Applied Energy 326 (2022) 119983

Contents lists available at ScienceDirect

Applied Energy

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

A novel method for favorable zone prediction of conventional hydrocarbon accumulations based on RUSBoosted tree machine learning algorithm

Kuiyou Ma [a,b], Xiongqi Pang [a,b], Hong Pang [a,b,*], Chuanbing Lv [c], Ting Gao [c], Junqing Chen [d], Xungang Huo [a,b], Qi Cong [d], Mengya Jiang [a,b]

[a] *State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum, Beijing 102249, China*
[b] *College of Geosciences, China University of Petroleum, Beijing 102249, China*
[c] *Huabei Oilfield Company of PetroChina, Renqiu 062552, China*
[d] *College of Sciences, China University of Petroleum, Beijing 102249, China*

ARTICLE INFO

Keywords:
Petroleum exploration
Conventional oil and gas
RUSBoost algorithm
Favorable zone forecast
Jin 93 Well Block
Shulu Sag

ABSTRACT

The prediction of favorable zone (FZ) is the most important step for conventional hydrocarbon accumulations (CHAs) exploration. Recently, the method of coupling multiple hydrocarbon accumulation (HA) elements is widely used to predict the distribution of FZ in the petroleum exploration field. However, the forming mechanism of CHAs is extremely complicated, which causes difficulty in accurately describing the relationship between multiple HA elements and HA probability (HAP). Hence, it is difficult to predict the distribution of FZ quantitatively and credibly using traditional methods. This study proposes a method for predicting FZ for CHAs based on random undersampling boosted (RUSBoosted) tree machine learning (ML) algorithm. First, the characteristics of data in the petroleum exploration field are clarified, and a suitable ML algorithm is selected. Second, the theory and knowledge of the petroleum exploration field is integrated into the data, a HAP prediction model for CHAs is constructed, and then the method for FZ prediction is proposed. Further, the method is applied to Jin 93 Well Block for predicting FZ of CHAs. Finally, this study discussed the difference in performance among models constructed by the RUSBoosted tree and other five ML algorithms and the difference in training results between the original geological data and preprocessed geological data on the RUSBoosted tree ML algorithm. Results show that, currently, datasets in the petroleum exploration field are small and unbalanced, and the RUSBoosted tree ML algorithm has excellent training results on it. Compared with the original geological data, the performance of the HAP prediction model constructed by preprocessed geological data is improved. On a Jin 93 Well Block dataset, the HAP prediction model constructed by the RUSBoosted tree ML algorithm belongs to a good prediction model, and FZ of CHAs predicted by this HAP prediction model agree well with CHAs discovered areas. The results of this study provide an idea for intelligently predicting the distribution of FZ of CHAs and are of great significance to the development of intelligent petroleum exploration technology.

1. Introduction

Conventional hydrocarbon accumulations (CHAs) result from the joining of multiple hydrocarbon accumulation (HA) elements [1,2]. Favorable zone (FZ) prediction of CHAs is a paramount step in petroleum exploration [1–3]. Moreover, the establishment of HA probability (HAP) prediction model based on real geological data of HA elements is the core of FZ prediction [4–6]. In statistics and petroleum geology, HAP exists and means the probability of hydrocarbon accumulating under certain geological conditions. But HAP only can be estimated manually based on petroleum exploration experience or using statistical methods based on enormous drilling data. High HAP areas will be predicted using geological data of unexplored areas once the HAP prediction model is established [2]. Petroleum system is the entirety of a source rock and all HAs generated by this source rock, encompassing all the HA elements and processes [7]. In theory, there are homologous processes and similar geological conditions for CHAs in a single petroleum system. Thus, there are also similar characteristics for conventional oil and gas accumulations [7]. Therefore, the HAP prediction model for CHAs is usually constructed and applied in a single petroleum system to improve the

* Corresponding author at: College of Geosciences, China University of Petroleum, Beijing 102249, China.
E-mail address: panghong19820107@126.com (H. Pang).

https://doi.org/10.1016/j.apenergy.2022.119983
Received 16 April 2022; Received in revised form 2 August 2022; Accepted 12 September 2022
Available online 30 September 2022

accuracy of FZ prediction [8]. Moreover, in petroliferous basins with multiple petroleum systems, it is necessary to divide them into several single petroleum systems, and then further predict FZ for CHAs in each single petroleum system [3,8]. Therefore, it is an essential basis for predicting the distribution of CHAs in complex petroliferous basins to accurately predict FZ for CHAs in a single petroleum system.

In a single petroleum system, traditional methods for constructing HAP prediction models based on multiple HA elements mainly include the expert scoring method [9], analytic hierarchy process [4], fuzzy comprehensive evaluation method [5], and multiple linear regression method [6]. These methods employ linear models to represent the relationship between HA elements and HAP [4–6]. The first three methods are qualitative analysis methods [4,5,9], whereas the last method is a quantitative analysis [6]. For the expert scoring method, the weight of each HA element is qualitatively evaluated based on the opinion of experts who are familiar with the geological conditions and oil and gas accumulation process in the study area [9]. The analytic hierarchy process and fuzzy comprehensive evaluation method are improvements to the expert scoring method, making the weights of each HA element reflect their importance more accurately [4,5]. Actually, the expert scoring method, analytic hierarchy process, and fuzzy comprehensive evaluation method are mostly used in the primary stage of HAs exploration, have a high degree of human interference and a high risk for actual exploration work. For multiple linear regression method, which is a representative data-based method, hydrocarbon reserve, hydrocarbon saturation, and hydrocarbon production always are used to characterize HAP and the weight of each HA element is calculated using the multiple linear regression method [6]. Nevertheless, owing to the strong heterogeneity of geology, the process of oil and gas from source to accumulation is extremely complicated [10–13]. There is no simple linear relationship between HA elements and hydrocarbon reserve. In addition, only a small amount of sample data is used to construct the linear equation; large amounts of drilled non-accumulation sample data are ignored in this method. Therefore, the multiple linear regression method also induces a high exploration risk in practical application.

As well-known data mining methods in the past decade, machine learning (ML) algorithms no longer aim at constructing neat and elegant mathematical function models but to establish a more complex and accurate mapping relationship between sample features and results [14–16]. The mapping relationship between HA elements and HPA is extremely complex and difficult to express intuitively. In addition, HPA prediction models constructed by conventional neat mathematical functions hardly realize an accurate result. In theory, it is completely feasible to use an ML algorithm to build a highly accurate HPA prediction model based on petroleum exploration data [14,15]. However, ML algorithms are data-driven models. To build a credible prediction model, it is necessary to select a suitable ML algorithm and training strategy according to the dataset characteristics [14,16].Currently, petroleum exploration datasets are small and imbalanced. Recently, due to the constrain of cost for exploration well drilling and geophysical technological survey and the impact of confidentiality rules, most researchers can obtain only a small dataset to conduct a study in the petroleum exploration field. For example, Wang et al. [3] constructed a HAP prediction model based on a 114 × 5 (114 samples with 5 features) dataset and predicted the favorable zone for oil and gas accumulation in Cambrian Longwangmiao Formation in Sichuan Basin, China. Zhao [17] used a 102 × 7 (102 samples with 7 features) dataset to predict gas content in shale reservoirs for Wufeng–Longmaxi Formation, west of Xuefeng Mountain, China. Sheremetov et al. [18] established a monthly oil production prediction model based on a 340 × 31 (340 samples with 31 features) dataset. In addition, due to the low success rate of exploration [19], the number of hydrocarbon layers drilled is far less than that of nonhydrocarbon layers, making the dataset of the petroleum exploration field imbalanced. In imbalanced datasets, samples of one category (majority category) extremely outnumber samples of other categories (minority categories), and the prediction accuracy of minority categories is relatively poor [20,21].

Limited and imbalanced datasets make the application of ML algorithms face two challenges. (1) Generally, the performance of a prediction model constructed using an ML algorithm based on a small dataset is always poor [16,22,23]. (2) For most ML algorithms, imbalanced datasets severely compromise their performance [20,21,24]. Regarding the first challenge, the application of ML algorithms to limited data has been explored and discussed extensively [24–28]. At present, advanced and feasible methods to solve limited data problems is that integrate knowledge and theory into ML algorithms or raw data [16,25–28]. Karniadakis et al. [22] reported a type of ML prediction model that uses physical knowledge and theory to solve physical problems, thereby solving the problem of low accuracy of ML algorithms based on small datasets in the physical field. Their primary idea is the complementarity of theory and knowledge and data size. Without theory and knowledge constrain, creating a good or excellent performance prediction model requires massive data. Meanwhile, the amount of necessary data decreases when theory and knowledge constrain is integrated into ML algorithms (Fig. 1) [22]. Likewise, to overcome the limitation of small datasets and build a high-performance prediction model for HAP, it is necessary to integrate theory and knowledge in the petroleum exploration field into ML algorithms. The second challenge also needs to be overcome in the petroleum exploration field. ML algorithms usually focus on the overall dataset prediction performance, resulting in insufficient attention and severely low prediction accuracy for the minority categories [21]. The same problem is also widely distributed in the credit forecasting and medical case diagnosis fields [20].

In a single petroleum system, how can a HAP prediction model be built based on multiple HA elements by ML algorithms for the CHAs exploration field? And how can FZ for CHAs be predicted automatically with the HAP prediction model? Aiming at this scientific question, this study proposed a data preprocessing method is proposed to integrate theory and knowledge of the petroleum exploration field into raw data. In addition, random undersampling boosted (RUSBoosted) ensemble ML algorithms are used to solve the low-performance problem due to imbalanced datasets. Finally, a HAP prediction model is established, and a strategy for automatically predicting FZ for CHAs is proposed. Further, all proposed methods are applied and validated for Jin 93 block in Shulu Sag in Jizhong Depression in Bohai Bay Basin. The results of this study provide ideas for constructing an intelligent and automatic module for predicting favorable zones for conventional oil and gas exploration. They are also of great significance for developing intelligent oil and gas exploration technology and constructing digital oil fields.

2. Method

The study methodology to establish HAP prediction model and predict FZ in a single petroleum system will be presented by following three parts: data preprocessing, HAP prediction model construction, and favorable zones prediction strategy. The strategy for constructing features with a strong relationship to HAP based on theory and knowledge is discussed in detail in the first part. The dataset construction method, principle of the RUSBoosted tree ML algorithm, and method to train the HAP prediction model are introduced in the second part. The specific workflow for predicting favorable zones is presented in the third part. Finally, a comprehensive technical guide for this study is shown.

2.1. Data preprocessing

Petroleum system theory provides essential theoretical guidance for oil and gas exploration, whose core idea is that the formation of CHAs is a dynamic process from source to trap [7]. In this process, the three most important elements are source, gather, and reserve, corresponding to source rock, gather environment, and reservoir conditions, respectively [7]. To integrate the knowledge of the petroleum exploration field into

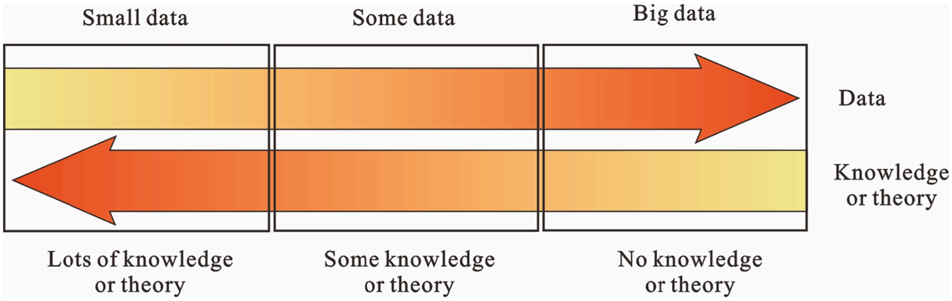


Fig. 1. Schematic of the complementary role of theory and knowledge and data size in the ML algorithm training process [adapted from Karniadakis et al. [22]].

raw data and improve the performance of ML prediction as much as possible, a raw data preprocessing method is proposed based on source controlling theory [29–32], fluid potential theory [7,33–39], and predominant reservoir theory [40–42]. Correspondingly, the source index (I_s), relative potential index (I_{rp}), porosity index (I_φ), and permeability index (I_k) are constructed to reflect the quality for source condition, gather environment, storage space condition, and percolation capacity, respectively. Each index has a close relationship with HAP.

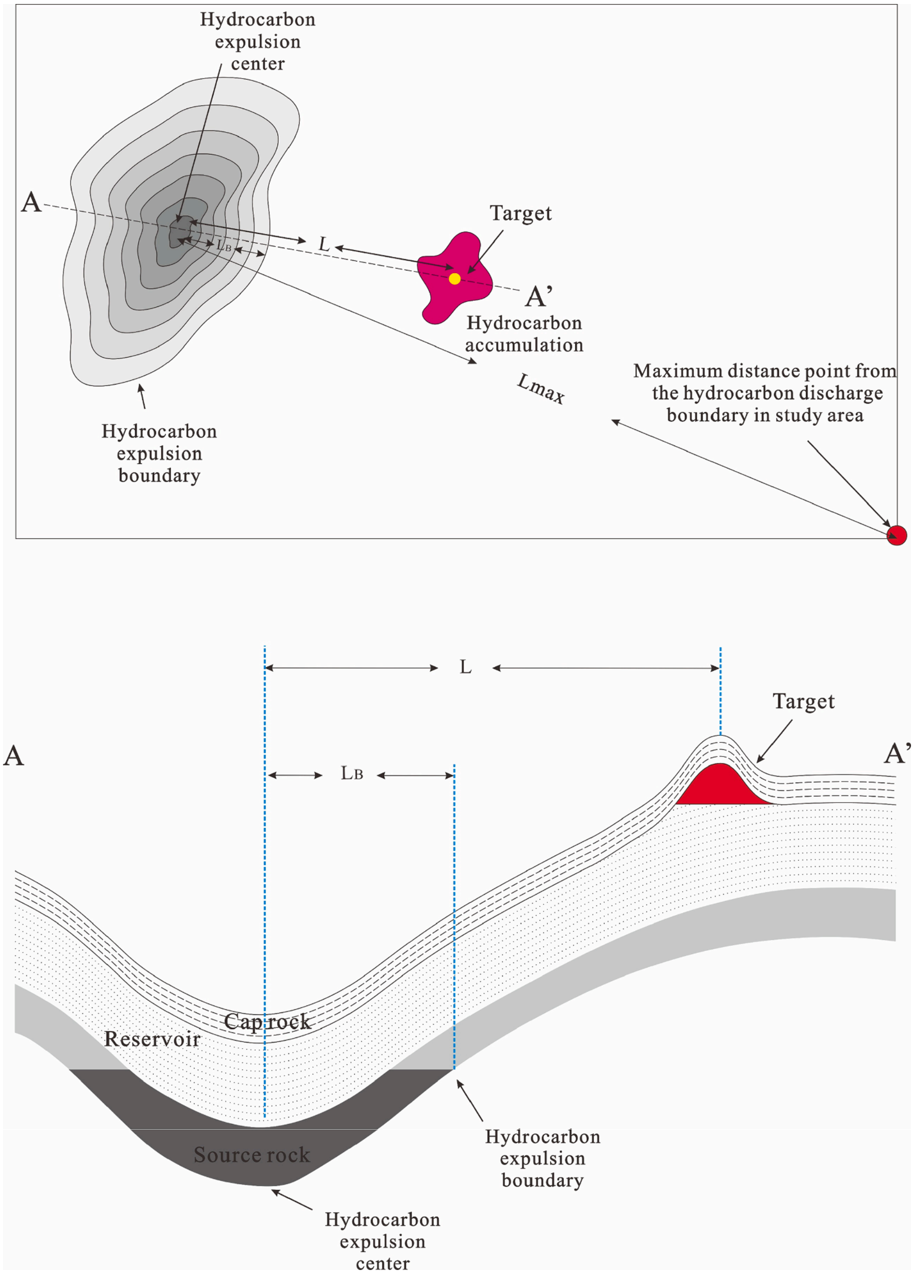

Hydrocarbon expulsion center
A
L
LB
Target
A'
Hydrocarbon accumulation
Lmax
Maximum distance point from the hydrocarbon discharge boundary in study area
Hydrocarbon expulsion boundary
L
LB
Target
A
A'
Reservoir
Cap rock
Source rock
Hydrocarbon expulsion boundary
Hydrocarbon expulsion center

Fig. 2. Schematic of source index constructs related concepts [adapted from Jiang et al. [44]].

2.1.1. Construction method for source index (I_s)

The supply condition for hydrocarbon sources is an essential material basis and prerequisite for CHAs in petroliferous basins [40]. Based on the study on the characteristics of CHAs distribution in Mesozoic and Cenozoic continental basins in eastern China, Hu [29] found that oil and gas accumulations are mainly distributed in a certain range around hydrocarbon-generating depressions, and only 10 %–20 % of the oil and gas accumulations are distributed far away from the hydrocarbon-generating depressions. Similar studies have recognized the "source controlling theory," i.e., oil and gas reservoirs are mainly distributed in adjacent source rock areas [29–32]. To accurately define the location of the source rock center and the distribution range of source rocks, the concept of "hydrocarbon expulsion intensity" is proposed and used to quantitatively characterize the hydrocarbon expulsion amount per unit area of source rock [32,43]. For a set of source rocks, the region with maximum hydrocarbon expulsion intensity is the source rock center, and the range of hydrocarbon expulsion intensity equal to 0 is the source rock boundary (Fig. 2). In terms of single HA element for source condition, the closer the reservoir is to the source rock, the greater the probability of being supplied by the oil source, which also means a higher HAP [44]. Thus, the index that can characterize the distance to source rock has an intrinsically strong relationship with HAP.

Based on the source controlling theory, we propose a method to construct I_s. First, the oil source correlation work is done for the target reservoir to clarify the source rock that provides the hydrocarbon source for each CHA. It is a noteworthiness point that we only consider the conditions within a single petroleum system in this section, which means all CHAs generated by one source rock. And the application of the method to basins with multiple petroleum systems will be discussed in the 'discussion' section. Second, the distribution characteristics of hydrocarbon expulsion intensity of the source rock that provides hydrocarbon source for oil and gas accumulations in the target reservoir are calculated, and the source rock center and boundary are determined. Finally, I_s at any point on the target reservoir is calculated using Eq. (1). Fig. 2 shows the key parameters in the formula.

$$I_s = \begin{cases} \dfrac{\mathrm{L}}{2\mathrm{L_b}} & (\mathrm{L} \leqslant \mathrm{L_b}) \\[2ex] \dfrac{\mathrm{L} + \mathrm{L_{max}} - 2\mathrm{L_b}}{2(\mathrm{L_{max}} - \mathrm{L_b})} & (\mathrm{L_b} < \mathrm{L} \leqslant \mathrm{L_{max}}) \end{cases} \tag{1}$$

where I_s denotes the source index, dimensionless; L denotes the distance between any point on the target reservoir and the source rock center, Km; L_b denotes the distance from the source rock center to the boundary in the L direction, Km; and L_{max} denotes the farthest distance from the source rock center to the target reservoir.

When the oil and gas accumulations are located within the source rock boundary, I_s is between 0 and 0.5, and the closer the accumulations are to the source rock center, the I_s is closer to 0. When the oil and gas accumulate outside the source rock boundary, I_s is between 0.5 and 1, and the closer the accumulations are to the source rock boundary, the closer I_s is to 0.5. I_s can accurately reflect the proximity of each location of the target reservoir to the source rock and hence can characterize the oil source supply conditions of each location of the target reservoir, i.e., the smaller the oil source index, the better the oil source conditions.

2.1.2. Construction method for relative potential index (I_{rp})

The secondary migration of oil and gas deeply affects the forming process of CHAs from the source rock to trap and directly determines the gather environment and distribution for CHAs [7,33,39]. Fluid potential, which was introduced in the petroleum geology field by Hubbert [34], is the sum of mechanical energy per unit of mass fluid relative to the reference plane. The distribution of CHAs in different fluid potential regions reflects different oil and gas gather environments. In the migration system with strong connectivity, oil and gas are unstable in the high fluid potential region and tend to migrate and gather in the low fluid potential region [13]. Under this case, a series of CHAs with anticline and buried hill trap as gather environment will be formed in the low fluid potential region [13]. It is difficult for oil and gas to migrate to the low fluid potential region in a migration system with poor connectivity [33]. In this case, a series of CHAs based on lithologic traps will be formed in the high potential region [33]. Notably, the low fluid potential region of oil and gas accumulation is a local and relative concept [35]. The region whose fluid potential is lower than that of any adjacent region will form a local low-energy barrier obstructing fluid flow [35], called the local low potential region (LLPR). The boundary of LLPR, called the local low potential trap boundary (LLPTB), is the closed fluid potential contour with the highest fluid potential in LLPR or the fluid potential contour with the highest fluid potential closed by combination with closed faults. The minimum fluid potential point in LLPR is called the local lowest potential point (LLPP). As shown in Fig. 3a, points A–D have different absolute fluid potentials, following the order point A > point B > point C > point D. However, these points are LLPPs and local low-energy barriers and can provide a low-energy area for oil and gas accumulations (Fig. 3a). Therefore, compared with the absolute fluid potential, the relative fluid potential can more effectively reflect the gather environment of oil and gas. For the hydrocarbon accumulating process in certain target reservoirs in a single petroleum system, all oil and gas accumulations formed in the target reservoir have similar hydrocarbon source and migration conditions. There is a large probability that these oil and gas accumulations form with similar gather environment. Therefore, a prior probability distribution for HAP related to the index that can accurately reflect relative fluid potential can be established based on numerous drilled samples, which can be used to predict HAP for unexplored areas [3].

Based on the above ideas of relative fluid potential, the relative potential index (I_{rp}) is constructed (Fig. 3). First, the fluid potential on the top surface of the study reservoir is calculated, and the contour map of fluid potential distribution is obtained. Second, each LLPP and LLPTB is recognized by a human or program module. Finally, I_{rp} is calculated in each LLPTB using Eq. (2) (including LLPTB) and outside all LLPTB using Eq. (3) (excluding LLPTB).

$$I_{rp} = \frac{2P_{lb}^{k} - P_{min}^{k} - P^{k}}{2\left(P_{lb}^{k} - P_{min}^{k}\right)}, \tag{2}$$

where P^k is the absolute fluid potential at any position in the kth LLPR, J/kg; P_{min}^{k} denotes the minimum fluid potential in the kth LLPR, J/kg; and P_{lb}^{k} denotes the fluid potential value of LLPTB in the kth LLPR, J/kg.

$$I_{rp} = \frac{P_{max} - P}{2(P_{max} - P_{min}^{lb})}, \tag{3}$$

where P denotes the fluid potential at any position outside LLPR, J/kg; P_{max} denotes the maximum fluid potential outside LLPR, J/kg; and P_{min}^{lb} denotes the minimum fluid potential of each LLPTB, J/kg.

Utilizing the above method, the I_{rp} of the profile in Fig. 3(a) is calculated, which is shown in Fig. 3b. The result shows that I_{rp} ranges from 0 to 0.5 between the maximum fluid potential point and LLPTB, that is, as the position is closer to LLPP, I_{rp} is closer to 1. Meanwhile, I_{rp} ranges from 0 to 0.5 outside LLPTB, that is, as the position is closer to LLPTB, I_{rp} is closer to 0.5. I_{rp} is a good index for describing the relative fluid potential of any position in the target reservoir. In addition, when sealing faults are developed in a study area, their influence should also be considered. Because of the barrier effect of a sealing fault, the petroleum potentials in the two walls of the fault have no communication [45]. In this case, I_{rg} should be calculated independently on both sealing fracture walls.

2.1.3. Construction method for porosity index (I_φ)

Reserve condition is a crucial element for CHAs forming [40–42]. Physical properties, such as porosity and permeability, are critical

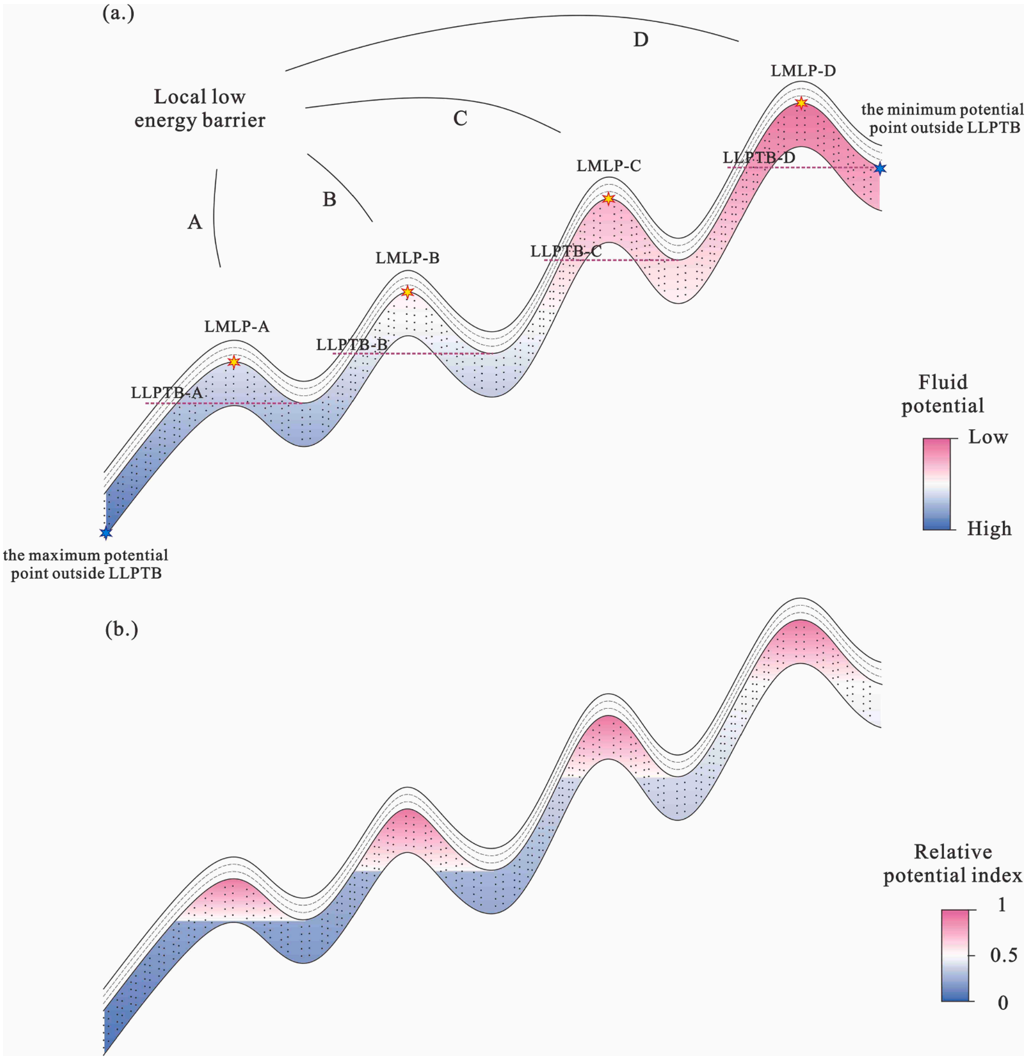


Fig. 3. (a) Relationship between absolute fluid potential distribution and local low potential barrier [adapted from Dahlberg [35]]. (b) Result diagram for relative potential index construction.

indexes commonly used to evaluate the goodness of a reservoir. However, there is almost no obvious correlation between physical properties and HAP because the physical property interval suitable for oil and gas accumulation changes with buried depth [40,41,46]. In terms of porosity, Pang et al. [40] established the statistical relationship between the porosity of oil and gas accumulation and their buried depth in eight major petroleum-bearing basins in China, showing that there is an obvious porosity lower limit for oil and gas accumulations, and the lower limit gradually decreases with an increase in burial depth. In addition, when porosity is greater than the porosity lower limit, some scholars have found that there is a good positive correlation between reservoir porosity and HAP [46–49] (Fig. 4a), whereas some others have shown that the HAP first increases and then decreases with an increase in reservoir porosity [41] (Fig. 4b). Therefore, in a relatively homogeneous reservoir section, the relationship between HAP and reservoir porosity may be one of the following two types of numerical models.

(1) Model a (labeled M_a). HAP gradually increases with an increase in reservoir porosity. In this numerical relationship, the shape of a statistical map for reservoir porosity versus HAP resembles a left-half-normal distribution (Fig. 4a).
(2) Model b (labeled M_b). HAP first increases and then decreases with an increase in reservoir porosity. In this numerical relationship, the shape of a statistical map for reservoir porosity versus HAP resembles an entire normal distribution (Fig. 4b). Moreover, the numerical relationship model between reservoir porosity and HAP may still differ in different reservoirs or the same reservoir with different depth ranges [10,11].

According to the numerical relationship between reservoir porosity and HAP, which have a good positive correlation, the porosity index (I_φ) is constructed based on drilled samples in the target reservoir (Fig. 4c). First, divide the target reservoir into several small depth interval reservoirs (reservoir unit) (Fig. 4c). The principle of dividing reservoir units is that lessening the depth interval of each reservoir unit as much as possible under the premise of the number of drilled samples in each reservoir unit is suitable to fit the normal distribution model. Then, fit the numerical relationship model between reservoir porosity and HAP by the normal distribution model [Eq. (4)] in each reservoir unit and identify the model type for the numerical relationship. In this process, the nonlinear fitting method used to obtain fitting parameters is the least square method in this study; still, other nonlinear fitting methods (such as the gradient descent method) may also be used. There are many parameters to characterize the fitting error. Coefficients of determination (R^2) are used to present the fitting error in this study. The numerical model of a reservoir unit will be identified as M_a when the maximum porosity of drilled samples in this reservoir unit is less than$\mu + \sigma$of its normal distribution; otherwise, it will be identified as M_b. The reservoir

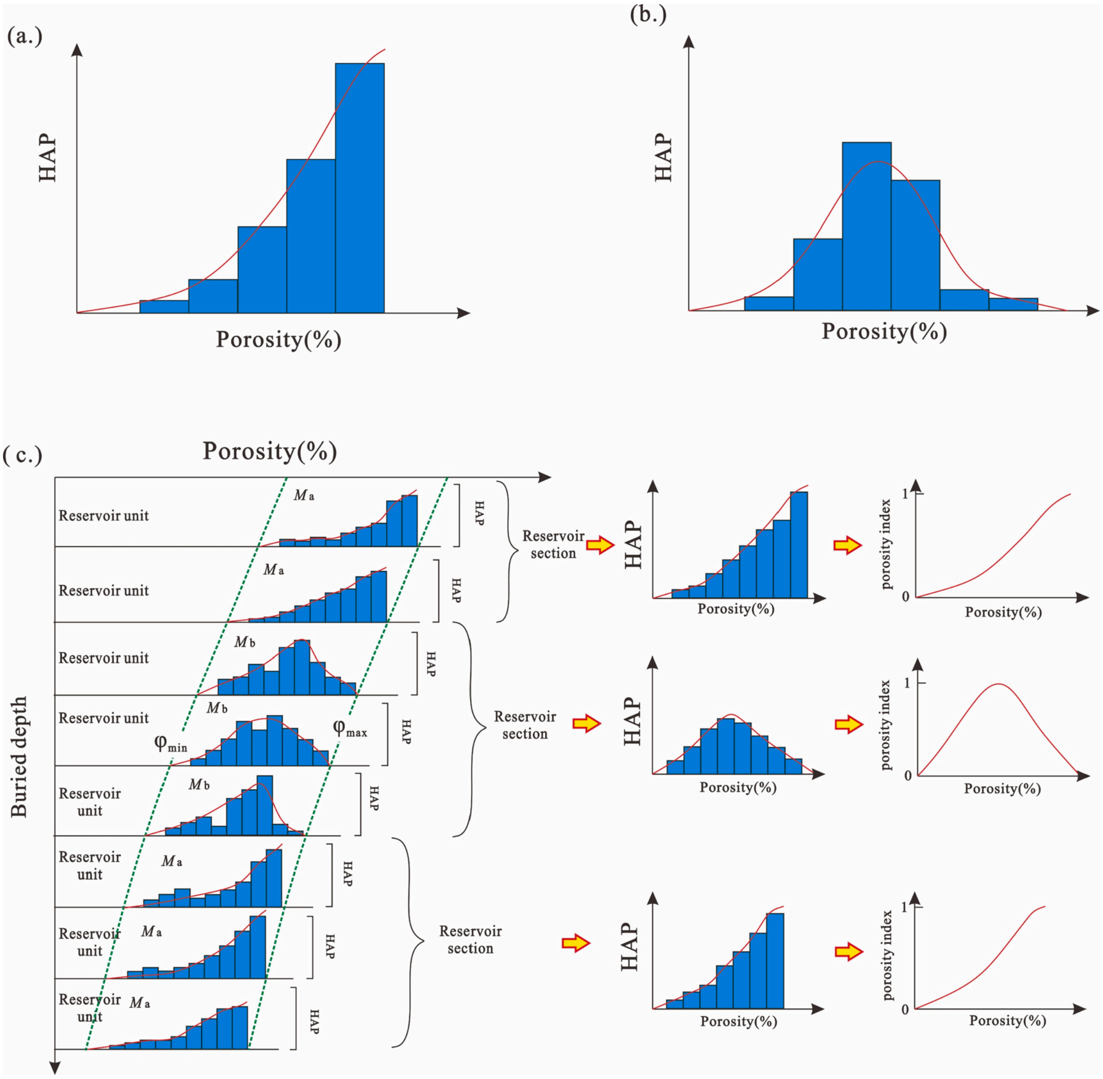


Fig. 4. (a) M_a and (b) M_b (adapted from Huo et al. [41]); (c) I_φ construction process including reservoir unit division, reservoir section distinguishment, and I_φ construction.

unit with continuous burial depth and the same type of numerical relationship model are combined into a reservoir section (Fig. 4c), and the numerical relationship model between reservoir porosity and HAP is constructed using the normal distribution model [Eq. (4)] in each reservoir section. Finally, the numerical relationship model between reservoir porosity and HAP in each reservoir section is normalized, which is then taken as I_φ, the normalize function, as shown in Eq. (5). It should be noted that the reservoir section is a combination of reservoir units with similar burial depth and similar relationships between porosity and HAP. Thus, in different reservoir sections, the relationships between reservoir porosity and HAP are obviously different. The purpose of the reservoir section division is the more accurate characterization of the relationship between porosity and HAP. However, this does not mean that the type of numerical relationship model of a reservoir section is necessarily the same as that of the reservoir units that compose it. Moreover, even if there are different types of numerical relationship models between them, this does not violate the original intention of the reservoir section division. In essence, the prior probability distribution for HAP related to porosity in each reservoir section is constructed by the above steps based on drilled samples. Because I_φ is the output of normalized prior probability distribution for HAP in each reservoir section, there is a good positive correlation between I_φ and HAP. In addition, the coefficient of determination (R^2) between I_φ and HAP is affected by the fit error of each prior probability distribution.

$$HAP_i = \frac{1}{\sqrt{2\pi}\sigma_i} \times e^{\left(-\left(\frac{\varphi-\mu_i}{\sqrt{2}\sigma_i}\right)^2\right)} \tag{4}$$

where HAP_i represents the HAP of the i^th reservoir unit or section, dimensionless; φ denotes the porosity of the i^th reservoir unit or section, %; and σ_i and μ_i are the parameters of the i^th reservoir unit or section to be fitted, representing the standard deviation and mean of φ, respectively, both in %.

$$I_\varphi = \frac{HAP_j}{HAP_{jmax}} \tag{5}$$

where HAP_j denotes the HAP of the j^th reservoir section, dimensionless, and HAP_{jmax} denotes the maximum HAP of the j^th reservoir section, dimensionless.

2.1.4. Construction method for permeability index (I_k)

Permeability is another paramount parameter for the physical properties of reservoirs [48]. In most instances, the porosity and the logarithm of permeability have a good linear relationship [48], which means that the relationship between the logarithm of reservoir permeability and HAP is similar to that between reservoir porosity and HAP in most cases. However, under the influence of diageneses, such as cementation and recrystallization, the throat in the pore network may be blocked, which will cause the deterioration of the linear correlation between porosity and the logarithm of permeability [50]. In this case, the relationship between pairs of porosity, permeability, and HAP will differ. Therefore, the relationship between permeability and HAP needs to be considered separately.

I_k is constructed using a method similar to the I_φ construction. The difference between the construction methods is that the former is completed by the logarithm of permeability. The numerical relationship model between reservoir permeability and HAP is fitted by Eq. (6) based on the normal distribution model in each reservoir unit or section. The final I_k construction formula is shown in Eq. (7).

$$HAP_i = \frac{1}{\sqrt{2\pi}\sigma_i} \times e^{\left(-\left(\frac{ln_{(k)}-\mu_i}{\sqrt{2}\sigma_i}\right)^2\right)} \tag{6}$$

where k denotes the permeability of the ith reservoir unit or section, mD, and σ_i and μ_i are the parameters of the ith reservoir unit or section to be fitted, representing the standard deviation and mean of k, respectively, both in mD.

$$I_k = \frac{HAP_j}{HAP_{jmax}} \tag{7}$$

where HAP_j denotes the HAP of the jth reservoir section, dimensionless, and HAP_{jmax} denotes the maximum HAP of the jth reservoir section, dimensionless.

2.2. HAP prediction model construction

2.2.1. Dataset construction

Although four features of drilled reservoir samples have been constructed, their HAP has not been quantitatively characterized. Some continuous variables, such as hydrocarbon saturation [17], hydrocarbon reserves [6], and hydrocarbon production [18], are considered to characterize HAP for drilled reservoir samples. This is a regression problem for ML algorithms trained on the dataset comprising four features and continuous HAP of drilled reservoir samples. However, a large proportion of data is a nonhydrocarbon layer whose HAP is zero in the exploration field dataset, which will significantly affect the performance of the regression ML prediction model [51]. In addition, the regression problem for imbalanced datasets is not mature, and only a few studies focus on this topic [51]. However, there are numerous study results of classification problems for imbalanced datasets [20,21,52]. Therefore, a dispersed variable is used to quantitatively characterize HAP, i.e., HAP is equal to 1 and 0 for hydrocarbon and nonhydrocarbon layer samples, respectively. The final dataset comprises four features, i.e., I_s, I_{rp}, I_φ, and I_k, and one label including two categories (0 and 1). A suitable ML algorithm will be selected to solve the classification problem on this dataset.

2.2.2. RUSBoosted tree ML algorithm principle

The application of the RUSBoosted tree method to unbalanced datasets has achieved excellent performance [53]. RUSBoost algorithm is an ensemble ML algorithm introduced by Seiffert et al. [21] to improve the accuracy of classifier problems on unbalanced datasets. The ensemble ML method accomplishes the classification task by training many base classifiers. Base classifiers are commonly trained using decision tree (DT), support vector machine (SVM), and neural network algorithms [21]. Compared with neural network algorithms, DT and SVM are more suitable for training small datasets of natural science. Because both DT and SVM have good interpretability and can easily find the decision boundaries of the prediction model, we can more easily understand the principle of the prediction model work through model decomposition [54,55]. But this is difficult when using multilayer neural network algorithms [56]. In addition, the application effect of SVM mainly depends on the type of kernel functions, and selecting the optimal kernel function is always a complex process [55]. Because DT has good decomposition and applicability, it is used as the base classifiers in most RUSBoost algorithm studies. Therefore, the DT algorithm is selected to train the base classifier in the RUSBoost algorithm in this study. Therefore, RUSBoosted tree is the name of the final algorithm. The DT algorithm solves the classification task according to the information gain principle, whose theoretical research is mature (see Ross Quinlan [54] for the detail). The RUSBoost algorithm trains an unbalanced dataset by the random undersampling method, which makes the dataset balanced by randomly selecting a certain number of majority category samples and mixing them with all minority category samples until the desired mixture rate is achieved. Many balanced subdatasets are selected randomly from the original unbalanced dataset, each of which is trained using the DT model as a base classifier iteratively. After each base classifier is trained, the weight of each sample will be changed to correctly classify samples in the base classifier. The final ensemble model is constructed by all base classifiers with a weighted vote. The RUSBoost algorithm's training is efficient on imbalanced datasets [20,21].

Considering the dataset of this study as an example, the specific implementation process of the RUSBoosted tree algorithm is as follows:

① Select DT as a base classifier and set the balance rate (BalRate) of subdataset and iteration time (T).

② Input original dataset: D = [(x_1, y_1), (x_2, y_2), ……(x_n, y_n)], where n represents the number of samples in the original dataset; x_i denotes a feature vector composed by I_s, I_{rp}, I_φ, and I_k of the ith sample in the original dataset; and y_i denotes the label of the ith sample in the original dataset (HAP equals 0 or 1).

③ Set initial weight for each sample of the original dataset:

$$w_0(i) = \frac{1}{n}$$

④ Begin iteration, assuming the iteration time is t, t ranges from 1 to T.

a. Construct subdataset (D_t) using the undersampling method; the subdataset comprises all hydrocarbon layer samples (the number of it is m) and (m × BalRate) nonhydrocarbon layer samples.
b. Train the tth base classifier (h_t) using DT on D_t and its samples weight (w_{t-1}).
c. Calculate the loss (ϵ_t) of h_t:

$$\epsilon_t = \sum_{(i,y):yi\neq y} w_{t-1}(i)(1 - h_t(x_i, y_i) + h_t(x_i, y))$$

d. Calculate the weight (α_t) of h_t:

$$\alpha_t = \frac{\epsilon_t}{1 - \epsilon_t}.$$

e. Update the weight of the original dataset (w_t):

$$w_t(i) = w_{t-1}(i)\alpha t^{0.5(1+h_t(x_i,y_i)-h_t(x_i,y:y\neq y_i))}.$$

f. Normalize the weight of w_t:

$$w_t(i) = \frac{w_t(i)}{\sum_i w_t(i)}$$

⑤ Finish iteration and establish the final ensemble classifier [H(x)]:

$$H(x) = sign\left(\sum_{t=1}^{T} h_t(x, y)^* log\frac{1}{\alpha_t}\right).$$

The DT and RUSBoost algorithms used in this study are both from modules of MATLAB R2019a. In order to avoid over-fitting of each DT base classifiers caused by too many branches, the maximum splitting number of the DT algorithm is set to 20. The number of base classifiers was set to 30 by comprehensively considering the calculation amount (optimization time) and the change trend of model loss with the number of base learners And balance rate takes the default value of the RUSBoost algorithm in MATLAB R2019a.

2.3. Favorable zones prediction strategy

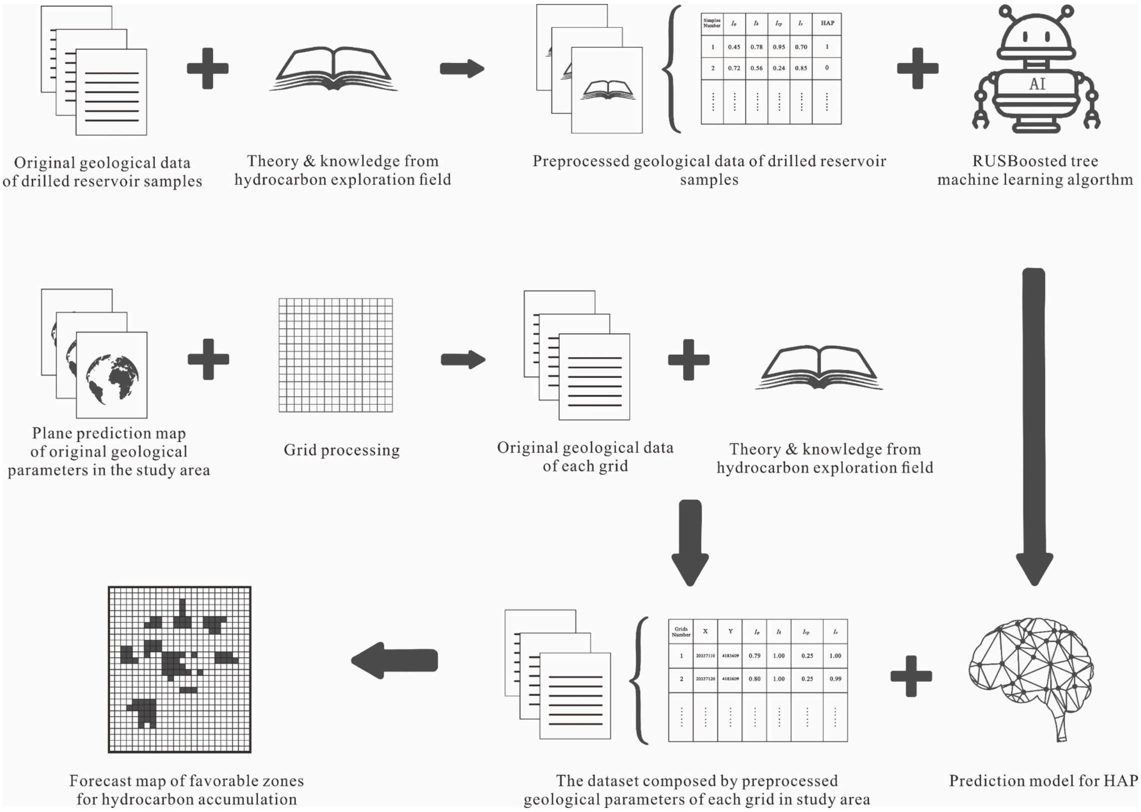
The HAP prediction model will be trained using the RUSBoosted tree algorithm on the final dataset. Based on the HAP prediction, a strategy of favorable zone prediction for unexplored areas in the target reservoir is proposed. First, the unexplored areas in the target reservoir were evenly divided into many grids. Second, according to the porosity distribution prediction map, permeability distribution prediction map, fluid potential distribution prediction map, and hydrocarbon expulsion intensity map, the predicted geological parameters of each grid are obtained, including the mean of porosity, permeability, fluid potential, and the distance from the source rock center and boundary. Finally, the HAP of each grid in the target reservoir is predicted using the HAP prediction model, and the prediction map for favorable zones is formed by filling color to the grids whose HAP prediction value is 1.

Fig. 5 shows the comprehensive technical guide of the entire research method in this study.

3. Geological setting

Jizhong Depression is located in the west of Bohai Bay Basin (Fig. 6a), which was formed during the Late Cretaceous to Paleogene [57,58]. Shulu Sag, located in the southeast corner of Jizhong Depression (Fig. 6a), is a paramount petroliferous sag in Jizhong Depression [59]. Shulu Sag, covering an area of approximately 700 Km^2, is adjacent to Xinhe Fault in the east and Leizjiazhuang Fault in the south with a northeast–southwest trend [60] (Fig. 6a). It can be divided into two structural units the western sag belt and eastern slope belt which have oil and gas accumulations discovered in the south (Fig. 6b). Paleogene to Quaternary Formations are widely deposited in Shulu Sag. From the bottom to top are Paleogene Shahejie Formation (Es), Paleogene Dongying Formation (Ed), Neogene Guantao Formation (Ng), Minghuazhen Formation (Nm), and Quaternary Pingyuan Formation (Qp). Es and Ed are the studied formations; Es can be further divided into three parts: the third member of Es (Es_3), the second member of Es (Es_2), and the first member of Es (Es_1), from bottom to top (Fig. 6c). In terms of lithology, the bottom of Es_3 is mainly dark mudstone accompanied by a small amount of carbonate rock and argillaceous limestone, and the top is mainly interbedded with sand and mudstone. Es_2 is filled with pebble sandstone interbedded rock of sandstone and thin mudstone. A small amount of carbonate rocks is deposited at the bottom of Es_1, which gradually transition upward to dark mudstone and sand mudstone interbedding. Ed mainly comprises thick sandstone and thin mudstone (Fig. 6c). The dark mudstone and carbonate rocks at Es_3 and the bottom of Es_1 are good rock sources in Shulu Sag [59,61] (Fig. 6c). However, Es_1 source rock has limited hydrocarbon generation due to its low maturity [58]. Therefore, Es_3 is the most important source of oil and gas in Shulu Sag [58,61]. A single petroleum system is composed by Es_3 source rock and Es and Ed reservoirs (Es_3-Es, Ed petroleum system). The tectonic evolution of Shulu Sag in the Paleogene can be divided into two stages: rifting period from 45 to 24.6 Ma and postrifting period after 24.6 Ma [62] (Fig. 6c).

In this study, Jin 93 Well Block in Shulu Sag was selected as the study area. Jin 93 Well Block is located in the CHAs enrichment area on the southern slope of Shulu Sag (Fig. 6b), and the CHAs whose type is mainly fault accumulation are mostly distributed in Ed and Es (Fig. 7). In the study area, many oil and gas reservoirs are continuously distributed along the footwall of the fault from shallow to deep (Fig. 7). These accumulations were mainly formed in the postrifting period, during which


Original geological data
of drilled reservoir samples
Theory & knowledge from
hydrocarbon exploration field
Preprocessed geological data of drilled reservoir
samples
AI
RUSBoosted tree
machine learning algorthm
Plane prediction map
of original geological
parameters in the study area
Grid processing
Original geological data
of each grid
Theory & knowledge from
hydrocarbon exploration field
Forecast map of favorable zones
for hydrocarbon accumulation
The dataset composed by preprocessed
geological parameters of each grid in study area
Prediction model for HAP

Simples Number	I_φ	I_k	I_{fp}	I_s	HAP
1	0.45	0.78	0.95	0.70	1
2	0.72	0.56	0.24	0.85	0
⋮	⋮	⋮	⋮	⋮	⋮

Grids Number	X	Y	I_φ	I_k	I_{fp}	I_s
1	20337110	4183609	0.79	1.00	0.25	1.00
2	20337120	4183609	0.80	1.00	0.25	0.99
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Fig. 5. Comprehensive technical guide for the research method.

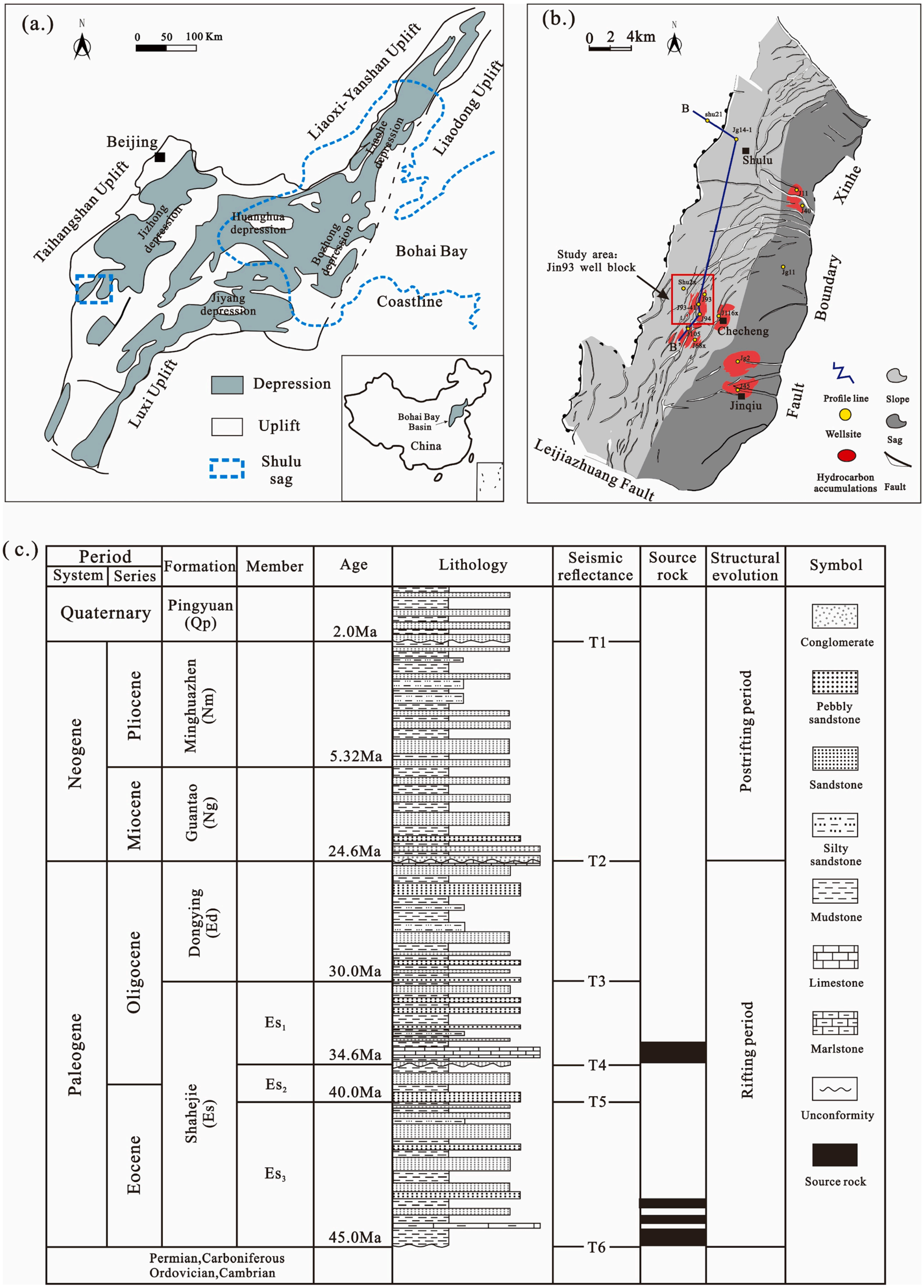


Fig. 6. (a) Distribution characteristics of main petroliferous depressions in Bohai Bay Basin [adapted from Jiang et al. [59]]; (b) distribution characteristics of structural units in Shulu Sag [adapted from Jiang et al. [59]]; (c) stratigraphic system and lithologic characteristics in Shulu Sag [adapted from Jiang et al. [59]].

the fault activity was weak and had excellent sealing performance [58]. Jin 93 Well Block in Shulu Sag has a high degree of exploration, considerable conventional oil and gas accumulation discovery, and rich data accumulation, which are suitable for ML research.

4. Data sources and characteristics

In this study, the HAP prediction model was established based on the drilling data of Es_1, Es_2, Es_3, and Ed in Jin 93 Well Block of Shulu Sag, and Es_2 is taken for favorable zone prediction. A total of 2,645 drilled

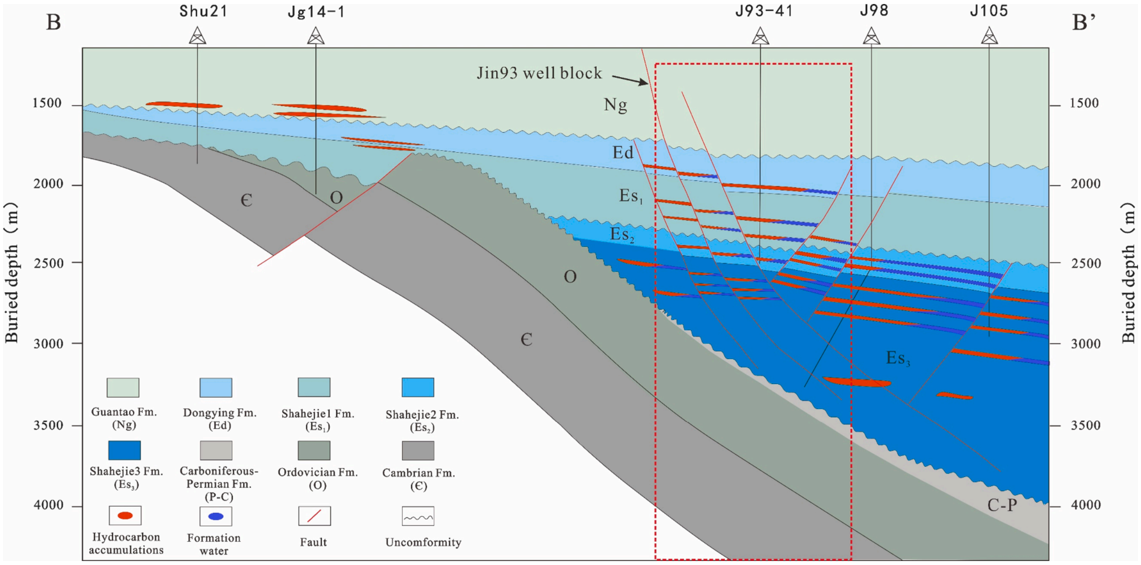


Fig. 7. The profile of Jin 93 Well Block in Shulu Sag [adapted from Jiang et al. [59]], the location of the profile is showed by B-B' line in Fig. 6.

reservoir samples from 47 exploration wells in Jin 93 Well Block in Shulu Sag were collected, and drilling data of each drilled reservoir sample include well name, drilling depth, drilling Formation, type of reservoir (hydrocarbon layer or nonhydrocarbon layer), porosity, permeability, argillaceous content, hydrocarbon expulsion intensity and distance from source rock center (Table S.1). In addition, contour structures maps, distribution maps of discovered accumulations, porosity distribution prediction maps, and permeability distribution prediction maps of Ed, Es_1, Es_2, and Es_3 were collected. These were collected from PetroChina Huabei Oilfield Company. The wells collecting drilling data were numerous and widely distributed in the study area (Fig. 8), and the data collected were representative. However, there were only 519 hydrocarbon layer samples, accounting for only 19.6 % of the entire collected data (Table S.1), attributable to the low success rate for oil and gas accumulation exploration. From 2009 to 2018, the exploration success rate of international oil companies was approximately 20 %–50 %, with mean of approximately 30 % [19]. Owing to the low success rate of exploration, the number of hydrocarbon layers drilled is far less than that of nonhydrocarbon layers, making the dataset of the petroleum exploration field imbalanced. In our dataset, the ratio of hydrocarbon layer samples to nonhydrocarbon layer samples is approximately 1:4 (Table S.1).

5. Result

The result of this study is also presented in three parts, i.e., data preprocessing, HAP prediction model construction, and favorable zone prediction results.

5.1. Data preprocessing result

5.1.1. Construction result for source index (I_s)

There are two sets of source rocks in Shulu Sag, among which the Es_3 source rock is the main source for oil and gas accumulations [61]. The shape of the hydrocarbon expulsion intensity contour is approximately a north–south ellipse (Fig. 9). Jin 93 Well Block is close to the hydrocarbon expulsion boundary, and the west and east are outside and inside the boundary, respectively (Fig. 9a and b). The characteristic of I_s distribution for the Es_3 source rock is acquired using the method described in Subsection 2.1.1. I_s within the hydrocarbon expulsion boundary of Jin 93 Well Block is less than 0.5, and I_s gradually decreased as closing to the hydrocarbon expulsion center and gradually increased as far away the source rock (Fig. 9c). There is a good consistency between I_s and the degree of adjacent source rocks.

5.1.2. Construction result for relative potential index (I_{rg})

The residual pressure of Es and Ed is mainly distributed in the sag center and gradually decreases to the slop region. Because Jin 93 Well Block is far from the sag center, the formation pressure of this area is normal to weak overpressure, indicating that the hydrodynamic forces have little effect on the fluid potential in this region [58]. The concept of petroleum energy proposed by England et al. [33] can quickly and accurately characterize the fluid energy of oil and gas under hydrostatic conditions without capillary force. The mathematical model is as follows:

$$E_o = v_0(\rho_w - \rho_o)gz \tag{8}$$

where E_o denotes the petroleum energy, J; v_0 denotes the volume of oil and gas, m^3; ρ_w denotes the density of formation water, kg/m^3; ρ_o denotes the density of oil and gas, kg/m^3; g denotes the gravitational acceleration, m/s^2; and z denotes the burial depth, m.

To obtain the potential of oil and gas per unit volume (called petroleum potential), we divide both sides of Eq. (8) by v_0, as follows:

$$\phi_o = (\rho_w - \rho_o)gz \tag{9}$$

where ϕ_o denotes the petroleum potential of oil and gas, J/m^3.

The absolute petroleum potential of Es_3, Es_2, Es_1, and Ed is calculated using Eq. (10), and the characteristic of I_{rg} distribution in these four formations is obtained according to method shown in Subsection 2.1.2. Owing to the barrier effect of sealing fault, the petroleum potential in the two walls of the fault cannot be communicated [45]. The faults in Jin 93 Well Block are mainly sealing faults during oil and gas accumulation [58]. Therefore, in this study, I_{rg} was calculated independently on both sealing fracture walls.

Compared with the absolute petroleum potential, I_{rg} can more accurately characterize the oil and gas accumulation conditions at each structural location. The petroleum potential of Es_2 is lower in the east and higher in the west. Six LLPRs are developed, namely A, B, C, D, E, and F regions (Fig. 10a). Most oil and gas wells are drilled in these six areas, suggesting that these six areas have good conditions for HA elements, including oil and gas gather environment (Fig. 10a and 10b). However, the absolute petroleum potential has a good matching relationship with LLPR only in A and B. In the other four LLPRs with high absolute petroleum potential, the absolute petroleum potential loses its indication to LLPR (Fig. 10a and 10b). The I_{rg} of these six LLPRs and trap

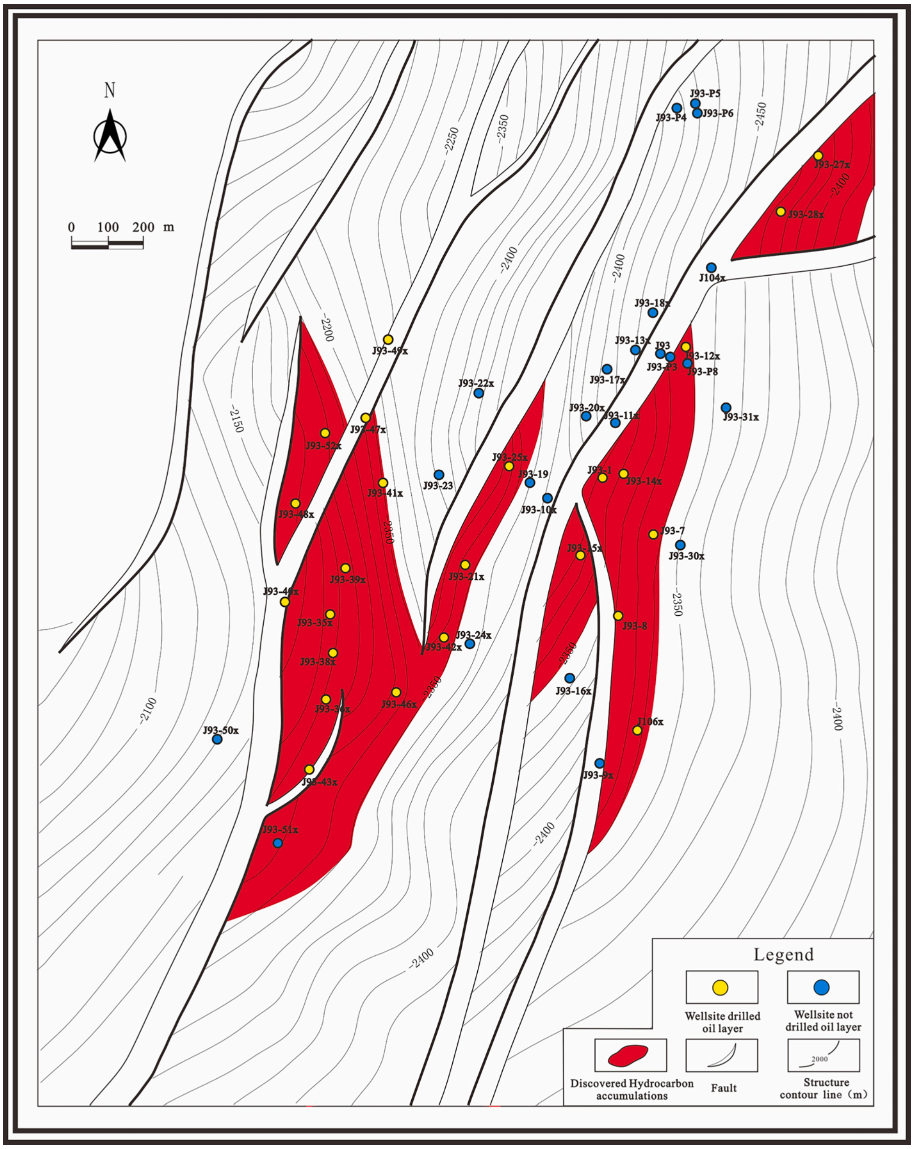


Fig. 8. Structural contours and oil and gas accumulations distribution map of Es_2 and sampling well location map of this study.

range of all low potential energy areas of the Es_2 reservoir are close to 1 and greater than 0.5, respectively, which has a good indicative function for LLPR and is a good index for oil and gas gather environment. Based on the drilled reservoir samples in the study area, the prior probability distribution for HAP related to the index is created by counting the HAP of each I_{rg} interval (Fig. 10). The real value of HAP can be approximated using a statistical method, and it is estimated using Eq. (10) in this study. Overall, the HAP of each I_{rg} interval increases gradually with an increase in I_{rg}. HAP reaches the maximum value when it is greater than 0.7, suggesting that there is a large probability of oil and gas gathering in the relatively low potential environment of the study area. Moreover, HAP ranges from 0.04 to 0.16 when I_{rg} ranges from 0.2 to 0.7, indicating the possibility of finding oil and gas accumulations in relatively high potential environments, such as lithologic traps, but its probability is lower.

$$HAP = \frac{N_{HL}}{N_{NHL} + N_{HL}} \tag{10}$$

where HAP represents the HAP of a reservoir, dimensionless; N_{HL} denotes the number of hydrocarbon layer discovered in the reservoir, unit; and N_{NHL} denotes the number of nonhydrocarbon layers discovered in the reservoir.

5.1.3. Construction result for porosity index (I_φ)

According to the distribution characteristics of the number of samples in different burial depth intervals, 50 m is taken as the depth interval for each reservoir unit. The numerical relationship between porosity and HAP in each reservoir unit is fitted. The type of each numerical model is then distinguished, and the reservoir sections are divided (Fig. 11a). Finally, the numerical model for I_φ construction is established.

Ed and Es can be divided into three reservoir sections in Jin 93 Well Block of Shulu Sag: 1,784.7–2,234.7 m with M_a, 2,234.7–2,734.7 m with M_b, and 2,734.7–3,084.7 m with M_a (Fig. 11a). In addition, there is a good Gaussian distribution relationship between I_φ and HAP in each reservoir section, and their coefficient of determination (R^2) is 0.99, 0.989, and 0.989, respectively (Fig. 11a). The numerical model for I_φ construction in each reservoir section is obtained by normalizing the numerical relationship model between reservoir porosity and HAP; the specific mathematical expression given by Eq. (11). The I_φ of each sample in the dataset is calculated using Eq. (11), showing a good positive correlation with HAP (Fig. 11b).

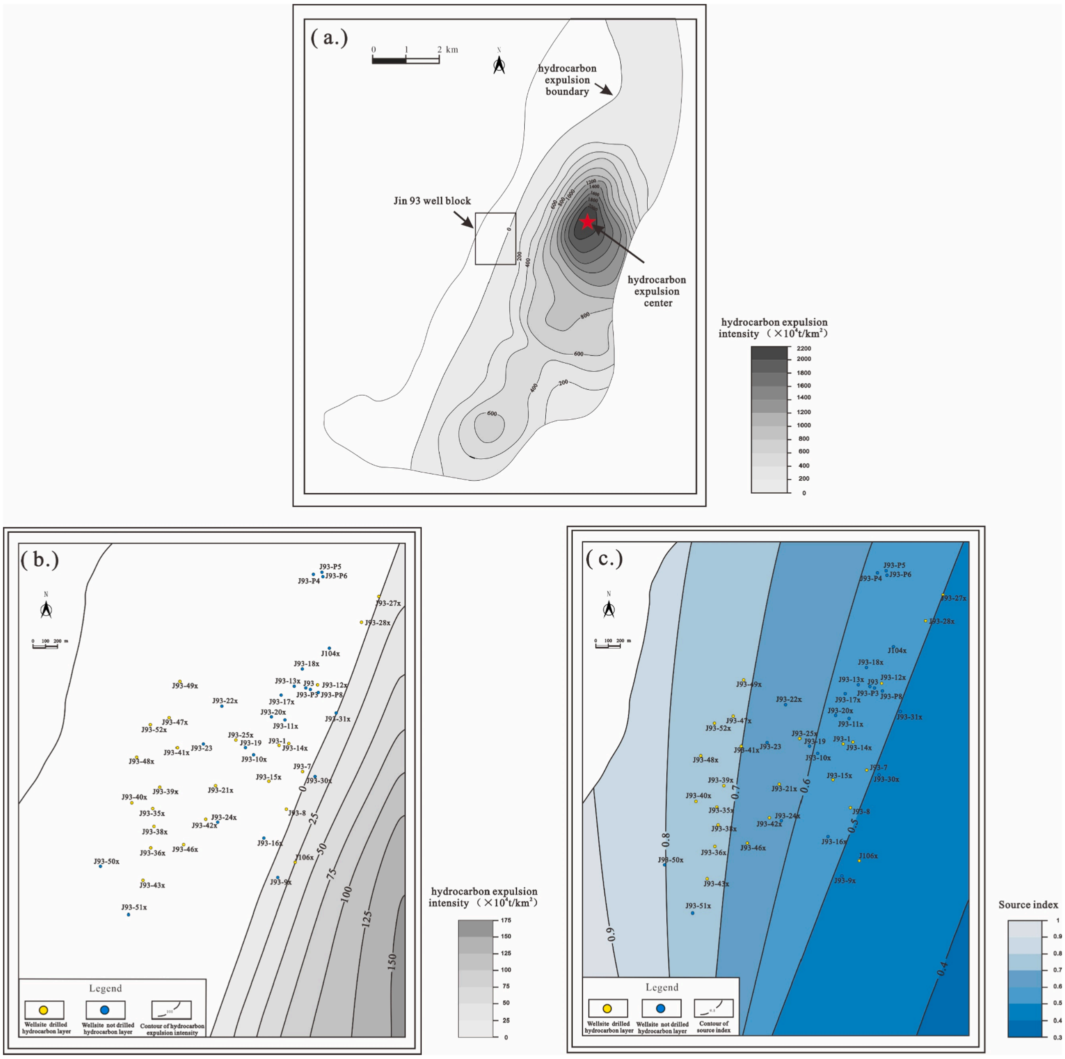

(a.)
0 1 2 km
N
hydrocarbon expulsion boundary
Jin 93 well block
hydrocarbon expulsion center
hydrocarbon expulsion intensity （×10⁴t/km²）
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
(b.)
N
0 100 200 m
hydrocarbon expulsion intensity （×10⁴t/km²）
175
150
125
100
75
50
25
0
Legend
Wellsite drilled hydrocarbon layer
Wellsite not drilled hydrocarbon layer
Contour of hydrocarbon expulsion intensity
(c.)
Source index
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Legend
Wellsite drilled hydrocarbon layer
Wellsite not drilled hydrocarbon layer
Contour of source index

Fig. 9. (a) Distribution map for hydrocarbon expulsion intensity of Es$_3$ source rocks of Shulu Sag [adapted from Huo et al. [61]]; (b) distribution map for hydrocarbon expulsion intensity of Jin 93 Well Block; (c) distribution map for I_s of Jin 93 Well Block.

$$I_\varphi = \begin{cases} e^{-\left(\frac{\varphi-28.93}{7.077}\right)^2} & (1784.7m \leqslant H < 2234.7m) \\ e^{-\left(\frac{\varphi-17.11}{7.128}\right)^2} & (2234.7m \leqslant H < 2734.7m) \\ e^{-\left(\frac{\varphi-21.16}{7.754}\right)^2} & (2734.7m \leqslant H \leqslant 3084.7m) \end{cases} \tag{11}$$

where H denotes the reservoir depth, m.

5.1.4. Construction result for permeability index (I_k)

The depth interval of each reservoir unit is still 50 m for I_k construction. There are also three reservoir sections for I_k, which are slightly different from reservoir sections of I_φ in depth interval (Fig. 12a). Reservoir sections for I_k are 1,784.7–2,284.7 m with M_a, 2,284.7–2,784.7 m with M_b, and 2,784.7–3,084.7 m with M_a (Fig. 12a). There is also a good Gaussian distribution relationship between I_k and HAP in each reservoir section, and their decision index (R^2) is 0.997, 0.989, and 0.989, respectively. The final calculation model of I_k is as follows:

$$I_\varphi = \begin{cases} e^{-\left(\frac{\ln k-2.806}{0.9952}\right)^2} & (1784.7m \leqslant H < 2284.7m) \\ e^{-\left(\frac{\ln k-1.576}{0.8486}\right)^2} & (2284.7m \leqslant H < 2784.7m) \\ e^{-\left(\frac{\ln k-2.348}{1.237}\right)^2} & (2784.7m \leqslant H \leqslant 3084.7m) \end{cases} \tag{12}$$

Similar to the relationship of I_φ versus AHP, I_k calculated using Eq. (12) also has a good correlation with AHP (Fig. 12b). Owing to the diagenesis, there is no good consistency between the porosity and logarithm of permeability in Jin 93 Well Block, causing a slight difference in the depth interval of the reservoir sections between I_φ and I_k. This also suggests that it is necessary to characterize the reserve condition for oil and gas accumulation by I_φ and I_k together.

5.2. HAP prediction model construction result

5.2.1. HAP prediction model construction

The 2,645 collected drilled reservoir samples in Jin 93 Well Block were preprocessed, and the dataset was established by considering I_s, I_{rp}, I_φ, and I_k of each drilled reservoir sample as features and considering the HAP (0 or 1) of each drilled reservoir sample as a label (Table S.2). Based on the dataset, the final prediction model for HAP is obtained by training the RUSBoosted tree ML algorithm. Further, how can the goodness of this prediction model be evaluated? Accuracy—the ratio of the number of samples correctly predicted by the prediction model to the total number of tested samples is commonly used to evaluate the classification prediction model. However, accuracy cannot represent the real performance of a prediction model trained on an imbalanced dataset in most cases [20,21,52]. For example, assuming that the HAP prediction model constructed in this study predicts the HAP of all test samples to be

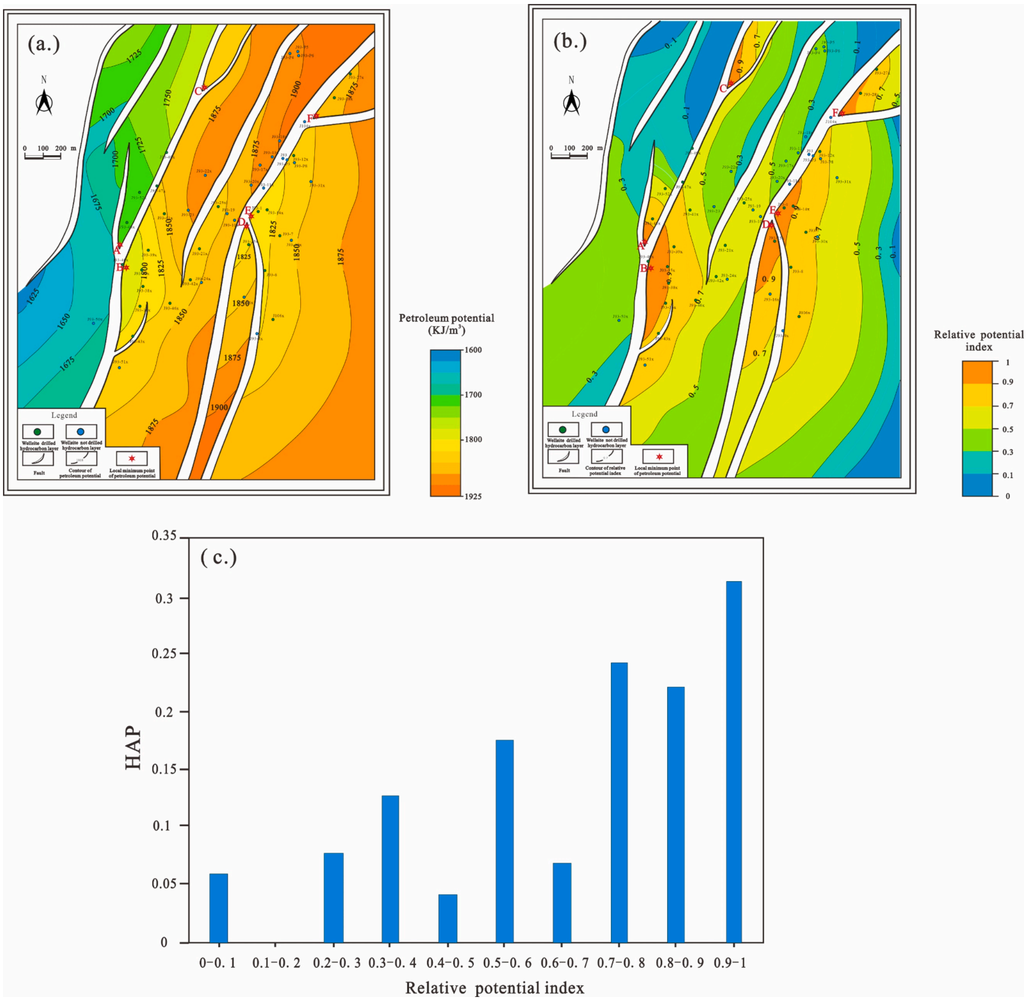


Fig. 10. (a) Absolute petroleum potential distribution of Es_2; (b) relative potential index distribution of Es_2; (c) relationship between relative potential index and AHP.

0, the model accuracy can reach 80.4 %, although the performance of this prediction model is terrible. Therefore, more effective indicators should be used to evaluate the performance of the HAP prediction model.

5.2.2. Evaluation system for HAP prediction model

The confusion matrix is introduced to display the test results for the HAP prediction model. The precision, recall, and F-measure method are employed to evaluate the prediction model. A confusion matrix is a matrix that intuitively shows the test results of a classification prediction model [52]. In this study, the confusion matrix divides the test results into four parts, i.e., true hydrocarbon layer (TH), true nonhydrocarbon layer (TNH), false hydrocarbon layer (FH), and false nonhydrocarbon layer (FNH) (Fig. 13). The TH layer refers to hydrocarbon layer samples predicted as hydrocarbon layer by the prediction model. The TNH layer refers to nonhydrocarbon layer samples predicted as nonhydrocarbon layer by the prediction model. The FH layer refers to hydrocarbon layer samples predicted as nonhydrocarbon layer. The FNH layer refers to nonhydrocarbon layer samples predicted as hydrocarbon layer by the prediction model (Fig. 13).

Precision and recall are obtained based on the number of the four parts in the confusion matrix, which are used to evaluate the ability of the HAP prediction model to predict hydrocarbon layers. Precision represents the probability that the predicted hydrocarbon layer is a real hydrocarbon layer, which can characterize the reliability of the HAP prediction model and can be calculated using Eq. (13) [20]. Recall represents the probability that the hydrocarbon layer sample is correctly predicted, which can characterize the proportion of the hydrocarbon layer samples predicted correctly to total hydrocarbon layer samples and can be calculated using Eq. (14) [20]. The precision and recall values range from 0 to 1, and the performance for the HAP prediction model is optimal when the precision and recall values are close to 1 simultaneously, which is impossible in most cases. Therefore, it is necessary to choose between a high recall rate and a high accuracy rate.

$$Precision = \frac{N_{TH}}{N_{TH} + N_{FH}} \tag{13}$$

$$Recall = \frac{N_{TH}}{N_{TH} + N_{FNH}} \tag{14}$$

where N_{TH}, N_{FH}, and N_{FNH} represent the numbers of TH, FH, and FNH layers, respectively, dimensionless.

F-measure can comprehensively evaluate the performance of the classification prediction model by integrating precision and recall and can be calculated using Eq. (15). Whether F-measure can effectively evaluate the performance of the classification prediction model depends on whether the appropriate weight (β) can be selected. Different fields pay different attention to recall and precision and select different β [20]. For example, in the medical field of tumor prediction, recall is much more imperative than precision, and β is greater than 1. For bad debt prediction in the field of bank loans, precision is much more imperative than recall, and β is less than 1. At present, there is no recognized value

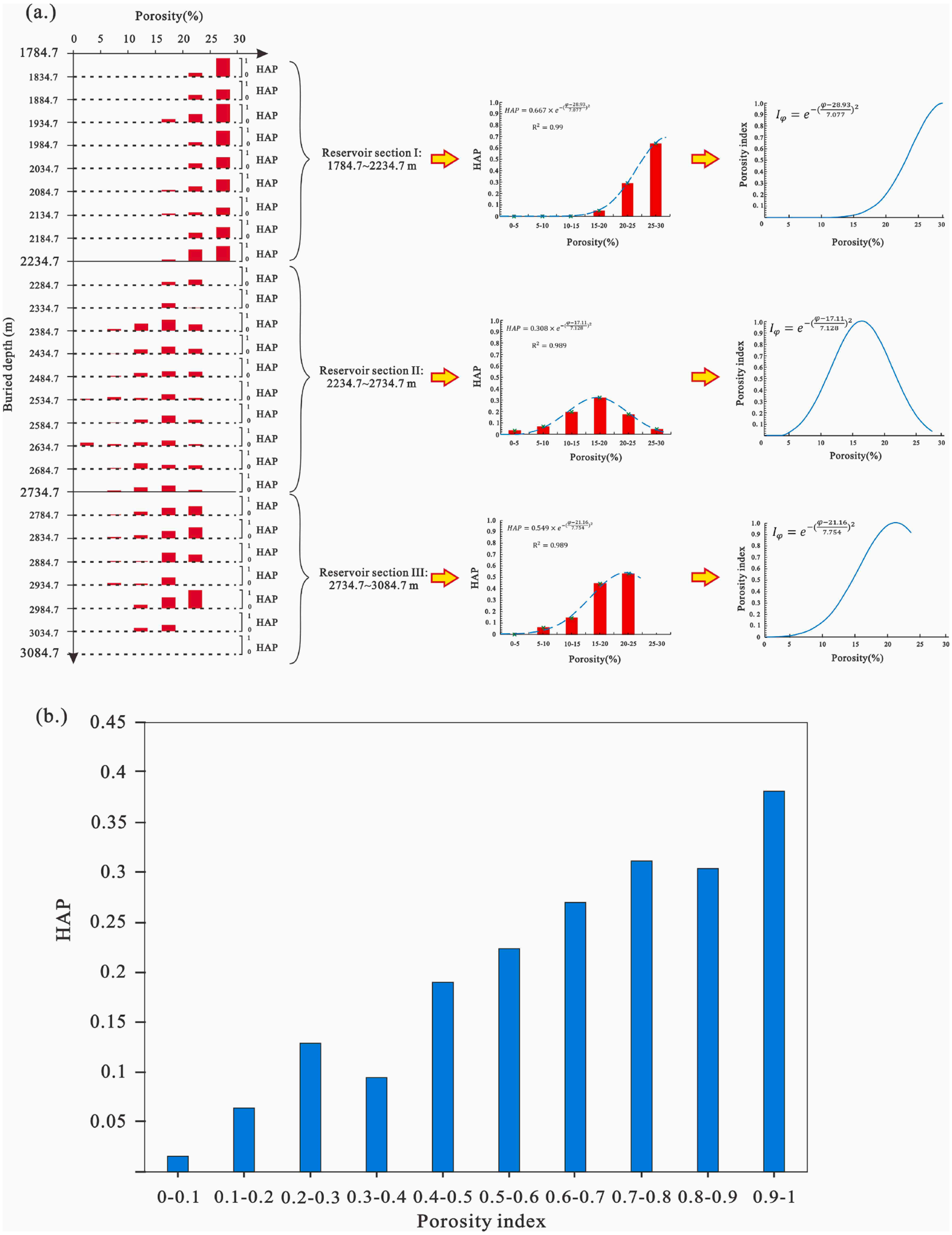


Fig. 11. (a) The numerical model between porosity and HAP in each reservoir unit and section in Jin 93 Well Block; (b) the relationship between porosity index and HAP in Jin 93 Well Block.

of β and evaluation system for F-measure in the petroleum exploration field, so it is necessary to select the appropriate value of β and build an evaluation system for F-measure according to the actual situation of the petroleum exploration field.

$$F_\beta = \left(1+\beta^2\right)\frac{Precision*Recall}{Recall+\beta^2 Precision} \tag{15}$$

where F_β denotes the F-measure value, dimensionless, and β denotes the weight between precision and recall, in which precision and recall have the same weight when $\beta = 1$, the weight of precision is greater than that of recall when $\beta < 1$, and the weight of precision is less than that of recall when $\beta > 1$, dimensionless.

From 2009 to 2018, the exploration success rate of international oil companies was approximately 20 %–50 %, with a mean of approximately 30 % [19]. Before drilling an exploration well, a favorable zone for oil and gas accumulation will be determined in the target reservoir, and the well set will be finally decided in the favorable zone. We assume that the favorable zone includes all oil and gas accumulations in the target reservoir and that the drilling exploration well in the favorable zone is a random process. Under these assumptions, the recall of current exploration technology is a constant 1, and the precision of it ranges from 20 % to 50 %. Let F_β equal to 0.8 as precision equal to 50 % and recall equal to 1, which represents the higher level for current exploration technology. According to this setting for F_β, β is calculated to be

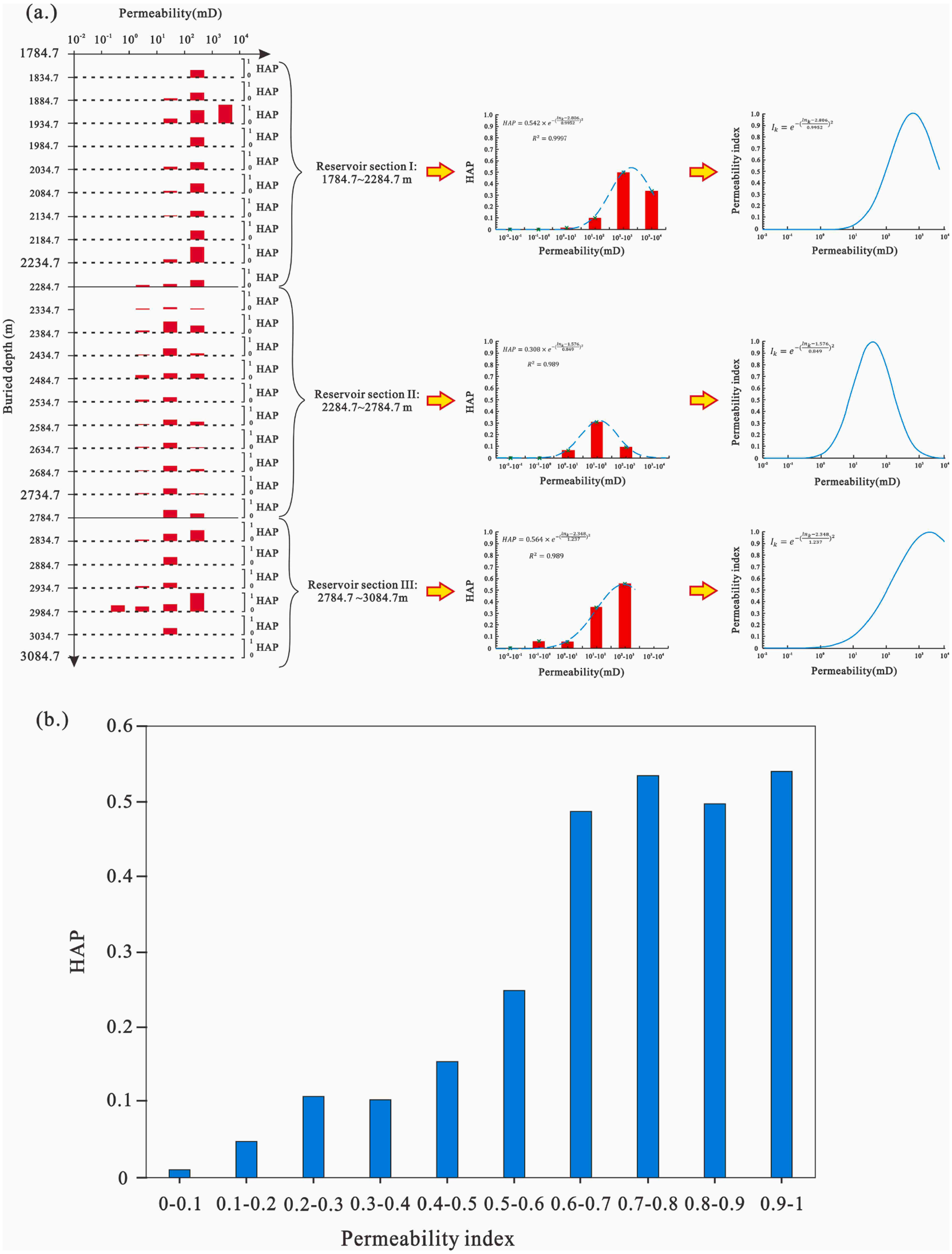


Fig. 12. (a) The numerical model between permeability and HAP in each reservoir unit and section in Jin 93 Well Block; (b) the relationship between permeability index and HAP in Jin 93 Well Block.

$\sqrt{3}$ using Eq. (15). Based on this weight (β), when recall equal to 1, 20 % precision and 30 % precision represent the lower and medium level for current exploration technology, corresponding F_β value of them is 0.5 and 0.63, respectively. Above all, the evaluation system of the petroleum exploration field for F-measure is established (Table 1). The prediction model with $F_\beta \geq 0.8$, $0.8 > F_\beta \geq 0.63$, $0.63 > F_\beta \geq 0.5$, and $0.5 > F_\beta$ will be evaluated as the excellent, good, fair, and poor model. However, the assumptions we make deviate from reality. Actually, the favorable zone selected by a human cannot cover all oil and gas accumulations, and the location of the exploration well set is carefully chosen in the favorable zone. All assumptions we make overestimate the level of current exploration technology; thus, an evaluation system for F-measure built under this assumption will be conservative and reliable, although it deviates from reality.

5.2.3. *Performance of HAP prediction model*

The k-fold cross-validation method, which is a well-known test method [63], is used to obtain the test result of the constructed HAP prediction model. In this method, the dataset is randomly divided into k portions first. One of the portions is then selected as test data (testing set), and the other k-1 portions are used as training data (training set). Further, the testing set is used to verify the prediction model trained by the training set. Finally, repeat the previous process k times to ensure

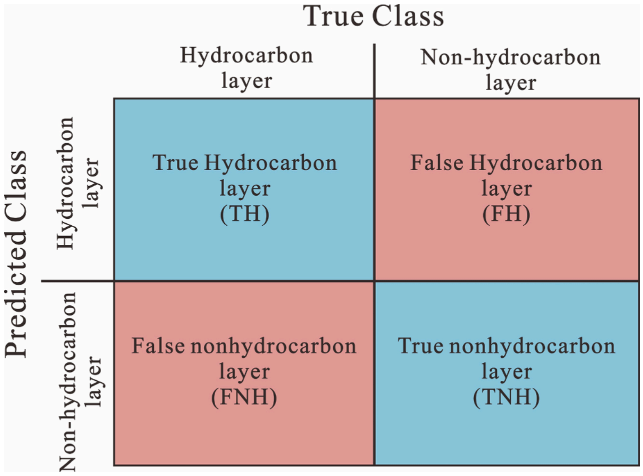

True Class
Hydrocarbon layer
Non-hydrocarbon layer
Predicted Class
Hydrocarbon layer
Non-hydrocarbon layer
True Hydrocarbon layer (TH)
False Hydrocarbon layer (FH)
False nonhydrocarbon layer (FNH)
True nonhydrocarbon layer (TNH)

Fig. 13. Schematic of the confusion matrix for this study.

Table 1

ML performance evaluation table for F-measure ($\beta = \sqrt{3}$).

F-measure value	$0 < F_\beta \leq 0.5$	$0.5 < F_\beta \leq 0.63$	$0.63 < F_\beta \leq 0.8$	$0.8 < F_\beta \leq 1$
ML performance evaluation	poor	fair	good	excellent

each portion of the dataset as the testing set verifies the prediction model trained by the other k-1 dataset portions, and the k times validation results are analyzed together [64]. Because the prediction model will be tested by all data in the dataset, the k-fold cross-validation method is an excellent test method when the dataset is small. However, the test result is affected by the division results of the dataset. To suppress this effect as much as possible, the test results of many k-fold cross-validations are considered [65].

In this study, the test results are obtained by ten times tenfold cross-validation method. Table 2 calculates and presents the evaluation indicators for the prediction model performance of each tenfold cross-validation. For the constructed HAP prediction model, its recall ranged from 0.8 to 0.82 with a mean of 0.81, its precision ranged from 0.44 to 0.45 with a mean of 0.45, and its F_β ranged from 0.66 to 0.68 with a mean of 0.67. According to the built performance evaluation system, the prediction model is good.

5.3. Favorable zone prediction result

Es_2 is selected for predicting a favorable zone. Es_2 is evenly divided into 61,919 grids of 0.01 × 0.01 km^2. In addition, according to the grid pattern, contour structure maps, porosity distribution prediction maps, and permeability distribution prediction maps are grided, and the distance from the center and boundary of the Es_3 source rock, which is mainly a hydrocarbon source for the Es_2 reservoir of each grid, is obtained. Because the area of the grid is very small, the geological parameters in each grid can be regarded as uniform distribution, and the geological parameters of each grid center are considered to represent the geological characteristics of the entire grid. Based on the distance from the source rock center and boundary of each grid, the I_S of each grid is calculated using Eq. (1). Based on the petroleum potential of each grid, the I_{rg} of each grid is calculated using Eqs. (2) and (3). Based on the porosity and burial depth, the I_φ of each grid is calculated using Eq. (11). Based on the permeability and burial depth, the I_k of each grid is calculated using Eq. (12). The data of each grid of Es_2 in Jin 93 Well Block, including grid center coordinates, I_S, I_{rg}, I_φ, and I_k, are assembled as a dataset (Table S.3). The favorable zone for oil and gas accumulations of Es_2 is obtained (Fig. 14a) by inputting the I_S, I_{rg}, I_φ, and I_k of each grid into the HAP prediction model and coloring the grid predicted as a hydrocarbon layer green.

The favorable zone predicted by the model is highly consistent with the discovered oil and gas accumulation range (Fig. 14a). The total area of discovered oil and gas accumulations in Es_2 in Jin 93 Well Block is 0.82 km^2. In the area of discovered oil and gas accumulations, a total of 5,942 grids are predicted as hydrocarbon layer, with a predicted area of 0.59 km^2. Using the favorable zone prediction method, 72.45 % area of discovered oil and gas accumulations are predicted as favorable zone and almost hydrocarbon wells within the prediction favorable zone (Fig. 14a). Overall, our model for predicting a favorable zone for oil and gas accumulation is effective, but it has some shortcomings. Some hydrocarbon wells and their adjacent areas are not predicted as favorable zones by the HAP prediction model, such as J93-52, J93-21X, and J93-25X wells (Fig. 14a). In addition, some water well distribution areas were predicted as favorable zones by the model, such as J93-19, J93-16X, J93-13, J93-22X, and J93-50X wells (Fig. 14a). Most of these wells are distributed in the favorable zone boundary, suggesting that the HAP model is not accurate and precise enough to predict the boundary conditions between oil and gas accumulations and non-oil and gas accumulations. In addition, the precision of the prediction map of geological parameters is a critical factor affecting the accuracy and precision of favorable zones. The final favorable zones for oil and gas accumulations of Es_2 are obtained by removing the favorable zone boundaries that do not agree with the drill result (Fig. 14b). The results show that the south of J93-46x, north of J93-41x, east of J93-7, and west of J93-48x have good exploration prospects (Fig. 14b).

Table 2

Ten times tenfold cross validation result for prediction model.

Validation times	Indicator	Value	Validation times	Indicator	Value
1	Recall	0.82	2	Recall	0.81
	Precision	0.45		Precision	0.44
	F_β	0.68		F_β	0.67
3	Recall	0.80	4	Recall	0.81
	Precision	0.44		Precision	0.45
	F_β	0.66		F_β	0.68
5	Recall	0.80	6	Recall	0.81
	Precision	0.44		Precision	0.44
	F_β	0.66		F_β	0.67
7	Recall	0.81	8	Recall	0.82
	Precision	0.45		Precision	0.44
	F_β	0.68		F_β	0.67
9	Recall	0.82	10	Recall	0.81
	Precision	0.45		Precision	0.45
	F_β	0.68		F_β	0.68

6. Discussion

6.1. ML model selection

In terms of the dataset in the petroleum exploration field, is the training result of the RUSBoosted tree ML algorithm better than other well-known ML algorithms? Five well-known ML algorithms were selected to train our preprocessed dataset (Table S.2), i.e., K-nearest neighbor (KNN) algorithm (for the specific algorithm, refer to Abeywickrama et al. [66]), SVM with Gaussian kernel algorithm (for the specific algorithm, refer to Wang et al. [55]), Naive Bayes algorithm (for the specific algorithm, refer to Langley P et al., [67]), logistic regression algorithm (for the specific algorithm, refer to Peduzzi et al., [68]), and DT algorithm (for the specific algorithm, refer to Ross Quinlan, [54]). Ten times tenfold cross-validation and F-measure with $\beta = \sqrt{3}$ are also used to evaluate the performance of the HAP prediction model constructed by the six ML algorithms. For conciseness, the HAP prediction model constructed by ML algorithms is abbreviated as the ML model. The test results of the six HAP prediction models significantly differ.

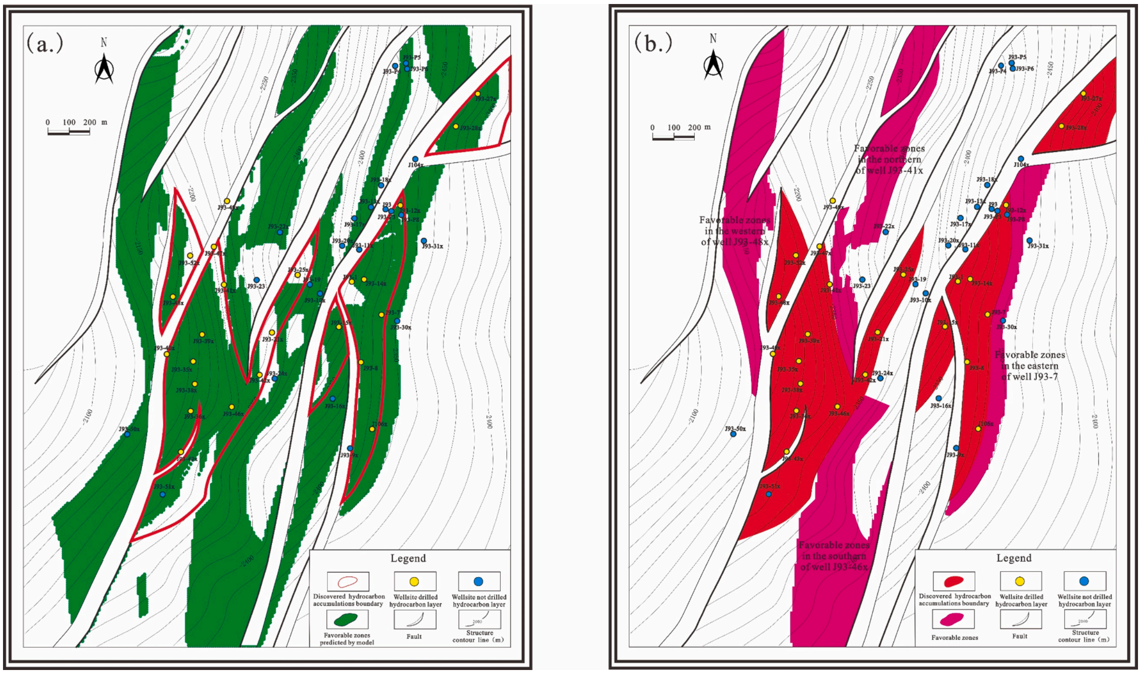

(a.)
N
0 100 200 m
Legend
Discovered hydrocarbon accumulations boundary
Wellsite drilled hydrocarbon layer
Wellsite not drilled hydrocarbon layer
Favorable zones predicted by model
Fault
Structure contour line (m)
(b.)
N
0 100 200 m
Favorable zones in the northern of well J93-41x
Favorable zones in the western of well J93-48x
Favorable zones in the eastern of well J93-7
Favorable zones in the southern of well J93-46x
Legend
Discovered hydrocarbon accumulations boundary
Wellsite drilled hydrocarbon layer
Wellsite not drilled hydrocarbon layer
Favorable zones
Fault
Structure contour line (m)

Fig. 14. The prediction result of favorable zones for oil and gas accumulations of Es_2 in Jin 93 Well Block.

According to the median or average of F_β, the following sequence is realized: RUSBoosted tree model (mean = 0.673, median = 0.675) > KNN model (mean = 0.610, median = 0.613) > Naive Bayes model (mean = 0.557, median = 0.556) > DT model (mean = 0.506, median = 0.509) > SVM with Gaussian kernel model (mean = 0.483, median = 0.479) > logistic regression model (mean = 0.296, median = 0.298) (Fig. 15 and Table S.4). The RUSBoosted tree model outperforms other models. In addition, the performance of the KNN model is the most similar to the RUSBoosted tree model. The mean precision and recall of the KNN model are 0.61 and 0.60, respectively, and the RUSBoosted tree model has a higher mean recall (0.81) and lower mean precision (0.45) (Table S.4). Thus, will the KNN model be better than the RUSBoosted tree model in predicting favorable zones? The KNN model is used to predict favorable Es_2 zones. The results show that the favorable zones predicted by the KNN model are very limited. A total of 9118 grids are predicted as hydrocarbon layers by the KNN model, which only accounts for 42.19 % of that predicted by the RUSBoosted tree model (Fig. 16). Moreover, the favorable zones predicted by the KNN model have poor continuity and are mainly distributed in local areas near the hydrocarbon well, and they have a low overlap rate with discovered oil and gas accumulation areas. This is most likely caused by the low recall, indicating that the model can correctly predict only a small part of hydrocarbon layer samples, which limits the favorable zones predicted. A conservative style is shown for this model. Compared with the KNN model, the RUSBoosted tree model, which has a higher recall, has a more radical style and is more suitable for favorable zone prediction. This also suggests that recall should be set at a higher weight than precision in the petroleum exploration field.

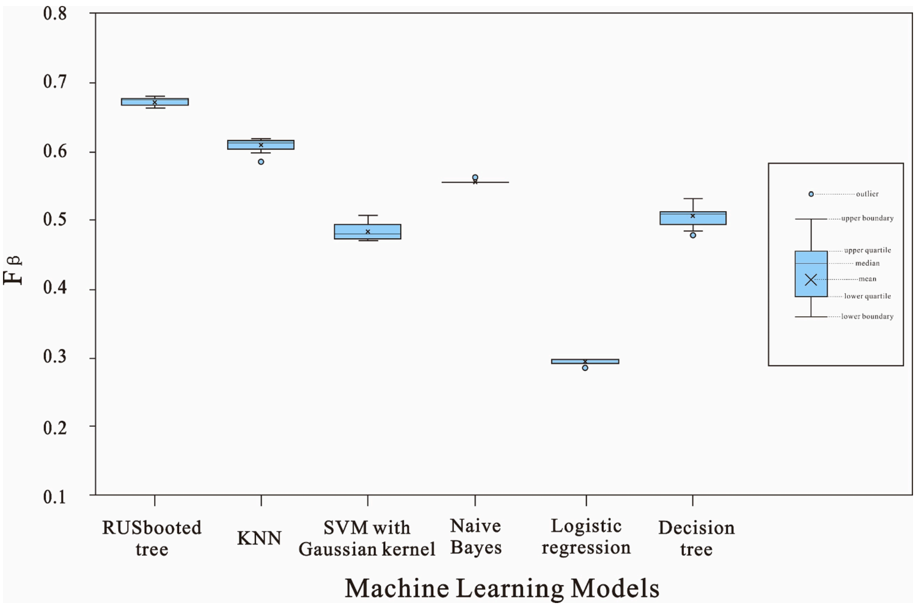


Fig. 15. Box figure for ten times tenfold cross-validation results of RUSBoosted tree model and other five well-known ML models on the dataset of this study.

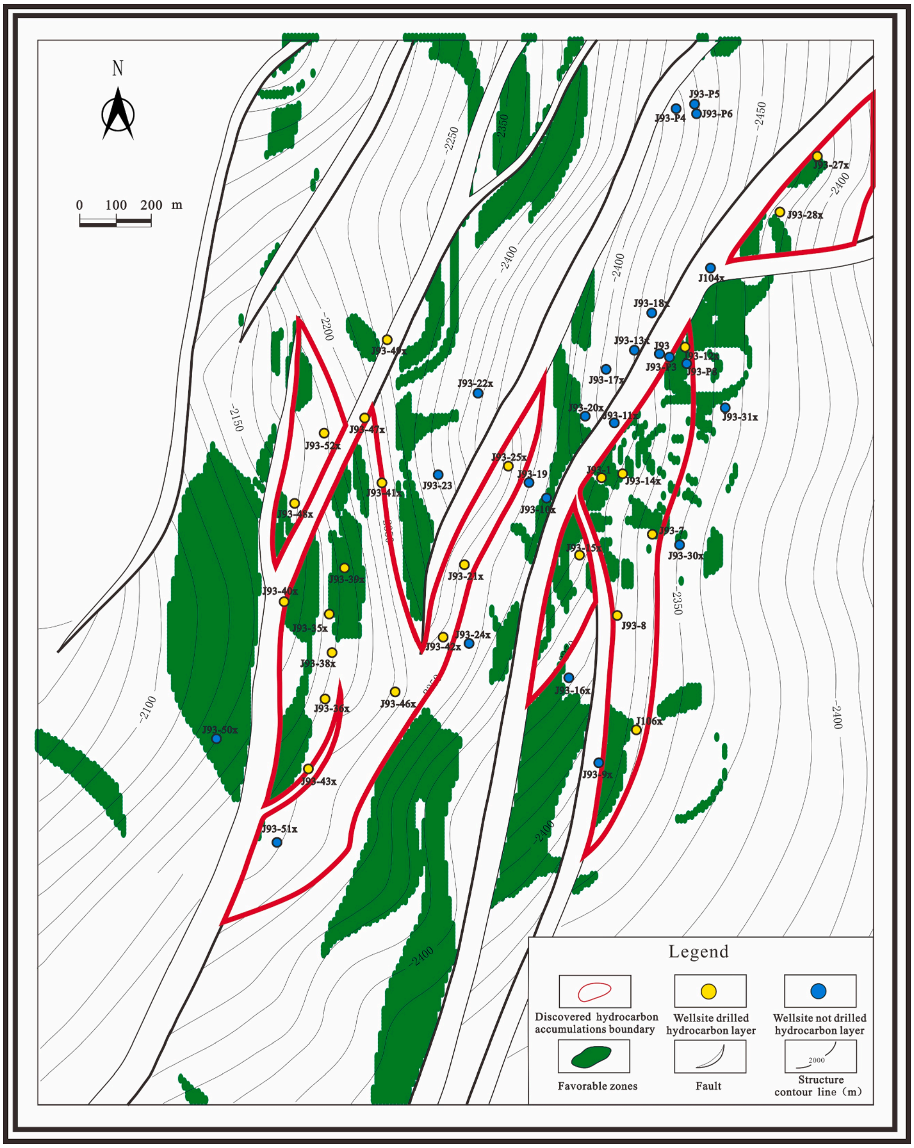


Fig. 16. Prediction results of favorable zones for oil and gas accumulations of Es_2 in Jin 93 Well Block based on the KNN model.

6.2. *Data preprocessing effect*

In terms of a small dataset, the performance of the HAP prediction model can be promoted by integrating professional knowledge into original data or embedding professional knowledge into the ML model as a regularization item or a constraint module [16,25–28]. In this study, professional knowledge of the petroleum exploration field is integrated into original data via a data preprocessing method based on source controlling theory [29–32], predominant reservoir theory [40–42], and fluid potential theory [7,35–39,69]. Nevertheless, owing to the heterogeneity of the geological body, the theories of the petroleum exploration field cannot accurately reflect the real underground situation in most cases. Can the constructed data preprocessing method effectively improve the performance of the ML model? To discuss this question, 20 types of sub-datasets are randomly selected from each of original geological dataset (Table S.1) and the preprocessed dataset (Table S.2), which contain 5 %-100 % data volume of the two datasets at 5 % intervals, respectively. In order to reduce the uncertainty caused by random selection, five sub-datasets were randomly selected for each type of sub-dataset. To maintain the same ratio of hydrocarbon layer to nonhydrocarbon layer as the original dataset in each sub-dataset, the same proportion of data is randomly selected from the hydrocarbon and nonhydrocarbon layer samples to form each of the sub-dataset. Further, the performances of the HAP prediction model trained by the sub-datasets selected from original geological dataset and the preprocessed dataset datasets under the same amount of data are compared. The mean of F_β for ten times tenfold cross-validation is used to evaluate their performance. In general, the performance of the RUSBoosted tree model trained on the preprocessed dataset is better than that trained on the original geological dataset, and the difference between them has an obvious decreasing trend as the amount of data increases (Fig. 17). This indicates that, in the case of sufficient data, the performance of the ML model trained on the original geological dataset can achieve similar results to that trained on the preprocessed dataset. This explanation is consistent with the complementary relationship between data volume and theoretical knowledge proposed by Karniadakis et al. [22].

6.3. *Applicability in complex petroliferous basin*

It is difficult to obtain good prediction results by directly applying FZ prediction method to complex petroliferous basin with multiple petroleum systems. The geological conditions of hydrocarbon accumulating in different petroleum systems are quite different, including the difference of source strata, the difference of hydrocarbon supply capacity, the

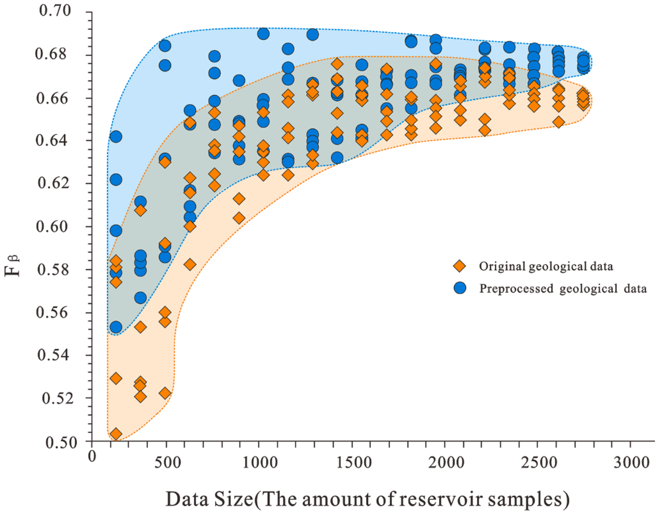


Fig. 17. Comparison map for the performance of RUSBoosted tree model trained on a preprocessed dataset and original geological dataset.

difference of migration system, the difference of reservoir distribution, the difference of trap type and the difference of source-reservoir contact relationship, etc [7]. The distribution characteristics of oil and gas accumulations in different petroleum systems are also different [7]. It is difficult to establish a reliable and accurate HAP prediction model based on HA elements through the unified analysis of CHAs formed in different petroleum systems [70]. Therefore, in complex petroliferous basins, the complex multiple petroleum systems should be divided into several single petroleum systems. Then, the HAP prediction model should be established in each single petroleum system. Thereafter, the FZ distribution for the petroleum system is predicted. Finally, the FZ predicted in each petroleum system are comprehensively analyzed to get the final FZ prediction result for the complex petroliferous basins. This strategy has been widely used in complex petroliferous basins and has achieved good results [3,8,70]. It can also help the method proposed by this study achieve good applicability in complex petroliferous basins with multiple source rocks.

6.4. Deficiencies and improvement

This study proposed a method of predicting favorable zones for conventional oil and gas accumulation based on ML, which provides an idea for the development of intelligent oil and gas exploration technology. However, at present, the results of this method are roughly equivalent to the average level of manual oil and gas exploration, which still cannot meet the requirements for industrial application. There are some deficiencies for application of this method:

(1) The number of features is small. The size of the original geological dataset and the preprocessed dataset used in this study is 2645 × 6 and 2645 × 4, respectively, making the performance of the ML model inadequate. Some studies show that the optimal performance of ML models tends to be stable when the amount of data reaches a certain threshold, which is related to the number of features of the dataset [24,71,72]. In this study, when the data sample exceeds 2,000, the performances of the model built based on the original geological and preprocessed datasets are generally similar and tend to be stable (Fig. 17), mainly attributable to the small number of features in the datasets. The only three most important HA elements with mature theory are considered in this study. However, besides the three HA elements, there are still many factors which have important influence on CHAs forming, such as cap conditions, preservation conditions, reservoir heterogeneity conditions and reservoir connectivity, etc. A study conducted by Lerche and Thomsen [73] showed that about 270 geological factors have an impact on the distribution and accumulation of CHAs. But how to use the theory of other elements to preprocess the raw data still needs further studies, which is also the next study direction.

(2) Strong constraints on the ML model. The method of data preprocessing essentially uses theory and knowledge to constrain the distribution of data features, making it difficult for ML models to exploit the relationship between original data and labels. In addition, directly preprocessing the data by a theory that is not entirely correct in the petroleum exploration field may introduce errors.

(3) Noise of data. Original geological data mainly come from geophysical instruments. Owing to the complicated lithologic composition and fluid distribution, geophysical data always include many noises.

(4) Hardly application in unexplored and low-explored areas. Presently, data-driven methods can hardly provide exploration directions for unexplored or low-explored areas. Theoretically, the HAP prediction model constructed on mature exploration blocks can be applied to unexplored or low-explored areas in a single petroleum system as they have similar geological conditions and conventional oil and gas formation processes. However, there are great uncertainties for geological data of reservoirs in unexplored or low-explored areas, which are primarily deduced from seismic, geomagnetic, and other geophysical data, especially in lacustrine basins with strong heterogeneity. The accuracy of input geological parameters is an important effect factor for the HAP prediction model. Therefore, the HAP prediction model constructed in mature exploration is still hardly applied to unexplored or low-explored areas in the same petroleum system. Moreover, in unexplored petroleum systems, the method proposed by this study is not applicable because of the lack of data.

To further improve the ML model performance, our next study targets are as follows: (a) Considering more HA elements and expanding the number of features in the dataset by collecting more independent features related to HAP, (b) exploiting the constraint effect of theory and data mine ability of ML by creating a theory and knowledge constrain module and embedding it into loss function of ML model, and (c) removing the obvious noisy data from the original dataset by establishing data cleaning standards for the petroleum exploration field.

7. Conclusion

(1) The proposed data preprocessing method based on source controlling theory, predominant reservoir theory, and fluid potential theory can effectively represent the quality of oil and gas storage space of reservoir, percolation capacity, accumulation conditions, and source conditions.

(2) Currently, the dataset of the petroleum exploration field is small and imbalanced. In addition, the training results of the RUSBoosted tree model are better than those of the other five ML models. Moreover, compared with the original geological data, the ML performance constructed by preprocessed data is improved.

(3) The prediction model built by the RUSBoosted tree algorithm and drill dataset of Jin 93 Well Block is a good model with recall and precision of 0.81 and 0.44, respectively. Favorable zones for oil and gas accumulation predicted by this prediction model agree well with discovered oil and gas accumulation areas.

(4) Considering more HA elements and expanding the number of features in the dataset, exploiting the constraint effect of theory and data mine ability of ML, and removing the obvious noisy data

from the original dataset may be three strategies to further enhance the performance of the prediction model.

CRediT authorship contribution statement

Kuiyou Ma: Methodology, Validation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Software. **Xiongqi Pang:** Methodology, Funding acquisition, Supervision. **Hong Pang:** Methodology, Funding acquisition, Project administration, Writing - Review & Editing. **Chuanbing Lv:** Funding acquisition, Resources. **Ting Gao:** Resources. **Junqing Chen:** Methodology, Formal analysis. **Xungang Huo:** Data Curation, Visualization. **Qi Cong:** Software, Visualization. **Mengya Jiang:** Software, Visualization.

Declaration of Competing Interest

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

Acknowledgements

This work was supported by Joint Funds of National Natural Science Foundation of China, China (Grant No. U19B6003-02) and Science Foundation of China University of Petroleum, Beijing, China (Grant number 2462020YXZZ022). We appreciate the PetroChina Huabei Oilfield Company for providing background geologic data and permission to publish the results.

References

[1] Pang X. Hydrocarbon distribution threshold and accumulation areas prediction. Science Press; 2015.

[2] Guo J, Xu J, Guo F, Li J, Pang X, Dong Y, et al. Functional-element constraint hydrocarbon distribution model and its application in the 3rd member of Dongying Formation, Nanpu Sag, Bohai Bay Basin, eastern China. J Pet Sci Eng 2016;139: 71–84. https://doi.org/10.1016/j.petrol.2015.12.017.

[3] Wang W, Pang X, Chen Z, Chen D, Yu R, Luo B, et al. Statistical evaluation and calibration of model predictions of the oil and gas field distributions in superimposed basins: a case study of the Cambrian Longwangmiao Formation in the Sichuan Basin, China. Marine Petroleum Geol 2019;106:42–61. https://doi.org/10.1016/j.marpetgeo.2019.04.032.

[4] Lewandowska-Śmierzchalska J, Tarkowski R, Uliasz-Misiak B. Screening and ranking framework for underground hydrogen storage site selection in Poland. Int J Hydrog Energy 2018;43:4401–14. https://doi.org/10.1016/j.ijhydene.2018.01.089.

[5] Rui Z, Lu J, Zhang Z, Guo R, Ling K, Zhang R, et al. A quantitative oil and gas reservoir evaluation system for development. J Nat Gas Sci Eng 2017;42:31–9. https://doi.org/10.1016/j.jngse.2017.02.026.

[6] Liu G, Hu S, Zhao W. Oil resource abundance of petroleum plays in Chinese basins and its prediction model. Pet Explor Dev 2007;33(759–761):775. https://doi.org/10.3321/j.issn:1000-0747.2006.06.022.

[7] Magoon LB. Petroleum system: Status of research and methods, 1992. Alexandria, VA (United States): Geological Survey; 1992.

[8] Pang H, Chen J, Pang X, Liu L, Liu K, Xiang C, et al. Key factors controlling hydrocarbon accumulations in Ordovician carbonate reservoirs in the Tazhong area, Tarim basin, western China. Mar Pet Geol 2013:88–1010. https://doi.org/10.1016/j.marpetgeo.2013.03.002.

[9] Liang F. The Research on Shale Gas Enrichment Pattern and the Favorable Area Optimizing of Wufeng-Longmaxi shale in middle and upper Yangtze Region. China university of mining and technology, 2018.

[10] Stephenson LP, Plumley WJ, Palciauskas VV. A model for sandstone compaction by grain interpenetration. SEPM J Sediment Res 1992;62(1):11–22. https://doi.org/10.1306/D4267875-2B26-11D7-8648000102C1865D.

[11] Houseknecht DW. Assessing the relative importance of compaction processes and cementation to reduction of porosity in sandstones. AAPG Bull 1987;71(6):633–42. https://doi.org/10.1306/9488787F-1704-11D7-8645000102C1865D.

[12] Abidoye LK, Das DB. Scale dependent dynamic capillary pressure effect for two-phase flow in porous media. Adv Water Resour 2014;74:212–30. https://doi.org/10.1016/j.advwatres.2014.09.009.

[13] Luo X, Hu C, Xiao Z, Zhao J, Zhang B, Yang W, et al. Effects of carrier bed heterogeneity on hydrocarbon migration. Mar Pet Geol 2015;68:120–31. https://doi.org/10.1016/j.marpetgeo.2015.08.015.

[14] Sircar A, Yadav K, Rayavarapu K, Bist N, Oza H. Application of machine learning and artificial intelligence in oil and gas industry. Pet Res 2021;6:379–91. https://doi.org/10.1016/j.ptlrs.2021.05.009.

[15] Bangert P, editor. Machine learning and data science in the oil and gas industry: best practices, tools, and case studies. Cambridge, MA Oxford: Gulf Professional Publishing; 2021.

[16] Von Rueden L, Mayer S, Beckh K, Georgiev B, Giesselbach S, Heese R, et al. Informed Machine Learning – A Taxonomy and Survey of Integrating Knowledge into Learning Systems 2019. https://doi.org/10.48550/ARXIV.1903.12394.

[17] 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.

[18] Sheremetov LB, González-Sánchez A, López-Yáñez I, Ponomarev AV. Time series forecasting: applications to the upstream oil and gas supply chain. IFAC Proc 2013; 46:957–62. https://doi.org/10.3182/20130619-3-RU-3018.00526.

[19] Sun C, Su X, Yang H, Li F. Fracture characteristics from outcrops and its meaning to gas accumulation in the Jiyuan Basin, Henan Province. China Open Geosci 2020; 12:1309–23. https://doi.org/10.1515/geo-2020-0199.

[20] He H, Garcia EA. Learning from Imbalanced Data. IEEE Trans Knowl Data Eng 2009;21:1263–84. https://doi.org/10.1109/TKDE.2008.239.

[21] Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. RUSBoost: a hybrid approach to alleviating class imbalance. IEEE Trans Syst Man Cybern - Part Syst Hum 2010;40:185–97. https://doi.org/10.1109/TSMCA.2009.2029559.

[22] Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. Nat Rev Phys 2021;3:422–40. https://doi.org/10.1038/s42254-021-00314-5.

[23] Mohammadpoor M, Torabi F. Big Data analytics in oil and gas industry: an emerging trend. Petroleum 2020;6:321–8. https://doi.org/10.1016/j.petlm.2018.11.001.

[24] Beleites C, Neugebauer U, Bocklitz T, Krafft C, Popp J. Sample size planning for classification models. Anal Chim Acta 2013;760:25–33. https://doi.org/10.1016/j.aca.2012.11.007.

[25] Diligenti M, Roychowdhury S, Gori M. Integrating Prior Knowledge into Deep Learning. In: 2017 16th IEEE Int. Conf. Mach. Learn. Appl. ICMLA, Cancun, Mexico: IEEE; 2017:920–923. https://doi.org/10.1109/ICMLA.2017.00-37.

[26] Xu J, Zhang Z, Friedman T, Liang Y, Broeck GV den. A Semantic loss function for deep learning with symbolic knowledge 2017. https://doi.org/10.48550/ARXIV.1711.11157.

[27] Daw A, Karpatne A, Watkins W, Read J, Kumar V. Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling 2017. https://doi.org/10.48550/ARXIV.1710.11431.

[28] Stewart R, Ermon S. Label-Free Supervision of Neural Networks with Physics and Domain Knowledge. ArXiv160905566 Cs 2016.

[29] Hu C. Oil source area controlling oil and gas field distribution: an effective theory for regional exploration in continental basins in eastern China. J Oil 1982:9–13. https://doi.org/10.7623/syxb198202002.

[30] Hu C, Liao X. The concept of petroleum system raised in china and its application. Editor Off ACTA Pet Sin 1996;17:10–6. https://doi.org/10.7623/syxb199601002.

[31] Jiang F, Pang X, Bai J, Zhou X, Li J, et al. Comprehensive assessment of source rocks in the Bohai Sea area, eastern China. AAPG Bull 2016;100:969–1002. https://doi.org/10.1306/02101613092.

[32] Pang X, Jin Z, Jiang Z. The evaluation of hydrocarbon resources in superimposed basins and its research significance. Pet Explor Dev 2002:9–13. https://doi.org/10.3321/j.issn:1000-0747.2002.01.003.

[33] England WA, Mackenzie AS, Mann DM, Quigley TM. The movement and entrapment of petroleum fluids in the subsurface. J Geol Soc 1987;144:327–47. https://doi.org/10.1144/gsjgs.144.2.0327.

[34] Hubbert MK. Entrapment of Petroleum Under Hydrodynamic Conditions. AAPG Bull 1953;37. https://doi.org/10.1306/5CEADD61-16BB-11D7-8645000102C1865D.

[35] Dahlberg EC. Applied Hydrodynamics in Petroleum Exploration. New York, NY: Springer US; 1982. 10.1007/978-1-4684-0144-8.

[36] Schowalter Tim T. Mechanics of Secondary Hydrocarbon Migration and Entrapment. AAPG Bull 1979;63. https://doi.org/10.1306/2F9182CA-16CE-11D7-8645000102C1865D.

[37] Hindle AD. Petroleum migration pathways and charge concentration: a three-dimensional model. AAPG Bull 1997;81:1451–81. https://doi.org/10.1306/3B05BB1E-172A-11D7-8645000102C1865.

[38] Berg RR. Capillary pressures in stratigraphic traps. AAPG Bull 1975;59. https://doi.org/10.1306/83D91EF7-16C7-11D7-8645000102C1865D.

[39] Hobson GD. Some Fundamentals of Petroleum Geology, Oxford University Press, London, 1-139. https://doi.org/10.1017/s001675680006653x.

[40] Pang X, Zhou X, Jiang Z. Hydrocarbon reservoirs formation, evolution, prediction and evaluation in the superimposed basins. J Geol 2012;86:1–103. https://doi.org/10.3969/j.issn.0001-5717.2012.01.001.

[41] 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;36:574-582+588. https://doi.org/10.11781/sysydz201405574.

[42] Wang W, Pang X, Chen Z, Chen D, Ma X, Zhu W, et al. Improved methods for determining effective sandstone reservoirs and evaluating hydrocarbon enrichment in petroliferous basins. Appl Energy 2020;261. https://doi.org/10.1016/j.apenergy.2019.114457.

[43] Pang X, Li P, Jin Z, Zhang S, Zuo S, Chen D. Research on hydrocarbon accumulation threshold and its application in Jiyang Depression. Oil Gas Geol 2003:204–9. https://doi.org/10.3321/j.issn:0253-9985.2003.03.003.

[44] Jiang F, Pang X, Guo J. Quantitative analysis model and application of the hydrocarbon distribution threshold. Acta Geol Sin - Engl Ed 2013;87:232–42. https://doi.org/10.1111/1755-6724.12044.

[45] Fu X, Jia R, Wang H, Wu T, Meng L, Sun Y. Quantitative evaluation of fault-caprock sealing capacity: a case from Dabei-Kelasu structural belt in Kuqa Depression, Tarim Basin. NW China Pet Explor Dev 2015;42:329–38. https://doi.org/10.1016/S1876-3804(15)30023-9.

[46] Walker RG. Facies Model – 3. Sandy Fluvial Systems. GeosclenceCanada 1976;3:10.

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

[48] Miall AD. Principles of Sedimentary Basin Analysis. Berlin, Heidelberg: Springer Berlin Heidelberg; 2000. 10.1007/978-3-662-03999-1.

[49] Ma Y. Sedimentology of carbonate reservoir. Geological Publishing House; 1999.

[50] Timmerman EH. Practical reservoir engineering. Pennwell Pubco; 1982.

[51] Yang Y, Zha K, Chen Y-C, Wang H, Katabi D. Delving into Deep Imbalanced Regression 2021. https://doi.org/10.48550/ARXIV.2102.09554.

[52] Longadge R, Dongre S. Class Imbalance Problem in Data Mining Review 2013. https://doi.org/10.48550/ARXIV.1305.1707.

[53] Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans Syst Man Cybern Part C Appl Rev 2012;42:463–84. https://doi.org/10.1109/TSMCC.2011.2161285.

[54] Quinlan JR. C4.5: Programs for Machine Learning. San Francisco, CA: Morgan Kaufmann; 1993. 10.1016/C2009-0-27846-9.

[55] Wang X, Chen J. Parameter Selection of SVM with Gaussian kernel. Comput Syst Appl 2014.

[56] Ibrahim M, Louie M, Modarres C, Paisley J. Global Explanations of Neural Networks: Mapping the Landscape of Predictions. arXiv preprint arXiv 2019; 1902.02384. https://doi.org/10.48550/arXiv.1902.02384.

[57] Ye H, Zhang B, Mao F. The Cenozoic tectonic evolution of the Great North China: two types of rifting and crustal necking in the Great North China and their tectonic implications. Tectonophysics 1987;133:217–27. https://doi.org/10.1016/0040-1951(87)90265-4.

[58] Li D. Study on hydrocarbon enrichment in gentle slope belt of faulted basin. China Univ Geosci (Beijing) 2010.

[59] Jiang Z, Chen D, Qiu L, Liang H, Ma J. Source-controlled carbonates in a small Eocene half-graben lake basin (Shulu Sag) in central Hebei Province. North China Sedimentol 2007;54:265–92. https://doi.org/10.1111/j.1365-3091.2006.00834.x.

[60] Ren C, He F, Gao X, Wu D, Yao W, Tian J, et al. Prediction of exploration targets based on integrated analyses of source rock and simulated hydrocarbon migration direction: a case study from the gentle slope of Shulu Sag, Bohai Bay Basin, northern China. Geosci J 2019;23:977–89. https://doi.org/10.1007/s12303-018-0078-0.

[61] Huo Z, Tang X, Meng Q, Zhang J, Li C, Yu X, et al. Geochemical characteristics and hydrocarbon expulsion of lacustrine marlstones in the Shulu Sag, Bohai Bay basin, eastern China: assessment of tight oil resources. Nat Resour Res 2020;29:2647–69. https://doi.org/10.1007/s11053-019-09580-8.

[62] Ye H, Shedlock KM, Hellinger SJ, Sclater JG. The North China Basin: an example of a Cenozoic rifted intraplate basin. Tectonics 1985;4:153–69. https://doi.org/10.1029/TC004i002p00153.

[63] Stone M. Cross-validatory choice and assessment of statistical predictions. J R Stat Soc Ser B Methodol 1974;36:111–33. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x.

[64] Wainer J, Cawley G. Nested cross-validation when selecting classifiers is overzealous for most practical applications. Expert Syst Appl 2021;182. https://doi.org/10.1016/j.eswa.2021.115222.

[65] Valente G, Castellanos AL, Hausfeld L, De Martino F, Formisano E. Cross-validation and permutations in MVPA: validity of permutation strategies and power of cross-validation schemes. NeuroImage 2021;238. https://doi.org/10.1016/j.neuroimage.2021.118145.

[66] Abeywickrama T, Cheema MA, Taniar D. k-Nearest neighbors on road networks: a journey in experimentation and in-memory implementation 2016. https://doi.org/10.48550/ARXIV.1601.01549.

[67] Langley P, Iba W, Thompson K. An Analysis of Bayesian Classifiers. Proc Tenth Natl Conf Artif Intell 1998;90.

[68] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373–9. https://doi.org/10.1016/S0895-4356(96)00236-3.

[69] Hubbert MK, Rubey WW. Role of fluid pressure in mechanics of overthrust faulting: I. Mechanics of fluid-filled porous solids and its application to overthrust faulting. Geol Soc Am Bull 1959;70(2):115–66. https://doi.org/10.1130/0016-7606(1959)70[115:ROFPIM]2.0.CO;2.

[70] Pang X, Zhou X, Yan S, Wang Z, Yang H, Jiang F, et al. Research advances and direction of hydrocarbon accumulation in the superimposed basins, China: take the Tarim Basin as an example. Pet Explor Dev 2012;39(6):692–9.

[71] Jain AK, Chandrasekaran B. 39 Dimensionality and sample size considerations in pattern recognition practice. Handb Stat 1982;2:835–55. https://doi.org/10.1016/S0169-7161(82)02042-2.

[72] Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. BMC Med Inform Decis Mak 2012;12(1):1–10. https://doi.org/10.1186/1472-6947-12-8.

[73] Lerche I, Thomsen RO. Hydrodynamics of Oil and Gas. Kluwer Academic Publishers, Hingham. Pet Explor Dev 2012 1994;39(6):692-699.