Journal of Applied Geophysics 194 (2021) 104443

Contents lists available at ScienceDirect

Journal of Applied Geophysics

journal homepage: www.elsevier.com/locate/jappgeo

Volcanic lithology identification based on parameter-optimized GBDT algorithm: A case study in the Jilin Oilfield, Songliao Basin, NE China

Zhichao Yu [a], Zhizhang Wang [a,*], Fancheng Zeng [b], Peng Song [b], Bestman Adjei Baffour [a], Peng Wang [c], Weifang Wang [a], Ling Li [a]

[a] College of Geosciences, China University of Petroleum, Beijing 102249, China
[b] PetroChina Jilin Oil Field E&P Research Institute, Songyuan, Jilin 138000, China
[c] Sinopec Petroleum Exploration and Production Research Institute, Beijing 10083, China

ARTICLE INFO

Keywords:
Volcanic
Lithology identification
Songliao Basin
Ensemble learning algorithm
GBDT

ABSTRACT

The reservoir rocks in the volcanic strata of the Jilin oil field are characterized by great complexity and diversity in composition and structure of lithology. To enhance the rate of lithology identification in subsurface is very laborious. However, lithology identification is often ignored in quantitative studies, though it is the basis for reservoir characterization. In this paper, an ensemble learning algorithm named gradient boosting decision tree (GBDT) was used to establish the classification model for the volcanic lithology identification of the Lower Cretaceous Yingcheng Formation in the Songliao Basin, NE China. At the same time, support vector machine (SVM), logistic regression (LR) and decision tree (DT) classification models were also adopted in contrast with the classification accuracy of GBDT model. Subsequently, the optimal key parameters for each model were determined by employing validation curves and GridSearchCv. These results indicate that the GBDT model is superior to the single classifier and can accurately distinguish the lithologic interface of breccia tuff and rhyolite. Moreover, it also has better recognition ability for thin layer. It was concluded that the ensemble learning algorithm GBDT has significantly enhanced the accuracy of lithology identification and can be used as a lithologic identification technology.

1. Introduction

According to history of petroleum, the major targets of petroleum exploration and exploitation around the globe are often clastic and carbonate reservoirs. Less concentration is given to the deeply buried volcanic reservoirs due to their complicated lithologies and lithofacies characteristics (He et al., 2020). Since the first discovery of volcanic hydrocarbon reservoir in San Joaquin basin, California, USA in 1887, significant advances have been made in volcanic hydrocarbon exploration. More than 300 volcanic or volcanic-related reservoirs have been found worldwide, of which 169 volcanic hydrocarbon reserves have been proven (Petford and Mccaffrey, 2003). Statistically, there exist abundant oil and gas resources in global volcanic reservoirs, containing a total of 65.5×10^8 tons of proven oil reserves and 36×10^8 tons of gas (Schutter, 2003). An example is the Cristales oil field of the North Cuba basin where more than 3425 tons of oil per day (t/d) were successfully extracted from depths exceeding 2000 m (Zou et al., 2008). Additionally, in the Yoshii-Kashiwazaki gas field in the Niigata Basin, single wells produce up to 49.5×10^4 cubic meters of gas per day (m^3/d) through the rhyolite reservoir (Zhang et al., 2008; Wang et al., 2015). China has also initiated many volcanic hydrocarbon explorations at depths greater than 3000 m (Zishu and Wu, 1994; Mao et al., 2015; Feng, 2008). China's volcanic hydrocarbon exploration target has the characteristics of larger size and deeper-layer formations compared with other analogous volcanic oil and gas fields throughout the world (Jia et al., 2016). Deep-layered formations, especially volcanic strata, represent the major targets of exploration in the Songliao Basin. Several huge gas fields with reserves of over 100 billion cubic meters have been found, including Changling, Yingtai, and Wangfu gas fields (Zhang et al., 2015; Zhang et al., 2017). As the fourth strategic energy succession region of 100 billion cubic meters of natural gas exploration of Jilin oilfield, industrial gas flow has been obtained from wells in the strata of Huoshiling formation, Shahezi formation and Yingcheng formation in Dehui fault depression (Fig. 1a, b). The main formation of this gas field is Cretaceous in age, buried at a depth of more than 3000 m. However, the complicated geological conditions of volcanic gas reservoir have resulted in

* Corresponding author.
E-mail address: wang_zhizhang@126.com (Z. Wang).

https://doi.org/10.1016/j.jappgeo.2021.104443
Received 9 December 2020; Received in revised form 3 July 2021; Accepted 24 August 2021
Available online 28 August 2021

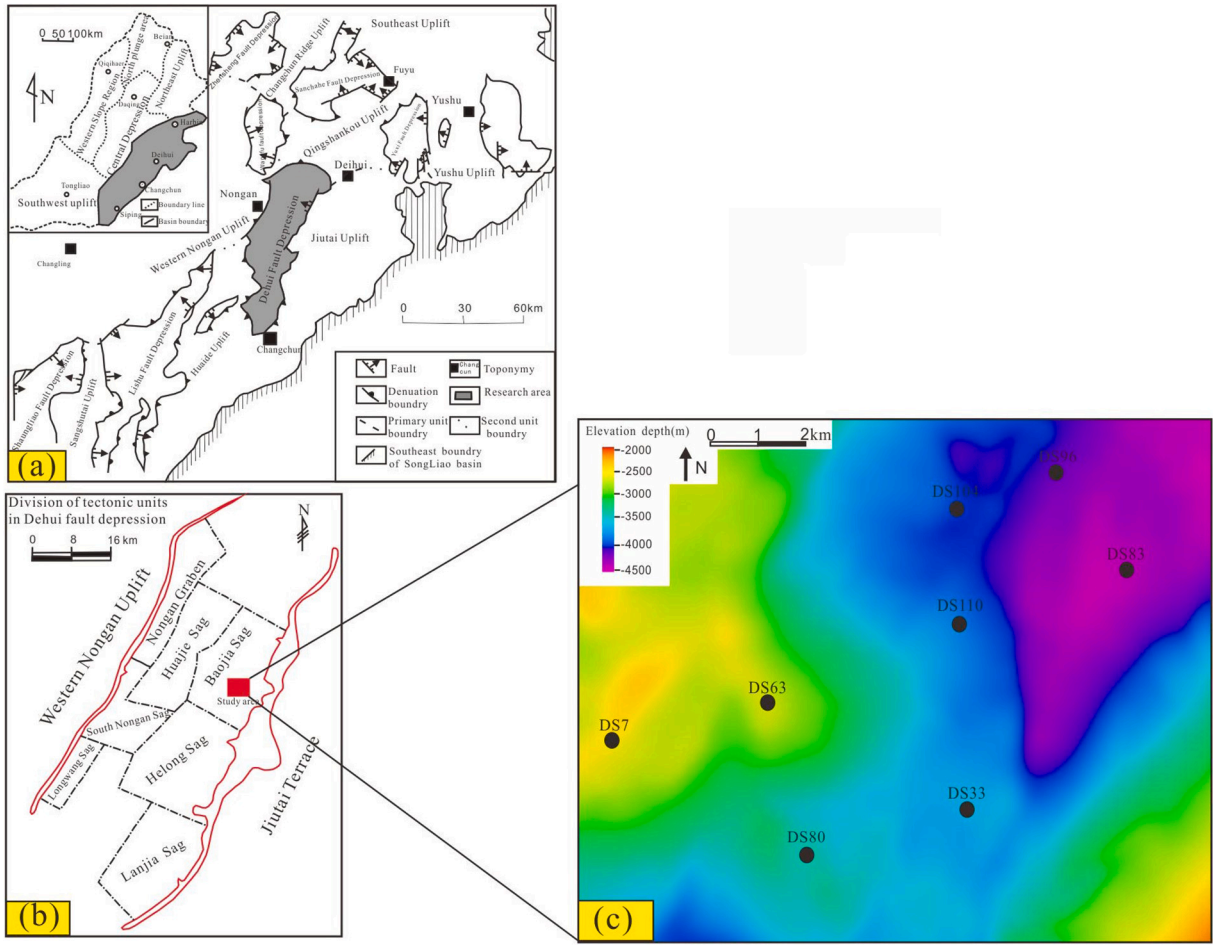


Fig. 1. Location of the Dehui Fault Depression and division of its tectonic units. (a) Tectonic components of the Songliao Basin, including two uplifts, two depressions, one plunge and one slope area. The Dehui Fault Depression is located in the Southeast Uplift of the Songliao Basin. (b) The Baojia Sag, located in the northeastern part of the Dehui Fault Depression of the Southeast Uplift(modified from Jilin Oilfied). (c) The depth-domain map of Cretaceous strata in the study area.

great complexity and diversity in composition and structure of volcanic lithology and lithofacies.

In recent years, researchers have reached broad agreement about the physical properties and reservoir space of volcanic reservoirs in different regions(Sun et al., 2019; Feng et al., 2018; Gong et al., 2017). However, for volcanic oil reservoirs, no practical mature methods or techniques to effectively identify the volcanic lithology are available. In the research of igneous reservoir, lithology identification is the basis of reservoir characterization (Ye et al., 2017; Han et al., 2018). Different types of volcanic lithology have distinct petrological characteristics and mineral assemblage types, and different logging numerical characteristics and pattern combination relationship of logging curves. It is very challenging to improve the rate of lithology identification in igneous reservoirs. Core analysis is the most direct and effective method for identifying the lithology of volcanic rocks, but due to the high cost of coring, it is almost impossible to take cores in every single well. It is, therefore, crucial to make full use of conventional logging data to identify the lithology of volcanic rocks. The lithology in the study area is predominantly acidic volcanic rocks, therefore, the change in composition is relatively small, while variation of structure exhibits different logging response of resistivity (RLLD, RLLS), density (DEN), acoustic slowness (AC), and compensated neutron log (CNL) values. A more traditional method of identifying volcanic lithology based on well log data is the use of cross plots (Zhang et al., 2017). At present, many machine learning methods have been introduced into volcanic lithology identification, including neural network, support vector machine (SVM), logistic regression (LR) and decision tree (DT). These methods have various applications in different research areas, but they all have their own limitations. Supported vector machine algorithm is difficult to implement for large-scale training samples and neural network is easy to fall into local optimum (Alpaydin, 2014; LeCun et al., 2015). Logistic regression and decision tree are easy to under fit, resulting in low classification accuracy (Guoyin et al., 2018; Guo and Liu, 2016; Camila et al., 2018).

The goal of this study is to distinguish the acidic volcanic rocks with similar composition but different structures in deep buried volcanic strata within the working area of Baojia sag, Dehui fault depression (Fig. 1b). Based on the conventional logging data of the working area, this paper innovatively introduces the ensemble learning algorithm named gradient boosting decision tree (GBDT) into the study of lithology identification. Experiments show that the GBDT algorithm has higher classification performance than the traditional classification model and can be used as a lithologic identification technology. The results allow us to improve our analysis of volcanic reservoir intervals and subsequent identification of volcanic lithofacies.

2. Geological setting

The Dehui Fault Depression is 4053 km^2 in size and is located in the middle of southeast uplift of Songliao Basin. The Dehui Fault Depression is a secondary tectonic unit of Songliao Basin, which is east of the Nongan Uplift, south of the Wangfu Fault Depression, west of the Jiutai Uplift, and north of the Huaide Uplift and the Lishu Fault Depression (Fig. 1a). The Dehui Fault Depression can be further divided into seven

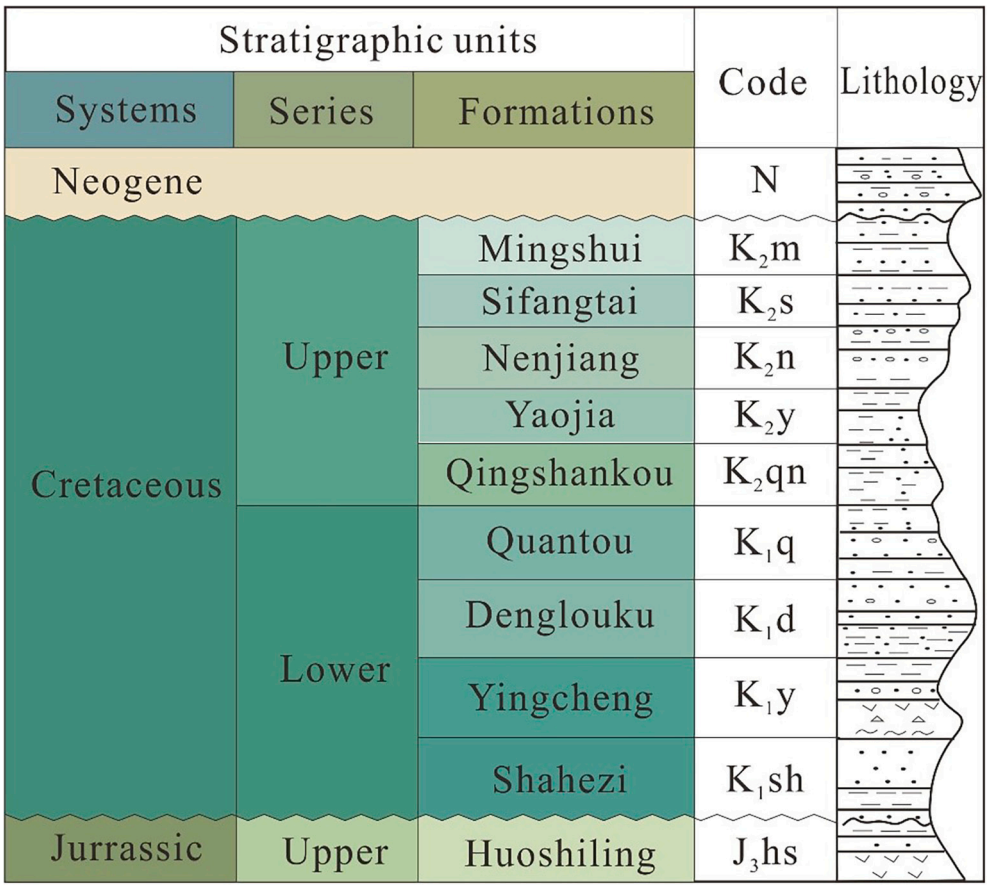


Fig. 2. Stratigraphic column of the Jilin Oilfield.

sub structural units, namely the Nongan graben, the Huajia sag, the Baojia sag, the Helong sag, the Lanjia sag, the Nongan South sag and the Longwang sag. The study area, which measures 300 km^2, is located in the Baojia Sag (Fig. 1b).

Within this gas field, hydrocarbon reservoirs have been identified in Cretaceous and Jurassic strata. Natural gas production from Cretaceous reservoirs accounts for a large amount of total gas production (Libin et al., 2006). The Cretaceous strata can be divided into two series (Upper and Lower Cretaceous) and nine formations (Fig. 2). From bottom to top, these Cretaceous strata are divided into the Shahezi Formation (K1sh), Yingcheng Formation (K1y), Denglouku Formation (K1d), Quantou Formation (K1q), Qingshankou Formation(K2qn), Yaojia Formation (K2y), Nenjiang Formation(K2n), Sifangtai Formation(K2s) and Mingshui Formation(K2m) (Yang et al., 2019; Jing and Liande, 2016).

The Baojia Sag experienced multiple stages of tectonic movements in the early Yingcheng formation(Shuangfang et al., 2010). Meanwhile, multi-phase volcanic eruption has generated large sets of volcanic construction accompanied with the strong tectonic movement (Huiguang et al., 2011),thus resulting in forming volcanic structural traps in the local area, which provide favorable reservoir for hydrocarbon enrichment of Yingcheng formation.

3. Data and method

The study area consists of the Baojia sag of the Jilin Oilfield (Figs.1b and 1c). To date, a total of 26 wells have been drilled in the Cretaceous strata, including 18 vertical wells, and 8 inclined wells, with depths exceeding 3000 m. Core samples (including thin sections) and imaging logging data (mostly in the form of formation microscanner images (FMI)) provided direct evidence for volcanic rock identification. Hence, conventional logging data, which are calibrated by cores and FMI, could serve as sufficient information for establishing machine learning model for identification of volcanic rocks.

Ensemble learning is a new machine learning paradigm, which

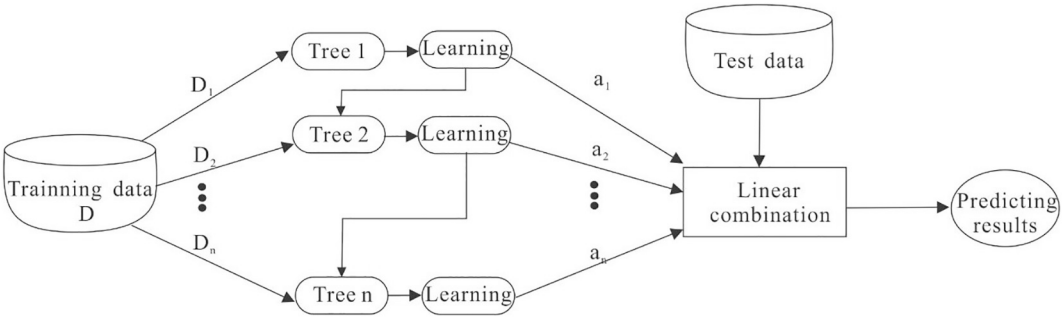


Fig. 3. Workflow of GBDT algorithm.

Fig. 4. Characteristics of volcanic reservoir cores: (a) Rhyolite, marked by vesicular structure, well DS63, 2394.60 m. (b) Dacite core interval, a series of steep fractures were encountered, well DS7, 2083.00–2085.70 m. (c) Tuff, steep fracture filling calcite, well DS83, 4086.08 m. (d) Tuffite, well DS80, 2562.00 m. (e) Ignimbrite, characterized by welded structure, with coarse grain size, well DS80, 2582.50 m. (f) Breccia tuff, finer grain size than ignimbrite, well DS80, 2543.00 m (g) Diabase, fine to medium grain size, well DS80, 4158.00 m.

constructs multiple learners to solve the same problem (Avnimelech and Intrator, 1999; Elghazel and Aussem, 2015; Miyoshi et al., 2006). By referring to the gradient descent method, its underlying principle which is training the newly added weak classifier according to the negative gradient information of the loss function of the current model is applied (Fig. 3; Li et al., 2018; Sakhnovich, 2007). Subsequently, the trained weak classifier will be appended to the existing model. The GBDT algorithm can be explained as the adoption of a decision tree as the weak classifier in a gradient boosting algorithm (Liao et al., 2016; Jin Yuan et al., 2018; Xin et al., 2019). The workflow of GBDT algorithm is as follows:

(1) Initializing the model with constant$\gamma 0$

$$F_0(x) = argmin_{\gamma 0} \sum_{i=1}^{n} L(yi\gamma 0) \tag{1}$$

(2) For m from 1 to M:

(a) The negative gradient of loss function is used to approximate the value of residual in the current model$Fm-1(x)$:

$$r_{im} = -\left[\frac{\partial L(yiF(xi))}{\partial F(xi)}\right]^{F(x)=Fm-1(x)} \quad for\ i = 1, 2, \ldots, n. \tag{2}$$

(b) In accordance with the training set$\{(xi, rim)\}_{i=1}^{n}$, a base learnerhm$(x)$is constructed to fit the pseudo residual.

(c) The multiplierγ_m is calculated by the following one-dimensional optimization problem:

$$\gamma_m = arg_{\gamma} min \sum_{i=1}^{n} L(yiF_{m-1}(xi) + \gamma h_m(xi)) \tag{3}$$

(d) The model is then updated

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{4}$$

(3) Output$FM(x)$stands for the prediction of a strong classifier composed of a series of weak decision tree models.

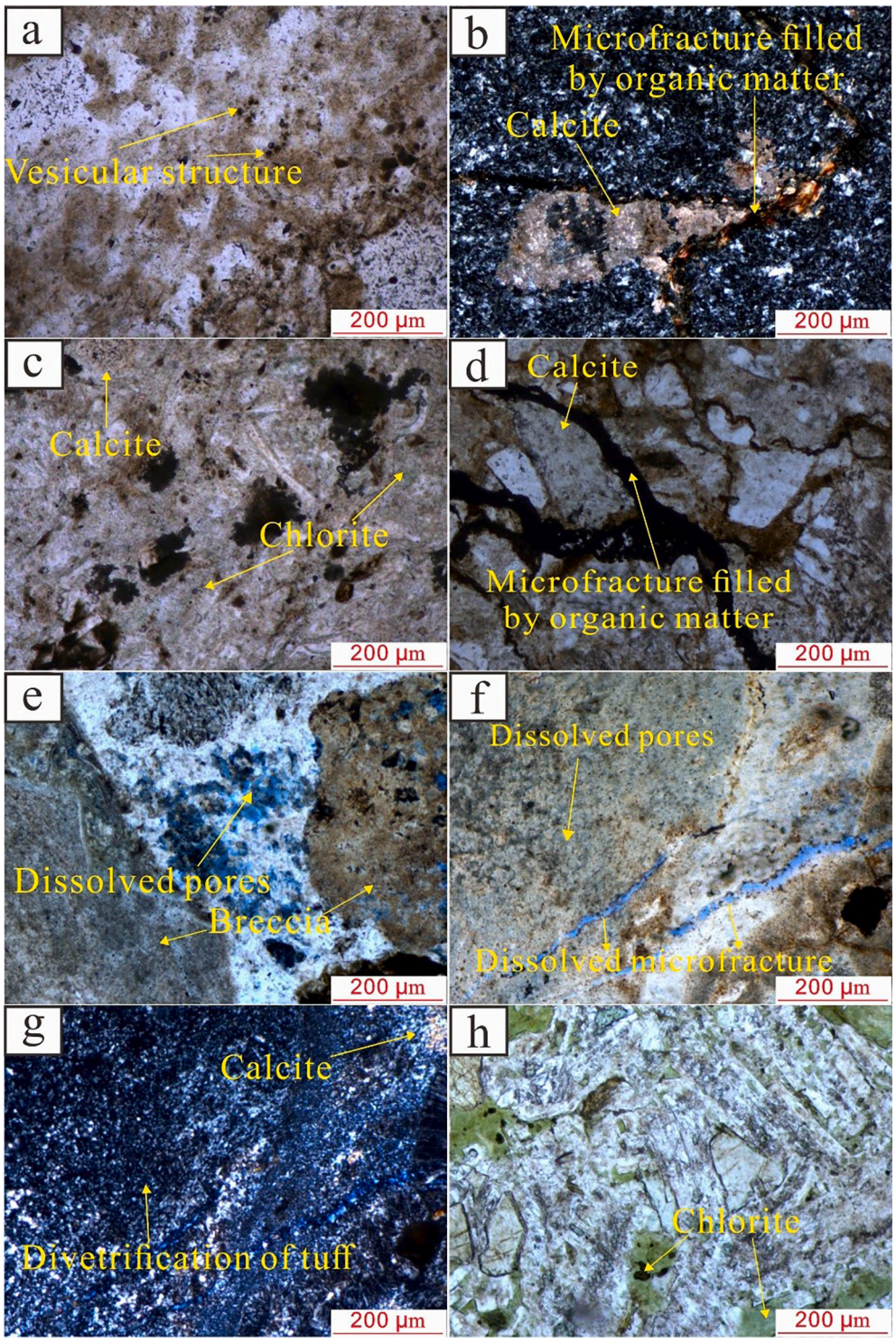


Fig. 5. Thin-section photomicrographs: (a) Spherulite rhyolite, developing vesicular structure, well DS110, 3220.00 m. (b) Dacite, micro-fractures filled by organic matter, carbonate metasomatism were displayed locally, well DS17, 2237.34 m. (c) Tuff, chloritization and carbonate metasomatism can be observed, well DS104, 3060.00 m. (d) Tuffite, no porosity were observed under microscope, carbonate metasomatism and micro-fractures filled by organic matter were recorded, well DS33, 2897.60 m. (e) Breccia tuff, tuff fillings were obviously dissolved, well DS80, 2740.60 m. (f) Ignimbrite, under plane polarized light, dissolved microfractures and dissolved pores (generated by devitrification of tuff, seen in Fig. 7g) were encountered, well DS80, 2522.00 m. (g) Ignimbrite, under perpendicular polarized light, devitrification of tuff and carbonate metasomatism were observed, well DS80, 2522.00 m. (h) Diabase,the chloritization in the feldspar was developed and micro pores were observed, well DS80, 4158.00 m.

4. Application and results

4.1. Rock type and characteristics

Detailed lithology analysis was performed on the studied wells on whole cores and core plugs of the Yingcheng Formation. Accordingly, these core samples were analyzed with casting and conventional thin section. Additionally, imaging logging, which can obtain high-resolution images (as data are collected by vertically scanning the formation at 2.5 mm intervals) and identify structure characteristics of volcanic rocks within the borehole, is an effective tool for us to better understand the development of volcanic reservoirs. Through the above dataset, there are mainly three categories (volcanic lava, pyroclastic rock and intrusive rock) and seven kinds (rhyolite, dacite, tuff, tuffite, breccia tuff, ignimbrite and diabase) of volcanic rock types in the study area.

Volcanic lava mainly consists of rhyolite and dacite. Core from (a) depth of 2394.6–2394.8 m in Well DS63 in the Yingcheng Formation represents a typical rhyolite development section, the core sample is generally greyish-white and features the development of vesicular

Table 1
Well log responses of different volcanic lithology.

Volcanic lithology	GR (API)	AC (μs/ft)	RHOB (g/cm^3)	CNL (%)	RLLD (Ω·M)	RLLS (Ω·M)
Rhyolite	128.4–217.8	53.0–59.3	2.06–2.66	1.05–13.02	96.1–1990.6	73.1–1686.8
	172.8	55.8	2.59	3.28	874.6	248.4
Dacite	130.8–201.1	54.2–63.7	2.50–2.67	3.02–8.57	35.8–332.7	43.5–610.3
	153.5	57.9	2.61	5.22	181.3	251.2
Tuff	80.3–321.3	50.2–73.7	2.04–2.71	3.54–20.05	44.4–1986.2	41.2–1981.5
	159.7	57.3	2.54	8.75	612.1	586.4
Tuffite	107.7–266.0	51.3–74.8	2.11–2.64	5.86–38.14	22.9–1961.5	27.4–1625.5
	167.2	57.6	2.53	14.0	346.6	314.6
Breccia tuff	76.8–324.1	51.8–67.1	2.02–2.70	4.14–23.14	30.6–1796.8	35.2–1852.0
	167.2	57.3	2.57	10.47	345.4	334.3
Ignimbrite	35.3–99.7	56.3–83.9	2.18–2.69	5.87–38.01	8.6–733.1	7.8–523.4
	606	71.9	2.54	23.46	48.2	36.9
Diabase	22.7–84.6	48.7–69.7	2.12–2.91	0.14–30.99	10.3–1948.4	7.9–1667.1
	39.7	55.4	2.71	11.00	257.4	149.9
Minimum-Maximum						
Average						

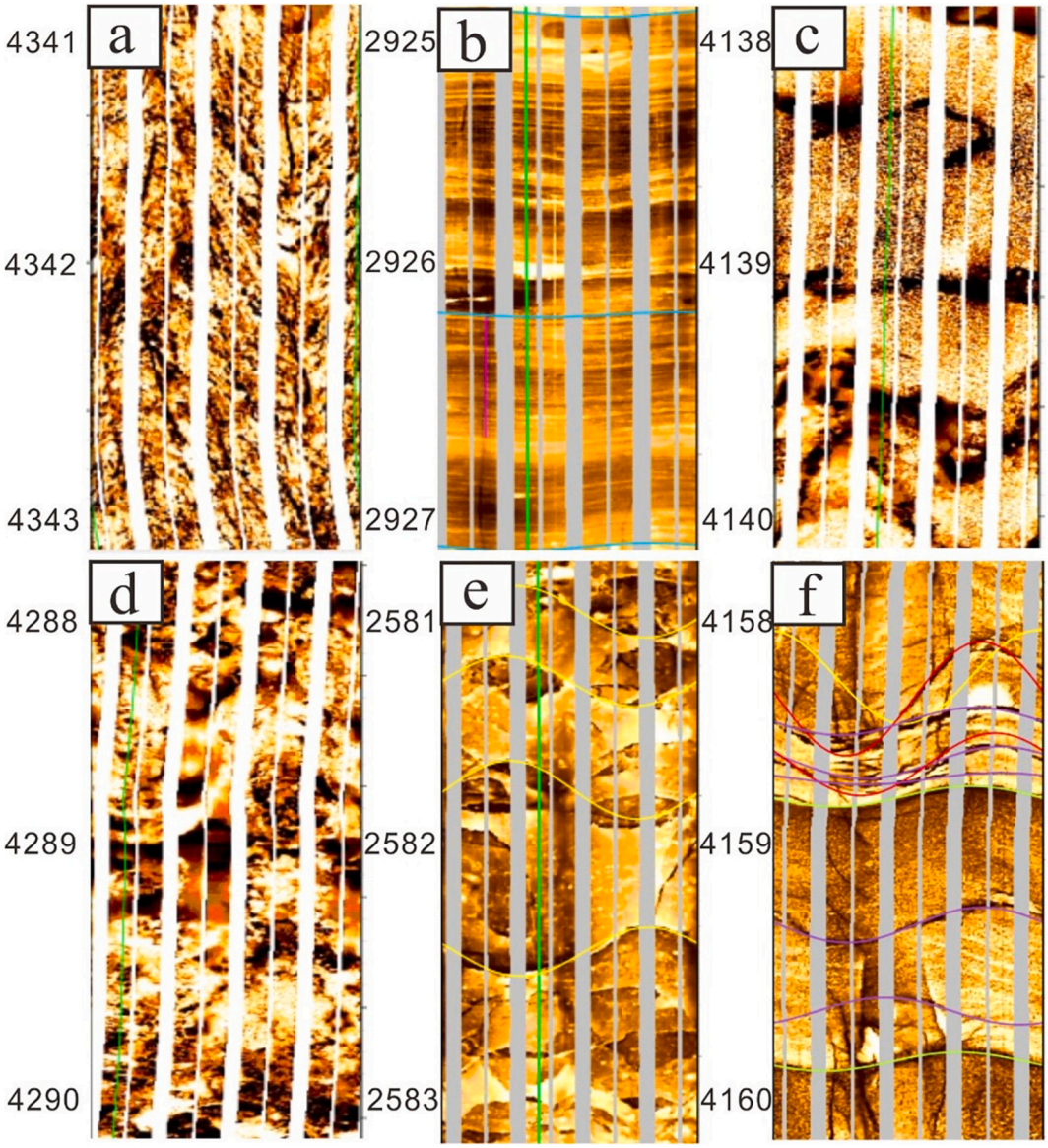


Fig. 6. FMI images of different volcanic lithology: (a) Rhyolite, showing the development of flow structure, well DS83,4341.00–4343.00 m. (b)Tuffite, featured by layering development, well DS80,2925.00–2927.00 m. (c)Tuff with tuffaceous structure, well DS83,4138.00–4140.00 m. (d) Breccia tuff, porphyritic structure, well DS83,4288.00–4290.00 m. (e) Ignimbrite, marked by blocky structure, well DS80,2581.00–2583.00 m. (f) Diabase, blocky structure with well-developed fracture, well DS80,4158.00–4160.00 m.

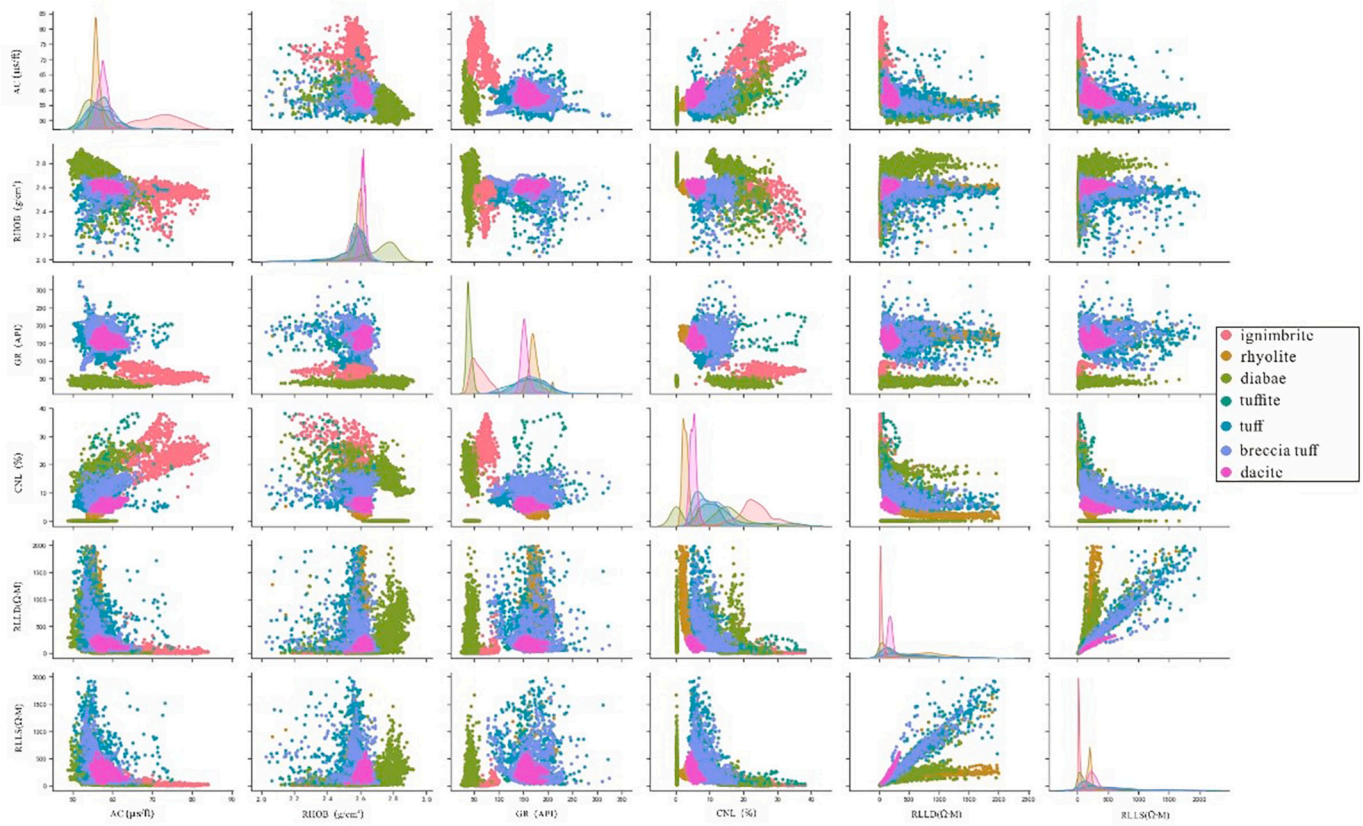


Fig. 7. The multi-curve cross plot of volcanic lithology.

structure (Fig. 4a). It shows a very characteristic flow structure (Fig. 6a) and is marked by vesicular characteristic under microscope (Fig. 5a). Conventional logging was characterized by high natural gamma ray radiation (>150API), low acoustic slowness, high density, low neutron porosity, low resistivity and moderate-to-low compensated neutron log values (Table 1). The dacite core was obtained from a depth of 2083.00 to 2085.70 m in Well DS7 and the core section contains a series of steep fractures (Fig. 4b). Additionally, micro-fractures filled by organic matter were also recorded during the thin section studies (Fig. 5b). Compared with rhyolite, which are relatively denser, the dacite interval exhibited lower resistivity (R_{LLD}, R_{LLS}), lower density (DEN), lower natural gamma ray radiation (GR), higher acoustic slowness (AC), and higher compensated neutron log (CNL) values (Table 1).

Pyroclastic rock comprises of tuff, tuffite, breccia tuff and ignimbrite. Their conventional well logging values are recoded in Table 1. Intersection of two kinds of conventional logging data shows that there are overlapping zones over different lithology (Fig. 7). The cores spanning a depth of 2500 to 4200 m were obtained from Well DS83 and DS80 in the Yingcheng Formation. Tuff and tuffite have greyish-white and grey color (Fig. 4c, d), with grain size less than 2 mm, and the tuff core has been partially cemented by calcite (Fig. 4c) while the tuffite has the characteristic of layering development (Fig. 4d). In contrast, breccia tuff and ignimbrite have greenish grey or dark color and breccia makes up 15% portion of the core sample (Fig. 4e, f). Petrographic studies on cores and thin section samples obtained from the Yingcheng Formation indicated that the Yingcheng Formation has been subjected to various diagenetic

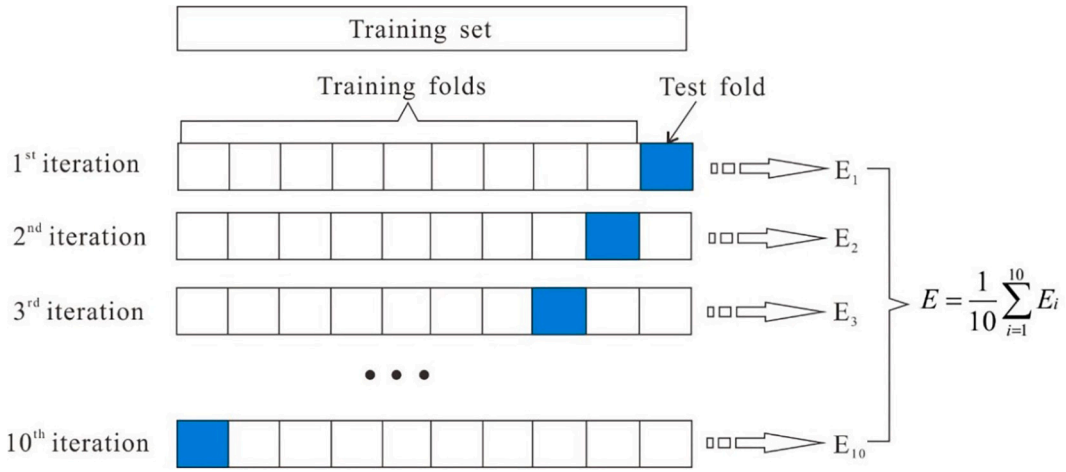
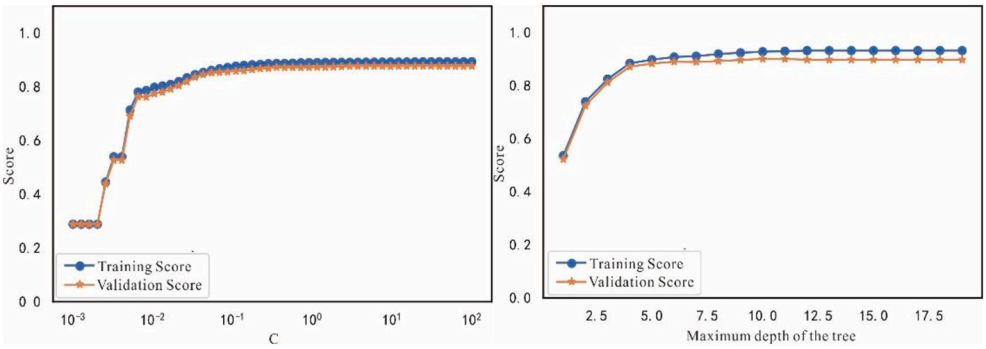
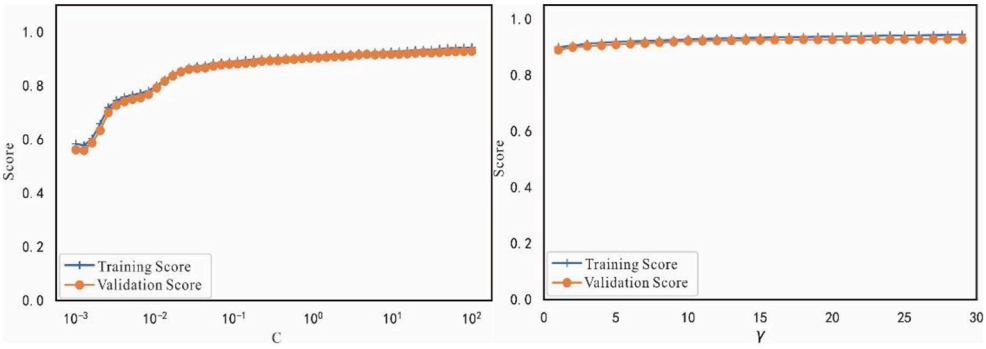


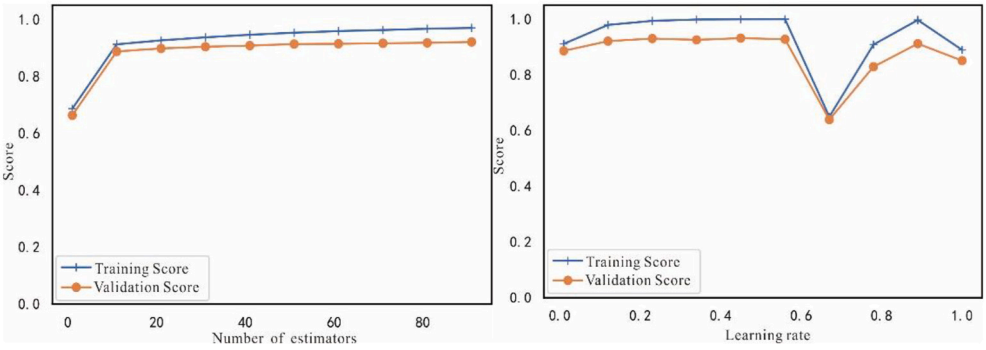
Fig. 8. Sketch of 10-folder cross validation.



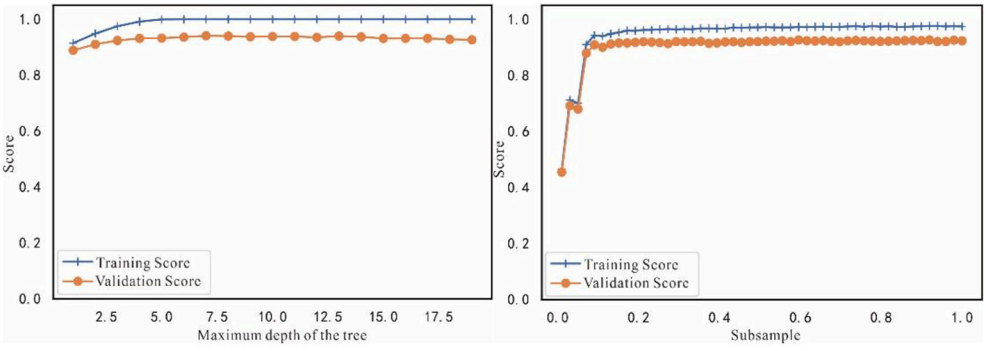
(a) Penalty C of LR algorithm (b) Maximum depth of DT algorithm



(c) Penalty C of SVM algorithm (d) Parameter γ of SVM algorithm



(e) Number of estimators of GBDT algorithm (f) Learning rate of GBDT algorithm



(g) Maximum depth of GBDT algorithm (h) Subsample of GBDT algorithm

Fig. 9. Validation curves for key parameters of different classifiers.

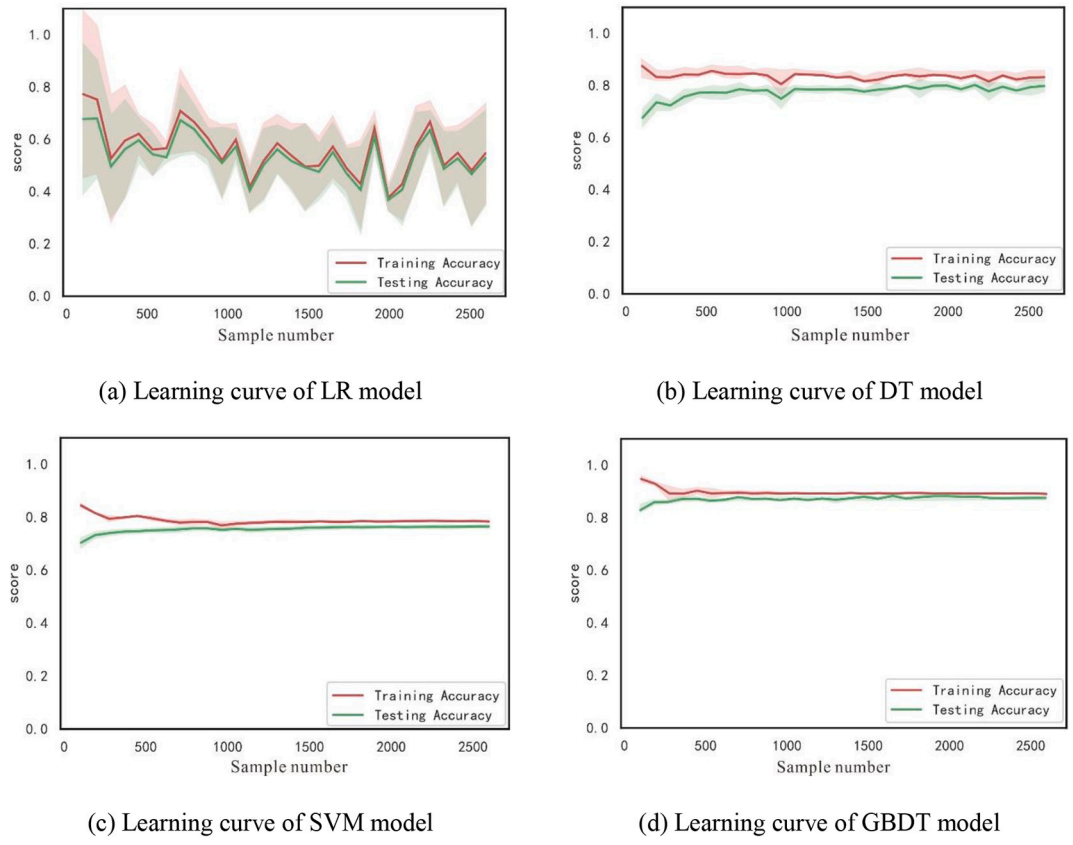


Fig. 10. Learning curve of different classifiers.

alterations. Dissolution is one of the most important diagenetic processes in terms of porosity generation in the Yingcheng Formation. It is recorded in most thin sections in the pyroclastic rock and is marked by feldspar dissolution facies (Fig. 5e, f). Moreover, the Yingcheng Formation has been greatly affected by metasomatism during the burial within the area. Carbonate metasomatism are the main features observed in thin sections of this formation (Fig. 5g). The pyroclastic rocks appear as distinct structure on FMI image. Tuffite, tuff, breccia tuff and ignimbrite are characterized by lamellar, tuffaceous, porphyritic, and blocky structure, respectively (Fig. 6b, c, d, e).

Diabase is the most common intrusive rock in the study area, which belongs to basic hypabyssal intrusive rock with fine to medium grain size through core observation (Fig. 4g). Also, the feldspar is alternated by chlorite and micro pores were observed under microscope (Fig. 5h). The diabase is imaged as blocky structure on FMI (Fig. 6f).

4.2. Model establishment of lithology identification

The fundamental data of lithology analysis is various logging curves related to lithology, therefore, conventional logging data, which are calibrated by cores and FMI, can serve as training data to identify lithology. On the basis of logging response mechanism of petrophysics, 6 conventional logs (GR, AC, R_{LLD}, R_{LLS}, DEN, CNL), which are sensitive to lithology, were selected to be the original input features of the model, whereas the output is volcanic lithology. Before the experiment, the training data were preprocessed to improve the quality of the data: the abnormal data caused by human or external factors were eliminated. Subsequently, the data were normalized from 0 to 1, avoiding the influence of dimension on prediction results. These 'calibrated' data were collected from Well DS80, Well DS83, Well DS7 and Well DS63. A total of 9724 data points were obtained from the Yingcheng formation. 6806 (approximately 70%) data points were randomly selected as training samples and the rest served as test samples to check the efficacy and reliability of the proposed method. The methods of DT, SVM, LR, and GBDT algorithms were used respectively to construct the model.

4.3. Parameter optimization

4.3.1. Cross validation

Cross validation is a common approach in machine learning to build models and verify model parameters. It can be explained as follows: the training data set were grouped into training and validation data sets. The training set was used to train the classifier while the validation set was employed to test the model trained by the training set (Krogh and Vedelsby, 1995). In this paper, 10-folder cross validation (Fig. 8) was adopted to conduct the experiment. The original training data were grouped into 10 sections (generally equally), a validation set was set for each subset of the data, and the rest of the 9 groups were taken as the training set. Consequently, we can get 10 models. Average classification accuracy of the validation set of the 10 models is regarded as the performance index of the classifier under the 10-folder cross validation. 10-folder cross validation can effectively avoid the occurrence of over fitting and under fitting. Furthermore, the final result is more convincing in the sense that almost all samples in each round have been used for training the model.

4.3.2. Validation curves

The validation curve determines whether the model is over-fitting or under-fitting by virtue of the values of some hyperparameters by drawing the curve of a single hyperparameter with training score and cross validation score (Airola et al., 2011). In instances where the training score and the cross validation score are very low, the classifier may not be suitable due to possible occurrence of under-fitting. Over-fitting may also occur if the training score is high and the validation score is low. The performance curves of vital hyperparameters for different model are drawn in Fig. 9.

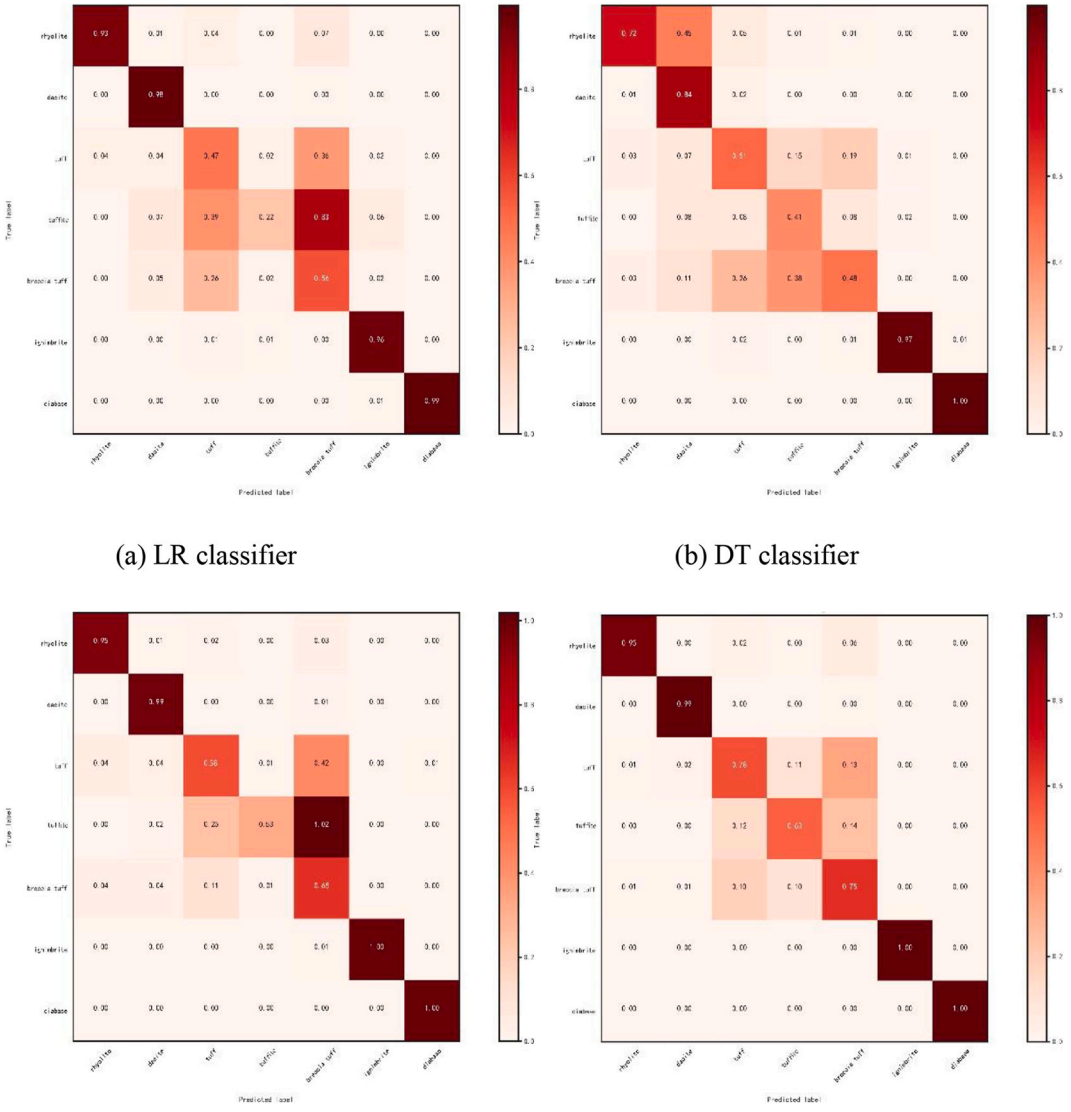


Fig. 11. Confusion matrix of $f1$ score over different classifiers.

4.3.3. Grid search

By selecting a reasonable parameter range, the optimal parameter set of the model can be sought out automatically by using Grid search (Ji et al., 2008). The essential idea is to divide the parameters to be optimized into grids in a certain space, searching for the optimal parameters by traversing all the intersections in the grid. In accordance with the given model, cross validation can be automatically carried out. Additionally, the validation curve can help specify the search scope, thus considerably increasing the efficiency. Table 2 shows the parameters requiring adjustment of each model and corresponding optimal parameters.

4.4. Model evaluation

There are many evaluation indexes in machine learning over classification problems, such as learning curve, F1 score, ROC curve. Fig. 10 shows the learning curve of the four classifiers. The result shows that SVM has the best performance in a single classifier, while LR and DT have the problem of underfitting. Meanwhile, the GBDT model outperforms the single classifier and the score can reach approximately 0.9.

With regard to binary classification, the concerned classes are ordinarily designated as positive class while other classes are designated as negative class. The prediction results of the classifier on the test set are only true or false. Confusion matrix is a visual tool in supervised learning, which is mainly employed to compare the prediction results and genuine results (Mitchell, 2003; Yang et al., 2019). Each row in the matrix represents the predicted label, and each column represents the true label. The confusion matrix of classification results is shown in Table 3.

However, accuracy is not the only index used in unbalanced binary classification. PrecisionPr, recallReand$f1$score are also employed as important indexes to evaluate the performance of the algorithm. They are defined as follows:

$$Pr = N_{TP}/(N_{TP} + N_{FP}) \tag{5}$$

$$Re = N_{TP}/(N_{TP} + N_{FN}) \tag{6}$$

$$f1 = 2(1/Re + 1/Pr)^{-1} = 2\frac{\mathrm{PrRe}}{Pr + Re} \tag{7}$$

Where precisionPr represents the true positive proportion of the samples predicted as positive (i.e. how many of the found were correct

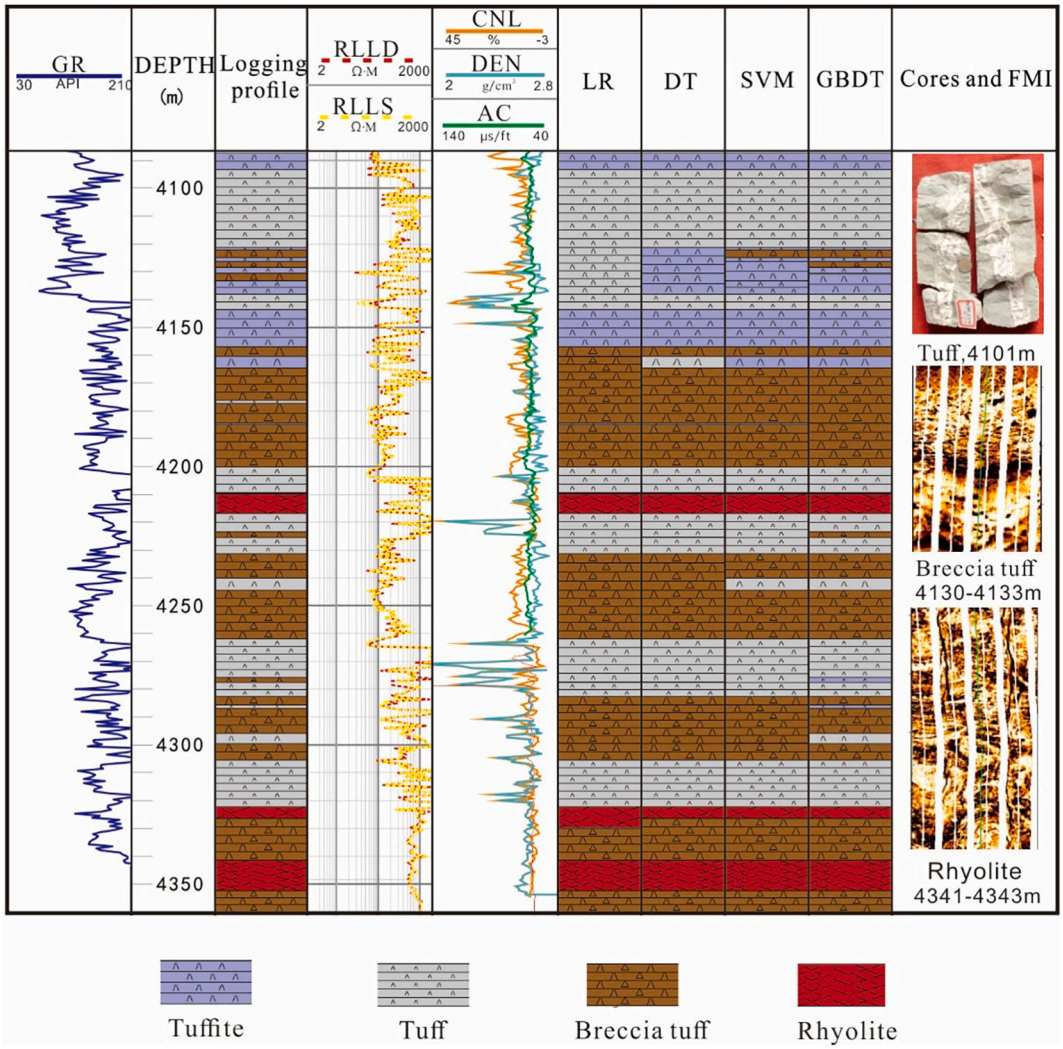


Fig. 12. Lithology identification results of well DS83.

hits). Recall literally is how many of the true positives were recalled (found), i.e. how many of the correct hits were found.*f*1 score is the harmonic average of precision and recall. It is commonly used as evaluation metrics in many cases because there is a reciprocal relationship between precision and recall. A stable algorithm should simultaneously maximize both precision and recall. Blindly increasing either precision or recall may be of no significance since *f*1 score will be high under circumstance of either high precision or recall.

Tables 4 and 5 show the precision, recall, and*f*1score for each type of lithology on the test set over different classifiers. The result showed that SVM has the best performance in a single classifier, and the average value of*f*1score reached 81.2%, while LR and DT can only reach 72.8% and 70.2%, respectively. Evidently, GBDT has significantly enhanced the efficiency of the classifier by increasing the *f*1 score up to 87.1%.

4.5. Results

Fig. 11 shows the confusion matrix of *f*1 score of 7 kinds of lithology over different classifiers on test samples. Among the single classifier, the *f*1 score of rhyolite and dacite are relatively high (exceeding 90%), while tuffite has the lowest, with 41%, 53% and 22% in the DT, SVM and LR, respectively. Evidently, the model constructed by GBDT algorithm performs superior to the three single classifiers, the *f*1 score of tuffite increased to 63%. The average accuracy of each lithology can reach up to 85.6%.

Fig. 12 shows the result of volcanic lithology identification of well DS83 from a depth of 4086 to 4396 m. Identification results of different classification models are contrasted with the logging profile (which has been accurately calibrated by FMI) of this well. It can be seen that the GBDT model has the best performance on lithology identification. It has a satisfactory application on breccia tuff, tuff and rhyolite, which can accurately distinguish their lithologic interface. Nevertheless, the tuff and tuffite are easily misjudged and intersected. The confusion matrix also indicates that the two lithologies tend to be obscure, which is mainly due to the little difference in mineral composition contents with distinct origin. Additionally, thin layers with small thickness can be identified by the GBDT model. Considering the breccia tuff interval developed within a buried depth of 4225.1–4226.7 m, only the recognition results of GBDT model is consistent with the lithologic interpretation results of logging. Fig. 13 shows the result of volcanic lithology identification of well DS107 from a depth of 2800 to 2950 m. It can be seen that the GBDT and SVM model has the best performance on lithology identification. It has a satisfactory application on breccia tuff and ignimbrite, while DT and LR model tend to misjudge the tuff and breccia tuff.

5. Discussion

Volcanic reservoirs in deep buried Cretaceous strata are heterogeneous, making effective exploration relatively challenging. Lithology

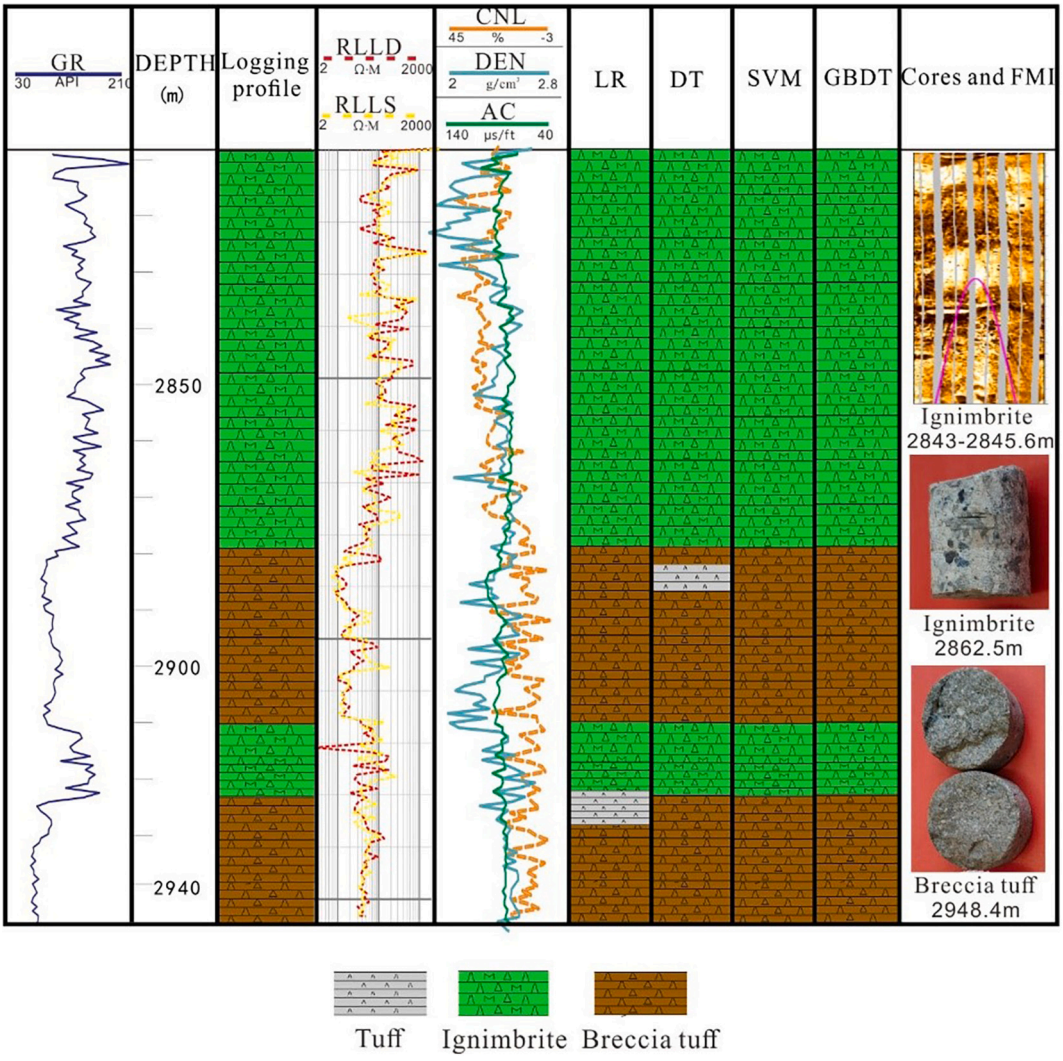


Fig. 13. Lithology identification results of well DS107.

Table 2
Critical hyperparameters and optimal parameter values for each classification model.

Classification model	Optimized parameter	Search range	Optimal parameter
DT	Feature selection criterion	Gini/ Entropy	Entropy
	Maximum depth of the tree	1– 20	10
SVM	Parameter γ of kernel function	1– 50	19
	Penalty C	10^{-3}– 10^{2}	82.86
LR	Regularization strategy	L1/L2	L1
	Penalty C	10^{-3}– 10^{2}	79.06
GBDT	Number of estimators	1– 100	56
	Learning rate	0.01– 1	0.23
	Subsample	0– 1	0.16
	Maximum depth of the tree	1– 20	7

Table 3
Confusion matrix of classification results.

True label	Predict label	
	Positive	Negative
Positive	N_{TP}	N_{FN}
Negative	N_{FP}	N_{TN}

identification is the groundwork in research of this kind of reservoir. Therefore, it is crucial to undertake related work to enhance the accuracy of lithology identification (Zhang et al., 2015). In this study, we presented a new ensemble learning algorithm (namely GBDT) to address this problem. Firstly, conventional logging data, which are calibrated by cores and FMI, can serve as training data to help establish the model. To avoid the occurrence of over-fitting and under-fitting, 10-folder cross validation was adopted to conduct the experiment so that all samples in each round could be used for training the model. It is critical and challenging to achieve the best accuracy without over-fitting and under-fitting. Validation curves and GridSearchCv were employed to determine the optimal key parameters for each model to find the balance between under-fitting and over-fitting. These results indicated that SVM and DT outperformed LR in the single classifier. Logging curves are highly nonlinear related. A strong classifier such as SVM and DT can better adapt to the irregular data while LR is a linear model. Therefore, they can achieve better classification results than LR. The GBDT model is superior to the single classifier and can accurately distinguish the lithologic interface of breccia tuff and rhyolite. Moreover, it also has better recognition ability for thin layer. Rhyolite and dacite have large quantity of samples, the number of which is approximately five times the tuffite. This makes the classification result of rhyolite obviously surpass that of

Table 4
Precision, recall and *f*1 score for 10-fold-cross validation over DTC and GBDT classifier.

Lithology	Sample number	DTC classifier			Sample number	GBDT classifier		
		Pr	Re	*f*1		Pr	Re	*f*1
rhyolite	462	0.94	0.58	0.72	462	0.98	0.92	0.95
dacite	664	0.75	0.97	0.84	664	0.98	1.00	0.99
tuff	202	0.49	0.53	0.51	202	0.73	0.83	0.78
tuffite	85	0.31	0.62	0.41	85	0.86	0.49	0.63
breccia tuff	268	0.66	0.38	0.48	268	0.66	0.86	0.75
ignimbrite	299	0.98	0.96	0.97	299	1	1	1
diabase	802	0.99	0.99	0.99	802	1	1	1

Table 5
Precision, recall and *f*1score for 10-fold-cross validation over SVM and LR classifier.

Lithology	Sample number	SVM classifier			Sample number	LR classifier		
		Pr	Re	*f*1		Pr	Re	*f*1
rhyolite	462	0.99	0.91	0.95	462	0.98	0.89	0.93
dacite	664	0.99	0.99	0.99	664	0.96	0.99	0.97
tuff	202	0.59	0.55	0.57	202	0.45	0.49	0.47
tuffite	85	0.49	0.58	0.53	85	0.50	0.14	0.22
breccia tuff	268	0.61	0.68	0.64	268	0.51	0.61	0.56
ignimbrite	299	1.00	1.00	1.00	299	0.94	0.97	0.96
diabase	802	1.00	1.00	1.00	802	0.99	0.99	0.99

the tuffite in the single classifier. However, for this kind of unbalanced data, the GBDT algorithm can still determine the features of samples fairly well. It can adaptively adjust the weight distribution of samples, thereby making it possible to deal with samples that are hard to train with.

The approach presented in this study could be used to improve the accuracy of volcanic lithology identification. However, the tuff and tuffite are usually misjudged on account of their similar mineral composition. Future work focused, in particular, on solving the problem of similar logging response with different lithology is recommended. Great deal of additional scientific research could be undertaken to determine other features which can distinguish them in origin.

6. Conclusions

(1) Multi-phase volcanic eruption has generated large sets of volcanic constructions in the Yingcheng Formation, making it relatively hard to identify volcanic lithology due to its strong heterogeneity. By using a variety of data, including cores, casting and conventional thin sections and imaging logging data, we were able to divide the volcanic lithology into three categories (volcanic lava, pyroclastic rock and intrusive rock) and eight kinds (rhyolite, dacite, tuff, tuffite, breccia tuff, ignimbrite, (and) diabase) in the Yingcheng Formation within the study area.

(2) In this paper, ensemble learning algorithm GBDT was used to establish a classification model for volcanic lithology with strong heterogeneity. The results showed that the GBDT model outperforms the single classifier. It can accurately distinguish the lithologic interface of breccia tuff and rhyolite, and also has better recognition ability for thin layer. Although the ensemble learning algorithm GBDT has significantly enhanced the accuracy of lithology identification, there still exist issues in distinguishing the different lithology with similar well logging response.

(3) To enhance the generalization ability and avoid under-fitting and over-fitting, parameter optimization becomes crucial and critical. Experiments were conducted on how to select the optimal parameters for each model. 10-folder cross validation curves and GridSearchCv were employed to determine the optimal key parameters for each model to find the balance between under-fitting and over-fitting. Eventually, these resultsrevealed that SVM has the best performance in a single classifier, and the average value of *f*1score reached 81.1%, while LR and DT reached 71.5% and 70.2%, respectively. GBDT significantly enhanced the efficiency of the classifier and increased the*f*1 score up to 87.1%.

Declaration of Competing 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.

Acknowledgements

This research was performed on Python 3.6 using machine learning package scikit-learn. This study was supported by the Strategic Priority Research Program of Jilin Oilfield (Grant No. 2019-FW-027). We are deeply grateful to the Exploration and Development Research Institute of Jilin Oilfield, CNPC, for providing research data and publication permission. The editor and anonymous reviewers are greatly appreciated for theirthorough and critical reviews and suggestions, which have significantly improved the quality of this paper.

References

Airola, A., Pahikkala, T., Waegeman, W., et al., 2011. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve[J]. Computat. Stat. Data Analysis 55 (4), 1828–1844.
Alpaydin, E., 2014. Introduction to Machine Learning. The MIT Press.
Avnimelech, R., Intrator, N., 1999. Boosted mixture of experts: an ensemble learning scheme[J]. Neural Comput. 11 (2), 483–498.
Camila, M.S., Leonardo, G.F., Egberto, P., Leonardo Costa, O., 2018. Machine learning approaches for petrographic classification of carbonate-siliciclastic rocks using well logs and textural information. J. Appl. Geophys. 155, 217–225.
Elghazel, H., Aussem, A., 2015. Unsupervised feature selection with ensemble learning [J]. Mach. Learn. 98 (1–2), 157–180.
Feng, Z.Q., 2008. Volcanic rocks as prolific gas reservoir: a case study from the Qingshen gas field in the Songliao Basin, NE China[J]. Mar. Pet. Geol. 25 (2008), 416–432.
Feng, J.W., Ren, Q.Q., Xu, K., 2018. Quantitative prediction of fracture distribution using geomechanical method within Kuqa Depression, Tarim Basin, NW China. J. Petrol. Sci. Eng. 162, 22–34.
Gong, L., Gao, S., Fu, X.F., Chen, S.M., Lyu, B.Y., Yao, J.Q., 2017. Fracture characteristics and their effects on hydrocarbon migration and accumulation in tight volcanic reservoirs: a case study of the Xujiaweizi fault depression, Songliao Basin, China. Interpretation 5 (4), 57–70.
Guo, Y., Liu, Y., 2016. Deep learning for visual understanding: a review. Neurocomputing 187 (C), 27–48.

Guoyin, Z., Zhizhang, W., Huaji, L., Yanan, S., Wei, C., 2018. Permeability prediction of isolated channel sands using machine learning. J. Appl. Geophys. 159, 605–615.
Han, Y., Yuan, C., Fan, Y., 2018. Identification of igneous reservoir lithology based on empirical mode decomposition and energy entropy classification: a case study of Carboniferous igneous reservoir in Chunfeng oilfield. Oil Gas Geol. 39 (4), 759–765 (in Chinese with English abstract).
He, H., Li, S., Liu, C., Kong, C.X., Jiang, Q.P., Chang, T.Q., 2020. Characteristics and quantitative evaluation of volcanic effective reservoirs:A case study from Junggar Basin, China. J. Pet. Sci. Eng. 195, 107723.
Huiguang, Li, Haitao, Xue, Wenbiao, Huang, et al., 2011. Hydrocarbon source rock exploration potential of deep layer in Dehui fault depression[J]. Sci. Technol. Eng. 11 (27), 6578–6582 (in Chinese with English abstract).
Ji, W., Shuli, K., Chuanning, T., et al., 2008. Meta-prediction of phosphorylation sites with weighted voting and restricted grid search parameter selection[J]. Nucleic Acids Res. 4, 4.
Jia, H., Ji, H., Wang, L., et al., 2016. Tectono-sedimentary and hydrocarbon potential analysis of rift-related successions in the Dehui Depression, Songliao Basin, Northeastern China[J]. Mar. Pet. Geol. 76, 262–278 (in Chinese with English abstract).
Jin Yuan, L., Yong, D., Tao, L.I., 2018. Classification of Flight Delay Based-on GBDT[J]. Mathemat. Pract. Theory 48 (4), 1–7.
Jing, Zhao, Liande, Bai, 2016. Main controlling factors of high-quality volcanic reservoir in southern Songliao basin[J]. Special Oil Gas Reser. 23 (3), 52–56 (in Chinese with English abstract).
Krogh, A., Vedelsby, J., 1995. Neural network ensembles, cross validation, and active Learning[J]. Adv. Neural Inf. Proces. Syst. 7 (10), 231–238.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436.
Li, L.J., Yu, Y., Bai, S.S., et al., 2018. Towards effective network intrusion detection: a hybrid model index and GBDT with PSO [J]. J. Sens. 6, 1–9.
Liao, Z., Huang, Y., Yue, X., et al., 2016. In silico prediction of gamma-aminobutyric acid type-a receptors using novel machine-learning-based SVM and GBDT Approaches[J]. Biomed. Res. Int. 2016 (6), 1–12.
Libin, Zhao, Zhilong, Huang, Ma, Yujie, et al., 2006. A study on geochemical character and origin of deep natural gas in Dehui fault depression of the southern Songliao basin. Nat. Gas Geosci. 17 (2), 176–182 (in Chinese with English abstract).
Mao, Z.G., Zhu, R.K., Luo, J.L., Wang, J.H., Du, Z.H., Su, L., Zhang, S.M., 2015. Reservoir characteristics, formation mechanisms and petroleum exploration potential of volcanic rocks in China. Pet. Sci. 12, 54–66.
Mitchell, Tom M., 2003. Machine learning[M]. McGraw-Hill.
Miyoshi, S., Uezu, T., Okada, M., 2006. Statistical mechanics of time domain ensemble learning[J]. J. Phys. Soc. Jpn. 75 (8), 2652–2674.
Petford, N., Mccaffrey, K.J.W., 2003. Hydrocarbons in Crystalline Rocks [M]. The Geological Society of London, London.
Sakhnovich, A., 2007. Nonisospectral integrable nonlinear equations with external potentials and their GBDT solutions[J]. J. Phys. A Math. Theor. 41(15).
Schutter, S.R., 2003. Occurrences of hydrocarbons in and around igneous rocks[J]. Hydrocarb. Crystalline Rocks 214 (1), 35–68.
Shuangfang, Lu, Hui, Sun, Weiming, Wang, et al., 2010. Key factors controlling the accumulation of volcanic gas reservoirs in the deep part of southern Songliao Basin [J]. J. Daqing Petrol. Instit. 34 (5), 42–47 (in Chinese with English abstract).
Sun, H.T., Zhong, D.K., Zhan, W.J., 2019. Reservoir characteristics in the cretaceous volcanic rocks of Songliao Basin, China: a case of dynamics and evolution of the volcano-porosity and diagenesis. Energy Explor. Exploit. 37 (2), 607–625.
Wang, Luo, Jianghai, Li, Yongmin, Shi, et al., 2015. Review and prospect of global volcanic reservoirs[J]. Geol. China 42 (5), 1610–1620 (in Chinese with English abstract).
Xin, Q., Hao Jue, H., Qing, Y., et al., 2019. Prediction of Temperature of Asphalt Pavement Surface Based on APRIORI-GBDT Algorithm[J]. J. Highway Transpor. Res. Develop. 36 (5), 1–10.
Yang, X., Wang, Z., Zhou, Z., et al., 2019. Lithology classification of acidic volcanic rocks based on parameter-optimized AdaBoost algorithm[J]. Acta Pet. Sin. 40 (4), 457–467.
Ye, T., Wei, A., Deng, H., 2017. Study on volcanic lithology identification methods based on the data of conventional well logging data: a case from Mesozoic volcanic rocks in Bohai bay area. Prog. Geophys. 32 (4), 1842–1848.
Zhang, J., Qin, L., Zhang, Z., 2008. Depositional fades, diagenesis and their impact on the reservoir quality of Silurian sandstones from Tazhong area in Central Tarim Basin, western China[J]. J. Asian Earth Sci. 33 (1–2), 42–60.
Zhang, D., Zou, N., Jiang, Y., 2015. Logging identification method of volcanic rock lithology: a case study from volcanic rock in Junggar Basin. Lithologic Reserv 27 (1), 108–114.
Zhang, L., Zhang, G., Qi, Y., 2017. Lithology identification of carboniferous volcanic rock with logging data in Xiquan Area. Junggar Basin. Xinjiang Petrol. Geol 38 (04), 427–431 (in Chinese with English abstract).
Zishu, Zhang, Wu, Banghui, 1994. Investigation of the research status and exploration technology at home and abroad about volcanic reservoir[J]. Natural Gas Explorat. Develop. 16 (1), 1–26 (in Chinese with English abstract).
Zou, C.N., Zhao, W.Z., Jia, C.Z., et al., 2008. Formation and distribution of volcanic hydrocarbon reservoirs in sedimentary basins of China[J]. Pet. Explor. Dev. 35 (3), 257–271. London: Geological Society, London, Special Publications, 2003, 214(1): 35–68(in Chinese with English abstract).