Applied Energy 351 (2023) 121836

Contents lists available at ScienceDirect

Applied Energy

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

A novel method for petroleum and natural gas resource potential evaluation and prediction by support vector machines (SVM)

Qiaochu Wang [a,b], Dongxia Chen [a,b,*], Meijun Li [a,b], Sha Li [a,b], Fuwei Wang [a,b], Zijie Yang [a,b], Wanrong Zhang [a,b], Shumin Chen [a,b], Dongsheng Yao [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*

HIGHLIGHTS

- A new method of petroleum resource potential prediction based on SVC and SVR.
- A combination of classification and regression models is applied for the first time.
- A systematical quantitative procedures for feature variable construction and selection.
- Discussion of the key factors for utilization of machine learning to energy research.

ARTICLE INFO

Keywords:
Petroleum and natural gas
Machine learning
Support vector machines
Hydrocarbon accumulation
Resource potential prediction

ABSTRACT

Petroleum and natural gas resources (PNGR) are some of the major forms of fossil energy that are important for the development of industry and energy security. Along with the growing demand of petroleum consumption and the requirement for enhancing drilling success rate, reducing the exploration risk and saving exploration cost, prediction method for PNGR potential with high accuracy and wide practicability is needed. However, the existing PNGR evaluation and prediction methods based on traditional statistical principles are far from meeting the requirements of the present petroleum exploration and exploitation. Therefore, this study introduces a novel method for PNGR potential prediction by applying support vector machines (SVM) in the context of the rapid development of artificial intelligence and machine learning. This novel machine learning methodology first proposed a combination of support vector classification (SVC) for hydrocarbon accumulation probability prediction and then support vector regression (SVR) for reserve abundance prediction. The combining use of classification and regression model can fully utilize the professional knowledge of petroleum geology and the powerful data processing capabilities of machine learning algorithms and hence significantly improve the performance of the method. Furthermore, the dataset is set based on petroleum geology knowledge with the feature variables of source rock, sandstone reservoirs, sealing capacity and hydrocarbon migration, whose distribution are predictable and thus ensures the predictive effect in practical petroleum exploration. The results show that the testing accuracy of the hydrocarbon accumulation probability evaluation model by SVC ranges from 80% to 100% with an average of 88.92%. The performance of the SVR model for evaluating reserve abundance also performs well with the highest correlation coefficient of 0.767. In addition, several validation ways are applied for testing the reliability and stability of the model. For a hold-out test for a new zone, the model provides precise prediction of hydrocarbon accumulation probability and reserve abundance with an accuracy of 72.5% and a correlation coefficient of 0.744. The evaluation metric of the F1-score shows an average of 0.91 for the SVC models, the 4-fold cross-validation shows an average correlation coefficient of 0.663 for SVR model, which indicates the good performance of the SVC and SVR model. To conclude, this study not only provides an intelligent ML method system for PNGR potential precisely evaluation and prediction with the combination of SVC and SVR which is firstly used by application of ML in petroleum industry field, but is also significant for the application of ML in petroleum and natural gas exploration and exploitation.

* Corresponding author at: 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.apenergy.2023.121836
Received 20 June 2023; Received in revised form 3 August 2023; Accepted 23 August 2023

1. Introduction

As major forms of fossil energy, petroleum and natural gas resources (PNGR) play crucial roles in the development of industry and people's lives worldwide. In recent years, the consumption of petroleum and natural gas has shown a straight increasing trend, although COVID-19 has constrained the demand for fossil energy [2]. Along with the Russia–Ukraine conflict this year, the energy shortage is becoming increasingly serious. In this context, the evaluation and prediction of PNGR potential are of great significance for both the development of the petroleum industry and the stabilization of the whole industry system [3–6,9]. The precise evaluation and prediction of PNGR has been demonstrated to largely enhance the efficiency and economic benefits of petroleum and natural gas exploration and exploitation [7,8,10,11,14]. While along with the development of petroleum industry, the previous evaluation methods are unable to meet the requirements of the exploration, especially for enhancing the success rate of drilling, reducing the exploration risk and saving exploration cost. In this case, a novel PNGR evaluation and prediction method with high accuracy, high efficiency and wide practicability is imperative.

The prediction of PNGR potential has existed since the initial development of the petroleum geology industry and discipline and is widely used in many petroliferous basins or districts [3,15–17,29]. After years of development, qualitative and semiquantitative evaluation and prediction methods based on mathematical geology and petroleum geology professional knowledge have formed with statistical principles as the main theoretical support [18,19,24,25,27,34]. Specifically, Dow, Singer and Reed et al. used petroleum systems analysis to evaluate the probability of hydrocarbon accumulation in Alaska and the Williston Basin to select favourable zones in the 1970s and 1980s [11,25,26]. McCammon et al., Cox and Freek et al. predicted the petroleum reserves of Montana, Venezuela and Puerto Rico using a professional evaluating system based on weights of evidence in the 1990s and 2000s [20,21,33]. Additionally, Agterberg, Jiang et al. and Chen et al. used quantitative calculations based on confidence function theory to predict oil saturation of reservoirs in the Dongpu Depression and Junggar Basin in China in recent years [13,14,39]. However, hydrocarbon accumulation has been confirmed to be a series of complex and interrelated geological processes, which are difficult to characterize quantitatively by statistical and regression principles [22,23,28]. Therefore, these methods often perform poorly in basins that have experienced multiple sedimentary and tectonic processes.

Along with the information technology revolution, machine learning (ML) algorithms have shown increasing superiority in data management and analysis and have been widely applied in transportation, medicine, agriculture, education and other fields [30–32,36]. Compared with traditional methods, evaluation and prediction methods based on ML can integrate the existing theoretical knowledge for model building and enhancement and can maximize objectivity in the decision-making process to reduce the subjectivity of human judgement [35,42–45].

In recent years, the utilization of ML in petroleum and natural gas exploration fields has been increasing. At present, the ML is mainly applied for petroleum production prediction, geological feature recognition and classification, and the resource potential prediction for a petroliferous basin or district (Table 1). The oil and gas production prediction of individual wells by ML algorithms is widely used with mature models. For production prediction, the size of data is often large enough. An individual well with 1-year production can provide over 10,000 couples of data including 8–20 variables [99–101]. Therefore, the prediction models with good performance are easily to be constructed especially when the advanced ML algorithms such as support vector machines (SVM), decision tree (DT), gradient boosting decision tree (GBDT), random forest (RF) and artificial neural network (ANN) are applied [102–104]. The application of ML in feature recognition and classification is mainly based on well logging and seismic data, which can also provide enough data size for the construction of the ML models. The recognition and classification of lithologies (shale, sandstone and carbonate, etc.) and geofluids (water, gas and oil) by ML models are widely used with high accuracy. The multiple well-log data with numerous wells and the interpretation results with the calibration of the drill tests provide data with high quality [40,105,106]. The development of the SVM and convolutional neural networks (CNN) provides a good way for seismic feature extraction and seismic interpretation [41,107–109]. The sedimentary facies and sandstone reservoirs identification by ML models can significantly improving the predicting accuracy and therefore the efficiency of petroleum exploration. However, comparing to the production prediction and feature recognition, the application of ML in PNGR potential is immature. The largest problem in resource potential prediction is that the data size based on the petroleum exploration is small (usually 100 to 300 couples of data) for the ML algorithms to construct an ideal model ([110]; [112]). Furthermore, there are many alternative feature variables owing to the complex geological processes which all have influences on the petroleum accumulation processes. With the different feature selection, feature fusion or dimensionality reduction processes, the final performances of the models are of large disparity. With the different output, the resource

Table 1
Application of ML methods on petroleum and natural gas exploration fields.

Output	Data source	Data size	ML algorithm	Performance	Representative References
Shale gas production	Production	18 variables 1 well	RF, SVM, ANN (individual)	Accuracy:96%	[99]
	Production	20 variables 10,000 wells	ANN	Corelation coefficient:0.94	[102];
	Production	9 variables 4256 wells	GBDT	Accuracy: 73%–79%	[129]
Lithologies	Well-log	7 variables 3232 samples	DT, SVM	Accuracy: 38%–62%	[105]
Oil and gas layer	Well-log	9 variables 3 wells	CNN	Accuracy:28%–82%	[106]
Sandstone reservoir distribution	Seismic	5 variables 2 seismic profiles	ANN	Corelation coefficient:0.83–0.86	Ariza et al., 2021
Sweet pot distribution	Seismic	8 variables 1 seismic profile	SVM	Identification of 8 favourable zone types	[110]
Sedimentary facies	Seismic	7 variables 2 seismic profiles	CNN	A division of 6 sedimentary facies	[108]
Hydrocarbon accumulation probability	Exploration	4 variables 519 samples	GBDT	72.45%	[113]
Gas content	Exploration	7 variables 102 samples	ANN	Corelation coefficient: >0.65	[114]

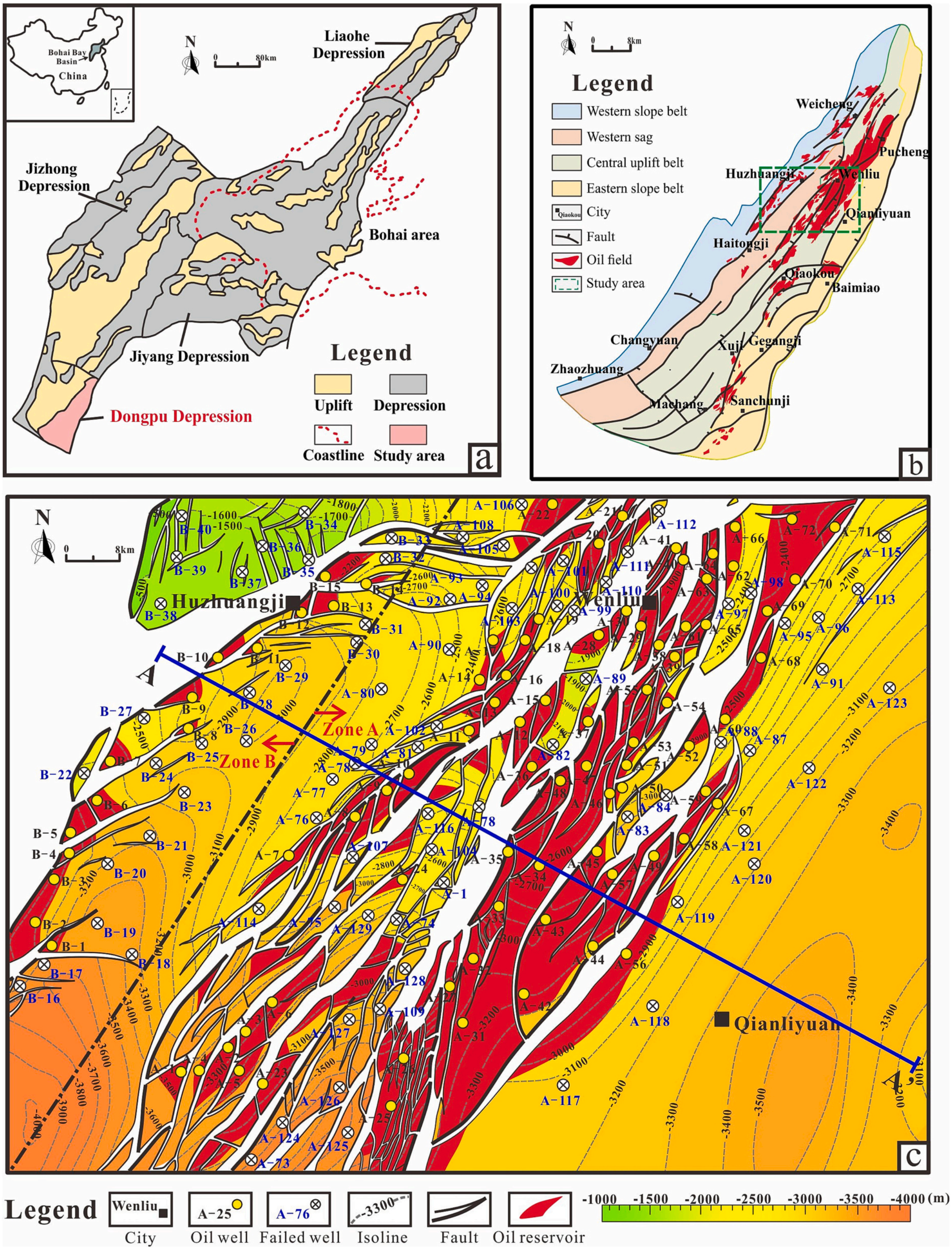


Fig. 1. Location, tectonic belts and oil reservoir distribution of the Es_3 Formation in the Dongpu Depression in the Bohai Bay Basin, eastern China. (a) Location of the Dongpu Depression. (b) Tectonic structures of the Dongpu Depression and location of the study area. (c) Oil reservoir distribution of the study area.

potential prediction model can be both classification and regression model. The classification models usually provide the probability of hydrocarbon accumulation or the location of the sweet pot for a district. While without the quantity evaluation of the petroleum resources, the classification model can only provide limited help in petroleum exploration practice [113]. The regression model often provides the distribution of oil saturation or oil and gas reserve which is useful for determining the location of exploration wells. However, the performances of the present regression models on resource potential prediction are not as good as the classification ones [114].

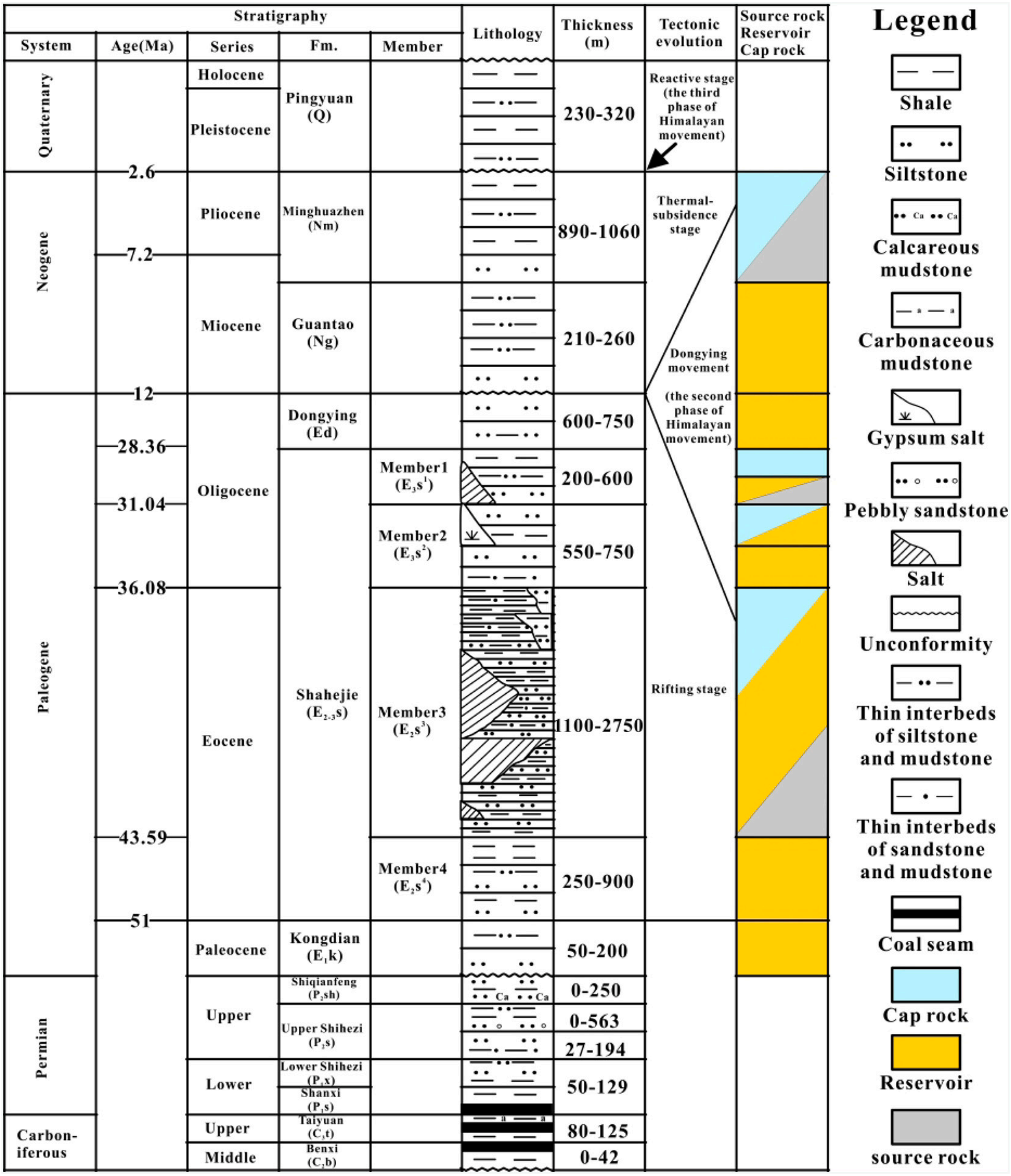


Fig. 2. Generalized stratigraphic column for the Dongpu Depression, Bohai Bay Basin (Edited by [49]).

To conclude, there are two main problems for the utilization of ML algorithms in PNGR exploration fields. The first is the difficulty in feature variable selection and characterization owing to the complex and interrelated geological processes, which often lead to low accuracy of the evaluation and prediction results. The second is the poor performance of the ML regression model (oil saturation, oil and gas reserves, etc.) which is attributed to the small dataset. These two problems are also the main challenges in PNGR potential evaluation and prediction.

To address these two problems, this study takes the Dongpu Depression in the Bohai Bay Basin as an example and proposes a quantitative evaluation and prediction method system for PNGR potential. To address the first problem, this method system uses a feature variable decision method based on petroleum geological knowledge to determine effective parameters and further enhance the quality of the dataset and predicting accuracy. For the second problem, this study uses a method containing two SVM models to enhance the final prediction accuracy. First, the probability of hydrocarbon accumulation is judged by the support vector classification (SVC) algorithm. Second, the reserve abundance in the favourable zones of hydrocarbon accumulation is determined by the application of the support vector regression (SVR) algorithm. This study provides an intelligent ML method system for PNGR potential precisely evaluation and prediction with the combination of SVC and SVR which is firstly used by application of ML in petroleum industry field. The construction of dataset with variable features is also predictable for the modelling application in practical. Furthermore, the construction of the novel method is also significant for the application of ML in petroleum and natural gas exploration and exploitation.

2. Geological setting

The Dongpu Depression is located in the southwest of the Bohai Bay Basin, which is one of the largest petroliferous basins in China (Fig. 1a). The Dongpu Depression is characterized by a rift basin with developed faults of different sizes. The depression is divided into 4 main tectonic zones, which are the western slope belt, western sag, central uplift belt

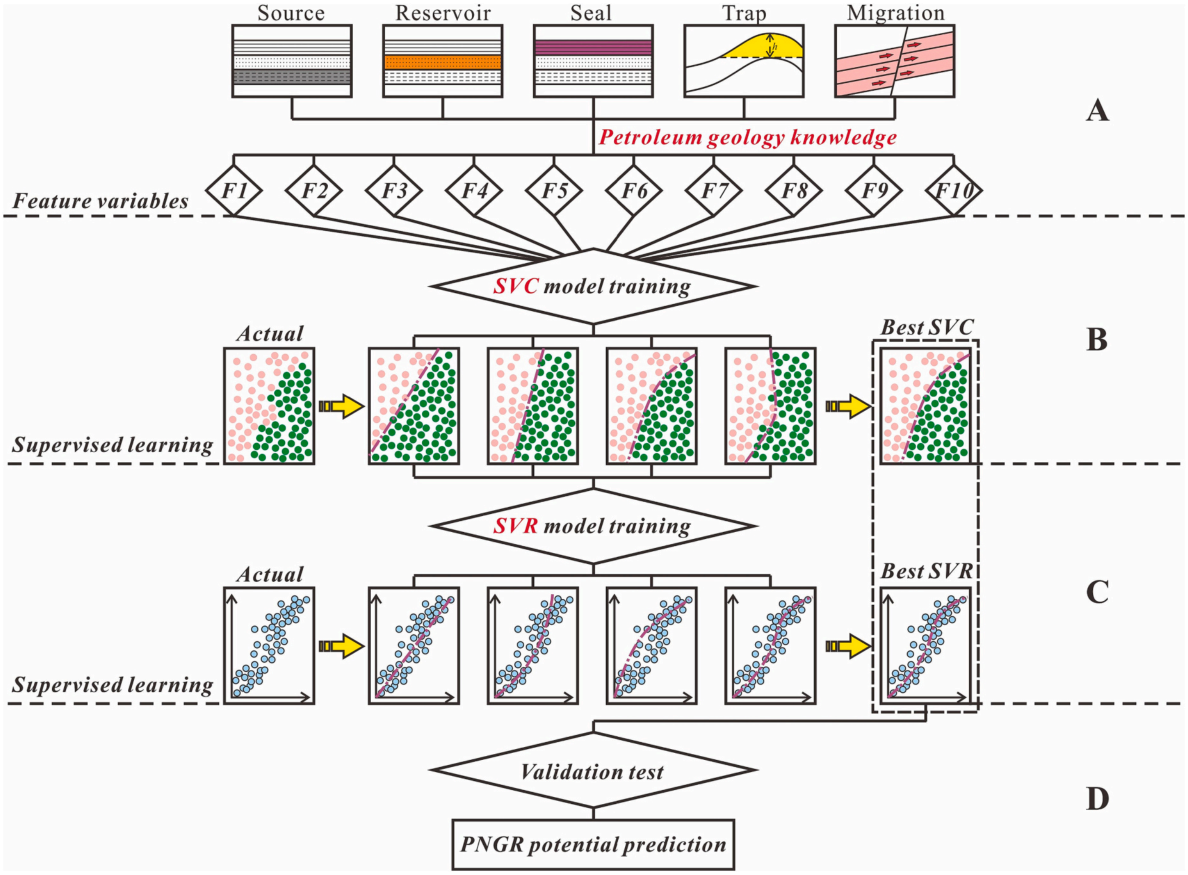


Fig. 3. Workflow of the PNGR potential evaluation and prediction model construction by SVC and SVR.

and eastern slope belt from east to west (Fig. 1b). Oil and gas resources are abundant in the Dongpu Depression and distributed mainly in the northern part; hence, this region is selected as the study area of this research. The oil reservoirs in the study area are dominated by fault block reservoirs, which are cut into hundreds of units of different sizes and shapes by faults (Fig. 1c). At present, 4 large oilfields with over 300 oil reservoirs have been discovered, which indicates the high resource potential of the study area.

From the bottom to the top, 12 formations developed in the Dongpu Depression with large disparities in sedimentary environment and lithologies (Fig. 2). The main source rocks are dark mudstone and shale in the 3rd member of the Shahejie Formation (Es_3 Fm). These two main source rocks are located in the southwest and southeast sags of the study area named the Haitongji and Qianliyuan Sags (Fig. 1b). The main reservoirs are the sandstone or siltstone reservoirs in the Es_3 Fm, 2nd member of the Shahejie Formation (Es_2 Fm), 1st member of the Shahejie Formation (Es_1 Fm), Dongying Formation (E_d Fm) and Guantao Formation (N_g Fm). The main caprocks are low-permeability mudstone and gypsum salt in Es_3 and Es_2 [47,48]. To conclude, there are multiple petroleum systems and hydrocarbon accumulating units consisting of different combinations of source rocks, reservoirs, and caprocks [48,49].

The Dongpu Depression experienced three main tectonic events from the Palaeogene period to the present. The whole Palaeogene period was dominated by a rifting stage with frequent fault movement [46,49]. Following deep burial of the sediments, the source rocks began to generate hydrocarbons. Oil charging and migration were active in the late Palaeogene period [46,48]. Then, the whole depression experienced a tectonic uplift and erosion process at the end of the Dongying period. From the Neogene period to the present, the depression experienced a stable thermal subsidence stage without large-scale fault movement (Fig. 2). The disparities in depositional processes and tectonic movements led to the complexity of the geological conditions in the Dongpu Depression and therefore a disparity in the distribution and enrichment of petroleum and natural gas resources.

3. Method and data

The construction of the evaluation and prediction model of PNGR potential can be divided into 4 steps. First, the dataset is constructed by data processing. Considering that the amount of data is relatively small, the original data are optimized based on petroleum geological knowledge. Second, the probability of hydrocarbon accumulation is determined based on the SVC algorithm. Third, the dataset for reserve abundance prediction is constructed and used to construct the SVR model for reserve abundance evaluation and prediction. Finally, the PNGR potential evaluation and prediction model based on two SVM models is tested using the data from the adjacent petroleum district (Fig. 3).

3.1. Data processing

The hydrocarbon accumulation probability and reserve abundance are influenced by multiple interrelated factors, which are characterized by both quality and quantity [12,50–52,57]. Owing to the differences in data resources, there are often problems associated with data mismatching and missing data. In addition, the amount of data is often too small to meet the requirements of ML. Therefore, it is necessary to identify independent and dependent variables based on professional knowledge and dimension reduction.

3.1.1. Feature variable setting

Based on petroleum geological principles, the hydrocarbon

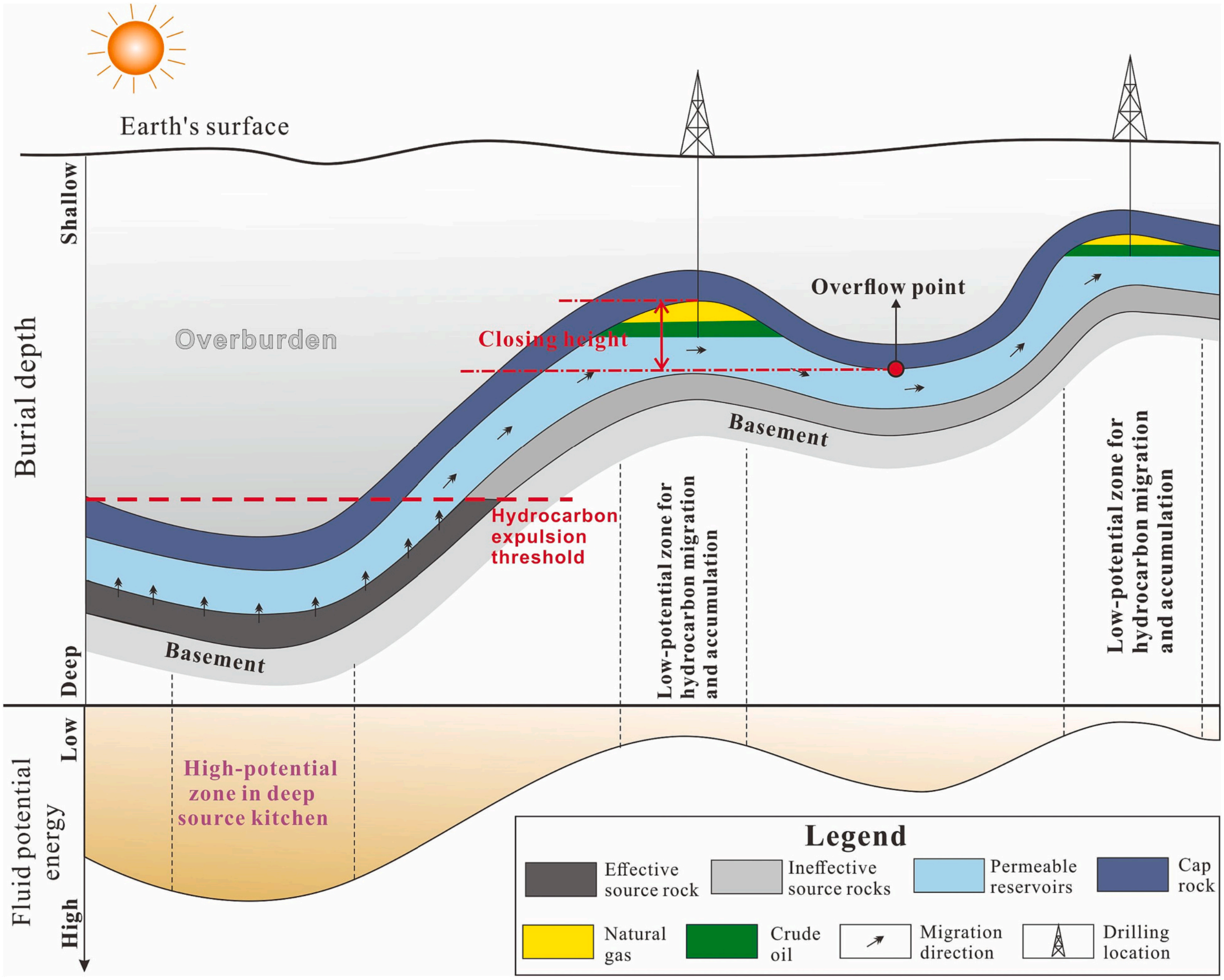


Fig. 4. Hydrocarbon accumulation processes and key parameters defined in petroleum geology.

accumulation process of conventional resources is controlled by source rock conditions, sedimentary reservoir conditions, seal conditions, trap conditions and migration conditions [19,22,23,53] (Fig. 4). The characterization methods for the hydrocarbon accumulation factors are listed below.

(1) Source rock conditions

The condition of the source rocks is considered one of the most important factors in hydrocarbon accumulation [54–56,60]. Previous studies have identified some quantitative parameters of source rocks, including the thickness, maturity, and total organic content (TOC) [9,50,57–59]. However, considering that only the hydrocarbons expelled from the source rocks can accumulate in reservoirs, the hydrocarbon expulsion intensity (HEI), which is defined as the quantity of hydrocarbon expulsion per unit area, is introduced to characterize the source rock conditions [61]. The HEI can be calculated as follows [61–63]:

$$HEI = \int_{z_0}^{z} 0.1 \times P_e \times H \times \rho \times TOC \times dz \tag{1}$$

$$P_e = P_{g-original} - P_{g-remain} \tag{2}$$

where *HEI* is the hydrocarbon expulsion intensity (t/km^2); *z* is the burial depth (m); and z_0 is the burial depth of the hydrocarbon expulsion threshold (m), which is determined by the crossplot of (S_1 + S_2)/TOC and Ro. P_e is the hydrocarbon expulsion ratio (mg/g), which can be calculated with Eq. (2) (Fig. 5a). Previous studies have shown such a relation and parameters in the Dongpu Depression [64]. *H* is the thickness of the source rocks (m), *ρ* is the density of source rock (g/cm^3), and TOC is the total organic content (%).

Note that in a simple petroleum system, there is always a single source rock (points A and D in Fig. 5b). However, in complex petroleum systems, there are always multiple source rocks (points B and C in Fig. 5b). In this case, the final HEI is the sum of the HEI values of each source rock.

(2) Reservoir conditions

The characteristics of sandstone reservoirs are another important factor for hydrocarbon accumulation. The distribution of sandstone reservoirs and their physical properties largely determine the probability of hydrocarbon accumulation and oil and gas reserves. The distribution of sandstone reservoirs can be described by sedimentary facies distribution or sandstone thickness evaluation [23,67], while the physical properties of reservoirs are always characterized by porosity and permeability [65]. The sedimentary facies and sandstone thickness are always determined by petroleum geological signals (core analysis, stratigraphy determination and well log data) and knowledge, and their distributions are generally continuous and thus predictable [66,68]. However, sedimentary facies descriptions in words need to be translated into digital data. In this study, we create the sedimentary facies index (*FIs*) to quantitatively characterize the sedimentary facies by Eq. (3):

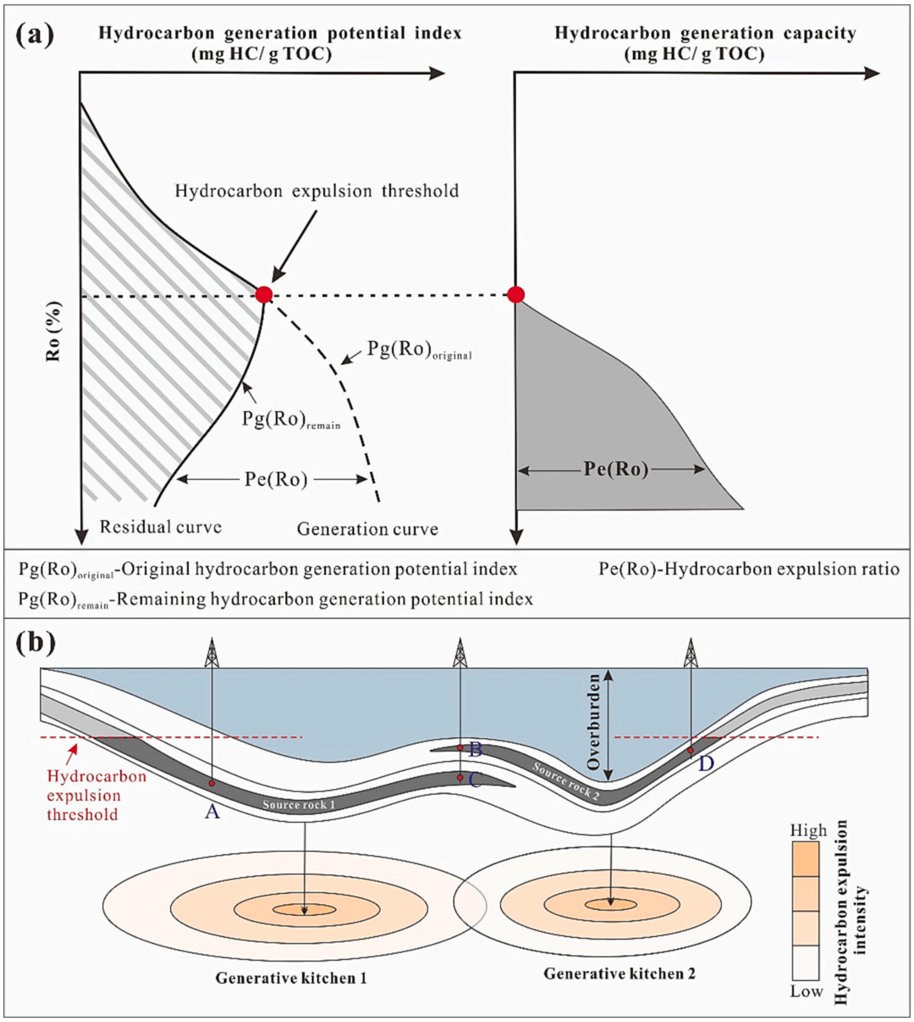

(a)
Hydrocarbon generation potential index (mg HC/ g TOC)
Hydrocarbon generation capacity (mg HC/ g TOC)
Hydrocarbon expulsion threshold
Ro (%)
Pg(Ro)original
Pg(Ro)remain
Pe(Ro)
Residual curve
Generation curve
Pg(Ro)original-Original hydrocarbon generation potential index
Pe(Ro)-Hydrocarbon expulsion ratio
Pg(Ro)remain-Remaining hydrocarbon generation potential index
(b)
Overburden
Hydrocarbon expulsion threshold
A
B
C
D
Source rock 1
Source rock 2
High
Low
Hydrocarbon expulsion intensity
Generative kitchen 1
Generative kitchen 2

Fig. 5. Definition and calculation of hydrocarbon expulsion intensity.

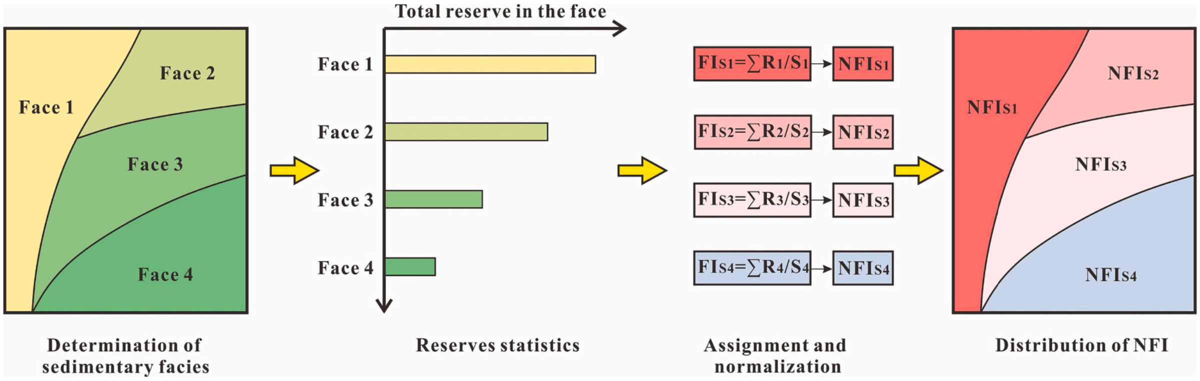


Fig. 6. Definition and characterization of the normalized facies index (*NFI*).

$$FI_{si} = \frac{\sum R_i}{S_i} \tag{3}$$

where FI_{si} is the sedimentary facies index for type i sedimentary facies, mt/km^2; $\sum R_i$ is the total reserve distributed in type i sedimentary facies, mt; and S_i is the total area of type i sedimentary facies, km^2.

Considering that the value range will influence the accuracy of the SVM algorithm ([79]), FI_{si} is normalized to NFI_{si}, which is calculated by Eq. (4):

$$NFI_{si} = \frac{FI_{si}}{FI_{si-max}} \tag{4}$$

where NFI_{si} is the normalized sedimentary facies index for type i sedimentary facies, dimensionless, and FI_{si-max} is the maximum of I_{si} in the study area, mt/km^2 (Fig. 6). There are 4 main sedimentary facies in the study area: fluvial facies, delta-plain facies, delta-front facies, and lacustrine facies. The fluvial facies has the largest total petroleum reserve with the highest NFI of 1.00, and the NFI values for the delta-

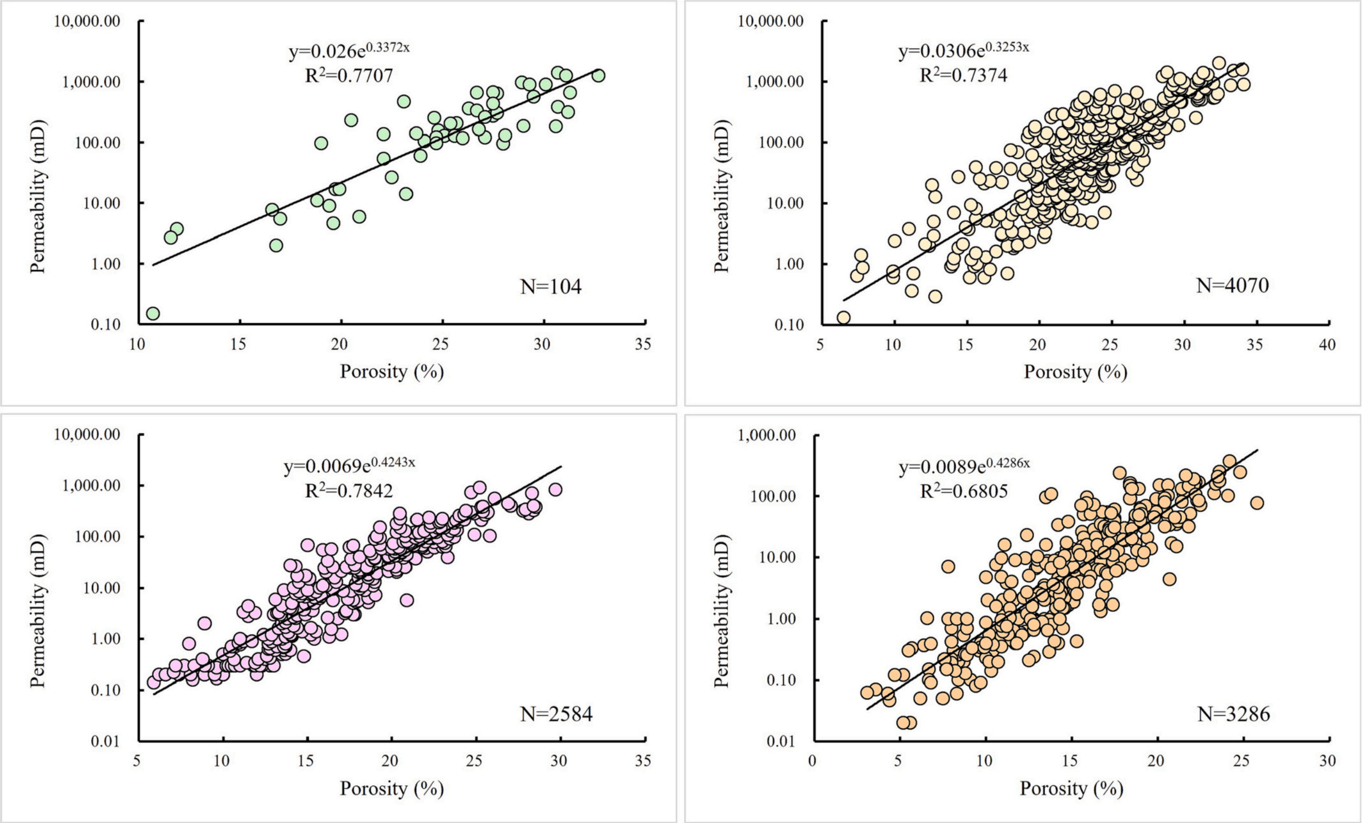


Fig. 7. Porosity–permeability relationships for different depth ranges: (a) <2000 m, (b) 2000–2500 m, (c) 500–3000 m and (d) >3000 m in the Dongpu Depression.

plain, delta-front, and lacustrine facies are 0.37. 0.25 and 0.01, respectively. The lateral distribution of the NFI is shown in Section 4.1.1.

The porosity and permeability of sandstone reservoirs can be obtained from laboratory experiments and well logging analysis. Previous studies have shown that the relationships between sandstone porosity and depth and between porosity and permeability provide a method to obtain continuous porosity and permeability distributions for prediction work [69,70]. In this study, the porosity and permeability distribution of the study area is determined based on the 10,044 pairs of tested porosity and permeability data from different depth ranges (Fig. 7).

(3) Caprock conditions

The caprocks of a petroleum system are generally a stably distributed low-permeability layer that prevents oil and gas from escaping [19]. Good caprocks often show obvious differences in well log data or seismic profiles. In this study, the caprock conditions are classified as good and poor, which are represented by 1 and 0 (1 for present and 0 for missing) for ML model construction and prediction.

(4) Trap conditions

The trap is a local structure underground that provides favourable hydrocarbon accumulation space [19,65]. The crucial parameter for a trap is the closing height, which is defined as the elevation difference between the highest point of the trap and the overflow point (Fig. 4). It is obvious that the greater the closing height is, the higher the probability of hydrocarbon accumulation.

(5) Migration conditions

The migration process of oil and gas from source rocks to sandstone reservoirs is crucial for hydrocarbon accumulation and directly influences the reserve abundance in reservoirs [71,72,77]. The hydrocarbon migration process is complicated and difficult to describe in detail. Hubbert [74] introduced the definition of fluid potential, which is applied for characterizing the possible migration direction and pathways and is widely used for petroleum migration analysis in many petroliferous basins worldwide [75,76]. The fluid potential is defined as the total mechanical energy of a unit underground relative to the set base level (usually sea level) [74]. In this case, the pore fluid will migrate from a high-fluid potential place to a low-fluid potential place. In this study, the fluid potential is also selected as the feature parameter for characterizing the migration conditions of petroleum and natural gas. The fluid potential is calculated by Eq. (5) as follows [73,74]:

$$\varphi_f = \rho_f g z + P \tag{5}$$

where φ_f is the fluid potential of geofluids, J/kg; ρ_f is the density of geofluid, kg/m^3; g is the gravitational acceleration, 9.8 m/s^2; z is the distance between the calculating point and the base level, m; and P is the fluid pressure, Pa. In this study, the geofluid is formation water with a density of 1.0 g/cm^3.

Other than the above feature variables, the reservoir temperature, pore pressure.

3.1.2. Target variable setting

In this study, the PNGR potential is determined by the hydrocarbon accumulation probability and reserve abundance, which are considered the target variables for ML model construction.

(1) Hydrocarbon accumulation probability

The hydrocarbon accumulation probability is divided into two labels of success and failure and encoded as 1 and 0, respectively. The

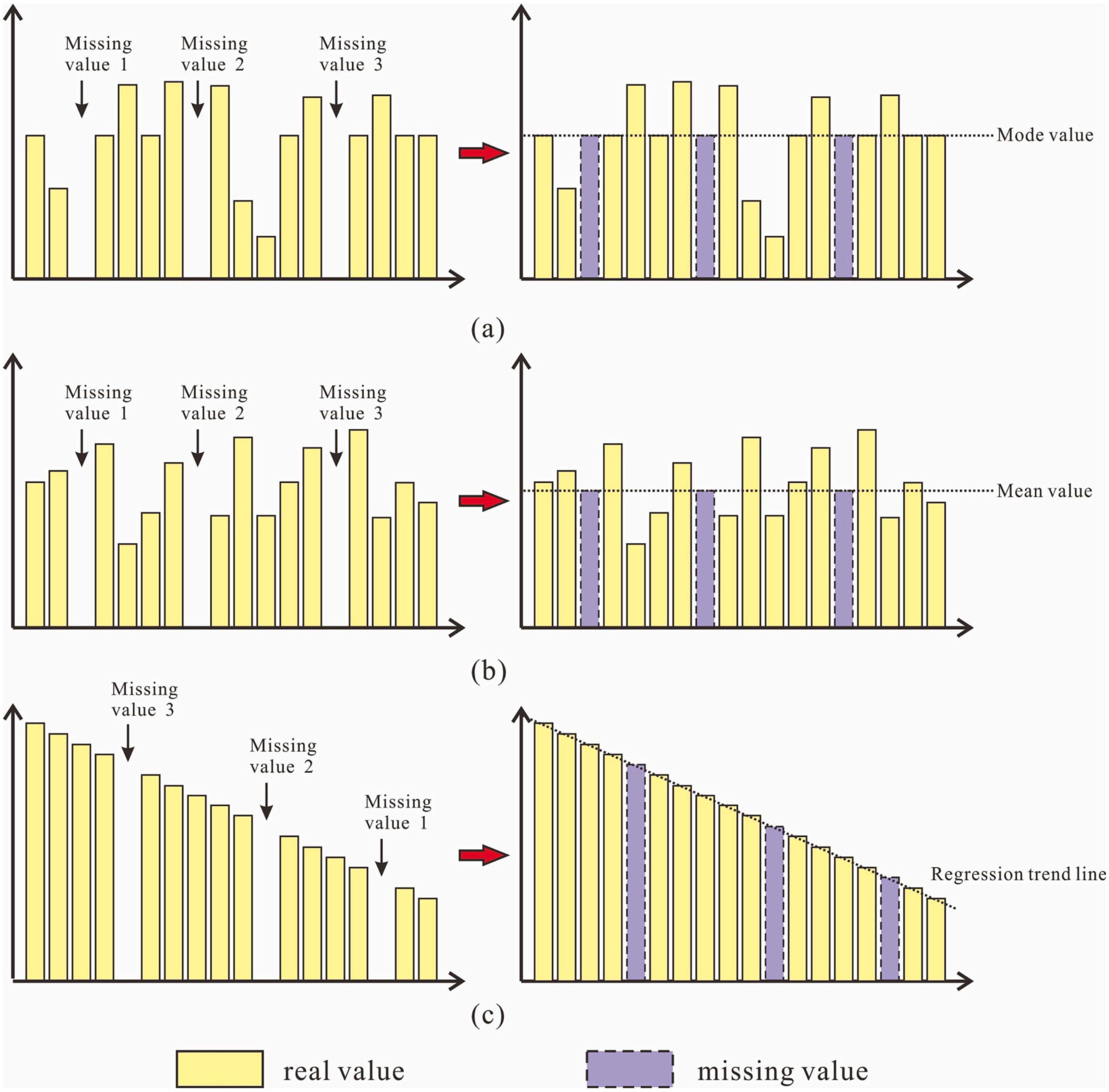


Fig. 8. Missing value compensation methods in ML. (a) Mode value replacement; (b) Mean value replacement (applied for sandstone thickness missing value); (c) Regression replacement (applied for porosity and permeability missing value).

hydrocarbon accumulation probability label is judged by oil and gas data from exploratory wells.

(2) Reserve abundance

The reserve abundance of a hydrocarbon reservoir is defined as the ratio of reserve and the area of the reservoir. The reserve abundance can reflect favourable zones of hydrocarbon accumulation and the economic value for the PNGR.

3.2. *Feature engineering*

The feature engineering processes are important for enhancing the quality of the data of feature variables and thus the final performance of the evaluation and prediction models. The feature engineering analysis in this paper includes standardization, missing value compensation and the noisy feature elimination.

3.2.1. *Standardization*

The standardization process is always the first step for the feature engineering and it can significantly improve the efficiency of the ML model by reducing the calculation time. In this research, the *Z*-score standardization is selected to make the value of the dataset to range from −1 to 1and conform to normal distribution.

3.2.2. *Missing value compensation*

The original dataset is usually incomplete which needs the compensation or fixing by several statistical methods, including mode value replacement, mean value replacement and regression replacement (Fig. 8) ([139]). In the original data applied in this research, 22 pairs of data are incomplete, including 9 sandstone thickness data, 6 porosity data and 7 permeability data. The sandstone thickness data is replaced by the mean value from the adjacent wells because of the gradual change of the sandstone thickness, while the porosity and permeability value are compensated by the regression replacement because the well relationship between porosity and permeability in different depth ranges (Fig. 7).

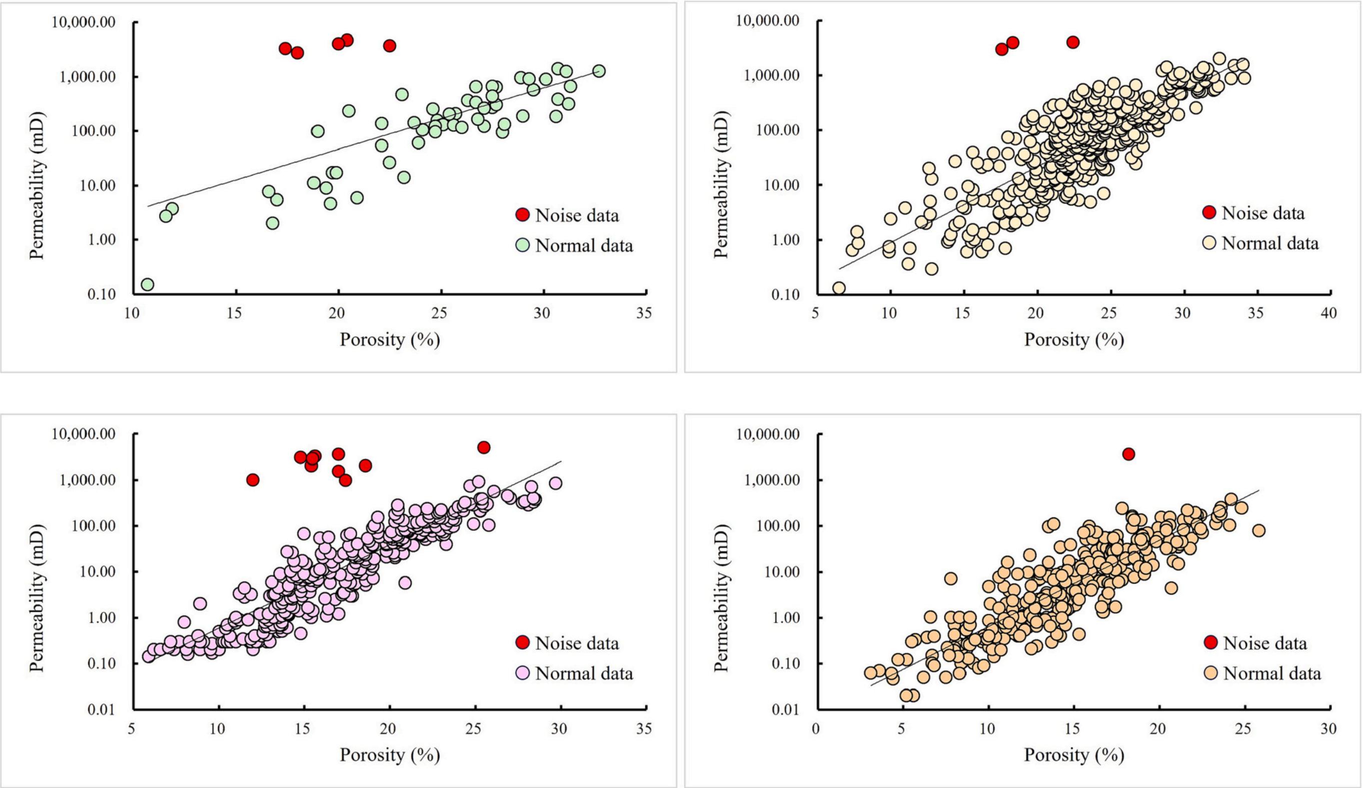


Fig. 9. Noisy feature value recognition for porosity-permeability feature variables for different depth ranges. (a) <2000 m, 5 pairs of noisy data are eliminated; (b) 2000–2500 m, 3 pairs of noisy data are eliminated; (c) 500–3000 m, 10 pairs of noisy data are eliminated; (d) >3000 m, 1 pair of noisy data are eliminated.

3.2.3. Noisy feature elimination

For constructing a prediction model with generality, some of the noisy feature value should be eliminated from the original dataset, which do not follow the general distributions or rules. In this research, the noisy feature values exist in porosity and permeability value. In this research, 19 pairs of porosity-permeability data are recognized as noisy feature because of the large deviation between these values and normal trend (Fig. 9). All the porosity and permeability values are higher than those of the normal trend because the development of the fracture (due to improper experimental operation) for these measured samples can cause an increase in the pore spaces and fluid pathways which will significantly improve the porosity and permeability values.

3.3. Evaluation and prediction models based on SVM

The support vector algorithm is a nonlinear generalization of the generalized portrait algorithm, which is firmly based on the framework of statistical learning theory [78,79]. The SVM method was largely developed and applied in optical character recognition [80–82]. SVC is effective for dealing with classification problems became competitive with the robust available systems [83]. Additionally, SVR with excellent performance is soon obtained [84,85]. As the application of kernel functions has matured, SVM has become the most powerful machine learning algorithm for both classification and regression problems [86–88].

3.3.1. Hydrocarbon accumulation probability prediction by SVC

The principle of the SVM is extremely complex. In brief, the support vector addresses classification problems by constructing hyperplanes in a high- or infinite-dimensional space [89,90,96]. The best classifier is determined with the largest distance to the nearest training-data point of any class [91–93]. In this study, the hydrocarbon accumulation probability prediction model based on the SVC algorithm is constructed as follows:

(1) The training dataset of n points is formed:

$$(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)$$

where x_i represents a p-dimensional vector (p is the number of feature variables) and y_i is labelled 1 or -1 for the success or failure of hydrocarbon accumulation, respectively.

(2) The hyperplanes for the SVC model are determined:

The hyperplane that can divide the group of points x_i for which $y_i = 1$ from the group of points for which $y_i = -1$ can be written as the set of points x satisfying:

$$w^T x - b = 0 \tag{6}$$

where w is the normal vector to the hyperplane and b is the intercept.

For a normalized or standardized dataset, each hyperplane must satisfy:

$$w^T x_i - b \geq 1, if\ y_i = 1 \tag{7}$$

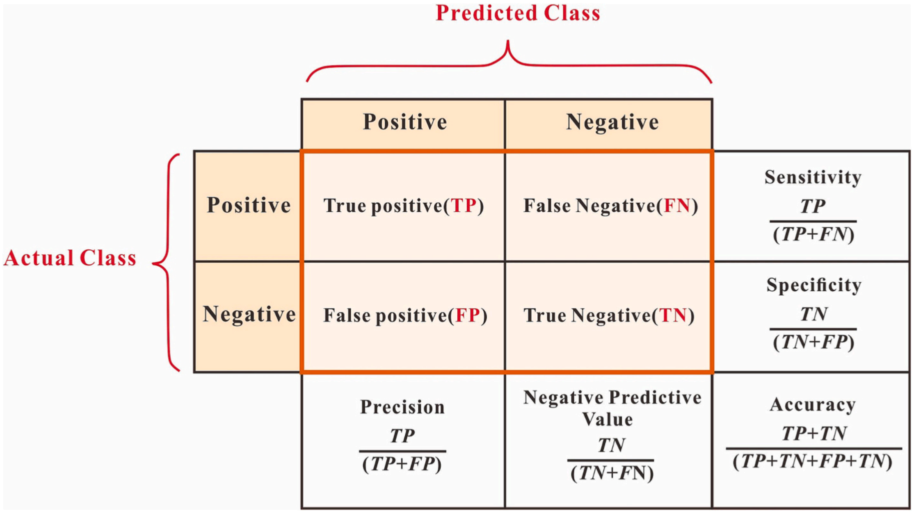
$$w^T x_i - b \leq -1, if\ y_i = -1 \tag{8}$$

Eq. (8). and Eq. (9) can be written as one equation:

$$y_i \left(w^T x_i - b \right) \geq 1, 1 \leq i \leq n \tag{9}$$

(3) The best hyperplane is determined:

The offset of the hyperplane from the origin along the normal vector



		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True positive(TP)	False Negative(FN)	Sensitivity $\frac{TP}{(TP+FN)}$
	Negative	False positive(FP)	True Negative(TN)	Specificity $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+TN)}$

Fig. 10. Evaluation metrics for classification problems in machine learning ().

w is $\frac{b}{\|w\|}$, which should be as large as possible. To conclude, the problem of finding the best SVC model for hydrocarbon accumulation probability prediction can be written as follows:

Minimize $\|w\|$ subject to $y_i(w^Tx_i - b) \geq 1$, for $1 \leq i \leq n$.

Moreover, for dealing with the nonlinear classification for the problem in this study, kernel functions are also applied. In this study, several functions are tested, and the radial basis function is used as the final kernel function.

3.3.2. *Reserve abundance prediction by SVR*

Similar to the classification problem, the support vector algorithm also addresses regression problems by constructing hyperplanes [94]. The goal of the SVR is to find a regression hyperplane that has the most deviation ε from the farthest training data point. In other words, the deviation of all the data points must be smaller than ε [89,95,96]. In this study, the reserve abundance prediction model based on SVR is constructed as follows:

(1) The training dataset of n points is formed:

$$(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)$$

(2) The hyperplanes for the SVR model are identified:

For a normalized or standardized dataset, each regression hyperplane must satisfy:

$$y_i - w^T x_i - b \leq \varepsilon \tag{10}$$

$$w^T x_i + b - y_i \leq \varepsilon \tag{11}$$

(3) The best regression hyperplane is determined:

To determine the best regression hyperplane, our goal is to find a hyperplane that has the highest ε from the actually obtained targets y_i and at the same time is as flat as possible [94,97]. In other words, we have to find the minimum of $\frac{1}{2}\|w\|^2$.

However, we cannot always obtain all the training data points by setting a small ε, and it is also unreasonable to enlarge ε for several abnormal points, which may interfere with the determination of the regression hyperplane. The slack variable ξ is then introduced to cope with otherwise infeasible constraints of the problem [82,98]. With the slack variable, the problem of finding the best SVR model for reserve abundance prediction can be written as follows:

$$Minimize\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{12}$$

$$\text{Subject to} \begin{cases} y_i - w^T x_i - b \leq \varepsilon + \xi_i \\ w^T x_i + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \tag{13}$$

where ξ_i is the upper bound slack variable and ξ_i^* is the lower bound slack variable.

Moreover, for demonstrating the well performance and highlighting the advantages of the method introduced in this paper, a comparison between our method and other classification or regression algorithms including DT, RF, naïve bayes (NB), k-nearest neighbour (KNN), multiple-liner regression (MLR), polynomial regression (PR) and GBDT algorithms is also used to construct PNGR prediction models with the same dataset.

3.4. *Model boosting and evaluation*

The constructed SVM models need calibration and boosting to overcome overfitting and enhance the performance for new data. In this research, the hyperparameter tuning is applied for both SVC and SVR model boosting. For evaluating the classification results and performance, the evaluation metrics are used for SVC models. While considering the relatively small data size for regression model construction, the cross validation is applied for evaluating the stability of the regression model and eliminating the influence of training and validating data composition on the model performance [133,134].

3.4.1. *Hyperparameter tuning for SVC and SVR model calibration*

The determination and tuning of hyperparameters are a key step for ML application for its significant influence on the final performance of the ML models. The most important hyperparameters for the SVM model are type of kernel function, gamma value and C value [135]. The kernel function is the way for SVM to map the input space to a high-dimensional feature space by nonlinear transformation [112]. The different kernel functions have different influences on linear and nonlinear problems. In this research, the linear, polynomial and radial basis function (RBF) are used as alternative functions for the determination of the kernel for both the SVC and SVR models. The gamma value is a coefficient of the kernel of the polynomial and RBF which can influence the model performance by influencing the kernel [112,136]. The

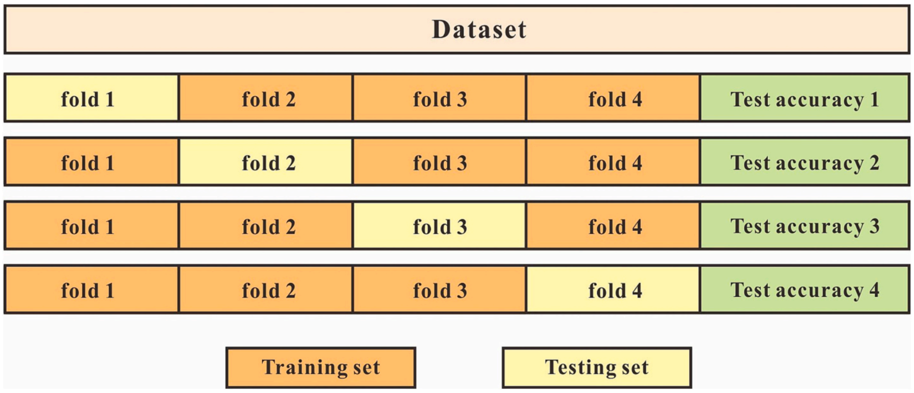


Fig. 11. Utilization of 4-fold cross validation (Edited by Wang et al. [141]).

C value is named as the penalty coefficient, which represents the tolerance of the model for errors [137]. The larger C value represents lower tolerance level of error which may lead to complex model with high training accuracy but also an overfitting trend; the smaller C value represents higher tolerance level of error which may lead to simple model with certain generalization ability but under-fitting trend. Therefore, the determination of the suitable composition of kernel, gamma value and C value is the key for boosting the evaluation and prediction model. In this study, the gamma and C value are determined by random sampling for enhancing the efficiency of the model construction process. The searching range of the random research for the gamma value is from 1 to 10 and 0.1 to 1 for SVC and SVR model, respectively, the searching range of the SVR model is from 0.01 to 0.1. Furthermore, in order to search for the best parameter combination, each model will undergo 100 iterations.

3.4.2. *SVC model evaluation by evaluation metrics*

The evaluation metrics of classification problems applied in this research because the testing accuracy cannot completely reflect the performance of the classification models ([43]). The output result for a binary classification model can be divided into 4 types owing to the different situations between the predicted class and actual class (Fig. 10). Correspondingly, there are 4 basic evaluation criteria called sensitivity, specificity, precision and negative predictive value (NPV) to

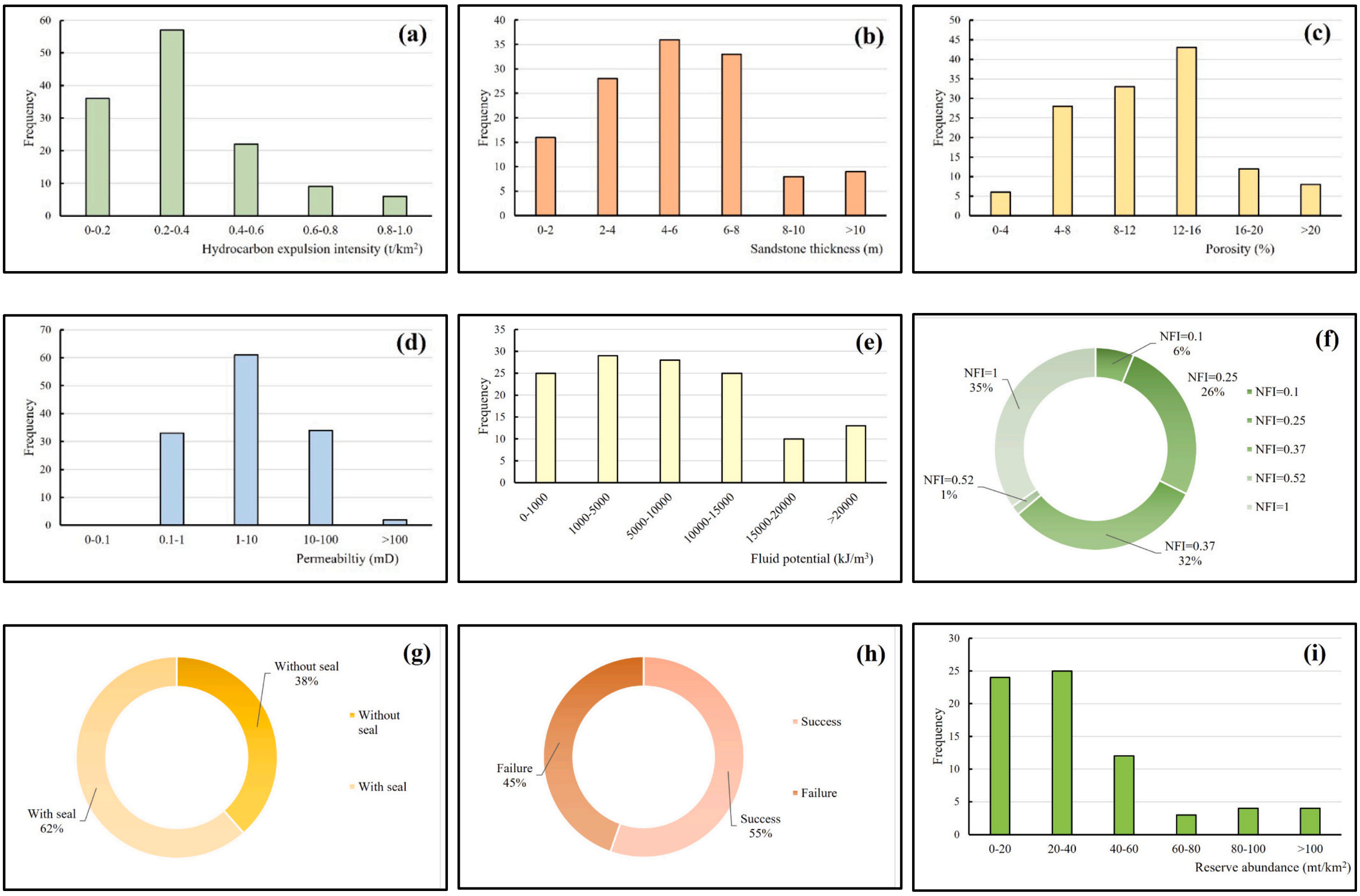


Fig. 12. Composition and distribution of training and testing data. (a) Distribution of the HEI. (b) Distribution of sandstone thickness in the Es_3 Formation. (c) Distribution of reservoir porosity. (d) Distribution of permeability. (e) Distribution of fluid potential. (f) Distribution of NFI. (g) Composition of seal development. (h) Composition of hydrocarbon accumulation probability. (i) Distribution of reservoir abundance.

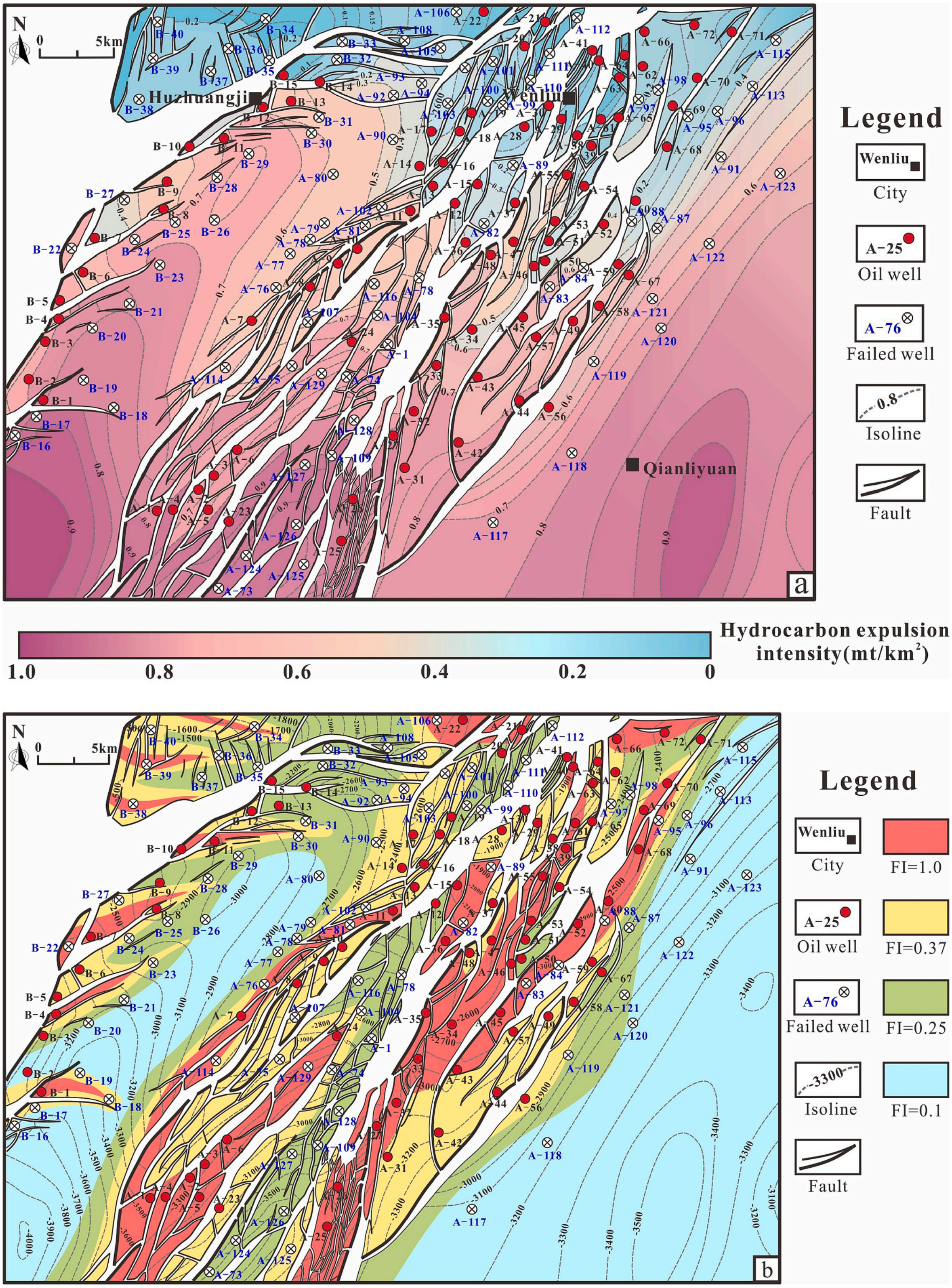


Fig. 13. Lateral distributions of feature variables: (a) HEI, (b) NFI, (c) sandstone thickness, (d) seal development and (e) fluid potential in the study area.

evaluate the performance of the classification model on the positive class and negative class (Fig. 10). Furthermore, the F1-score is also introduced to evaluate both sensitivity and precision:

$$F1 - score = \frac{2*Precision*Sensitivity}{(Precision + Sensitivity)} \tag{14}$$

3.4.3. SVR model evaluation by 4-fold cross validation

The cross-validation is an effective validation method to estimate the performance of the reliability and generalization of models [137]. The cross-validation is a resampling method which equally divides the original data to several parts, and uses the different parts of data to train and test a model on different iterations [133]. When one subgroup is

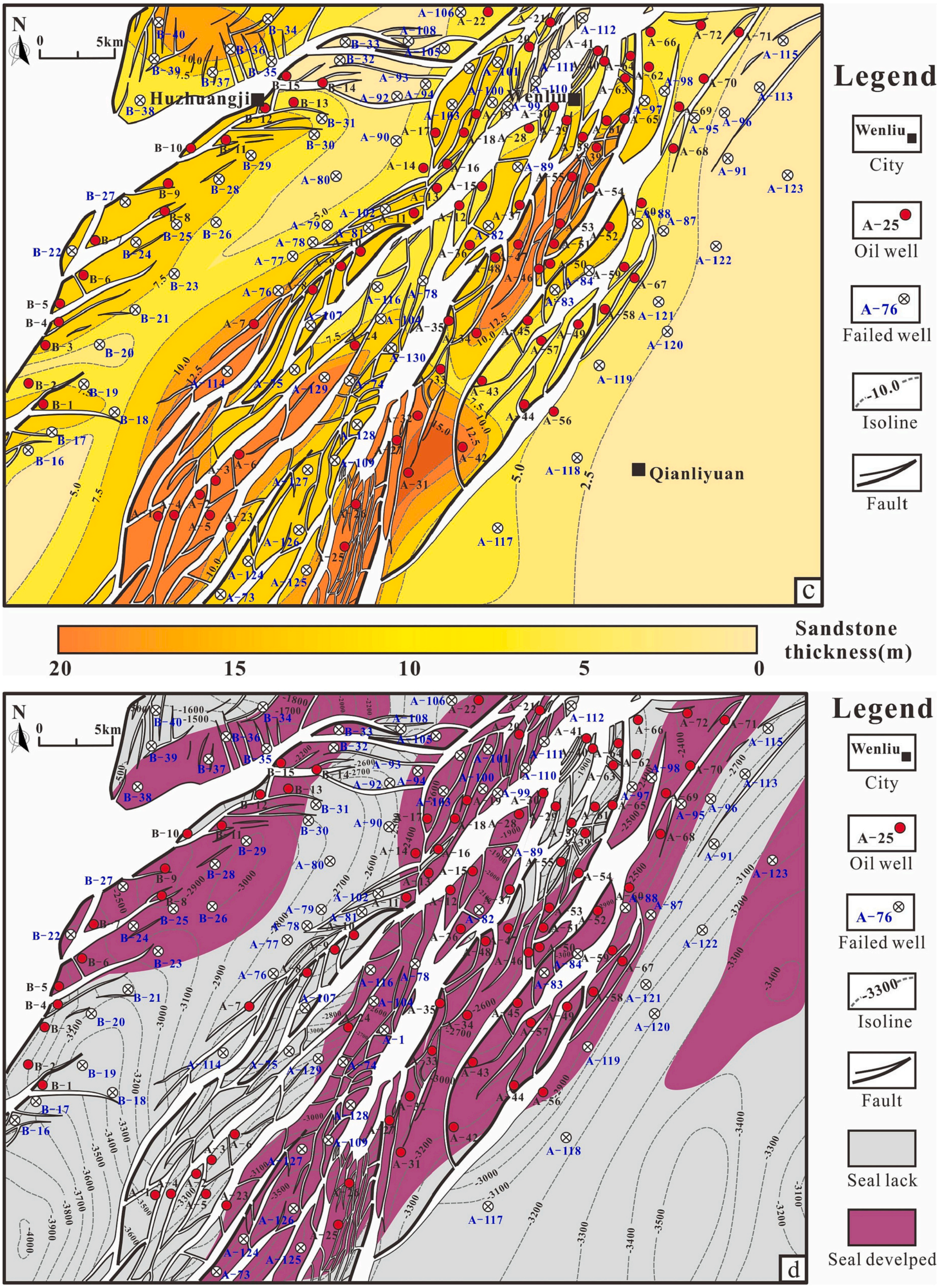


Fig. 13. (*continued*).

selected as testing data, the other subgroups are selected as the training data (Fig. 11). In this study, a 4-fold cross validation is applied, through which the original dataset for RA prediction is divided into 4 subgroups. In the process of the 4-fold cross-validation, each part of the data is used as testing data exactly once, which will overcome the repeat random sampling for certain data [134].

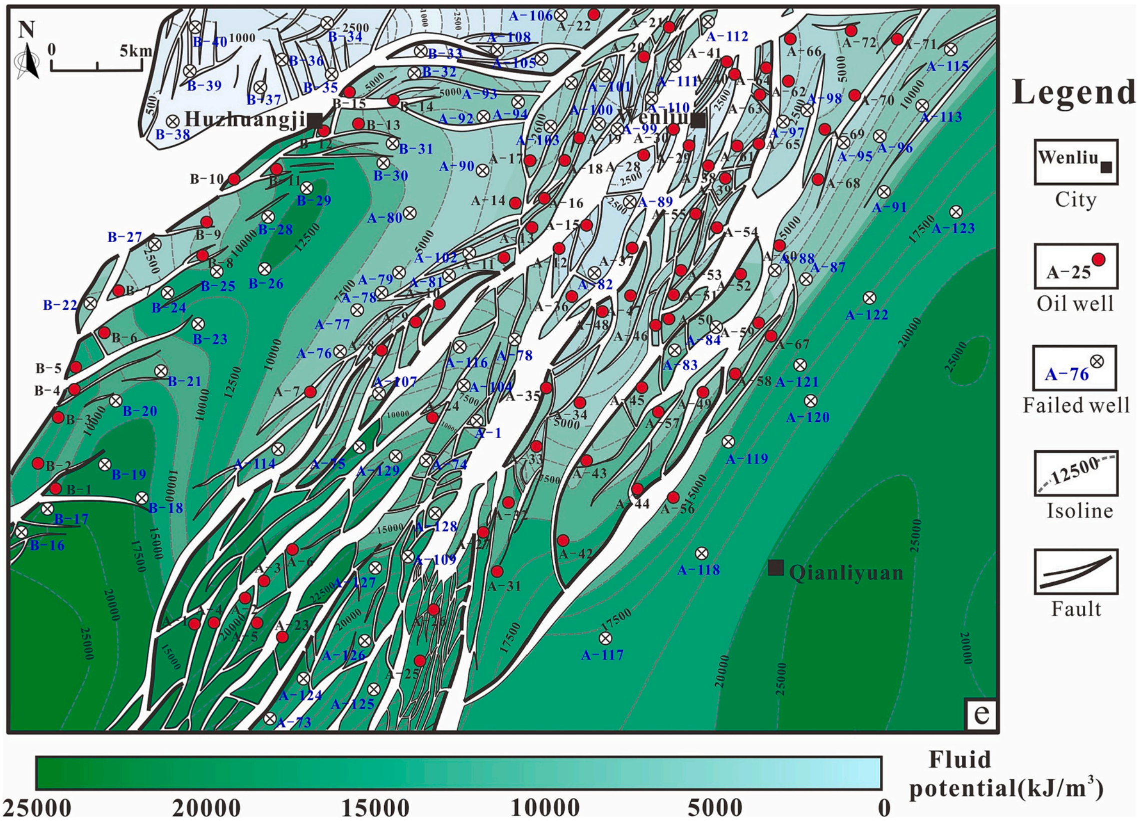


Fig. 13. (*continued*).

3.5. Peculiarity of the method

For dealing with the small data size and lower regression accuracy for PNGR potential prediction, there are two main innovations of the new ML method in this research:

(1) The feature variables are constructed and calibrated by integration and noise reduction.

Feature integration is an effective method to compensate the insufficiency of data size. Based on the petroleum geological knowledge, the feature variables of HEI and FI are firstly introduced for ML application in the PNGR prediction model. The new feature variables contain many messages of source rocks and sedimentary facies which can significantly influence the hydrocarbon generation, migration and accumulation processes and therefore can improve the performance of the prediction model.

(2) The PNGR is predicted by classification and regression processes by the combination of SVC and SVR model.

Considering the poor performance by a single regression ML model, the PNGR is predicted first by the hydrocarbon accumulation probability prediction by the SVC model and then the reserve abundance prediction by the SVR model. Based on the result of SVC model, the distribution of favourable zones for hydrocarbon accumulation is determined. Constrained by the range of favourable zones, the accuracy of the reserve abundance prediction by SVR model increases obviously.

4. Result

4.1. Data processing result

In this study, the whole dataset is divided into two parts: one is applied for model training and testing and covers the middle and eastern portions of the study area named zone A; the other is used for validation and covers the western portion of the study area named zone B (Fig. 1).

4.1.1. Training and testing data

Zone A is one of the most favourable hydrocarbon accumulation zones in the Dongpu Depression, with 130 exploration and exploitation wells and 102 discovered hydrocarbon reservoirs. As previously stated, the feature variables selected by this study are characterized by a continuous distribution and are therefore predictable. In this research, a total of 130 × 8 = 1040 pieces of data were collected and used as training and testing data in zone A. Different feature variables are characterized by different distributions. The porosity, permeability and sandstone thickness are characterized by a nearly normal distribution, while the HEI shows unbalanced distributions concentrated at small values, and the distribution of fluid potential is characterized by a uniform distribution (Fig. 12a-g). The target variable of hydrocarbon accumulation probability is determined by oil and gas shows and production signals from exploration and exploitation wells. Among the total 130 wells, 72 are oil wells, and the other 58 are failed wells (Fig. 12h). The distribution of the other target variable, reserve abundance, is characterized by large disparities, ranging from 4.62 to 165.73 mt/km^2, while over 70% of the reserve abundance values range from 0 to 60 mt/km^2 (Fig. 12i).

For the feasibility of further prediction of hydrocarbon accumulation probability and reserve abundance, continuous lateral distributions of the HEI, sandstone thickness, NFI, caprocks and fluid potential are needed. The lateral distributions of these feature variables are obtained

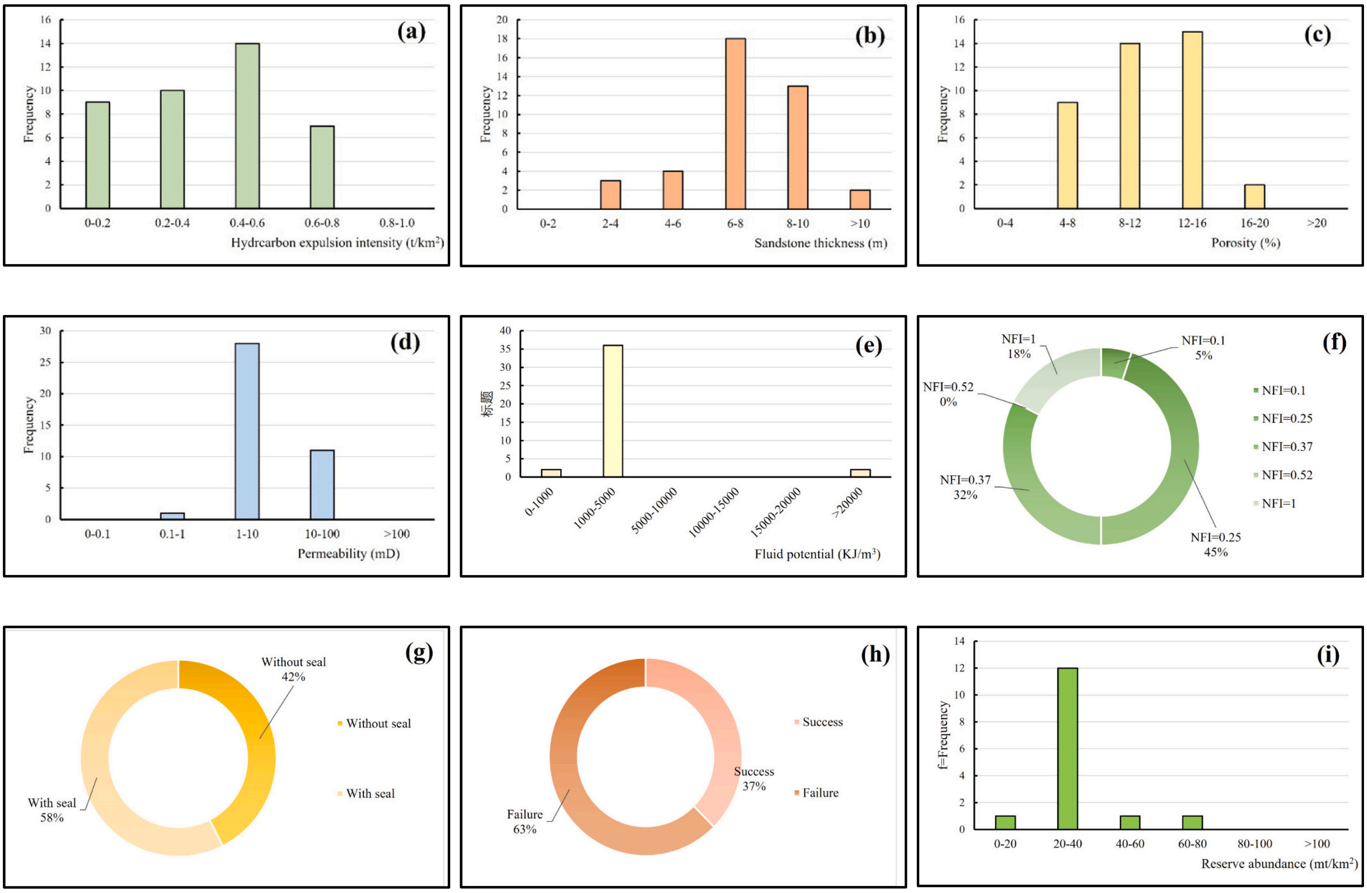


Fig. 14. Composition and distribution of training and testing data. (a) Distribution of the HEI. (b) Distribution of sandstone thickness. (c) Distribution of reservoir porosity. (d) Distribution of permeability. (e) Distribution of fluid potential. (f) Distribution of NFI. (g) Composition of seal development. (h) Composition of hydrocarbon accumulation probability. (i) Distribution of reservoir abundance.

by interpolation based on the measured data or well log and seismic data (Fig. 13). The distribution of HEI is characterized by a gradually increasing trend from the northeast to the southwest of the study area with two hydrocarbon generation centres in the southwest and southeast of the study area (Fig. 13a). The distribution of NFI shows different characteristics from the HEI distribution. The wells with high NFI value are located in the central portion of the study area and characterized by a gradually decreasing trend from the central edge area. While there are also some high NFI which are locate in the western edge of the study area (Fig. 13b). The sandstone thickness distribution is similar to the NFI distribution, which is also characterized by gradually decreasing trend from the central to the edge of the study area. The sandstone thickness of the central portion can reach 15 to 20 m (Fig. 13c). The caprocks is mainly distributed in the northwest, central, and northeast regions. Note that the central portion with high NFI and large sandstone thickness are also characterized by good sealing condition for oil and gas storage (Fig. 13d). The distribution of fluid potential is similar to that of the HEI, the hydrocarbon generation centres are characterized by the largest fluid potential values of approximate 25,000 kJ/m^3. From the northeast to the southwest, the fluid potential shows a gradually decreasing trend (Fig. 13e). In this study, the lateral distribution of the favourable zones for hydrocarbon accumulation and the distribution of reserve abundance are predicted based on the constructed model of SVC and SVR, which are shown and discussed in Section 4.4.

4.1.2. Validating data

Zone B is selected for validation of the evaluation models of SVC and SVR. There are 40 exploration wells and a total of 40 × 8 = 320 pieces of data that have been collected for the validation test. The feature variables for validation are also characterized by different distributions. The HEI and closing height of traps are characterized by unbalanced distributions and concentrated distributions with small values, while the sandstone thickness is characterized by concentrated distributions with larger values. Porosity shows a concentrated distribution ranging from 8 to 16%, permeability is mostly distributed from 1 to 10 mD, and fluid potential mainly ranges from 1000 to 5000 kJ/m^3 (Fig. 14a to g). For the target variable of hydrocarbon accumulation probability, there are 15 oil and gas wells and 25 failed wells (Fig. 14h). The target variable of reserve abundance shows an unbalanced distribution, with over 75% of the data ranging from 0 to 40 mt/km^2 (Fig. 14i). To conclude, the distribution of validation data is very different from the training and testing data, and this situation is suitable for validating the evaluation models.

4.2. Hydrocarbon accumulation probability evaluation model by SVC

4.2.1. Training and validating of the SVC model

The hydrocarbon accumulation prediction model is considered the first step in evaluating or predicting the PNGR potential. In this study, 75% of the data from the original dataset are set as the training data, and the other 25% of the data are set as validating data for searching for the best hyperparameters (Kernel, gamma and C value). In this research, the best hyperparameters are determined by random search. The whole process is achieved by SPSS modeller software. With a random sampling process, a certain composition of training and validating dataset is separated. After the determination of hyperparameters, a model with the highest validating accuracy of 82.14% and the corresponding training accuracy of 95.1% is constructed (Fig. 15). In this model, only 3 oil wells and 5 failed wells are predicted by mistake. The 3 wrongly predicted oil

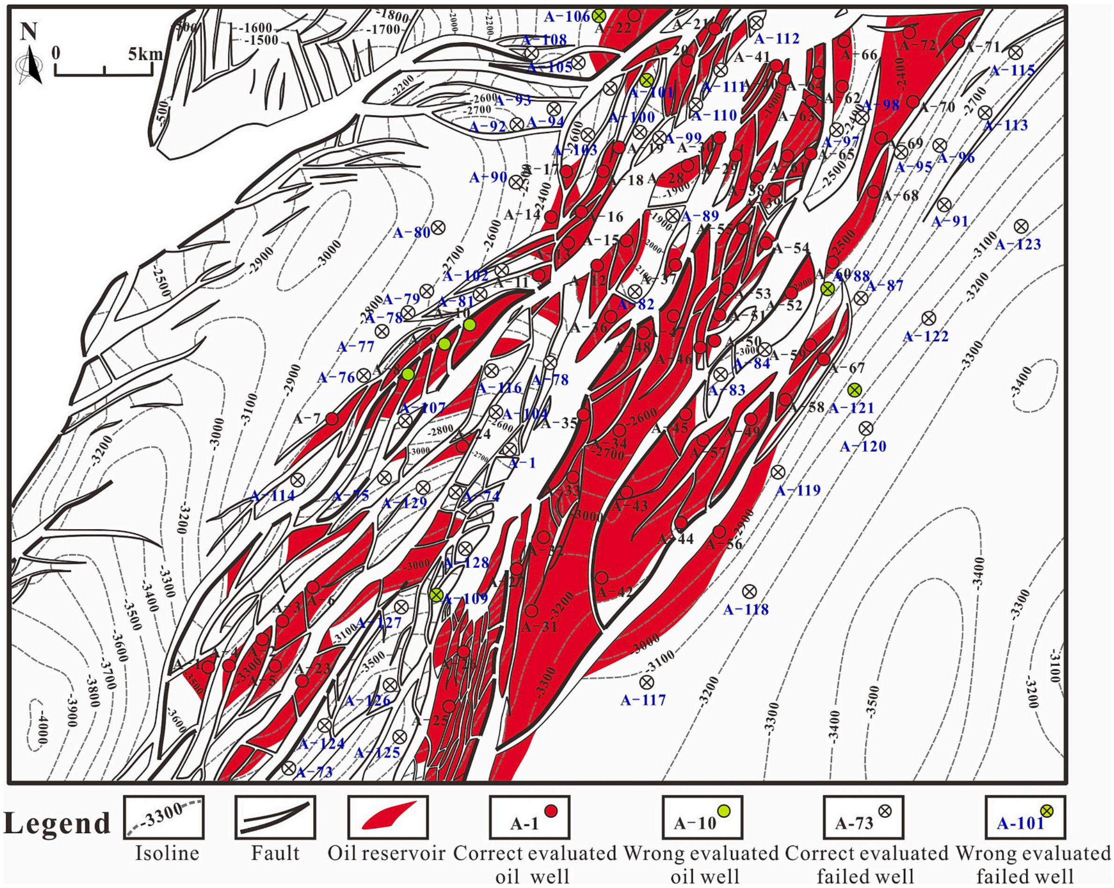


Fig. 15. Training and testing results of hydrocarbon accumulation probability evaluation by the SVC model.

Table 2
Training and testing accuracy for 20 SVC models.

Number	Kernel			Gamma	C value	Validating accuracy (%)	Training accuracy (%)
	RBF	Polynomial	Linear				
1	√			1.02	0.01	82.14	95.1
2	√			1.52	0.16	83.78	96.77
3	√			1.17	0.011	93.55	93.94
4				1.10	0.14	84.21	93.48
5				1.00	0.17	90.48	92.05
6	√			1.69	0.014	100	91.26
7	√			2.28	0.011	83.87	95.96
8	√			1.04	0.016	88.24	95.83
9	√			1.84	0.18	84.62	93.41
10	√			1.15	0.19	96.43	94.12
11			√	/	0.11	87.1	90.91
12		√		1.78	0.015	85.71	92.16
13	√			1.01	0.13	94.29	92.63
14	√			1.06	0.19	92.68	93.26
15	√			1.74	0.10	96.55	92.08
16			√	/	0.017	80	92.63
17			√	/	0.018	82.5	92.22
18		√		1.57	0.09	88.24	93.75
19	√			1.39	0.017	87.86	95.06
20	√			1.00	0.016	96.15	91.35

wells are located in the middle part of the Zone A with a concentrated distribution, while the 5 wrongly predicted failed wells are characterized by a discrete distribution at the edge of the study area.

Considering that our dataset is small and the prediction model is easily influenced by the composition of training and validating data, we constructed 20 models based on different compositions of training and testing data by random sampling. For achieving higher validating accuracy, the kernel function of the RBF, polynomial function and linear function are all used in this study. For searching the best gamma value and C value, the random search is applied in the hyperparameter determination processes. The training accuracy of the 20 models ranges from 90.91% to 96.77%, and the validating accuracy of the models ranges from 80.0% to 100.0%. Table 2 shows the constructing result of the 20 models with the best validating accuracy after the determination

Table 3
Evaluation metric values of SVC models for the testing process.

Number	Precision	Sensitivity	Specificity	NPV*	F1-score	Testing accuracy (%)
1	0.88	0.83	0.80	0.73	0.86	82.14
2	0.78	0.95	0.72	0.93	0.86	83.78
3	0.88	1.00	0.88	1.00	0.93	93.55
4	0.75	1.00	0.70	1.00	0.86	84.21
5	0.95	0.88	0.94	0.85	0.91	90.48
6	1.00	1.00	1.00	1.00	1.00	100
7	0.93	0.82	0.93	0.81	0.88	83.87
8	0.91	0.91	0.83	0.83	0.91	88.24
9	0.86	0.86	0.82	0.82	0.86	84.62
10	0.95	1.00	0.90	1.00	0.97	96.43
11	0.79	1.00	0.75	1.00	0.88	87.1
12	0.85	0.94	0.70	0.88	0.89	85.71
13	0.94	0.94	0.94	0.94	0.94	94.29
14	0.89	1.00	0.82	1.00	0.94	92.68
15	1.00	0.93	1.00	0.94	0.96	96.55
16	0.80	0.84	0.75	0.80	0.82	80
17	0.82	0.96	0.69	0.92	0.88	82.5
18	0.87	0.95	0.77	0.91	0.91	88.24
19	0.87	0.93	0.81	0.89	0.90	87.86
20	0.93	1.00	0.92	1.00	0.97	96.15
Average	0.88	0.94	0.83	0.91	0.91	88.92

NPV*: Negative predictive value.

Table 4
Evaluation metric values of SVC models for the testing processes.

Number	Precision	Sensitivity	Specificity	NPV	F1-score	Testing accuracy (%)
6	0.55	0.40	0.80	0.69	0.46	65.0
10	0.58	0.47	0.80	0.71	0.52	67.5
14	0.63	0.67	0.76	0.79	0.65	72.5
15	0.44	0.27	0.80	0.65	0.33	60.0
20	0.64	0.60	0.80	0.77	0.62	72.5
Average	0.57	0.48	0.79	0.72	0.52	67.5

of the hyperparameters (the corresponding training accuracies of the models are also shown). The determination result of Kernel function shows the obvious advantage of the RBF among the three functions. The gamma value ranges from 1.00 to 2.62, while the C value ranges from 0.01to 0.19. Note that when the linear function is selected as the kernel, the gamma value is not available. With the constraint of gamma and C values, the overfitting of models is corrected with high validating accuracy. Furthermore, the result also shows the influence of data composition on the performance of the model (Table 2).

4.2.2. *Testing of the SVC model*

4.2.2.1. Model testing by evaluation metrics. As previously stated, the evaluation metrics are applied in this research for further analysing the classification performance of the SVC models. The results of the evaluation metrics show that the precision value of the 20 SVC models ranges from 0.86 to 1.00, with an average of 0.88; the sensitivity value ranges from 0.82 to 1.00, with an average of 0.94; the specificity value ranges from 0.69 to 1.00, with an average of 0.83; and the negative predictive value ranges from 0.73 to 1.00, with an average of 0.91. The F1-score value ranges from 0.82 to 1.00, with an average of 0.91. The SVC models show better performance on Sensitivity and NPV than that on precision and specificity, which means most SVC models are better at oil well location prediction than the failed oil well location prediction. The possible reason for the performance is that the amount of oil well is far more than the failed well. In total, the SVC models show good performance in the evaluation metric parameters (Table 3).

The value of the F1-score can reflect the performance of the SVC model in a more subjective way [94]. Therefore, the F1-score is selected as the evaluation criterion for the SVC model. Accordingly, models 6, 10, 14, 15 and 20 with the highest F1-scores are selected as the best models for the validation test using the dataset of zone B.

4.2.2.2. Model testing by hold-out test dataset in zone B. In this study, zone B is selected for testing the performance of the SVC model. The results of the best 5 SVC models with the highest F1-scores are applied for the validation test. The results show that the testing accuracies of the 5 models are 65%, 67.5%, 72.5%, 60% and 72.5%, with an average of 67.5% (Table 4). The results of the evaluation metrics show that the specificity and NPV are better than the precision and sensitivity for all the models, which also indicates the better performance of these models to predict oil well location than the failed ones. Models 14 and 20 show the best performances of hydrocarbon accumulation probability prediction with F1-scores of 0.65 and 0.62, respectively. The successfully evaluated data points are distributed in the middle part of zone B, while the incorrectly evaluated data points are distributed in the northwestern and southwestern portions. The 5 wrongly predicted oil wells (Well B-33, B-37, B-38, B-39 and B-40) are characterized by a concentrated distribution on the north edge of the Zone B for all the 5 models. The failed wells of Well B-1 to Well B-6, which are concentrated on the southwest part of Zone B, are also easily to be predicted by mistake (Fig. 16). Comparing to the training and validating dataset, the testing data is characterized by larger proportion of failed wells, which may lead to the deviation of the predicting results.

4.3. *Reserve abundance evaluation model by SVR*

4.3.1. *Training and validating of the SVR model*

The reserve abundance prediction model is constructed by the SVR algorithm with the kernel function, which is applied to enhance the performance of the SVR model in regression. By random sampling, 10 pairs of SVR models are constructed. For achieving higher validating accuracy, the determination of hyperparameters is still required. The RBF, polynomial function and linear function are used for kernel function selection. Also, the hyperparameters of gamma and C value for each model are determined by random search.

The training and validating results show large disparity of the correlation coefficients (Ccoe) with different dataset and hyperparameters. The kernel functions of RBF and linear perform better than the polynomial. With the constraint of the gamma and C value, the final validating Ccoe ranges from 0.564 to 0.862 with an average of 0.7042, and the training Ccoe ranges from 0.566 to 0.721 with an average of 0.6395, which is slightly lower than that of the validating Ccoe (Table 5). On the one hand, the higher Validating Ccoe indicates the effect of the hyperparameter of gamma and C value for overcoming the overfitting of the model which will lead to better training performance and worse generalization performance of the model. On the other hand, the unstable performances for the 10 models also show the large influence of the original dataset on the final performance of the models, especially when the size of the dataset is not large enough. The models with RBF have an average validating Ccoe of 0.7403 and training Ccoe of 0.6514, while the models with Polynomial are characterized by the average validating Ccoe of 0.6200 and training Ccoe of 0.6117. The Linear function has not provided a good model with high validating Ccoe, which indicates the nonlinear characteristic of the problem. Considering both the stability and the reliability of the predicting model, the fifth model with the best training Ccoe and the secondary highest validating Ccoe is determined as the best model for RA evaluation and prediction. The training and testing results of the fifth model using the RBF are shown in Fig. 17.

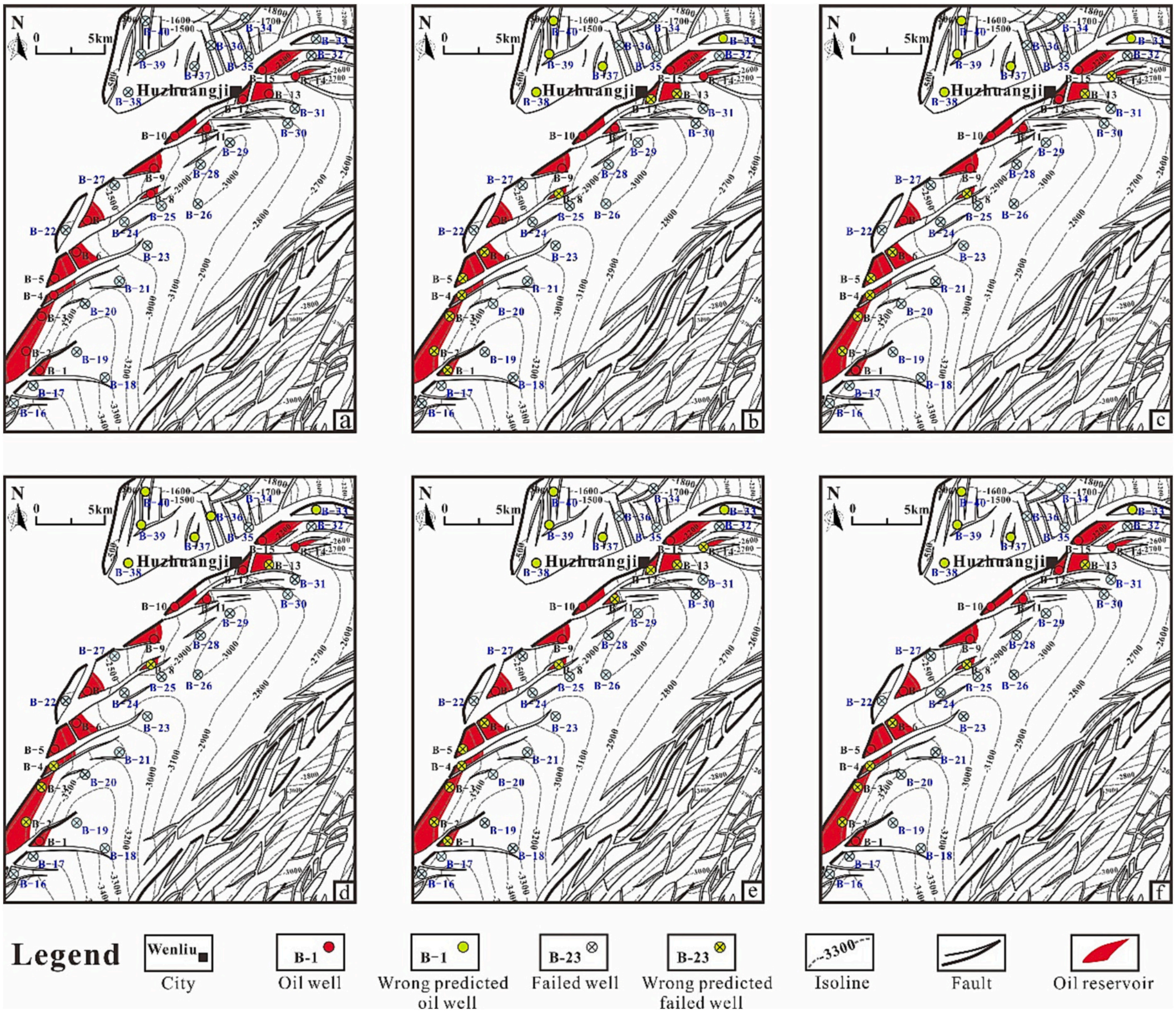


Fig. 16. Testing results in zone B by the best 5 SVC models. (a) Actual distribution of oil wells and failed wells in zone B. (b) Validation test result of the No. 6 SVC model. (c) Validation test result of the No. 10 SVC model. (d) Validation test result of the No. 14 SVC model. (e) Validation test result of the No. 15 SVC model. (f) Validation test result of the No. 20 SVC model.

Table 5
Correlation coefficients of the SVR models with different kernel functions.

Number	Kernel			Gamma	C value	Validating Ccoe	Training Ccoe
	RBF	Polynomial	Linear				
1	√			0.17	0.011	0.791	0.622
2	√			0.11	0.016	0.862	0.596
3		√		0.12	0.014	0.622	0.678
4	√			0.12	0.014	0.613	0.651
5	√			0.11	0.017	0.803	0.721
6		√		0.19	0.010	0.564	0.591
7	√			0.12	0.012	0.671	0.682
8	√			0.14	0.018	0.739	0.611
9	√			0.16	0.011	0.703	0.677
10		√		0.17	0.014	0.674	0.566

4.3.2. *Testing results of the SVR model*

4.3.2.1. Model testing by 4-fold cross validation. For eliminating the influence of dataset on the model performance, the 4-fold cross validation is applied. The whole dataset is equally divided into 4 subgroups, and 4 SVR models are constructed by the application of RBF with each subgroup set as the testing data. The result of the 4 models shows similar Ccoe with an average of 0.663. The predicted reserve abundance by the 4 models also show the similar distribution to the model by random sampling (Fig. 18). In addition, the average Ccoe of the 4 models is close to the average Ccoe of the SVR model by random sampling, which indicates good reliability and stability performance of the SVR model for reserve abundance prediction.

4.3.2.2. Model testing by hold-out dataset in zone B. In this study, zone B is also selected for testing the performance of the SVR model. The

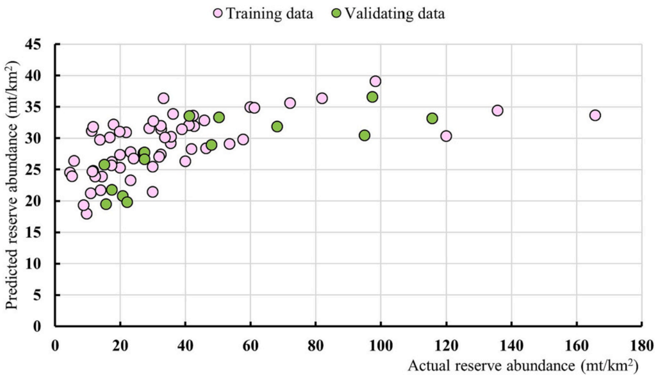


Fig. 17. Training and validating results of reserve abundance evaluation by the SVR model.

selected SVR model shows good performance in evaluating the reserve abundance for 15 data points with a Ccoe of 0.744. The testing result shows that the SVR model can provide a good estimate of the distribution of reserve abundance and performs better in the evaluation of relatively small reserve abundance with values close to the real ones (Fig. 19). However, the model cannot precisely evaluate a large reserve abundance in terms of absolute value. The relatively weak ability for the model to provide the precise prediction of the large reserve abundance attributes to the composition of the dataset, in which the reservoirs with larger reserve abundance (>50mt/km^2) are rare (Fig. 14h).

4.4. *Prediction performance of PNGR potential by SVC and SVR*

The models constructed show good performance with respect to the dataset from exploration and exploitation wells. However, during the PNGR exploration process, the prediction of PNGR potential must be performed without enough well data, which means that ML models should provide the lateral continuous distribution of PNGR potential, which is hard to achieve. Based on the method of feature variable characterization introduced in this study, prediction of the lateral continuous distribution of hydrocarbon accumulation probability and reserve abundance can be accomplished. To further enhance the prediction accuracy of the PNGR potential, we used the SVC and SVR models in combination to first predict the hydrocarbon accumulation probability and then the reserve abundance.

Based on the lateral continuous feature variable distribution, the hydrocarbon accumulation probability of the study area is predicted. The whole study area is gridded to 191 × 140 = 26,740 prediction units, and the values of the feature variables in each prediction unit are obtained by interpolation (Fig. 20a). The prediction results show that most of the discovered oil reservoirs are successfully predicted by the 14th SVC model selected in this study. The predicted oil reservoir range is also similar to the actual range (Fig. 20b). The final statistical result shows that the prediction accuracy of the hydrocarbon accumulation probability reaches 84.3%.

The reserve abundance prediction is based on the hydrocarbon accumulation probability prediction result. In other words, only the area with positive hydrocarbon accumulation probability prediction results is selected as the data points of reserve abundance prediction. In this study, the predicted reserve abundance results for the study area also show a similar distribution to the actual reserve abundance distribution (Fig. 20c). Based on these prediction results, the value of reserve abundance for every grid can be provided for comparison and further determination of exploration and exploitation strategies (Fig. 20d).

5. Discussion

5.1. *Effectiveness evaluation of the combination of SVC and SVR*

In this study, the PNGR potential is predicted first by hydrocarbon accumulation probability prediction by SVC and then by reserve

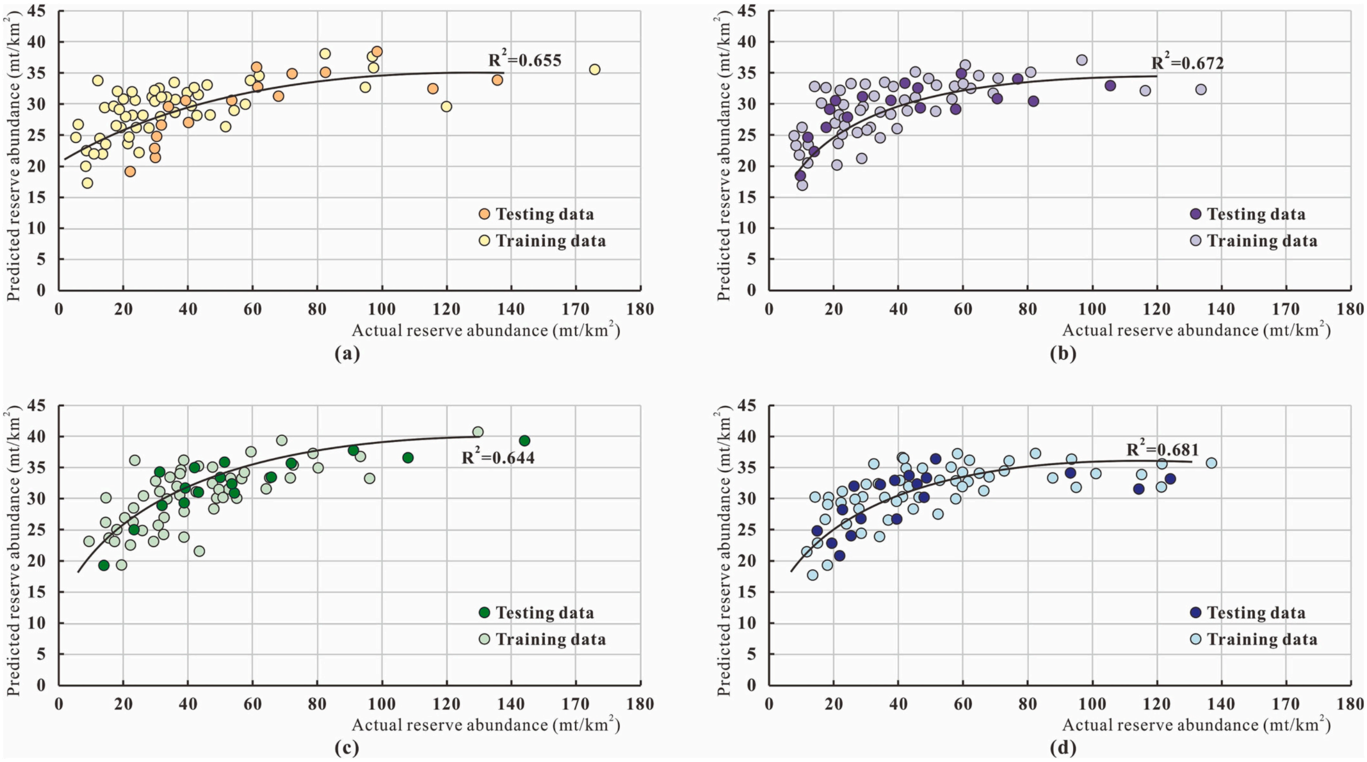


Fig. 18. 4-fold cross validation test results for the reserve abundance prediction models by SVR. Figure (a) to (d) shows the 4 SVR models of the 4-fold cross validation.

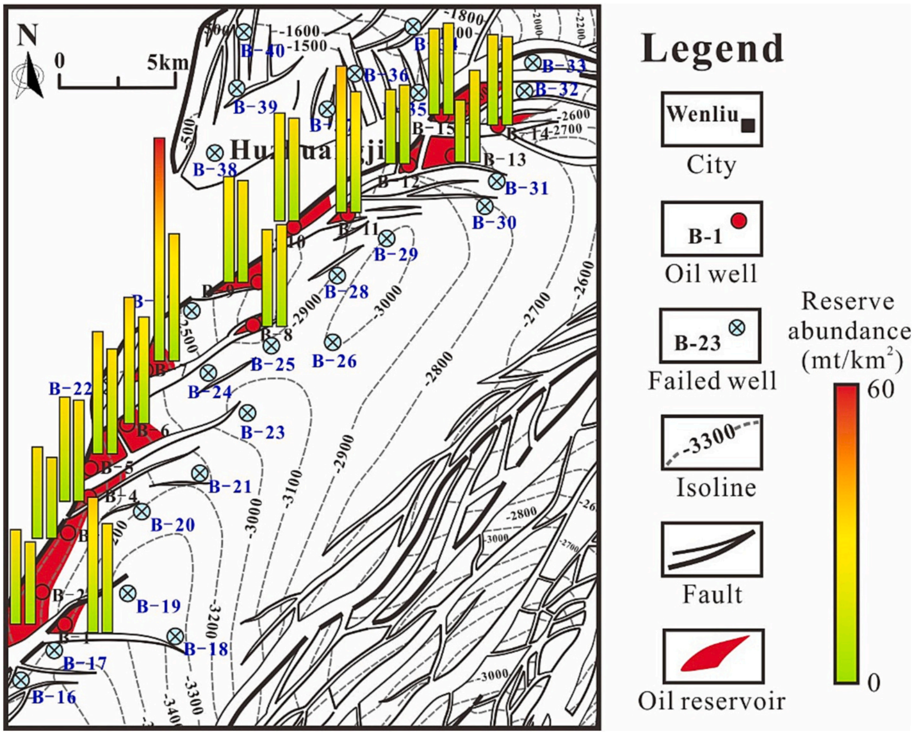


Fig. 19. Testing result of zone B by the selected SVR model.

abundance prediction by SVR. One of the most important reasons for this two-step prediction process is that the PNGR potential prediction only by the regression process always leads to a low accuracy, which will increase the risk of a failed well. However, the SVC model shows good performance. Constrained by the result of the SVC model, RA prediction by SVR is accomplished in an area with a high probability of hydrocarbon accumulation.

To prove the effectiveness of the combination of the SVC and SVR models in this study, the reserve abundance prediction model without the constraint of SVC results is also constructed. The result of reserve abundance prediction shows that the favourable hydrocarbon accumulation zone with different RA values is much larger than the actual range of oil reservoirs and the prediction results by the combination of SVC and SVR models (Fig. 21). Additionally, the model also provides a number of reserve abundance values in the area of actual failed wells (Fig. 21b and Fig. 1). Furthermore, the model using only SVR failed to provide an reserve abundance distribution similar to that of the model based on SVC and SVR, which means that the model cannot provide the best location of exploration wells for obtaining the most PNGR (Fig. 21b). To conclude, without the constraint of hydrocarbon probability prediction by SVC, the regression model overestimates the PNGR potential in the study area, which enhances the risk of failed wells.

5.2. Comparison between multiple machine learning algorithms

5.2.1. Classification model comparison

Although previous studies have shown that the SVC algorithm is one of the best algorithms for solving classification problems by determining the best class boundaries (SVMs), there are also several popular algorithms for dealing with classification problems. In this study, hydrocarbon accumulation prediction models based on decision tree (DT), random forest (RF), naïve Bayes (NB) and K-nearest neighbour (KNN) are also constructed for comparison.

Using the same dataset, 20 models for each algorithm are constructed and tested. The results show that the performance of the SVC model is indeed the best, with the best performance in all the evaluation parameters (Fig. 22). The KNN model and NB model are characterized by good stability with an accuracy $>60\%$, while the maximum accuracy of the KNN and NB model can reach only 80.4% and 81.8%, respectively. The DT and RF models are characterized by instability. In other words, the maximum accuracy of the DT and RF models can reach 100%; however, the minimum accuracies of these two models are also $<50\%$ and 60%, respectively.

5.2.1.1. Comparison between KNN and SVC. The KNN algorithm is a typical unsupervised learning algorithm which divides data into several groups based on the requirements of the practical problems. The KNN model is always used for the classification of data without label and thus the classification result of the KNN may or may not match the actual situation of the original data (Fig. 23a). In this study, the data for the classification processes has the label, which means KNN is unable to exert its advantages of unsupervised learning comparing to SVC or other supervised learning algorithms. Furthermore, there are only two classification results for the hydrocarbon accumulation probability, which also increases the difficulty for KNN to make a good classification of the dataset which contains 7 feature variables. The comparison between the KNN and SVC shows that almost every SVC model has higher testing accuracy than the corresponding KNN model with the same dataset, which indicates the advantage of SVC model (Fig. 23b).

5.2.1.2. Comparison between DT and SVC. The DT algorithm is one of the most popular supervised learning algorithms which is both suitable for classification and regression problems ([140]). In this research, the binary classification tree is applied for hydrocarbon accumulation probability prediction (Fig. 24a). The hyperparameters including the maximum of the layer of a decision tree, the minimum of the samples in a subgroup and the minimum impurity decrease are determined for improving the performance of DT by overfitting calibration ([49,121]). The advantage and disadvantage of the DT algorithm are equally prominent: hhe DT model may provide an excellent performance with the testing accuracy as high as 100% (4 of 20 models in this research); however, 5 models show the test accuracy of $<50\%$ with the minimum

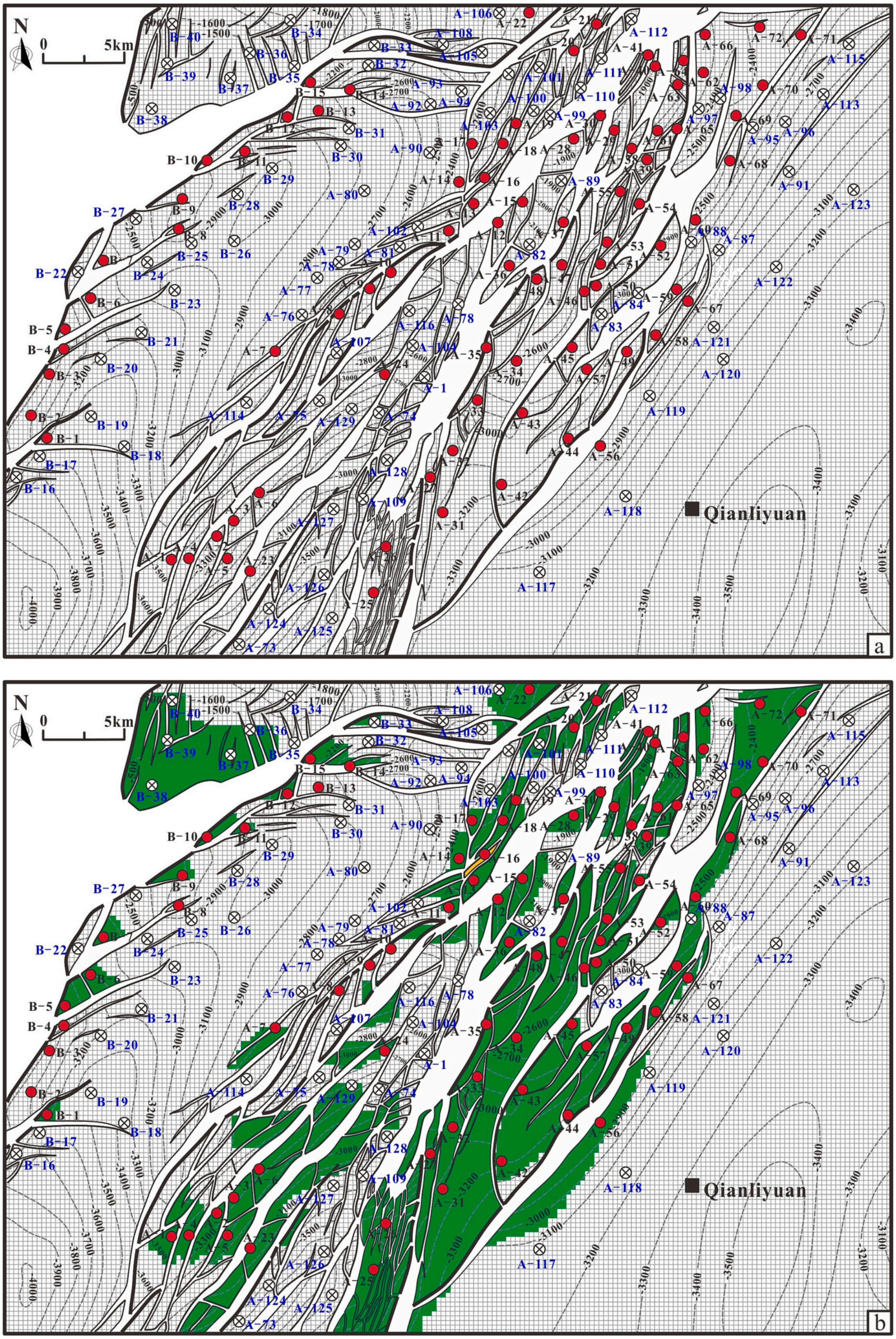


Fig. 20. PNGR potential prediction processes. (a) Gridding of the study area. (b) Hydrocarbon accumulation probability prediction by SVC. (c) RA prediction by SVR. (d) Precise prediction of reserve abundance for the selected grid.

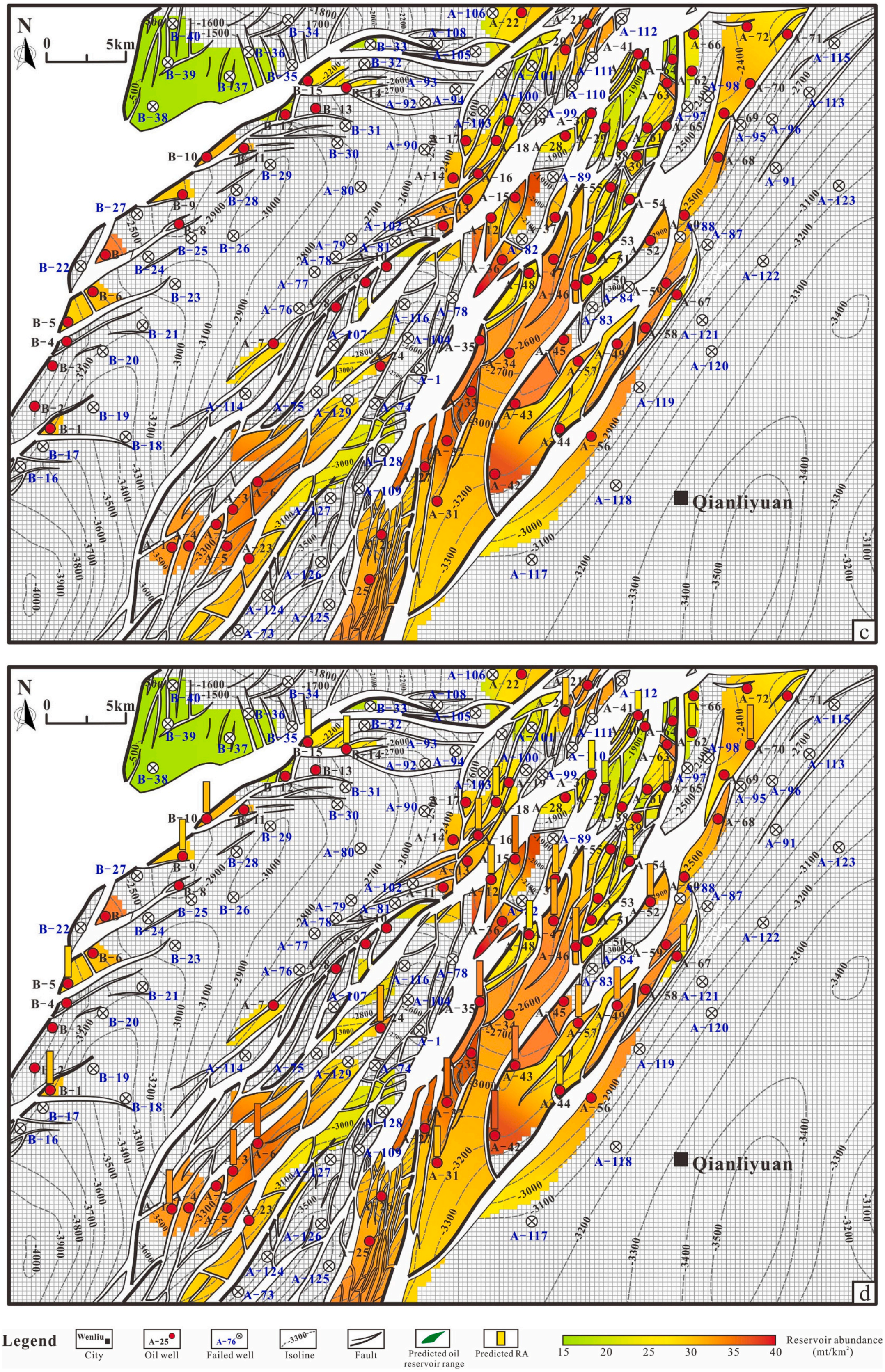


Fig. 20. (*continued*).

of approximate 35%, which indicates the intensive overfitting of the models (Fig. 24b). In other words, the performance of DT model is unstable and can be easily affected by few features of the dataset ([141]).

5.2.1.3. Comparison between RF and SVC. The RF algorithm is an ensemble learning algorithm based on decision tree. Comparing to the DT algorithm, the RF algorithm can significantly improve the accuracy

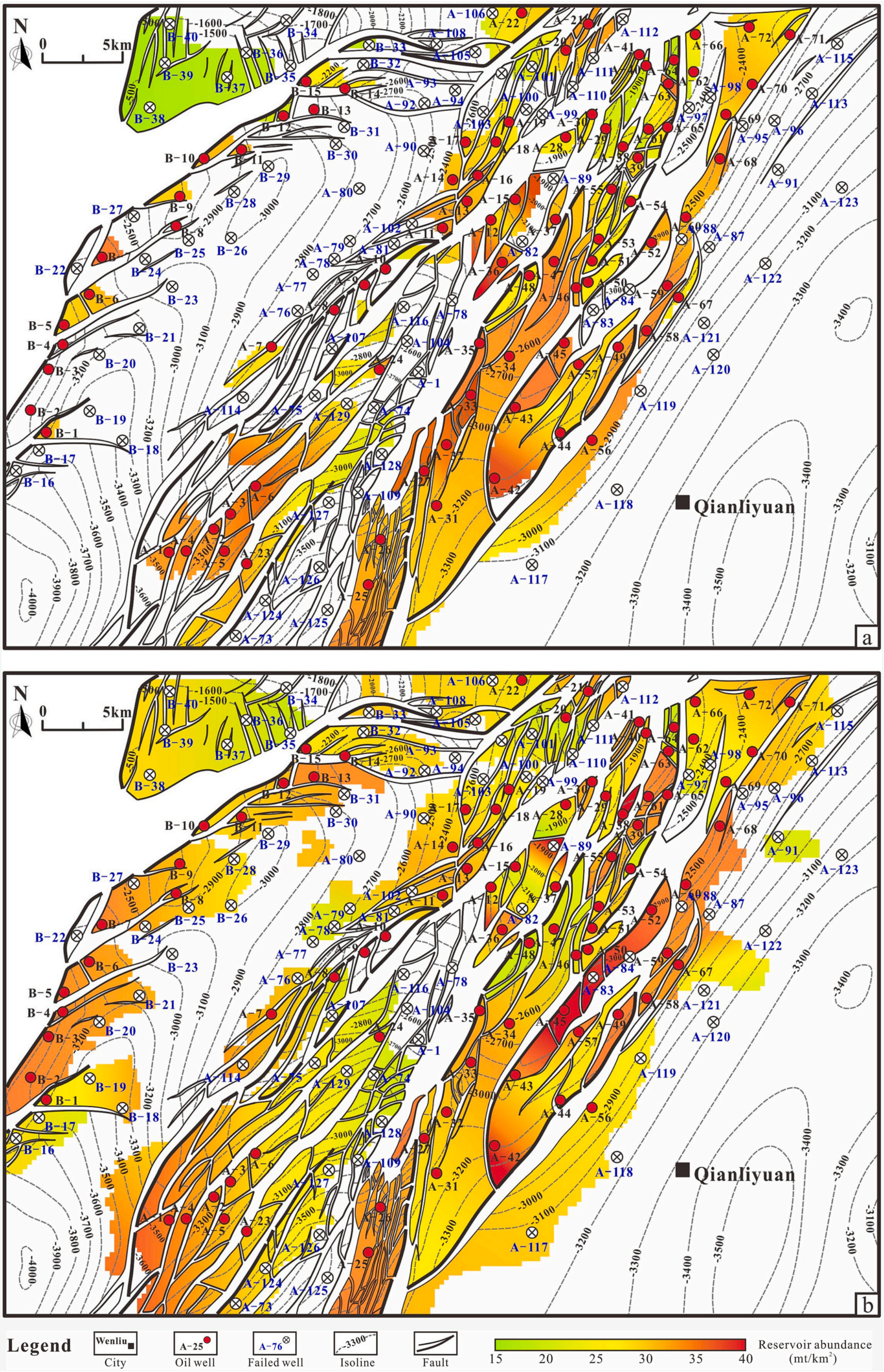


Fig. 21. Reserve abundance prediction by (a) the combination of SVC and SVR models and (b) single SVR model.

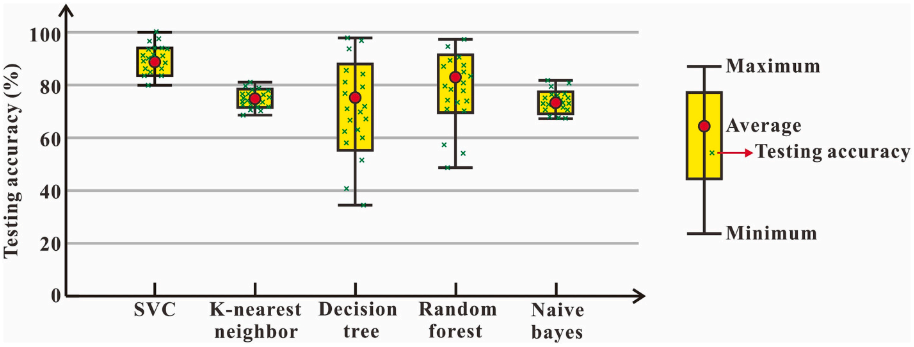


Fig. 22. Performance of different classification models in hydrocarbon accumulation probability prediction.

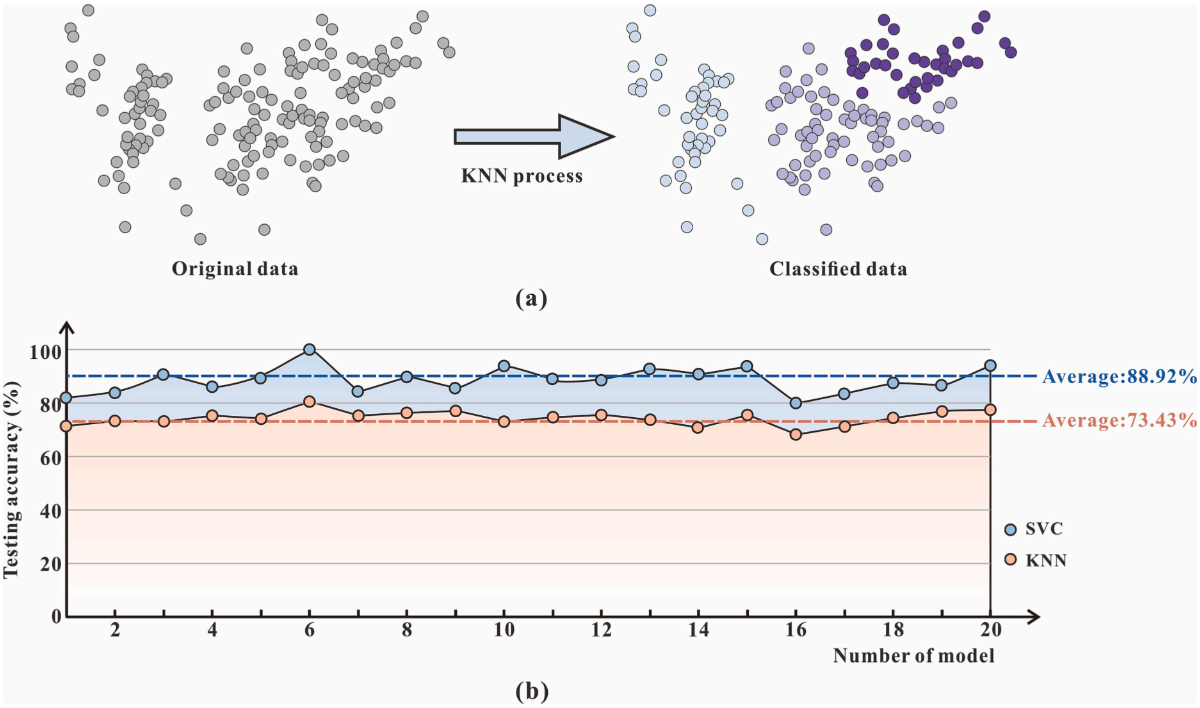


Fig. 23. The comparison between the KNN and SVC algorithms for the prediction of hydrocarbon accumulation probability. (a) The unsupervised process of KNN without labels; (b) The prediction results of the KNN and SVC algorithms for 20 models using the same data.

and stability by making decision from multiple single prediction DT models [116,122,123]. In this research, every RF model is constructed by 10 individual binary decision trees (Fig. 25a). The testing result shows the obvious improvement of the stability by RF models comparing to the corresponding DT models (Fig. 22). Almost all the models show the testing accuracy of higher than 60%. However, the SVC models still show better performance than that of the corresponding RF models in general by the higher average testing accuracy (Fig. 25b). On the one hand, the overfitting may still exist for the trees which compose the RF and may lead to the deviation of the classification; on the other hand, the size of the dataset is small for an ensemble learning algorithm to show the advantage.

5.2.1.4. Comparison between NB and SVC. The NB algorithm is a supervised learning algorithm that directly measures the probability relationship between labels and features [124]. Comparing to the traditional Bayes method, the naïve bayes algorithm presumes that all the feature variables are independent of each other [125]. While as preciously stated, the feature variables in this research are related and can influence each other. Therefore, it is reasonable that the NB model cannot provide a good performance for the hydrocarbon accumulation probability prediction with the lowest testing accuracy maximum of 79.4%. Comparing to the NB model, the SVC models also provide better performance in both testing accuracy and stability (Fig. 26). To conclude, the SVC model is the most suitable model for the mode construction of hydrocarbon accumulation in this study.

5.2.2. *Regression model comparison*

In addition to the SVR algorithm, there are several other popular algorithms for solving regression problems. In this study, we also selected several regression algorithms, including MLR, PR and GBDT, to compare to the SVR model. The training and testing datasets are also the same as those of the SVR model used in this study. The result shows that the SVR model is also the best model, with Ccoe values ranging from 0.502 to 0.767 with an average of 0.602. Among the other models, the GDBT model also performs well, with the highest Ccoe of 0.735 and average Ccoe of 0.582. The MLR and PR models are characterized by poor performance with low Ccoe values (Fig. 27). Although the models show large disparities, the results of the comparison represent the performance of the models only in terms of the problems examined in this study instead of all the actual problems.

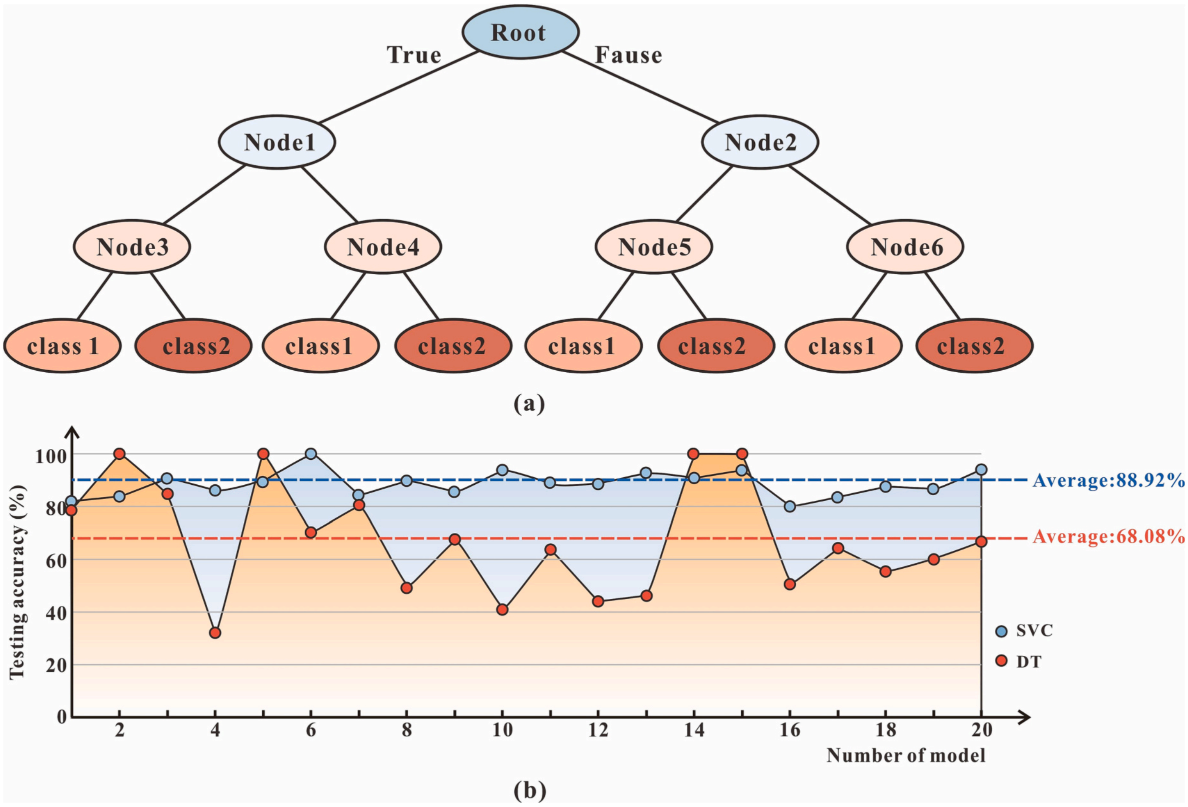


Fig. 24. The comparison between the DT and SVC algorithms for the prediction of hydrocarbon accumulation probability. (a) The structure of a binary decision tree; (b) The prediction results of the DT and SVC models for 20 algorithms using the same data.

5.2.2.1. Comparison between MLR and SVR. The MLR algorithm is the simplest regression algorithm which characterizes the linear correlation between an independent variable and multiple dependent variables. The MRL model has a wide application range except for the multicollinearity problem, which means the explanatory variables are characterized by precisely or highly correlated relationship [127]. In this research such precisely correlated relationship does not exist between the variables. The testing result shows that the testing Ccoes of most of the MLR models are lower than the corresponding SVR models with a relatively large disparity (Fig. 28). 8 of the 10 models are characterized by the testing correlation coefficient lower than 0.5 with the minimum of 0.272. The result also indicates the nonlinear characteristic of the reserve abundance prediction process.

5.2.2.2. Comparison between PR and SVR. The PR algorithm is an improvement of the linear regression algorithm and can be applied to deal with the nonlinear regression problems by ascending dimension [128]. The final polynomial will be complex without unified expression by ascending dimension and multiple features in the dataset. The testing result shows that the PR models perform better than the MLR models in average, which proves the advantage of the nonlinear algorithm for the reserve abundance prediction in this research. While the SVR models also perform better than the PR models for most groups (Fig. 29). The average testing Ccoe of the PR models is 0.564, which is also lower for the requirement of the RA prediction in practical.

5.2.2.3. Comparison between GDBT and SVR. The GDBT algorithm is an ensemble learning algorithm which is applied for dealing with regression problems [130,132]. While different from RF, the GBDT uses boosting instead of bagging to construct an ensemble learning model. In the GBDT, every decision tree model is constructed for improving the prior DT model [129,131]. In this situation, the single DT model in the GBDT is not independent but related. The testing results show that the GBDT models have the similar performance both in the accuracy and the stability with the similar average Ccoe and the Ccoe distribution (Fig. 30). Therefore, the GBDT may have the potential to be a better choice because the advantages of the ensemble learning model may be more obvious with the increase of the data size. However, the computation of the ensemble learning model will also increase rapidly, so the SVR model with relatively small computation and therefore high efficiency is still competitive, especially when the data size is difficult to be enlarged significantly in practical.

5.2.2.4. Comparison between the linear and nonlinear regression models. The regression algorithms can be simply divided into two types, which are liner regression and nonlinear regression. While, most of the problems which need the ML methods are nonlinear problems in practical. Therefore, the nonlinear seems to have better applicability. The nonlinear regression models can always show good performance on linear data, but the linear regression models always have poor performance on nonlinear data (Fig. 31a and b). However, the linear regression also has its advantage of the higher efficiency and smaller computation than the nonlinear regression algorithms. Different from most ML algorithms, the SVM can provide both linear and nonlinear models with the different kernel functions (Fig. 31c). Therefore, the SVM algorithm has the potential to provide the evaluation and prediction model with both high efficiency and high accuracy, which makes the SVM the preferred algorithm for solving practical problems.

5.3. Improvement of the model

There are multiple factors that influence the performance and practicability of an ML model. In other words, the performance of the ML model can be improved in many ways. In this study, we have checked the whole process of PNGR resource prediction and provided the following ways to improve the performance of the model:

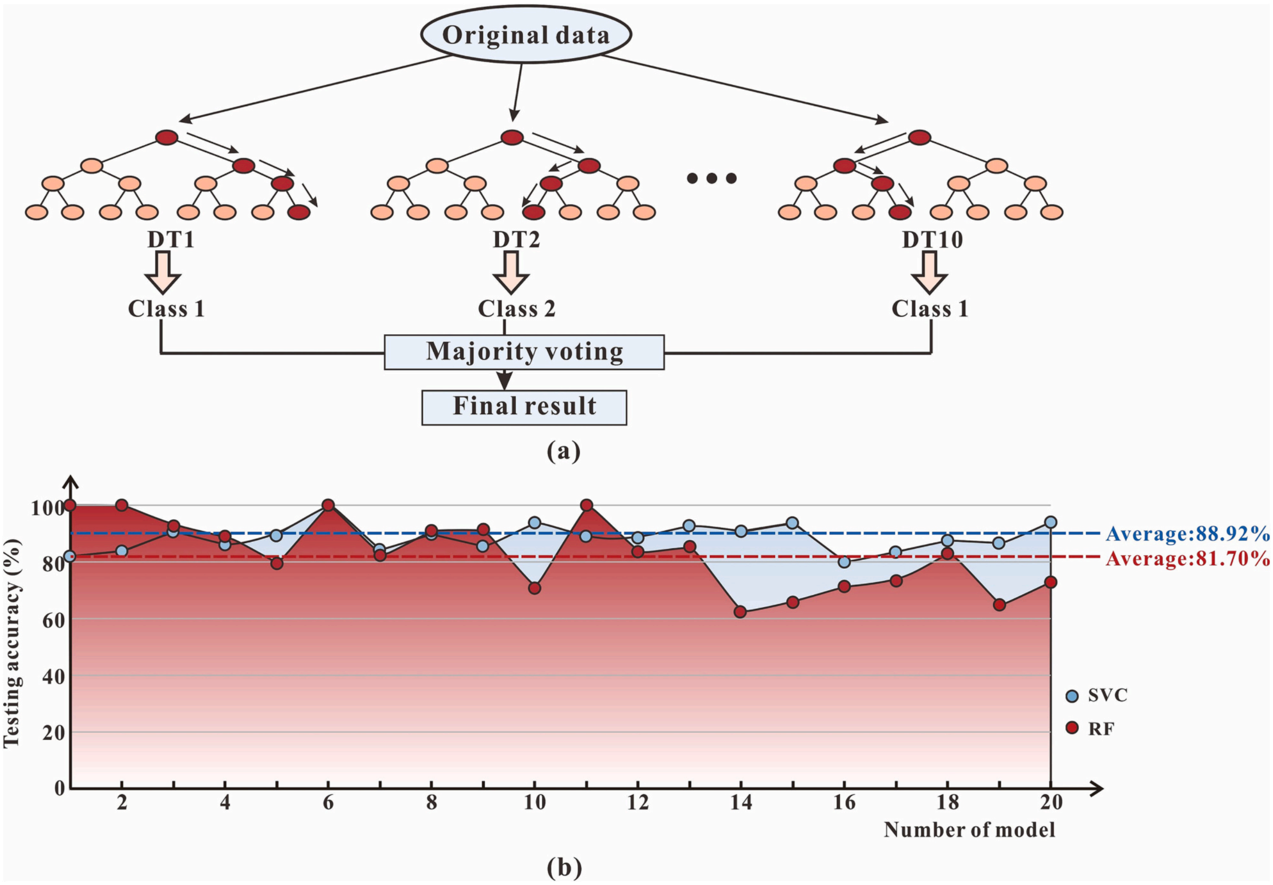


Fig. 25. The comparison between the RF and SVC algorithms for the prediction of hydrocarbon accumulation probability. (a) The structure of a RF constructed by 10 binary decision trees; (b) The prediction results of the RF and SVC algorithms for 20 models using the same data.

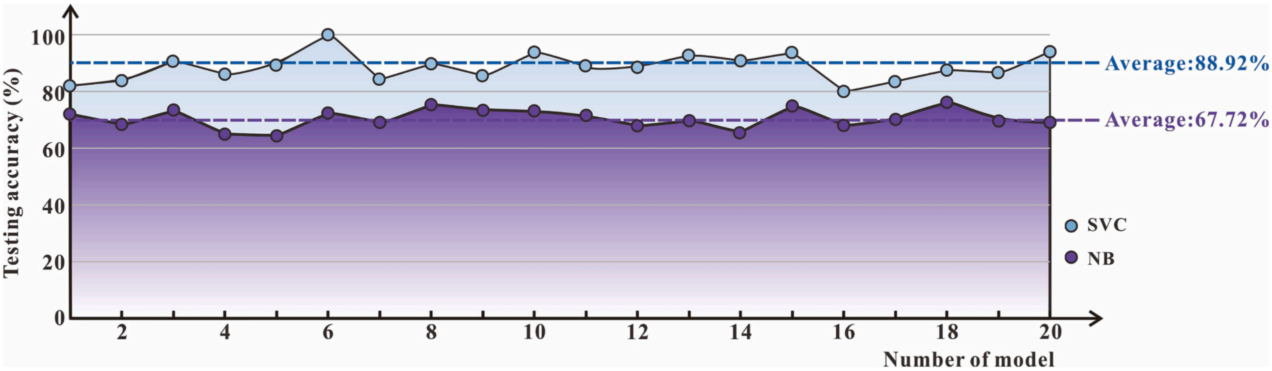


Fig. 26. The comparison between the NB and SVC algorithms for the prediction of hydrocarbon accumulation probability by 20 couples of models.

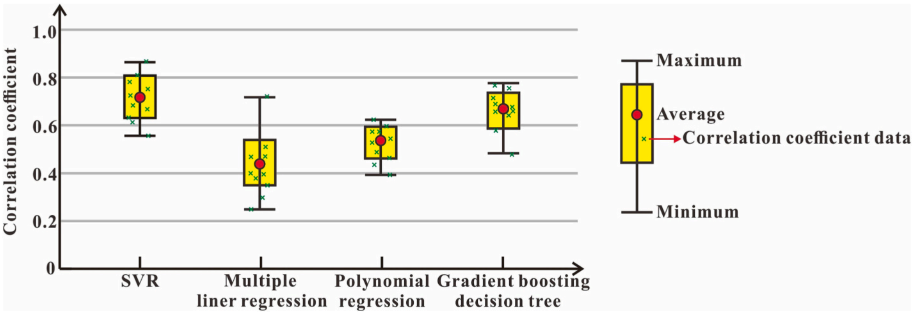


Fig. 27. Performance of different regression models in reserve abundance prediction.

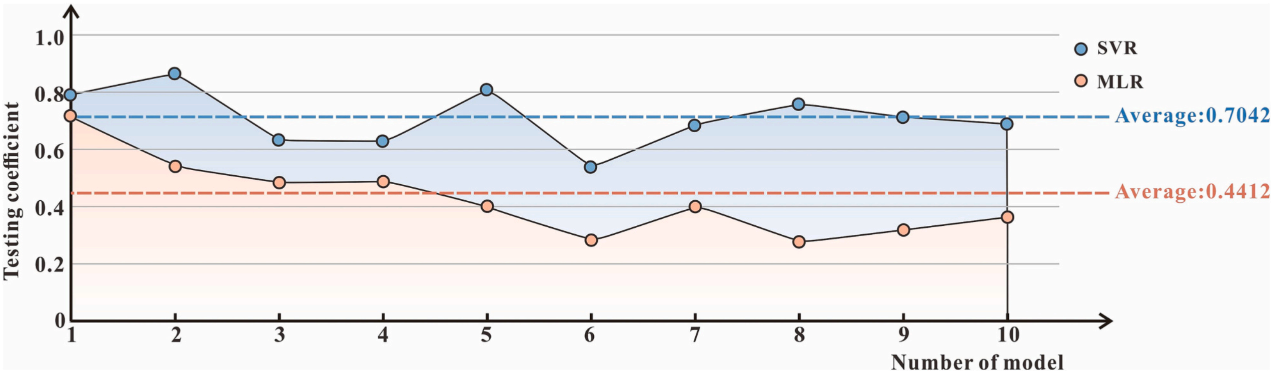


Fig. 28. The comparison between the MLR and SVR algorithms for the prediction of reserve abundance by 10 couples of models.

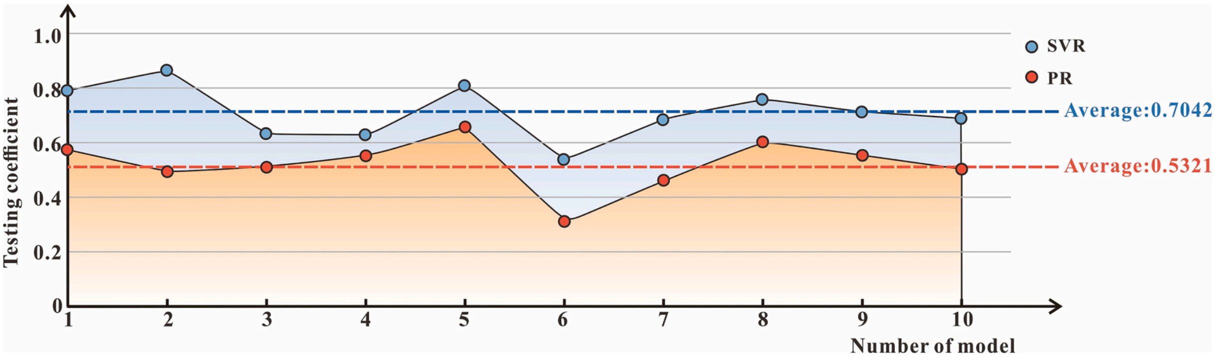


Fig. 29. The comparison between the PR and SVR algorithms for the prediction of reserve abundance by 10 couples of models.

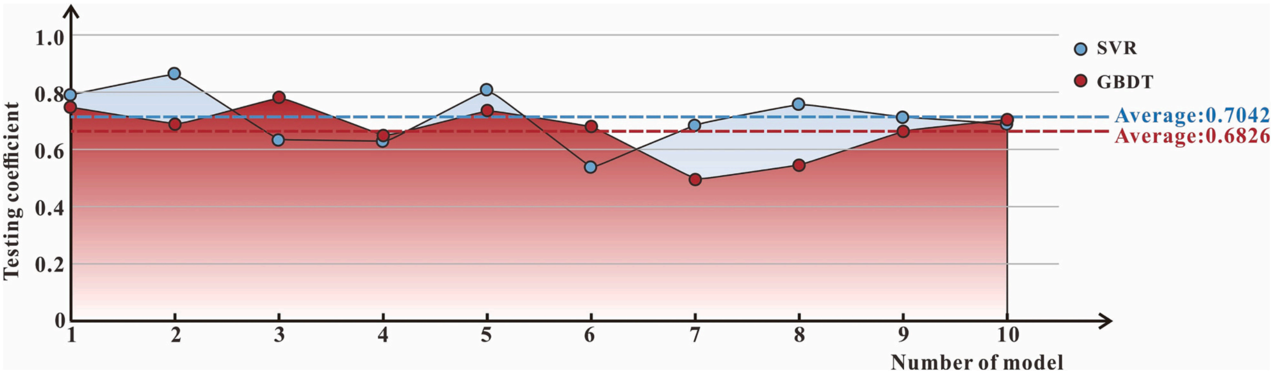


Fig. 30. The comparison between the GBDT and SVR algorithms for the prediction of reserve abundance by 10 couples of models.

(1) Enlarging the data size. The data are the foundation of the ML process. Compared to the application of ML in medical treatment, transportation, and commerce, the data size of the petroleum exploration industry is small. Owing to the lack of a larger reserve abundance value for model training, the final prediction results of reserve abundance show poor performance in PNGR potential prediction for oil and gas reservoirs with large reserve abundance. If the data size is large enough, the evaluation and prediction performance of the ML model will be better and could meet the requirements of precise petroleum exploration and exploitation.

(2) Optimizing the feature variable construction process. Feature variables based on petroleum geological knowledge can compensate for the influence of a small amount of data. An appropriate scheme of feature variable construction can make the dataset reflect more useful information and therefore improve the performance of the ML model. Furthermore, the advanced and reasonable construction of feature variables can also compensate for the defects of insufficient data.

(3) Using the ensemble learning model. The ensemble learning model is proved to improve the accuracy and stability of the single evaluating and predicting models especially when the data size is large enough. When the single model by SVC and SVR in this research cannot meet the requirement, an ensemble learning model based on multiple SVC or SVR models may improve the performance of the evaluating and predicting work.

Furthermore, this model can be employed in other rift basins with similar tectonic and sedimentary evolution processes. Under similar geological backgrounds, the controlling factors of hydrocarbon accumulation are always consistent. Otherwise, this evaluation and prediction model should be modified. For example, for unconventional

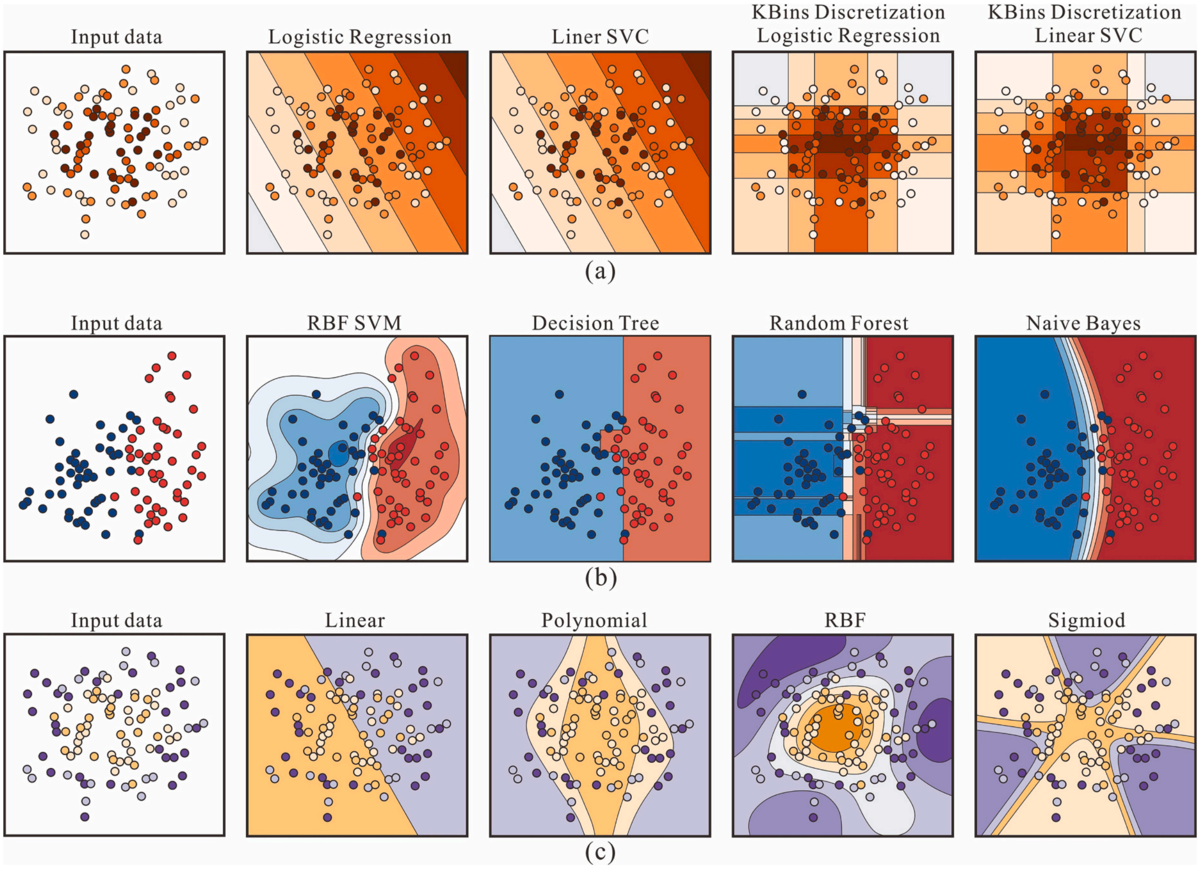


Fig. 31. Comparison of the linear and nonlinear regression algorithms for dealing with different kind of problems. (a) The poor performance of the linear regression models to deal with nonlinear problems. (b) The good performance of the non-linear regression models to deal with linear problems. (c) The SVM can provide the linear models by linear kernel and the nonlinear models by polynomial, RBF and sigmoid kernels.

resources, the development of caprock is not a necessary condition for hydrocarbon accumulation, and the decisive factors may be the development of fractures and well pore connections that construct the "sweet spots" of tight gas sandstone or shale gas reservoirs. Overall, the construction of PNGR resource potential prediction models should be guided by petroleum geological knowledge.

6. Conclusion

(1) This study introduces a PNGR potential prediction method, which is based on ML and firstly applies the SVC and SVR as a combination for dealing with a practical problem in petroleum exploration and exploitation field. Based on the data preprocessing and construction of feature variables, which are constrained and fixed by petroleum geology knowledge, the PNGR potential is first predicted by hydrocarbon accumulation probability prediction by SVC and then reserve abundance prediction by the SVR algorithm. Furthermore, the predictability of the feature variables can support the model to make lateral continuous distribution of PNGR potentials, which is essential for the model application in practical.

(2) The two-step SVM prediction model of PNGR shows good performance with high accuracy, high stability and good generalization. The classification model based on SVC provides a hydrocarbon accumulation probability evaluation with tested accuracies ranging from 80% to 100% and an average of 88.92%. Based on the constrain of SVC model, the SVR model provides the reserve abundance evaluation with the tested Ccoe ranging from 0.502 to 0.767, with an average of 0.602. The validation results show that the model can predict the hydrocarbon accumulation probability with an accuracy of 72.5% and the reserve abundance with a Ccoe of 0.744, which is <0.4 without the constrain of SVC model. Therefore, the proper classification processes based on the professional knowledge system can significantly improve the performance of the regression model in this research, which is also referable for the application of ML. In total, the predicted results by the SVC and SVR models match well with the actual exploration situations in the Dongpu Depression. The result indicates the effectiveness of the combination of SVR and SVC, which is firstly applied in the PNGR prediction.

(3) The PNGR prediction model based on SVC and SVR also shows better performance comparing to single classification or regression algorithms. The DT and RF model show similar upper limit of accuracy as the SVC model but also show instability with lowest accuracy of 36.3% to 49.4%. While the KNN and NB model show good stability but lower upper limit of the predicting accuracy. The advantage of the SVR model constrained by the SVC model is obvious comparing to the other single regression models both in accuracy and stability. Furthermore, the ensemble learning model of GBDT shows obvious improvement comparing to the single model both in the improvement of upper limit and the stability.

(4) There are several ways to improve the performance of the prediction model, including enlarging the data size, optimizing the feature variable construction process and using the ensemble learning model. The utilization of ML in PNGR potential prediction has broad prospects and will be helpful in guiding the exploration and exploitation of petroleum and natural gas resources.

CRediT authorship contribution statement

Qiaochu Wang: Writing – original draft, Methodology, Data curation. **Dongxia Chen:** Supervision, Funding acquisition, Conceptualization. **Meijun Li:** Formal analysis, Conceptualization. **Sha Li:** Investigation, Data curation. **Fuwei Wang:** Software, Formal analysis. **Zijie Yang:** Software, Resources. **Wanrong Zhang:** Validation, Resources. **Shumin Chen:** Writing – review & editing, Investigation. **Dongsheng Yao:** Visualization, Data curation.

Declaration of Competing Interest

None.

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 gratefully acknowledge the Dongpu Oilfield Branch of the China Petroleum & Chemical Corporation for providing field test data and reservoir data.

References

[2] EIA. International energy outlook. Washington: US Energy information Administration; 2020.

[3] White IC. The geology of natural gas. Science 1885;5(125):521–2.

[4] Hunt JM. Petroleum geochemistry and geology. San Francisco: W. H. Freeman and Company; 1979.

[5] Masters JA. Deep basin gas trap, western Canada. AAPG Bull 1979;63(2):152–81.

[6] Jia CZ. Breakthrough and significance of unconventional oil and gas to classical petroleum geology theory. Pet Explor Dev 2017;44(1):1–10.

[7] Rose PR, Everett JR, Merin IS. Possible basin centered gas accumulation, Roton Basin, Southern Colorado. Oil Gas J 1984;82(10):190–7.

[8] Spencer CW. Geologic aspects of tight gas reservoirs in the rocky mountain region. J Petrol Tech 1985;37(7):1308–14.

[9] Collett TS. Energy resource potential of natural gas hydrates. AAPG Bull 2002;86 (11):1971–92.

[10] Levorsen AI. Geology of petroleum. San Francisco: W. H. Freeman and Company; 1956.

[11] Dow WG. Application of oil-correlation and source-rock data to exploration in Williston Basin. AAPG Bull 1974;58(7):1253–62.

[12] Peerodon A. Petroleum systems: models and applications. J Pet Geol 1992;15(2): 319–25.

[13] Jiang FJ, Pang XQ, Li LL, et al. Petroleum resources in the Nanpu sag, Bohai Bay Basin, eastern China. AAPG Bull 2018;102(7):1213–37.

[14] Chen JQ, Pang XQ, Wang XL, et al. A new method for assessing tight oil, with application to the Lucaogou formation in the Jimusaer depression, Junggar Basin, China. AAPG Bull 2020;104(6):1199–229.

[15] Wandrey CJ, Law BE, Shan HA. Patala-nammal composite total petroleum system, Kohat Potwar geologic province, Pakistan. Reston: U. S. Geological Survey; 2004.

[16] Schmoker JW. Resource-assessing perspectives for unconventional gas systems. AAPG Bull 2002;86(11):1993–2000.

[17] Li ZX, Bogdanova SV, Collins AS, et al. Assembly, configuration and break-up history of Rodinia: A synthesis. Precambrian Res 2008;160(1):179–210.

[18] Sonnenberg SA, Pramudito A. Petroleum geology of the giant elm coulee field, Williston Basin. AAPG Bull 2009;93(9):1127–53.

[19] Magoon LB, Dow WG. The petroleum system-status of research and methods. USGS Bull 1992;20(7):98.

[20] McCammon RB, Briskey JA. A proposed national mineral resource assessment. Nonrenew Resour 1992;1(4):259–65.

[21] Cox DP. Estimation of undiscovered deposits in quantitative mineral resource assessments-examples from Venezuela and Puerto Rico. Nonrenew Resour 1993;2 (2):82–91.

[22] Magoon LB, Schmoker JW. The total petroleum system: The natural fluid network that constrains the assessment unit. Reston: U. S. Geological Survey World Petroleum Assessment; 2000.

[23] Perrodon A, Masse P. Subsidence, sedimentation and petroleum systems. J Pet Geol 1984;7(1):5–25.

[24] Singer DA. Basic concepts in three-part quantitative assessments of undiscovered mineral resources. Nonrenew Resour 1993;2(2):69–81.

[25] Reed BL, Menzie WD, McDermott M, et al. Undiscovered lode tin resources of the Seward peninsula, Alaska. Econ Geol 1989;84:1936–47.

[26] Singer DA, Ovenshine AT. Assessing Metallic Resources in Alaska: American Scientist67; 1979. p. 582–9.

[27] Zou CN, Zhang GS, Yang Z, et al. Geological concepts, characteristic, resource potential and key techniques of unconventional hydrocarbon: on unconventional petroleum geology. Pet Explor Dev 2013;40(4):385–99.

[28] Meneley RA, Calverley AE. Resource assessment methodologies: current status and future direction. AAPG Bull 2003;87(4):535–40.

[29] Klett TR. United States geological Survey's reserve growth models and their implementation. Nat Resour Res 2005;14(3):249–64.

[30] Wang J, Heng J, Xiao L, et al. Research and application of a combined model based on multi-objective optimization for multi-step ahead wind speed forecasting. Energy 2017;125:591–613.

[31] Zamo M, Mestre O, Arbogast P, et al. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part I: deterministic forecast of hourly production. Energy 2014;105:792–803.

[32] Ozgoren M, Bilgili M, Sahin B. Estimation of global solar radiation using ANN over Turkey. Expert Syst Appl 2012;2012(39):5043–51.

[33] Freek VDM, Dijk PV, Werff HVD, et al. Remote sensing and petroleum seepage: a review and case study. Terra Nova 2002;14(1):1–17.

[34] Lee PJ, Wang PC. Probabilistic formulation of a method for the evaluation of petroleum resources. Math Geol 1983;15(1):163–81.

[35] Piryonesi SM, EI-Diraby TE. Role of data analytics in infrastructure asset management: overcoming data size and quality problems. J Transp Eng 2021;146 (2). https://doi.org/10.1061/JPEODX.0000175.

[36] Kreimeyer K, Dang O, Spiker J, Munoz AM, Rosner G, Ball R, et al. Feature engineering and machine learning for causality assessment in pharmacovigilance: lessons learned from application to the FDA adverse event reporting system. Comput Biol Med 2021;135:104517.

[39] Agterberg FP. A modified weights-of-evidence method for regional mineral resource estimation. Nat Resour Res 2011;20(2):95–101.

[40] Fatai A, Abdulazeez A. Fuzzy logic-driven and SVM-driven hybrid computational intelligence models applied to oil and gas reservoir characterization. J Nat Gas Sci Eng 2011;3(3):505–17.

[41] Yenugu M, Fisk JC, Marfurt KJ, et al. Probabilistic neural network inversion for characterization of coalbed methane. Soc Explor Geophys 2010. https://doi.org/10.1190/1.3513449.

[42] Mitchell TM. Machine learning. New York: McGraw-Hill; 1997.

[43] Quinlan JR. C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann; 2003. p. 302.

[44] Pradhan B. 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 2013;51:350–65.

[45] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. Springer; 2008 [ISBN 0-387-95284-5.].

[46] Jiang YL, Fang L, Liu JD, et al. Hydrocarbon charge history of the Paleogene reservoir in the Northern Dongpu depression, Bohai Bay Basin, China. Pet Sci 2016;13(4):625–41.

[47] Wang M, Sherwood N, Li ZS, et al. Shale oil occurring between salt intervals in the Dongpu depression, Bohai Bay Basin, China. Int J Coal Geol 2015;152:100–12.

[48] Xu TW, Zhang HA, Li JD, et al. Characters of hydrocarbon generation and accumulation of salt-Lake facies in Dongpu Sag, Bohai Bay Basin. Oil Gas Geol 2019;40(2):248–61 [in Chinese with English Abstract].

[49] Liu XY, Chen HH, Xiao XW, et al. Overpressure evolution recorded in fluid inclusions in the Dongpu depression, Bohai Bay basin, North China. J Earth Sci 2022;33(4):916–32.

[50] Wang QC, Chen DX, Gao XZ, Wang FW, Li S, Tian ZY, et al. Evolution of abnormal pressure in the Paleogene Es3 formation of the Huimin depression, Bohai Bay bBasin, China. J Petrol Sci Eng 2021;203. https://doi.org/10.1016/j.petrol.2021.108601.

[51] Amyx JW, Bass DM, Whiting RL. Petroleum reservoir engineering, physical properties. New York: McGraw-Hill; 1960. p. 610.

[52] Gvirtzman H, Stanislavsky E. Palaeohydrology of hydrocarbon maturation, migration and accumulation in the Dead Sea rift. Basin Res 2000;12(1):79–93.

[53] Jia CZ, Pang XQ, Song Y. The mechanism of unconventional hydrocarbon formation: hydrocarbon self-containtment and intermolecular forces. Pet Explor Dev 2021;48(3):437–52.

[54] Tissot BP, Welte DH. Petroleum formation and occurrence. New York: Springer-Verlag; 1978.

[55] Dai JX, Ni YY, Zou CN. Stable carbon and hydrogen isotopes of natural gases sourced from the Xujiahe formation in the Sichuan Basin, China. Org Geochem 2012;43:103–11.

[56] Durand B. Present trends in organic geochemistry in research on migration of hydrocarbons. In: Advances in organic chemistry. New York: John Wiley; 1983. p. 117–28.

[57] Wang QC, Chen DX, Wang FW, et al. Origin and distribution of an under-pressured tight sandstone reservoir: the Shaximiao formation, Central Sichuan Basin. Mar Pet Geol 2021;132(3). https://doi.org/10.1016/j.marpetgeo.2021.105208.

[58] Hu TL, Ge BX, Zhang YG. The development and application of fingerprint parameters for hydrocarbons absorbed by source rocks and light hydrocarbons in natural gas. Pet Geol Exp 1990;12(4):375–93.

[59] Peters KE. Guidelines of evaluating petroleum source rock using programmed pyrolysis. AAPG Bull 1986;70(3):318–29.

[60] Mukhopadhyay PK, Wade JA, Kruge MA. Organic facies and maturation of Jurassic/cretaceous rocks, and possible oil-source rock correlation based on pyrolysis of asphaltenes, Scotia Basin, Canada. Org Geochem 1995;22:85–104.

[61] Pang XQ. Theory and application of the hydrocarbon expulsion threshold controlling petroleum distribution. Beijing: Petroleum Industry Press; 1995. p. 88–92.

[62] Bai H, Pang XQ, Kuang LC, et al. Hydrocarbon expulsion potential of source rocks and its influence on the distribution of lacustrine tight oil reservoir, middle Permian Lucaogou formation, Jimsar Sag, Junggar Basin, Northwest China. J Petrol Sci Eng 2017;149:740–55.

[63] Zheng TY, Ma XH, Pang XQ, et al. Organic geochemistry of the upper Triassic T_3x^5 source rocks and the hydrocarbon generation and expulsion characteristics in Sichuan Basin, Central China. J Petrol Sci Eng 2019;173:1340–54.

[64] Tang L, Pang XQ, Xu TW, et al. Hydrocarbon generation thresholds of Paleogene Shahejie Fm source rocks and their north–south differences in the Dongpu Sag, Bohai Bay Basin. Nat Gas Ind 2017;37(2):26–37 [in Chinese with English Abstract].

[65] Hunt JM. Petroleum geochemistry and geology. 2nd ed.296-298. New York: Freeman; 1996.

[66] Miall AD. Principles of sedimentary basin analysis. New York: Springer-Verlag; 1984.

[67] Friedman GM, Sanders JE. Principles of sedimentology. New York: Wiley; 1978.

[68] Roger GW. Facies models. Ottawa: Geological Association of Canada; 1992.

[69] Huo Z, Pang X, Fan K, Chen D, Zhang J. Analysis and application of facies and potential coupling control of typical lithologic hydrocarbon reservoirs in Jiyang depression. Pet Geol Exp 2014;36. https://doi.org/10.11781/sysydz201405574. 574-582+588.

[70] Canham A. Reservoir quality prediction in sandstones and carbonates. J Petrol Sci Eng 2001;30:260–1. https://doi.org/10.1016/S0920-4105(01)00117-6.

[71] Catalan LFXW, Chatzis I, Dullien FAL. An experimental study of secondary oil migration. AAPG Bull 1992;76:638–50.

[72] Dembicki Jr H, Anderson MJ. Secondary migration of oil: experiments supporting efficient movement of separate, buoyant oil phase along limited conduits. AAPG Bull 1989;73:1018–21.

[73] England WA, Mackenzie AS, Mann DM. Movement and entrapment of petroleum fluid in the subsurface. J Geol Soc London 1987;144(2):327–47.

[74] Hubbert MK. Entrapment of petroleum under hydrodynamic conditions. AAPG Bull 1953;37(8):1954–2026.

[75] Liu GD. Petroleum geology (in Chinese). Beijing: Petroleum Industry Press; 2018.

[76] Marine IW, Fritz SJ. Osmotic model to explain anomalous hydraulic heads. Water Resour Res 1981;17:73–82.

[77] Yang H, Jin FU, Liu X, Meng PL. Accumulation conditions and exploration and development of tight gas in the upper Paleozoic of the Ordos Basin. Pet Explor Dev 2012;39(3):315–24 [in Chinese with English Abstract].

[78] Vapnik V, Lerner A. Pattern recognition using generalized portrait method. Autom Remote Control 1963;24:774–80.

[79] Vapnik VN. Estimation of dependences based on empirical data. Berlin: Springer; 1982.

[80] Boser BE, Guyon IM, Vapnik VN. Atraining algorithm for optimal margin classifiers. In: Haussler D, editor. Proceedings of the annual conference on computational learning theory. Pittsburgh, PA: ACM Press; 1992. p. 144–52.

[81] Guyon I, Boser B, Vapnik V. Automatic capacity tuning of very large VC-dimension classifiers. In: Hanson SJ, Cowan JD, Giles CL, editors. Advances in neural information processing systems 5. Morgan Kaufmann Publishers; 1993. p. 147–55.

[82] Vapnik V, Golowich S, Smola A. Support vector method for function approximation, regression estimation, and signal processing. In: Mozer MC, Jordan MI, Petsche T, editors. Advances in neural information processing systems 9, MA. Cambridge: MIT Press; 1997. p. 281–7.

[83] Burges CJC. A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 1998;2(2):121–67.

[84] Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V. Support vector regression machines. In: Mozer MC, Jordan MI, Petsche T, editors. Advances in neural information processing systems 9. Cambridge, MA: MIT Press; 1997. p. 155–61.

[85] Stitson M, Gammerman A, Vapnik V, Vovk V, Watkins C, Weston J, et al. Support vector regression with ANOVA decomposition kernels. In: Advances in kernel methods—Support vector learning. Cambridge, MA: MIT Press; 1999. p. 285–92.

[86] Haykin S. Neural networks: A comprehensive foundation. 2nd ed. New York: Macmillan; 1998.

[87] Cristianini N, Shawe-Taylor J. An introduction to support vector machines. Cambridge, UK: Cambridge University Press; 2000.

[88] Herbrich R. Learning kernel classifiers: Theory and algorithms. MIT Press; 2002.

[89] Cortes C, Vapnik V. Support vector networks. Mach Learn 1995;20:273–97.

[90] Chang CC, Lin CJ. Training ν-support vector classifiers: theory and algorithms. Neural Comput 2001;13(9):2119–47.

[91] Huber PJ. Robust statistics. New York: John Wiley and Sons; 1981.

[92] Keerthi SS, Shevade SK, Bhattacharyya C, Murty KRK. Improvements to platt's SMO algorithm for SVM classifier design. Neural Comput 2001;13:637–49.

[93] Lee YJ, Mangasarian OL. SSVM: A smooth support vector machine for classification. Comput Optim Appl 2001;20(1):5–22.

[94] Vapnik V. The nature of statistical learning theory. New York: Springer; 1995.

[95] Bennett KP, Mangasarian OL. Robust linear programming discrimination of two linearly inseparable sets. Optim Methods Softw 1992;1:23–34.

[96] Cortes C, Vapnik V. Support vector networks. Mach Learn 1995;20:273–97.

[97] McCormick GP. Nonlinear programming: Theory, algorithms, and applications. New York: John Wiley and Sons; 1983.

[98] Wahba G. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In: Schölkopf B., Burges C.J.C., and Smola A.J., editor. Advances in kernel methods— Support Vecto learning. Cambridge, MA: MIT Press; 1999. p. 69–88.

[99] Bhattacharya S, Ghahfarokhi PK, Carr TR, Pantaleone S. Application of predictive data analytics to model daily hydrocarbon production using petrophysical, geomechanical, fiber-optic, completions, and surface data: A case study from the Marcellus Shale, North America. J Petrol Sci Eng 2019;176:702–15.

[100] Mudunuru MK, Hoover KF. Machine learning models for real-time forecasting of shale gas production (no. LA-UR-20-25822). Los Alamos: National Lab (LANL), Los Alamos, NM (United States); 2020.

[101] Huang R, Wei C, Wang B, Yang J, Xu X, Wu S, et al. Well performance prediction based on long short-term memory (LSTM) neural network. J Petrol Sci Eng 2022;208:109686.

[102] Liu YY, Ma XH, Zhang XW, Guo W, Kang LX, Yu RZ, et al. A deep-learning-based prediction method of the estimated ultimate recovery (EUR) of shale gas wells. Pet Sci 2021;18(5):1450–64.

[103] Rahmanifard H, Alimohamadi H, Gates I. Well performance prediction in montney formation using machine learning approaches. In: SPE/AAPG/SEG Unconventional Resources Technology Conference; 2020.

[104] Niu W, Sun Y, Zhang X, Lu J, Liu H, Li Q, et al. An ensemble transfer learning strategy for production prediction of shale gas wells. Energy 2023;275:127443.

[105] Gao F, Qu ZP, Wei Z, Zhu JB, Cheng YF. A study on well-log lithofacies classification based on machine learning methods. In: Progress in geophysics; 2023 (in Chinese), https://kns.cnki.net/kcms2/detail/11.2982.P.20230801.1451.034.html.

[106] Luo G, Xiao LZ, Shi YQ, Shao RB. Machine learning for reservoir fluid identification with logs. Pet Sci Bull (in Chin) 2022;7(1):24–33.

[107] Ariza Ferreira DJ, Dias RM, Lupinacci WM. Seismic pattern classification integrated with permeability ‐ porosity evaluation for reservoir characterization of Presalt carbonates in the Buzios field, Brazil. J Petrol Sci Eng 2021;201:1–12.

[108] Chevitarese DS, Szwarcman D, E Silva RMG, et al. Deep learning applied to seismic facies classification: A methodology for trainingvol. 01. European Association of Geoscientists & Engineers; 2018. p. 1–5.

[109] Colombera L, Mountney NP. Accommodation and sediment ‐ supply controls on clastic Parasequences: A Meta ‐ analysis. Sedimentology 2020;67:1667–709.

[110] Convers C, Hanitzsch C, Curia D, et al. Elastic parameter estimation for the identification of sweet spots, Vaca Muerta formation, Neuquén Basin, Argentina. Lead Edge 2017;36(11). 948a1-948a10.

[112] Li W, Yue D, Colombera L, et al. Quantitative prediction of fluvial Sandbodies by combining seismic attributes of neighboring zones. J Petrol Sci Eng 2021;196:107749.

[113] Ma KY, Pang XQ, Pang H, Lv CB, Gao T, Chen JQ, et al. A novel method for favorable zone prediction of conventional hydrocarbon accumulations based on RUSBoosted tree machine learning algorithm. Appl Energy 2022;326:119983.

[114] Zhao A. Quantitative screening method for shale gas favorable area of Wufeng-Longmaxi formation in the tectonic complex area west of Xuefeng Mountain. Chengdu University of Technology; 2019.

[116] Han D, Jung J, Kwon S. Comparative study on supervised learning models for productivity forecasting of shale reservoirs based on a data-driven approach. Appl Sci 2020;10(4):1267.

[121] Ghane M, Ang MC, Nilashi M, Sorooshian S. Enhanced decision tree induction using evolutionary techniques for Parkinson's disease classification. Biocybernetics Biomed Eng 2022;42:902–20.

[122] Harris JR, Ayer J, Naghizadeh M, Smith R, Snyder D, Behnia P, et al. A study of faults in the superior province of Ontario and Quebec using the random forest machine learning algorithm: spatial relationship to gold mines. Ore Geol Rev 2023;157:105403.

[123] Hoang N, Bérengère SS, Virginie P, Arnt J, Robin K, Alexandre B. Application of random forest algorithm to predict lithofacies from well and seismic data in balder field, Norwegian North Sea. AAPG Bull 2022;106(11):2239–57.

[124] Ilić M, Srdjević Z, Srdjević B. Water quality prediction based on Naïve Bayes algorithm. Water Sci Technol 2022;85(4):1027–39.

[125] Alizadeh SH, Hediehloo A, Harzevili NS. Multi independent latent component extension of naive Bayes classifier. Knowl Based Syst 2021;213:106646.

[127] Ciulla G, D'Amico A. Building energy performance forecasting: A multiple linear regression approach. Appl Energy 2019;253:113500.

[128] Davut O, Ali OP, Ali U. Creating a non-linear total sediment load formula using polynomial best subset regression model. J Hydrol 2016;539:662–73.

[129] Vikara D, Remson D, Khanna V. Gaining perspective on unconventional well design choices through play-level application of machine learning modeling. Upstream Oil Gas Technol 2020;4:100007.

[130] Jamei M, Ali M, Karbasi M, Xiang Y, Ahmadianfar I, Yaseen ZM. Designing a multi-stage expert system for daily ocean wave energy forecasting: A multivariate data decomposition-based approach. Appl Energy 2022;326:119925.

[131] Pandey M, Jamei M, Ahmadianfar I, et al. Assessment of scouring around spur dike in cohesive sediment mixtures: A comparative study on three rigorous machine learning models. J Hydrol 2022;606:127330.

[132] Pamidi VDK, Neetish KM. Machine learning approaches for formation matrix volume prediction from well logs: insights and lessons learned. Geoenergy Sci Eng 2023;229:212086.

[133] Kelter R. 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 2021;100:102474.

[134] Wainer J, Cawley G. Nested cross-validation when selecting classifiers is overzealous for most practical applications. Expert Syst Appl 2021;182:115222.

[135] Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physicsinformed machine learning. Nat Rev Phys 2021;3:422–40.

[136] Saeed B, Masoud M, Adel N. Review of application of artificial intelligence techniques in petroleum operations. Pet Res 2023;8(2):167–82.

[137] Elise LG, Martin F, Kristin T. Current overview and way forward for the use of machine learning in the field of petroleum gas hydrates. Fuel 2023;334:126696.

[139] Breiman L. Random Forests. Machine Learning 2001;45(1):5–32.

[140] Quinlan JR. C4.5: Programs for machine learning. San Mateo, California: Morgan Kaufmann; 1993.

[141] Wang QC, Chen DC, Li MJ, Wang FW, Wang Y, Du WL, et al. Application of machine learning for evaluating and predicting fault seals: A case study in the Huimin Depression, Bohai Bay Basin, Eastern China. Geoenergy Science and Engineering 2023;228:212064.