Petroleum Science 19 (2022) 570–580

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

Identification of low-resistivity-low-contrast pay zones in the feature space with a multi-layer perceptron based on conventional well log data

Lun Gao, Ran-Hong Xie*, Li-Zhi Xiao, Shuai Wang, Chen-Yu Xu

State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum (Beijing), Beijing, 102249, China

ARTICLE INFO

Article history:
Received 22 December 2020
Accepted 26 September 2021
Available online 9 December 2021

Edited by Jie Hao

Keywords:
Low-resistivity-low-contrast (LRLC) pay zones
Conventional well logging
Machine learning
DBSCAN algorithm
Multi-layer perceptron

ABSTRACT

In the early exploration of many oilfields, low-resistivity-low-contrast (LRLC) pay zones are easily overlooked due to the resistivity similarity to the water zones. Existing identification methods are model-driven and cannot yield satisfactory results when the causes of LRLC pay zones are complicated. In this study, after analyzing a large number of core samples, main causes of LRLC pay zones in the study area are discerned, which include complex distribution of formation water salinity, high irreducible water saturation due to micropores, and high shale volume. Moreover, different oil testing layers may have different causes of LRLC pay zones. As a result, in addition to the well log data of oil testing layers, well log data of adjacent shale layers are also added to the original dataset as reference data. The density-based spatial clustering algorithm with noise (DBSCAN) is used to cluster the original dataset into 49 clusters. A new dataset is ultimately projected into a feature space with 49 dimensions. The new dataset and oil testing results are respectively treated as input and output to train the multi-layer perceptron (MLP). A total of 3192 samples are used for stratified 8-fold cross-validation, and the accuracy of the MLP is found to be 85.53%.

1. Introduction

A low-resistivity-low-contrast (LRLC) pay zone refers to a pay zone with a resistivity R_t close to that of its adjacent water zone R_0, i.e., the formation resistivity index I ($I = R_t/R_0$, Archie, 1942) is low, and, usually, $I < 2$. An LRLC pay zone is better considered as a relative concept than an absolute concept, as pay zones in different oilfields can have very different resistivity ranges. For example, the resistivities of LRLC pay zones and water zones in the Attaka Oilfield in Indonesia range from 2 to 5 Ohmm and 1 to 5 Ohmm, respectively, but the resistivities of LRLC pay zones and water zones in the Lakhmani Oilfield in India range from 10 to 30 Ohmm and 15 to 25 Ohmm, respectively (Worthington, 1997). LRLC pay zones are widely distributed all over the world, including in Louisiana and Oklahoma in the United States, Hainan and Tarim in China, the Gulf of Mexico, the Middle East, etc. However, they are also easily overlooked by conventional log analysis due to the low contrast in resistivity.

In recent years, some novel well logging methods have been demonstrated to perform well in the identification of LRLC pay zones, such as NMR well logging (Guru et al., 2008; Rima et al., 2012; Belevich and Bal, 2018; Heidary et al., 2019) and pulse neutron well logging (Simpson and Menke, 2010). However, the data acquired from these novel well logging methods are usually absent in many oilfields, and conventional well log data are the only complete data that can be used.

Many researchers have sought model-driven solutions to the identification of LRLC pay zones for decades. One kind of method is to calculate the water saturation of a formation to determine whether or not it is hydrocarbon-bearing. The most widely used model for the calculation of water saturation from conventional well log data is the Archie's equation (Archie, 1942). However, if an LRLC pay zone has similar resistivity to its adjacent water zone, the water saturation calculated by the Archie's equation will be overestimated. Hill and Milburn (1956) found that the relationship between rock resistivity and formation water resistivity is nonlinear due to the effect of clay. Based on this relationship, Waxman and Smits (1968) developed a model that explains the electrical conductivity of shaly sand. Givens (1987) proposed a

* Corresponding author.
E-mail address: xieranhong@cup.edu.cn (R.-H. Xie).

https://doi.org/10.1016/j.petsci.2021.12.012

conductive rock matrix model that separates the rock into two parallel components, i.e., a conductive pore network containing fluids that are free to move, and the remainder of the rock, which may be conductive due to conductive minerals and immobile conductive water. With the deepening of the study on LRLC pay zones, recent methods tend to identify them based on the causes of the LRLC phenomenon. Worthington (1997) proposed a systematic workflow to interpret LRLC pay zones based on different causes, and different models were deployed for water saturation calculation. To identify LRLC pay zones caused by fresh drilling mud invasion, Li et al. (2010) established a three-step well logging identification method that contains conventional well logging cross-plot, drilling mud invasion factor analysis and well comparison; Pratama et al. (2017) proposed an integrated workflow that combines petrographical analysis, rock typing, and petrophysical analysis to calculate the shale volume, water saturation, and effective porosity, and eventually identifies the LRLC pay zones. Mashaba and Altermann (2015) considered clay-bound water and silt-bound water, which were the main causes of low resistivity in their study area, and improved the Archie's equation to calculate the water saturation. However, most of these model-driven methods are only locally applicable, and cannot yield satisfactory results when the causes of LRLC pay zones are complicated and the types of hydrocarbon layers (e.g., gas layer, gas/oil layer, oil layer, etc.) are various.

With the increasing scale and sophistication of data, the development of machine learning algorithms has shed light on complicated geophysical problems from a data-driven perspective (Bergen et al., 2019). Unsupervised algorithms like clustering and graphical models are usually used for well log data or seismic data preprocessing (Kang et al., 2019), reservoir multiscale modeling (Esmaeilzadeh et al., 2020), lithology identification (Feng et al., 2018; Li et al., 2021), and production schedule optimization for naturally fractured reservoirs (Liu and Forouzanfar, 2017). Regarding supervised algorithms, artificial neural networks (ANNs) are a popular tool for solving both regression and classification problems. ANNs can be categorized into different types based on their structures. Convolutional neural networks (CNNs) have a unique convolutional layer and a pooling layer, which are suitable for processing image data. CNNs are therefore used for seismic horizon tracking (Yang and Sun, 2020), seismic impedance inversion (Das et al., 2019), low-frequency noise suppression (Zhao et al., 2020), and seismic fault detection (Cunha et al., 2020; Yang et al., 2021). Recurrent neural networks (RNNs), including the long short-term memory (LSTM) network, have an advantage in dealing with sequential data, and are used to generate NMR T_2 distributions (Li and Misra, 2019), optimize seismic waveform inversions (Sun et al., 2020), pore fluid identification (Zhou et al., 2021) and obtain pore size distributions (Li et al., 2020). The multi-layer perceptron (MLP) is widely applied together with some heuristic algorithms to bridge scales for seismic phase picking (Chai et al., 2020), seismic facies analysis (Bagheri and Riahi, 2015), and shear wave travel time estimation (Anemangely et al., 2017).

In the present work, the powerful data processing and information extraction abilities of machine learning algorithms are employed in the construction of an implicit relationship between conventional well log data and LRLC pay zones. The density-based spatial clustering algorithm with noise (DBSCAN) is first applied to cluster the input dataset. By calculating the Euclidean distances from the data to the cluster centroids, the dataset can be projected into a feature space with a higher dimension. Eventually, the optimally structured MLP is deployed to train the data for the identification and classification of LRLC pay zones.

2. Methodology

2.1. DBSCAN algorithm

DBSCAN is a clustering algorithm based on the density distribution of a dataset, and it can exclude noise points in the dataset. Before the introduction of the DBSCAN algorithm, the following definitions are provided.

Given a dataset $D = \{x_1, x_2, \ldots, x_n\}$, $N_\varepsilon(x_i)$, the ε neighborhood of point x_i, is defined as the collection of points in dataset D whose distance to point x_i is less than ε, i.e., $N_\varepsilon(x_i) = \{x_j \in D | \mathrm{dist}(x_i, x_j) \le \varepsilon\}$. Point x_i is called the core point when its ε neighborhood contains no fewer than n points, i.e., $|N_\varepsilon(x_i)| \ge n$. Point x_j is directly density-reachable to point x_i (denoted as $x_j \Rightarrow x_i$) when point x_j is in the ε neighborhood of point x_i. Point x_j is density-reachable to point x_i (denoted as $x_j \rightarrow x_i$) when $x_j \Rightarrow x_p \Rightarrow \cdots \Rightarrow x_q \Rightarrow x_i$. Based on these definitions, the DBSCAN algorithm is described as follows:

1) Determine the parameters ε and n;
2) Find all the core points in dataset D and collect them into another dataset Ω;
3) Randomly choose one core point x_i in Ω, find its density-reachable points in D, and gather x_i and its density-reachable points in dataset C_i as one single cluster;
4) Update Ω by excluding the core points contained in cluster C_i, i.e., $\Omega = \Omega - C_i$;
5) Go to step 3 if Ω is not empty.

2.2. Multi-layer perceptron

The basic structure of an MLP includes an input layer, several hidden layers, and an output layer. The more hidden layers an MLP has, the more powerful its learning ability will be, and the more risk of overfitting it will have. Fig. 1 presents the structure of an MLP with two hidden layers. Given a labeled dataset $A = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, its training steps in this MLP are described as follows.

1) Input layer → hidden layer 1 → hidden layer 2

$$v_i = f(w_1 x_i + b_1),\ i = 1,\ 2,\ \ldots,\ n \tag{1}$$

where v_i is the output of hidden layer 1 and also the input of hidden layer 2, and f is the activation function of the neurons, common choices of which are the rectified linear unit, sigmoid, and tanh

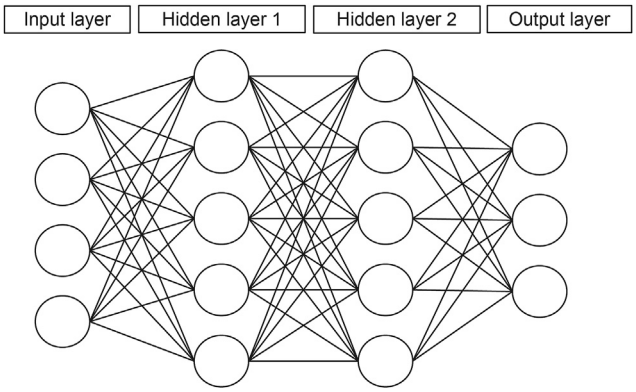


Fig. 1. Structure of an MLP with two hidden layers.

functions.

2) Hidden layer 1→hidden layer 2→output layer

$$u_i = f(w_2 v_i + b_2),\ i = 1,\ 2,\ \ldots,\ n \tag{2}$$

For a regression problem, $u_i = y_i^*$. However, for a classification problem, y_i^* is a probability that is generally expressed as the softmax function:

$$y_{i,j}^* = \exp(u_{i,j}) \Big/ \sum_j \exp(u_{i,j}) \tag{3}$$

After all the sample inputs x_i have been trained, their corresponding predicted outputs y_i^* are obtained. For a regression problem, the mean squared error (MSE) between the predicted output y_i^* and the real output y_i can be calculated as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_j (y_i^* - y_i)^2 \tag{4}$$

The MSE function is then minimized to find the optimal weight and bias parameters w_1, w_2, b_1, and b_2 of the MLP.

For a classification problem, the cross-entropy (CE) between the predicted output y_i^* and the real output y_i can be calculated as follows:

$$\mathrm{CE} = -\sum_i (y_i^* \cdot \log(y_i)) \tag{5}$$

The CE function is then minimized to find the optimal weight and bias parameters w_1, w_2, b_1, and b_2 of the MLP.

To avoid the overfitting problem in the training steps, dropout regularization is applied to randomly neglect some neurons in the hidden layers by setting the weights of these neurons to 0.

3. Case application

The study area is located in W Oilfield, China. It includes two sets of Paleogene hydrocarbon-bearing strata. The formation lithology is dominated by pebbled sandstones with low porosity, low permeability, and a complex pore structure. Fig. 2 presents the resistivity histogram of the water layers and oil layers from 15 wells

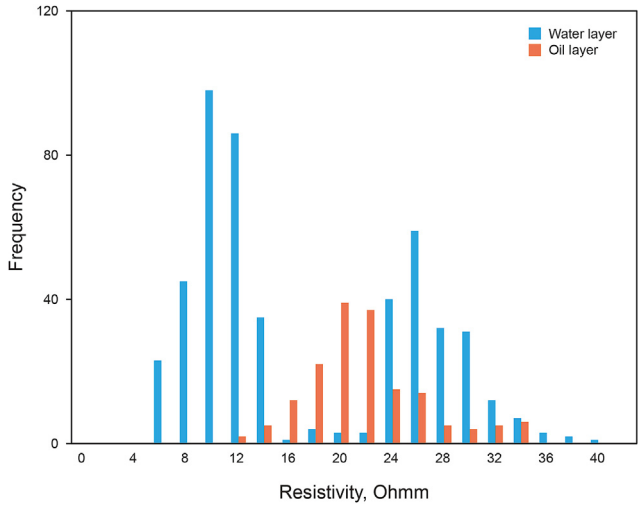


Fig. 2. The resistivity histogram of water layers and oil layers from 15 wells.

in the study area. It is evident that the resistivities of the water layers are in the range of 6 Ohmm - 40 Ohmm, and the resistivities of the oil layers are in the range of 12 Ohmm - 36 Ohmm. Fig. 3 exhibits the well logs of a water layer and an oil layer from well #2. The 1st track contains spontaneous potential log (SP), gamma ray log (GR) and caliper log (CAL). The 2nd track contains deep investigation induction log (ILD) and laterolog 8 log (LL8). The 3rd track contains density log (DEN), acoustic log (AC) and compensated neutron porosity log (CNL). The 4th track is the oil testing results. The contrast between the resistivity logs of these layers is low. Overall, it can be concluded that the reservoirs in the study area are typical LRLC pay zones.

The workflow of this case study is shown in Fig. 4. In section A, the causes of the LRLC pay zones in the study area are analyzed, which offers instructions for the data preprocessing described in section B. In section C, by deploying the DBSCAN algorithm, the dataset is projected into the feature space. In section D, the new dataset and the oil testing results are respectively treated as the input and output to train an MLP.

3.1. Cause analysis of LRLC pay zones

The water salinity in the study area has a very complicated distribution, and ranges from 421 mg/L – 89997.5 mg/L. Fig. 5 shows the water salinity histogram of the study area, from which we can see the water salinity distributes widely and unevenly. A water layer whose formation water has a very low water salinity will have a relatively high resistivity, whereas a hydrocarbon layer whose formation water has a high water salinity will have a relatively low resistivity. Fig. 6 exhibits an example to show the influence of the water salinity on the resistivities of difference layers. In Fig. 6, the layers of well #6 are tested as gas/oil layer and the water salinity in this depth interval is 12000 mg/L. The layers of well #13 are tested as water layer and the water salinity in this depth interval is 3381 mg/L. However, the resistivity ranges of well #6 and #13 are similar. Therefore, we can see the influence of the complex distribution of formation water salinity on the formation resistivity.

From the cast thin sections in Fig. 7, we can see there are many dissolution pores in the core samples. The pore size distributions show that there are two kinds of pore sizes. The first peak is around 0.01 μm, and the second peak is around 1 μm. Core porosity of the study area is in the range of 4.54%–12.7% and the porosity of most core samples is around 8%. Core permeability of the study area is in the range of 0.20 mD - 2.49 mD and the permeability of most core samples is around 0.5 mD. Fig. 8 is the porosity and permeability histograms that exhibits the specific porosity and permeability distribution of the core samples. Due to the small pore size, low porosity and low permeability, the irreducible water saturation can be very high. A high irreducible water saturation can result in the high conductivity of the pay zone. Based on the clay mineral X-ray diffraction analysis presented in Fig. 9, the clay content is around 20%, and the mixed-layer illite/smectite, which has a very high cation exchange capacity, composes about 50% of all the clay mineral content.

Above all, it can be concluded that the causes of the LRLC pay zones in this area are the complex distribution of formation water salinity, high irreducible water saturation due to micropores, and high shale volume. Due to the complicated causes, common methods cannot obtain satisfactory classification results. Fig. 10 shows the cross-plots of ILD with other well logs. We can see that it is difficult to divide the oil layers and water layers in these cross-plots. Therefore, it is necessary to apply the machine learning methods to establish an explicit and robust model for LRLC pay zone identification and classification.

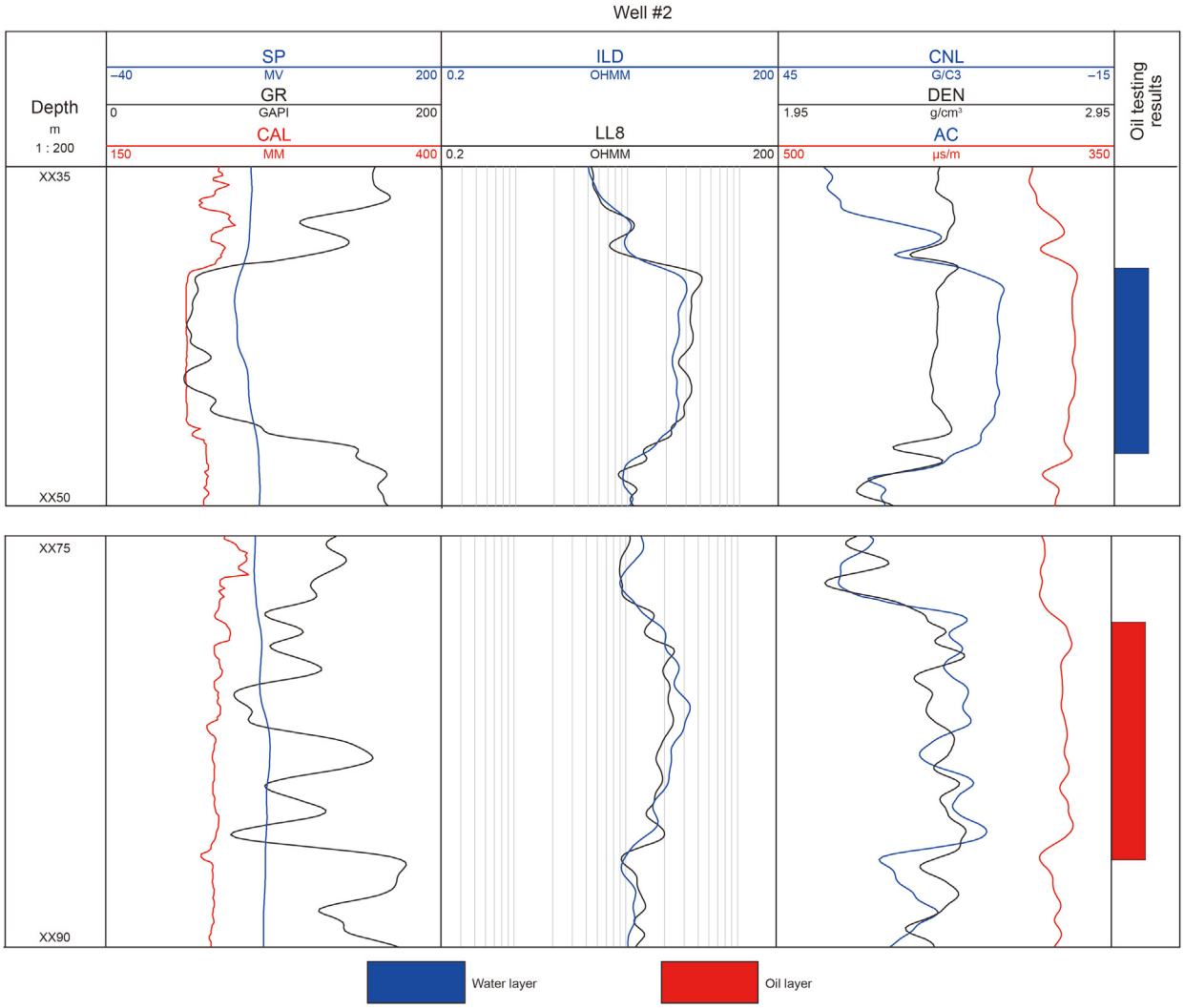


Fig. 3. The well logs of a water layer and an oil layer from well #2.

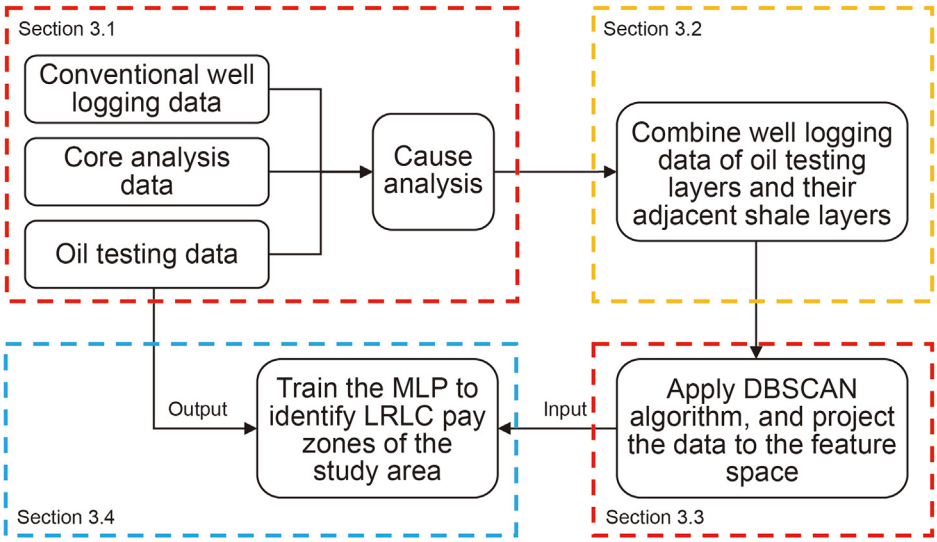


Fig. 4. The workflow of this case study.

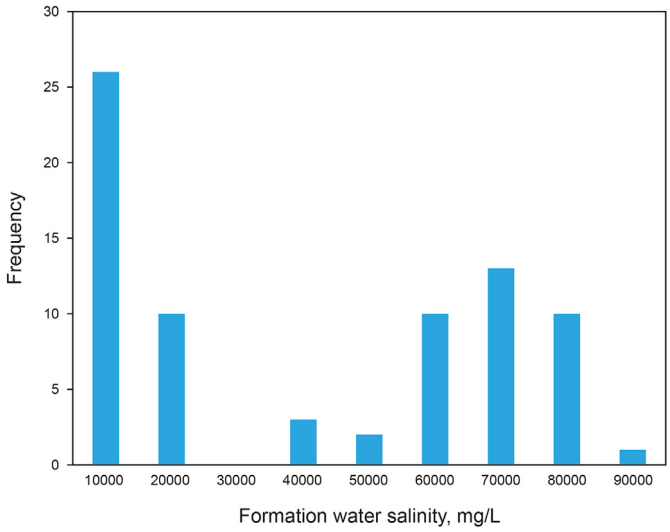


Fig. 5. The formation water salinity histogram.

3.2. Data preprocessing

Considering the complicated water salinity distribution and pore structure, different oil testing layers may have different LRLC causes. Therefore, it is reasonable to find the shale layers adjacent to the oil testing layers based on the gamma ray (GR) well logs and use them as reference data. As a result, 14 dimensions are included in the dataset, which includes GR_{ot}, GR_{sh}, SP_{ot}, SP_{sh}, AC_{ot}, AC_{sh}, CNL_{ot}, CNL_{sh}, DEN_{ot}, DEN_{sh}, ILD_{ot}, ILD_{sh}, $LL8_{ot}$, and $LL8_{sh}$. Footnote 'ot' stands for 'oil testing layer', and 'sh' stands for 'adjacent shale layer'.

Another problem is well logging data size mismatch between the oil testing layers and its adjacent shale layers. For example, an oil testing layer is 3 m thick (30 samples), whereas its adjacent shale layer is 2 m thick (20 samples). Due to the size mismatch, they cannot be combined as different dimensions of a dataset. To solve this problem, the well log value histogram of the shale layer is created and fitted with a Gaussian distribution. The mean value of the Gaussian distribution is then taken as the representative value and duplicated to make it have the same size as the oil testing layer.

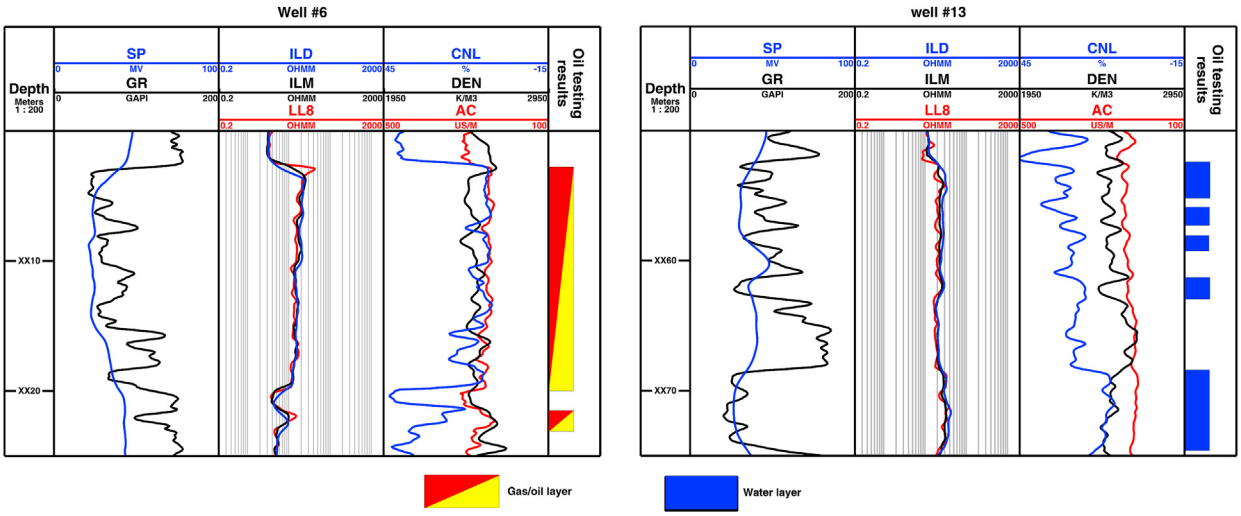
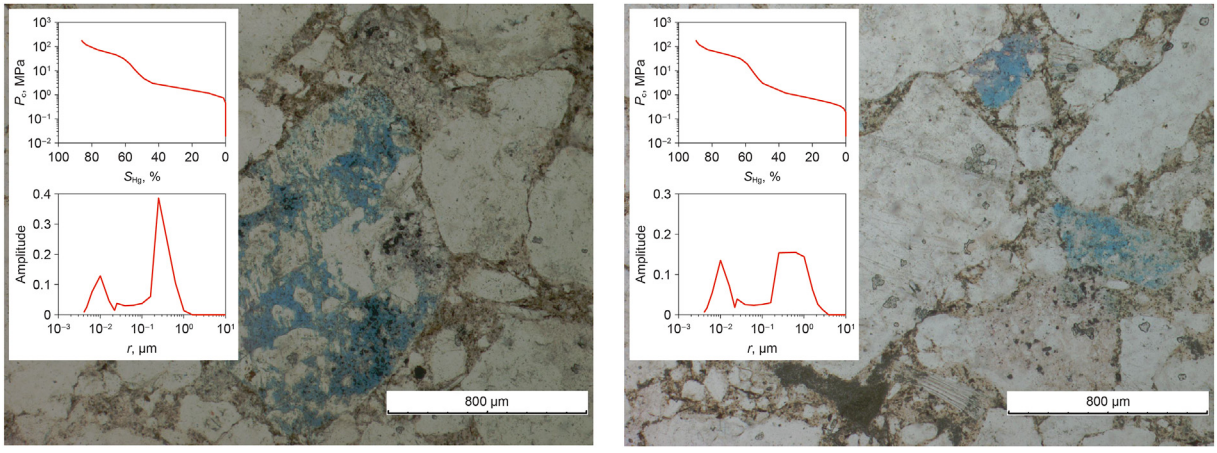


Fig. 6. Comparison of the gas/oil layers of well #6 and the water layers of well #13.



(a) Core sample No. 21 **(b)** Core sample No. 22

Fig. 7. Cast thin sections of two core samples with their corresponding MICP curves and pore size distributions.

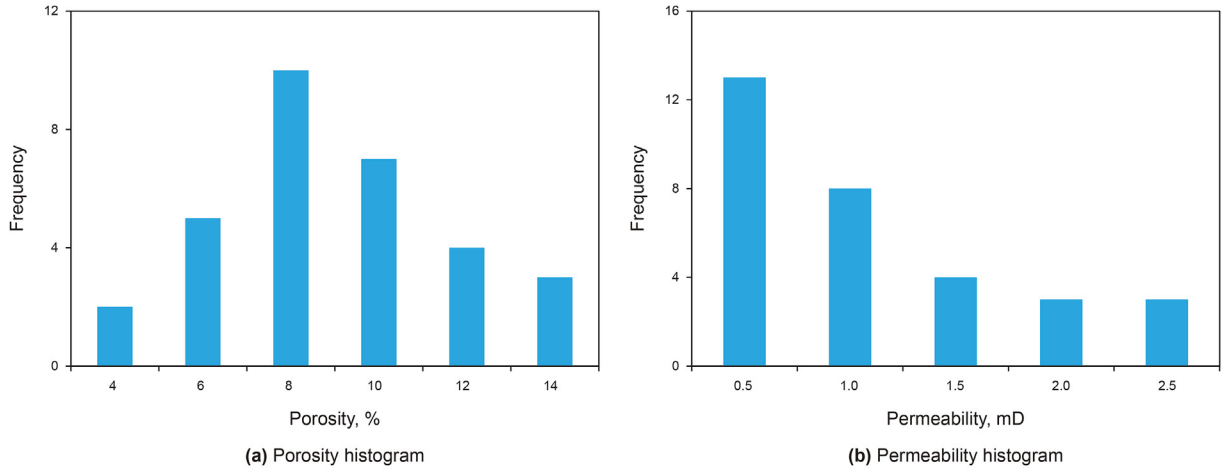


Fig. 8. The porosity and permeability histograms of the core samples in the study area.

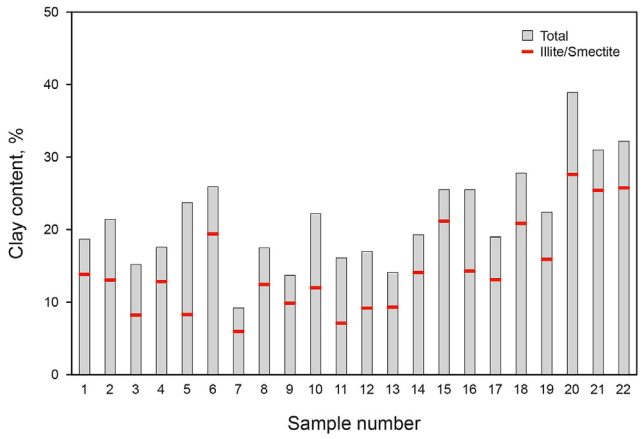


Fig. 9. Clay mineral X-ray diffraction analysis of 22 core samples in well #7.

Eventually, the well log dataset can be obtained, as partly shown in Table 1.

The dataset is normalized base on Equation (6):

$$x' = (x - x_{\min})/(x_{\max} - x_{\min}) \tag{6}$$

where x is the sample to be normalized, and x' is the normalized sample.

There are 10 types of layers in the study area based on the oil testing results, namely dry layers (DL), water layers (WL), oily water layers (OWL), low-production gas layers (LPGL), low-production oil layers (LPOL), low-production gas/oil layers (LPGOL), water/gas/oil layers (WGOL), water/oil layers (WOL), gas/oil layers (GOL), and oil layers (OL). With 3192 samples in the dataset, the oil testing results can be expressed as a matrix with a size of 3192 × 10. In the matrix, each type of layer has a unique corresponding vector. The vector of a dry layer, for example, is (1,0,0,0,0,0,0,0,0,0), which means that the probability of a dry layer is 1 and the probability of other types is 0. Table 2 reports the number of samples, vectors, and descriptions of each type of layer.

Overall, the normalized well log dataset (denoted as dataset P) with a size of 3192 × 14 can be obtained, as can the matrix of the oil testing results with a size of 3192 × 10.

3.3. Application of the DBSCAN algorithm

To apply the DBSCAN algorithm, the neighborhood parameters ε and n must first be determined. Generally, $n = d + 1$, where d is the dimension of samples in dataset P; therefore, in this work, $n = 15$. Moreover, ε can be determined by the silhouette coefficient:

$$S_i = (b_i - a_i)/\max(a_i, b_i) \tag{7}$$

where S_i is the silhouette coefficient of x_i, a_i is the mean Euclidean distance from x_i to the other samples in its cluster, and b_i is the shortest Euclidean distance from x_i to the centroids of other clusters. After the silhouette coefficients of all the samples in dataset P are calculated, a mean silhouette coefficient $\overline{S}$ can be obtained, which can be used to evaluate the clustering result. A higher $\overline{S}$ indicates a better clustering result. Fig. 11 presents the $\overline{S}$ curve with different values of ε in the range of 0.1–0.5. The optimal value of $\varepsilon =$ 0.19 can be obtained when $\overline{S}$ reaches its maximum.

By applying $\varepsilon = 0.19$ and $n = 15$ to the DBSCAN algorithm, dataset P can be clustered into 49 clusters. The centroid c_j, $j = 1, 2, \ldots, 49$, of each cluster is the arithmetic mean of the data samples in the cluster.

Dataset P is then projected into the feature space of the clusters by calculating the Euclidean distance from each sample in dataset P to the centroids. The sample in the feature space can be expressed as

$$z_i = (z_{i,1}, z_{i,2}, \ldots, z_{i,49}),\ i = 1, 2, \ldots, 3192 \tag{8}$$

$$z_{i,j} = \mathrm{dist}(x_i, c_j) \tag{9}$$

From Equation (8), it can be seen that the samples in the feature space have 49 dimensions, whereas the samples in dataset P have 14 dimensions. It is known that a dataset projected into a higher dimension is more easily classified than the dataset with a lower dimension. Therefore, the dataset in the feature space is used as the input of the MLP in the next section.

3.4. Application of the MLP

The performances of MLPs with different structures are investigated to find the optimal structure, as shown in Table 3. Stratified 8-fold cross-validation is applied to split the input and output data into 8 folds. 7 folds are treated as the training dataset, and the final

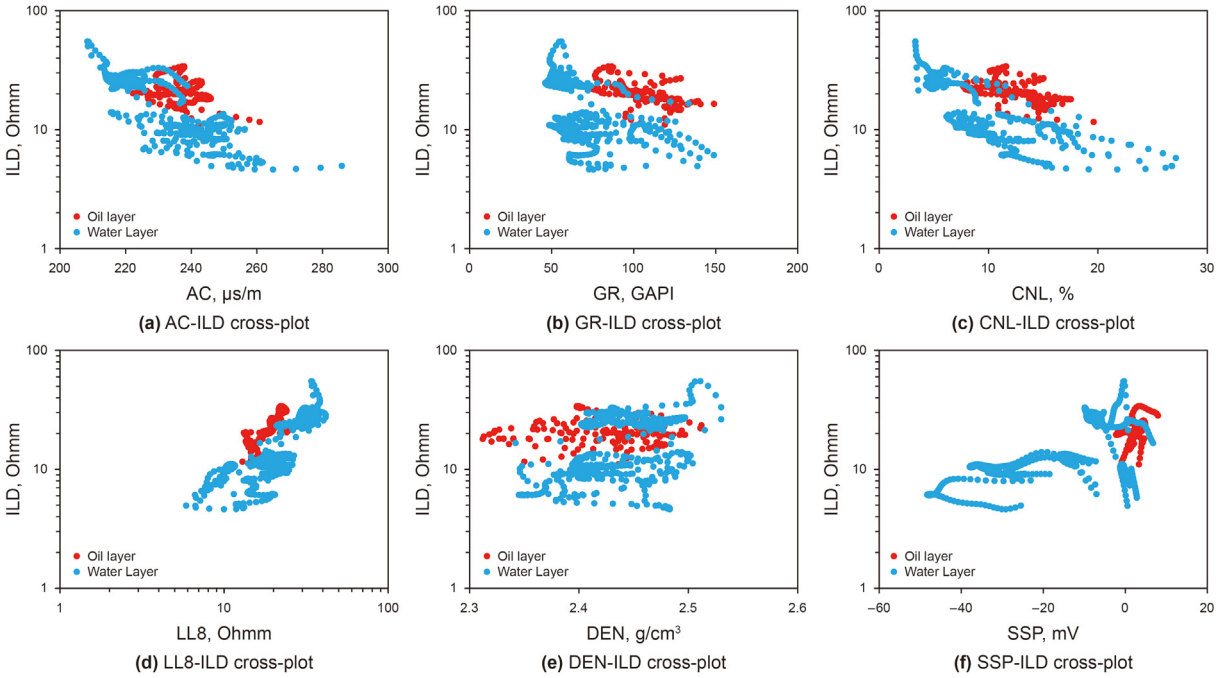


Fig. 10. Cross-plots of ILD with other well logs.

Table 1
Part of the well log dataset of an oil testing layer (only the AC, CNL, DEN logs are displayed).

AC_{ot}, μs/m	AC_{sh}, μs/m	CNL_{ot}, %	CNL_{sh}, %	DEN_{ot}, g/cm^3	DEN_{sh}, g/cm^3
332.4	315	55.85	52.5	1.56	1.9
332.2	315	52.82	52.5	1.63	1.9
328.5	315	48.21	52.5	1.72	1.9
319.7	315	42.71	52.5	1.83	1.9
308.6	315	37.94	52.5	1.96	1.9

fold is treated as the test dataset. After 8 loops, every fold has been treated as the test dataset, and the test accuracy of the whole dataset can be obtained. To be specific, the weight and bias parameters of the MLP in each loop are trained from the beginning to make sure they are not affected by the other loops. The CE function is used as the loss function, and the steepest gradient descent method is applied to minimize it. The second column in Table 3 reports the structures of the MLPs. As an example, "49-128-64-32-10" means that this MLP has 5 layers, and the numbers of neurons in each layer are 49, 128, 64, 32, and 10, respectively. The complexity can be calculated based on the structure via Equation (10):

$$\text{complexity} = \sum_{i=1}^{n-1}(l_i + 1)l_{i+1}r_{\text{dr}} \tag{10}$$

where n is the number of layers in the MLP, l_i is the number of neurons in the ith layer, and r_{dr} is the dropout rate in each layer. Because the complexity is equal to the number of coefficients to be trained in an MLP, the higher the complexity, the longer the run time required by the MLP will be. For structures No. 1 — No. 9, each MLP is run five times, and the mean accuracy is calculated. From Table 3, it can be seen that the accuracy increases with the complexity. Structure No. 9 has the highest mean accuracy of 85.97%. However, structure No. 4 has a mean accuracy of 85.53%, which is very close to that of structure No. 9. Moreover, structure No. 4 has lower complexity than structure No. 9. Therefore, for a slight sacrifice of accuracy in exchange for greater efficiency, structure No. 4 is assumed to have the best performance.

To further analyze the prediction result, a confusion matrix is obtained to show the detailed numbers of actual and predicted types of samples. The precision and recall of each type of sample can be calculated from the confusion matrix, and are respectively defined as Equations (11) and (12):

Table 2
Number of samples, vectors, and descriptions of each type of layers.

Type	Number of samples	Corresponding vector	Description
DL	137	(1,0,0,0,0,0,0,0,0,0)	Fluid production <0.4 t/d.
WL	991	(0,1,0,0,0,0,0,0,0,0)	Water cut >98%.
OWL	110	(0,0,1,0,0,0,0,0,0,0)	Water cut >90%.
LPGL	43	(0,0,0,1,0,0,0,0,0,0)	Gas/oil ratio >890, and gas production <3 × 10^4 m^3/(km·d).
LPOL	92	(0,0,0,0,1,0,0,0,0,0)	Water cut <2%, gas/oil ratio <534, and 1 t/(km·d) < oil production <5 t/(km·d).
LPGOL	141	(0,0,0,0,0,1,0,0,0,0)	Water cut <2%, 534 < gas/oil ratio <890, 1 t/(km·d) < oil production <5 t/(km·d), and gas production <3 × 10^4 m^3/(km·d).
WGOL	177	(0,0,0,0,0,0,1,0,0,0)	10% < water cut <90%, and 534 < gas/oil ratio <890.
WOL	361	(0,0,0,0,0,0,0,1,0,0)	10% < water cut <90%.
GOL	733	(0,0,0,0,0,0,0,0,1,0)	Water cut <2%, and 534 < gas/oil ratio <890.
OL	407	(0,0,0,0,0,0,0,0,0,1)	Water cut <2%, and gas/oil ratio <534.

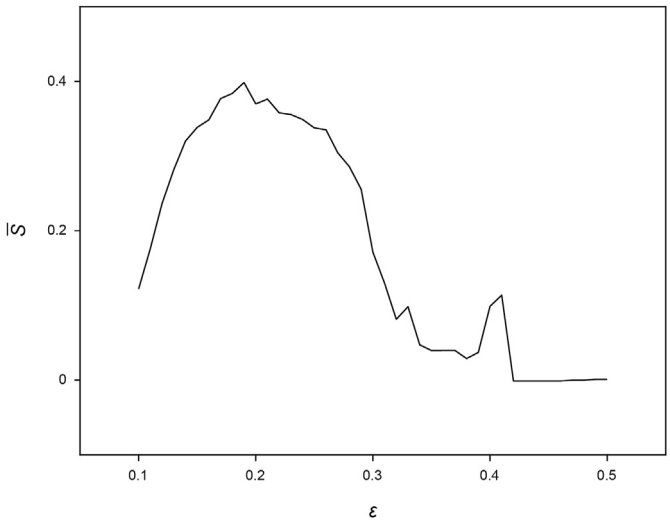


Fig. 11. Silhouette coefficient with different values of ε.

$$\text{precision} = TP/(TP + FP) \tag{11}$$

$$\text{recall} = TP/(TP + FN) \tag{12}$$

where *TP* is the number of true positive samples, *FP* is the number of false positive samples, and *FN* is the number of false negative samples. Table 4 presents the confusion matrix of the 2nd MLP from structure No. 4. Taking the water layer in Table 4 as an example, the *TP* of the water layer is 761 (the 4th row of the 3rd column in Table 4), the *FP* of the water layer is 59 (the sum of the 3rd column in Table 4 excluding its 4th row), and the *FN* of the water layer is 230 (the sum of the 4th row in Table 4 excluding its 3rd column).

From Table 4, it can be seen that the recalls of the water layer and water/oil layer are relatively low, because many water layer samples are predicted as low-production gas layer, low-production oil layer, and water/oil layer samples. Consequently, the precisions of these three types of samples are affected.

To intuitively show the effectiveness of this method, we exhibit the well logs, oil testing results and predicted results of well #2, as shown in Fig. 12. In Fig. 12, the 1st track contains SP, GR and CAL. The 2nd track contains ILD and LL8. The 3rd track contains DEN, AC and CNL. The 4th track is the oil testing results. The 5th track is the predicted results. Layers where the predicted results are inconsistent with the oil testing results are marked by symbol '*'. We can see the oil testing results of most layers have a very good agreement with its corresponding predicted results. Thus, the effectiveness of the proposed method is well demonstrated.

4. Discussions

4.1. The necessity of including the well log data of adjacent shale layers and projecting the dataset into the feature space

Two cases are designed for comparison with the proposed method:

Case 1: Only the well log data of the oil testing layers are included in the dataset;
Case 2: The well log data of the oil testing layers and adjacent shale layers are included in the dataset.

The difference between Case 1 and Case 2 is that the well log data of adjacent layers are included in Case 2. The difference between Case 2 and the proposed method is that the dataset is not projected into the feature space in Case 2.

Table 3
Performances of MLPs with different structures.

No.	Structure	Complexity	Accuracy, %					
			1st	2nd	3rd	4th	5th	Mean
1	49-128-64-32-10	8533	71.77	74.28	76.07	72.43	73.56	73.62
2	49-128-64-32-16-10	8717	74.09	72.18	66.29	69.61	67.95	70.02
3	49-256-128-64-32-10	28,181	79.45	81.52	79.51	80.95	83.11	80.91
4	49-512-256-128-64-10	99,365	85.15	87.41	84.84	84.34	85.9	85.53
5	49-1024-512-256-128-10	370,757	85.68	85.28	85.59	86.34	86.69	85.92
6	49-512-256-128-64-32-10	100,245	83.15	85.31	85.96	85.40	84.27	84.82
7	49-512-256-128-64-32-16-10	374,565	85.06	81.58	86.94	82.24	81.80	83.52
8	49-1024-512-256-128-64-10	100,429	85.49	85.20	85.27	85.68	85.25	85.38
9	49-1024-512-256-128-64-32-10	375,445	88.35	85.24	85.18	86.40	84.68	85.97

Table 4
Confusion matrix of the 2nd MLP from structure No. 4.

Actual	Predicted										Recall
	DL	WL	OWL	LPGL	LPOL	LPOGL	WGOL	WOL	OGL	OL	
DL	135	0	1	0	0	0	0	1	0	0	0.99
WL	12	761	11	27	40	2	8	95	0	35	0.77
OWL	0	1	107	0	0	0	0	2	0	0	0.97
LPGL	0	0	0	43	0	0	0	0	0	0	1.00
LPOL	0	0	0	0	91	0	0	1	0	0	0.99
LPGOL	0	0	0	0	0	141	0	0	0	0	1.00
WGOL	0	0	0	0	2	0	175	0	0	0	0.99
WOL	0	46	16	0	1	0	0	243	0	55	0.67
OGL	0	0	0	0	0	0	0	0	733	0	1.00
OL	0	12	1	0	2	0	0	2	0	390	0.96
Precision	0.92	0.93	0.79	0.61	0.67	0.99	0.96	0.71	1.00	0.81	

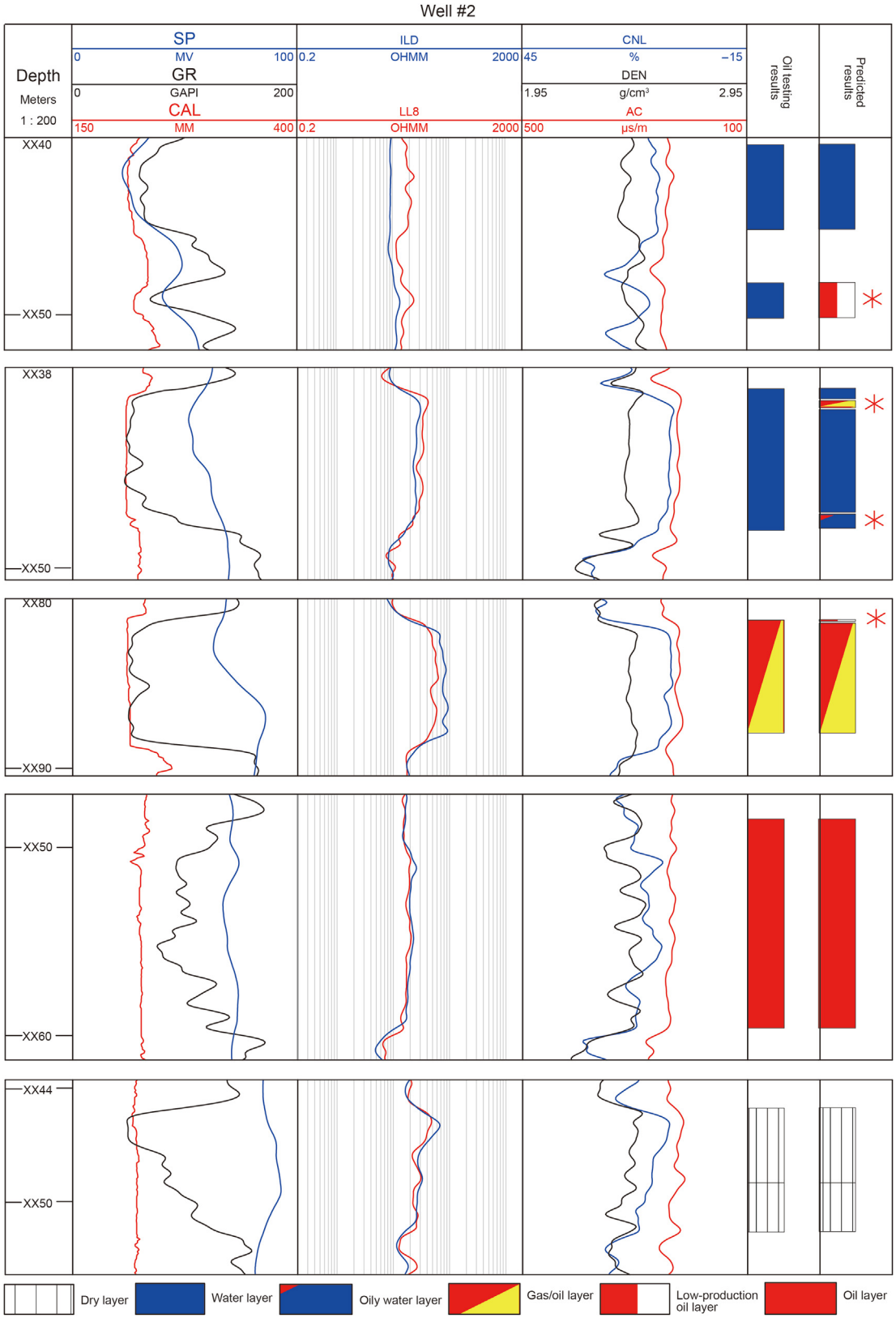


Fig. 12. The well logs, oil testing results and predicted results of well #2.

Table 5
Performances of MLP with different structures in Case 1.

No.	Structure	Complexity	Accuracy, %					
			1st	2nd	3rd	4th	5th	Mean
1	7-128-64-32-10	11,690	64.29	60.93	61.34	60.46	63.19	62.04
2	7-128-64-32-16-10	12,058	58.96	59.87	60.43	62.27	61.25	60.56
3	7-256-128-64-32-10	45,610	65.19	67.61	65.82	67.67	68.33	66.92
4	7-512-256-128-64-10	177,226	68.73	70.11	67.98	66.92	70.8	68.91
5	7-1024-512-256-128-10	698,506	72.06	73.03	74.15	73.15	73.09	73.10
6	7-512-256-128-64-32-10	178,986	65.95	71.02	70.8	66.92	67.2	68.38
7	7-512-256-128-64-32-16-10	179,354	65.19	70.8	66.32	68.3	66.76	67.47
8	7-1024-512-256-128-64-10	706,122	71.77	70.39	74.69	71.27	72.49	72.12
9	7-1024-512-256-128-64-32-10	707,882	72.81	74.22	73.18	72.59	71.77	72.91

Table 6
Performances of MLP with different structures in Case 2.

No.	Structure	Complexity	Accuracy, %					
			1st	2nd	3rd	4th	5th	Mean
1	14-128-64-32-10	12,586	73.97	66.82	68.52	69.92	71.15	70.08
2	14-128-64-32-16-10	12,954	71.46	69.71	66.35	64.97	69.77	68.45
3	14-256-128-64-32-10	47,402	71.71	68.17	65.91	69.96	66.54	68.46
4	14-512-256-128-64-10	180,810	81.58	81.17	80.01	79.86	77.54	80.03
5	14-1024-512-256-128-10	705,674	81.64	80.80	80.70	82.42	81.55	81.42
6	14-512-256-128-64-32-10	182,570	78.04	79.64	80.42	80.08	79.51	79.54
7	14-512-256-128-64-32-16-10	182,938	73.81	74.75	76.88	79.32	77.63	76.48
8	14-1024-512-256-128-64-10	713,290	81.17	80.26	81.20	81.30	81.92	81.17
9	14-1024-512-256-128-64-32-10	715,050	78.82	81.05	77.38	79.07	78.95	79.05

Tables 5 and 6 exhibit the performances of MLPs with different structures for Case 1 and Case 2. The highest mean accuracy of Case 1 achieved by the MLP with structure No. 5 is 73.10%, and the highest mean accuracy of Case 2 achieved by the MLP with structure No. 5 is 81.42% from structure No. 5, which is 8% higher than that for Case 1. Also, the mean accuracy of the proposed method is 85.53%, which is 4% higher than that of Case 2. Overall, the necessity of including the well log data of the adjacent shale layers and projecting the dataset into the feature space is revealed.

4.2. Results comparison with support vector classification

Support vector classification (SVC) is a kind of support vector machine (SVM), but SVC is used to solve classification problems. SVC can be classified into two categories, linear SVC and kernel SVC. The common choices of kernel functions are polynomial function, radial basis function (RBF) and sigmoid function. We apply these methods above to compare their predicted results with the results of the proposed method, as shown in Table 7, where the regularization parameter used in the SVC methods is 0.1. From Table 7, we can see that the accuracy of these method is smaller than the accuracy of the proposed method, which is 85.53%.

4.3. Analysis of the reason for the poor prediction results of water layers and water/oil layers

The recalls of the water layers and water/oil layers are around 70%, which are relatively low compared to the recalls of other types of layers. Specifically, there are 991 samples of water layers, 95 of which are predicted as water/oil layers; moreover, there are 361 samples of water/oil layers, 46 of which are predicted as water layers, and 55 of which are predicted as oil layers.

Table 7
Predicted results of SVC.

SVC	Linear	Polynomial	RBF	Sigmoid
Accuracy	63%	69%	80%	30%

The possible reason for this phenomenon is the scale difference between the well log data and oil testing results. The depth interval of well log data in the study area is 0.1 m, whereas the oil testing results are an overall evaluation of a whole layer, which is usually several meters thick. For instance, the oil testing results in well #2 indicate that a water/oil layer has an oil production of 0.5 m^3 per day, a water production of 3.07 m^3 per day, and is 3.2 m thick, which means that this layer contains 32 pieces of well log data. Therefore, the well log data below the water/oil interface are very likely to be predicted as water layers, while the well log data above the water/oil interface are very likely to be predicted as oil layers.

4.4. Problems to be optimized

The determination of neighborhood parameters is very essential for DBSCAN algorithm. This paper offers a general but not optimal procedure to calculate these parameters. There are also infinite choices for the number of layers and neurons in an MLP. Perfect balance between accuracy and efficiency is still needed to be found.

Another problem is the category imbalance in the training dataset. The water layer samples and gas/oil layer samples take up half of the whole sample numbers, which may affect the predicted results. Oversampling the minority categories or undersampling the majority categories is a feasible way to adjust the imbalance.

5. Conclusions

A large amount of core analysis shows that the causes for those LRLC pay zones include complex distribution of formation water salinity, high irreducible water saturation due to micropores, and high shale volume. A machine learning method for LRLC pay zones identification and classification is proposed. After the result analysis and discussion, several conclusions are drawn: The proposed machine learning methods can effectively learn the relation

between conventional well logging data and oil testing results. This method is applicable when the causes are complicated; Considering the complexity and accuracy, the optimal structure of the MLP is '49-512-256-128-64-10' with a mean accuracy of 85.53%; By comparing the results of Case 1 method to the results of Case 2 method, we can see there is a great accuracy improvement after adding the data of adjacent shale layers into the dataset; By comparing the results of proposed method to the results of Case 2 method, we can see the necessity of projecting the dataset into feature space.

Acknowledgements

The work was funded by the Strategic Cooperation Technology Projects of CNPC and CUPB (ZLZX2020-03).

Nomenclature

AC	Acoustic logs
ANN	Artificial neural network
CE	Cross entropy
CNL	Compensated neutron porosity logs
CNN	Convolutional neural network
DBSCAN	Density-based spatial clustering algorithm with noise
DEN	Density logs
DL	Dry layer
FN	False negative
FP	False positive
GOL	Gas/oil layer
GR	Gamma ray logs
ILD	Deep investigation induction logs
LL8	Laterolog 8 logs
LPGL	Low-production gas layer
LPGOL	Low-production gas/oil layer
LPOL	Low-production oil layer
LRLC	Low-resistivity-low-contrast
LSTM	Long short-term memory
MICP	Mercury injection capillary pressure
MLP	Multi-layer perceptron
MSE	Mean squared error
OL	Oil layer
OWL	Oily water layer
RBF	Radial basis function
RNN	Recurrent neural network
SP	Spontaneous potential logs
SVC	Support vector classification
SVM	Support vector machine
TP	True positive
WL	Water layer
WGOL	Water/gas/oil layer
WOL	water/oil layer

References

Anemangely, M., Ramezanzadeh, A., Tokhmechi, B., 2017. Shear wave travel time estimation from petrophysical logs using ANFIS-PSO algorithm: a case study from Ab-Teymour Oilfield. J. Nat. Gas Sci. Eng. 38, 373–387. https://doi.org/10.1016/j.jngse.2017.01.003.

Archie, G.E., 1942. Electrical resistivity log as an aid in determining some reservoir characteristics. Trans. AIME 146 (1), 54–62.

Bagheri, M., Riahi, M.A., 2015. Seismic facies analysis from well logs based on supervised classification scheme with different machine learning techniques. Arabian J. Geosci. 8 (9), 7153–7161. https://doi.org/10.1007/s12517-014-1691-5.

Belevich, A., Bal, A.A., 2018. The problem with silt in low-resistivity low-contrast (lrlc) pay reservoirs. Petrophysics 59 (2), 118–135. https://doi.org/10.30632/PJV59N2-2018a1.

Bergen, K.J., Johnson, P.A., de Hoop, M.V., et al., 2019. Machine learning for data-driven discovery in solid Earth geoscience. Science 363 (6433), 1299. https://doi.org/10.1126/science.aau0323.

Chai, C., Maceira, M., Hector, J., et al., 2020. Using a deep neural network and transfer learning to bridge scales for seismic phase picking. Geophys. Res. Lett. 47 (16). https://doi.org/10.1029/2020GL088651.

Cunha, A., Pochet, A., Lopes, H., et al., 2020. Seismic fault detection in real data using transfer learning from a convolutional neural network pre-trained with synthetic seismic data. Comput. Geosci. 135, 104344. https://doi.org/10.1016/j.cageo.2019.104344.

Das, V., Pollack, A., Wollner, U., et al., 2019. Convolutional neural network for seismic impedance inversion. Geophysics 84 (6), 869–880. https://doi.org/10.1190/GEO2018-0838.1.

Esmaeilzadeh, S., Salehi, A., Hetz, G., et al., 2020. Multiscale modeling of compartmentalized reservoirs using a hybrid clustering-based non-local approach. J. Petrol. Sci. Eng. 184, 106485. https://doi.org/10.1016/j.petrol.2019.106485.

Feng, R.H., Luthi, S.M., Gisolf, D., et al., 2018. Reservoir lithology determination by hidden Markov random fields based on a Gaussian mixture model. IEEE T. Geosci. Remote 56 (11), 6663–6673. https://doi.org/10.1109/TGRS.2018.2841059.

Givens, W.W., 1987. A conductive rock matrix model (crmm) for the analysis of low-contrast resistivity formations. Log. Anal. 28 (2), 138–151.

Guru, U., Heaton, N., Bachman, H.N., et al., 2008. Low-resistivity pay evaluation using multidimensional and high-resolution magnetic resonance profiling. Petrophysics 49 (4), 342–350.

Heidary, M., Kazemzadeh, E., Moradzedeh, A., et al., 2019. Improved identification of pay zones in complex environments through wavelet analysis on nuclear magnetic resonance log data. J. Petrol. Sci. Eng. 172, 465–476. https://doi.org/10.1016/j.petrol.2018.09.092.

Hill, H.J., Milburn, J.D., 1956. Effect of clay and water salinity on electrochemical behavior of reservoir rocks. Trans. AIME 207 (1), 65–72.

Kang, B., Kim, S., Jung, H., et al., 2019. Efficient assessment of reservoir uncertainty using distance-based clustering: a review. Energies 12 (10), 1859. https://doi.org/10.3390/en12101859.

Li, C., Shen, A.J., Chang, S.Y., et al., 2021. Application and contrast of machine learning in carbonate lithofacies log identification: a case study of Longwangmiao Formation of MX area in Sichuan Basin. Petrol. Reserv. Eval. Dev. 11 (4), 586–596. https://doi.org/10.13809/j.cnki.cn32-1825/te.2021.04.015 (in Chinese).

Li, C.X., Shi, Y.J., Zhou, C.C., et al., 2010. Evaluation of low amplitude and low resistivity pay zones under the fresh drilling mud invasion condition. Petrol. Explor. Dev. 37 (6), 696–702. https://doi.org/10.1016/S1876-3804(11)60004-9 (in Chinese).

Li, H., Misra, S., He, J.B., 2020. Neural network modeling of in situ fluid-filled pore size distributions in subsurface shale reservoirs under data constraints. Neural Comput. Appl. 32, 3873–3885. https://doi.org/10.1007/s00521-019-04124-w.

Li, H., Misra, S., 2019. Long short-term memory and variational autoencoder with convolutional neural networks for generating NMR T2 distributions. IEEE Geosci. Remote S. 16 (2), 192–195. https://doi.org/10.1109/LGRS.2018.2872356.

Liu, Z., Forouzanfar, F., 2017. Ensemble clustering for efficient robust optimization of naturally fractured reservoirs. Comput. Geosci. 22 (1), 283–296. https://doi.org/10.1007/s10596-017-9689-1.

Mashaba, V., Altermann, W., 2015. Calculation of water saturation in low resistivity gas reservoirs and pay-zones of the Cretaceous Grudja formation, onshore Mozambique basin. Mar. Petrol. Geol. 67, 249–261. https://doi.org/10.1016/j.marpetgeo.2015.05.016.

Pratama, E., Mohd, S.I., Syahrir, R., 2017. An integrated workflow to characterize and evaluate low resistivity pay and its phenomenon in a sandstone reservoir. J. Geophys. Eng. 14 (3), 513–519. https://doi.org/10.1088/1742-2140/aa5efb.

Rima, C., Datta, G.S., Farooqui, M.Y., 2012. Application of nuclear magnetic resonance logs for evaluating low-resistivity reservoirs: a case study from the Cambay basin, India. J. Geophys. Eng. 9 (5), 595–610. https://doi.org/10.1088/1742-2132/9/5/595.

Simpson, G., Menke, J.Y., 2010. Identifying low contrast-low resistivity pay zones with pulsed neutron capture logs in shaly sand Miocene formations of South Louisiana. In: SPWLA 51st Annual Logging Symposium.

Sun, J., Niu, Z., Innanen, K.A., et al., 2020. A theory-guided deep learning formulation and optimization of seismic waveform inversion. Geophysics 85 (2), 87–99. https://doi.org/10.1190/GEO2019-0138.1.

Waxman, M.H., Smits, L.J.M., 1968. Electrical conductivities in oil-bearing shaly sands. Soc. Petrol. Eng. J. 8 (2), 107–122.

Worthington, P.F., 1997. Recognition and development of low-resistivity pay. In: SPE Asia Pacific Oil and Gas Conference and Exhibition.

Yang, L.X., Sun, S.Z., 2020. Seismic horizon tracking using a deep convolutional neural network. J. Petrol. Sci. Eng. 187, 106709. https://doi.org/10.1016/j.petrol.2019.106709.

Yang, W.Y., Yang, J.R., Chen, S.Q., et al., 2021. Seismic data fault detection based on U-net deep learning network. Oil Geophys. Prospect. 56 (4), 688–697. https://doi.org/10.13810/j.cnki.issn.1000–7210.2021.04.002 (in Chinese).

Zhao, Y.X., Li, Y., Yang, B.J., 2020. Low-frequency desert noise intelligent suppression in seismic data based on multiscale geometric analysis convolutional neural network. IEEE T. Geosci. Remote 58 (1), 650–665. https://doi.org/10.1109/TGRS.2019.2938836.

Zhou, X.Q., Zhang, Z.S., Zhu, L.Q., et al., 2021. A new method for high-precision fluid identification in bidirectional long short-term memory network. J. China Univ. Petrol (Edition of Natural Science) 45 (1), 69–76, 0.3969/j.issn.1673-5005.2021.01.008. (in Chinese).