ISSN: 2688-8645

Lithology Identification Method for Sandstone and Mudstone Based on the ADASYN-IRMO-CatBoost Combined Model

Xinyu Wang [1, 2, *]

[1] School of Earth Sciences and Engineering, Xi'an Shiyou University, Xi'an Shaanxi, 710065, China

[2] Shaanxi Key Laboratory of Oil and Gas Accumulation Geology, Xi'an Shiyou University, Xi'an Shaanxi, 710065, China

* Corresponding author Email: 1092677263@qq.com

Abstract

Accurate lithology identification is a crucial prerequisite for sedimentary environment analysis, oil and gas exploration, and development. Traditional identification methods have problems such as low efficiency, high cost, or limited applicability. This paper proposes an ADASYN-IRMO-CatBoost combined model. Firstly, the ADASYN algorithm is used to perform adaptive sampling on the logging datasets of three wells in the southern Ordos Basin to solve the problem of data imbalance. Then, the Improved Radial Movement Optimization (IRMO) algorithm is utilized to optimize the hyperparameters of the CatBoost model. Finally, CatBoost is used as the core classifier for lithology identification. Experimental results show that the overall accuracy of the combined model on the test set reaches 92%, which is 13% higher than that of the single CatBoost model. The F1 scores of various lithologies and the AUC values of the ROC curves are significantly better than those of the single model, demonstrating stronger classification performance and robustness. It provides an efficient and accurate new method for lithology identification of sandstone and mudstone reservoirs and has good application prospects in the field of geological exploration.

Keywords

Lithology Identification; Logging Curves; Sample Imbalance; LightGBM Model; IRMO Algorithm.

1. Introduction

Accurately distinguishing lithology is a crucial prerequisite for sedimentary environment analysis, stratigraphic framework construction, oil and gas exploration and development, as well as reservoir evaluation. In actual production operations, traditional lithology identification methods mainly include cross-plot methods, drilling coring techniques, and imaging logging techniques[1-2]. Cross-plots largely rely on the knowledge reserves and practical experience accumulated by individual geologists and require a great deal of manpower and time investment[3]; although drilling coring can directly obtain relatively complete core samples, it is extremely costly and cannot acquire continuous geological profile-related information for the entire well area[4]; imaging logging technology is carried out synchronously with the drilling process, and it can directly perform physical imaging of the wellbore and its surrounding reservoirs. The description of reservoir characteristics is both intuitive and highly accurate. However, this technology is very expensive and not suitable for widespread practical application[5-6].

In recent years, with the development of computers, machine learning can efficiently and accurately mine and extract key features from a large amount of logging data and has been applied in the field of lithology identification. Guo Yushan introduced a feature-weighted KNN model into lithology identification, assigning different weights to different features, which improved the identification accuracy; Liu Junyi adopted the random forest algorithm to identify complex lithologies in the buried hill section of the Huizhou structure in the Pearl River Mouth Basin in 2025 and achieved good results; Zhang Fengbo proposed using a genetic algorithm (GA) to optimize a hybrid neural network (CNN-LSTM) in 2024, further enhancing the accuracy of lithology identification. Machine learning has made significant progress in the field of lithology identification, but there are still some problems. Most models ignore the impact of unbalanced datasets on performance, and the models will be biased towards the majority class, which seriously affects the performance of the entire model[7]. In addition, the parameter configuration of the model also affects the identification accuracy, and a single optimization algorithm is difficult to quickly find the optimal combination in a huge parameter space, which affects the model performance.

To address the above problems, an ADASYN-IRMO-CatBoost combined model is constructed based on the actual logging data of the study area. The ADASYN algorithm is used to balance sample data, the Improved Radial Movement Optimization (IRMO) algorithm is introduced to optimize and adjust the hyperparameters of the CatBoost model, and CatBoost is used as the core classifier for classification. The results show that the ADASYN-IRMO-CatBoost combined model has strong reliability and generalization ability for lithology identification.

2. Key Algorithmic Principles of the Combined Model

2.1. ADASYN Adaptive Sampling Technique

During the training of lithology identification models, there is often a significant imbalance in the number of samples across different lithology categories. Although the traditional SMOTE method can generate minority-class samples through linear interpolation, it uniformly produces a large number of samples with blurred boundaries (Liu Zhihan et al., 2024), which instead increases the difficulty of classification. The ADASYN (Adaptive Synthetic Sampling) algorithm improves this limitation by introducing an adaptive mechanism, whose core innovation lies in dynamically adjusting the generation strategy based on the sample distribution density.

ADASYN first calculates the distribution density of majority-class samples around each minority-class sample, and determines the number of new samples to be generated in each region based on this[8]. In specific implementation, the algorithm uses the K-nearest neighbor algorithm to calculate the learning difficulty weight of each minority-class sample, then allocates the number of samples to be generated according to the weight, and finally generates new samples in the feature space through linear interpolation. Different from the uniform sampling of SMOTE, this adaptive mechanism of ADASYN enables it to more targeted enhance the sample density at the classification boundary while avoiding the generation of meaningless noise samples.

2.2. IRMO Algorithm

To improve the performance of the CatBoost model in lithology identification tasks, this study adopts the IRMO (Improved Radial Movement Optimization) algorithm for automatic hyperparameter optimization of the CatBoost model. This algorithm iteratively updates the optimal position and central position of the particle swarm, and continuously adjusts the search strategy based on fitness feedback results, gradually converging the results to the global optimal solution.

(1) Population Initialization: M N-dimensional particles are randomly generated in the solution space, and the fitness value of each particle is calculated. In this study, the weighted

F1-score under 5-fold cross-validation is used as the fitness evaluation index [10]. The particle with the optimal fitness is set as the initial central particle and the global optimal solution, which participates in the subsequent iteration process.

(2) Particle Position Update: During the K-th iteration, the position of particles is updated according to Equations (1) and (2). After each iteration, the current optimal particle is selected; if its fitness is better than the historical global optimal, the global optimal solution is updated.

$$Y_{i,\ j}^{k} = [rand(-0.5,0.5)] \times \left(X_j^{max} - X_j^{min}\right) \times w^k + C_j^k \tag{1}$$

$$w^k = 1.009 \times exp\left(-\frac{(k \times G + 0.03307)^2}{0.2454}\right) \tag{2}$$

(3) Central Position Update: The position of the central particle is adjusted based on the deviation between the optimal solution of the previous generation population and the global optimal solution. The adjustment factors C_1 and C_2 are used to control the movement amplitude. This process continues until the iteration ends, and the final global optimal solution is the optimal hyperparameter combination for the CatBoost model.

$$C_{center}^{k+1} = C_{center}^{k} + C_1\left(R_{best}^k - C_{center}^k\right) + C_2\left(G_{best}^k - C_{center}^k\right) \tag{3}$$

Among them, the adjustment factors C_1 and C_2 in this study are set to 0.5 and 0.4, respectively.

2.3. CatBoost Algorithm

CatBoost is a novel ensemble algorithm based on gradient-boosted decision trees proposed by the Yandex team in 2017, and it is a member of the Boosting family. This algorithm exhibits significant advantages in handling categorical features, avoiding overfitting, and enhancing the model's generalization ability. Its core features lie in the adoption of the Ordered Boosting mechanism and a special processing method for categorical features, which effectively address the problems of gradient bias and prediction shift.

The Ordered Boosting mechanism constructs the model through a permutation-based gradient estimation method. This mechanism effectively overcomes the issues of gradient bias and prediction shift caused by target variable leakage in traditional gradient boosting, making model training more stable, significantly improving generalization performance, and suppressing overfitting. Secondly, CatBoost possesses native and efficient categorical feature processing capabilities. It automatically converts categorical features into numerical representations without relying on manual encoding (such as one-hot encoding or label encoding), which not only avoids the problem of high-dimensional sparsity but also significantly improves computational efficiency and feature expression ability.

Thanks to the integration of these two key technical advantages—Ordered Boosting and intelligent categorical feature processing—CatBoost can effectively handle heterogeneous features in complex datasets across various machine learning tasks. While maintaining efficient training, it greatly reduces the need for parameter tuning and has become one of the important tools for processing industrial-scale data.

3. Case Application

3.1. Data Preprocessing

This study selects lithologies from parts of the Permian System in three wells in the southern Ordos Basin as the research object. Based on existing logging data, a sample database is established, incorporating seven logging curves (SP, GR, AC, CNL, AC, DEN, LLD, LLS) from the study interval. The main lithologies include fine sandstone, siltstone, medium-coarse sandstone, argillaceous siltstone, and mudstone, with corresponding labels assigned as 0, 1, 2, 3, and 4, respectively. The dataset is divided into training and test sets using an 8:2 stratified sampling method. The logging curve data in the sample database undergoes preprocessing: due to inconsistent dimensions across different logging curves (which hinder direct comparison), the samples are standardized. Given the imbalanced distribution of sample data in the database, the ADASYN adaptive sampling technique is applied to resample the training set, balancing the dataset.Subsequently, the CatBoost model is trained, with the Improved Radial Movement Optimization (IRMO) algorithm introduced for parameter tuning. The fitness function is set as the F1-score from five-fold cross-validation, with a population size of 20 and 15 iterations. Through continuous iterative optimization, the optimal model parameters are obtained to maximize performance.

Table 1. Results of Dataset Partitioning

Lithological classification	Training sample	Test sample	Lithological label
Lithology Classification	**Training Samples**	**Test Samples**	**Lithology Label**
Fine Sandstone	668	167	0
Siltstone	220	55	1
Medium-Coarse Sandstone	301	75	2
Argillaceous Siltstone	571	143	3
Mudstone	974	244	4

3.2. Parameter Setting Based on IRMO

Model parameters determine model performance, and different algorithms have different parameters. The processed training samples with 7 logging curve features were used to train the CatBoost model. In the IRMO optimization algorithm, the optimal parameter combination for the CatBoost model was found according to the above settings, achieving efficient optimization of the parameters for the lithology identification model. After multiple iterations, the optimized parameters of the model were determined as follows: iterations = 208, learning_rate =0.3, depth=10.

3.3. Model Test Results

The global optimal parameters of the combined model were used for training and testing on the test set. The single CatBoost model was compared with the CatBoost model optimized by the IRMO algorithm in terms of prediction and classification on the test set to evaluate their identification effects.

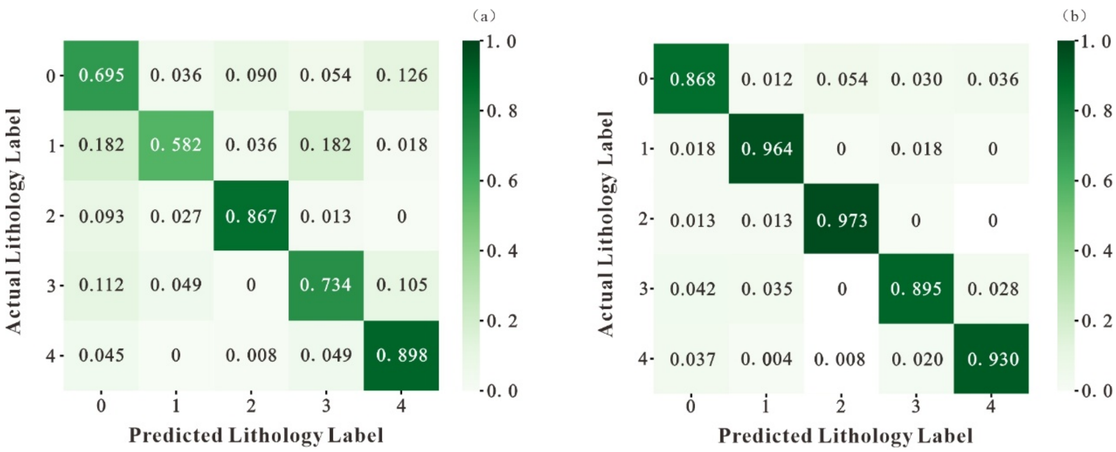

Figure 1. Confusion Matrices of the Two Models on the Test Set
Note:(a)—CatBoost;(b)ADASYN-IRMO-CatBoost

ISSN: 2688-8645

The confusion matrices of the accuracy of the two models on the test set are shown in Figure 1, where the diagonal elements represent the prediction accuracy of various lithologies. The single CatBoost model achieved an identification accuracy of 0.695 for fine sandstone, 0.582 for siltstone, 0.867 for medium-coarse sandstone, 0.734 for argillaceous siltstone, and 0.898 for mudstone, with obvious misjudgment in fine sandstone and siltstone. In comparison, the ADASYN-IRMO-CatBoost combined model performed better, with identification accuracies of 0.868 for fine sandstone, 0.964 for siltstone, 0.973 for medium-coarse sandstone, 0.895 for argillaceous siltstone, and 0.930 for mudstone. The misjudgment ratio of various lithologies was significantly reduced. Overall, the ADASYN-IRMO-CatBoost model had smaller errors, reflecting its stronger discrimination ability and stability.

As shown in Table 2, the overall accuracy ($E_{Accuracy}$) of the ADASYN-IRMO-CatBoost combined model reaches 92%, which is a 13% improvement compared to the 79% accuracy of the single CatBoost model. The combined model achieves a systematic increase in F1 scores for all lithology categories, and particularly addresses the insufficient ability of the original single model to identify minority-class samples. In terms of lithology identification for sandstone-mudstone reservoirs, the overall performance of the combined model far surpasses that of the single model. This fully demonstrates that the integration of ADASYN (for solving data imbalance), IRMO (for parameter optimization), and the inherent advantages of CatBoost significantly enhances the accuracy of lithology identification in sandstone-mudstone reservoirs, thereby verifying the effectiveness of the proposed combined model.

Table 2. Classification Report

Model	Lithological classification	ADASYN-IRMO-CatBoost	CatBoost
E_{F1}	Fine Sandstone	0.88	0.71
	Siltstone	0.91	0.63
	Medium-Coarse Sandstone	0.92	0.82
	Argillaceous Siltstone	0.91	0.75
	Mudstone	0.94	0.88
$E_{Accuracy}$		0.92	0.79

To comprehensively verify the lithology identification performance of the combined model, ROC curves and AUC values were introduced for auxiliary analysis. The ROC curve takes the false positive rate (FPR) as the horizontal axis and the true positive rate (TPR) as the vertical axis, intuitively presenting the classification performance of the model under different thresholds (Zhao Ranlei et al., 2025); the AUC value quantifies the area under the ROC curve, and the closer the value is to 1, the better the discrimination efficiency of the model.

To further evaluate the classification performance of the models, ROC curves of the two models were plotted and compared. As shown in Figure 2, the ROC curve of the ADASYN-IRMO-CatBoost combined model is closer to the upper-left corner, indicating extremely high generalization ability. In contrast, although the single CatBoost model also exhibits good performance, its AUC values for all lithology categories are significantly lower than those of the combined model—with a particularly notable gap in the identification of fine sandstone. The overall results of the ROC curves show that the proposed combined model not only performs prominently in terms of classification accuracy but also is significantly superior to the traditional CatBoost method in terms of category discrimination ability and model robustness. This further verifies the effectiveness and necessity of the IRMO algorithm in improving lithology identification performance.

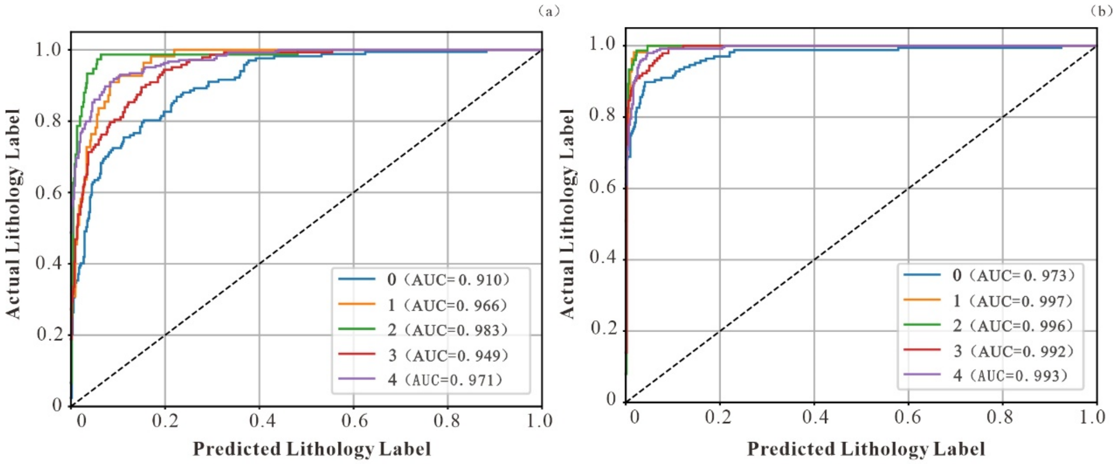

Figure 2. ROC Curves and AUC Values of the Two Models on the Test Set

Note: (a)—CatBoost;(b)ADASYN-IRMO-CatBoost

4. Conclusion

The ADASYN-IRMO-CatBoost combined model proposed in this study effectively addresses the problems of data imbalance and model parameter optimization for lithology identification in sandstone-mudstone reservoirs. The ADASYN algorithm balances the dataset through adaptive sampling, which enhances the basis for identifying minority-class lithology samples; the optimization of CatBoost model hyperparameters by the IRMO algorithm gives full play to CatBoost's advantages in categorical feature processing and anti-overfitting.

Case application results show that the overall accuracy of the combined model on the test set reaches 92%, which is a 13% improvement compared with the single CatBoost model. Additionally, the F1 scores and AUC values for all lithology categories of the combined model are significantly better than those of the single model, and its ROC curve is closer to the upper-left corner—demonstrating stronger classification performance and robustness.

This combined model provides an efficient and accurate method for lithology identification in sandstone-mudstone reservoirs and holds promising application prospects in fields such as geological exploration.

References

[1] Zhao Y, Guan Y, Li P, et al. Reservoir lithology identification based on quantum-derived vortex algorithm and T-S fuzzy reasoning model [J]. Petroleum Geophysical Exploration, 2024, 59(01): 23–30.

[2] Li J, Qian L, Qian L, et al. Multi-attribute basement lithology prediction and effect based on rock classification analysis [J]. Petroleum Geophysical Exploration, 2022, 57(S2): 70–73+227.

[3] Li C, Jia J, Shen A, et al. A new method and application of identifying microbial carbonate lithofacies based on conventional logging [J]. Progress in Geophysics, 2020, 35(05): 1792–1802.

[4] Cheng C, Li P, Chen Y, et al. Research progress on reservoir logging evaluation based on machine learning [J]. Progress in Geophysics, 2022, 37(01): 164–177.

[5] An P, Cao D. Research and application of logging lithology identification method based on deep learning [J]. Progress in Geophysics, 2018, 33(03): 1029–1034.

[6] Hu H, Zeng H, Liang H, et al. Lithology identification method based on principal component analysis and learning vectorized neural network [J]. Well Logging Technology, 2015, 39(05): 586–590.

[7] Chen Y, Wang P. Establishment of three-dimensional solid model of blasting rock mass based on intelligent lithology identification [J]. Coal Geology & Exploration, 2022, 52(04): 92–97.

[8] Xu T, Ren Q, Li J, et al. Analysis of frequent congestion causes based on ADASYN-Optuna-CatBoost and SHAP interpretation algorithm [J]. Highway, 2025, 70(08): 228–239.

[9] Jin L, Luo C, Wei J, et al. Reservoir slope stability analysis under dynamic fluctuating water level using improved radial movement optimisation (IRMO) algorithm [J]. Mathematics, 2024, 12(13): 2055.

[10] Wang J, Wang H, Zhang W, et al. Study on surface subsidence prediction model based on IRMO–XGBoost [J/OL]. Journal of Safety and Environment, 2025: 1–12 [2025-07-28].