228 (2023) 212064

Contents lists available at ScienceDirect

Geoenergy Science and Engineering

journal homepage: www.sciencedirect.com/journal/geoenergy-science-and-engineering

Application of machine learning for evaluating and predicting fault seals: A case study in the Huimin depression, Bohai Bay Basin, Eastern China

Qiaochu Wang [a,b], Dongxia Chen [a,b,*], Meijun Li [a,b], Fuwei Wang [a,b], Yu Wang [a,b], Wenlei Du [a,b], Xuebin Shi [a,b]

[a] State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum (Beijing), Beijing, 102249, China
[b] College of Geosciences, China University of Petroleum (Beijing), Beijing, 102249, China

ARTICLE INFO

Keywords:
Fault seal
Decision tree
Random forest
Intelligent exploration and exploitation
Cross validation
Huimin depression

ABSTRACT

Fault seal is of great significance for hydrocarbon migration, accumulation, and further hydrocarbon reservoir production. Approximate 32% petroleum resource is confirmed to be related to the faults. However, the existing fault seal evaluation methods based on statistical analysis cannot accurately predict fault seal which is influenced by multiple factors in a complex way. It is necessary to improve the fault seal evaluation methods for enhancing the exploration success rate. In this study, a new fault seal evaluation and prediction method based on decision tree (DT) and random forest (RF) is introduced. First, the original dataset was set by quantification and feature engineering work. Second, the nonlinear classification models for fault seal evaluation and prediction using a binary decision tree named the classification and regression tree (CART) were constructed and improved by overfitting calibration. Third, the random forest algorithm was selected as an ensemble learning method to improve the faut seal evaluation and prediction accuracy. Third, the evaluation metrics and cross-validation were used to evaluate the performance of the model. Finally, the validation test is applied for testing the reliability of the model. The result showed that among the 100,000 models constructed in this study, the DT best model could evaluate and predict the fault seal with a cross-validation accuracy of 80.60% after overfitting calibration by pruning. The best RF model showed the highest test accuracy of 86.54%, which is higher than that of the DT model. The models were used for predicting fault seals in another district in the Huimin Depression, and the prediction accuracy reached 90% and 95% for the DT and RF model, respectively. This study not only introduced a new method for fault seal evaluation and prediction, but also provided guidance for the application and development of machine learning in petroleum exploration and exploitation field and industry.

1. Introduction

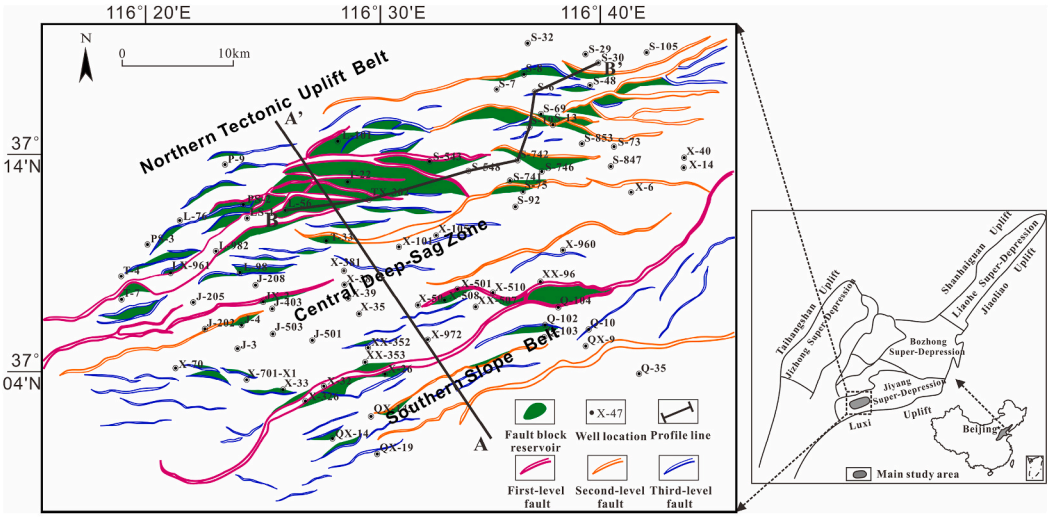
In a sedimentary basin, the properties of faults control tectonic and sedimentary processes, as well as the transport geofluids, such as oil and gas, which show the significance of faults in petroleum industry (Hindle, 1997; Hao et al., 2007; Alvar et al., 2009; Pei et al., 2015). Previous statistical studies on the petroleum industry have shown that among the discovered oil and gas reservoirs around the world, 32% are related to faults (Tan et al., 2019; Dong et al., 2021). During the exploration and exploitation of petroleum, fault seal evaluation is dominant in fault property analysis owing to its controlling effects on hydrocarbon migration and accumulation (Lindsay et al., 1993; Yielding et al., 1997; Eichhubl and Boles, 2000; Lao et al., 2020). Faults with poor sealing capacity could provide migrating pathways for geofluids, while faults with good sealing capacity could provide preserving and accumulating conditions for oil and gas by blocking effects (Alan et al., 2012; Wang et al., 2020).

Previous studies have shown several methods for fault seal evaluation. Knipe (1997) introduced triangle maps of lithologies juxtaposed in the hanging wall and footwall for qualitative evaluation of fault seals. Regarding quantitative evaluation methods, some researchers have introduced different parameters, such as clay smear potential (CSP), shale smear factor (SSF), shale gouge ratio (SGR) and normal stress of faults, for evaluating fault seals based on statistical principles (Fulljames et al., 1997; Zhou et al., 2000; Fu et al., 2005; Vrolijk et al., 2016).

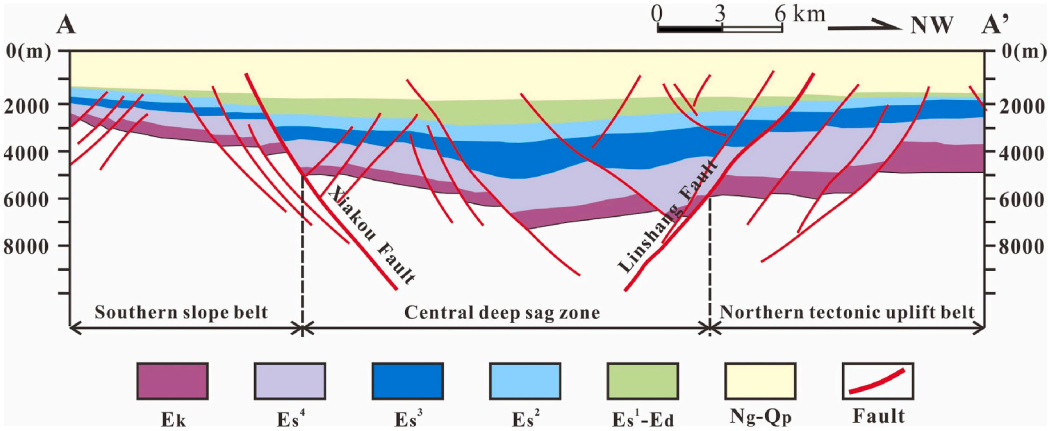
Although there are several methods for fault seal evaluation, they are based on a single fault seal mechanism or they only consider one controlling factor of the fault seal. Recent studies have demonstrated that

* Corresponding author. State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum (Beijing), Beijing, 102249, China.
E-mail address: lindachen@cup.edu.cn (D. Chen).

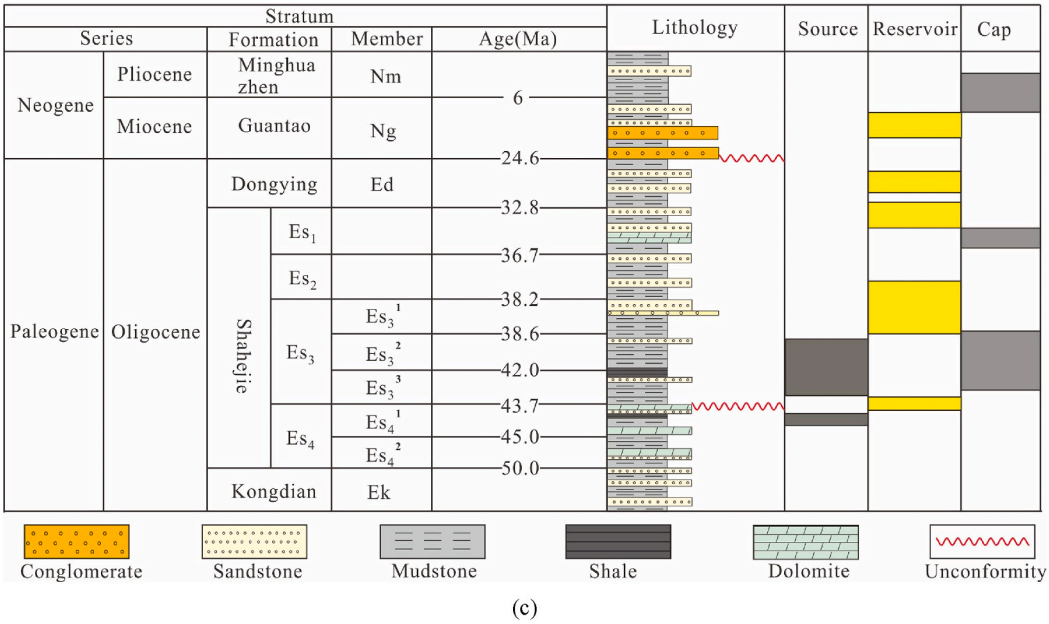
https://doi.org/10.1016/j.geoen.2023.212064
Received 4 September 2022; Received in revised form 12 April 2023; Accepted 2 May 2023
Available online 30 June 2023



(a)



(b)



(c)

Fig. 1. Location, oil reservoir distribution, fault development and stratigraphy of the Huimin Depression (a) The location of the Huimin Depression with the tectonic units, distribution of main faults, well location and hydrocarbon reservoir locations. (b) The cross section profile of the Huimin Depression. The location of the profile was showed in Fig. 1(a). (c) Generalized stratigraphic column for the Huimin Depression (Edited by Wang et al., 2020).

the sealing capacity of faults is controlled by multiple factors, including the lithology of the hanging wall and footwall of the fault, mud content, fault throw, fault-dip angle, normal stress, shale-smear indices, microfracture and carbonate cements. (Knott, 1993; Yielding et al., 1997; Baudon and Cartwright, 2008; Pei et al., 2015). In different geological backgrounds, the dominant controlling factors of fault seals are different (Caillet and Batiot, 2003; Zhang et al., 2011). In this case, fault seals could not show a good correlation to a single controlling factor based on linear regression using statistical analysis (Knipe et al., 1997; Færseth et al., 2007; Choi et al., 2016).

The Huimin Depression is one of the major petroliferous depressions in the Bohai Bay Basin, which is one of the largest rift basins in eastern China. There are over 90 faults in the Huimin Depression, which has led to widely distributed fault-block, hydrocarbon reservoirs. Among the 113 discovered hydrocarbon reservoirs, 86 are fault-block reservoirs. Thus, the evaluation of fault seals is of great significance for the analysis of hydrocarbon reservoir distribution and hence exploration and exploitation strategies. However, previous studies have shown that the fault seals in the Huimin Depression are complex and influenced by multiple factors (Chen et al., 2010; Feng et al., 2010). Furthermore, the evaluation of fault seals using existing methods (such as SGR or triangle maps) often provides inconsistent conclusions regarding real fault seal conditions (Gao et al., 2003; P.J. Ma et al., 2023). Thus, a new reliable recognition and classification method for weighting the analysis of the main controlling factors of fault seals is required.

Machine learning is the study of computer algorithms that can improve automatically through experience and by the use of data (Mitchell, 1997). In recent years, machine learning has been employed in many fields and subjects and has shown significant potentials for dealing with complex scientific problems. In this research, the decision tree (DT) and random forest (RF) algorithms were used for constructing a model for fault seal evaluation and prediction.

A decision tree is a set of procedures for classifying input training data into more homogeneous subgroups using generated rules or decisions called nodes (Friedl and Brodley, 1997; Quinlan, 2003). By training a DT, the input data can be divided into generated subsets with maximum information and minimum entropy (Bravo et al., 2014; Mnih et al., 2016). DTs have been widely used in geological analysis owing to the complexity of geological processes, including mineral identification, landslide susceptibility map determination and petroleum production prediction. (Li et al., 2013; Pradhan, 2013; Akkas et al., 2015). The analyzing results show the advantages of DT in weighting analysis and attribute prediction. The random forest is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time (Breiman, 2001; Martens, 2010). For classification tasks, which were used in this study, the output of the random forest is the class selected by most decision trees (Hastie et al., 2008; Piryonesi and EI-Diraby, 2021). Therefore, the random forest method usually can further improve the prediction and evaluation accuracy of the DT model. The algorithm of RF was applied for favourable reservoir prediction, pore fluid identification and seismic inversion (Asim et al., 2017; Rakers et al., 2017; Javier and Miguel, 2018). While, the utilization of machine learning in petroleum exploration is limited and unmature with several problems.

There are two main problems for the fault seal evaluation using machine learning method. The first one is the difficulty for the quantitative characterization of the geological features which control the fault seal (Piryonesi and EI-Diraby, 2021). The second one is the concerns about the reliability of the machine learning methods which are only based on data analysis and lack the theoretical support of petroleum geology (Caillet and Batiot, 2003; Zhang et al., 2011). For the first problem, this study provided a systematic approach of dataset construction including quantitative characterization of feature variables and feature engineering processes for determination of feature variables. For the second question, we applied the correlation analysis of DTs and RFs to analyse the dominate controlling factors of fault seal and validated their geological rationality on the basis of petroleum geological knowledge and theory (Huo et al., 2021; Chan et al., 2021). This study not only provided a new method for fault seal evaluation and prediction with the utilization of DT and RF, but also proofed the rationality and reliability of the machine learning methods for the petroleum exploration, and further discussed the key factors for the application of machine learning in the petroleum exploration and exploitation, which is of great significance for the development of intelligent petroleum exploration technology.

2. Geological setting

The Huimin Depression is located in the southwestern region of the Bohai Bay Basin, which is one of the largest rift basins in China. The whole depression is divided into three main tectonic zones by two first-level faults: the Linshang Fault in the north and the Xiakou Fault in the south (Fig. 1a). The three tectonic zones are the Northern Tectonic Uplift Belt (NTUB), the Central Deep-Sag Zone (CDSZ), and the Southern Slope Belt (SSB) (Fig. 1a) (Wang et al., 2019).

The Huimin Depression is a NE-SW-trending graben that underwent intensive rifting during the Cenozoic (approximately 64 Ma) Era. The extensional stress oriented N–S led to the formation of a large number of faults at different levels during the Paleogene Kongdian period to the Dongying period (Fig. 1b) (Wang et al., 2020). In addition to the Linshang Fault and Xiakou Fault, there are over 10 s-level faults and over 100 third-level faults developed in the Huimin Depression. The entire depression experienced two-stage tectonic evolution from the Cenozoic to the present: The first stage is the rift stage from the Kongdian period to the Dongying Period. During this stage the multi-scale faults were active with large-scale movements. Then the uplift and erosional processes occurred at the end of the Dongying period and lasted during approximately 24.6 Ma to 14 Ma. Since 14 Ma, the whole depression experienced the second tectonic stage of subsidence and depression without intensive tectonic movement and activity of faults (Fig. 1c).

The main sedimentary strata in the Huimin depression are Paleogene Kongdian Formation (E_k), Shahejie Formation (subdivided into 4 members of Es_4, Es_3, Es_3 and Es_1 from the bottom to the top), Dongying Formation and Neogene Guantao Formation (N_g) and Minghuazhen Formation (N_m) (Fig. 1c). The Shahejie Formation, which is the dominant source rock and reservoir, is characterized by dark mudstone and porous sandstone in lacustrine and river-delta facies (Wang et al., 2019). The petroleum system of the Huimin Depression contains the Es_3 source rock of lacustrine dark mudstone with the thickness of 200–400 m, the Es_3 and Es_2 river-delta sandstone reservoirs and the dark gray mudstone seals in Es_2 and Es_1 corresponding to the Es_3 and Es_2 reservoirs (Fig. 1c).

The Huimin Depression is characterized by abundant oil and gas resources. At present, a total of 113 hydrocarbon reservoirs have been discovered in the Huimin Depression, among which 86 are fault-block reservoirs, which shows that the faults control hydrocarbon accumulation in the study area (Fig. 1a). The distribution, size and oil saturation of these oil and gas reservoirs are different owing to the complexity of fault distribution and features.

3. Data and methods

3.1. Identification of fault states

To construct the evaluation model of fault seals, the states of the faults should be known first. Previous studies have shown that diagenetic minerals and fluid inclusions can provide evidence for fluid flow through faults and thus the identification of open or closed faults (Sorkhabi and Tsuji, 2005). However, it is unrealistic to collect sufficient core samples along a fault. Furthermore, for fault-state identification to determine hydrocarbon migration and accumulation, the scope of mineral and fluid-inclusion analysis is too small.

Both closed and open faults influence hydrocarbon migration and

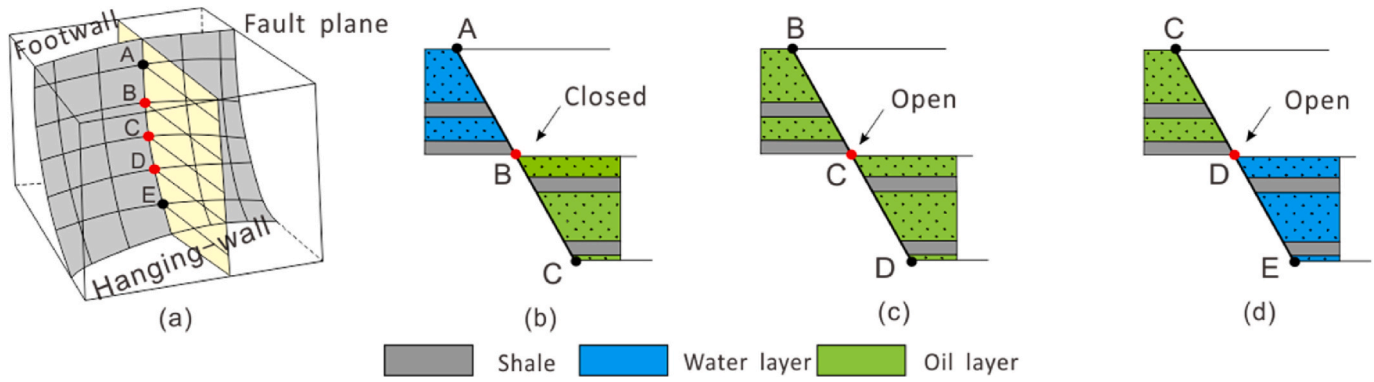


Fig. 2. Fault-state identification based on well-interpretation data (Edited by Zhang et al., 2011).

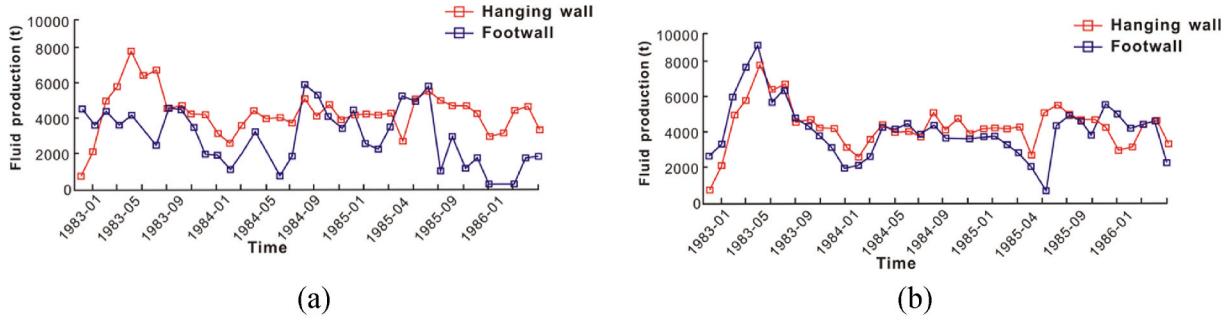


Fig. 3. Fault-state identification based on well productivity (Data from the Shengli Oilfield Branch of the China Petroleum and Chemical Corporation).

accumulation. An open fault serves as a pathway for hydrocarbon migration, while a closed fault provides a lateral barrier for hydrocarbon accumulation (Lindsay et al., 1993; Cartwright et al., 2007; Frery et al., 2015). Therefore, the distribution of discovered hydrocarbon reservoirs and well-production data can be indicators for fault states (Zhang et al., 2011; Caro et al., 2023). In this study, faults with or without oil and gas reservoirs and their related data regarding both the hanging wall and footwall were selected for fault-state identification and fault seal evaluation model construction.

Specifically, hydrocarbon reservoir distribution based on well-log interpretation of fluid properties (water layer, gas layer, oil layer and dry layer) and the productivity curves of wells were used as fault-state indicators in this study. Using the normal fault as an example (Fig. 2) and assuming that oil and gas can only migrate upward from the hanging wall to the footwall along the fault plane, the fault state at node B depends on the distribution of hydrocarbon reservoirs at node A and node C. If the well-log interpretation shows oil reservoirs are only distributed at node C, it indicates that hydrocarbon is blocked by the fault and cannot migrate from the hanging wall to the footwall through node B. Therefore, the fault at node B is closed (Fig. 2b). If well-log interpretation shows that the oil reservoirs are distributed at both node A and node C or only at node C, then hydrocarbons can partly or completely migrate from the hanging wall to the footwall through node B; hence, the fault at node B is open (Fig. 2c and d).

Productivity curves can also be used for fault-state identification (Luo et al., 2012; Lei et al., 2013; S.J. Ma et al., 2023). Additionally, for nodes A, B and C, if the wells at nodes A and C show similar productivity curves, the fluid flow of the hanging wall and footwall are connected, and the fault at node B is open (Fig. 3a). Otherwise, the fault at node B is closed (Fig. 3b).

For the construction of the CART, the status of the fault needs to be assigned. Considering that there are only two states (open or closed) for fault seals in this study, the Boolean value assignment was used. If the fault status is open, then the Boolean value is 1; otherwise, the Boolean value is 0.

3.2. Quantitative characterization of controlling factors of fault seals

As previously stated, the sealing capacity of the fault may be controlled by many geological factors, which determine the states of the fault (Karlsen and Skeie, 2006; Cartwright et al., 2007; Cipr et al., 2017). To construct a fault seal evaluation model by applying a decision tree, these geological factors should be characterized by representative parameters for quantitative analysis. Based on a previous study, the fault-dip angle, fault throw, normal stress in the fault plane, shale smear, development of microfractures and diagenesis are possible controlling factors of fault seals in the Huimin Depression (Zhao et al., 2004; Du, 2005; Chen et al., 2010). The controlling effects and quantitative characterizing parameters of these factors are discussed below.

3.2.1. Fault-dip angle

The dip angle of the fault plane can influence the normal stress and buoyancy on the fault plane (Hindle, 1997; Choi et al., 2016). Large dip angles lead to smaller normal stress and larger buoyancy. These changes cause an increase in the driving force and a decrease in the resistance of hydrocarbon migration and hence poor sealing capacity and the well-connecting capacity of the fault. The dip angle of a fault is not constant and always changes with depth (Torabi et al., 2019). In this study, the fault-dip-angle value is obtained from seismic interpretation data provided by the Shengli Oilfield Branch of the China Petroleum and Chemical Corporation (SINOPEC-SOB).

3.2.2. Fault throw (vertical)

Previous studies have shown that fault throws have a positive correlation with fault connectivity (Smit et al., 2023). A large fault throw can cause the development of fractures and microfractures near the fault plane or in the fault zone and therefore provide a large number of hydrocarbon migration pathways (Dewey et al., 1998). However, some

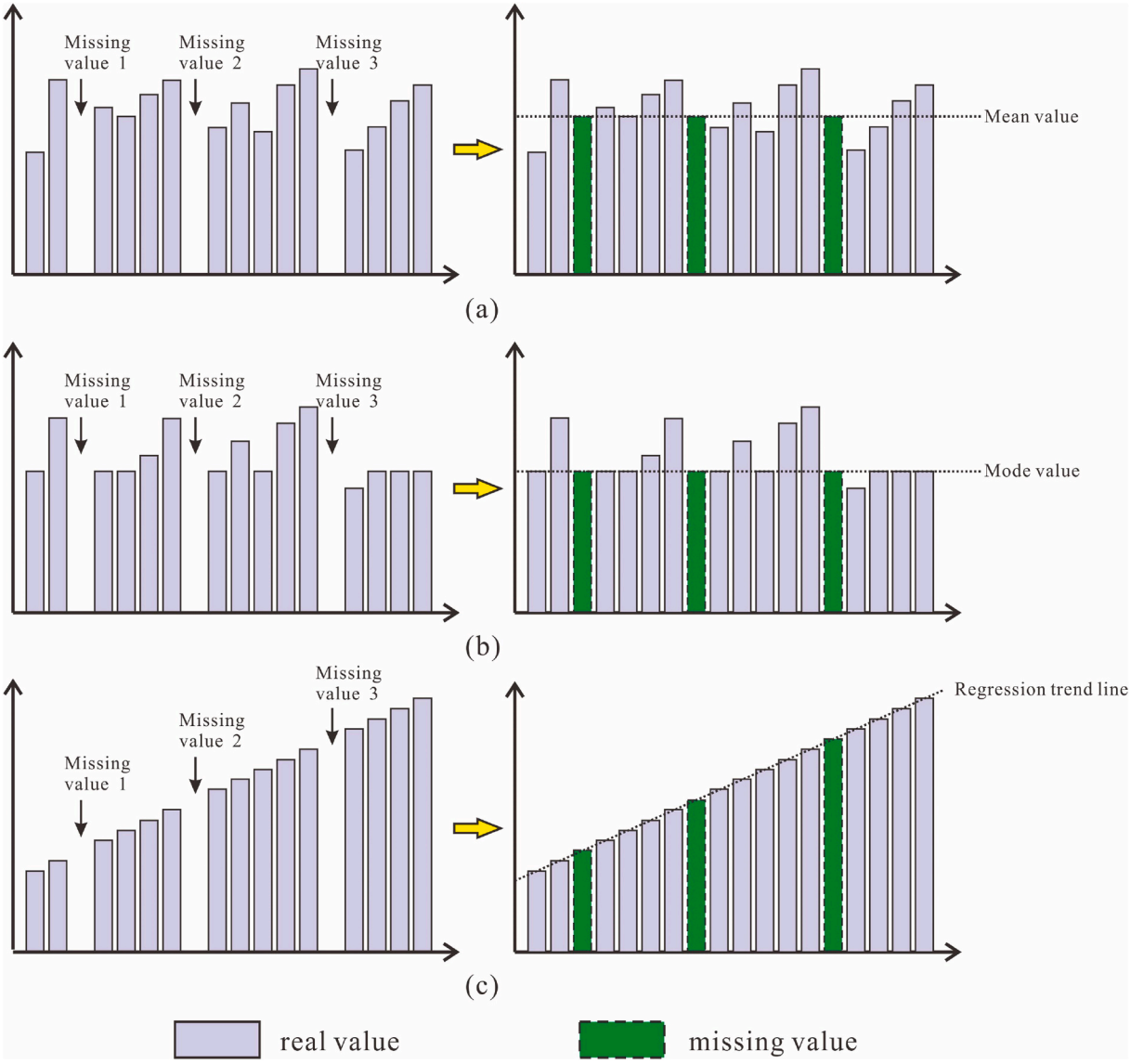


Fig. 4. Missing value imputing methods in machine learning. (a) Imputing value by mean value. (b) Imputing value by mode value. (c) Imputing value by constructing regression relation.

researchers have concluded that a large fault throw might cause strong cataclasis in rocks and thus sufficient pore space for rock-water interactions to form carbonate cements. These fine-grained rock particles and cements might occupy pore space and block pore throats, which tend to make the fault a barrier with a good-sealing capacity (Lindsay et al., 1993; Jobe et al., 2022). The fault-throw data were obtained from outcrop observations, reservoir reports and seismic interpretations in this study.

3.2.3. Fault length

The fault length represents the size of a fault, which also contains information on the intensity of tectonic movement (Hao et al., 2010; Morris et al., 2016). Faults with longer lengths tend to be more open than those with shorter lengths (Liu et al., 2018; Dong et al., 2022). Fault length was measured by outcrop observations and seismic interpretation in this study.

3.2.4. Normal stress of the fault plane

The normal stress of the fault plane is considered as a major controlling factor for fault seals (Harper and Lundin, 1997; Lu and Wang, 2010). Previous studies have shown that the normal stress determines the threshold pressure of lateral hydrocarbon migration through the fault plane (Lu et al., 1996; Lyu et al., 2023). A larger normal stress indicates a better lateral-sealing capacity of the faults. In general, the normal stresses of reverse faults are larger than those of normal faults (Linjordet and Skarpnes, 1992; Lao et al., 2022). Many researchers have introduced methods for calculating normal stress (Cowie and Scholz, 1992b; Fu et al., 2005, 2015). In this study, Fu et al.'s (2015) calculation model was used:

$$P = P_1 + P_2 = \delta \sin\beta \sin\alpha + H(\rho_b - \rho_w) \times 0.009876 \cos\alpha \quad (1)$$

where P represents the normal stress (MPa), P_1 represents the lateral stress from tectonic activity (MPa), P_2 is the overburden stress (MPa), δ is the regional tectonic principal stress (MPa), β is the angle between the direction of the fault strike and δ (°), α is the fault-dip angle (°), H is the burial depth (m), and ρb and ρw are the densities of overburden sediments and formation water (g/cm^3), respectively.

In this research, the value of δ is from published articles (Hou et al., 2006; Guo et al., 2009; Wang et al., 2020), and the values of β and α are from reservoir reports and seismic profiles provided by SINOPEC-SOB. The values of ρb and ρw are calculated by using density well-log data.

Table 1
Results of chi square test by feature selection process in sklearn software.

parameter	Fault throw (m)	Fault-dip angle (°)	Fault length (m)	Normal stress (MPa)	SGR	Ccarbonate[a] (%)	Dsf (cm/cm^2)
Chi square	0.511	0.580	0.525	0.508	0.528	0.685	0.702

[a] Ccarbonate: content of the carbonate cements.

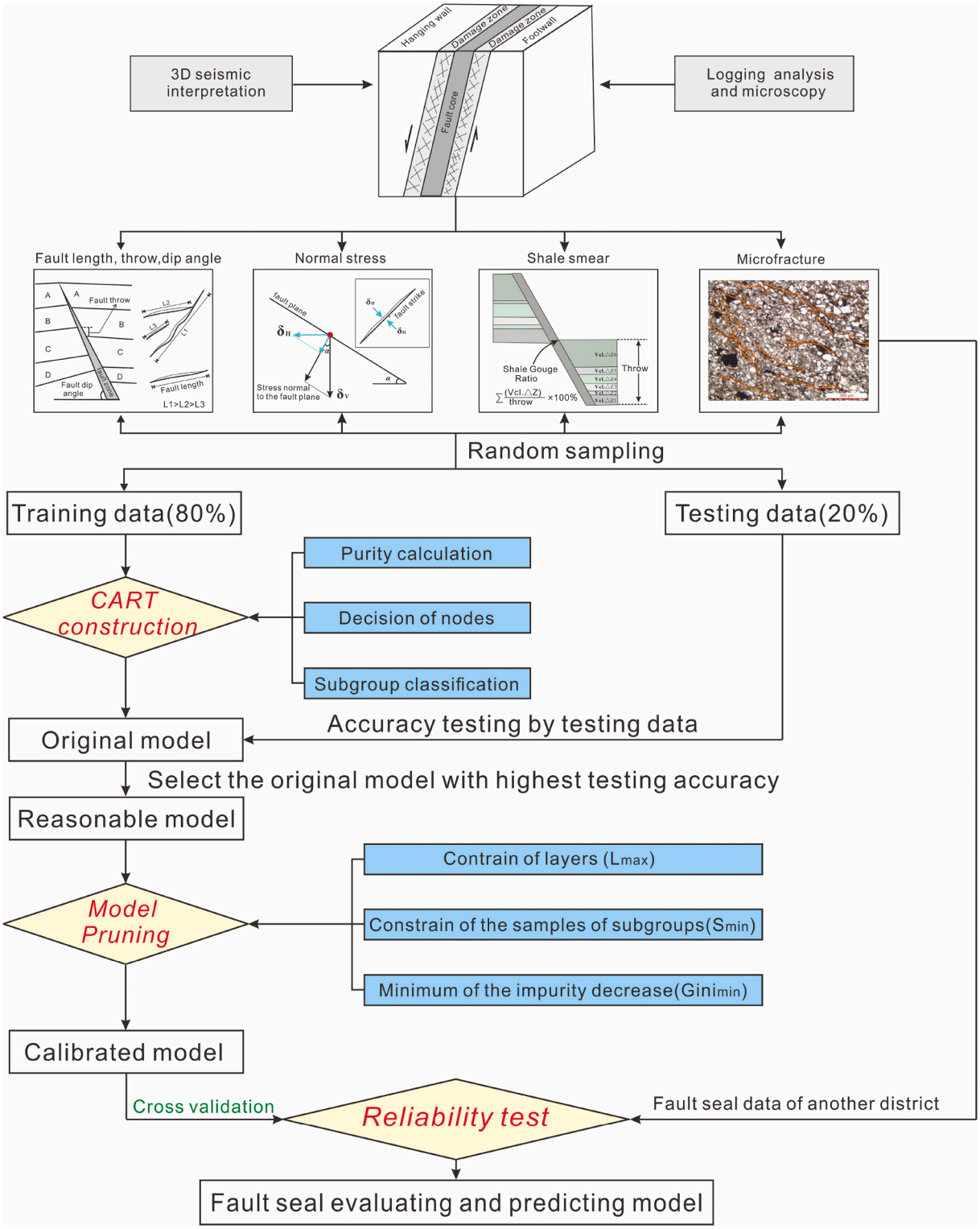


Fig. 5. Workflow chart of the construction and application of evaluation and prediction model by a decision tree.

3.2.5. Shale smear

Shale smears have been observed in outcrop studies and experiments by many researchers (Lindsay et al., 1993). They have noted that the faulting of sand-shale sequences can form a continuous, multilayered clay gouge along the fault plane (Caro et al., 2023). The structure of the clay smear is an effective seal to fluid flow. Several algorithms for characterizing the shale-smear effect have been introduced, including the CSP, SSF and SGR (Fulljames et al., 1997; Torabi et al., 2019). Considering that the SGR is the most commonly used parameter and is in the database we obtained, the SGR value was selected as the characterizing parameter for the effect of shale smears:

$$\mathrm{SGR} = \frac{\sum[(\text{Zone thickness}) \times (\text{Zone shale fraction})]}{\text{fault throw}} \times 100\% \quad (2)$$

3.2.6. Development of microfractures

The cataclasis of sand grains can produce a fault gouge of finer-grained material that increases the sealing capability of the fault. However, cataclasis can also lead to fractures and microfractures along the fault (Hull, 1988). The development of microfractures can increase the permeability of fault zones and hence provide conduits for hydrocarbon migration (Kalani et al., 2015; Liu et al., 2017; Teixeira et al., 2017; Panahi et al., 2019). In this study, the surface density of microfractures (D_{sf}) was used to characterize the development of microfractures. D_{sf} is defined as the total length of microfractures in a unit area observed by microscopy (Bergosh and Lord, 1987):

$$D_{sf} = \frac{1}{A_s}\sum_{i=1}^{n} L_i \quad (3)$$

where D_{sf} is the surface density of the microfracture, cm/cm^2; A_s is the unit area under the scope of microscopy, cm^2, which was calculated by Microimage software; and L_i is the length of the microfracture, cm, which was measured in the selected photo of the thin section.

3.2.7. Diagenesis

The movement of the fault and cataclasis of sand grains can provide an open environment and connect deep-burial and shallow-fluid flows (Bruna et al., 2021; Schultz and Hofmann, 2021). Frequent fluid flow may provide good conditions for water-rock interactions and the generation of cementation, especially carbonate cements, which can partially or completely fill pore space and pore throats, ultimately creating a hydraulic seal in the fault zone (Hodson et al., 2016; Michie et al., 2021). The development of carbonate cements is characterized by cement content observed and calculated by Zeiss Merlin microscopy with the image analysis software Microimage. The software can recognize certain cements by identifying their colour, shape and texture and can then calculate the content of the cement.

3.3. Feature engineering

The feature engineering process is required for determining the characterization parameters and improving the quality of the original dataset (Dai et al., 2020). In this paper, the standardization, imputing for the missing value and feature selection processes were used for dealing with the original dataset.

3.3.1. Standardization

For the convenience of model construction and calculation, all the original data needs standardization. In this study, we used the tool of standardization in sklearn software and chose Z-score standardization to make all the collected data to range from −1 to 1 and conform to normal distribution (Botu and Ramprasad, 2015).

3.3.2. Imputing for the missing value

The missing value existed in a total of 19 pairs of the data in the original dataset. The missing value includes fault dip angle, content of carbonate cements and microfracture. As usual, the imputing methods for the missing value include mean value replacement, mode replacement and regression replacement (Fig. 4) (Breiman, 2001; Caillet and Batiot, 2003). In this study, we used the mean of the existed value to replace the missing value of content of carbonate cement and microfracture, and the missing value of fault dip angle was imputed by the mode.

3.3.3. Feature selection by correlation analysis

Previous studies have shown that fault dip angel, fault throw, fault length, normal stress, shale smear, carbonate cements and development of microfracture are main controlling factors to fault seal. However, the correlation between these characterization parameters and the fault seal needs further analyses to select the proper features for machine learning. In this study, feature selection process of “filter” was employed for feature selection (Kreimeyer et al., 2021). The chi square test showed that the content of carbonate cement and the development of microfracture had highest chi square (Table 1). Furthermore, all the parameters had the chi square lager than 0.5. Considering that the fault state was influenced by multiple factors which may influence each other, we concluded that all the 7 characterization parameters had influences on the fault seal and should be used as the feature value of the DT and RF.

3.4. Fault seal prediction using decision trees

A decision tree is a hierarchical model that splits independent input variables into homogeneous subgroups by finding decision rules called nodes (). DTs can be divided into a classification tree or a regression tree based on whether the input variables are discrete or continuous, respectively (). Previous studies have introduced many algorithms for constructing decision-tree models, including ID3, C4.5 and classification and regression tree (CART) (Breiman et al., 2007; Quinlan, 1993). ID3 cannot overcome the overfitting of the model compared with C4.5 and CART. Compared with C4.5, CART is a binary tree, which is more suitable because the target variable contains only two categories. In this study, the target variable-fault, state-only contains two categories: open and closed. Therefore, the CART was selected for this research. Considering that the required outcome of fault seal prediction is a discrete category rather than a numerical value, the classification tree is more comfortable for predicting fault seals. Therefore, the classification decision tree based on CART algorithms was used to perform the decision-tree algorithm by using sklearn software in this study.

3.4.1. Preparation of training and testing datasets

For the prediction model constructed by CART, pairs of data containing the feature variables and target variable are required (Fig. 5). Specifically, for a pair of data, the feature variables are fault-dip angle, fault throw, normal stress, SGR value, surface density of microfractures and carbonate cements. The target variable is the fault state, which contains two attributes: open or closed. In this study, 171 pairs of data were used to form a total dataset. The total dataset is then divided into two subsets: (1) part one, which contains 80% of the data that are used in the training phase of the fault seal prediction model and provides the decision tree of the CART, and (2) part two, which contains the remaining 20% of the data that are used for validation of the prediction model and to confirm its accuracy. Considering that different sampling compositions of the training and testing datasets will lead to different DT prediction models with different accuracies, 100,000 compositions of the training and testing datasets by random sampling were used to find the best prediction model with the highest accuracy.

3.4.2. CART construction

The fault seal evaluation model by CART is constructed in four steps presented below (Fig. 5):

Fig. 6. Utilization of 10-fold cross validation (Edited by Kang et al., 2019; T = Average test accuracy, Ti = iterated test accuracy).

(1) Purity calculation

The training data for CART construction are divided into different subgroups with the outcomes of fault seals (open or closed) following the decision rules, which are called nodes in the decision-tree algorithm. The node represents the method of partitioning, which is selected by a key parameter named "node purity" (Nikolaev and Slavov, 1998; Nock and Jappy, 1999). In a certain group of data, node purity is defined as the ratio of data with a certain attribute to the total data (Quinlan, 2003). In the CART, impurities are more commonly used and characterized by the Gini index (Quinlan, 2003; Leibovici et al., 2011):

$$Gini = 1 - \sum_{i=0}^{c-1}[p(i|t)] \quad (4)$$

where c is the category of the attributes of the target variable, which is 2 in this study (open and closed), and $p(i|t)$ is the ratio of data with attribute i to the total data for given node t. In other words, $p(i|t)$ is the purity of node t.

(2) Decision of nodes

A higher purity or a lower Gini index indicates that the node is effective for the classification of target variables. Therefore, the target of CART construction is to find the nodes with the highest total Gini index of all nodes (Itani et al., 2020; Huang et al., 2020).

(3) Subgroup determination

The first selected node with the lowest purity or highest Gini index is used to binarily divide the total training data into two subgroups. Then, the purity of the two subgroups is calculated again for selecting new nodes, which will divide each subgroup (which is called the parent subgroup) into another two subgroups (which is called the descendant group). When every descendant subgroup contains only one attribute of the target variable, which indicates that the given decision tree provides the unique outcome of the fault status for each pair of input parameters of fault-seal-controlling factors, the partition is terminated, and the fault-seal evaluation model based on CART is constructed.

Note that the nodes were formed by setting the critical values of selected feature variables, which are the controlling factors of fault seals in this study. Therefore, the Gini indices can also reflect the information gain of all feature variables and can make the correlation analysis between fault seals and the controlling factors available.

3.4.3. Reasonable model selection

The accuracy of the prediction model constructed by the DT is determined by the training error and generalization error (Kim, 2016). The training error represents the misclassification of the training data by the training model, while the generalization error represents the misclassification of the testing data. An ideal prediction model must be characterized by both low training error and low generalization error (Li et al., 2009; AbouEisha et al., 2016). In this study, a 100,000 CART

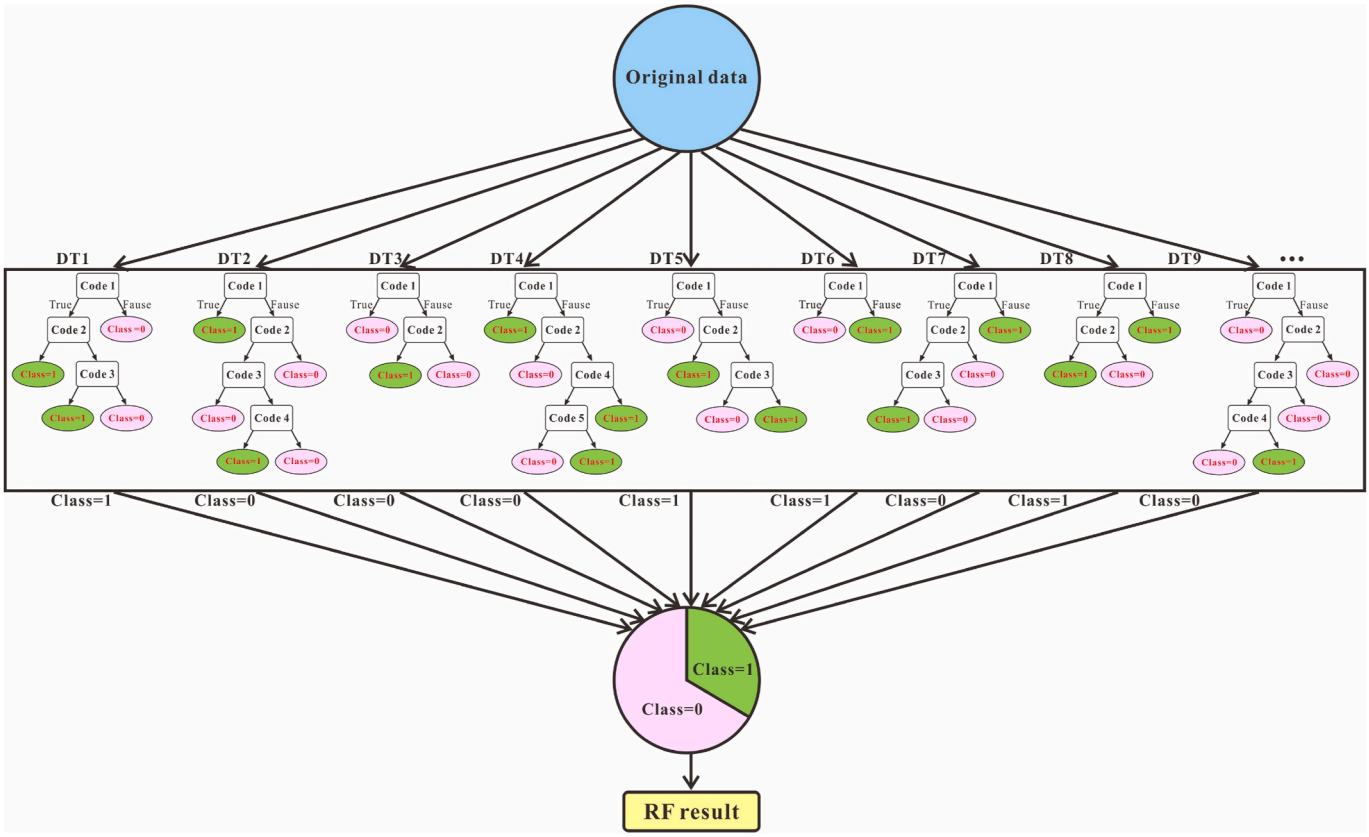


Fig. 7. Workflow chart of the construction and application of evaluation and prediction model by random forest method.

model was constructed to select the best model with the lowest training error and generalization error as the reasonable model.

3.4.4. *Overfitting calibration by model pruning*

During the training process, the training error of a predicting model can be minimized by the information gain algorithm of the DT (Jegadeeshwaran and Sugumaran, 2013; Huo et al., 2021). Therefore, the methods for decreasing the training error include enlarging the sample size of training data or changing the type of DT (Chandra et al., 2010; Saettler et al., 2017).

However, the decrease in training error always leads to an increase in generalization error because the training model may only fit the specific given training data instead of providing a universal prediction model for any input data. The training model with low training error and high generalization error is called an overfitting model (Quinlan, 2003). The calibration of overfitting DTs is called pruning. In this study, we pruned the overfitting trees by the following steps:

(1) Constrain of layers (L_{max})

The layer of the DT represents the time of decision of descendant subgroups. Too many subgroup divisions will decrease the generalization of the model.

(2) Constrain of the samples of subgroups (S_{min})

The number of samples in a subgroup showed the effectiveness of the nodes. A subgroup containing only a few samples indicates that this subgroup and its nodes only focus on a specific dataset.

(3) Minimum of the impurity decrease ($Gini_{min}$)

The impurity also shows the effectiveness of nodes, and a lower Gini index corresponds to a lower effectiveness of the nodes. Nodes with extremely low Gini indices should be abandoned to improve the generalization of the evaluating model.

3.4.5. *Reliability test*

The calibrated model should be tested by new testing data. While considering the amount of data is not enough, the cross-validation is used in this research. Cross-validation is a resampling method that uses different portions of a dataset to test and train a model on different iterations (Geisser and Eddy, 1979; Kelter, 2021). The goal of cross-validation is to estimate the performance of reliability and generalization of a model (Burman, 1989). In this paper, we selected k-fold cross-validation for reliability test of the model. In 10-fold cross-validation, the original sample is randomly partitioned into 10 equal sized subsamples, of which a single subsample is retained as the test data for testing the model, the remaining 9 subsamples are used as training data. The cross-validation is then repeated 10 times. With each of the 10 subsamples used exactly once as the test data, the 10 results can then be averaged to produce a single estimation (Fig. 6). The advantage of this method over repeated random sub-sampling is that all observations are used for both training and testing, and each observation is used for testing exactly once. Furthermore, cross-validation is more effective when the amount of data is small (Kang et al., 2019; Wainer and Cawley, 2021).

3.5. *Fault seal prediction using random forest method*

Based on the CART model, the random forest method was used to improve the rationality of the model by a single decision tree (Asim et al., 2017; Rakers et al., 2017). The random forest algorithm is an ensemble learning method for classification and regression (Javier and Miguel, 2018). It works by constructing multiple decision trees and giving the final output by combining all the results of the decision trees it

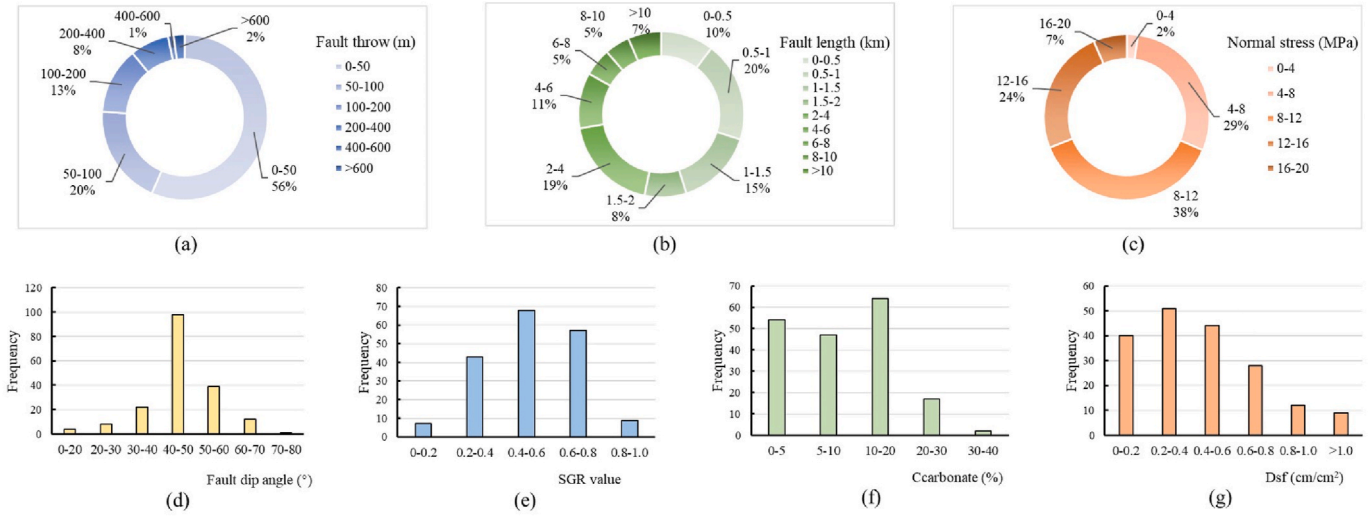


Fig. 8. Distribution of the feature variables of fault seal. (a) Distribution of fault throw; (b) Distribution of fault length; (c) Distribution of fault normal stress; (d) Distribution of fault dip angle; (e) Distribution of SGR value; (f) Distribution of carbonate cement content; (g) Distribution of surface density of the microfracture;

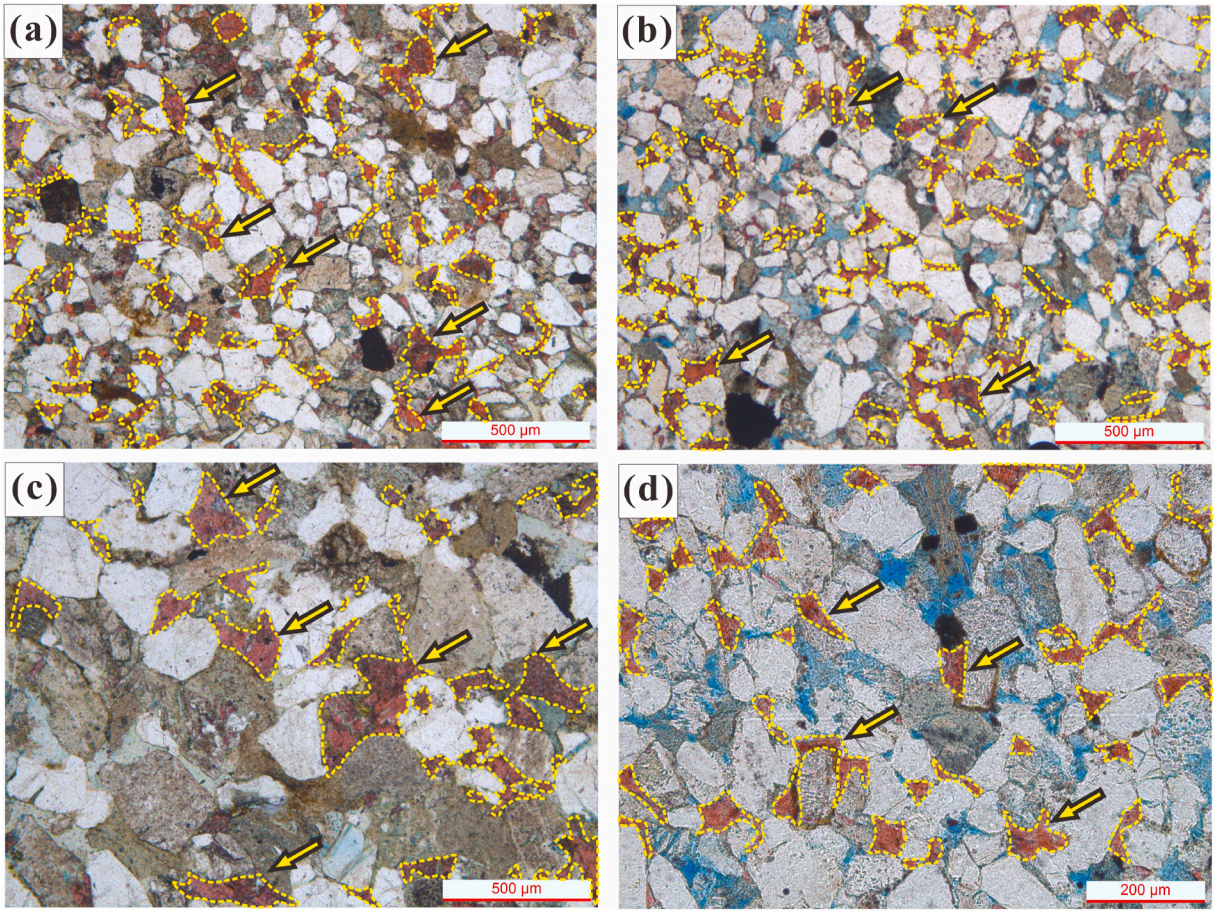


Fig. 9. Thin section photos of carbonate cements at fault zone in the Huimin Depression (The well locations are showed in Fig. 1c). The carbonate cement of calcite is dyed red and highlighted with yellow dotted circles and arrow, the pore space is dyed blue. (a) Developed calcite cement which occupied pore spaces (Ccarbonate = 20.4%), Well X-326, 3130 m. (b) Developed calcite cement filled-in primary pores, (Ccarbonate = 15.5%), Well X-507, 2819 m. (c) Less calcite cement filled-in intragranular pores, (Ccarbonate = 7.7%), Well X-510, 3032 m. (d) Less calcite cement between quartz, (Ccarbonate = 5.5%), Well S-543, 3322 m.

contains (Kam, 1995). In the random forest method used in this study, 20 CART models were held for judging the state of the fault. The output of the random forest was the class selected by most decision trees among the 20 DTs (Fig. 7). As an algorithm of ensemble learning, random forest usually can improve the evaluation accuracy and rationality compared with the DT.

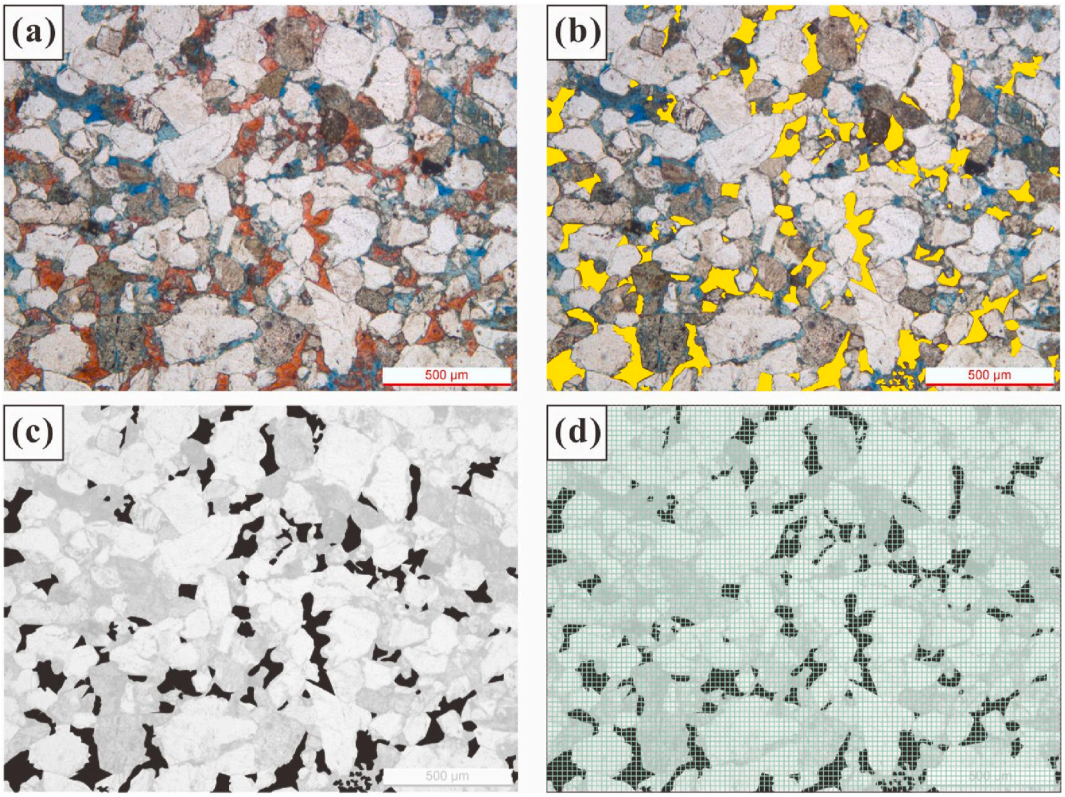


Fig. 10. Measurement of the content of calcite cements by image analysis software Microimage (The well locations are showed in Fig. 1c). (a) Thin section photo of optical microscopy with calcite cements dyed in red, Well X-33, 3302 m. (b) Calcite cements recognized (in yellow) by image software Microimage. (c) Calcite cements determined and marked in black by image software Microimage. (d) Determination of calcite cement content by gridding and calculation of black part.

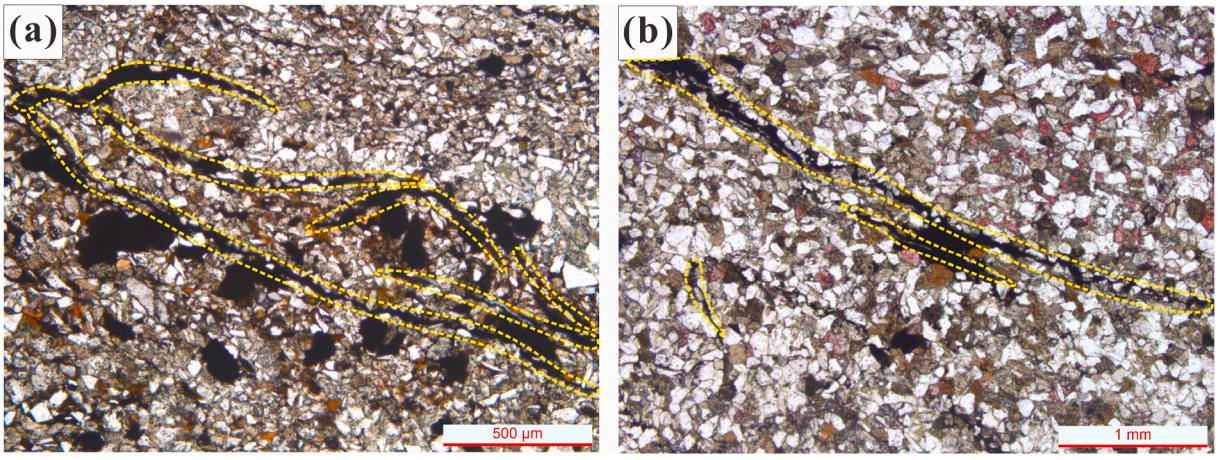


Fig. 11. Microfractures observed by optical microscopy. The area of microfractures were highlighted with yellow dotted circles. (a) Developed microfractures filled by asphalt and connected pore spaces, Well X-510, 3032 m. (b) Less microfractures filled by asphalt, Well X-501, 3503 m.

4. Results

4.1. Training and testing datasets

4.1.1. Original dataset

A total of 184 pairs of data of the characteristics and status of faults in the Huimin Depression were obtained for the construction and calibration of the CART. The complete dataset was showed in the appendix of the paper (see Appendix A). Considering that the fault throw and fault-dip angle for a fault are not constant, the median of the fault throw and fault-dip angle were used. The distribution of the feature variables of the fault seal showed large disparities. The distribution of fault dip angle and SGR value was characterized by normal distribution, while the development of carbonate cements and microfracture showed unbalanced distribution (Fig. 8). In addition, based on the observation of thin sections and previous studies, the carbonate cements in the Huimin Depression were dominated by calcite cements with intensive heterogeneity along the faults (Fig. 9). The content of calcite cements ranged from 0.27 to 31.22% with an average of 5.76% according to the image analysis software (Fig. 10). Calcite cements developed more near the fault zone and showed strong heterogeneity in different area (Fig. 9). Microfractures were also widely distributed in the Huimin Depression. The microfractures were characterized by lengths ranging from approximately 13.4 to 146.2 μm and an average width of approximately

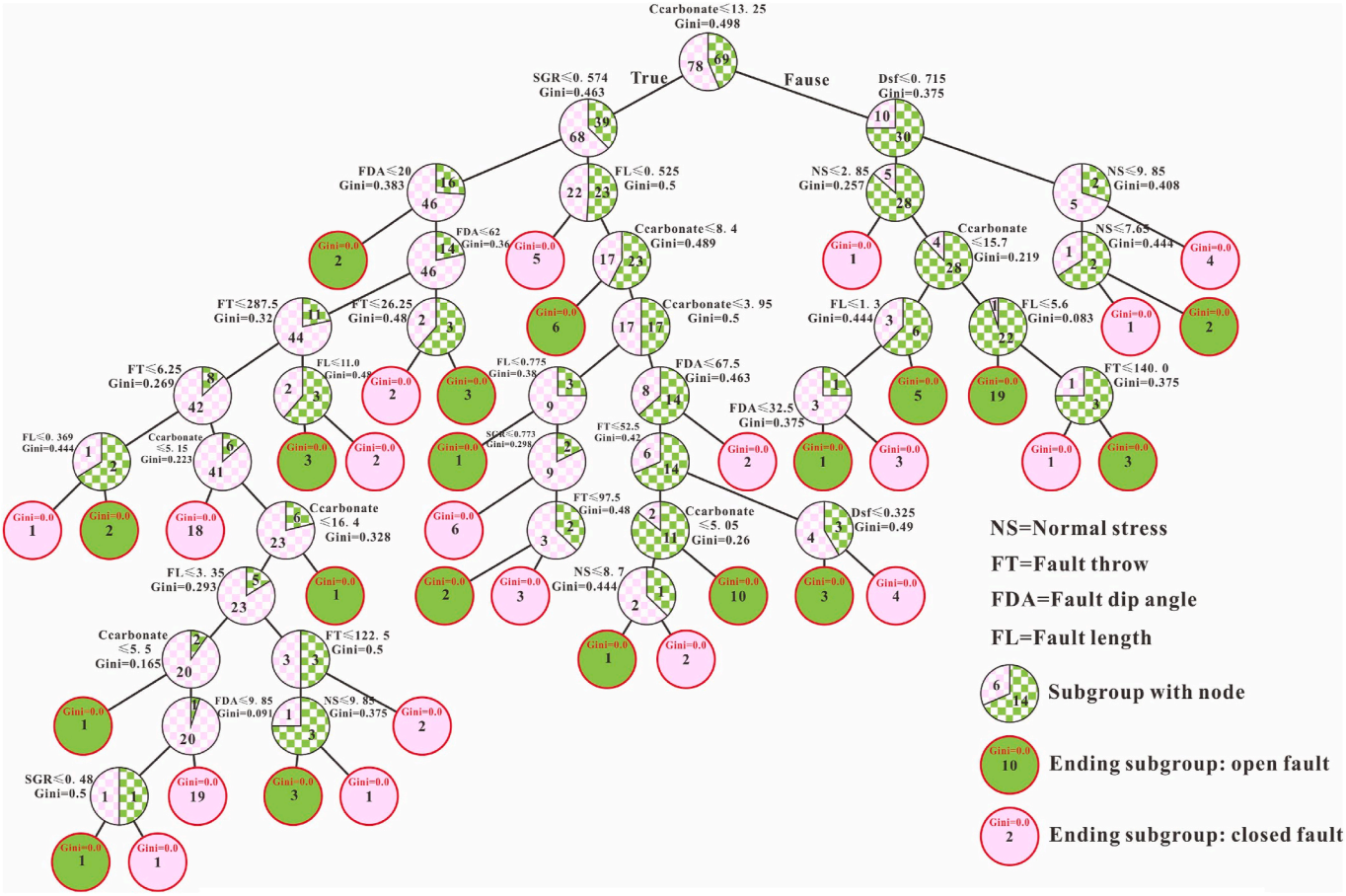


Fig. 12. CART fault seal evaluation and prediction model based on Python code using sklearn software.

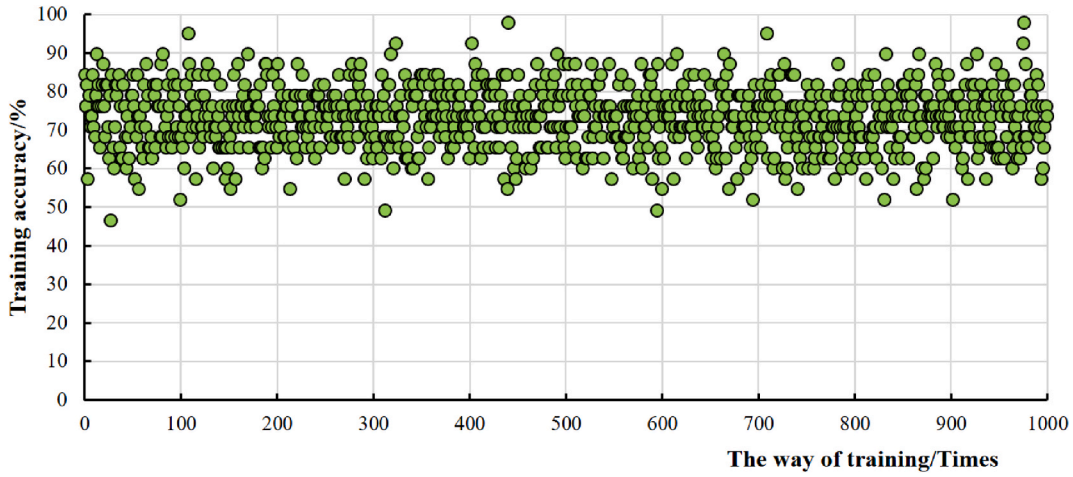


Fig. 13. Records of training accuracy for 1000 constructed CARTs.

3.3 μm (Fig. 11). Microfractures were more developed near the fault zones with asphalt filling, which indicated that the hydrocarbons had migrated through the microfractures of the faults (Fig. 11). The calculation of D_{sf} also showed that the microfractures were more developed near the faults. However, the microfractures also showed a complex distribution with heterogeneity, with D_{sf} ranging from 0 to 3.22 cm/cm^2 in the Huimin Depression. To conclude, the possible controlling factors showed complex relationships to the distribution of faults and hence led to difficulty in analysing and predicting the capability of fault seals. Considering the heterogeneity of the distribution of carbonate cements and microfractures, 10 thin section samples were collected for each test point of the fault seal, and the content of carbonate cements and value of D_{sf} were the average from the 10 thin sections.

4.1.2. *Determination of training dataset and testing dataset*

For the construction of the CART, 181 pairs of data were divided into two attributes of training data and testing data by random sampling. In each dataset, 145 data points were used as training data for constructing the fault-seal evaluation model by CART. Another 36 data points were applied to test the accuracy of the model. In this study, a total of 100,000

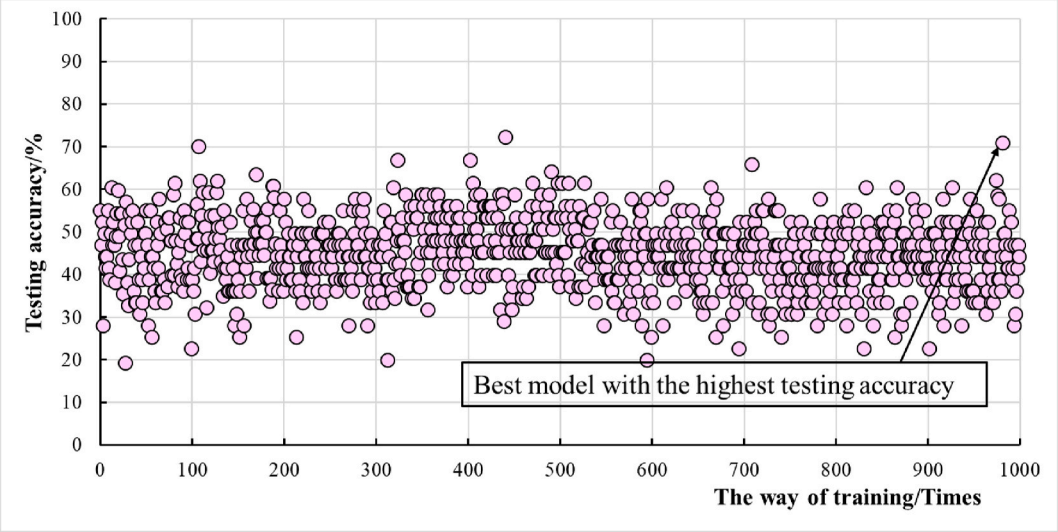


Fig. 14. Records of testing accuracy for 1000 constructed CARTs.

Table 3
Records of hyperparameters and testing accuracy.

L_{max}	Testing accuracy (%)	S_{min}	Testing accuracy (%)	$Gini_{min}$	Testing accuracy (%)
2	50.5	4	60.5	0.04	62.2
3	55.2	8	76.3	0.08	65.4
4	70.1	12	77.7	0.12	58.9
5	62.4	16	80.2	0.16	78.5
6	92.1	20	68.2	0.2	80.2
7	82.2	24	65.5	0.24	85.4
8	79.5	28	65.5	0.28	69.5
9	71.1	32	63.2	0.32	64.8
10	53.2	36	63.2	0.36	72.2
11	52.1	40	58.1	0.4	63.2
12	53.1	44	58.1	0.44	58.1
13	50.1	48	58.1	0.48	58.1

pairs of training and testing data were determined to find the best fault seal evaluation model by CART.

4.2. Model training

4.2.1. Construction of CART

In this study, sklearn software was applied to construct the CART decision tree based on Python code. The CART with the highest accuracy (smallest training error) is shown in Fig. 12. This CART was a 13-level decision tree with 38 ending subgroups in total in this research, which were highlighted in pure coloured ellipses with red circles. The ending group in pink represented the output of closed fault, while the ending group in green indicated the output of open fault, The subgroups were marked in mixed colour and showed the fault state composition of the data points in certain subgroup. The node and the Gini index were also showed near the subgroup.

With the classification of the group and generation of subgroups, the Gini index of the nodes gradually decreased; if the Gini index for a node reached 0, the classification stopped. When all Gini indices reached 0, the construction of the CART was completed. The ending subgroup for a node could appear at any layer of the decision tree. In the given CART in Fig. 12, the ending subgroups appeared in the third to the thirteenth layer of the decision tree.

4.2.2. Training accuracy and training error

The training accuracy and training error represent the evaluation accuracy of the CART for the training data. In this study, an iteration procedure was applied to search for the best evaluation model. During this iteration process, 100,000 models were compared one by one to determine the model with the highest training accuracy. The training accuracies of the last 1000 models are shown in Fig. 13. The training accuracy for the CART sample mainly ranged from approximately 65 to 80%, with maximum and minimum values of 97.96% and 46.63%, respectively.

4.3. Model testing

The high training accuracy of a CART might indicate that this decision tree was overfitting to a certain training dataset and hence lost generalization for the testing data and other groups of data. The overfitting of a decision tree would then lead to low accuracy for the testing data and high generalization error. For the 1000 constructed CARTs shown in Fig. 14, the results of the testing accuracy mainly ranged from approximately 40%–55% with maximum and minimum values of 19.23% and 72.28%, respectively. The average testing accuracy was 44.87%, and the generalization error reached 45.13%. Therefore, the best model was selected with the highest testing accuracy of 72.28%. Compared with the high training accuracy, the results of testing accuracy indicated that most CARTs were overfitted, which indicates that overfitting calibration by pruning is necessary.

4.4. Overfitting calibration and cross validation

As shown above, the CART was overfitted, and overfitting calibration by pruning was required. In this study, the best selected model shown in Fig. 12 was calibrated by the process of pruning and setting three hyperparamters in the following ways: (1) constraining of the layers, which means that the maximum of the layer of a decision tree (L_{max}) will be defined and the subgroups over the maximum will be pruned. (2) Constraining of the samples of subgroups, which defined the minimum of the samples in a subgroup (S_{min}). Subgroups with samples less than the defined minimum will be pruned. (3) Minimum impurity decrease ($Gini_{min}$), which means that the classification of the group (subgroups) will not appear when the Gini index is less than the defined minimum.

In this study, the grid search method was used for selecting the best hyperparameters. The grid search was an exhaustive searching through a manually specified subset of the hyperparameter space of a learning algorithm. Considering the large time consumption of the grid search method, an estimated value range of each hyperparameter was defined.

The value of L_{max} was firstly tested and the L_{max} of 6, 7, 8 and 9

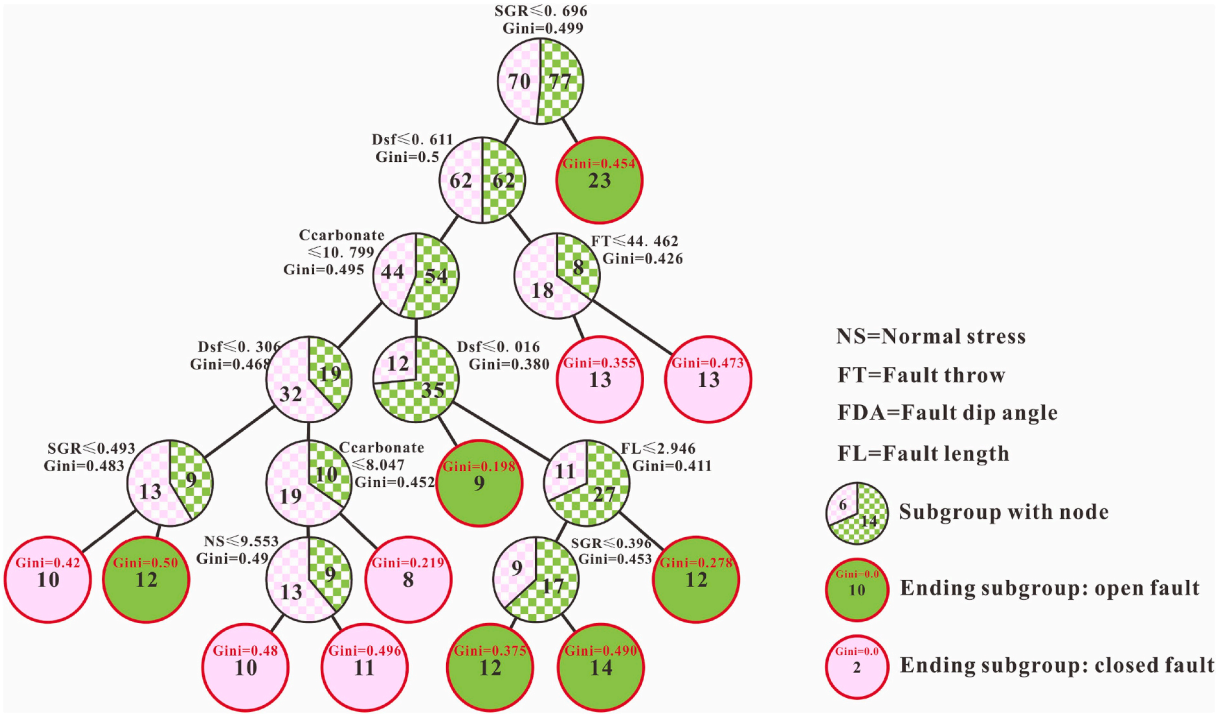


Fig. 15. The CART sample with the highest testing accuracy after overfitting calibration by pruning.

Table 4
Testing accuracy of 10-fold cross-validation for the CART model.

Iteration	1	2	3	4	5
Testing accuracy (%)	83.3	72.2	77.8	88.9	72.2
Iteration	6	7	8	9	10
Testing accuracy (%)	94.4	88.9	66.7	77.8	83.3

showed better performance on testing accuracy. Then the value of S_{min} was tested with the L_{max} of 7. The testing result showed higher testing accuracy when the S_{min} ranged from 8 to 36. Finally the $Gini_{min}$ was defined when L_{max} and S_{min} equalled to 7 and 12, respectively. The result showed that the testing accuracy was higher when $Gini_{min}$ ranged from 0.16 to 0.28 (Table 3).

According to the hyperparameter test, the grid search programme for L_{max}, S_{min} and $Gini_{min}$ was determined: The searching range of the L_{max} was from 6 to 9 with the step width of 1, the searching range of the S_{min} was 8–36 with the step width of 1, the searching range of the $Gini_{min}$ was from 0.16 to 0.28 with the step width of 0.01. The iteration method was also used for searching the best couple of hyperparameters.

After the iteration and testing, the best couple of hyperparameters was found: $L_{max} = 7$, $S_{min} = 10$, and $Gini_{min} = 0.21$. The pruned model was shown in Fig. 15. After pruning, the prediction model was tested again with the testing data. The results showed that the testing accuracy of the pruned model improved from 72.28% to 80.60%, which showed the necessity of overfitting calibration.

For further testing the generalization and reliability of the CART model and avoiding the selection bias of the test dataset, 10-fold cross-validation was used in this research. The result showed that the average of the testing accuracy by 10 times cross-validation was 80.6% (Table 4), which showed the well generalization and high reliability of the model.

4.5. Improvement of the model by using random forest

Based on the CART model by decision tree algorithm, the random

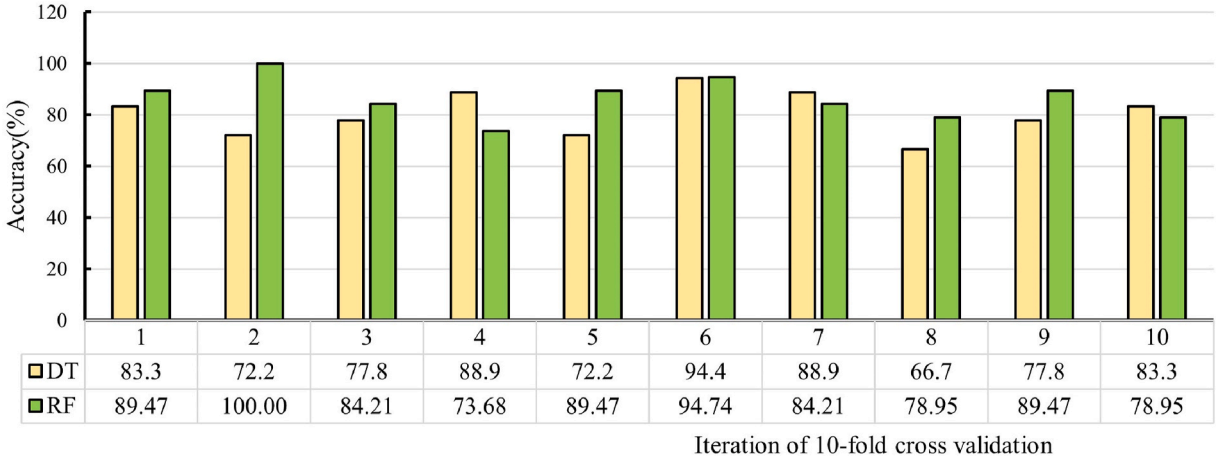


Fig. 16. Comparison of performance between DT and RF model trough 10-fold cross validation.

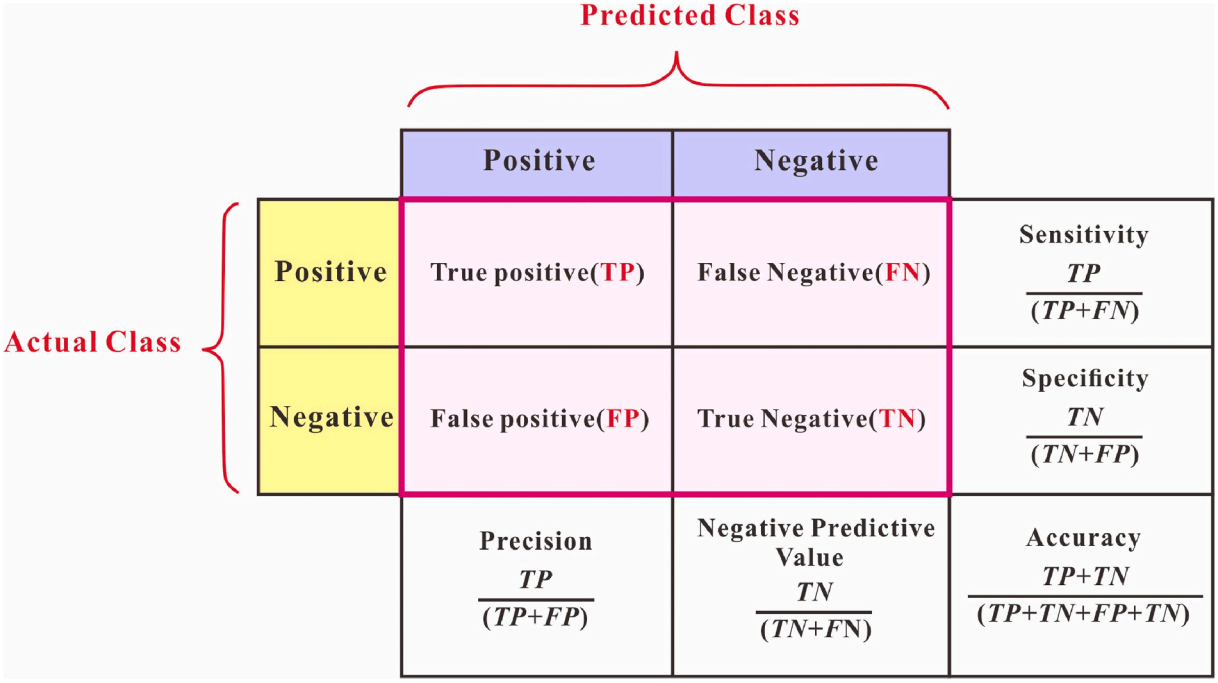


Fig. 17. Classification of evaluation metrics for classification problems in machine learning.

Table 5
Evaluation metrics for the Testing accuracy of 10-fold cross-validation.

Iteration	Accuracy	Sensitivity	Precision	specificity	negative predicted value	F1-score
1	0.833	0.800	0.875	0.889	0.778	0.842
2	0.722	0.600	0.875	0.857	0.636	0.706
3	0.778	0.667	0.800	0.400	0.923	0.500
4	0.889	0.857	1.000	1.000	0.667	0.923
5	0.722	0.600	0.875	0.857	0.636	0.706
6	0.944	0.667	1.000	1.000	0.938	0.800
7	0.889	0.600	1.000	1.000	0.867	0.750
8	0.667	0.714	0.500	0.833	0.333	0.769
9	0.778	0.800	0.667	0.923	0.400	0.857
10	0.833	0.833	0.833	0.714	0.909	0.769
Average	0.806	0.714	0.843	0.847	0.709	0.775

forest method was used for further improving the accuracy and rationality of the model. In this study, 20 CARTs were used for random forest construction by sklearn software. The dataset for every CART were still made by random sampling. For each CART in the random forest, the ratio of the training data and testing data was 70%–30%. The constructing process of each tree was the same as the constructing model by CART which was shown above. In this study, we totally constructed 100 groups of random forest for finding the best random forest. The evaluating results by RFs showed that the highest accuracy of the model among the 100 RFs was 86.54%, which is indeed slightly higher than the CART model. The 10-fold cross validation of the RF model showed that the accuracy ranged from 73.68% to 100% with an average of 86.32%, which was also slightly higher than that of the CART model (Fig. 16).

4.6. Evaluation metrics of the testing accuracy

The testing accuracy was easily to be affected by the composition of training and testing dataset and hence cannot reflect the real reliability of the evaluation model. In this research, sensitivity (recall), precision, specificity, negative predictive value and F1-score were used as the metrics for the evaluation of the model accuracy.

For a binary classification model, the output result could be divided into four types (Fig. 17). The case that actual class of positive was correctly predicted as positive by the model was called true positive (abbreviated as TP), the case that actual class of negative was correctly predicted as negative was called true negative (abbreviated as TN). While, the case that actual class of positive was falsely predicted as negative was called false negative (abbreviated as FN), the case that actual class of negative was falsely predicted as positive was called false negative (abbreviated as FP). The sensitivity reflected the evaluation accuracy of the actual positive class for a dataset and was defined as the ratio of TP and the sum of TP and FN:

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{5}$$

The precision reflected the evaluating accuracy of the predicted positive class for a dataset and was defined as:

$$Precision = \frac{TP}{(TP + FP)} \tag{6}$$

The specificity reflected the evaluating accuracy of the actual negative class for a dataset and was defined as:

$$Specificity = \frac{TN}{(TN + FP)} \tag{7}$$

The negative predicted value reflected the evaluating accuracy of the

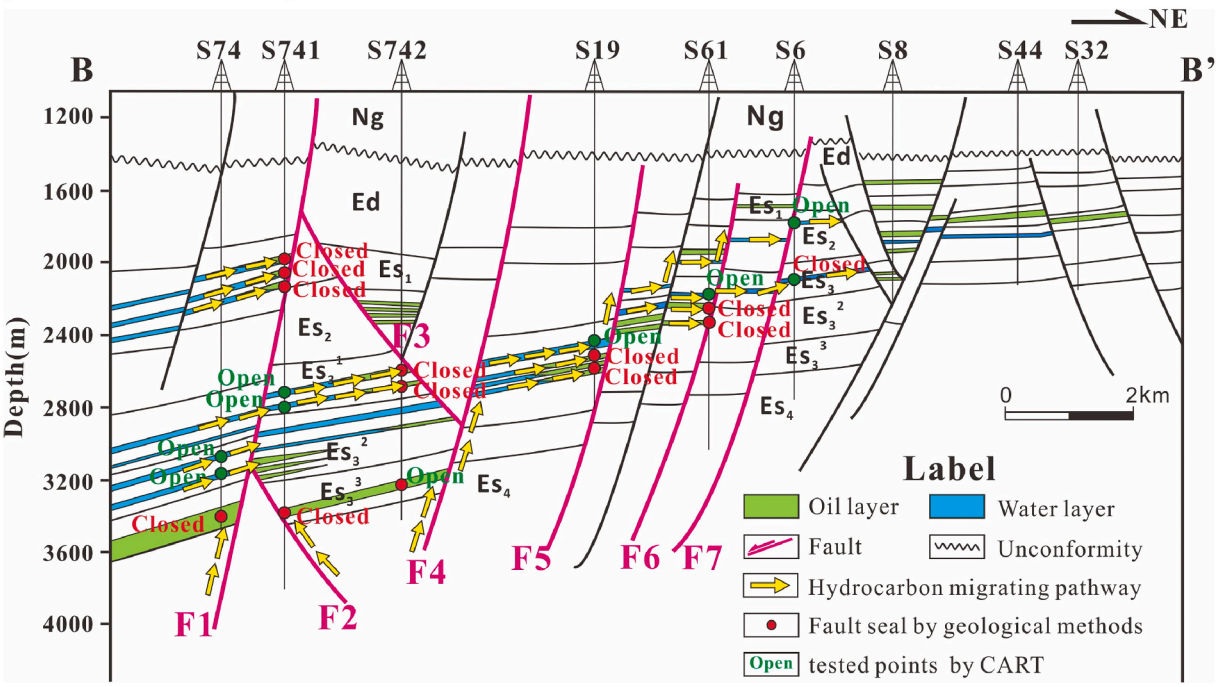


Fig. 18. Profile B–B': Fault seals determined by geological analysis and CART in the Huimin Depression.

Table 6
Fault-state evaluation results of profile A-A' in the Huimin Depression.

Testing points	Depth (m)	Fault number	Fault state	CART	RF	SGR value	SGR result*	SSF value	SSF result*	Normal Stress value	Normal stress result*
1	3101.2	F1	Open	Open	Open	0.722	Closed*	1.22	Closed	8.3	Closed
2	3211.9	F1	Open	Open	Open	0.231	Open	8.13	Open	8.5	Closed
3	3484.1	F1	Closed	Closed	Closed	0.322	Open	7.32	Open	8.9	Closed
4	2062.2	F1	Closed	Closed	Closed	0.292	Open	7.11	Open	6.9	Uncertain
5	2143.2	F1	Closed	Closed	Closed	0.277	Open	8.14	Open	7.1	Uncertain
6	2240.8	F1	Closed	Closed	Closed	0.454	Uncertain	2.98	Closed	7.3	Uncertain
7	2798.4	F1	Open	Open	Open	0.475	Uncertain	5.32	Uncertain	7.8	Closed
8	2832.9	F1	Open	Open	Open	0.112	Open	8.99	Open	7.8	Closed
9	3421.1	F2	Closed	Closed	Closed	0.677	Closed	2.23	Closed	9.0	Closed
10	2730.2	F3	Closed	Closed	Closed	0.311	Open	7.75	Open	7.7	Closed
11	2810.3	F3	Closed	Closed	Closed	0.862	Closed	6.88	Open	7.9	Closed
12	3300.4	F4	Closed	**Open**	Open	0.143	Open	7.98	Open	8.6	Closed
14	2683.9	F5	Open	Open	Open	0.302	Open	7.32	Open	7.5	Closed
15	2765.8	F5	Closed	Closed	Closed	0.318	Open	0.71	Closed	7.5	Closed
16	2794.3	F5	Closed	Closed	Closed	0.235	Open	7.49	Open	7.8	Closed
17	2253.9	F6	Open	Open	Open	0.343	Open	8.75	Open	6.5	Uncertain
18	2399.1	F6	Closed	Closed	Closed	0.355	Open	7.15	Open	6.8	Uncertain
19	2471.6	F6	Closed	Closed	Closed	0.283	Open	7.35	Open	7.0	Uncertain
20	2200.2	F7	Open	**Closed**	Open	0.214	Open	8.54	Open	6.7	Uncertain

*: The incorrect evaluation results by fault seal evaluation methods are in bold.
*: The criteria of the SGR, SSF and normal stress methods are listed in Table 7.

Table 7
Criteria of SGR, SSF and normal stress methods.

Fault state	Open	Uncertain	Closed	References
SGR method	<0.45	0.45–0.75	>0.75	Yielding et al., 1997
SSF method	>6.5	3–6.5	<3	Yielding et al., 1997
Normal stress method	<5	5–7.5	>7.5	Fu et al., 1996

predicted negative class for a dataset and was defined as:

$$Negative\ predictive\ value = \frac{TN}{(TN + FN)} \tag{8}$$

The F1-score was an evaluation metric for testing both of sensitivity and precision:

$$F1 - score = \frac{2 * Precision * Sensitivity}{(Precision + Sensitivity)} \tag{9}$$

In this paper, the fault state of open was defined as positive, and the state of closed was defined as negative. The value of the evaluation metrics for the testing accuracy of the 10-fold evaluation was listed in Table 5. The value of sensitivity ranged from 0.600 to 0.833 with an average of 0.714, the value of precision ranged from 0.667 to 1.000 with the average of 0.843, the value of specificity ranged from 0.400 to 1.000 with the average of 0.847, the value of negative predicted value ranged from 0.333 to 0.909 with the average of 0.709, the value of F1-score ranged from 0.500 to 0.923 with the average of 0.775. The distribution of sensitivity and precision showed the high efficiency and reliability of the model in predicting open state of faults, while the

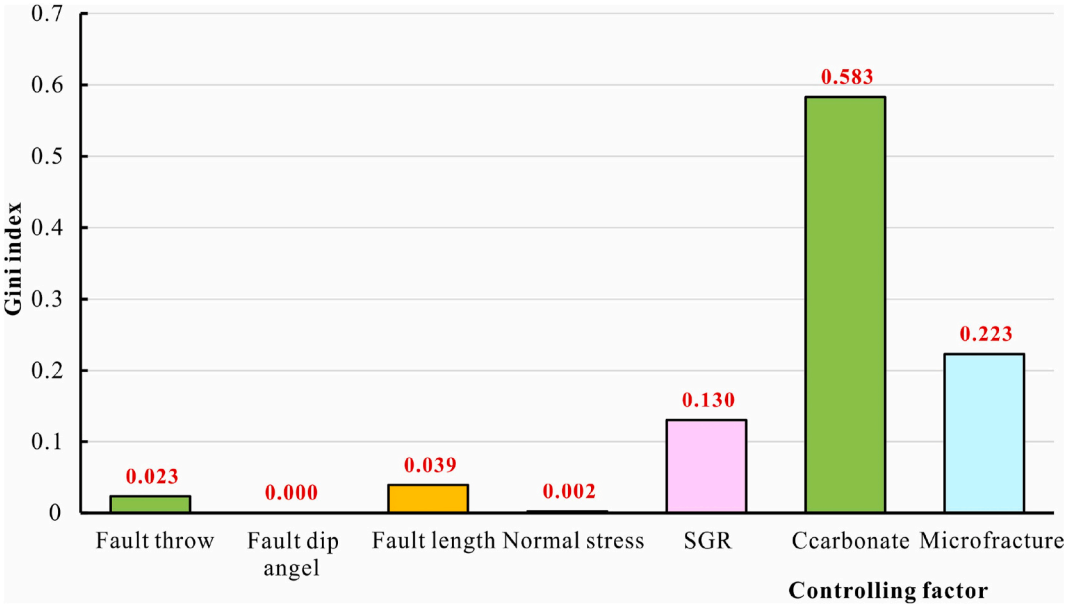


Fig. 19. The Gini index value calculated by DT.

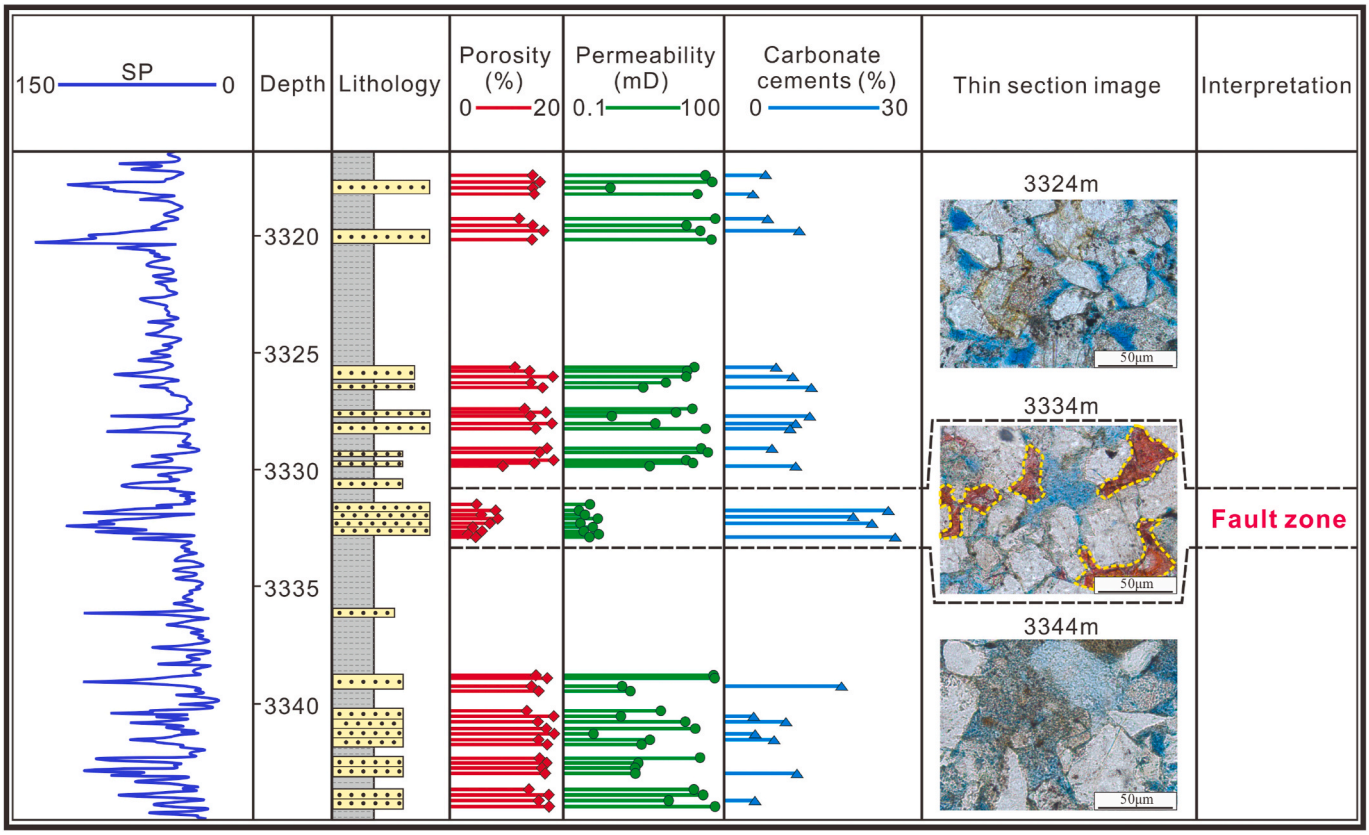


Fig. 20. Well-log profile of Well X-503: Calcite is more developed in fault zones with low porosity and low permeability and hence good seal capacity. The carbonate cement was dyed in red and marked by yellow dotted circles in the thin section photo. The location of the well is shown in Fig. 1.

distribution of specificity and negative predicted value showed the high efficiency and reliability of the model in predicting closed state of the faults. The distribution of F1-score indicated that the model performed stable during the evaluating processes. Therefore, the evaluation metrics showed the high reliability of the fault seal evaluating model. For the petroleum exploration and exploitation, the different metrics represent different meanings. The selection and prior evaluation metrics in different situations will be discussed in section 5.2.1.

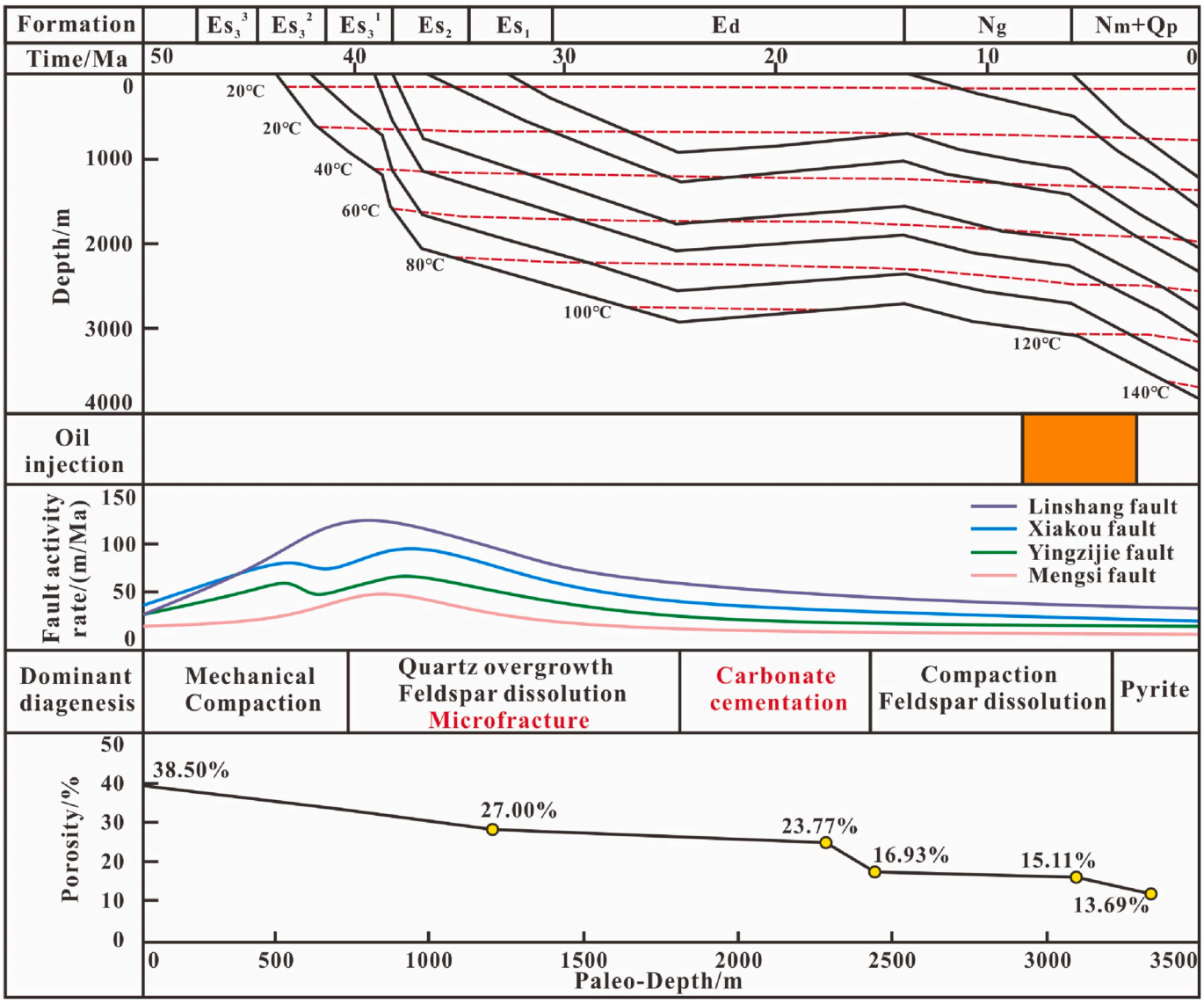


Fig. 21. Matches of burial history, oil injection period, evolution of fault, diagenesis history and evolution of porosity (Modified by Shi et al., 2017).

5. Discussion

5.1. Comparison between decision tree and random forest

5.1.1. Algorithm and accuracy

A DT is based on a set of features or attributes present in the dataset and combined with a series of sequential decisions made to reach a specific result, the construction of a DT only requires one set of training data (Quinlan, 1993). Therefore, the model of decision tree is easily affected by certain training data. Without the constraint of hyper-parameters, the model of DT learns particular features of one dataset, which often leads to overfitting and poor generalization (Fratello and Tagliaferri, 2019).

The random forest is a tree-based machine learning algorithm that uses the power of multiple DTs for making decisions (Kam, 1995). Comparing to DT algorithm, the random forest algorithm overcomes overfitting by randomly splitting and selecting training and testing datasets for constructing different DTs and hence generate the final output by combining the output of each individual DT. Therefore, the random forest algorithm showed the characterization of "random" during the model training process, which led to better performance in model accuracy and reliability (Speiser et al., 2019). In this research, the model of random forest indeed showed higher accuracy than that of the DT model. However, constrained by the amount of data, the advantage of random forest algorithm was not fully reflected on testing accuracy.

5.1.2. Effectiveness

As is stated above, the algorithm of DT is simple, the training and testing of a DT costs less time than almost all other machine learning algorithms. Furthermore, the calibration of a DT model is also convenient by setting the key superparameters (Kaminski et al., 2017). While the RF algorithm requires large amount of data to construct a "forest", and also requires much more training and testing time than the DT algorithm (Smith et al., 2013). The calibration of the RF model also costs much time comparing to the overfitting calibration of DT. To conclude, the DT algorithm shows higher effectiveness in utilization and time consumption.

5.1.3. Interpretability and visualization

Previous studies have shown that the greatest advantage of the DT algorithm is the interpretability and visualization (Quinlan, 2003). So far, most machine learning algorithms can only provide a black box model for users. The constructing processes and the logic of judgements cannot be clearly showed (Tom, 1997; Bishop, 2006). The

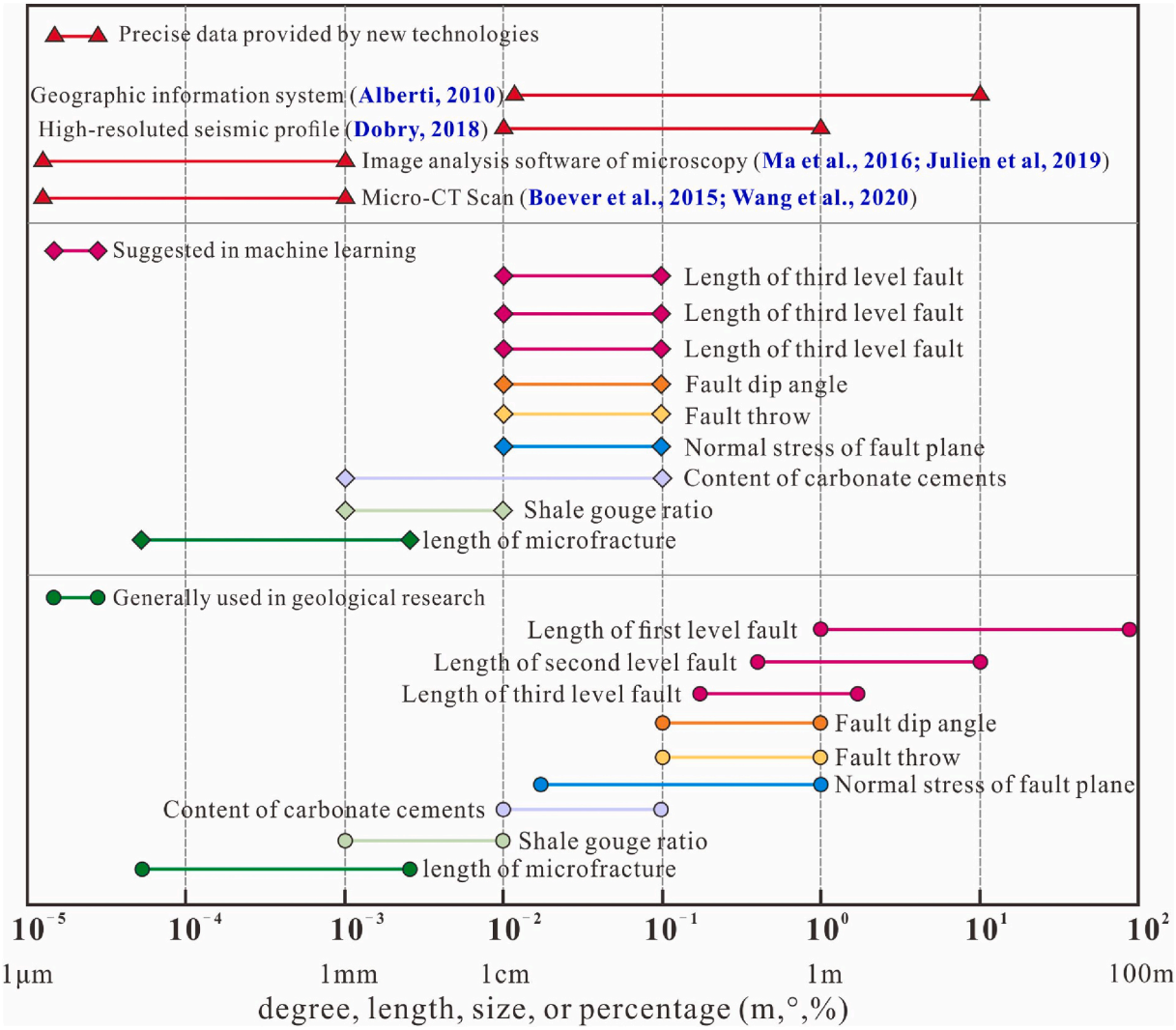


Fig. 22. Precision of data generally used in geological research and suggested in machine learning in fault seal analysis.

interpretability or visualization is one of the most important problems when the machine learning methods are used in other research fields, including petroleum exploration and exploitation. Although random forest seems to just combine multiple DTs, the calculating and judging processes of random forest are difficult to interpret (Denisko and Hoffman, 2018). The advantages of interpretability and visualization make DT algorithm easier to use and understand.

In summary, DT performs better in interpretability and visualization, while RF indeed showed slightly higher testing accuracy. However, constrained by the amount of data, the disparities on reliability and effectiveness between the two algorithms were not reflected in this case.

5.2. The reliability of the predicting model by machine learning

5.2.1. Application of the machine learning for fault seal prediction in the Huimin depression

The constructed prediction model was applied to evaluate fault seals in the Huimin Depression to test the accuracy and reliability of the model under the geological background. The fault state was judged by both geological methods (stated in section 3.1) and the prediction model based on machine learning, and the results are shown in Fig. 18 and Table 6. The results showed that the DT model provided 18 correct judgements at all 20 testing points; the accuracy of the model in the Huimin Depression reached 90%. The RF model provided 19 correct judgements of all the test points with the accuracy of 95%.

Table 6 also shows the results of some previous fault seal evaluation methods, which contained the SGR, SSF, and normal stress of the fault plane. The calculating formula of SGR is shown in equation (2), the calculating formula of normal stress of the fault plane is shown in equation (1), the calculation of SSF is listed below:

$$\mathrm{SSF} = \frac{\textit{Fault throw}}{\textit{Shale layer thickness}} \tag{10}$$

According to the criteria of the SGR, SSF and normal stress methods (Table 7), the evaluation results showed that the three methods provided 10, 11 and 8 correct judgements among the 20 test points, which corresponded to accuracies of 50%, 55% and 40%, respectively. Therefore, the prediction model based on machine learning was more appropriate for evaluating fault seals in the Huimin Depression. Note that the two test points along fault F4 and fault F7 (test points 12 and 20 shown in Table 5, respectively) at which the prediction model provided an incorrect judgement of the fault state were relatively farther from the fault plane. Therefore, the data collected by these two points and the real characteristics of the fault may have some differences, which indicates that the selection of the test points may influence the accuracy of the fault state by using machine learning methods. The test points on the fault plane or near the fault plane are appropriate for the application of the prediction models.

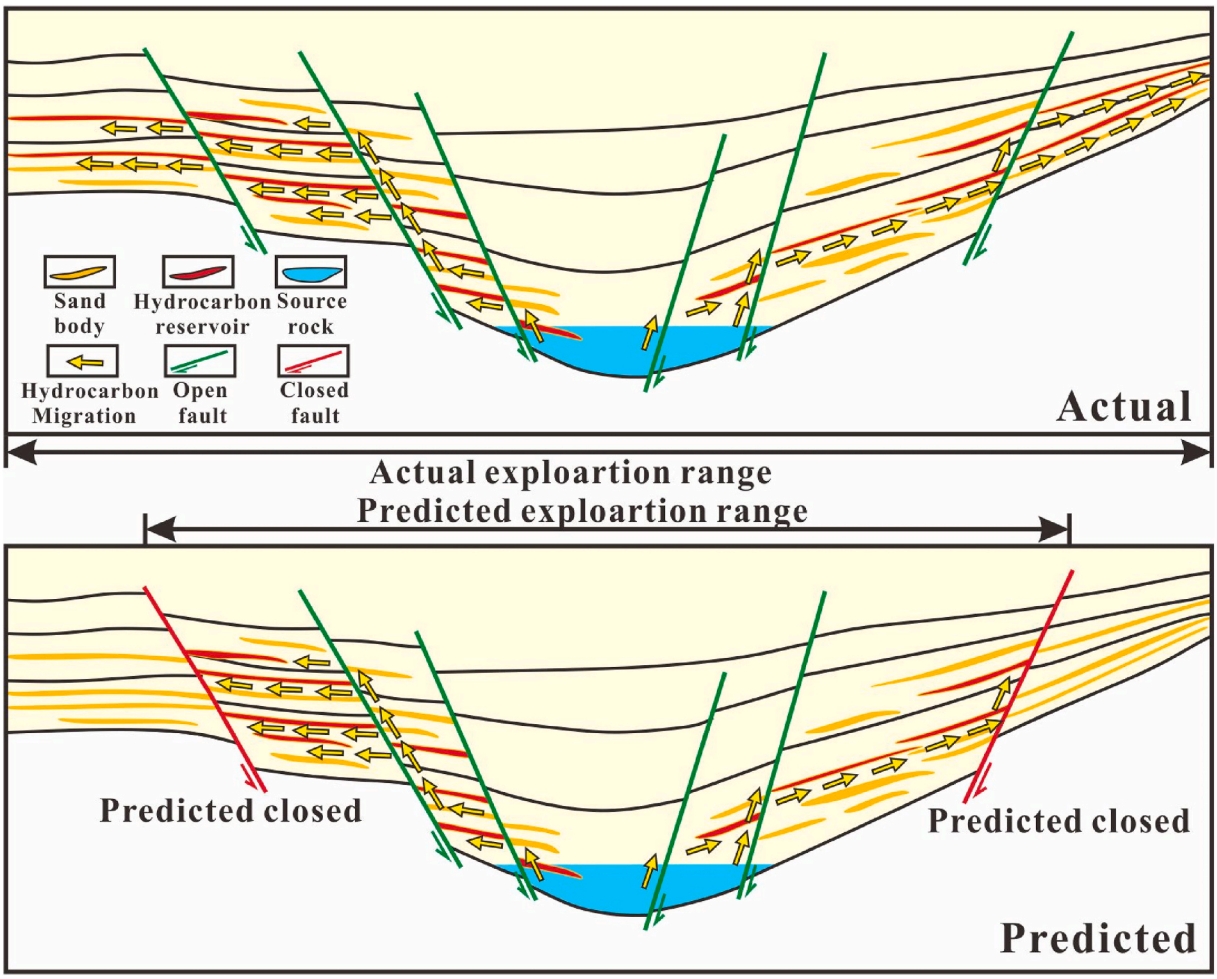


Fig. 23. Influence of evaluation metric of sensitivity on petroleum exploration. The lower profile shows the low sensitivity may lead to underestimation of exploration range and potential of the petroliferous district.

5.2.2. Geological rationality of the controlling factor analysis by DT

The prediction model not only predicts fault seals but also dominates the controlling factors of fault seals in a certain study area by calculating the information gain by the Gini index. A higher Gini index indicates a more dominant controlling factor of the fault seal capacity. Note that the value of the Gini index provides the relative importance of the controlling factors instead of the absolute importance. According to the calculation results, the carbonate cement content is the most important controlling factor on fault seals in the Huimin depression. In addition, the development of microfractures and the shale-smear effect also have some influence on faut seals. The fault throw and fault length show limited controlling effects on fault seals, while the controlling effects of fault-dip angle and fault length on fault seals are not obvious in the Huimin Depression in this study (Fig. 19).

5.2.2.1. Carbonate cements. Observation of core samples and the analysis of thin sections also showed the significant influence of carbonate cements on fault seals. Previous studies have shown that volcanism in the Huimin Depression was active from the Mesozoic Era to the Quaternary Period (Song et al., 2007; Sun et al., 2008). Hydrothermal fluids with higher temperatures deep in the Earth could have migrated to shallower depths along the faults and provide abundant carbon dioxide (CO_2). CO_2 first dissolved in water and then generated H_2CO_3, which could then become HCO_3^- and H^+:

$$CO_2(g)+H_2O=H_2CO_3 \quad (11)$$

$$H_2CO_3=H^++HCO_3^- \quad (12)$$

The Ca^{2+}, Mg^{2+} and Fe^{2+} in pore fluids could interact with HCO_3^- and generate calcite and dolomite cements along fault zones:

$$Ca^{2+}+2HCO_3^- = Ca(HCO_3)_2 \quad (13)$$

$$Ca(HCO_3)_2 = CaCO_3\ (\text{calcite})+HCO_3^- + H^+ \quad (14)$$

$$2H_2CO_3+Ca^{2+}+Mg^{2+}\rightarrow CaMg(CO_3)_2\ (\text{dolomite})+4H^+ \quad (15)$$

The calcite and dolomite cements filled in pore space and blocked pore throats, which could have largely decreased the permeability of fault zones. Furthermore, the sealing capacity of fault zones could have been strengthened. Previous studies have shown that the average contents of dolomite and calcite cements near fault zones were 19.7 and 30.5%, respectively. However, these cements far from the faults were only 5.3% and 4.4%, respectively. Furthermore, the permeability of the rocks near and far from fault zones also showed large disparities. The permeability of rocks near faults ranged from approximately 0.1 to 5.5 mD with an average of 3.7 mD. However, the permeability of rocks far from the faults ranged from 7.7 to 97.4 mD with an average of 68.4 mD (Fig. 20). Therefore, the development of carbonate cements could have largely enhanced the sealing capacity of the faults and could have blocked oil and gas from migrating through the faults.

The history of fault movement and diagenesis showed the faults in the Huimin Depression were active from approximate 40 to 25 Ma (Wang et al., 2019), during which period the hydrothermal fluids with large amount of CO_2 migrated to shallower depth along the faults. Then,

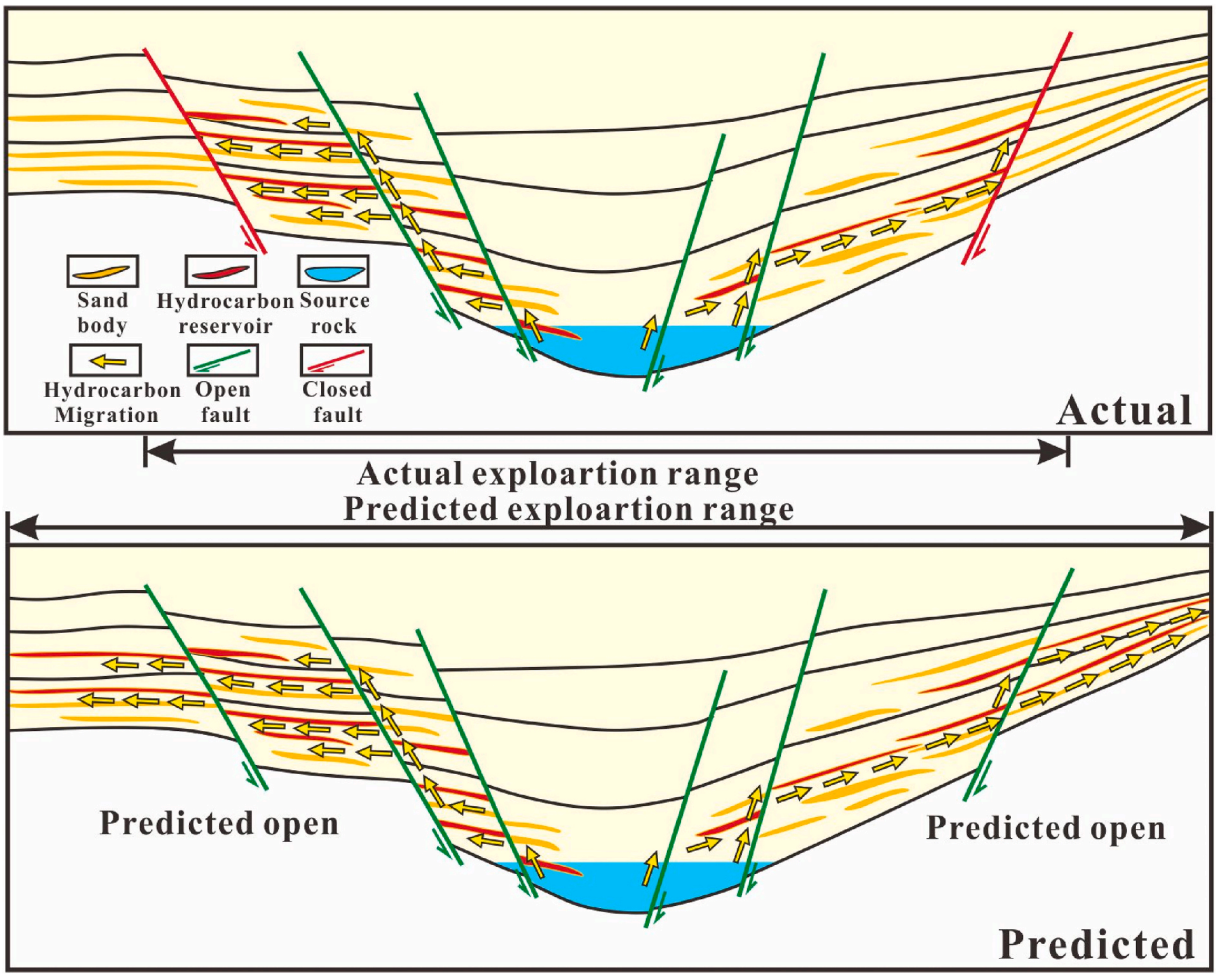


Fig. 24. Influence of evaluation metric of precision on petroleum exploration. The lower profile shows the low precision may lead to overestimation of exploration range and potential of the petroliferous district.

the carbonate cementation mainly developed from 25 to 15 Ma (Shi et al., 2017). During this period, the porosity of the sandstone reservoirs decreased from approximate 25%–17%, which led to the enhancement of fault seal. Finally, oil and gas migrated from source rocks to hydrocarbon traps and were blocked by faults with well sealing capacity from approximate 10 to 2 Ma (Fig. 21).

5.2.2.2. Microfracture. The generation of microfractures near the faults was the second most important controlling factor on fault seals in the Huimin Depression. Previous studies have shown that friction along fault planes due to the movement of faults can lead to widely distributed microfractures in the fault zone (Kalani et al., 2015). Microfractures can largely change the physical properties of faults, especially in increasing the permeability of faults (Hodson et al., 2016; Teixeira et al., 2017). In this study, microfractures were observed to be more developed near faults filled with asphalt, which was evidence of hydrocarbon migration (Fig. 11). The results indicated that the microfractures could have enhanced the migrating capacity of the faults.

The evolutionary processes of fault movement and diagenesis also showed that the microfractures mainly developed from approximate 40 Ma to 25 Ma as the main faults of the Huimin Depression moved frequently (Wang et al., 2019). The diagenetic history indicated that the microfractures mainly developed before the formation of carbonate cements, therefore, the former generated hydrocarbon migrating paths by microfractures could be filled by carbonate cements and hence became barriers for hydrocarbon migration (Fig. 21). The result of Gini index also showed similar message: Carbonate cementation and development of microfractues are two main controlling factors on fault seal, while carbonate cementation is the dominant one.

5.2.2.3. Fault-dip angle. The statistical data showed that the distribution of the fault-dip angle of the faults in the Huimin Depression showed a concentrated distribution ranging from 40 to 50° (Fig. 8). Additionally, measurements of the fault-dip angle in geological analyses are generally in a range of approximately 5–10°. Therefore, the measurement of the fault-dip angle probably leads to a large error in the construction of CART and hence hides the real relationship between the fault seal and the fault-dip angle. Improving the accuracy of fault-dip angle measurements by using high-resolution seismic technology would be helpful for dealing with this problem.

Essentially, as shown above, the key to the successful employment of artificial intelligence or machine learning in petroleum geology is the quality of the dataset. First, an ideal model based on machine learning requires at least hundreds or thousands of data and even millions of data, which is sometimes difficult to satisfy in a petroliferous region. Second, the accuracy requirements of computer science and petroleum geology are often different. The requirement of high-precision data in machine learning will not often be provided in geological research (Fig. 22). In this case, some new technology should be used for providing precise geological data, such as image software in microscopy, micro- or nano-CT Scans, imaging logging, etc. Combined with multiple technologies or methods, machine learning will have broad application prospects in geological research.

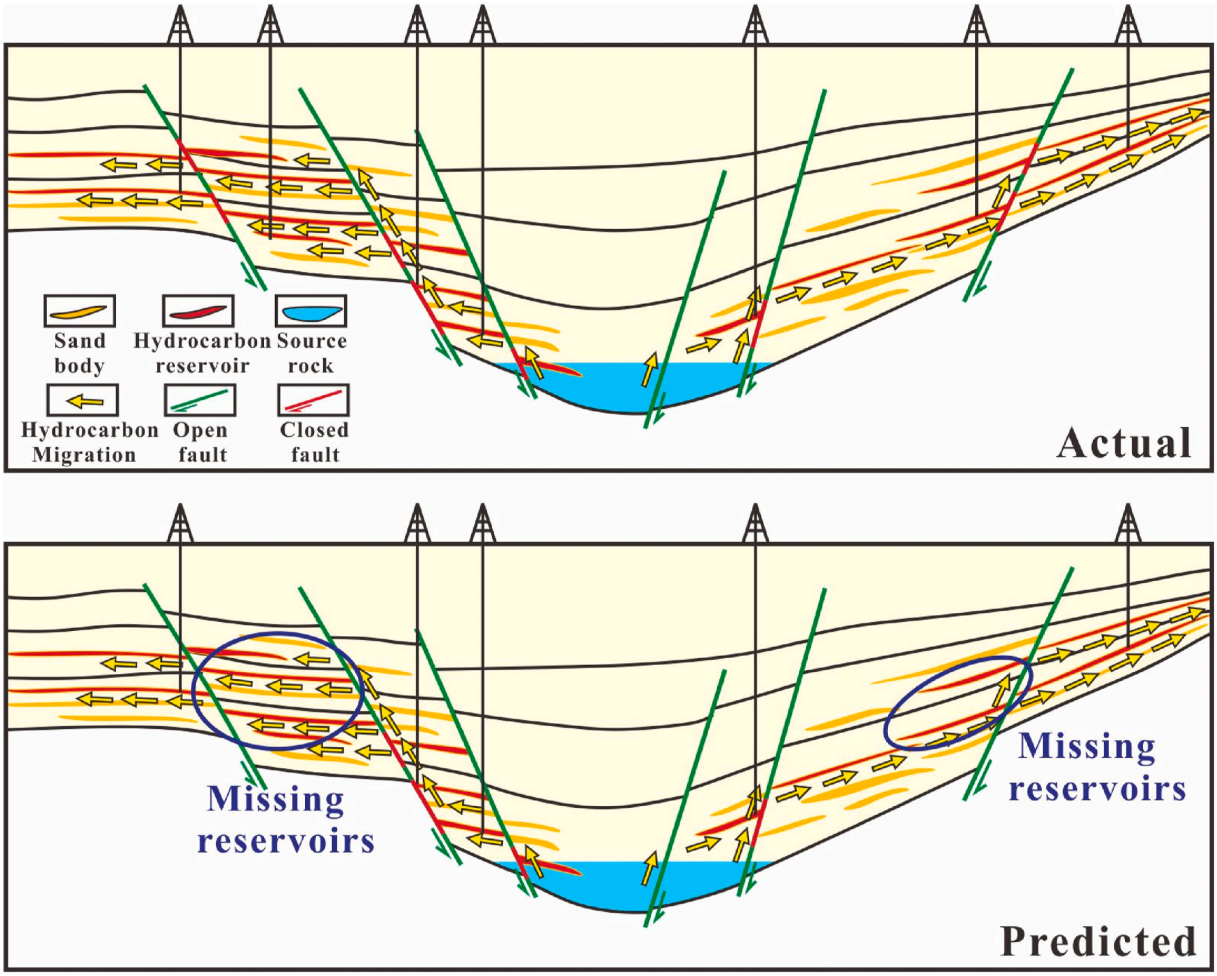


Fig. 25. Influence of evaluation metric of specificity on petroleum exploration. The lower profile shows the low specificity may lead to missing of reservoirs.

5.3. *Evaluation metrics and predicting model selection*

The evaluation metrics including sensitivity, precision, specificity, negative predicted value and F1-score could reflect the reliability and stability of a classification model. While, different evaluation metrics represent different situations of fault seal, hydrocarbon migration and accumulation and have different meanings for petroleum exploration and exploitation.

5.3.1. *Sensitivity*

In this research, the sensitivity of the evaluation model is the ratio of the open faults correctly identified by the model to the actual open faults among the whole dataset. Therefore, the sensitivity is important for the decision of exploration range and evaluation of the exploration potential of a petroliferous district. If the model shows low sensitivity, the migrating capability of the fault systems might be underestimated. Therefore, the exploration range might be restricted and hence the exploration potential might be underestimated (Fig. 23).

5.3.2. *Precision*

The precision of the evaluation model is the ratio of the open faults correctly identified by the model to all the open faults identified by the model. The precision directly reflects the ability of the model to predict the open state of the faults. If the model shows low precision, the migrating capability of the fault systems might be overestimated. Therefore, the exploration range might be larger than the actual one. On the one hand, part of hydrocarbon reservoirs might be missed; On the other hand, the exploration range might be wrongly expanded, which might lead to the waste of time and wrong guidance (Fig. 24).

5.3.3. *Specificity*

The specificity of the evaluation model is the ratio of the closed faults correctly identified by the model to the actual closed faults among the whole dataset. The specificity can reflect the ability of the model to evaluate the sealing capacity of faults. The low specificity of the faults might lead to underestimation of the sealing capacity of faults and hence the missing of the fault block reservoirs (Fig. 25).

5.3.4. *Negative predicted value*

The negative predicted value of the evaluation model is the ratio of the closed faults correctly identified by the model to all the closed faults identified by the model. The negative predicted value is of great significance for the well location design and exploitation strategy design. If the model shows low negative predicted value, the sealing capacity of faults might be overestimated and directly lead to mistake of well deployment (Fig. 26).

In conclusion, the evaluation metrics showed different importance and priority in different stage of petroleum exploration and exploitation. If the primary goal of exploration and exploitation is to find as many as resource potential and hydrocarbon reservoirs, the sensitivity and specificity might be prior evaluation metrics. However, if the primary goal of exploration and exploitation is to reduce exploration risk and improve economic efficiency, the precision and negative predicted value should be firstly considered. While, the selection of prior evaluation metrics and final evaluation models depend on the actual situations and dominate problems we face.

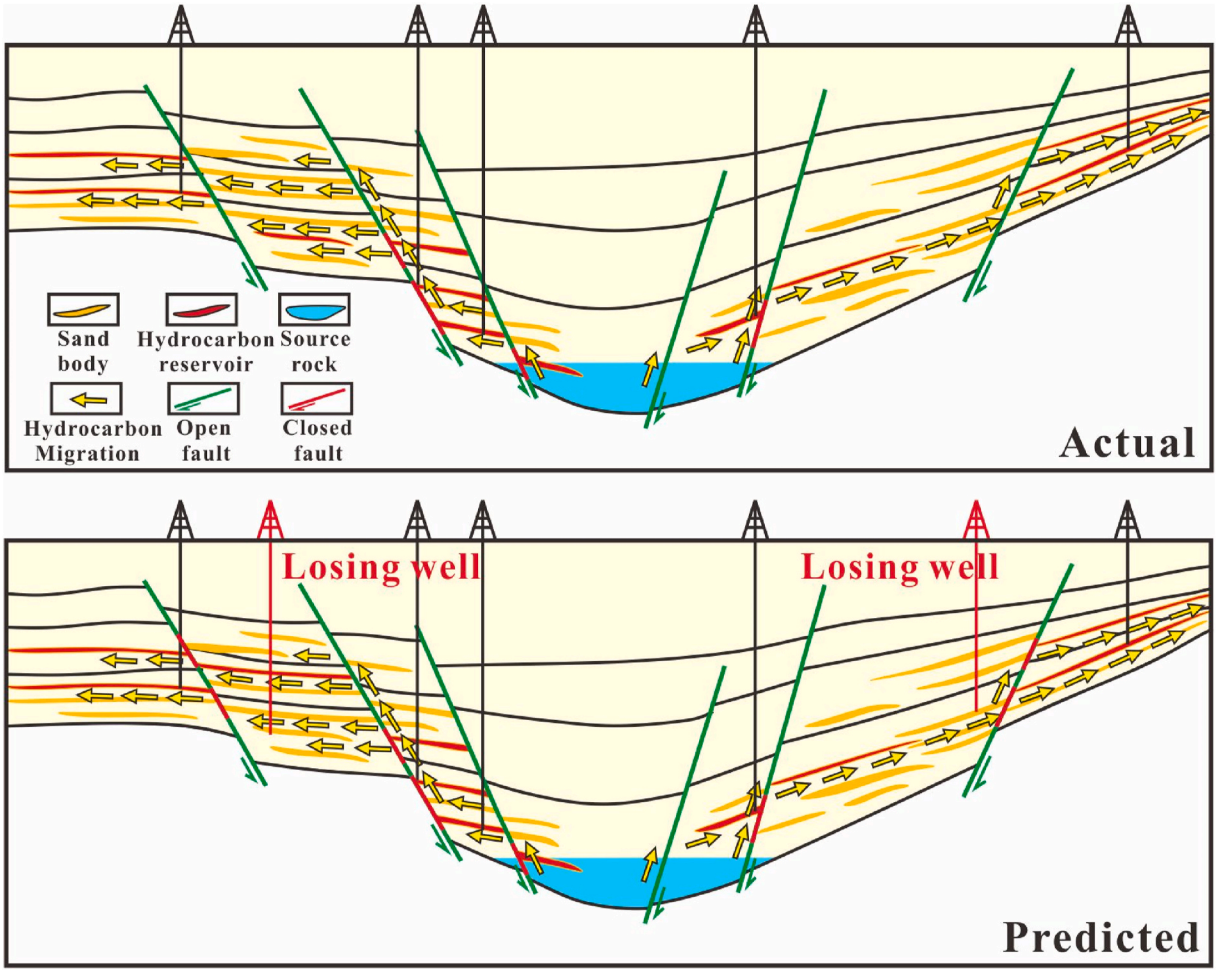


Fig. 26. Influence of evaluation metric of negative predicted value on petroleum exploration. The lower profile shows the low negative predicted value may lead to losing of wells.

6. Conclusion

Fault seals are influenced by multiple factors in a complex way. Formal fault seal evaluating methods, which focused on a single fault seal controlling factor show relative low accuracy. Therefore, the decision tree and random forest algorithms, which are characterized by nonliner regression analysis of multiple factors, are appropriate for evaluating fault seals. In this research, the evaluation model based on the CART and RF were used to evaluate fault seals in the Huimin Depression, Bohai Bay Basin. The results showed that the accuracy of the evaluation CART and RF models were 90% and 95%, respectively. The testing accuracies are higher than those of methods based on a single controlling factor.

The consistency between analytical results by DT and the geological analysis showed the rationality of the fault seal evaluating model. The Gini index calculated by CART showed that the development of carbonate cements and microfractures are the dominant controlling factors on fault seals in the Huimin Depression, which is consistent with geological analysis. Analysis of the dominant controlling factors also indicates the possible reasons for the low accuracy of previous fault seal evaluation methods, which only consider the shale-smear effect, fault-dip angle, fault throw or normal stress of the fault plane.

DT and RF algorithms showed advantages in different aspects. As one of the ensemble learning methods, RF usually has higher accuracy and reliability than the DT when the amount of data is large. However, the algorithm of DT has short training time and high efficiency. Furthermore, DT is easy to understand and shown in figures or work flows. The high interpretability and visualization are necessary in some fields or situations.

Errors in the evaluation and prediction by machine learning might be attributed to quality of data and calibration processes. Higher accuracy of the original dataset and a proper strategy for pruning of the decision tree can improve the reliability for a machine learning model. The testing accuracy usually needs evaluation metrics for further evaluating the reliability of the model. Different evaluation metrics including sensitivity, precision, specificity, negative predicted value and F1-score have different priorities according to the goals of petroleum exploration and exploitation. Furthermore, addressing the disparity in the amount and precision of the data between geological research and computer science may be the key to the employment of machine learning in petroleum geology.

Credit author statement

Qiaochu Wang: Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Dongxia Chen: Supervision, Conceptualization, Methodology, Meijun Li: Methodology, Software, Fuwei Wang: Data curation, Writing – review & editing, Yu Wang: Formal analysis, Wenlei Du: Data curation, Validation, Xuebin Shi: Writing – review & editing.

Declaration of competing interest

No conflict of interest exits in the submission of this manuscript, and

manuscript is approved by all authors for publication. I would like to declare on behalf of my co-authors that the work described was original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part. All the authors listed have approved the manuscript that is enclosed.

Data availability

Data will be made available on request.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (Grant No. 41972124). We also acknowledge the Shengli Oilfield Branch of the China Petroleum and Chemical Corporation for providing testing data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoen.2023.212064.

References

AbouEisha, H., Chikalov, I., Moshkov, M., 2016. Decision trees with minimum average depth for sorting eight elements. Discrete Appl. Math. 204, 203–207.
Akkas, E., Akin, L., Cubukcu, H.A., 2015. Application of decision tree algorithm for classification of natural Minerals using SEM-EDS. Comput. Geosci. 80, 38–48.
Alan, P.M., Kevin, J.S., David, A.F., Nathaniel, E.R., Peter, F.C., 2012. Production-induced fault compartmentalization at Elk Hills field, California. AAPG Bulletin 96 (6), 1001–1015.
Alvar, B., Jan, T., Haakon, F., Tore, S., Nestor, C., Semshaug, E.B., Einar, S., 2009. Fault facies and its application to sandstone reservoirs. AAPG Bulletin 93 (7), 891–917.
Asim, K.M., Awais, M., Martínez–Álvarez, F., et al., 2017. al . Seismic activity prediction using computational intelligence techniques in northern Pakistan. Acta Geophys. 65 (5), 919–930.
Baudon, C., Cartwright, J., 2008. The kinematics of reactivation of normal faults using high resolution throw mapping. J. Struct. Geol. 30 (8), 1072–1084.
Bergosh, J.L., Lord, G.D., 1987. New Developments in the Analysis of Cores from Naturally Fractured Reservoirs. SPE Annual Technical Conference and Exhibition, Dallas, Texas, US, pp. 27–30.
Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.
Botu, V., Ramprasad, R., 2015. Adaptive machine learning framework to accelerate abinitio molecular dynamics. Int. J. Quant. Chem. 115, 1074–1083.
Bravo, C.E., Saputelli, L., Rivas, F., et al., 2014. State of the art of artificial intelligence and predictive analytics in the E&P industry: a technology survey. SPE J. 19 (4), 547–563.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Breiman, L., Friedman, J.H., Olshen, J.A., Stone, C.J., 2007. Linking clinical measurements and kinematic gait patterns of toe-walking using fuzzy decision trees. Gait Posture (25), 475–484.
Bruna, V.L., Bezerra, F.H.R., Souza, V.H.P., Maia, R.P., Auler, A.S., Araujo, R.E.B., Cazarin, C.I., Rodrigues, M.A.F., Vieira, L.C., Sousa, M.O.L., 2021. High-permeability zones in folded and faulted silicified carbonate rocks – implications for karstified carbonate reservoirs. Mar. Petrol. Geol. 128 https://doi.org/10.1016/j.marpetgeo.2021.105046.
Burman, 1989. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. Biometrika 76, 503.
Caillet, G., Batiot, S., 2003. 2D modeling of hydrocarbon migration along and across growth faults. an example from Nigeria. Petrol. Geosci. 9, 113–124.
Caro, D.M.R., Batezelli, A., Leite, E.P., 2023. Fault reactivation potential in a carbonate field in Brazil based on geomechanical analysis. Mar. Petrol. Geol. 150, 106131.
Cartwright, J., Huuse, M., Aplin, A., 2007. Seal bypass systems. AAPG (Am. Assoc. Pet. Geol.) Bull. 91 (8), 1141–1166.
Cipr, U., Bergen, N., Earth, A.S., 2017. Fault visualization and identification in fault seismic attribute volumes: implications for fault geometric characterization, 1 Interpretation 5 (2), 16–17.
Chan, P.K.K., Zheng, J., Liu, H., Tsang, E.C.C., Yeung, D.S., 2021. Robustness analysis of classical and fuzzy decision trees under adversarial evasion attack. Applied Soft Computing Journal 107. https://doi.org/10.1016/j.asoc.2021.107311.
Chandra, B., Kothari, R., Paul, P., 2010. A new node splitting measure for decision tree construction. Pattern Recogn. 43 (8), 2725–2731.
Chen, W., Wu, Z.P., Hou, F., 2010. Relationship between hydrocarbon accumulation and Linshang fault zone in Linnan area, Huimin Depression. Petroleum Geology and Recovery Efficiency 17 (2), 25–28 (In Chinese).
Choi, J.H., Edwards, P., Ko, K., Kim, Y.S., 2016. Definition and classification of fault damage zones: a review and a new methodological approach. Earth Sci. Rev. 152, 70–87.
Cowie, P.A., Scholz, C.H., 1992b. Displacement-length scaling relationship for faults: data synthesis and discussion. J. Struct. Geol. 14 (10), 1149–1156.
Dai, D.B., Xu, T., Wei, X., Ding, G.T, Xu, Y., Zhang, J.C., Zhang, H.R. Using machine learning and feature engineering to characterize limited material datasets of high-entropy alloys. Comput. Mater. Sci. 2020 175 (109618). https://doi.org/10.1016/j.commatsci.2020.109618.
Denisko, D., Hoffman, M.M., 2018. Classification and interaction in random forests. Proc. Natl. Acad. Sci. U.S.A. 115 (8), 1690–1692.
Dewey, J.F., Holdsworth, R.E., Strachan, R.A., 1998. Transpression and transtension zones. Geological Society of London Special Publications 135 (1), 1–14.
Dong, Y.T., Ju, B.S., Yang, Y., Wang, J., Ma, S., Brantson, E.T., 2021. A semi- analytical method for optimizing the gas and water bidirectional displacement in the tilted fault block reservoirs. J. Petrol. Sci. Eng. 198 https://doi.org/10.1016/j.petrol.2020.108213.
Dong, Y.Y., Zeng, J.H., Dong, X.Y., Li, C.M., Liu, Y.Z., 2022. The control effect of normal faults and caprocks on hydrocarbon accumulation: a case study from the Binhai fault nose of the Huanghua Depression, Bohai Bay Basin, China. J. Petrol. Sci. Eng. 218, 110918.
Du, Y.M., 2005. Effect of Xiakou fault on field distribution and petroleum migration in Linnan Slope area. Xinjing Pet. Geol. 26 (5), 525–528 (In Chinese).
Eichhubl, P., Boles, J.R., 2000. Rates of fluid flow in fault systems-evidence for episodic rapid fluid flow in the Miocene Monterey Formation, coastal California. Am. J. Sci. 300, 571–600.
Færseth, R.B., Johnsen, E., Sperrevik, S., 2007. Methodology for risking fault seal capacity: implications of fault zone architecture. AAPG Bulletin 91, 1231–1246.
Feng, D.X., Tian, M.R., Zhang, H.C., Cao, A.F., 2010. Research on the characters of Tenseshearing geological structure and hydrocarbon accumulation in western of Huimin Depression. Shanghai Geol. 31 (1), 217–221 (In Chinese).
Fratello, M., Tagliaferri, R., 2019. Decision trees and random forests. Encyclopedia Bioinf. Comput. Biol. 1, 374–383.
Frery, E., Gratier, J.P., Zimmerman, N.E., Loiselet, C., Braun, J., Deschamps, P., Blamart, D., Hamelin, B., Swennen, R., 2015. Evolution of fault permeability during episodic fluid circulation: evidence for the effects of fluid–rock interactions from travertine studies (Utah–USA). Tectonophysics 651–652, 121–137.
Friedl, M.A., Brodley, C.E., 1997. Decision tree classification of land cover from remotely sensed data. Remote Sens. Environ. 61, 399–409.
Fu, G., Liu, H.X., Duan, H.F., 2005. Seal mechanism of different transporting passways of fault and their research methods. Petroleum Geology & Experiment 27 (4), 404–408 (In Chinese).
Fu, X.F., Jia, R., Wang, H.X., Wu, T., Meng, L.D., Sun, Y.H., 2015. Quantitative evaluation of fault-caprock sealing capacity: a case from Dabei-Kelasu structural belt in Kuqa Depression, Tarim Basin, NW China. Petrol. Explor. Dev. 42 (3), 329–338.
Fulljames, J.R., Zijerveld, L.J.J., Franssen, R.C.M.W., 1997. Fault seal processes: systematic analysis of fault seals over geological and production time scales. In: Møller-Pedersen, P., Koestler, A.G. (Eds.), Hydrocarbon Seals: Importance for Exploration and Production. Norwegian Petroleum Society, 7. Special Publication, Trondheim, Norway, pp. 51–59.
Gao, X.Z., Du, Y.M., Zhang, B.S., 2003. The sealing of Xiakou Fault and its model of controlling on the petroleum accumulation. Petrol. Explor. Dev. 30 (3), 76–78 (In Chinese).
Geisser, S., Eddy, W.F., 1979. A predictive approach to model selection. J. Am. Stat. Assoc. 74, 153–160.
Guo, X.W., Wu, Z.P., Yang, X.Q., Xu, H.H., Zhang, Z.X., Shi, X.B., Sun, Z., 2009. The evolution of transtensional structure and numerical modeling of stress field, Linnan subsag , Bohai Bay Basin. Mar. Geol. Quat. Geol. 29 (6), 79–86 (In Chinese).
Hao, F., Zou, H.Y., Gong, Z.S., Deng, Y.H., 2007. Petroleum migration and accumulation in the Bozhong sub-basin, Bohai Bay Basin, China: significance of preferential petroleum migration pathways (PPMP) for the formation of large oilfields in lacustrine fault basins. Mar. Petrol. Geol. 24 (1), 1–13.
Hao, F., Zou, H.Y., Gong, Z.S., 2010. Preferential petroleum migration pathways and prediction of petroleum occurrence in sedimentary basins: a review. Petrol. Sci. 7 (1), 2–9.
Harper, T.R., Lundin, E.R., 1997. Fault seal analysis: reducing our dependence on empiricism. In: Møller-Pedersen, P., Koestler, A.G. (Eds.), Hydrocarbon Seals: Importance for Exploration and Production. Norwegian Petroleum Society7, Special Publication, Trondheim, Norway, pp. 149–165.
Hastie, T., Tibshirani, R., Friedman, J., 2008. The Elements of Statistical Learning, second ed. Springer. 0-387-95284-5.
Hindle, A.D., 1997. Petroleum migration pathways and charge concentration: a three dimensional model. AAPG (Am. Assoc. Pet. Geol.) Bull. 81, 1451–1481.
Hodson, K.R., Crider, J.G., Huntington, K.W., 2016. Temperature and composition of carbonate cements record early structural control on cementation in a nascent deformation band fault zone: Moab Fault. Tectonophysics 690, 240–252. Utah, USA. 2016.
Hou, Q., Zhao, J., Hui, C., Nie, M.L., 2006. Simulation analysis of the tectonic stress field and the oil-gas migration-accumulation in Huimin Sag. 01. Petroleum Geology and Recovery Efficiency 13, 66–69 (In Chinese).
Huang, H.K., Wang, H.Z., Sun, M., 2020. Incomplete data classification with view-based decision tree. Appl. Soft Comput. 94 https://doi.org/10.1016/j.asoc.2020.106437.
Hull, J., 1988. Thickness-displacement relationships for deformation zones. J. Struct. Geol. 10, 431–435.
Huo, Y.C., Bouffard, F., Joós, G., 2021. Decision tree-based optimization for flexibility management for sustainable energy microgrids, 2021 Appl. Energy 290. https://doi.org/10.1016/j.apenergy.2021.116772.

Itani, S., Lecron, F., Fortemps, P., 2020. A one-class classification decision tree based on kernel density estimation. Appl. Soft Comput. 91 https://doi.org/10.1016/j.asoc.2020.106250.
Javier, H.T., Miguel, C.B., 2018. Using smart persistence and random forests to predict photovoltaic energy production. Energies 12 (1), 1–12.
Jegadeeshwaran, R., Sugumaran, V., 2013. Comparative study of decision tree classifier and best first tree classifier for fault diagnosis of automobile hydraulic brake system using statistical features. Measurement 46 (9), 3247–3260.
Jobe, J.T., Briggs, R., Gold, R., DeLong, S., Hille, M., Delano, J., Johnstone, S.A., Pickering, A., Phillips, R., Calvert, A.T., 2022. The Pondosa fault zone: a distributed dextral-normal-oblique fault system in northeastern California, USA. Geosphere 19 (1), 179–205.
Kalani, M., Jahren, J., Mondol, N.H., Faleide, J.I., 2015. Petrophysical implications of source rock microfracturing, 2015 Int. J. Coal Geol. 143, 43–67.
Kam, T., 1995. Random decision forests. In: Proceedings of the 3rd International Conference on Document Analysis and Recognition, pp. 278–282. Montreal, QC.
Kaminski, B., Jakubczyk, M., Szufel, P., 2017. A framework for sensitivity analysis of decision trees. Cent. Eur. J. Oper. Res. 26 (1), 135–159.
Kang, K., Qin, C.Y., Lee, B., Lee, I., 2019. Modified screening-based Kriging method with cross validation and application to engineering design. Appl. Math. Model. 70, 626–642.
Karlsen, D.A., Skeie, J.E., 2006. Petroleum migration, faults and overpressure: calibrating basin modeling using petroleum in traps-a review. J. Petrol. Geol. 29 (3), 227–256.
Kelter, R., 2021. Bayesian model selection in the M-open setting—approximate posterior inference and subsampling for efficient large-scale leave-one-out cross-validation via the difference estimator. J. Math. Psychol. 100 (102474).
Kim, K., 2016. A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree. Pattern Recogn. 60, 157–163.
Knipe, R.J., 1997. Juxtaposition and seal diagrams to help analyze fault seals in hydrocarbon reservoirs. AAPG (Am. Assoc. Pet. Geol.) Bull. 81 (2), 187–195. Alan, 1999.
Knipe, R.J., Fisher, Q.J., Jones, G., Clennell, M.R., Farmer, A.B., Harrison, A., Kidd, B., McAllister, E., Porter, J.R., White, E.A., 1997. Fault seal analysis: successful methodologies, application and future directions. In: Møller-Pedersen, P., Koestler, A.G. (Eds.), Hydrocarbon Seals: Importance for Exploration and Production, Norwegian Petroleum Society (NPF), vol. 7. Special Publication, Trondheim, Norway, pp. 15–40.
Knott, S.D., 1993. Fault seal analysis in the north sea. AAPG Bulletin 77, 778–792.
Kreimeyer, K., Dang, O., Spiker, J., Munoz, A.M., Rosner, G., Ball, R., Botsis, T., 2021. Feature engineering and machine learning for causality assessment in pharmacovigilance: lessons learned from application to the FDA Adverse Event Reporting System. Comput. Biol. Med. 135, 104517.
Lao, H.G., Shan, Y.X., Wang, Y.S., Wu, Z.H., 2020. Characteristics of growth fault architecture and its evolution in mudstone strata: evidence from the core of Bohai bay basin. Mar. Petrol. Geol. 119, 104503.
Lao, H.G., Wang, Y.S., Li, J.Y., Shan, Y.X., Song, L.L., 2022. Normal fault transmissibility characteristics under the transition condition of fault conduction and sealing observed in simulation experiments. Mar. Petrol. Geol. 143, 105779.
Lei, Y.H., Luo, X.R., Zhang, L.K., Song, C.P., Cheng, M., 2013. Quantitative characterization of Shahejie Formation sandstone carrier connectivity of the eastern part of the slope in Dongying sag. Acta Petrol. Sin. 34 (4), 692–700 (In Chinese).
Leibovici, D.G., Bastin, L., Jackson, M., 2011. Higher-order co-occurrences for exploratory point pattern analysis and decision tree clustering on spatial data. Comput. Geosci. 37 (3), 382–389.
Li, X., Chan, C.W., Nguyen, H.H., 2013. Application of the Neural Decision Tree approach for prediction of petroleum production. J. Petrol. Sci. Eng. 104, 11–16.
Li, W.W., Yang, C.W., Sun, D.L., 2009. Mining geophysical parameters through decision-tree analysis to determine correlation with tropical cyclone development. Comput. Geosci. 35 (2), 309–316.
Liu, N., Qiu, N.S., Chang, J., Shen, F.Y., Wu, H., Lu, X.S., Wang, Y.J., Jiao, Y.X., Feng, Q. Q., 2017. Hydrocarbon migration and accumulation of the suqiao buried-hill zone in wen'an slope, Jizhong subbasin, Bohai Bay Basin, China. 2017. Mar. Petrol. Geol. 86, 512–525.
Liu, Q.G., Lu, H., Li, L.X., Mu, A.T., 2018. Study on characteristics of well-test type curves for composite reservoir with sealing faults. Petroleum 4 (3), 309–317.
Lindsay, N.G., Murphy, F.C., Walsh, J.J., Watterson, J., 1993. Outcrop Studies of Shale Smears on Fault Surfaces, vol. 15. International Association of Sedimentologists Special Publication, pp. 113–123.
Linjordet, A., Skarpnes, O., 1992. Application of horizontal stress directions interpreted from borehole breakouts recorded by four arm dipmeter tools. In: Vorren, T.O. (Ed.), Arctic Geology and Petroleum Potential. Norwegian Petroleum Society 2, Special Publication, pp. 681–690.
Lu, Y.F., Li, G.H., Wang, Y.W., Song, G.J., 1996. Quantinative analyses in fault sealing properties. Acta Pet. Sin. 17 (3), 39–44 (In Chinese).
Lu, Y.F., Wang, S., 2010. Quantitative evaluation of fault seal. J. Daqing Pet. Inst. 34 (5), 38–44 (In Chinese).
Luo, X.R., Lei, Y.H., Zhang, L.K., 2012. Characterization of carrier formation for hydrocarbon migration: comcepts and approaches. Acta Petrol. Sin. 33 (3), 428–436 (In Chinese).
Lyu, X.Y., Yun, L., Xu, J.G., Liu, H., Yu, X.A., Peng, P., Ouyang, M.K., Luo, Y., 2023. Sealing capacity evolution of gypsum salt caprocks under multi-cycle alternating stress during operations of underground gas storage. J. Petrol. Sci. Eng. 220, 111244. Part A.
Ma, P.J., Lin, C.Y., Dong, C.M., Ren, L.H., Jahren, J., Hellevang, H., Lin, J.L., 2023. Effect of faulting on diagenetic processes in the silicate-sulfate-carbonate system: a case study from the Bonan sag of Jiyang depression, Bohai Bay Basin. Mar. Petrol. Geol. 147, 105985.
Ma, S.J., Zeng, L.B., Tian, H., Shi, X.W., Wu, W., Yang, S.H., Luo, L., Xu, X., 2023. Fault damage zone and its effect on deep shale gas: insights from 3D seismic interpretation in the southern Sichuan Basin, China. J. Struct. Geol. 170, 104848.
Martens, J., 2010. Deep learning via hessian-free optimization. ICML 27, 735–742.
Morris, A.P., Ferrill, D.A., McGinnis, R.N., 2016. Using fault displacement and slip tendency to estimate stress states. J. Struct. Geol. 83, 60–72.
Mitchell, T.M., 1997. Machine Learning. McGraw-Hill, New York.
Michie, E.A.H., Kaminskaite, I., Cooke, A.P., Fisher, Q.J., Yielding, G., Tobiss, S.D., 2021. Along-strike permeability variation in carbonate-hosted fault zones. J. Struct. Geol. 142 https://doi.org/10.1016/j.jsg.2020.104236.
Mnih, V., Badia, A.P., Mirza, M., et al., 2016. Asynchronous Methods for Deep Reinforcement Learning. International conference on machine learning, pp. 1928–1937.
Nikolaev, N.I., Slavov, V., 1998. Inductive genetic programming with decision trees. Intell. Data Anal. 2 (1–4), 31–44.
Nock, R., Jappy, P., 1999. Decision tree based induction of decision lists. Intell. Data Anal. 3 (3), 227–240.
Panahi, H., Kobchenko, M., Meakin, P., Dysthe, D.K., Renard, F., 2019. Fluid expulsion and microfracturing during the pyrolysis of organic rich shale, 2019 Fuel 235, 1–16.
Pei, Y.W., Paton, D.A., Knipe, R.J., Wu, K.Y., 2015. A review of fault sealing behaviour and its evaluation in siliciclastic rocks. Earth Sci. Rev. 150, 121–138.
Piryonesi, S.M., Ei-Diraby, T.E., 2021. Role of data analytics in infrastructure asset management: overcoming data size and quality problems. J. Transport. Eng. 146 (2) https://doi.org/10.1061/JPEODX.0000175.
Pradhan, B., 2013. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Comput. Geosci. 51, 350–365.
Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, California.
Quinlan, J.R., 2003. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA, p. 302.
Rakers, C., Reker, D., Brown, J.B., 2017. Small random forest models for effective chemogenomic active learning. J. Comput. Aided Chem. 18, 124–142.
Saettler, A., Laber, E., Pereira, F., 2017. Decision tree classification with bounded number of errors. Inf. Process. Lett. 127, 27–31.
Schultz, C.D., Hofmann, M.H., 2021. Facies, stratigraphic architecture, and faults - the controls on the cement distribution in the Devonian Sappington Formation in southwestern Montana. Mar. Petrol. Geol. 124 https://doi.org/10.1016/j.marpetgeo.2020.104806.
Smit, F.W.H., Stemmerik, M.E., Smith, P.T., Staudigel, M.L.M., Welch, F.S.P., Buchem, P. K.S., 2023. The importance of fault damage zones for fluid flow in low-permeable carbonate rocks – fault-related compaction fronts in the Danish North Sea. Mar. Petrol. Geol. 148, 105993.
Smith, P.F., Ganesh, S., Liu, P., 2013. A comparison of random forest regression and multiple linear regression for prediction in neuroscience. J. Neurosci. Methods 220 (1), 85–91.
Shi, S.G., Yang, T., Cao, Y.C., Zhang, H.N., Yuan, G.H., 2017. Diagenesis and pore evolution of turbidite reservoir in the linnan depression. Special Oil Gas Reservoirs 24 (2), 57–62 (in Chinese).
Song, Z.D., Zha, M., Zhao, W.W., Zhou, Y.C., 2007. Characteristics of igneous rocks and their effects on hydrocarbon accumulation in Ynagxin subsag of Huimin Sag. Journal of China University of Petroleum 31 (2), 1–8.
Sorkhabi, R., Tsuji, Y., 2005. The place of faults in petroleum traps. In: Sorkhabi, R., Tsuji, Y. (Eds.), Faults, Fluid Flow, and Petroleum Traps, vol. 85. AAPG Memoir, pp. 1–31.
Speiser, J.L., Miller, M.E., Tooze, J., Edward, I., 2019. A comparison of random forest variable selection methods for classification prediction modelling. Expert Syst. Appl. 134 (15), 93–101.
Sun, Y., Zhong, J.H., Yuan, X.C., Jiang, Z.X., Yang, W.L., Li, S.Y., 2008. Analysis on sequence stratigraphy of lacustrine carbonate in the first member of Shahejie Formation in Huimin Sag. Acta Pet. Sin. 29 (2), 213–218 (in Chinese).
Tan, X.Q., Liu, Y.Y., Zhou, X.Z., Liu, J.D., Zheng, R.C., Jia, C., 2019. Multi-parameter quantitative assessment of 3D Geological models for complex fault-block oil reservoirs. Petrol. Explor. Dev. 46 (1), 194–204.
Teixeira, M.G., Donzé, F., Renard, F., Panahi, H., Papachristos, E., Scholtès, L., 2017. Microfracturing during primary migration in shales. Tectonophysics 694, 268–279.
Tom, M., 1997. Machine Learning. McGraw Hill, New York.
Torabi, A., Alaei, B., Libak, A., 2019. Normal fault 3D geometry and displacement revisited: insights from faults in the Norwegian Barents Sea. Mar. Petrol. Geol. 99, 135–155.
Vrolijk, P.J., Janos, L.U., Kettermann, M., 2016. Clay smear: review of mechanisms and applications - sciencedirect. J. Struct. Geol. 86, 95–152.
Wainer, J., Cawley, G., 2021. Nested cross-validation when selecting classifiers is overzealous for most practical applications. Expert Syst. Appl. 182 (115222).
Wang, F.W., Chen, D.X., Wang, Q.C., Shi, X.B., Xie, G.J., Wang, Z.Y., Li, J.H., Liao, W.H., 2020. Evolution characteristics of transtensional faults and their impacts on hydrocarbon migration and accumulation: a case study from the Huimin Depression, Bohai Bay Basin, eastern China. Mar. Petrol. Geol. 120 https://doi.org/10.1016/j.marpetgeo.2020.104507.
Wang, Q.C., Chen, D.X., Wang, F.W., Li, J.H., Liao, W.H., Wang, Z.Y., Xie, G.J., Shi, X.B., 2019. Underpressure characteristics and origins in the deep strata of rift basins: a case study of the Huimin Depression, Bohai Bay Bain, China. Geol. J. https://doi.org/10.1002/gj.3651.

Yielding, G., Freeman, B., Needham, D.T., 1997. Quantitative fault seal prediction. AAPG (Am. Assoc. Pet. Geol.) Bull. 81, 897–917.
Zhang, L.K., Luo, X.R., Vasseur, G., Yu, C.H., Wang, Y., Lei, Y.H., Song, C.P., Yu, L., Yan, J.Z., 2011. Evaluation of geological factors in characterizing fault connectivity during hydrocarbon migration: application to the Bohai Bay Basin. Mar. Petrol. Geol. 28, 1634–1647.
Zhao, Y., Liu, Z., Dai, L.C., 2004. The analysis of the characteristics of subnormal pressure and hydrocarbon accumulation in Linnan Subsag, Huimin Sag. J. NW Univ. 34 (6), 713–716 (In Chinese).
Zhou, X.G., Sun, B.S., Tan, C.X., Tan, H.B., Zheng, R.Z., Ma, C.X., 2000. State of current geo-stress and effect of fault sealing. Petrol. Explor. Dev. 27 (5), 127–131 (in Chinese).