International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 8, Number 6, Year 2026
DOI: https://doi.org/10.62051/ijcsit.v8n6.03
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Carbonate lithology Identification Based on Genetic Algorithm Adaptive LightGBM

Jingyue Zhang *, Renfeng Zhang, Siyu Zhang, Sijie Li

College of Computer and Information, College of Arts and Sciences of Hubei Normal University, Huangshi, China

ABSTRACT

Lithology identification utilizing logging data stands as a pivotal research area within reservoir prediction, integral to the broader context of oil and gas exploration and development. In order to solve the problem that the traditional logging lithology identification method and machine learning method do not have high accuracy for the lithology identification of complex carbonate rocks. In view of the powerful performance of deep learning methods in feature extraction and data analysis, this paper proposes a lithology recognition method based on the Genetic Algorithm-adaptive LightGBM model. Due to the small number of features of the original logging data, the feature shift method is first used to expand the features of each data. Subsequently, to optimize the classification performance of our model, we leverage the Genetic Algorithm (GA) to fine-tune the hyperparameters of the LightGBM model. With the optimal hyperparameter configuration in place, we evaluate the model's performance through rigorous training with logging data. The adaptive LightGBM lithology recognition model based on GA is compared with the lithology identification results of GBDT, Bayes, DNN and DT, and the overall prediction performance is evaluated by using the accuracy and its macroscopic mean. Through the test and verification of the actual well data, the proposed method has achieved very considerable results in lithology identification, showing the broad prospect of LightGBM technology in the field of geophysics.

KEYWORDS

Lithological identification; LightGBM; GBDT; Geophysical logging

1. INTRODUCTION

Logging lithology identification is the basis for carbonate oil and gas reservoir research, which is of great significance in oil and gas development and mineral resources exploration. Drilling cores is the most direct and effective method for lithological identification. However, due to the high coring cost and low coring rate, it is not easy to fully obtain the required core. Therefore, the use of cores to identify strata in the actual exploration process is not sufficient for widespread use in all wells [1]. The classic lithological identification tool is the cross plot. The cross plot is generally composed of 2 or 3 kinds of logging curves, and the selected curve needs to have a unique response range for a variety of lithology that needs to be identified, because the lithological composition of unconventional reservoirs is complex, and most of the lithology has similar response characteristics on the logging curve, resulting in the classical identification tools represented by the cross plot are difficult to apply, so the use of deep learning methods to carry out lithology identification research on the logging curve is an important direction at present. Moreover, traditional well logging analysis techniques are not systematic and depend on the experience of the interpreter, so multiple interpretations can be generated [2, 3]. With the rapid development of artificial intelligence, machine learning methods have begun to be widely used in logging identification, and many studies use logging data for lithology

identification based on machine learning-related algorithms [4, 5]. An FL-XGBoost model was proposed by Peng et al. [6]. Based on the logging and logging data, select the appropriate logging parameters. The complete training dataset is constructed through multiple preprocessing operations, and the Fcoal Loss function is introduced to solve the problem of log label sparsity. In view of the powerful performance of machine learning in data analysis, Shi et al. [7] combined SMOTE, PSO-optimized GBDT and logic-CNN to address logging data imbalance and improve lithology identification accuracy for uranium drilling. Ren et al. [8] proposes a novel hybrid method of lithology identification based on k-means++ algorithm and fuzzy decision tree. It can be seen from the above models that machine learning algorithms such as random forests and deep neural networks have achieved significant results in solving related geological problems, but there are still some difficulties in the field of sandstone and mudstone identification: the preprocessing of sample sets has a great impact on the performance of machine learning algorithms, such as missing values and outliers. Moreover, carbonate lithology data are complex and diverse, and there are similarities between different lithology.

Due to the heterogeneity and complexity of the geological environment, the size of various samples of logging data is uneven. There may be strong similarities between different samples, making it difficult for various models to divide them. Not only that, but deep learning models have many hyperparameters. The same method is applied to different regions, and the hyperparameters need to be adjusted so that the model can still have good recognition results in different regions. In order to solve the problem that hyperparameters are difficult to determine, this paper proposes an optimization method with complementary genetic algorithm (GA) and GridSearch, which eliminates the need to manually adjust model hyperparameters. This parameter optimization method is combined with the LightGBM model, so that the proposed method can be better applied to different regions.

LightGBM is a framework that implements GBDT algorithms and supports efficient parallel training [9-11]. GBDT is an excellent machine learning model, its calculation principle is based on probability and statistics, through the regression tree between the residual between the calculated value and the target value to classify and analyze, and then use the stepwise improvement algorithm to continuously reduce the residual so that the calculated value gradually approximates the target value [12, 13]. However, GBDT needs to traverse the entire training data multiple times at each iteration. If you put the entire training data into memory, it will limit the size of the training data; If it is not loaded into memory, repeatedly reading and writing training data will consume a lot of time. Especially in the face of massive data, ordinary GBDT algorithms cannot meet their needs. Therefore, LightGBM is optimized on the traditional GBDT algorithm. The aim of this study is to use this model to identify carbonate lithology. In the following, the computational principles and experimental verification of the proposed model will be elaborated.

2. DATA ACQUISITION

The logging data in this article comes from the Hugoton and Panoma oil and gas fields in North America. The Panoma Council Grove Field is predominantly a carbonate gas reservoir encompassing 2700 square miles in Southwestern Kansas. The dataset includes 10 wells with determined lithology, totaling 4149 samples. The sampling interval is 0.5m and consists of a set of 5 predictors and lithofacies of each sample vector. Different logging methods have different response characteristics to the same lithology. Therefore, five variable features are selected in this paper, including gamma ray (GR), resistivity logging (ILD_log10), photoelectric effect (PE), neutron-density porosity difference (Delta PHI) and average neutron-density porosity (PHIND). However, PE data for some wells are missing. Missing data is cleared in the experiment. The logging data is shown in Table 1.

Table 1. Well Logging Data

row_id	Facies	well_id	GR	ILD_log10	DeltaPHI	PHIND
0	3	9	77.45	0.664	9.9	11.915
1	3	9	78.26	0.661	14.2	12.565
2	3	9	79.05	0.658	14.8	13.050
3	3	9	86.10	0.655	13.9	13.115
4	3	9	74.58	0.647	13.5	13.300
…	…	…	…	…	…	…
4144	5	1	46.719	0.947	1.828	7.254
4145	5	1	44.563	0.953	2.241	8.013
4146	5	1	49.719	0.964	2.925	8.013
4147	5	1	51.469	0.965	3.083	7.708
4148	5	1	50.031	0.970	2.609	6.668

The natural gamma curve reflects the intensity of gamma rays released when radioactive elements in a rock are metamorphosis. The intensity of gamma rays in general mudstones is high. The intensity of gamma rays in sandstone is weak. Moreover, the higher the argillaceous content in the sandstone, the greater the gamma ray intensity. Natural gamma logging is therefore sensitive to argillaceous content. Resistivity logging is a logging method that uses the principle of electromagnetic induction to measure the conductivity of dielectric. According to the different properties of different dielectric conductivity, the lithology of the formation is judged. The photoelectric effect (PE) is one of the effective parameters used in oil logging to determine the lithology of formations. The PE value is meant to highlight the average atomic number of the rock. PE is a parameter defined to better reflect lithology in proportion to the electron photoelectric absorption cross-section. The methods for determining the porosity of the formation mainly include acoustic, neutron and density logging. Neutron-density porosity logging can determine the skeleton value by intersection diagram, and then determine the lithology. The average neutron-density porosity logging is a weighted embodiment of neutron-density porosity logging. The response characteristics of carbonate stratigraphic logging are shown in Table 2.

Table 2. Logging Response Characteristics of Some Carbonate Formations

Facies	Density (g/cm3)	Neutron porosity (%)	Acoustic (μs/m)	Gamma ray (API)	PE (b/e)
Limestone	2.71	0-1	150-160	1-10	5-5.1
Dolomitic limestone	2.7-2.79	0-2	150-160	5-15	3.5-5
Gilliferous dolomitic limestone	2.65-2.73	5-7	170-175	10-75	3-4
Marl	2.65-2.78	2-15	170-230	40-110	3.3-4.5
Dolomite	2.87	1-3	140-146	15-20	3.1-3.2
Limy dolomite	2.79-2.87	1-9	150-160	10-20	3-3.7
Argillaceous dolomite	2.6-2.83	2-13	150-160	15-75	1.9-2.6

From Table 2, it can be seen that different curves of distinct geological strata exhibit variations in electrical and physical properties. By contrasting these logging curves, lithology discrimination can be achieved. Employing the morphological characteristics and values of the curves, reservoir rock types are identified, enabling the calculation of primary mineral and shale content, as well as the analysis of clay mineral composition within the shale.

3. DATA ACQUISITION

3.1. Rendezvous Diagram

The cross plot method is a technique for statistically recording characteristic values. Commonly used rendezvous methods are two-dimensional, three-dimensional and even multi-dimensional. It is the intersection of various logging data on a floor plan or space. According to the coordinates of the meeting point, determine the value, range or interval of the parameter sought. General steps of intersection diagram: firstly, the logging curve parameters with a large influence on lithology are preferred, such as natural gamma, deep lateral resistivity, compensation for acoustic time difference, compensation density, compensation neutrons, etc. Then select two or more logging data for two-dimensional or multi-dimensional intersection in the coordinate system. The data points in the coordinate system are calibrated according to the existing core lithological data. Prepare the corresponding rendezvous plates.

Because the cross plot method uses few parameters, it is relatively simple and convenient to identify. If the mineral composition and pore structure of the reservoir are complex, there will be serious crossover of logging parameters of different lithology. It is difficult to identify lithology using the plate method or the intersection diagram method. By employing suitable mathematical-physical approaches, the aforementioned challenges can be notably mitigated, thereby enhancing the accuracy of lithological identification to a certain extent. Notably, in scenarios where core data and comprehensive logging information are scarce, these mathematical-physical methods exhibit superior efficacy as identification tools. Furthermore, machine learning methods are based on robust mathematical physics theories. The purpose of this paper is to conduct in-depth analysis and optimization of various machine learning-based recognition methods. Specifically, it emphasizes strategies for promptly and accurately assessing complex lithological environments, ultimately achieving superior identification performance.

As can be seen from Figure 1, neither of the intersection of the two curves can effectively divide carbonate lithology. Because the intersection diagram can only represent a fine granular composition. The complexity of carbonate rocks stems from their physicochemical properties and genesis. This complexity cannot be expressed in finely granular composition.

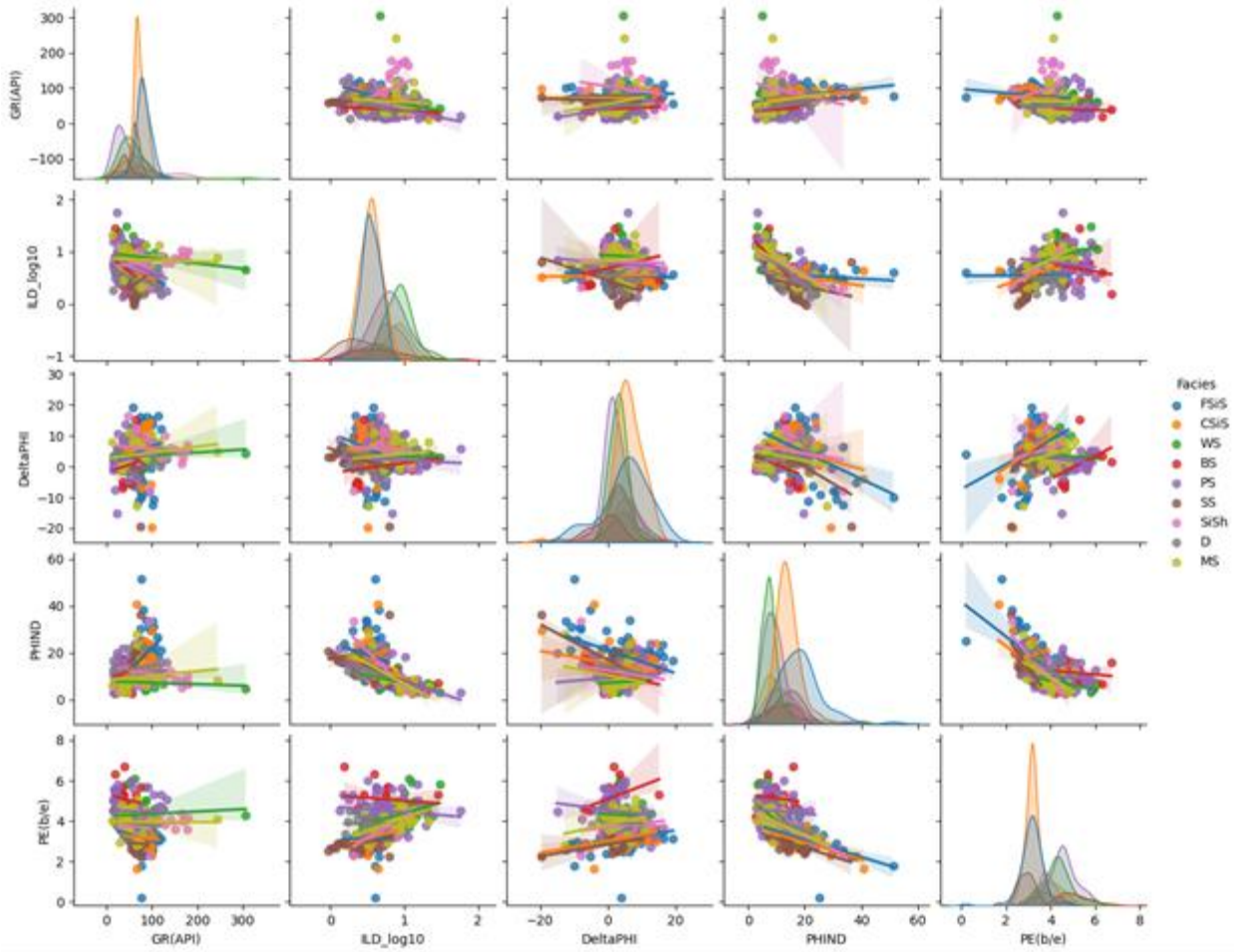

Figure 1. Two-Dimensional Intersection Diagram For Lithology Identification

3.2. LightGBM

3.2.1. Histogram Algorithm

The basic idea of the histogram algorithm is to first discretize continuous floating-point eigenvalues into k integers. Also construct a histogram with width k. As you traverse the data, accumulate statistics in the histogram based on the discretized values as indexes. When the data is traversed once, the histogram accumulates the required statistics. Then, according to the discrete values of the histogram, traverse to find the optimal split point.

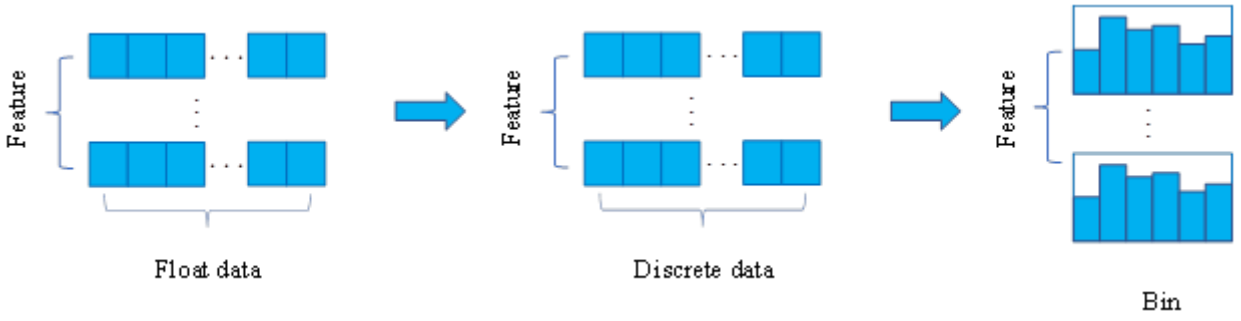

Figure 2. Histogram-Based Decision Tree Algorithm

Figure 2 shows the binning process of the model. The histogram algorithm is simply understood as: first determine how many bins are needed for each feature and assign an integer to each box. The range of floating-point numbers is then divided into intervals. The number of intervals is equal to the number of boxes. Update the sample data belonging to the box to the value of the bin. Finally, it is

represented by a histogram (bins). Feature discretization has many advantages, such as convenient storage, faster operation, strong robustness, and more stable models.

Another LightGBM optimization is Histogram Difference Acceleration. The histogram of a leaf can be obtained by differing the histogram of its parent node from the histogram of its brother. It can be doubled in speed. Usually when constructing a histogram, you need to iterate through all the data on that leaf. But to make a difference in the histogram, you only need to traverse the k Bins of the histogram. In the actual process of building the tree, LightGBM can also calculate the leaf nodes with small histograms, and then use the histogram to make a difference to obtain the leaf nodes with large histograms, so that you can get the histograms of its brother leaves at a very small cost. The histogram does the difference acceleration process, as shown in Figure 3.

Figure 3. The Histogram Does Differential Acceleration

3.2.2. Leaf-Wise Algorithm With Depth Limit

Level-wise once data can split leaves in the same layer at the same time. Easy multithreading optimization. It is also good to control the complexity of the model, and it is not easy to overfit. But in reality, Level-wise is an inefficient algorithm. Because it treats leaves of the same layer indiscriminately. It brings a lot of unnecessary overhead. In fact, many leaves have a lower split gain. There is no need to search and split. The Level-wise growth process is shown in Figure 4.

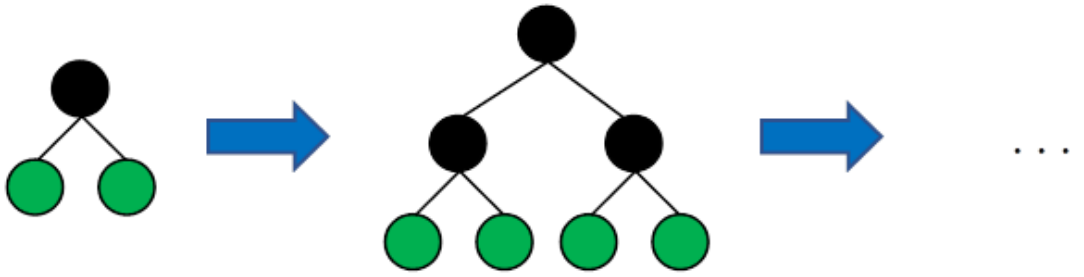

Figure 4. Level-Wise Tree Growth

Leaf-wise is a more efficient strategy. Each time from all the current leaves, find the one with the largest split gain and split. So compared to Level-wise, in the case of the same number of splits. Leaf-wise can reduce more errors and get better accuracy. The disadvantage of Leaf-wise is that it may grow deeper decision trees and overfit. So LightGBM adds a maximum depth limit on top of Leaf-wise. Prevent overfitting while maintaining high efficiency [14]. The Lesf-wise growth process is shown in Figure 5.

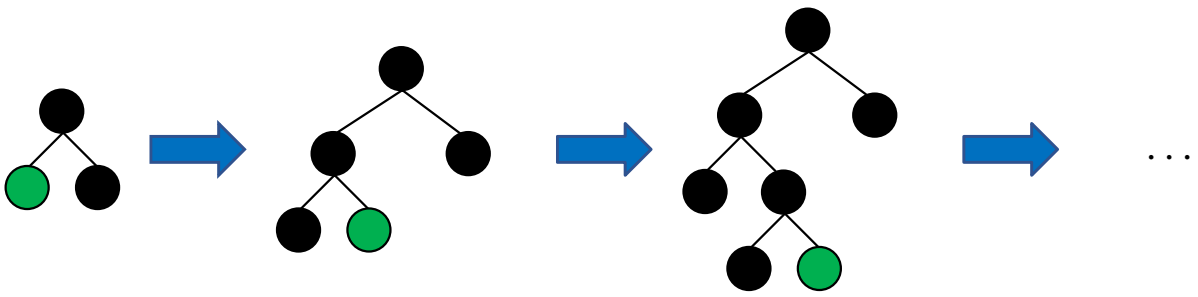

Figure 5. Leaf-Wise Tree Growth

3.2.3. Leaf-Wise Algorithm With Depth Limit

An example of sampling from GOSS is shown in Figure 6. Unilateral gradient sampling is undersampling of a sample. When GOSS undersampling, it mainly revolves around the following questions: how to take a more "informative" sample. How to ensure consistent sample distribution before and after sampling. To the first question, LightGBM's answer is to prefer samples with gradient values in the top a%. The reason for this is because under the framework of Gradient Boosting, the larger the gradient, the greater the gain when splitting. Samples with small gradients say that they have been trained better.

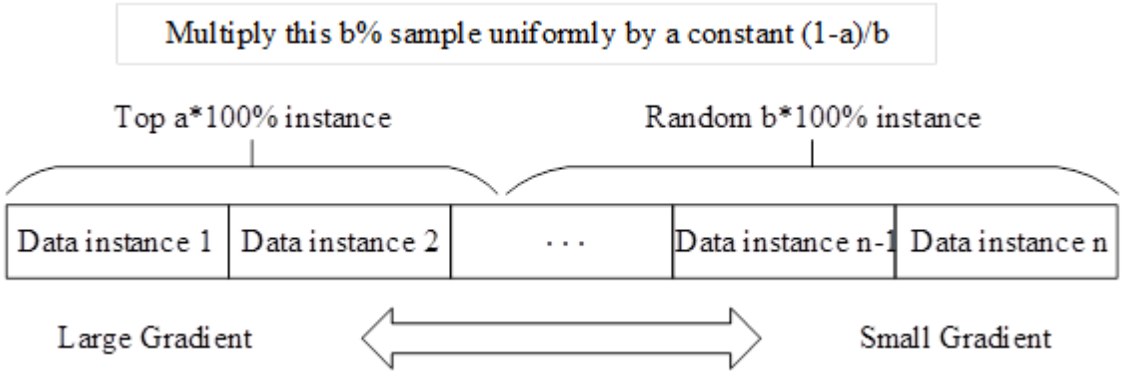

Figure 6. Gradient-Based One-Side Sampling

For the second problem, LightGBM first randomly samples the b% ratio sample in the remaining small gradient samples. Secondly, when calculating the weight, LightGBM weights this b% sample by (1-a)/b times. This further avoids the problem of inconsistent sample distribution.

3.3. Hyperparameter Optimization

To build the best machine learning model, you must explore a range of possibilities. The process of designing an ideal model architecture using an optimal hyperparameter configuration is called hyperparameter tuning [15]. Tuning hyperparameters is considered a key component of building effective ML models. Especially for tree-based ML models and deep neural networks with many hyperparameters [16]. Choosing the appropriate optimization technique to detect the best hyperparameters is critical. Genetic algorithm is a common traditional optimization algorithm. Hyperparameter optimization can use the genetic algorithm GA. The genetic algorithm has good global search capabilities. It is possible to quickly search out the whole solution in the solution space without falling into the trap of rapid descent of the local optimal solution. However, because the genetic algorithm jumps out of the local optimal by cross-mutation. A gene point variation can jump a long distance in space. This leads to poor local search ability and low efficiency in later search. grid search (GS) is a decision theory approach that exhaustively searches for the best configuration in a fixed domain of hyperparameters.Therefore, when the optimal solution obtained by the genetic algorithm is close to stable or the accuracy has reached the demand, GS can be used to find a better solution near the optimal solution of the genetic algorithm. Table 3 shows the optimal hyperparameters of LightGBM.

Table 3. Model Parameterss

Model	Hyperparameters	Search Range	Optimal Values
LightGBM	learning_rate	(0.01, 0.1)	0.09
	max_depth	(1, 10)	8
	num_leaves	(60, 80)	60
	n_estimators	(50, 1000)	200

4. EXPERIMENTS

4.1. Model Training

First, the null values in the data are eliminated, and the data features are enriched by feature shifting. Second, genetic algorithms and grid search are used to tune model hyperparameters. Third, the optimal hyperparameter configuration is used to fit the predictive model based on the training set. Fourth, according to the overall prediction results and the prediction ability of each characteristic, the test set is used to evaluate the model performance. Finally, by comparing the combined performance of these models, the optimal model is determined. If the predictive performance of the model is acceptable, it can be deployed with the model.

4.2. Model Testing

In order to highlight the validation effect, some other machine learning models were added to the experiment, namely DNN, Bayes, DT, GBDT and LightGBM. Deep neural networks use feed-forward propagation to input training data into the network, training layer by layer to the output layer. Then use the backpropagation algorithm to get the error value of each neuron. Next, using a gradient-based optimization algorithm, the gradient of each parameter of the model is calculated first based on the cost function. Calculate the "penalty value" based on the gradient and adjust the model parameters. Bring the prediction closer to the optimization goal. Bayes network uses the probability value of predicted samples in the training sample set of each recognition pattern for pattern judgment, so the main task of the training process is to construct the probability density distribution of each recognition pattern. DT is adecision analysis method that intuitively uses probability analysis to obtain the probability that the expected value of the net present value is greater than or equal to zero by forming a decision tree on the basis of knowing the probability of occurrence of various situations, evaluating the project risk, and judging its feasibility.

In addition to calculating the probability residuals of each sample, GBDT also needs to reasonably classify these residuals in the regression tree, so constructing a series of efficient regression trees is also an important task in the validation of the model. Both LightGBM and GBDT use the negative gradient of the loss function as an approximation of the residual of the current decision tree to fit the new decision tree. However, LightGBM adds histogram difference and unilateral gradient sampling methods on the basis of GBDT to improve the data processing speed without reducing the accuracy as much as possible.

In this paper, the best-in-one hyperparameters of LightGBM are obtained by comparing the given parameters with other methods, and then the hyperparameter optimization method is used to obtain the best hyperparameters of LightGBM. Comparing the prediction results of these algorithms, the results show that GBDT and LightGBM have better performance. GA_LightGBM accuracy reached the highest 87.3%, followed by GBDT with 86.7% accuracy. The result is shown in Figure 7.

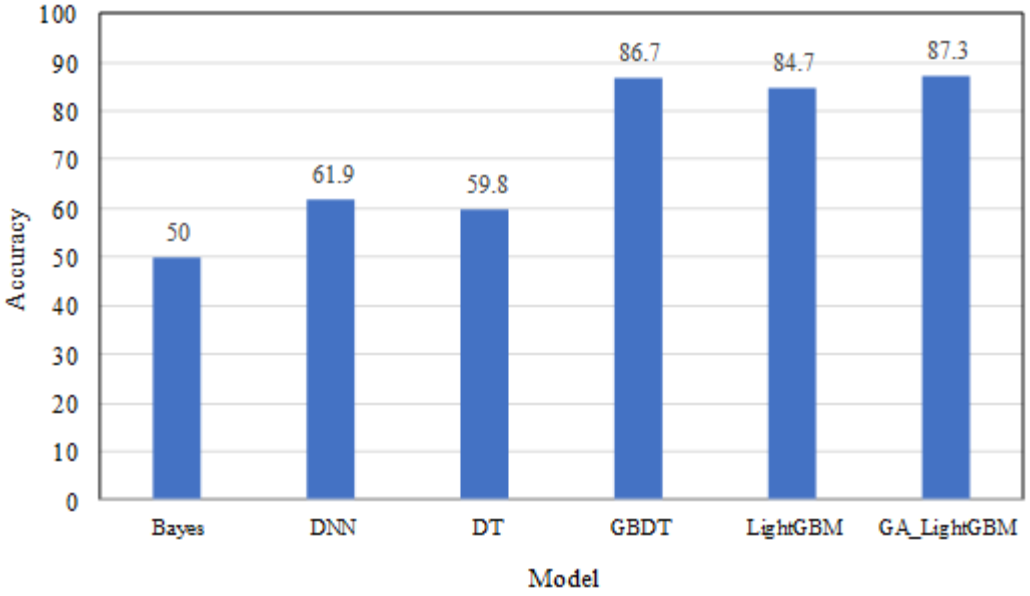

Figure 7. Accuracy of each comparison model

4.3. Model Evaluation Indexes

Table 4 shows the confusion matrix for GA_LightGBM: In the confusion matrix, each column represents the predicted category. The total number of columns represents the sample size predicted for that category. Each row represents the actual category of the sample, and the total number of rows represents the number of data instances for that category. where precision represents the ratio of the number of positive cases of the prediction pair to the total number of predicted positive cases. Recall represents the ratio of the number of positive cases of the predicted pair to the true number of positive cases. However, sometimes the accuracy and recall will be contradictory, so the F1 value will be weighted to reflect the accuracy of the model. The value on the main diagonal represents the number of correctly predicted samples. As you can see from the table, the overall accuracy is 86.9%. Due to the uneven distribution of lithological data, this paper divides the training set and the test set according to the label ratio. To ensure that the proportion of various lithology in the divided test set and the training set is the same. However, the accuracy of WS is still low, and the reason is that lithological MS and lithological PS are similar to lithological WS, so it is easy to have deviations and misidentify them as WS.

Table 4. The confusion matrix for GA_LightGBM

Pred	SS	CSiS	FSiS	SiSh	MS	WS	D	PS	BS	Total
SS	44	9	1							54
CSiS		168	14					1		183
FSiS		12	135					4		151
SiSh				46	3	4		1		54
MS			1		41	11	1	1		55
WS	1		1	3	2	97		9	1	113
D						2	25	1		28
PS				1		12		121		135
BS						1		6	30	37
Precision	0.98	0.89	0.88	0.92	0.89	0.76	0.96	0.84	1.00	0.87
Recall	0.81	0.92	0.89	0.85	0.75	0.86	0.89	0.90	0.81	0.87
F1	0.89	0.90	0.89	0.88	0.81	0.81	0.93	0.87	0.90	0.87
Facies classification accuracy =0.873										

Table 5. Overall prediction results of each model

Model	Precision	Reacll	F1	Accuracy
Bayes	0.53	0.44	0.45	0.499
DNN	0.63	0.61	0.59	0.605
DT	0.61	0.60	0.60	0.598
GBDT	0.87	0.87	0.87	0.867
LightGBM	0.85	0.85	0.85	0.848
GA_LightGBM	0.87	0.87	0.87	0.873

As can be seen from Table 5, traditional shallow models fail to capture complex nonlinear correlations among features, resulting in poor performance across all evaluation indicators. In contrast, ensemble tree-based models including GBDT and LightGBM deliver remarkably better comprehensive performance than Naive Bayes, deep neural networks and single decision trees, which verifies the superior fitting capability of gradient boosting decision trees on structured datasets. GA-LightGBM, whose hyperparameters are globally optimized via genetic algorithm, achieves the highest overall accuracy of 0.873, with precision, recall and F1-score all reaching 0.87. This model balances outstanding recognition accuracy and stable discrimination between positive and negative samples. The results demonstrate that genetic algorithms can efficiently realize global hyperparameter search for LightGBM and effectively mitigate the performance limitations caused by random hyperparameter initialization of the original LightGBM model.

5. CONCLUSIONS

Experiments show that the lithology recognition model based on LightGBM algorithm is better than the traditional machine recognition method for the identification of complex carbonate lithology, and because the LightGBM algorithm adopts multi-threaded and distributed computing methods, the training time is greatly shortened, so the method can be applied to models with large data.

(i) The preprocessing of data and the hyperparameter optimization of the model have a great impact on the accuracy of the model. Due to the small number of features of the original logging data, the features of each data are expanded by feature shift, and the part of the data with incomplete features is eliminated. Subsequently, genetic algorithm and grid search are adopted for hyperparameter tuning, followed by validation of the prediction model.

(ii) Since the dataset is unbalanced, especially for samples with unstable levels. An unbalanced dataset can adversely affect the performance of a machine learning model. When the sample classes of the dataset are significantly skewed, the machine learning model will tend to predict more sample classes, resulting in poor precision and underfitting.

(iii) Based on the analysis accuracy, recall and average accuracy of F1, the accuracy of GBDT and LightGBM is above 85%, and the accuracy of LightGBM after genetic algorithm optimization is the highest, which can reach 87.3%. Compared with intersection plots and other ML algorithms (DNN, DT, Bayes), GBDT and LightGBM perform better, further verifying their reliability and validity in carbonate lithology identification.

CONFLICTS OF INTEREST

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

REFERENCES

[1] Zhang, C., Zhang, Y., Shi, X., Almpanidis, G., Fan, G., & Shen, X. (2019). On incremental learning for gradient boosting decision trees. Neural Processing Letters, 50(1), 957–987. https://doi.org/10.1007/s11063-019-09999-3

[2] Lee, S. H., Kharghoria, A., & Datta-Gupta, A. (2002). Electrofacies characterization and permeability predictions in complex reservoirs. SPE Reservoir Evaluation & Engineering, 5(3), 237–248. https://doi.org/10.2118/78662-P

[3] Puskarczyk, E. (2020). Application of multivariate statistical methods and artificial neural network for facies analysis from well logs data. Energies, 13(7). https://doi.org/10.3390/en13071548

[4] Li, J., Bao, H., Liu, S., Shi, P., Gao, F., Wei, L., Wang, L., & Yu, H. (2026). Machine learning for lithology identification in electrical image logging: A case study of the Chang 73 Sub-Member, Yanchang Formation, East Gansu region. Earth Science Informatics, 19, 139. https://doi.org/10.1007/s12145-026-02191-x

[5] Liu, X., Zhao, Y., Dai, Y., Wang, W., Cao, Z., Wang, P., & Wu, H. (2026). A Bayesian-optimized hybrid transformer-long short-term memory framework for lithology identification. Journal of Applied Geophysics, 253, 106400. https://doi.org/10.1016/j.jappgeo.2026.106400

[6] Peng, Y., Li, K., Zhu, Y., Xu, Z., Yang, P., & Sun, X. (2023). FL-XGBoost algorithm-based method for identifying sandstone and mudstone: a case study of Niuzhuang area in Shengli Oilfield. Petroleum Geology and Recovery Efficiency, 30(1), 76–85.

[7] Shi, S. M. (2024). Lithology identification of uranium drilling based on SMOTE algorithm logic-CNN. Journal of Biotech Research, 18, 53–64.

[8] Quan, R., Zhang, H., Zhang, D., Zhao, X., & Yan, L. Z. (2022). A novel hybrid method of lithology identification based on k-means++ algorithm and fuzzy decision tree. Journal of Petroleum Science and Engineering, 208. https://doi.org/10.1016/j.petrol.2021.10968

[9] Liang, W., Luo, S., Zhao, G., & Wu, H. (2020). Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM. Mathematics, 8(5), 765. https://doi.org/10.3390/math8050765

[10] Ali, S., Alshboul, O., Al Mamlook, R. E., & Hamedat, O. (2021). Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. Automation in Construction, 129. https://doi.org/10.1016/j.autcon.2021.103827

[11] Wang, D. N., Li, L., & Zhao, D. (2022). Corporate finance risk prediction based on LightGBM. Information Sciences, 602, 259–268. https://doi.org/10.1016/j.ins.2022.04.058

[12] Li, L., Yu, Y., Bai, S., & Chen, X. (2018). Towards effective network intrusion detection: A hybrid model integrating Gini index and GBDT with PSO. Journal of Sensors, 2018, 1–9. https://doi.org/1155/2018/1578314

[13] Henan University, Education Information Centre of Henan Province. (2019). On incremental learning for gradient boosting decision trees. Neural Processing Letters, 50(1), 957–987. https://doi.org/10.1007/s11063-019-09999-31

[14] Weng, T., Liu, W., & Xiao, J. (2019). Supply chain sales forecasting based on lightGBM and LSTM combination model. Industrial Management & Data Systems, 120(2), 265–279. https://doi.org/10.1108/imds-03-2019-0170

[15] Zhang, Q., Wang, H., Liu, Z., & Liao, W. (2026). Predicting rare earth extraction efficiency with organophosphorus ligands: A data-driven machine learning approach. Journal of Rare Earths, 44(7), 2204–2214.

[16] Wang, Y. F. (2026). Model, data, and knowledge-driven geophysical inverse problems and intelligent computing. Science China Earth Sciences, 1–38.