamma) g(x)} \propto \left[\gamma + \frac{g(x)}{e(x)} (1-\gamma)^{-1}\right] \tag{12}$$

From equation (12), it can be seen that in order to obtain the maximum desired boost, the probability of the hyperparameter x in $e(x)$ should be as large as possible, while the probability in $g(x)$ should be as small as possible. By evaluating each hyperparameter x by $g(x)/e(x)$, in each iteration, the algorithm will return the hyperparameter value with the maximum EI, and thus the optimal value will be obtained.

3. Experiments and Results

3.1. Data Source

The stratigraphy below 2800 m in 10 wells in the Bozhong area was selected as the study object, and the corresponding logging data were collected and compiled to cover mainly five lithologies: mudstone, sandstone, glutenite, granite and limestone. Quantitative measurements of elements were made using the logging-with-drilling technique to obtain the percentage values of various elements, including Na, Mg, Al, Si, P, S, Cl, K, Ca, Ba, Ti, Mn, Fe, V, Ni, Sr, and Zr. A total of 17 elements were observed, and the percentage values of some elements were found to be extremely low. In order to reduce the dimensionality of the model input, elements with a percentage content of less than 0.01 were first eliminated, and 10 elements, Na, Mg, Al, Si, S, K, Ca, Ti, Mn, and Fe, were finally filtered out. In addition, the rock drillability index is also a specific manifestation of formation lithology. The nature of mud content and drillability varies among different lithologies and thus directly reflects the differences in rate of Penetration (ROP), torque (Tor), drilling speed (Rpm), drilling pressure (Wob), and fracture pressure gradient (Fra).

After completion, the number of experimental samples was 6383, including 1828 mudstone, 1026 sandstone, 1488 glutenite, 1167 granite and 874 limestone, and some samples are shown in Table 1. The above 10 elements and ROP, Rpm, Wob, Tor and Fra were used as the input of the overall model, and the output was the lithology, and two experiments were designed for comparison and verification.

Table 1. Partial learning samples

Sample No.	Na (%)	Mg (%)	Al (%)	Si (%)	…	Rop Min/m	Wob KN	Rpm r/min	Tor KN·m	Lithology[1]
1	0.85	1.37	5.22	21.08	…	3.75	7.85	60	10.7	0
2	0.15	1.63	6.26	25.92	…	3.67	7.79	60	10.2	0
3	0.41	1.25	6.15	25.12	…	4.91	8.93	60	9.7	0
4	0.18	1.51	6.72	28.03	…	4.03	7.76	60	10.1	0
5	0.15	0.61	11.46	39.12	…	1.88	9.59	68.4	15.6	1
6	0.21	0.84	10.24	37.37	…	3.05	9.52	52.4	13.9	1
7	0.16	0.78	12.31	37.41	…	2.82	9.67	49.5	14.2	1
8	0.14	2.55	13.19	33.18	…	3.27	9.55	50.8	13.2	1
9	2.55	0.68	7.05	36.61	…	0.93	3.9	80	21.9	2
10	2.29	0.3	4.15	36.79	…	1.05	3.69	80	21.8	2
11	2.44	0.39	6.07	34.71	…	1.66	3.02	80	21.8	2
12	2.41	0.43	5.84	35.91	…	1.02	3.6	80	21.5	2
13	2.41	1.36	11.8	35.12	…	12.67	6.7	61	10.8	3
14	2.49	0.85	12.2	35.43	…	10.16	6.69	61	10.5	3
15	2.25	1.05	12.0	33.94	…	11.27	6.8	60.2	11.4	3
16	2.21	0.88	12.4	35.76	…	11.24	7.78	60.6	11.3	3
17	0.88	2.19	7.79	25.37	…	5.09	9.41	50.9	22.7	4
18	0.86	2.27	7.53	25.59	…	4.34	9.16	49.9	23.3	4
19	0.92	2.24	7.39	26.04	…	4.79	9.56	50.2	22.9	4
20	1.15	1.99	7.72	26.61	…	5.43	9.51	51	21.4	4

1. Meaning of each number: 0- Mudstone; 1- Sandstone; 2- Glutenite; 3- Granite; 4- Limestone

3.2. Results

In the first experiment, the effectiveness of integrating KPCA with Bo was verified by comparing four models, CatBoost, Bo-CatBoost, PCA-Bo-CatBoost, and KPCA-Bo-CatBoost. According to the computational process, the samples were divided into training and test sets at a ratio of 7:3, since there is a special label coding pattern in CatBoost, manual lithology coding is not required. The original samples were subjected to PCA and KPCA, respectively, and Fig. 1 shows the cumulative contribution of principal components and eigenvalues after KPCA and PCA treatments, combined with the threshold value set in equation (11), it can be seen that: the cumulative contribution of the first 8 principal components reached 80.84% after PCA dimensionality reduction, and the cumulative contribution of the first 5 principal components reached 85.39% after KPCA dimensionality reduction. Obviously, the lithological feature information extracted by KPCA was more comprehensive and lower dimensional, reflecting the great advantage of KPCA in dealing with nonlinear data. The principal components extracted from the two models were entered into the Bo-CatBoost model for further comparison, and in addition, the CatBoost and Bo-CatBoost models were built separately for the original sample data, and the optimal hyperparameter combination results of each model are shown in Table 2.

Table 2. Model hyperparameter optimization results

Model	CatBoost	PCA-CatBoost	KPCA-CatBoost
Initialization parameter range (left boundary, right boundary)	lr(0.01,1) l2_leaf_reg(1,15) subsample(1,100) depth(1,20) iterations(100,1000)	lr (0.01,1) l2_leaf_reg(1,15) subsample(1,100) depth(1,20) iterations(100,1000)	lr (0.01,1) l2_leaf_reg(1,15) subsample(1,100) depth(1,20) iterations(100,1000)
Bo optimization results	lr =0.06 l2_leaf_reg=6 subsample=66 depth=12 iterations=840	lr =0.08 l2_leaf_reg=4 subsample=55 depth=8 iterations=630	lr =0.09 l2_leaf_reg=4 subsample=50 depth=7 iterations=450

lr indicates the learning rate; l2_leaf_reg is the L2 regularization factor; subsample is the selected sample rate; depth indicates the depth of the tree; iterations indicate the maximum number of trees.

The accuracy of different recognition models on the test set is shown in Fig. 2. The comparison shows that: the direct use of raw data for lithology recognition is less effective, but the classification effect of CatBoost after Bayesian optimization is improved, indicating that the integration of Bo is effective. In addition, the evaluation index of the Bo-CatBoost model after dimensionality reduction by PCA is significantly improved, while KPCA again improves the recognition accuracy of the model on each lithology class compared to PCA, further proving the effectiveness of KPCA integration.

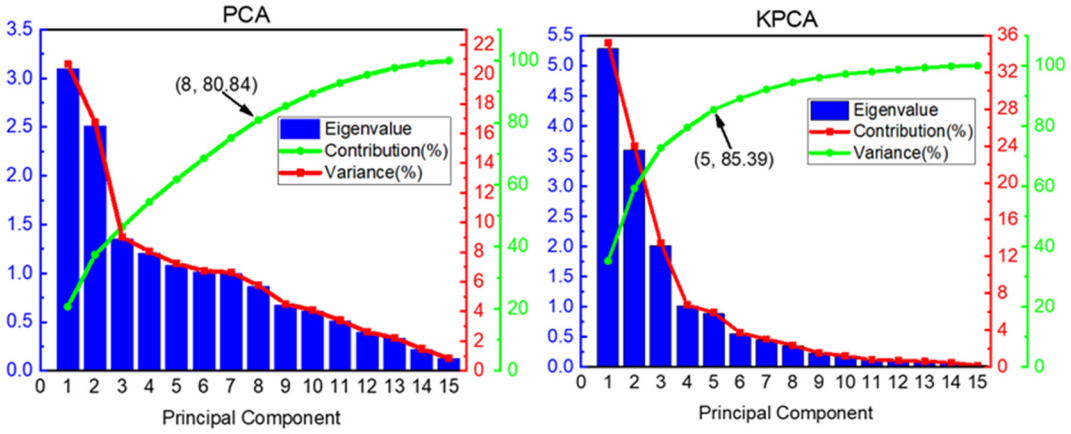

Fig 1. Graph of PCA and KPCA analysis results

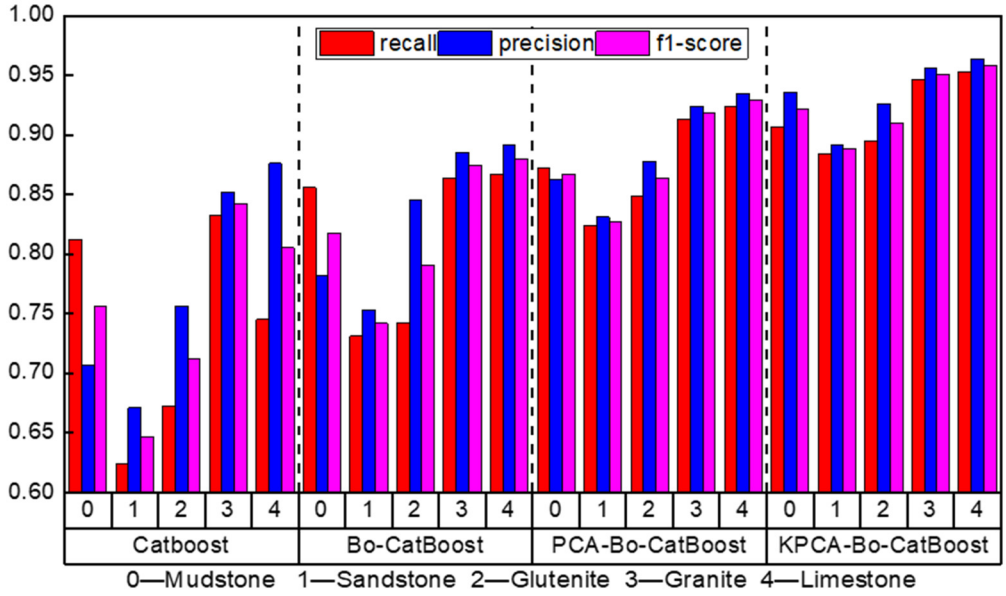

Fig 2. Indicators of each model

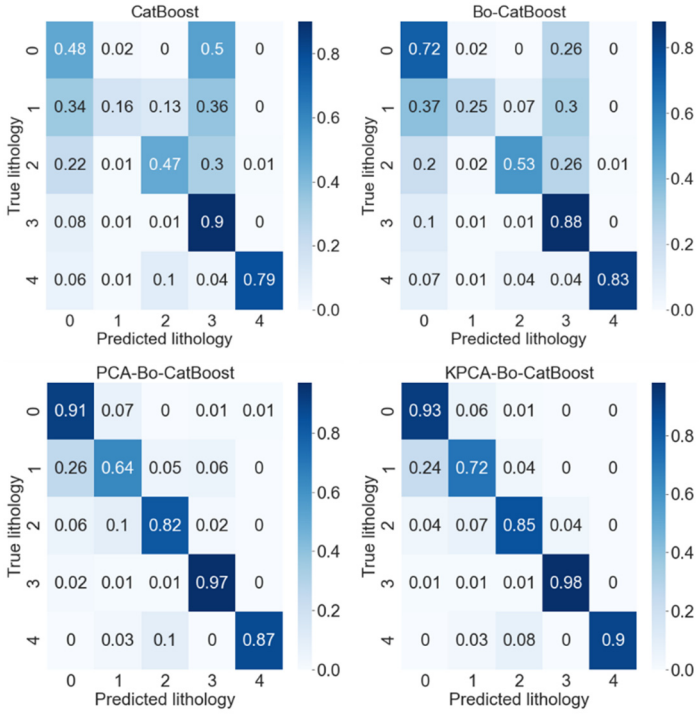

Fig 3. Confusion matrix diagram for each model (0- Mudstone; 1- Sandstone; 2- Glutenite; 3- Granite; 4- Limestone)

Fig. 3 shows the confusion matrix of different classification models after normalization. It can be seen that the performance of CatBoost is weaker in distinguishing the difference between positive and negative samples, and the recognition ability of positive samples is improved to some extent by CatBoost after Bayesian optimization, but both models do not perform well on sandstone and sandy mudstone. After dimensionality reduction by PCA and KPCA, the model prediction is significantly improved with high confidence, with KPCA-Bo-CatBoost performing best. In addition, the results of comparing the true values with the predicted values are shown in Figure 4 for 100 samples randomly selected from the test set, 20 for each lithology category, further confirming the strong predictive ability of the proposed model.

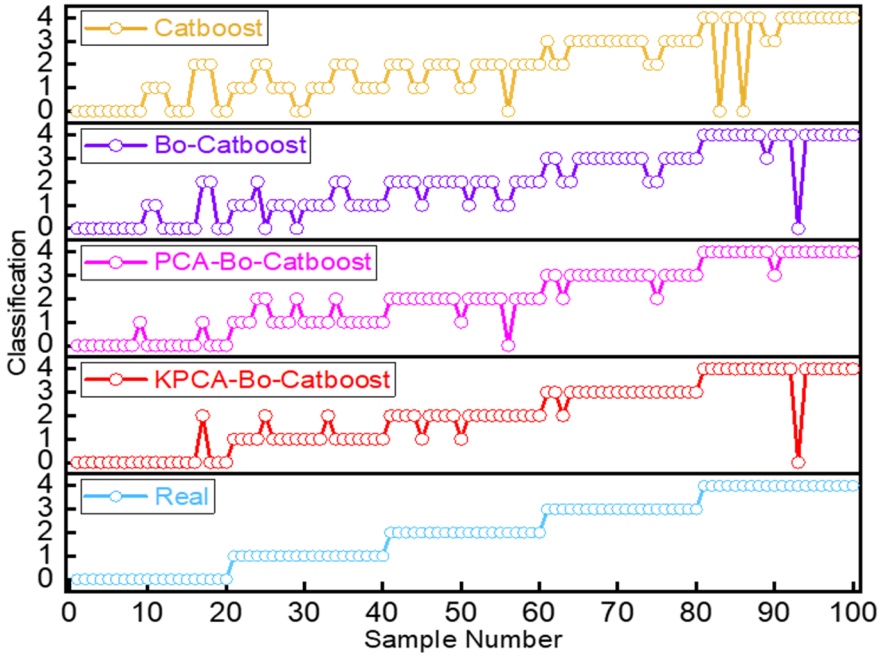

Fig 4. Comparison of the true and predicted values of each model (0- Mudstone; 1- Sandstone; 2- Glutenite; 3- Granite; 4- Limestone)

In the next second experiment, in order to prove the robustness of the model, SVM and GBDT were added for comparison, the original data was compulsorily discarded part of the logging data, because in the actual logging process, some objective reasons will cause some data to be missing, As an example, the Na element was selected, and 20% of its samples were randomly discarded, because CatBoost has the ability to handle sparse data, so the discarded samples were still available, but SVM and GBDT are unable to handle sparse data. In order not to affect the model training, the complete samples without Na elements were discarded, and the remaining 80% of complete samples (5106 samples) were used for training.The comparison models also need hyperparameters in modeling, and to ensure the fairness of the experiment, hyperparameters are also optimized for SVM and GBDT so that the models are all in the optimal state, and the final comparison of KPCA-Bo-SVM, KPCA-Bo-GBDT, KPCA-Bo-CatBoost1 (under sparse data conditions) and KPCA-Bo-CatBoost2 (80% of the complete data condition), additional logging data (1149 samples) from 2 wells in the area were selected for prediction, and one-hot coding was used for lithology because SVM and GBDT do not have coding capability by themselves. All the respective data were used as learning samples, and five principal components were obtained for both samples after KPCA dimensionality reduction, which were input to each model for training and fivefold cross-validation. The parameter settings of KPCA-Bo-CatBoost in both conditions followed the first experiment, and the results of the optimal hyperparameter combinations of the SVM and GBDT models are shown in Table 3. After the training was completed, the predicted data was fed into the models, and again the recognition ability of the four models was evaluated using the above metrics.

Table 3. Model hyperparameter optimization results

Model	SVM	GBDT
Initialization parameter range (left boundary, right boundary)	c(1,10) d (1,10) coef0 (0, 5)	iterations (100,1000) lr (0.001,1) subsample(0.1,1) depth (3,10) min_sample_split (1,20)
Bo optimization results	c=5.8 degree=4 coef0=3.7	iterations=750 lr=0.06 subsample=0.7 depth=6 min_sample_split=5

Table 4. Different model evaluation indicators

Model	Indicators	Mudstone	Sandstone	Glutenite	Granite	Limestone
Number of samples		277	243	216	265	148
KPCA-Bo-SVM	recall	0.912	0.546	0.722	0.851	0.863
	precision	0.876	0.642	0.764	0.828	0.884
	F1-score	0.894	0.590	0.742	0.839	0.873
KPCA-Bo-GBDT	recall	0.941	0.756	0.843	0.91	0.872
	precision	0.876	0.825	0.882	0.863	0.895
	F1-score	0.907	0.789	0.862	0.886	0.883
KPCA-Bo-CatBoost1	recall	0.927	0.824	0.783	0.854	0.892
	precision	0.912	0.916	0.812	0.891	0.976
	F1-score	0.919	0.868	0.811	0.872	0.932
KPCA-Bo-CatBoost2	recall	0.927	0.914	0.895	0.966	0.953
	precision	0.956	0.892	0.846	0.976	0.934
	F1-score	0.941	0.903	0.870	0.971	0.943

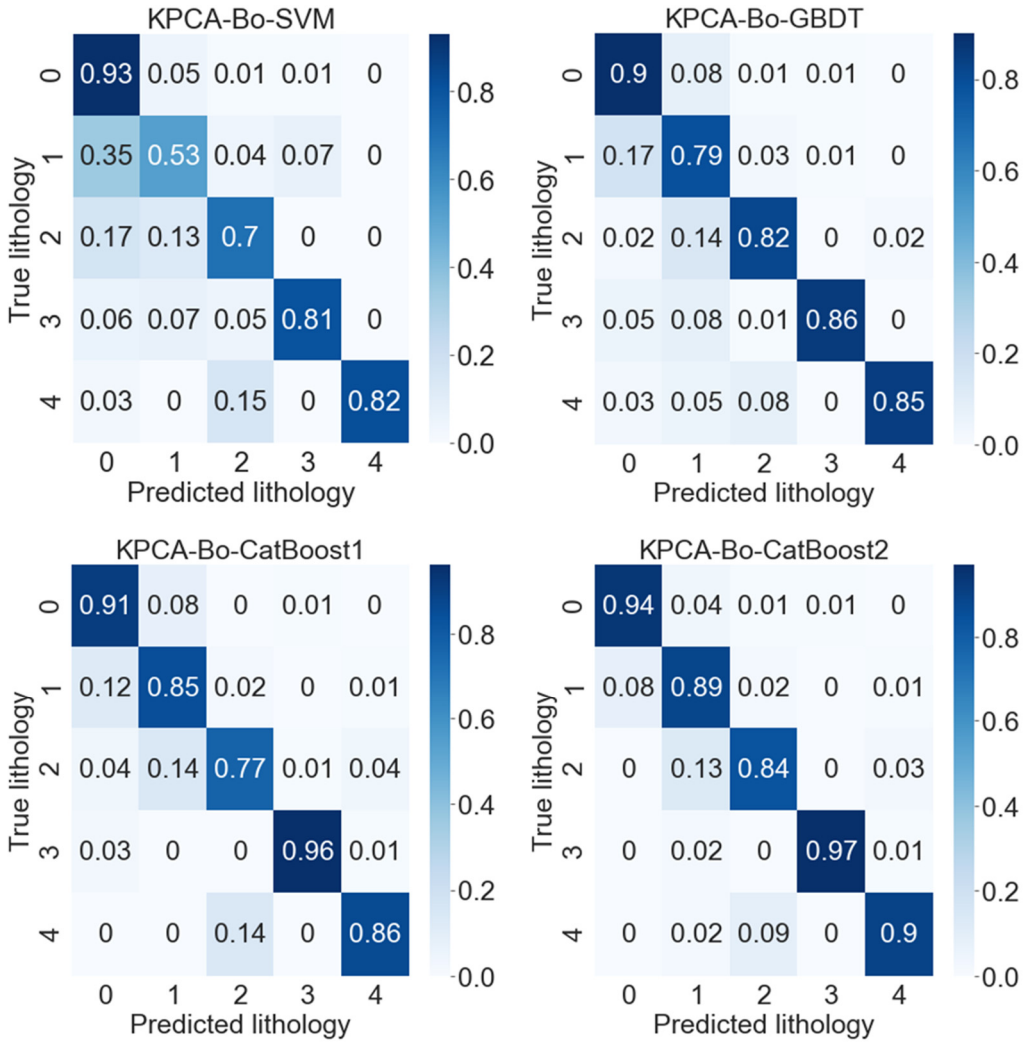

Fig 5. Confusion matrix diagram for each model (0- Mudstone; 1- Sandstone; 2- Glutenite; 3- Granite; 4- Limestone)

The accuracy information of each model is shown in Table 4, and it can be seen that the best prediction performance is KPCA-Bo-CatBoost2, and KPCA-Bo-SVM has worse prediction, indicating that the tree integration-based machine learning method is better than the traditional machine learning method after optimization. In addition, a comparison with the confusion matrix of the model (Fig.5) and the ROC plot (Fig.6) shows that the prediction effect of KPCA-Bo-CatBoost1 is reduced, but only by about 4.8% on average, but the effect is still better compared with the other two models, indicating that the proposed model can solve the lithology prediction problem even for sparse logging data, which fully reflects This shows that the proposed model can solve the lithology prediction problem even for sparse logging data, which fully reflects the advanced and superiority of the model in lithology identification.

4. Discussion

(1) In the first experiment, even only for the original data, the recognition accuracy of CatBoost is improved by Bayesian optimization, it is difficult to make the model optimal by manual parameter tuning, and Bayesian combines the information of the previous parameters so as to better adjust the current parameters, which proves the validity of the

integration of Bayesian optimization, moreover, it is found by comparison with the KCA that KPCA is more effective for nonlinear data processing is more effective, Without losing the information of the original samples, the main component information is filtered out by KPCA to reduce the coupling and redundancy between the sample data, which further improves the prediction efficiency and accuracy of the model, and the combined use of KPCA and Bo is considered to be beneficial for the calculation of CatBoost, as the model will be more practical in the prediction of lithology by combining the two techniques, which is fully demonstrated in the first experiment was fully demonstrated.

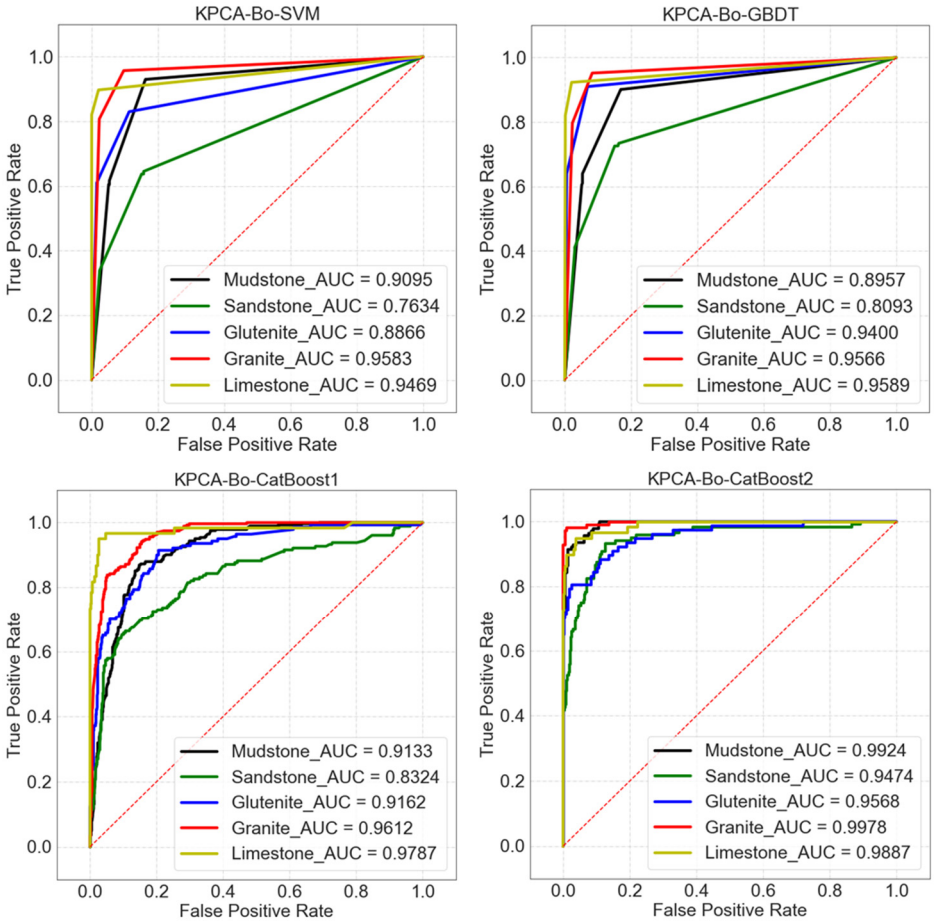

Fig 6. ROC curves for each model

(2) In the second experiment, KPCA-Bo-SVM and KPCA-Bo-GBDT predict relatively poorly, which is due to the differences in their intrinsic computational reasoning, in addition, the poor prediction also means that they don't have the desirable generalization performance because they can't deal with the part of the data that is lost, However, the comparison of the accuracy of the prediction using the KPCA-Bo-CatBoost shows that it is only reduced by about 6%, which is a good demonstration of the robustness of the model to the logging data under the sparse condition.

(3) In the second experiment, KPCA-Bo-CatBoost achieves good results regardless of whether the data is in a sparse state or not. This is due to the special training method of CatBoost, which can be analyzed even when multiple features are missing, and greatly improves the generalization performance of the model. Meanwhile, compared with the first experiment, KPCA-Bo-CatBoost has significantly improved the sandstone recognition ability under 80% complete data, which fully illustrates that more learning samples will make the model's prediction ability stronger during the training process, because more learning samples bring more opportunities for the model to correct the mapping relationship between the independent and dependent variables.

(4) Although instruments such as modern pulsed neutron logging can respond to mineral content as well as elemental concentration, it does not respond to formation lithology in real time, the method proposed in this paper is precisely a technique based on real-time lithology reflection with drilling measurement, which at the same time ensures drilling efficiency with high accuracy, so that KPCA-Bo-CatBoost is an efficient predictor when considered comprehensively in the experiments.

5. Conclusion

In order to overcome the shortcomings of traditional lithology identification methods, this paper proposes a hybrid KPCA-Bo-CatBoost model, which uses KPCA to extract fewer and more important features from the original data, combines Bayesian fast optimization hyperparameters, and establishes a more robust real-time lithology identification model by combining AI algorithms with drilling and logging technology, which not only eliminates the shortcomings of traditional methods such as poor identification accuracy and large workload, but also greatly improves the identification accuracy. The traditional method not only eliminates the defects of poor recognition accuracy and large workload, but also greatly improves the recognition accuracy, which not only conforms to the development trend of intelligent lithology recognition technology and promotes the

transformation and upgrading of petroleum drilling to intelligence, but also reveals the potential of the application of AI algorithms in petroleum drilling technology, which has the potential value of popularization.

Acknowledgments

This work was supported in part by the Science and Technology Cooperation Project of the CNPC-SWPU Innovation Alliance(2020CX020204), the Sichuan Science and Technology Program(2023NSFSC1981), the Opening Foundation of Agile and Intelligent Computing Key Laboratory of Sichuan Province (H23007).

References

[1] Abbey CP, Okpogo EU, Atueyi IO, Application of rock physics parameters for lithology and fluid prediction of 'TN' field of Niger Delta basin, Nigeria (Article) %J Egyptian Journal of Petroleum. Vol.27 (2018) No. 4, p. 853-866.

[2] Agbadze OK, Qiang C, Jiaren Y, Acoustic impedance and lithology-based reservoir porosity analysis using predictive machine learning algorithms %J Journal of Petroleum Science and Engineering. Vol.208(2022).

[3] Chongwain GC, Osinowo OO, Ntamak-Nida MJ, et al. Lithological typing, depositional environment, and reservoir quality characterization of the "M-Field," offshore Douala Basin, Cameroon %J Journal of Petroleum Exploration and Production Technology. Vol.9(2019) No.3, p. 1705-1721.

[4] Errachdi A, Benrejeb M, Online identification using radial basis function neural network coupled with KPCA %J International Journal of General Systems. Vol.46(2017) No.1, p. 52-65.

[5] Liang H, Chen H, Guo J, et al. Research on lithology identification method based on mechanical specific energy principle and machine learning theory. %J Expert Systems with Applications. Vol.189(2022).

[6] Okeugo CG, Onuoha KM, Ekwe AC. Lithology and fluid discrimination using rock physics-based modified upper Hashin–Shtrikman bound: an example from onshore Niger Delta Basin %J Journal of Petroleum Exploration and Production. Vol.11 (2021) No.2, p. 569-578.

[7] Middlebrook ML, Aud WW, Harkrider JD. An Evolving Approach in the Analysis of Stress-Test Pressure-Decline Data %J SPE Production & Facilities. Vol.12(1997) No.3, p. 187-194.

[8] Sun J, Li Q, Chen M, Ren L, et al. Optimization of models for a rapid identification of lithology while drilling - A win-win strategy based on machine learning. %J Journal of Petroleum Science & Engineering. Vol.14 (2019) No.17, p.321-341.

[9] Yang R, Wang P, Qi J. A novel SSA-CatBoost machine learning model for credit rating. %J Journal of Intelligent & Fuzzy Systems. Vol.44 (2022) No.2, p. 1-16.

[10] Zhang X, Sun Q, He K, et al. Lithology identification of logging data based on improved neighborhood rough set and AdaBoost %J Earth Science Informatics. Vol.15(2022) No.2, p. 1201-1213.

[11] Zhu X, Zhang H, Ren Q, et al. A Tri-Training method for lithofacies identification under scarce labeled logging data %J Earth Science Informatics. Vol.16(2021) No.2, p.1-13.

[12] Zhao Z, Du J, Zou C, et al. Geological exploration theory for large oil and gas provinces and its significance %J Petroleum Exploration and Development. Vol.38 (2011) No.5, p.513-522.

[13] Zou Y, Chen Y, Deng H. Gradient Boosting Decision Tree for Lithology Identification with Well Logs: A Case Study of Zhaoxian Gold Deposit, Shandong Peninsula, China %J Natural Resources Research. Vol.30(2021) No.5, p.3197-3217.

[14] Saporetti CM, Goliatt L, Pereira E. Neural network boosted with differential evolution for lithology identification based on well logs information %J Earth Science Informatics. Vol.14 (2021) No.1, p. 133-140.

[15] Gu Y, Zhang D, Bao Z. Lithology identification in tight sandstone reservoirs using CRBM-PSO-XGBoost; %J Oil and Gas Geology. Vol.42 (2021) No.5, p. 1210-1222.

[16] Santos DTd, Roisenberg M, Nascimento MdS. Deep Recurrent Neural Networks Approach to Sedimentary Facies Classification Using Well Logs %J IEEE Geoscience and Remote Sensing Letters. Vol.19 (2022) No.3, p.1-5.

[17] Simmini F, Rampazzo M, Peterle F, et al. A Self-Tuning KPCA-Based Approach to Fault Detection in Chiller Systems %J IEEE Transactions on Control Systems Technology. Vol.30 (2022) No.4, p. 1359-1374.

[18] Sun J, Li Q, Chen M, et al. Optimization of models for a rapid identification of lithology while drilling - A win-win strategy based on machine learning. %J Journal of Petroleum Science & Engineering. Vol.17 (2019) No6, p. 321-341.