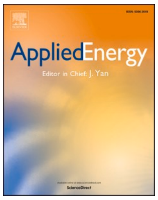
Contents lists available at ScienceDirect

Applied Energy

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



# A novel method for predicting shallow hydrocarbon accumulation based on source-fault-sand (S-F-$S_d$) evaluation and ensemble neural network (ENN)

Fuwei Wang [a,b], Dongxia Chen [a,b,*], Meijun Li [a,b], Zhangxin Chen [c], Qiaochu Wang [a,b], Mengya Jiang [a,b], Lanxi Rong [a,b], Yuqi Wang [a,b], Sha Li [a,b], Khawaja Hasnain Iltaf [d], Renzeng Wanma [a,b], Chen Liu [a,b]

[a] National Key Laboratory of Petroleum Resources and Engineering, China University of Petroleum (Beijing), Beijing 102249, China
[b] College of Geosciences, China University of Petroleum (Beijing), Beijing 102249, China
[c] Chemical and Petroleum Engineering, Schulich School of Engineering, University of Calgary, Calgary T2N 1N4, Canada
[d] Department of Earth and Environmental Sciences, The University of Texas at Arlington, TX 76019, USA

## HIGHLIGHTS

- A novel method for shallow hydrocarbon accumulation prediction based on S-F-$S_d$ evaluation and ENN.
- A new way to estimate predrilling hydrocarbon volume and geological reserves in offshore exploration.
- Systematic quantification procedure of key factors affecting shallow hydrocarbon accumulation.
- Discussion of key uncertainties in predicting hydrocarbon accumulation using machine learning.

## ARTICLE INFO

*Keywords:*
Shallow hydrocarbon accumulation
Quantitative prediction
Machine learning
Ensemble neural network
Source–fault–sand evaluation

## ABSTRACT

Shallow hydrocarbon accumulation (SHA) and predrilling volume prediction are important components of offshore oil and gas exploration. However, SHA prediction is complex and involves geological and technical uncertainties. Despite advances in available technology, reliable and convenient methods for predicting SHA are urgently needed by oil companies to avoid costly drilling mistakes. This study proposes a novel method for SHA prediction by combining source–fault–sand (S-F-$S_d$) evaluation and ensemble neural network (ENN) algorithms. First, twelve main controlling factors affecting SHA, which predominantly included geological parameters related to source rocks (S), fault zones (F) and sandstone reservoirs ($S_d$), were screened and quantified. Second, the six principal components obtained after the dimensionality reduction of the main control factors were selected as the model inputs. Then, using the BP neural network (BPNN), bagged neural network ensemble (Bagged-NNE) and boosted neural network ensemble (Boosted-NNE) algorithms, three different SHA prediction models with hydrocarbon column height (HCH) as the output were constructed. These models were applied to the K gasfield in the Xihu Depression, East China Sea Basin, to evaluate and optimize the model performance. Finally, the variable importance and the possible uncertainties in SHA prediction were discussed. The results show that the Boosted-NNE model is superior to the Bagged-NNE and BPNN models in SHA prediction. Moreover, the geological reserves of sandstone reservoirs calculated using the predicted HCH are close to the existing evaluation, which proves the effectiveness of the model output. In terms of variable importance, the synthetic parameters $F_1$, $F_2$, $F_5$ and $F_4$ obtained after dimensionality reduction are the four top principal components contributing to the model output. Under single-factor control, the HCH is positively correlated with the hydrocarbon expulsion rate, shale gouge ratio, sandstone thickness, porosity and permeability, but the relationship between the HCH and other controlling factors tends to be complicated. In addition, the model accuracy is affected by the uncertainties arising from the quantification and screening of the main controlling factors, as well as the dataset size and the machine learning algorithm selection. This contribution provides a reliable method for SHA prediction and corresponding predrilling volume evaluation, which can help avoid costly drilling mistakes and advance intelligent exploration techniques.

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

https://doi.org/10.1016/j.apenergy.2024.122684
Received 27 October 2023; Received in revised form 1 January 2024; Accepted 15 January 2024
Available online 25 January 2024

## 1. Introduction

In the 21st century, despite the active discovery of unconventional resources worldwide, conventional resources have remained an essential part of the energy supply because they are easier to exploit [1–5]. In 2013, conventional oil and gas reserves in China, the world's largest energy consumer, accounted for 25% of China's total fossil energy (Fig. 1) [6]. In 2021, the U.S. Geological Survey (USGS) estimated that China's undiscovered, technically recoverable conventional oil resources, i.e., 13.4 billion barrels and 244.4 trillion cubic feet of natural gas reserves, were spread across nine geological provinces [7]. With China's onshore exploration entering a highly mature stage, offshore shallow hydrocarbon accumulation (SHA) has gradually become the focus of China's conventional hydrocarbon exploration to ensure energy security [8]. During the "13th Five-Year Plan" period (2016–2020), the Chinese government vigorously promoted offshore SHA exploration and successfully discovered 26 large-medium oil–gas fields in the Bohai Bay Basin, East China Sea Basin, Pearl River Mouth Basin, etc. [9]. Typically, these types of oil–gas fields have mostly been discovered in shallow buried reservoirs that are vertically away from source rocks (Fig. 2a), where fault zones behave as vertical seepage conduits [8,10–12]. In the process of vertical migration along faults, hydrocarbons tend to diverge laterally into adjacent sandstone reservoirs and are eventually captured by different fault-bounded traps (Fig. 2b) [8,13,14]. However, heterogeneous source rocks (S), fault zones (F), sandstone reservoirs ($S_d$) and their diverse configurations make SHA and its corresponding volume challenging to predict, posing potential risks for future offshore drilling and exploration. Thus, reliable and convenient methods for predicting SHA remain the goal pursued by China's oil explorers and companies to avoid costly drilling mistakes.

In shallow oil–gas systems, various source–fault–sand (S-F-$S_d$) configurations constitute the basic geological framework for SHA. Specifically, source rocks control the material conditions of SHA depending on their geochemical features [14,15] and hydrocarbon expulsion characteristics [16,17]. According to petrophysical properties, faults and sandstones can behave as vertical and lateral conduits (or barriers), respectively, or can be combined into different fault-bounded traps capable of capturing and storing shallow hydrocarbons [18–20]. At present, there are many methods that can independently evaluate the hydrocarbon generation/expulsion of source rocks [16,17,21], the transport (or sealing) capacity of faults [12,22–27], and the transport (or storage) capacity of sandstones [20,28–31]. By constructing linear models, expert scoring methods [32], analytic hierarchy processes [33] and multiple regression methods [34] have been applied to realize hydrocarbon accumulation prediction by combining multiple geological factors. However, due to the strong geological heterogeneity and the multifactor joint contribution, the hydrocarbon accumulation process is actually complicated and shows an uncertain nonlinear relationship with geological elements. To this end, some scholars recently proposed nonlinear models based on major geological elements to predict the plane favorable zone of hydrocarbon accumulation [35,36]. Nevertheless, highly uncertain hydrocarbon migration progress within S-F-$S_d$ configurations, which is the key issue for SHA prediction, is not accounted for in these models. In addition, these studies only focus on the location of hydrocarbon accumulation without revealing the corresponding oil–gas volume, which is crucial for determining whether offshore shallow oil–gas reservoirs are economically recoverable. Thus, existing methods also induce a high risk of offshore shallow hydrocarbon exploration in practical applications.

With the rapid development of artificial intelligence (AI), evaluation (or prediction) models based on machine learning (ML) algorithms are widely used in transportation, healthcare, electricity and other fields, and have made positive significance [37–40]. The application of ML in the oil-gas industry is also increasing rapidly, involving geological feature identification (or classification), oil-gas content evaluation, and production prediction [41,42]. In previous studies, several ML techniques have been introduced to solve classification and regression problems in petroleum engineering and geology. Among them, decision tree (DT), random forest (RF), support vector machine (SVM) and artificial neural network (ANN) are commonly used due to their remarkable performance. The DT and RF techniques outperform other ML algorithms in solving classification problems, and have been applied in lithology identification [43,44], reservoir classification [45], and fault sealing evaluation [46]. The SVM technique presents unique advantages in dealing with nonlinear, high-dimensional, and potentially sparse datasets. This technique has been applied to forecast reservoir properties (e.g., porosity and permeability) [47–51], petroleum properties (e.g., interfacial tension) [52] and resource potential [53]. The ANN is the most popular ML model widely used for complex regression and optimization problems [42,54]. By simulating the basic principles of the biological nervous system and network topology, ANN can establish a mathematical model capable of nonlinear representation and logical operation [55–57], thereby solving complicated geological problems that cannot be addressed by traditional methods. The most widely used ANN model is the multi-layer neural network using the backpropagation algorithm, also known as BP neural network (BPNN). BPNN has the advantages of excellent fault tolerance, spontaneous learning and adaptability [58,59]. It has been proven to perform well in predicting total organic carbon (TOC) [60], reservoir properties [61], $CO_2$ geological storage [62], etc. Nevertheless, the single BPNN model also faces the challenges of slow convergence, low generalization, and unstable output. To this end, ensemble learning algorithms have been proposed to improve the prediction performance by integrating multiple basic (or weak) models [63–65]. Two popular methods for constructing ensembles are bagging (bootstrap aggregating) [66] and boosting [67]. Bagging only generates different bootstrap samples from the original

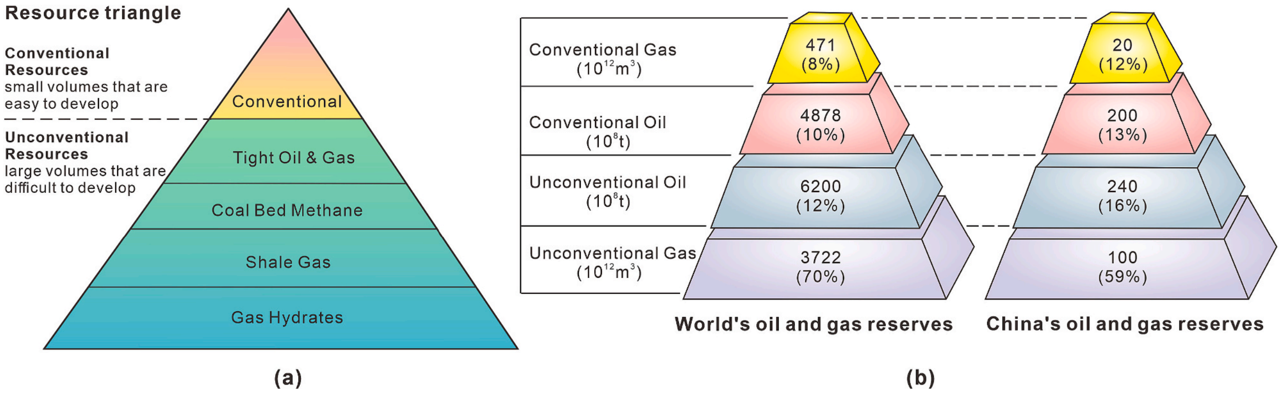

Resource triangle
Conventional Resources
small volumes that are easy to develop
Unconventional Resources
large volumes that are difficult to develop
Conventional
Tight Oil & Gas
Coal Bed Methane
Shale Gas
Gas Hydrates
(a)
Conventional Gas ($10^{12}m^3$)
Conventional Oil ($10^8t$)
Unconventional Oil ($10^8t$)
Unconventional Gas ($10^{12}m^3$)
471 (8%)
4878 (10%)
6200 (12%)
3722 (70%)
20 (12%)
200 (13%)
240 (16%)
100 (59%)
World's oil and gas reserves
China's oil and gas reserves
(b)

**Fig. 1.** (a) Resource triangle showing the relative volumes of conventional and unconventional hydrocarbon resources. (b) Global and Chinese oil and gas reserves in 2013 [6].

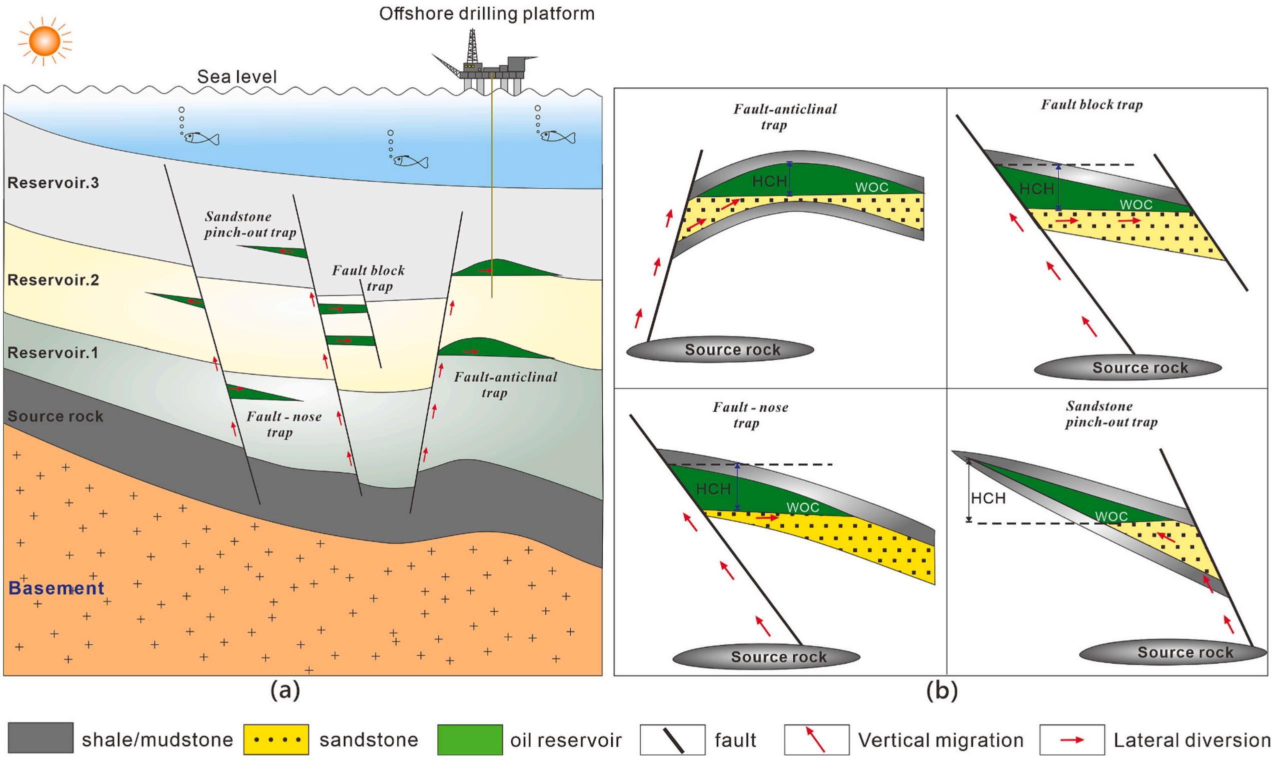


**Fig. 2.** (a) Conceptual model showing typical shallow hydrocarbon accumulation in offshore areas within China. (b) A schematic diagram of different fault-bounded hydrocarbon traps. HCH = hydrocarbon column height, WOC = water–oil contact.

training set for training different components. Boosting builds an ensemble by integrating multiple serially trained components, where each weak component can adaptively adjust the training set based on the performance of previously added components. Both bagging and boosting techniques can divide difficult prediction tasks into some relatively easy prediction subtasks and produce consistent prediction results for the original data [65]. Thus, ensemble learning tends to be an effective way to improve the prediction accuracy of complex regression problems in petroleum geology, especially for SHA prediction with a limited number of instances, high-dimensional feature sets, and highly complex trends.

The objective of this study is to introduce a reliable method for SHA prediction based on source-fault-sand (S-F-$S_d$) evaluation and ensemble neural network (ENN), which can help petroleum explorers or companies avoid costly drilling mistakes caused by uncertain shallow hydrocarbon accumulation within offshore reservoirs. In this investigation, an SHA prediction model based on a BPNN is first established on the basis of the systematic quantification of geological parameters related to S-F-$S_d$. Then, the bagged neural network ensemble (Bagged-NNE) and boosted neural network ensemble (Boosted-NNE) algorithms are used to improve the model performance. Finally, the optimal ensemble model is adopted to predict the SHA and hydrocarbon column height (HCH). Compared with previous reports, the innovations of this paper can address the following problems: (1) optimizing the quantitative characterization of geological factors controlling SHA, including source–fault–sand (S-F-$S_d$)-related geological and fluid characteristics, (2) constructing a fast and reliable model for SHA and its volume prediction to meet the decision-making requirements of offshore hydrocarbon exploration, and (3) providing an idea for transforming digital oil fields into intelligent oil fields by combining quantitative geological evaluation with an ensemble neural network. This contribution provides an efficient and rapid way to achieve quantitative prediction of SHA and the HCH, thereby reducing the risk of offshore hydrocarbon exploration and avoiding costly drilling mistakes.

## 2. Materials and methods

### 2.1. Data preprocessing

The data for this study were collected from the SINOPEC Shanghai Offshore Oil & Gas Company and China National Offshore Oil Corporation (CNOOC) Research Institute. These data consist of the 3D seismic volume after poststack geostatistical inversion, as well as logging, oil––gas testing, production and experimental analysis data of 22 drilling wells. These data were utilized to quantitatively characterize six sets of geological factors that control SHA associated with S–F–$S_d$, including hydrocarbon charging conditions, fault zone characteristics, sandstone reservoir characteristics, fault–sand intersection geometry, trap characteristics, and fluid properties. All of these input parameters were standardized before being input to the model. Moreover, the HCH was used as a quantitative indicator (output variable) to represent the disparity of the hydrocarbon accumulation volume in different traps. The quantification procedure of the geological parameters is as follows.

### 2.1.1. Hydrocarbon charging condition assessment

Source rock is an essential material basis for SHAs in petroliferous basins. Theoretically, only source rocks of sufficient geological and geochemical quality can support hydrocarbon expulsion. For a set of source rocks, the hydrocarbon expulsion capacity tends to decrease from the center of the basin or depression to the shallow buried marginal area. Thus, only the faults connected with source rocks with hydrocarbon expulsion ability can provide vertical hydrocarbon charging for SHA. In this study, four parameters, namely, the hydrocarbon expulsion rate (HER), charging duration (CD), fault–source contact thickness (FSCT) and vertical charging distance (VCD), were quantified to evaluate the hydrocarbon supply conditions of source rocks. The HER is the comprehensive reflection of source rock thickness, organic matter type, maturity, total organic carbon content (TOC), etc., and typically reflects the hydrocarbon quantities produced by source rock per unit area (square kilometers) in each million years. Since HER values vary with

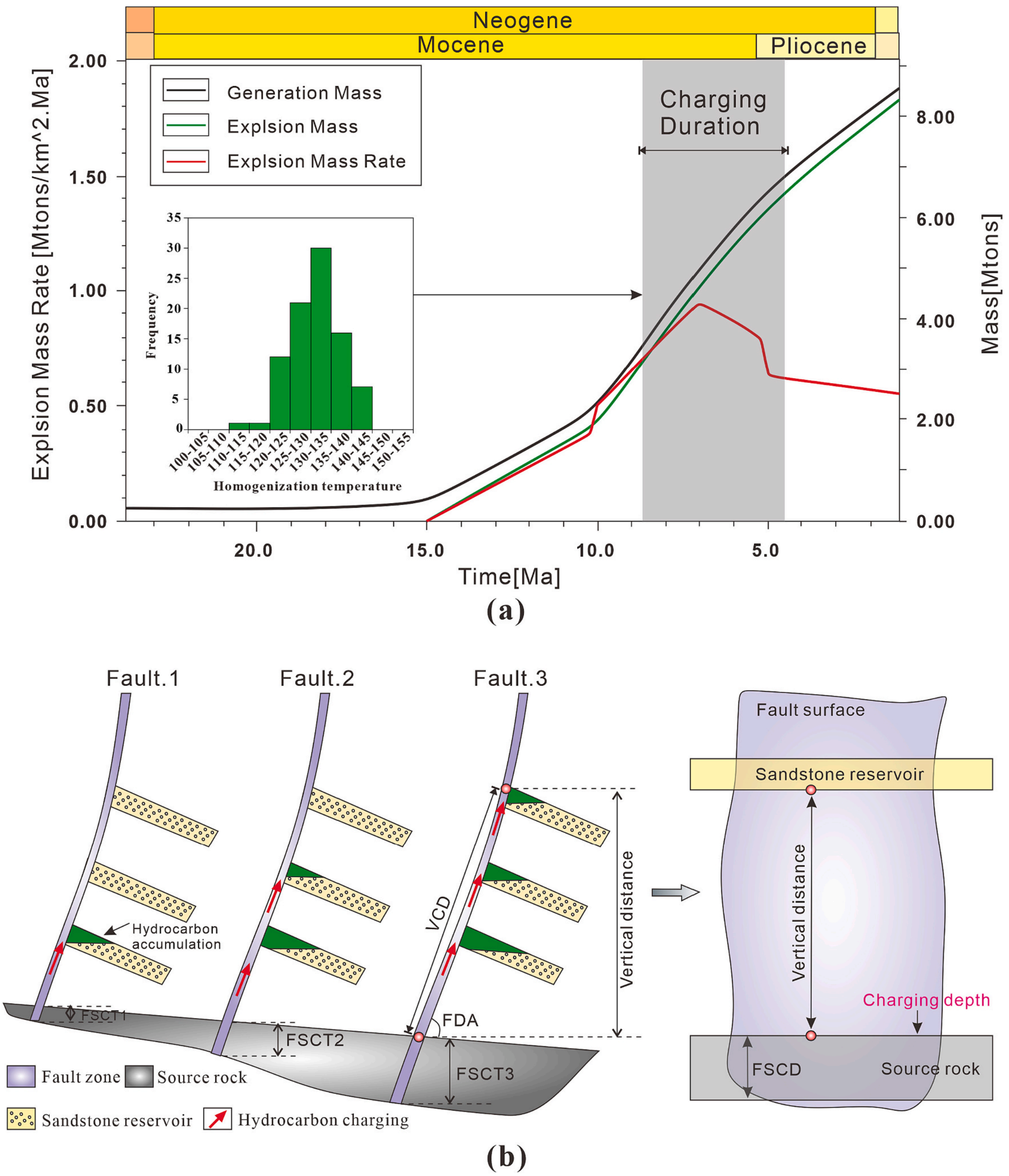


**Fig. 3.** Evaluation methods for parameters related to hydrocarbon charging conditions. (a) Combination of basin simulation with fluid inclusion homogenization temperature analysis to determine the charging duration (CD) and corresponding hydrocarbon expulsion rate (HER). (b) Use of the source–fault–sand configuration to determine the fault–source contact thickness (FSCT) and vertical charging distance (VCD).

the geological evolution history, Petro-mod software provided by Schlumberger was used to obtain the HER values during the major charging period (Fig. 3a). The CD refers to the duration of large-scale hydrocarbon charging, which can be determined by the fluid inclusion homogenization temperature and the corresponding basin thermal evolution history. The larger the CD value is, the longer the continuous hydrocarbon charging. The FSCT is defined as the vertical contact thickness between the fault zone and source rock, which physically represents the fault–sand contact area per unit width (Fig. 3b). Faults with larger FSCT values tend to receive more hydrocarbon supply when the hydrocarbon expulsion rate remains constant. In addition, the VCD represents the linear distance needed for hydrocarbons to migrate from the source rock to the target sandstone reservoir (Fig. 3b). The larger the VCD value is, the more material and dynamic support is needed for vertical hydrocarbon charging. The VCD can be calculated by the following equation:

$$VCD = VD \times \arcsin(FDA) \tag{1}$$

where VCD is the vertical charging distance, m; VD is the vertical distance between the source rock and the target sandstone reservoir, m; and FDA is the dip angle of the fault, °.

### 2.1.2. *Quantification of fault zone conduits*

As a complex three-dimensional geological body, faults can act as either conduits or barriers for SHA. Unlike normally deposited strata, fault rocks are formed by the mixing of host rocks that are fractured and

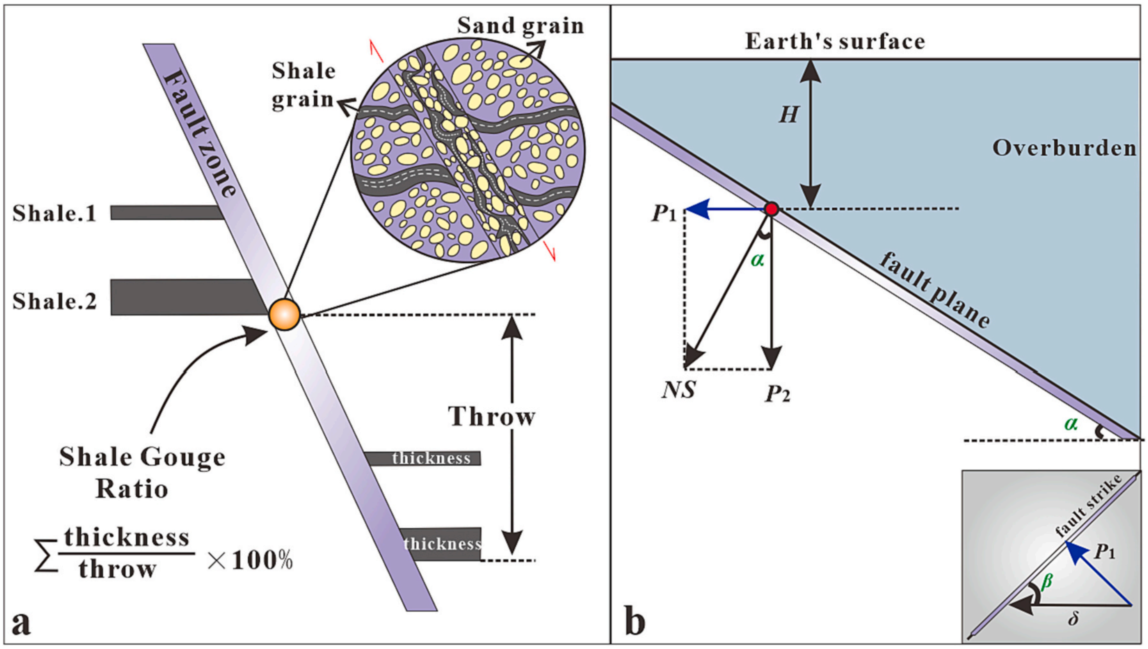


**Fig. 4.** Evaluation methods related to the fault zone conduit. (a) Conceptual map for the quantitative evaluation of the shale gouge ratio (SGR) [24]. (b) Conceptual diagram of stress normal to the fault plane (NS) [27]. where $P_1$ is the horizontal component of tectonic stress, $P_2$ is the vertical stress component from overburden, δ is the regional tectonic principal stress, β is the intersection angle of fault strike and δ, α is the fault dip angle, and H is the burial depth.

dragged into the fault zones [68–71]. It is difficult to directly characterize the physical properties of fault rock because the corresponding cores are rarely captured during drilling. According to existing studies, the main controlling factors of transport (or sealing) performance of faults can be summarized as spatial width [72–75], mudstone content [12,24,71] and stress conditions [12,25–27]. Thus, three parameters affecting the petrophysical properties of fault rocks, namely, fault throw (FT), shale gouge ratio (SGR) and normal stress to fault plane (NS), are selected to characterize the fault transport capacity. The FD value is usually positively correlated with the fault zone width [72,73] and negatively correlated with the fault rock granularity [75]. The increasing FD value often contributes to the widening of the fault zone width and the reduction in the fault rock pore space. Moreover, the mudstone content of fault rock is negatively correlated with its physical properties, which can be quantitatively characterized by the shale gouge ratio (SGR) (Fig. 4a) [24], as shown in Eq. (2).

$$\mathrm{SGR} = \frac{\sum \text{shale bed thickness}}{\text{fault throw}} \times 100\%$$

or

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

With respect to stress fields, the NS determines the tightness of the fault surface and the compression deformation of the fault rock. The NS is derived from both the overlying vertical stress and regional horizontal tectonic stress (Fig. 4b) [25–27], which can be quantitatively calculated by Eq. (3).

$$NS = H(\rho_b - \rho_w)*0.009876 cos\alpha + \delta sin\beta sin\alpha \qquad (3)$$

where NS is the normal stress to the fault plane, MPa; $H$ is the burial depth, m; $\rho_b$ and $\rho_w$ are the densities of overburden sediments and formation water, g/cm$^3$, respectively; $\alpha$ is the fault dip angle, °; $\delta$ is the principal tectonic stress, MPa; and $\beta$ is the intersection angle between the principal tectonic stress and fault strike,°.

### 2.1.3. *Quantification of sandstone reservoir*

Sandstone can act as a lateral conduit while providing storage space for hydrocarbon accumulation. For the SHA, the lateral hydrocarbon diversion at the fault–sand contact site determines whether the sandstone reservoir is charged. In sandstone control, the thickness, porosity and permeability properties affect the lateral fluid flow and redistribution. Generally, thick-bedded sandstone is more conducive to lateral diversion than thin-bedded sandstone due to its larger contact cross-sectional area with the fault zone (Fig. 5a). On the other hand, lateral hydrocarbon diversion is more likely to occur when the physical properties of sandstone are superior to those of fault rocks, and sandstone with larger porosity/permeability tends to be a preferential seepage conduit (Fig. 5b). Meanwhile, sandstone with greater thickness/porosity is capable of providing more storage space for SHA. Thus, sandstone thickness (ST), porosity (POR) and permeability (PERM) were selected as parameters to characterize the sandstone reservoir in this study.

### 2.1.4. *Quantification of fault–sand intersection geometry*

Buoyancy-driven hydrocarbon migration makes fault–sand intersection geometries critical for SHA. In this study, the fault dip angle (FDA), sandstone dip angle (SDA) and fault–sand intersection angle (FSIA) are selected to characterize the geometric features of the fault–sand configuration (Fig. 5c). The FDA, distributed in the range of 0° to 90°, can control the hydrocarbon flow velocity by changing the parallel fault buoyancy component. The SDA value is defined between −90° and 90°, where the up-dip and down-dip sandstones correspond to SDA values of 0–90° and − 90°-0°, respectively. Because the buoyancy remains vertically upward, the up-dip sandstones are more prone to lateral hydrocarbon divergence than the down-dip sandstones in the fault–sand intersection site. The FSIA is measured from the fault hanging wall along the dip direction and has a value of 0° ~ 360°, where a value of 0° ~ 180° represents sandstone on the hanging wall, whereas a value of 180° ~360° corresponds to sandstone on the footwall. Once the FDA, SDA and FSIA are determined, the spatial geometric characteristics of the fault–sand configuration can be accurately described.

### 2.1.5. *Quantification of trap characteristics*

Effective trapping conditions are the basis for capturing shallow oil

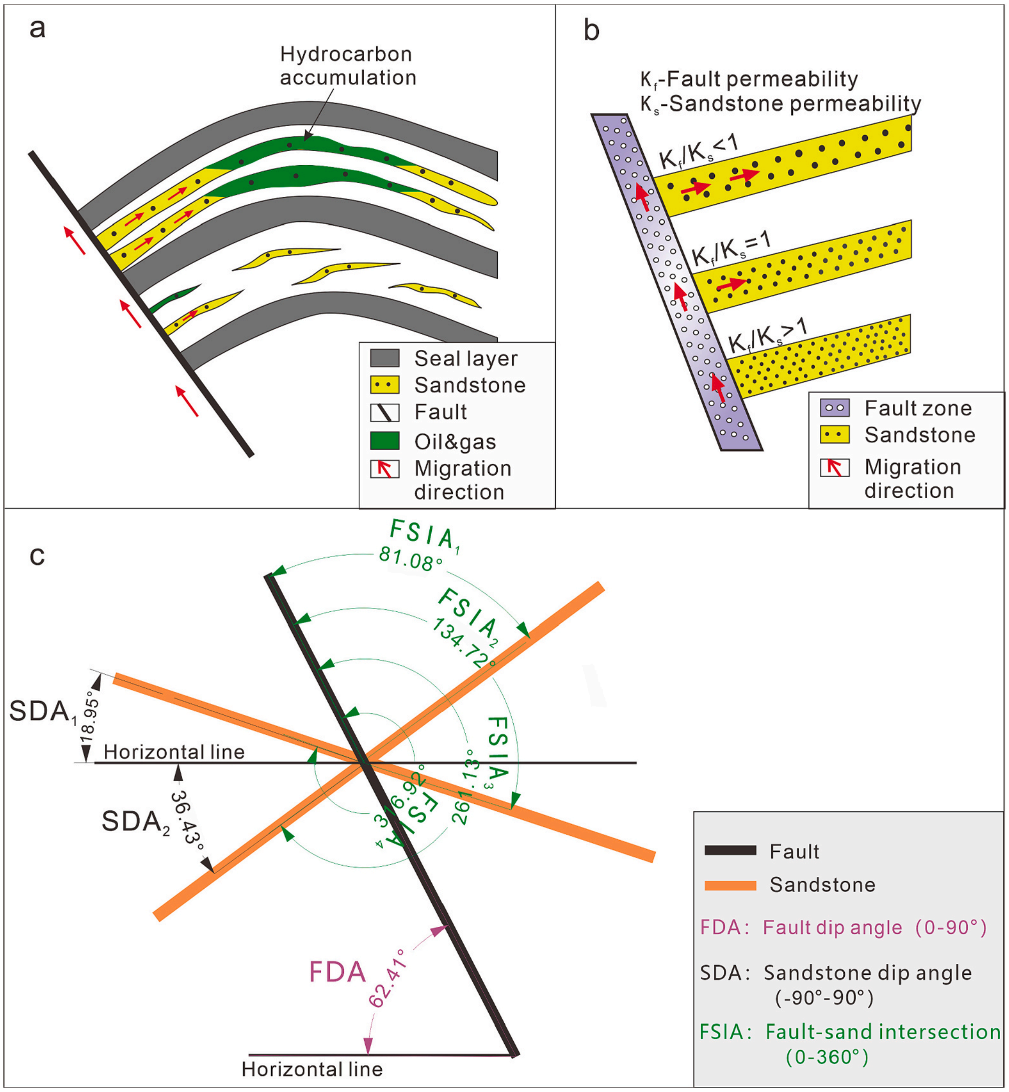


**Fig. 5.** Conceptual map showing the control of sandstone reservoirs and fault–sand intersection geometry in SHA. (a) Lateral hydrocarbon diversion controlled by sandstone thickness; (b) lateral hydrocarbon diversion controlled by sandstone physical properties; $K_s$ = sandstone permeability; $K_f$ = fault permeability, which was evaluated indirectly by using FT, SGR and NS; and (c) measurement of the fault dip angle (FDA), sandstone dip angle (SDA) and fault–sand intersection angle (FSIA).

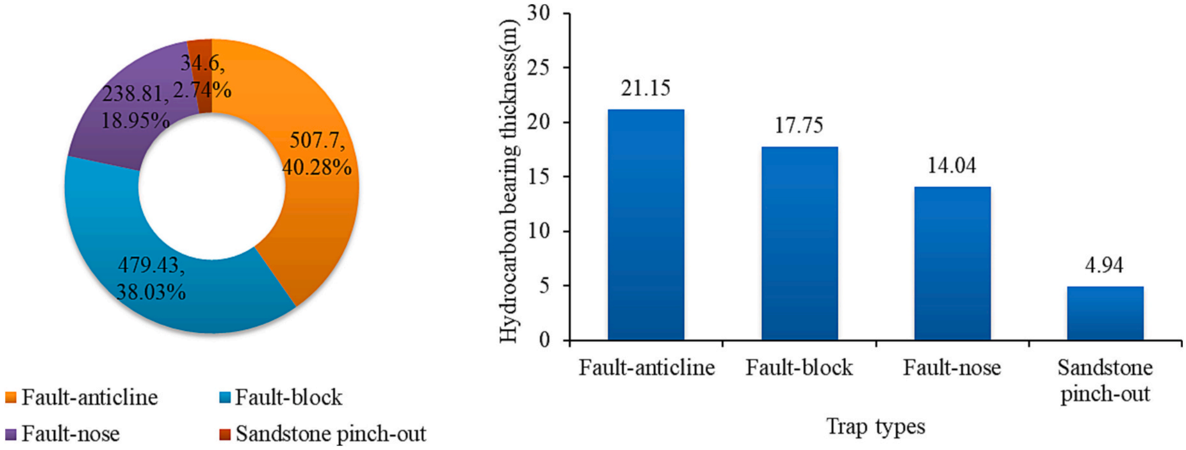


**Fig. 6.** Statistical results of hydrocarbons in different fault-bounded traps in the K gasfield of the Xihu Depression, East China Sea Basin. (a) Cumulative oil–gas show thickness percentage in different traps; and (b) average hydrocarbon-bearing thickness of different trap types.

**Table 1**
Hydrocarbons exhibit characteristics of different fault-bounded traps and their assigned values (*TT*) in the K gasfield of the Xihu Depression, East China Sea Basin.

| Content | Trap types of oil and gas reservoirs | | | | |
|---|---|---|---|---|---|
| | Fault-anticline | Fault-block | Fault-nose | Sandstone pinch-out | No trap |
| Number of discovered reservoirs | 25 | 28 | 18 | 8 | 0 |
| Cumulative oil-gas show thickness (m) | 507.70 | 479.43 | 238.81 | 34.6 | 0 |
| Percentage of hydrocarbon accumulation (%) | 40.28 | 38.03 | 18.95 | 2.74 | 0 |
| Average hydrocarbon-bearing thickness (m) | 21.15 | 17.75 | 14.04 | 4.94 | 0 |
| Assignment (TT) | 1.00 | 0.84 | 0.66 | 0.23 | 0 |

and gas. The target traps in this study are typically fault-bounded traps, which can be divided into four types: fault-anticline, fault-block, fault-nose and sandstone pinch-out traps. Fault-anticlinal traps can capture hydrocarbons that diverge laterally along sandstone by both anticlinal structure and fault sealing. Thus, even if a fault is not capable of lateral sealing, the anticlinal structure can still store hydrocarbons. In contrast, fault-block and fault-nose traps rely entirely on fault lateral sealing to capture hydrocarbons, which makes it difficult for weakly sealed traps to accumulate hydrocarbons. Sandstone pinch-out traps mainly rely on lithologic sealing to capture oil and gas. Only when the accumulated hydrocarbon column reaches the spill point will hydrocarbon accumulation be affected by the fault sealing ability. In this investigation, the constraints of traps on SHA were quantified by counting the size of the discovered oil–gas reservoirs in various traps (Fig. 6). Specifically, the fault-anticline trap with the highest average hydrocarbon-bearing thickness was assigned a trap type (TT) value of 1, representing the most favorable trap type for SHA in terms of statistical hydrocarbon-bearing thickness (Table 1). Other trap types were assigned TT values depending on the ratio of their average hydrocarbon-bearing thickness to the average hydrocarbon-bearing thickness of the fault-anticline trap. For cases where no effective traps were developed, we assigned a TT value of 0, indicating that there is no condition for capturing hydrocarbons.

### 2.1.6. *Quantification of fluid properties*

The fluid properties within the source–fault–sand geological framework, including pore fluid pressure (P), density (DEN), and viscosity (VIS), can control hydrocarbon migration efficiency to some extent. For basins that exhibit overpressure, excess pressure provides an extra driving force for hydrocarbon migration in addition to buoyancy. P can be effectively evaluated by combining measured pressure data with the equivalent depth method based on logging [12]. On the other hand, density/viscosity tend to affect the hydrocarbon flow velocity within the fault–sand carrier system as the temperature and pressure conditions change. The lower the density/viscosity of hydrocarbons is, the faster the flow velocity. In this study, the density/viscosity of hydrocarbons under different temperature and pressure conditions were obtained by pressure–volume–temperature (PVT) experimental data.

### 2.1.7. *Hydrocarbon column height assessment*

Estimating hydrocarbon volume is an important part of the exploration process because it determines whether a prospect contains enough hydrocarbons to justify drilling an exploration or appraisal well. However, obtaining hydrocarbon volumes is often challenging due to the complex trap morphology and diverse geological parameters needed. To simplify quantitative work, the hydrocarbon column height (HCH) is utilized in this study to approximately characterize the volume disparity of SHAs. The HCH is defined as the height of the continuous hydrocarbon column measured from the structural apex down to the hydrocarbon–water contact (Fig. 7). This parameter is the most critical parameter affecting volume calculation and is more convenient to obtain [76]. For fault-bounded hydrocarbon traps, the trap height and hydrocarbon column height were determined by measuring the depth of the structure apex, fluid contact and spill point. Once the HCH value was determined, the SHA and its corresponding volume in sandstone reservoirs adjacent to the fault can be quantitatively reflected. More specifically, the predicted HCH and the corresponding available trap geometry can be employed to estimate the horizontal gas-bearing area. On this basis, hydrocarbon volumes of different SHAs can be estimated using the volumetric method of gas reservoir geological reserves calculation (see details in the section "4.4. SHA prediction result").

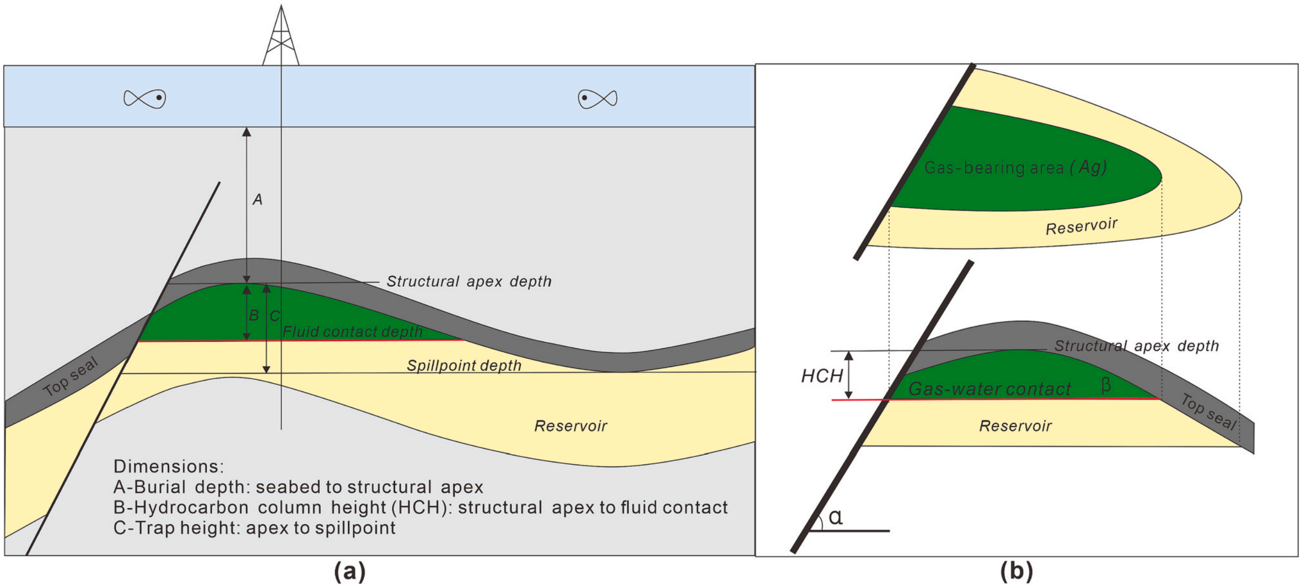


**Fig. 7.** A schematic diagram of a fault-bounded hydrocarbon trap: (a) hydrocarbon column height and (b) hydrocarbon-bearing area [76].

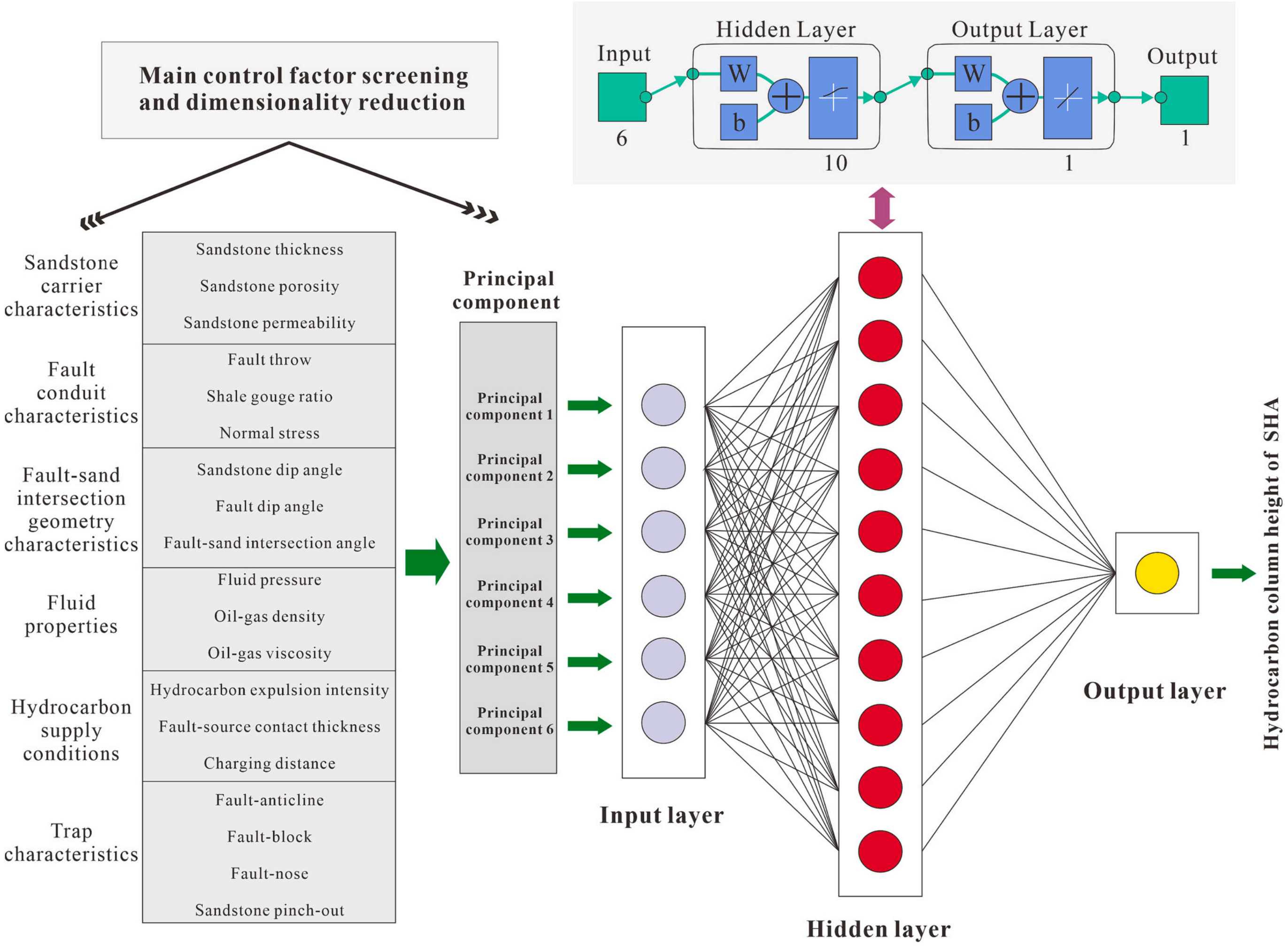


**Fig. 8.** Procedure of main control factor screening, dimensionality reduction processing and artificial neural network construction.

## 2.2. Main control factor screening and dimensionality reduction

Main control factor screening and dimensionality reduction processing are effective approaches to improve the rationality and accuracy of the SHA prediction model to be constructed. However, the screening of the main controlling factors of SHA in existing reports is highly subjective. On the basis of the established database, the gray correlation analysis, Pearson correlation coefficient, Spearman correlation coefficient and Kendall correlation coefficient were used in this study to rank the influencing factors of SHA [54,77–80]. These approaches were combined with the significance test to determine the main controlling factors [81]. In the gray correlation analysis, the similarity or dissimilarity between the development trend of each factor is determined by using the HCH as the reference sequence and all parameters related to S–F–S$_d$ as the comparison sequence. The Pearson correlation coefficient measures the linear correlation between two groups of data, whereas the Spearman and Kendall correlation coefficients evaluate the correlation degree between two variables according to the ranking position or order of the original data. According to the above evaluation methods, the ranking of each geological influencing factor in multiple correlation parameters can be obtained respectively. In practice, the ranking of a geological parameter tends to vary among different correlation parameters, while the corresponding weights are challenging to be artificially assigned due to complex nonlinear geological trends. In this study, the weights of multiple correlation parameters are treated as the same, and then the arithmetic average of the five correlation parameters mentioned previously is utilized as the basis for the comprehensive ranking. However, controlled by the number of spatial samples, a large correlation coefficient does not necessarily correspond to a high correlation degree. To this end, the significance test of the correlation coefficients based on IBM SPSS software was adopted to eliminate unimportant parameters and thereby determine the main controlling factors. During the significance test, IBM SPSS software initially assumes that the correlation between the two variables is zero and then calculate the corresponding probability value (abbreviated as *P* value or Sig value) for this hypothesis to be true. If the Sig value is small, it implies that the probability of the two variables being uncorrelated is low, and therefore the original hypothesis tends to be rejected and the alternative hypothesis (i.e., the two variables are correlated) is accepted. Thus, a smaller Sig value usually indicates a higher correlation. More specifically, 0.05 and 0.01 are universally recognized as two Sig thresholds (also known as significance levels) for judging the correlation degree between two variables. Sig $> 0.05$ indicates no significant correlation between the two variables, while Sig $<0.05$(*) reflects a significant correlation and sig $< 0.01$ (**) demonstrates an extremely significant correlation. In addition, to eliminate the possible multicollinearity between different factors, the principal component analysis (PCA) method was used to transform the main controlling factors from the high-dimensional space to the low-dimensional principal components [82]. In this way, the obtained principal components retain the meaningful properties of the original data and ideally approximate its intrinsic dimension.

### 2.3. BP neural network (BPNN)

#### 2.3.1. Artificial neural network structure

In this study, IBM SPSS Modeler (Version 18.0) was used to build the basic artificial neural network (ANN) model for SHA prediction. The ANN model consists of three parts: an input layer, a hidden layer, and an output layer. In this work, the six principal components obtained by screening and dimensionality reduction of S–F–$S_d$-related parameters were input into the input layer. Meanwhile, the HCH is chosen as the result of the output layer of the neural network model to reveal the volume variation of SHA. In subsequent work, the number of hidden layers and the corresponding number of nodes are determined and optimized. A typical three-layer neural network structure with a hidden laying containing 10 nodes is shown in Fig. 8.

#### 2.3.2. Basic principles of the BPNN algorithm

As a nonlinear transformation system, the ANN can perform nonlinear operations on complex problems and establish input–output prediction models [83–85]. To date, various ANN models with different learning rules have been proposed, among which the BPNN has been widely used due to its excellent fault tolerance, spontaneous learning and adaptability [58,59]. The BPNN is essentially a multilayer feedforward ANN based on error backpropagation algorithm training [86]. The working process of the BPNN follows the steps of forward propagation, error backpropagation, loop iteration and learning convergence.

During forward propagation, the input data are processed and transmitted through the hidden layer to the output layer, where the state of each layer of neurons is only affected by the previous layer of neurons [54]. The difference between the output value and the actual value (expected value) is defined as an error signal, and when it fails to meet the accuracy requirements of the model, error backpropagation will occur. At this stage, the error signal is propagated from the output layer to the input layer, during which the weights and thresholds are updated to bring the output closer to the actual value [87]. After many loop iterations, the neural network gradually converges and finally yields the desired output. Once the structure and connection functions of the BPNN are defined, the implementation process of the BP neural network algorithm is as follows [42,54,88,89]:

**Step 1: Network initialization.** Suppose that there are $n$, $p$, and $q$ nodes in the input, hidden and output layers, respectively. The network input is defined as $X_k = (x_1, x_1 \cdots, x_n)$, the connection weight between the input layer and the hidden layer is $\{w_{ij}\}$, where $i$=1, 2, ⋯, $n$ and $j$=1, 2, ⋯, $p$. The threshold for the hidden layer is represented by $\{\theta_j\}$, where $j$=1, 2, ⋯, $p$. The connection weight between the hidden layer and the output layer is expressed as $\{v_{jt}\}$, where $j$=1, 2, ⋯, $p$ and $t$=1, 2, ⋯, $q$. The threshold for the output layer is defined as $\{\gamma_t\}$, where $t$=1, 2, ⋯, $q$. The output of the neural network is $C_k = (c_1, c_1 \cdots, c_q)$.

**Step 2: Forward propagation.** In the process of forward propagation, the input set is processed sequentially by the hidden layer and the output layer. First, by utilizing the input $X_k$, the connection weight $w_{ij}$ and the threshold $\theta_j$, the hidden layer input $S_j^k$ and output $b_j^k$ can be calculated as follows:

$$S_j^k = \sum_{i=1}^{n} w_{ij} x_i - \theta_j \ k = 1, 2, \cdots, n; j = 1, 2, \cdots, p \tag{4}$$

$$b_j^k = f\left(S_j^k\right) = \frac{1}{1 + e^{-S_j^k}} \tag{5}$$

where $n$ is the number of nodes of the input layer, $p$ is the number of nodes of the hidden layer, and $f(\cdot)$ is the nonlinear activation function (sigmoid) for the hidden layer.

Subsequently, according to the output $b_j^k$ from the hidden layer, the connection weight $v_{jt}$ between the hidden layer and the output layer, and the threshold $\gamma_t$ of the output layer, the output layer input $L_j^k$ and output $C_j^k$ can be calculated as follows:

$$L_t^k = \sum_{i=1}^{P} v_{jt} b_j - \gamma_t \ j = 1, 2, \cdots, p; t = 1, 2, \cdots, q \tag{6}$$

$$C_t^k = f\left(L_t^k\right) = L_t^k \tag{7}$$

where $p$ is the number of nodes of the hidden layer, $q$ is the number of nodes of the output layer, and $f(\cdot)$ is the linear activation function for the output layer.

**Step 3: Error back propagation.** The backpropagation of the error adjusts the weights and thresholds of each layer according to the gradient descent principle. For the k-th input sample, the network prediction error $E_k$ can be determined by the network output $C_t^k$ and the actual value $y_t^k$:

$$E_k = \frac{1}{2} \sum_{t=1}^{q} \left(y_t^k - C_t^k\right)^2 \tag{8}$$

By using the gradient descent method, the connection weight $v_{jt}$ and threshold $\gamma_t$ associated with the output layer can be updated based on the error:

$$v'_{jt} = v_{jt} - \eta \left(\frac{\partial E_k}{\partial v_{jt}}\right) \tag{9}$$

$$\frac{\partial E_k}{\partial v_{jt}} = \frac{\partial E_k}{\partial C_t} \frac{\partial C_k}{\partial v_{jt}} = -\delta_t^k b_j \tag{10}$$

$$\gamma'_t = \gamma_t - \eta \left(\frac{\partial E_k}{\partial \gamma_t}\right) \tag{11}$$

$$\frac{\partial E_k}{\partial \gamma_t} = \frac{\partial E_k}{\partial C_t} \frac{\partial C_k}{\partial L_t} \frac{\partial L_t}{\partial \gamma_t} = \delta_t^k \tag{12}$$

where $v'_{jt}$ and $\gamma'_t$ are the updated weight and threshold associated with the output layer, respectively, $\eta$ is the learning rate of the neural network, and $\delta_t^k$ is the error gradient of the output layer.

The connection weight $w_{ij}$ and threshold $\theta_j$ related to the hidden layer are updated according to the following calculation.

$$w'_{ij} = w_{ij} - \eta \left(\frac{\partial E_k}{\partial w_{ij}}\right) \tag{13}$$

$$\frac{\partial E_k}{\partial w_{ij}} = \left(\sum_{t=1}^{q} \frac{\partial E_k}{\partial C_t} \frac{\partial C_t}{\partial L_t} \frac{\partial L_t}{\partial b_t}\right) \frac{\partial b_j}{\partial S_j} \frac{\partial S_j}{\partial w_{ij}} = -\delta_j x_i \tag{14}$$

$$\theta'_j = \theta_j - \eta \left(\frac{\partial E_k}{\partial \theta_j}\right) \tag{15}$$

$$\frac{\partial E_k}{\partial \theta_j} = \left(\sum_{t=1}^{q} \frac{\partial E_k}{\partial C_t} \frac{\partial C_t}{\partial L_t} \frac{\partial L_t}{\partial b_t}\right) \frac{\partial b_j}{\partial S_j} \frac{\partial S_j}{\partial \theta_j} = \delta_j \tag{16}$$

where $w'_{ij}$ and $\theta'_j$ are the updated weight and threshold associated with the hidden layer, respectively, $\eta$ is the learning rate of the neural network, and $\delta_j$ is the error gradient of the hidden layer.

**Step 4: Loop iteration.**

Determine whether the network error or the number of algorithm iterations meets expectations. If not, go back to Step 2 and repeat the steps above.

#### 2.3.3. Determination of the numbers of layers and nodes

In this investigation, the principal components ($F_i$) obtained after dimensionality reduction are used as the input variables of the BPNN model. Thus, the number of input layer nodes is equal to the number of principal components (6 in this paper). The output of the BPNN model is

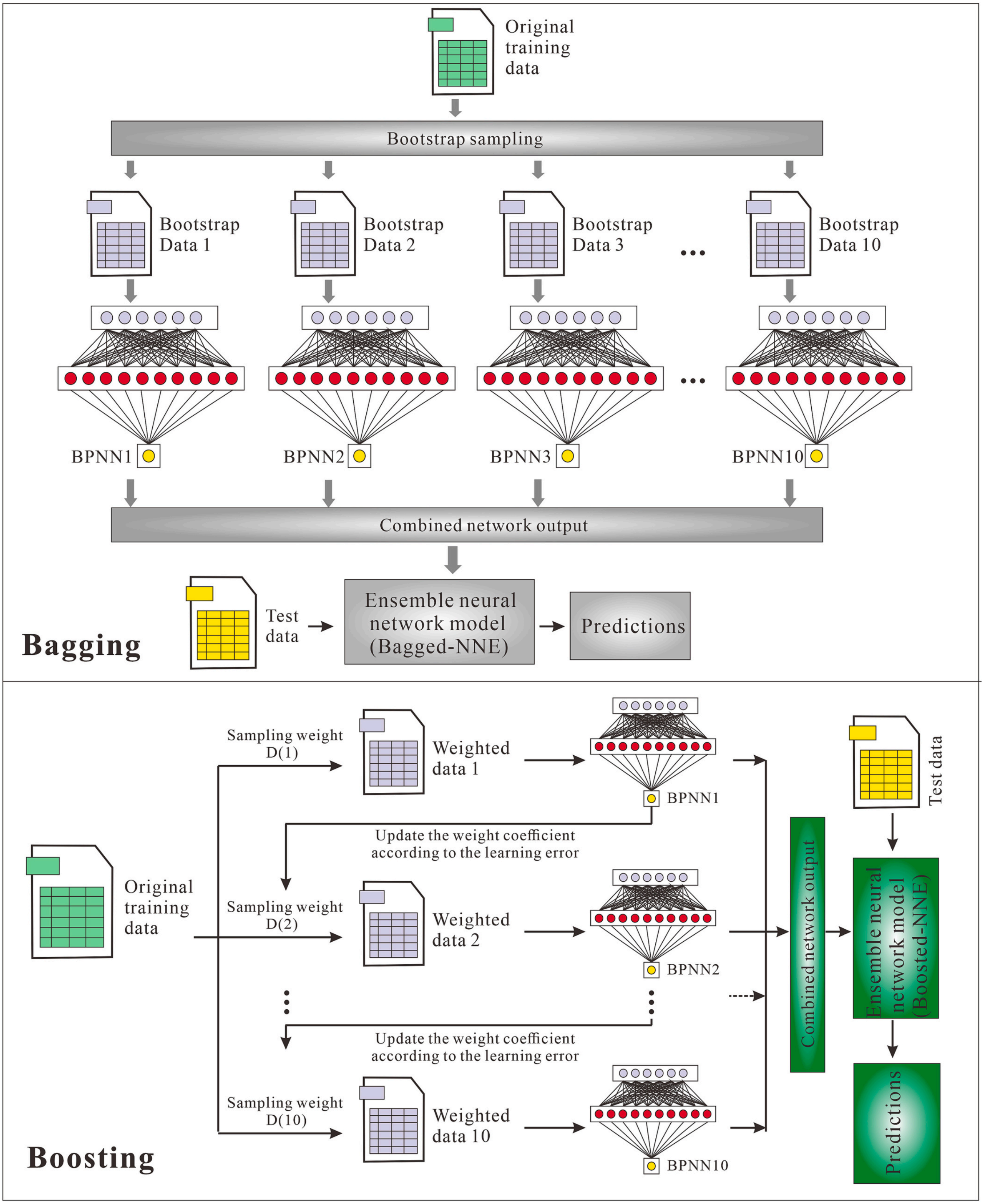


**Fig. 9.** Schematic of the Bagged-NNE model and Boosted-NNE model with 10 components in the testing phase.

the hydrocarbon column height (HCH), which can reflect the SHA volume in sandstone reservoirs. Therefore, the output layer node is 1. The BPNN model can theoretically have multiple hidden layers. However, the use of too many hidden layers may lead to problems such as overfitting and nonconvergence. This study selects a neural network structure with only one hidden layer, which can approach most nonlinear continuous functions with arbitrary accuracy [54]. Similarly, multiple hidden layer nodes can effectively reduce the training error but can also lead to an increase in training time and a decrease in convergence speed. According to the empirical formula [54,90–92], the following equation can be used to calculate and optimize the number of hidden layer nodes:

$$m = \sqrt{p+q} + a \tag{17}$$

where $m$ represents the number of hidden layer nodes, $p$ is the number of input layer nodes, $q$ is the number of output layer nodes, and $a$ is a constant between 1 and 10. For the BPNN model in this study, $p$ is 6 and $q$ is 1, so the number of hidden layer nodes is selected in the range of 3–13. By training BPNN models with different hidden layer nodes, the mean absolute error (MAE) and root mean square error (RMSE) can be used as accuracy indicators to optimize the number of hidden layer nodes.

### 2.3.4. Optimization of the network connection function

The selection of the activation (or transfer) function affects the learning rate and prediction output of the BPNN model [93,94]. In general, the activation function should be continuously differentiable to meet the requirement of gradient sensitivity during reverse error descent. In this study, the sigmoid function is adopted as the activation function of the hidden layer because its input set can be an arbitrary real number and can meet the large-scale and multiparameter requirements of the prediction model [54]. The activation function of the output layer varies according to the classification and regression problems. For the regression problem, the linear function can effectively preserve the format and range of the output values of the previous layer, thus ensuring that the output can be arbitrarily predicted values. Thus, the linear function is chosen as the activation function of the output layer of the BPNN model [54,95]. In this way, a three-layer BPNN model is constructed, as shown in Fig. 8.

## 2.4. Bagged neural network ensemble (bagged-NNE)

When using a single BPNN model for SHA prediction, there are often challenges such as slow convergence, low generalizability, and unstable output. Ensemble learning algorithms can effectively improve the prediction performance by integrating multiple basic (or weak) models [63–65]. The bagging (bootstrap aggregating) ensemble algorithm improves the unstable output of the predictive model by integrating multiple parallel-trained weak learners into one strong learner [66]. To construct the Bagged-NNE model, the bootstrapping sampling technique is applied to randomly select different training subsets for basic component model training (Fig. 9). The basic components are the typical three-layer BPNN model described above. After completing the training process, the outputs of all BPNN models need to be integrated according to specific strategies to build a powerful prediction model. For the SHA predictions in this study, the final output of the Bagged-NNE model is generally obtained by averaging the prediction results of all individual components, as shown in Eq. (18):

$$H(x) = \frac{1}{T}\sum_{i=1}^{T} h_i(x) \tag{18}$$

where $h_i$ is the component model, $h_i(x)$ represents the output of concrete instance x in $h_i$, $T$ is the number of components of Bagged-NNE model, and $H(x)$ is the final output of the Bagged-NNE model.

With BPNNs as components of the Bagged-NNE model, the pseudocode of the bagging algorithm is as follows:

```
Step 1: Training phase.
Input:
S: An original training sample set $\{(x_1,y_1),(x_2,y_2),\bullet\bullet\bullet,(x_N,y_N)\}$ with input space $X = (x_1,x_2\cdots x_D)$, where N is the number of training samples.
$T$: The number of components of Bagged-NNE model.
Output: The Bagged-NNE model
For t =1, 2, ···T do
    (1) Randomly select m inputs from the training sample set S as a subspace $S_t$.
    (2) Use the BP algorithm to train the component network $h_t$.
    (3) Put the training samples back into the original training set.
Obtain the network components of the Bagged-NNE model ($h_1$, $h_2\cdots h_t$).
Step 2: Prediction phase.
Project the given $x$ into different subspaces ($S_1,S_2\cdots S_T$).
Obtain predictions of different network components in the Bagged-NNE model: $h_t(S_t(x))(t = 1,2,\cdots T)$.
Calculate the prediction result of the Bagged-NNE model using the averaging method: $H(x) = \frac{1}{T}\sum_{t=1}^{T} h_t(S_t(x))$.
```

## 2.5. Boosted neural network ensemble (boosted-NNE)

Boosting is another popular way to build ensembles by integrating multiple serially trained weak learners, where each weak learner can adaptively adjust the training set based on the performance of previously added components [67,96]. Thus, these component models tend to be highly dependent on each other. By increasing the number of iterations, the boosting ensemble algorithm can produce a strong learner with near-perfect performance, thereby showing great advantages in improving the model accuracy. The construction process of Boosted-NNE is shown in Fig. 9. After training the first randomly generated BPNN model, the training set of each subsequent component is continuously adjusted according to the performance of the previous BPNN model. Generally, samples with larger errors tend to appear with greater weight in the updated training subset to build the new component. In this way, a strong Boosted-NNE model that satisfies the accuracy requirements can be obtained after many repeated iterations [97]. Using BPNNs as components of the Boosted-NNE model, the implementation process of the boosting algorithm is as follows:

```
Step 1: Training phase.
Input:
S: An original training sample set $\{(x_1,y_1),(x_2,y_2),\bullet\bullet\bullet,(x_N,y_N)\}$, where N is the number of training samples.
$T$: The number of components of the Boosted-NNE model.
$D_t$: The sample weight vector for the $t$-th iteration, where each sample has an initial weight of 1/N, and $D_1 = 1/N$.
Output: The Boosted-NNE model
For t =1, 2, ···T do
    (1) Use the BP algorithm to train the sample set under weights $D_t$ to obtain the weak neural network model $h_t$.
    (2) Calculate the error rate of neural network $h_t$:
```

$$E_t = \sum_{i=1}^{N} D_t^{\,2}(i)\left(y_i - h_t(x_i)\right)^2 \tag{19}$$

where $y_i$ and $h_t(x_i)$ are the true value and predicted value of the $i$-th sample of the neural network, respectively.

```
    (3) Update the weights of the training samples based on the error rate:
```

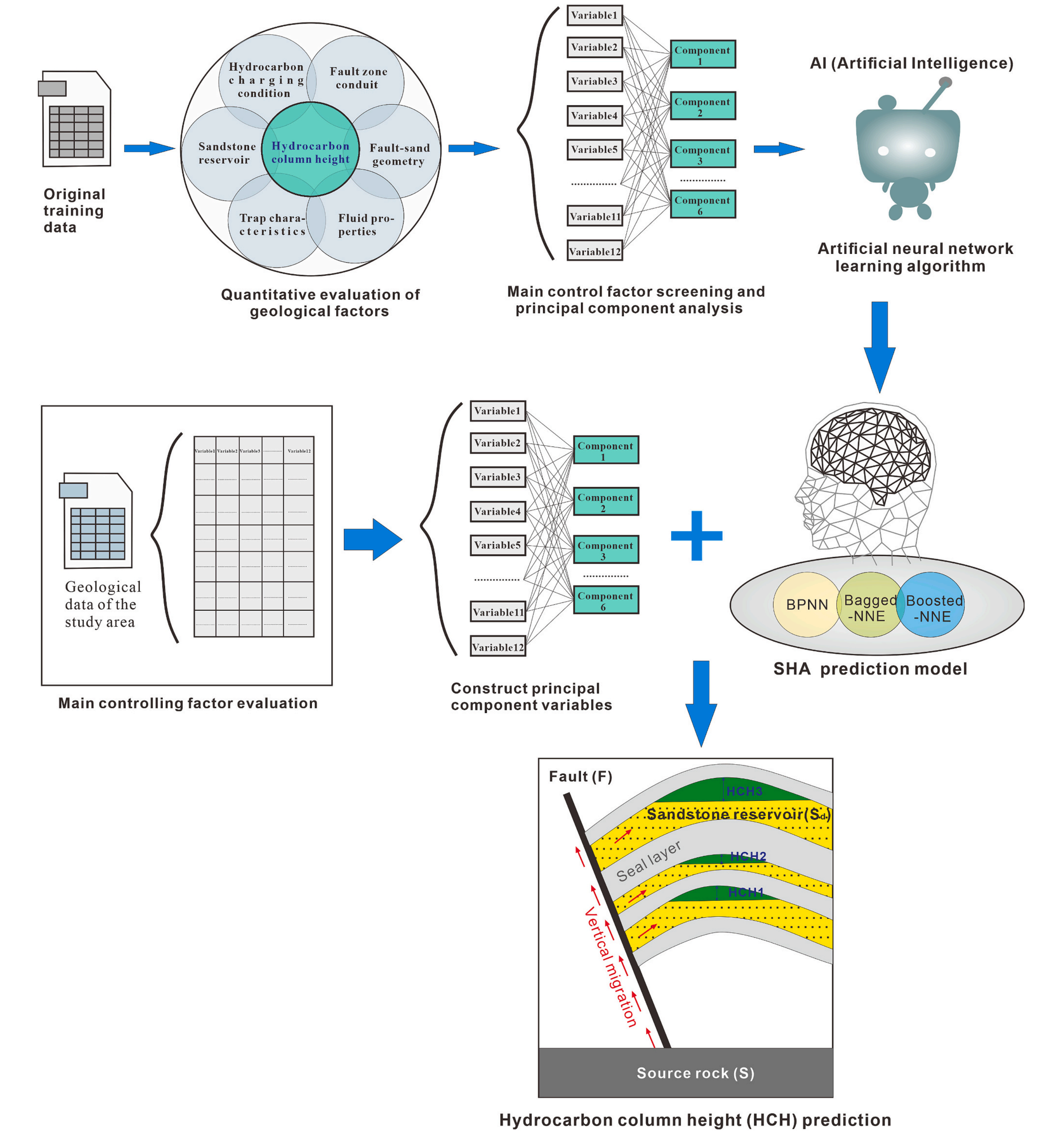

Original training data
Hydrocarbon charging condition
Fault zone conduit
Sandstone reservoir
Hydrocarbon column height
Fault-sand geometry
Trap characteristics
Fluid properties
Quantitative evaluation of geological factors
Variable1
Variable2
Variable3
Variable4
Variable5
...............
Variable11
Variable12
Component 1
Component 2
Component 3
...............
Component 6
Main control factor screening and principal component analysis
AI (Artificial Intelligence)
Artificial neural network learning algorithm
Geological data of the study area
Variable1
Variable2
Variable3
Variable12
Main controlling factor evaluation
Construct principal component variables
BPNN
Bagged -NNE
Boosted -NNE
SHA prediction model
Fault (F)
HCH3
Sandstone reservoir(S$_a$)
Seal layer
HCH2
HCH1
Vertical migration
Source rock (S)
Hydrocarbon column height (HCH) prediction

**Fig. 10.** Comprehensive technical guide for the shallow hydrocarbon accumulation and subsequent hydrocarbon column height prediction.

$$D_{t+1}(i) = \frac{D_t(i)\beta_t^{1-\frac{|y_i-h_t(x_i)|}{max(|y_i-h_t(x_i)|)}}}{Z_t} \tag{20}$$

where $\beta_t = E_t/(1-E_t)$ and $Z_t$ is the normalization factor.

(4) Calculate the weight of the neural network: $\alpha_t = ln(1/\beta_t)$

Obtain the network components of the Boosted-NNE model ($h_1$, $h_2 \cdots h_t$).

**Step 2: Prediction phase**

(1) Input the given x to obtain the predictions for the network components in the Boosted-NNE model: $h_t(x)$ $(t = 1, 2, \cdots T)$.

(2) Calculate the predictive output value of the Boosted-NNE model: $H(x) = \sum_{t=1}^{T} h_t(x) \times \alpha_t$ $(t = 1, 2, \cdots T)$

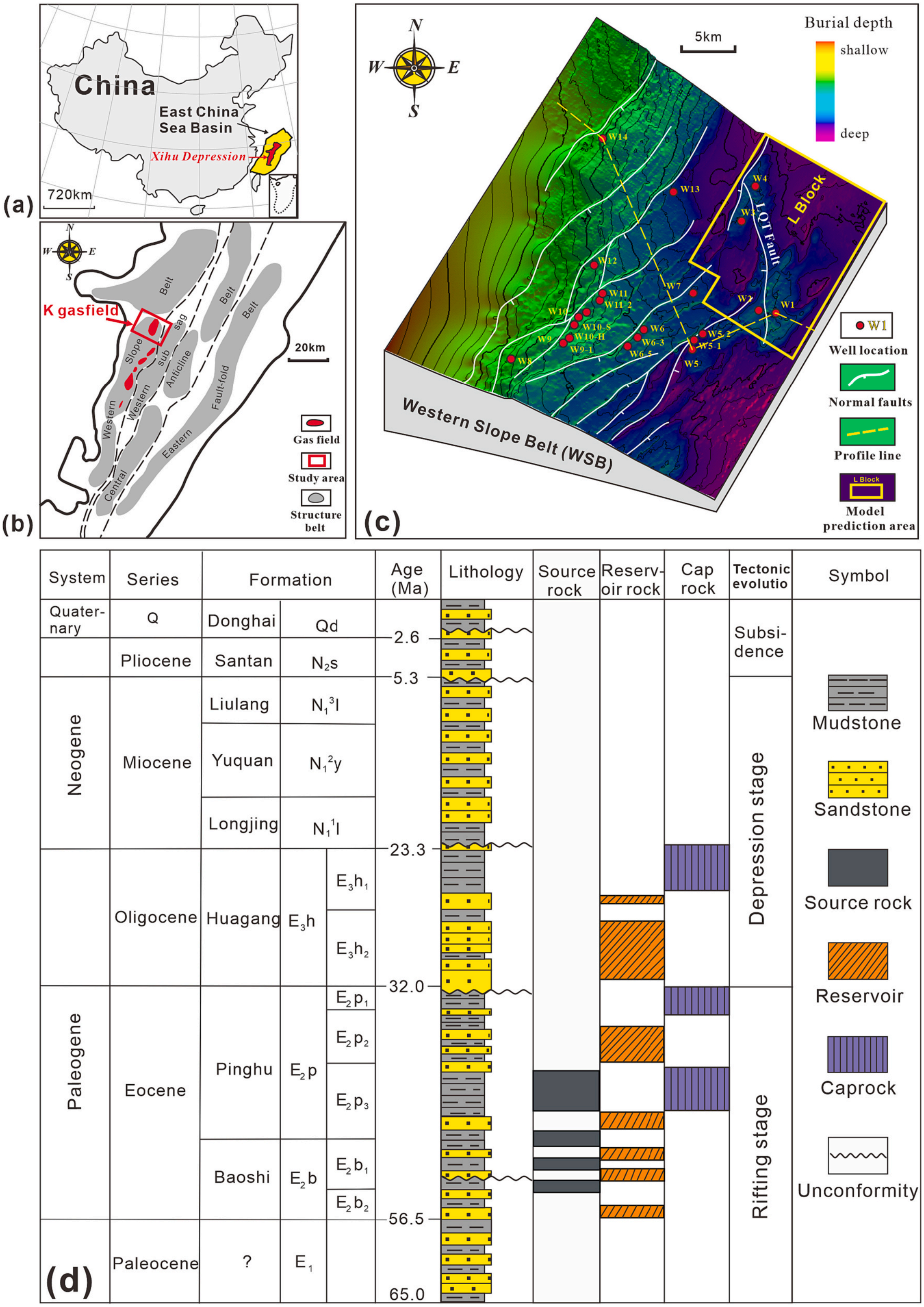


**Fig. 11.** (a) Structural maps showing the location of the East China Sea Basin and the Xihu Depression. (b) Tectonic map of the Xihu Depression showing the location of the K gasfield. (c) Structural map of the K gasfield showing the normal fault systems, exploration wells, profile location and modeling range. (d) Stratigraphic histogram showing the lithology, source–reservoir–cap configuration and tectonic stages in the K gasfield of the Xihu Depression.

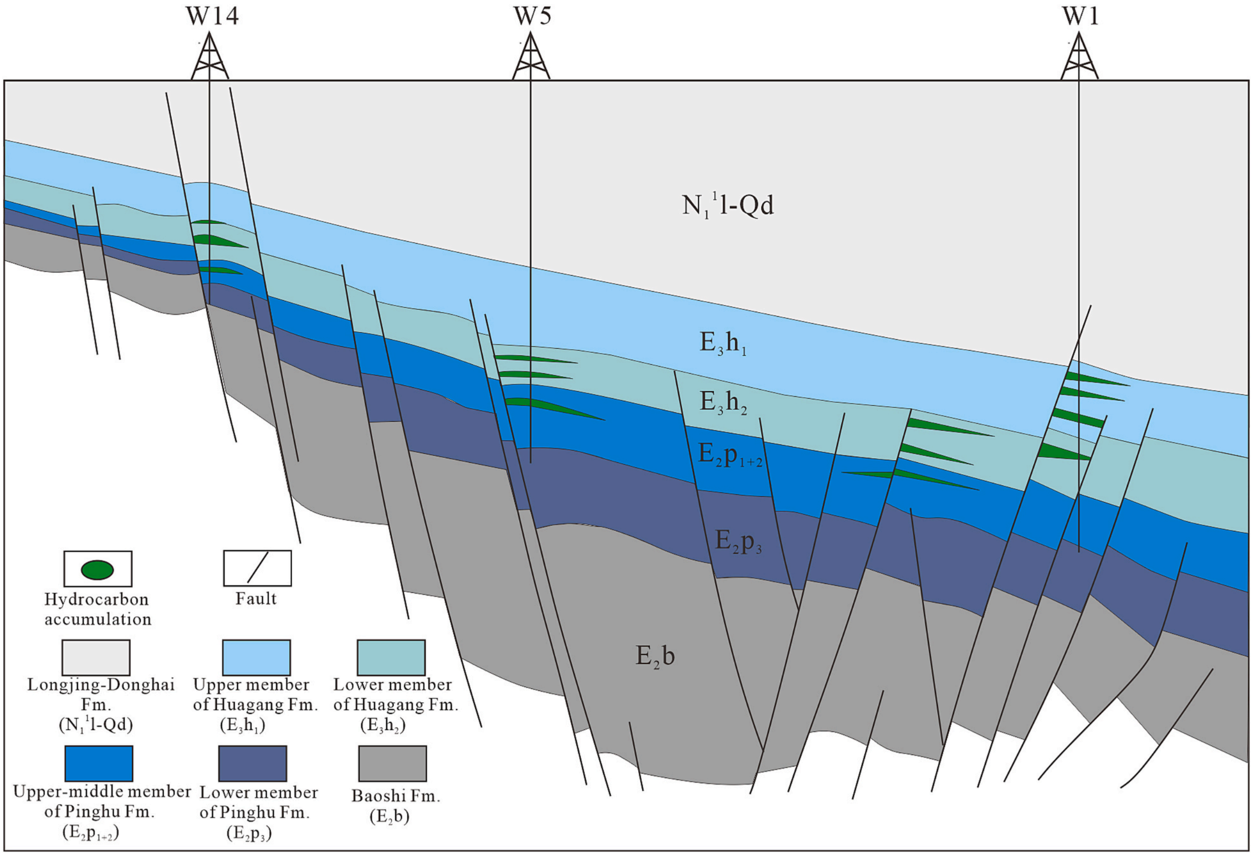

W14
W5
W1
$N_1^1l$-Qd
$E_3h_1$
$E_3h_2$
$E_2p_{1+2}$
$E_2p_3$
$E_2b$
Hydrocarbon accumulation
Fault
Longjing-Donghai Fm. ($N_1^1l$-Qd)
Upper member of Huagang Fm. ($E_3h_1$)
Lower member of Huagang Fm. ($E_3h_2$)
Upper-middle member of Pinghu Fm. ($E_2p_{1+2}$)
Lower member of Pinghu Fm. ($E_2p_3$)
Baoshi Fm. ($E_2b$)

**Fig. 12.** The profile of the K gasfield of the Xihu Depression. See Fig. 11c for the location of the profile.

**Table 2**
The distribution range of parameters related to source–fault–sand (S-F-$S_d$) evaluations.

| Geological factor | No. | Parameter | Minimum | Maximum | Average |
|---|---|---|---|---|---|
| Hydrocarbon charging condition | 1 | HER (Mtons/ $km^2$.Ma) | 0.10 | 1.20 | 0.686 |
| | 2 | CD (Ma) | 4.0 | 6.1 | 4.58 |
| | 3 | FSCT (m) | 150.00 | 500.00 | 346.86 |
| | 4 | VCD (m) | 2.10 | 1706.00 | 737.75 |
| Fault zone conduit | 5 | FT (m) | 3.00 | 337.76 | 147.48 |
| | 6 | SGR (%) | 8.00 | 87.00 | 43.16 |
| | 7 | NS (MPa) | 23.53 | 123.40 | 80.11 |
| Sandstone reservoir | 8 | ST (m) | 0.648 | 108.200 | 9.365 |
| | 9 | POR (%) | 0.001 | 21.244 | 9.928 |
| | 10 | PERM (mD) | 0.004 | 120.671 | 8.623 |
| fault-sand intersection geometry | 11 | SDA (°) | −23.15 | 25.67 | 0.70 |
| | 12 | FDA (°) | 64.28 | 87 | 74.362 |
| | 13 | FSIA (°) | 51.92 | 302.4 | 160.415 |
| Trap characteristics | 14 | TT | 0 | 1 | 0.625 |
| Fluid properties | 15 | P (MPa) | 23.96 | 69.5 | 45.029 |
| | 16 | DEN (g/$cm^3$) | 0.13 | 0.30 | 0.21 |
| | 17 | VIS ($10^{-1}$ mPa.s) | 0.21 | 0.33 | 0.27 |

### 2.6. Model performance evaluation and optimization

For the regression problem in this study, three indicators, the mean absolute error (MAE), root mean square error (RMSE) and R-squared ($R^2$), were selected to assess the performance of different predictive models [45]. The MAE indicates the average absolute value of the deviation between the predicted value and the actual value and can accurately reflect the actual prediction error without canceling each other out. The MAE is expressed as:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \widehat{y}_i| \tag{21}$$

where $y_i$ is the true value, $\widehat{y}_i$ is the predicted value, and $m$ is the number of samples.

The RMSE represents the square root of the mean squared error of the predicted value and the true value, allowing the dimension of the predicted value to be consistent with the original value. The RMSE is given by:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \widehat{y}_i)^2} \tag{22}$$

The $R^2$ evaluates the quality of the predictive model by analyzing the difference between the predicted value and the actual value [98], and the closer to 1 this value is, the higher the accuracy of the model. The $R^2$ can be expressed as:

$$R^2 = 1 - \frac{\sum_{i}^{m}(y_i - \widehat{y}_i)^2}{\sum_{i}^{m}(y_i - \overline{y})^2} \tag{23}$$

where $\overline{y}$ denotes the average of the observed data.

In this way, by comparing the performance of the BPNN, Bagged-NNE, and Boosted-NNE models, the most suitable model for SHA prediction can be determined.

**Table 3**
Correlation ranking of controlling factors for hydrocarbon column height of SHA.

| Parameter | | HER | CD | FSCT | VCD | FT | SGR | NS | ST | POR | PERM | SDA | FDA | FSIA | TT | P | DEN | VIS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Correlation coefficient | Gray | 0.993 | 0.989 | 0.662 | 0.591 | 0.827 | 0.995 | 0.893 | 0.997 | 0.990 | 0.991 | 0.986 | 0.916 | 0.817 | 0.992 | 0.942 | 0.891 | 0.885 |
| | Pearson | −0.356** | −0.459** | −0.009 | −0.341** | 0.036 | 0.488** | −0.187** | 0.709** | 0.280** | 0.228** | −0.277** | −0.127* | 0.037 | −0.345** | −0.039 | −0.04 | −0.04 |
| | Kendall | −0.448** | −0.489** | −0.013 | −0.302** | 0.063 | 0.389** | −0.226** | 0.510** | 0.353** | 0.526** | −0.347** | −0.043 | 0.266** | −0.365** | −0.042 | −0.036 | −0.036 |
| | Spearman | −0.601** | −0.619** | −0.021 | −0.422** | 0.091 | 0.538** | −0.333** | 0.536** | 0.499** | 0.720** | −0.480** | −0.063 | 0.364** | −0.456** | −0.059 | −0.05 | −0.05 |
| Relevance ranking | Lowest ranking | 4 | 7 | 17 | 17 | 16 | 5 | 11 | 5 | 7 | 9 | 8 | 13 | 15 | 8 | 14 | 15 | 16 |
| | Highest ranking | 1 | 3 | 16 | 6 | 12 | 2 | 10 | 1 | 6 | 1 | 7 | 10 | 10 | 4 | 9 | 12 | 13 |
| | Average ranking | 3 | 4 | 16.75 | 10.25 | 13.5 | 3.25 | 10.75 | 2.25 | 6.5 | 4.25 | 7.75 | 11.75 | 12.5 | 5.75 | 12.75 | 13.5 | 14.5 |

Note: Sig > 0.05 indicates no significant correlation; sig < 0.05 (*) indicates a significant correlation; sig < 0.01 (**) indicates an extremely significant correlation.

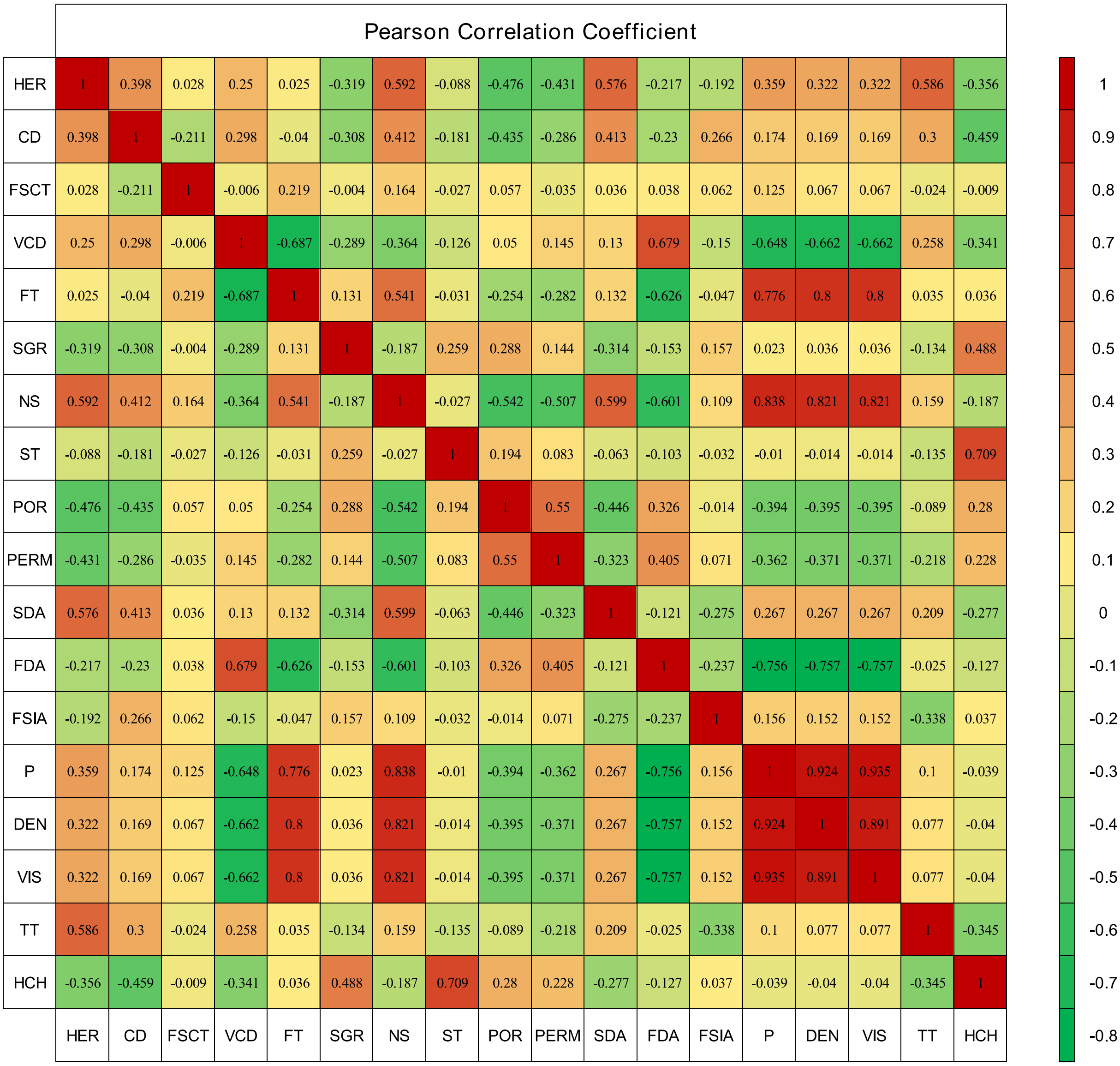
### 2.7. *Workflow to predict SHA*

The above ANN models were applied to quantitatively predict SHA in the three steps. First, the main controlling factors affecting SHA in the target area were systematically quantified. These factors mainly include geological conditions related to source rocks (S), fault zones (F) and sandstone reservoirs ($S_d$). Second, all the main controlling factors were standardized and dimensionally reduced to obtain the principal components that retain meaningful properties of the original data. Finally, these principal components were input into the optimal neural network model for SHA prediction, thereby obtaining the HCH of sandstone reservoirs within the fault-bounded traps. Prospectively, the predicted HCH can be used for predrilling volume assessment and to determine if a prospect contains enough hydrocarbons to justify drilling an exploration or appraisal well. Fig. 10 is a comprehensive technical guide for the entire research methodology in this contribution.

## 3. Geological setting

The Xihu Depression is the focus of conventional hydrocarbon exploration in the central East China Sea Basin, occupying a total area of 5.18 × $10^4$ $km^2$ [99,100] (Fig. 11a). This area can be divided into four structural subunits, from east to west, namely, the Eastern Fault-Fold Belt, the Central Anticline Belt, the Western Sub Sag and the Western Slope Belt (Fig. 11b). On the whole, the Xihu Depression has undergone three tectonic evolution stages: the rifting stage (65.0–32.0 Ma), thermal depression stage (32.0–5.3 Ma) and regional subsidence stage (5.3 Ma-present) [101–103] (Fig. 11d). In the process of depression evolution, the stratigraphic units were mainly filled by marine-continental transitional deposits [104]. These units can be divided from top to bottom into the Quaternary Donghai Group (Qd), Pliocene Santan Formation ($N_2s$), Miocene Liulang Formation ($N_1^3l$), Yuquan Formation ($N_1^2y$), Longjing Formation ($N_1^1l$), Oligocene Huagang Formation ($E_3h$), Eocene Pinghu Formation ($E_2p$) and Baoshi Formation ($E_2b$). According to the source–reservoir–cap rock units, the dark mudstone and coal seam developed in the lower member of the Pinghu Formation ($E_2p_3$) act as the main source rocks, while the sandstone distributed in the $E_2b$, $E_2p$ and $E_3h$ formations serves as the primary set of oil–gas reservoirs. The continuous thick mudstone on the top of the $E_2p$ and $E_3h$ formations is the regional caprock to prevent hydrocarbons from vertical leakage. Thus, the unique source rock and overlying multilayered sandstone reservoirs constitute the petroleum system within the $E_2p$ and $E_3h$ formations [11], covering most of the oil–gas accumulations discovered thus far.

In this investigation, the K gasfield in the Western Slope Belt is selected as the study area (Fig. 11c). In this area, the southwest- to northeast-trending normal faults and sandstone reservoirs deposited under marine-continental transition environments are developed. These normal faults can spatially connect the source rocks of the $E_2p_3$ member of the Pinghu Formation with overlying sandstone reservoirs, thus providing potential vertical seepage conduits for SHA [11,12]. In addition, due to the complex depositional environment, sandstone reservoirs adjacent to faults differ greatly in thickness and connectivity. The $E_3h$ formation is filled with sandstone bodies in lacustrine, fluvial, and deltaic environments with good spatial connectivity. The upper member ($E_2p_1$) and middle member ($E_2p_2$) of the Pinghu Formation are characterized by thick-bedded distributary channel sandstones in a deltaic environment, whereas the $E_2p_3$ member is dominated by thin-bedded sand bodies in a tidal flat environment. Thus, controlled by the heterogeneous source–fault–sand configuration, the hydrocarbons in the K gasfield are mostly found in fault-bounded traps after long-range vertical migration along the faults [105,106] (Fig. 12) concentrated in the shallowly buried sandstone reservoirs of the $E_2p_{1,}$ $E_2p_2$ and $E_3h$ formations. Therefore, SHA control factor analysis and quantitative prediction are essential to reduce future exploration risks in the K gasfield and Xihu Depression. The K gasfield is the most highly explored and data-rich area



| | HER | CD | FSCT | VCD | FT | SGR | NS | ST | POR | PERM | SDA | FDA | FSIA | P | DEN | VIS | TT | HCH |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HER | 1 | 0.398 | 0.028 | 0.25 | 0.025 | -0.319 | 0.592 | -0.088 | -0.476 | -0.431 | 0.576 | -0.217 | -0.192 | 0.359 | 0.322 | 0.322 | 0.586 | -0.356 |
| CD | 0.398 | 1 | -0.211 | 0.298 | -0.04 | -0.308 | 0.412 | -0.181 | -0.435 | -0.286 | 0.413 | -0.23 | 0.266 | 0.174 | 0.169 | 0.169 | 0.3 | -0.459 |
| FSCT | 0.028 | -0.211 | 1 | -0.006 | 0.219 | -0.004 | 0.164 | -0.027 | 0.057 | -0.035 | 0.036 | 0.038 | 0.062 | 0.125 | 0.067 | 0.067 | -0.024 | -0.009 |
| VCD | 0.25 | 0.298 | -0.006 | 1 | -0.687 | -0.289 | -0.364 | -0.126 | 0.05 | 0.145 | 0.13 | 0.679 | -0.15 | -0.648 | -0.662 | -0.662 | 0.258 | -0.341 |
| FT | 0.025 | -0.04 | 0.219 | -0.687 | 1 | 0.131 | 0.541 | -0.031 | -0.254 | -0.282 | 0.132 | -0.626 | -0.047 | 0.776 | 0.8 | 0.8 | 0.035 | 0.036 |
| SGR | -0.319 | -0.308 | -0.004 | -0.289 | 0.131 | 1 | -0.187 | 0.259 | 0.288 | 0.144 | -0.314 | -0.153 | 0.157 | 0.023 | 0.036 | 0.036 | -0.134 | 0.488 |
| NS | 0.592 | 0.412 | 0.164 | -0.364 | 0.541 | -0.187 | 1 | -0.027 | -0.542 | -0.507 | 0.599 | -0.601 | 0.109 | 0.838 | 0.821 | 0.821 | 0.159 | -0.187 |
| ST | -0.088 | -0.181 | -0.027 | -0.126 | -0.031 | 0.259 | -0.027 | 1 | 0.194 | 0.083 | -0.063 | -0.103 | -0.032 | -0.01 | -0.014 | -0.014 | -0.135 | 0.709 |
| POR | -0.476 | -0.435 | 0.057 | 0.05 | -0.254 | 0.288 | -0.542 | 0.194 | 1 | 0.55 | -0.446 | 0.326 | -0.014 | -0.394 | -0.395 | -0.395 | -0.089 | 0.28 |
| PERM | -0.431 | -0.286 | -0.035 | 0.145 | -0.282 | 0.144 | -0.507 | 0.083 | 0.55 | 1 | -0.323 | 0.405 | 0.071 | -0.362 | -0.371 | -0.371 | -0.218 | 0.228 |
| SDA | 0.576 | 0.413 | 0.036 | 0.13 | 0.132 | -0.314 | 0.599 | -0.063 | -0.446 | -0.323 | 1 | -0.121 | -0.275 | 0.267 | 0.267 | 0.267 | 0.209 | -0.277 |
| FDA | -0.217 | -0.23 | 0.038 | 0.679 | -0.626 | -0.153 | -0.601 | -0.103 | 0.326 | 0.405 | -0.121 | 1 | -0.237 | -0.756 | -0.757 | -0.757 | -0.025 | -0.127 |
| FSIA | -0.192 | 0.266 | 0.062 | -0.15 | -0.047 | 0.157 | 0.109 | -0.032 | -0.014 | 0.071 | -0.275 | -0.237 | 1 | 0.156 | 0.152 | 0.152 | -0.338 | 0.037 |
| P | 0.359 | 0.174 | 0.125 | -0.648 | 0.776 | 0.023 | 0.838 | -0.01 | -0.394 | -0.362 | 0.267 | -0.756 | 0.156 | 1 | 0.924 | 0.935 | 0.1 | -0.039 |
| DEN | 0.322 | 0.169 | 0.067 | -0.662 | 0.8 | 0.036 | 0.821 | -0.014 | -0.395 | -0.371 | 0.267 | -0.757 | 0.152 | 0.924 | 1 | 0.891 | 0.077 | -0.04 |
| VIS | 0.322 | 0.169 | 0.067 | -0.662 | 0.8 | 0.036 | 0.821 | -0.014 | -0.395 | -0.371 | 0.267 | -0.757 | 0.152 | 0.935 | 0.891 | 1 | 0.077 | -0.04 |
| TT | 0.586 | 0.3 | -0.024 | 0.258 | 0.035 | -0.134 | 0.159 | -0.135 | -0.089 | -0.218 | 0.209 | -0.025 | -0.338 | 0.1 | 0.077 | 0.077 | 1 | -0.345 |
| HCH | -0.356 | -0.459 | -0.009 | -0.341 | 0.036 | 0.488 | -0.187 | 0.709 | 0.28 | 0.228 | -0.277 | -0.127 | 0.037 | -0.039 | -0.04 | -0.04 | -0.345 | 1 |

**Fig. 13.** The Pearson correlation coefficient heatmap showing the relationships between the 18 variables.

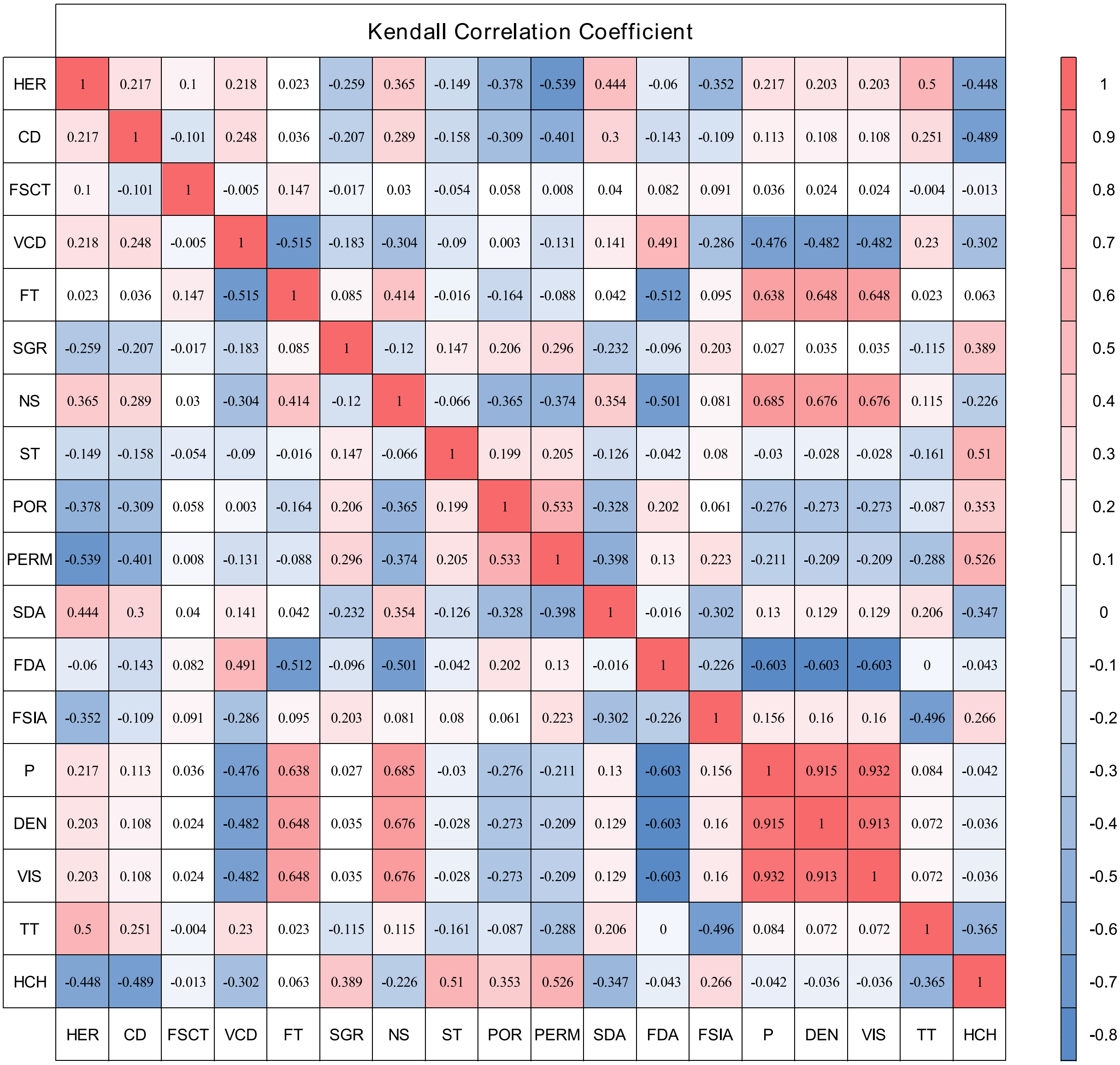
in the Western Slope Belt and is suitable for machine learning, SHA and corresponding HCH prediction studies.

## 4. Results

### *4.1. Main control factor screening results*

In this investigation, seventeen geological factors affecting SHAs were systematically and quantitatively characterized. These parameters can be grouped into six categories: hydrocarbon supply conditions, fault zone conduits, sandstone reservoirs, fault–sand intersection geometry, trap characteristics, and fluid properties. In the hydrocarbon supply evaluation, the hydrocarbon expulsion rate of source rocks was obtained by numerical simulation using Petro-mod software, the hydrocarbon charging duration was determined by combining the homogenization temperature of fluid inclusion and the corresponding thermal evolution history. Meanwhile, the source-fault contact thickness and vertical charging distance were measured using interpreted geological or seismic profiles. On the basis of measuring the fault distance and dip angle, the shale gouge ratio and normal stress in the fault zone were calculated by Eq. (2) and Eq. (3). In terms of sandstone reservoir, their thickness, porosity and permeability properties were characterized by the logging interpretation of 22 drilling wells. According to the sonic time curve of drilling wells, the equivalent depth method was employed to determine the formation fluid pressure at different burial depths. On the other hand, the changes in hydrocarbon viscosity and density under different temperature and pressure conditions were obtained by PVT experiment data from 5 drilling wells. For the predicted target parameter, the hydrocarbon column height, is determined by calculating the vertical distance from the gas-water interface depth derived from logging interpretation or reservoir profile to the apex depth of the trap structure. In this way, a total of 313 sets of sample data for the main control factor screening and model building were obtained.

Table 2 shows all controlling factors and corresponding parameter ranges related to S–F–$S_d$. By evaluating the gray correlation degree, Pearson correlation coefficient, Spearman correlation coefficient and Kendall correlation coefficient between each parameter and the corresponding HCH, the control factors were comprehensively ranked by the arithmetic average of multiple correlation parameters. Table 3 and Figs. 13–15 show the correlation coefficients and corresponding average rankings between 17 geological control factors and the HCH. To eliminate unimportant factors, a significance test was adopted to



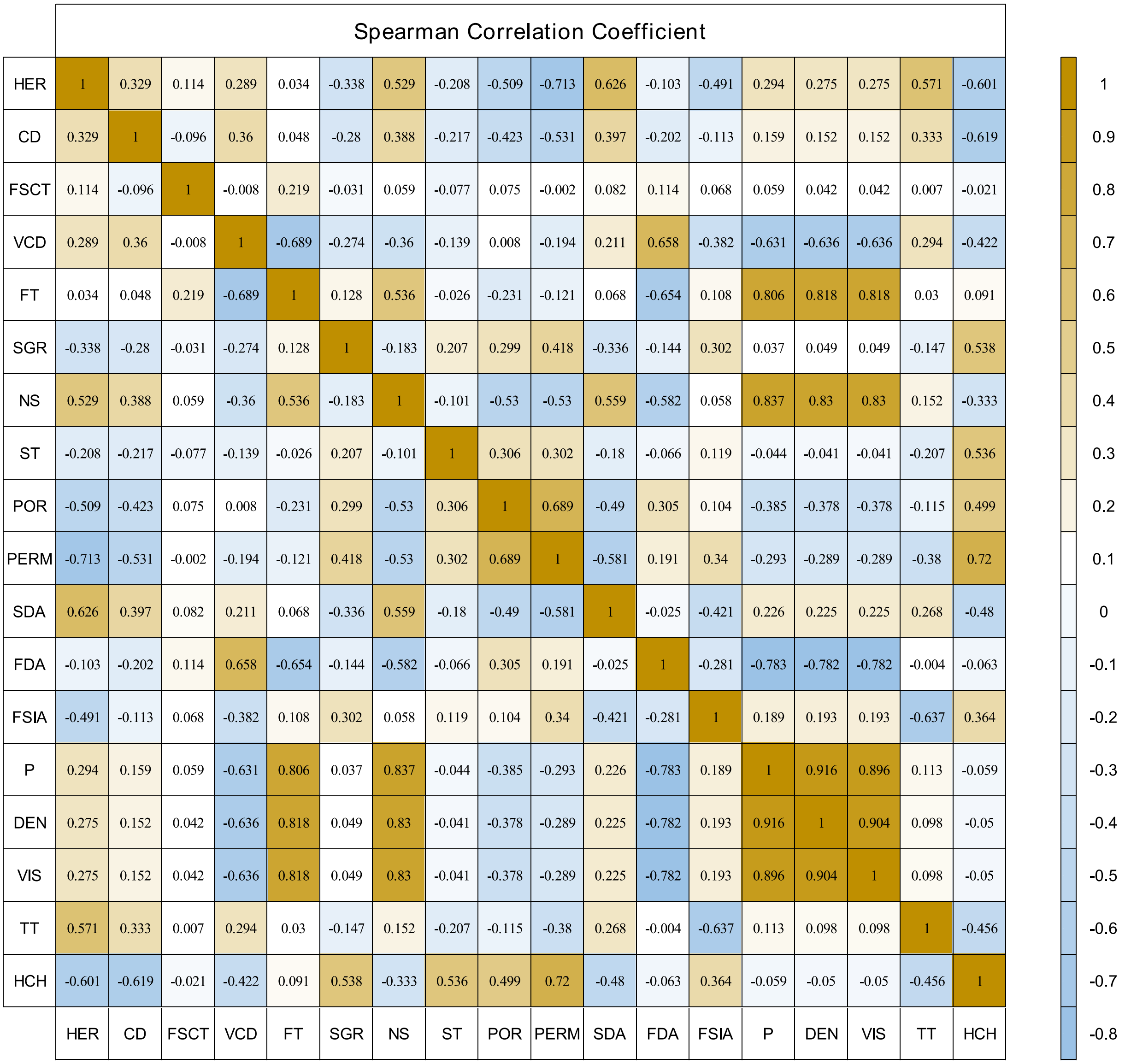
| | HER | CD | FSCT | VCD | FT | SGR | NS | ST | POR | PERM | SDA | FDA | FSIA | P | DEN | VIS | TT | HCH |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HER | 1 | 0.217 | 0.1 | 0.218 | 0.023 | -0.259 | 0.365 | -0.149 | -0.378 | -0.539 | 0.444 | -0.06 | -0.352 | 0.217 | 0.203 | 0.203 | 0.5 | -0.448 |
| CD | 0.217 | 1 | -0.101 | 0.248 | 0.036 | -0.207 | 0.289 | -0.158 | -0.309 | -0.401 | 0.3 | -0.143 | -0.109 | 0.113 | 0.108 | 0.108 | 0.251 | -0.489 |
| FSCT | 0.1 | -0.101 | 1 | -0.005 | 0.147 | -0.017 | 0.03 | -0.054 | 0.058 | 0.008 | 0.04 | 0.082 | 0.091 | 0.036 | 0.024 | 0.024 | -0.004 | -0.013 |
| VCD | 0.218 | 0.248 | -0.005 | 1 | -0.515 | -0.183 | -0.304 | -0.09 | 0.003 | -0.131 | 0.141 | 0.491 | -0.286 | -0.476 | -0.482 | -0.482 | 0.23 | -0.302 |
| FT | 0.023 | 0.036 | 0.147 | -0.515 | 1 | 0.085 | 0.414 | -0.016 | -0.164 | -0.088 | 0.042 | -0.512 | 0.095 | 0.638 | 0.648 | 0.648 | 0.023 | 0.063 |
| SGR | -0.259 | -0.207 | -0.017 | -0.183 | 0.085 | 1 | -0.12 | 0.147 | 0.206 | 0.296 | -0.232 | -0.096 | 0.203 | 0.027 | 0.035 | 0.035 | -0.115 | 0.389 |
| NS | 0.365 | 0.289 | 0.03 | -0.304 | 0.414 | -0.12 | 1 | -0.066 | -0.365 | -0.374 | 0.354 | -0.501 | 0.081 | 0.685 | 0.676 | 0.676 | 0.115 | -0.226 |
| ST | -0.149 | -0.158 | -0.054 | -0.09 | -0.016 | 0.147 | -0.066 | 1 | 0.199 | 0.205 | -0.126 | -0.042 | 0.08 | -0.03 | -0.028 | -0.028 | -0.161 | 0.51 |
| POR | -0.378 | -0.309 | 0.058 | 0.003 | -0.164 | 0.206 | -0.365 | 0.199 | 1 | 0.533 | -0.328 | 0.202 | 0.061 | -0.276 | -0.273 | -0.273 | -0.087 | 0.353 |
| PERM | -0.539 | -0.401 | 0.008 | -0.131 | -0.088 | 0.296 | -0.374 | 0.205 | 0.533 | 1 | -0.398 | 0.13 | 0.223 | -0.211 | -0.209 | -0.209 | -0.288 | 0.526 |
| SDA | 0.444 | 0.3 | 0.04 | 0.141 | 0.042 | -0.232 | 0.354 | -0.126 | -0.328 | -0.398 | 1 | -0.016 | -0.302 | 0.13 | 0.129 | 0.129 | 0.206 | -0.347 |
| FDA | -0.06 | -0.143 | 0.082 | 0.491 | -0.512 | -0.096 | -0.501 | -0.042 | 0.202 | 0.13 | -0.016 | 1 | -0.226 | -0.603 | -0.603 | -0.603 | 0 | -0.043 |
| FSIA | -0.352 | -0.109 | 0.091 | -0.286 | 0.095 | 0.203 | 0.081 | 0.08 | 0.061 | 0.223 | -0.302 | -0.226 | 1 | 0.156 | 0.16 | 0.16 | -0.496 | 0.266 |
| P | 0.217 | 0.113 | 0.036 | -0.476 | 0.638 | 0.027 | 0.685 | -0.03 | -0.276 | -0.211 | 0.13 | -0.603 | 0.156 | 1 | 0.915 | 0.932 | 0.084 | -0.042 |
| DEN | 0.203 | 0.108 | 0.024 | -0.482 | 0.648 | 0.035 | 0.676 | -0.028 | -0.273 | -0.209 | 0.129 | -0.603 | 0.16 | 0.915 | 1 | 0.913 | 0.072 | -0.036 |
| VIS | 0.203 | 0.108 | 0.024 | -0.482 | 0.648 | 0.035 | 0.676 | -0.028 | -0.273 | -0.209 | 0.129 | -0.603 | 0.16 | 0.932 | 0.913 | 1 | 0.072 | -0.036 |
| TT | 0.5 | 0.251 | -0.004 | 0.23 | 0.023 | -0.115 | 0.115 | -0.161 | -0.087 | -0.288 | 0.206 | 0 | -0.496 | 0.084 | 0.072 | 0.072 | 1 | -0.365 |
| HCH | -0.448 | -0.489 | -0.013 | -0.302 | 0.063 | 0.389 | -0.226 | 0.51 | 0.353 | 0.526 | -0.347 | -0.043 | 0.266 | -0.042 | -0.036 | -0.036 | -0.365 | 1 |

**Fig. 14.** The Kendall correlation coefficient heatmap showing the relationships between the 18 variables.

quantitatively identify if each parameter was significantly correlated with HCH, where a smaller Sig value indicates a higher correlation degree. More specifically, Sig $> 0.05$ illustrates no significant correlation between the two variables, which is represented by the only correlation coefficient value in the table. Sig $< 0.05$ and Sig $< 0.01$ indicate significant and extremely significant correlations between the two variables, respectively, and are represented by "*" and "**" above the correlation coefficients tested. According to Table 3 and Fig. 16, the average rankings of fluid pressure (P), fault throw (FT), hydrocarbon density (DEN), hydrocarbon viscosity (VIS), and fault-source contact thickness (FSCT) were the lowest and were not significantly correlated with HCH (sig $> 0.05$). Thus, these 5 parameters were identified as nonmain controlling factors. Instead, the remaining 12 geological parameters were identified as the main controlling factors for SHA in this study area. It should be noted that the main controlling factors of different petroliferous basins or depressions tend to vary due to the complexity and heterogeneity of geological conditions.

### 4.2. Principal component analysis results

In this study, the principal component analysis (PCA) method was applied to reduce the dimensionality of the main controlling factors after standardization. According to the critical condition that the cumulative variance contribution is greater than or equal to 80%–85% (Table 4), the top six principal components were determined in this study to replace the original 12 main controlling factors. These six principal components are independent of each other and ideally approximate the information of all main controlling factors affecting SHA. By using the principal component loading matrix and eigenvalues, six principal components ($F_1$, $F_2$, $F_3$, $F_4$, $F_5$, and $F_6$) can be conveniently expressed through the loading coefficient of the main controlling factors. As shown in Table 5, each principal component focuses on conveying different geological information. For example, the absolute values of the load coefficients of HER, NS, POR and SDA corresponding to F1 are significantly greater than those of the other parameters, suggesting that $F_1$ is more sensitive to changes in these four parameters. In contrast, the loading coefficients of VCD and FDA in $F_2$ are larger than those of the other parameters, suggesting that $F_2$ is more likely to reflect changes in the vertical charging distance and fault dip angle.



Spearman Correlation Coefficient

| | HER | CD | FSCT | VCD | FT | SGR | NS | ST | POR | PERM | SDA | FDA | FSIA | P | DEN | VIS | TT | HCH |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HER | 1 | 0.329 | 0.114 | 0.289 | 0.034 | -0.338 | 0.529 | -0.208 | -0.509 | -0.713 | 0.626 | -0.103 | -0.491 | 0.294 | 0.275 | 0.275 | 0.571 | -0.601 |
| CD | 0.329 | 1 | -0.096 | 0.36 | 0.048 | -0.28 | 0.388 | -0.217 | -0.423 | -0.531 | 0.397 | -0.202 | -0.113 | 0.159 | 0.152 | 0.152 | 0.333 | -0.619 |
| FSCT | 0.114 | -0.096 | 1 | -0.008 | 0.219 | -0.031 | 0.059 | -0.077 | 0.075 | -0.002 | 0.082 | 0.114 | 0.068 | 0.059 | 0.042 | 0.042 | 0.007 | -0.021 |
| VCD | 0.289 | 0.36 | -0.008 | 1 | -0.689 | -0.274 | -0.36 | -0.139 | 0.008 | -0.194 | 0.211 | 0.658 | -0.382 | -0.631 | -0.636 | -0.636 | 0.294 | -0.422 |
| FT | 0.034 | 0.048 | 0.219 | -0.689 | 1 | 0.128 | 0.536 | -0.026 | -0.231 | -0.121 | 0.068 | -0.654 | 0.108 | 0.806 | 0.818 | 0.818 | 0.03 | 0.091 |
| SGR | -0.338 | -0.28 | -0.031 | -0.274 | 0.128 | 1 | -0.183 | 0.207 | 0.299 | 0.418 | -0.336 | -0.144 | 0.302 | 0.037 | 0.049 | 0.049 | -0.147 | 0.538 |
| NS | 0.529 | 0.388 | 0.059 | -0.36 | 0.536 | -0.183 | 1 | -0.101 | -0.53 | -0.53 | 0.559 | -0.582 | 0.058 | 0.837 | 0.83 | 0.83 | 0.152 | -0.333 |
| ST | -0.208 | -0.217 | -0.077 | -0.139 | -0.026 | 0.207 | -0.101 | 1 | 0.306 | 0.302 | -0.18 | -0.066 | 0.119 | -0.044 | -0.041 | -0.041 | -0.207 | 0.536 |
| POR | -0.509 | -0.423 | 0.075 | 0.008 | -0.231 | 0.299 | -0.53 | 0.306 | 1 | 0.689 | -0.49 | 0.305 | 0.104 | -0.385 | -0.378 | -0.378 | -0.115 | 0.499 |
| PERM | -0.713 | -0.531 | -0.002 | -0.194 | -0.121 | 0.418 | -0.53 | 0.302 | 0.689 | 1 | -0.581 | 0.191 | 0.34 | -0.293 | -0.289 | -0.289 | -0.38 | 0.72 |
| SDA | 0.626 | 0.397 | 0.082 | 0.211 | 0.068 | -0.336 | 0.559 | -0.18 | -0.49 | -0.581 | 1 | -0.025 | -0.421 | 0.226 | 0.225 | 0.225 | 0.268 | -0.48 |
| FDA | -0.103 | -0.202 | 0.114 | 0.658 | -0.654 | -0.144 | -0.582 | -0.066 | 0.305 | 0.191 | -0.025 | 1 | -0.281 | -0.783 | -0.782 | -0.782 | -0.004 | -0.063 |
| FSIA | -0.491 | -0.113 | 0.068 | -0.382 | 0.108 | 0.302 | 0.058 | 0.119 | 0.104 | 0.34 | -0.421 | -0.281 | 1 | 0.189 | 0.193 | 0.193 | -0.637 | 0.364 |
| P | 0.294 | 0.159 | 0.059 | -0.631 | 0.806 | 0.037 | 0.837 | -0.044 | -0.385 | -0.293 | 0.226 | -0.783 | 0.189 | 1 | 0.916 | 0.896 | 0.113 | -0.059 |
| DEN | 0.275 | 0.152 | 0.042 | -0.636 | 0.818 | 0.049 | 0.83 | -0.041 | -0.378 | -0.289 | 0.225 | -0.782 | 0.193 | 0.916 | 1 | 0.904 | 0.098 | -0.05 |
| VIS | 0.275 | 0.152 | 0.042 | -0.636 | 0.818 | 0.049 | 0.83 | -0.041 | -0.378 | -0.289 | 0.225 | -0.782 | 0.193 | 0.896 | 0.904 | 1 | 0.098 | -0.05 |
| TT | 0.571 | 0.333 | 0.007 | 0.294 | 0.03 | -0.147 | 0.152 | -0.207 | -0.115 | -0.38 | 0.268 | -0.004 | -0.637 | 0.113 | 0.098 | 0.098 | 1 | -0.456 |
| HCH | -0.601 | -0.619 | -0.021 | -0.422 | 0.091 | 0.538 | -0.333 | 0.536 | 0.499 | 0.72 | -0.48 | -0.063 | 0.364 | -0.059 | -0.05 | -0.05 | -0.456 | 1 |

**Fig. 15.** The Spearman correlation coefficient heatmap showing the relationships between the 18 variables.

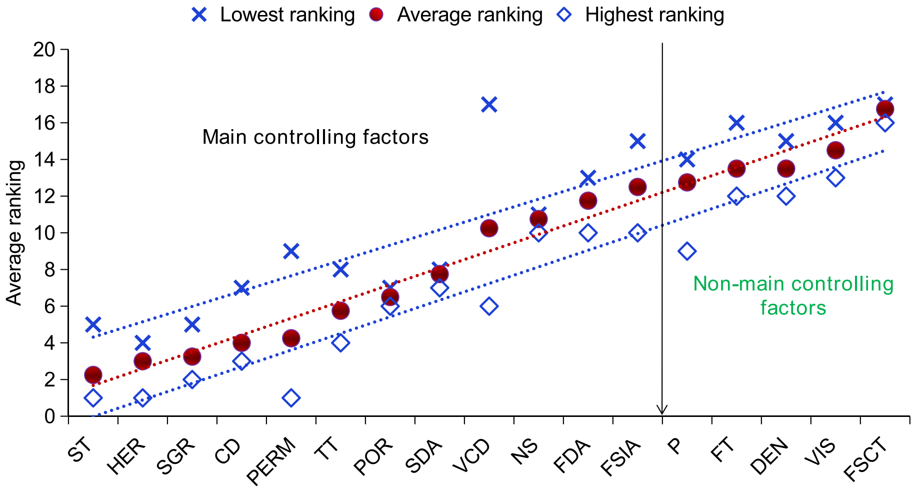


**Fig. 16.** Screening results of the main controlling factors of SHAs in the K gasfield.

**Table 4**
Principal component analysis information.

| Component | Initial eigenvalues | | | Extraction sums of squared loadings | | |
|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative/% | Total | % of Variance | Cumulative/% |
| 1 | 3.923 | 32.691 | 32.691 | 3.923 | 32.691 | 32.691 |
| 2 | 2.315 | 19.288 | 51.979 | 2.315 | 19.288 | 51.979 |
| 3 | 1.379 | 11.489 | 63.468 | 1.379 | 11.489 | 63.468 |
| 4 | 0.964 | 8.037 | 71.505 | 0.964 | 8.037 | 71.505 |
| 5 | 0.929 | 7.74 | 79.246 | 0.929 | 7.74 | 79.246 |
| 6 | 0.67 | 5.58 | 84.826 | 0.67 | 5.58 | 84.826 |
| 7 | 0.578 | 4.815 | 89.641 | | | |
| 8 | 0.424 | 3.534 | 93.175 | | | |
| 9 | 0.379 | 3.16 | 96.335 | | | |
| 10 | 0.222 | 1.853 | 98.188 | | | |
| 11 | 0.155 | 1.29 | 99.478 | | | |
| 12 | 0.063 | 0.522 | 100 | | | |

**Table 5**
The rotated component loading.

| Number | Parameter | Component loading | | | | | |
|---|---|---|---|---|---|---|---|
| | | $F_1$ | $F_2$ | $F_3$ | $F_4$ | $F_5$ | $F_6$ |
| 1 | HER | 0.804 | 0.248 | 0.229 | 0.211 | −0.006 | 0.049 |
| 2 | CD | 0.652 | 0.149 | −0.489 | 0.367 | 0.094 | 0.005 |
| 3 | VCD | 0.012 | 0.881 | −0.203 | 0.205 | 0.155 | −0.215 |
| 4 | SGR | −0.474 | −0.409 | 0.306 | 0.346 | −0.182 | −0.123 |
| 5 | NS | 0.808 | −0.398 | 0.049 | −0.008 | 0.116 | 0.235 |
| 6 | ST | −0.198 | −0.259 | 0.42 | 0.343 | 0.706 | −0.234 |
| 7 | POR | −0.746 | 0.112 | 0.173 | 0.28 | −0.01 | 0.318 |
| 8 | PERM | −0.67 | 0.217 | −0.131 | 0.119 | 0.19 | 0.555 |
| 9 | SDA | 0.723 | 0.179 | 0.153 | −0.187 | 0.361 | 0.26 |
| 10 | FDA | −0.454 | 0.773 | −0.076 | −0.156 | 0.121 | −0.109 |
| 11 | FSIA | −0.075 | −0.443 | −0.744 | 0.365 | 0.044 | −0.015 |
| 12 | TT | 0.441 | 0.422 | 0.379 | 0.47 | −0.41 | 0.081 |

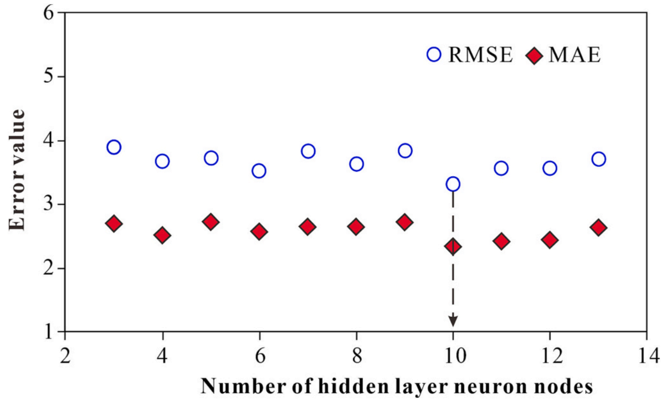


**Fig. 17.** Comparison of the error values of BPNN models with different neuron nodes.

### 4.3. Model construction results and performance

#### 4.3.1. BPNN model

In this study, 313 sets of geological data from the K gasfield were utilized to characterize the geological parameters associated with S–F–$S_d$. The six principal components obtained after dimensionality reduction of the main controlling geological factors were set as the input variables, and the HCH was selected as the output for the SHA prediction model. When using the BP algorithm for model construction, the quantified geological data were randomly divided into training, testing and validation samples according to the ratio of 70%:15%:15%. The maximum training times, minimum gradient and learning rate of the BPNN model were set at 1000, $1 \times 10^{-6}$ and 0.001, respectively. Fig. 17 shows the mean absolute error (MAE) and root mean square error (RMSE) of the BPNN model with different neuron nodes in the hidden layer. The MAE and RMSE values can be calculated by Eqs. (21) and (22), respectively. The results suggest that the MAE and RMSE values of the model were both minimum when the neuron node of the unique hidden layer was 10. Thus, the BPNN model with a 6 × 10 × 1 structure was established to predict the HCH captured in shallow fault-bounded traps by inputting a series of related geological variables.

#### 4.3.2. Bagged-NNE model

By using the bagging ensemble algorithm, the Bagged-NNE model was constructed by integrating 10 BPNN components with a 6 × 10 × 1 structure. Training, testing, and validation samples were divided according to the ratio of 70%:15%:15%. For the training stage, different training subsets were obtained through the bootstrapping sampling

**Table 6**
Comparison of performance evaluation indices of different neural network models.

| Metrics | Training | | | Validation | | | Testing | | |
|---|---|---|---|---|---|---|---|---|---|
| | BPNN | Bagged-NNE | Boosted-NNE | BPNN | Bagged-NNE | Boosted-NNE | BPNN | Bagged-NNE | Boosted-NNE |
| MAE | 1.937 | 1.414 | 0.939 | 1.861 | 0.785 | 1.373 | 1.521 | 1.237 | 0.906 |
| RMSE | 3.560 | 2.253 | 1.426 | 2.579 | 0.967 | 1.981 | 2.076 | 1.935 | 1.206 |
| $R^2$ | 0.9064 | 0.9365 | 0.9724 | 0.8429 | 0.9334 | 0.9714 | 0.8387 | 0.9217 | 0.9423 |

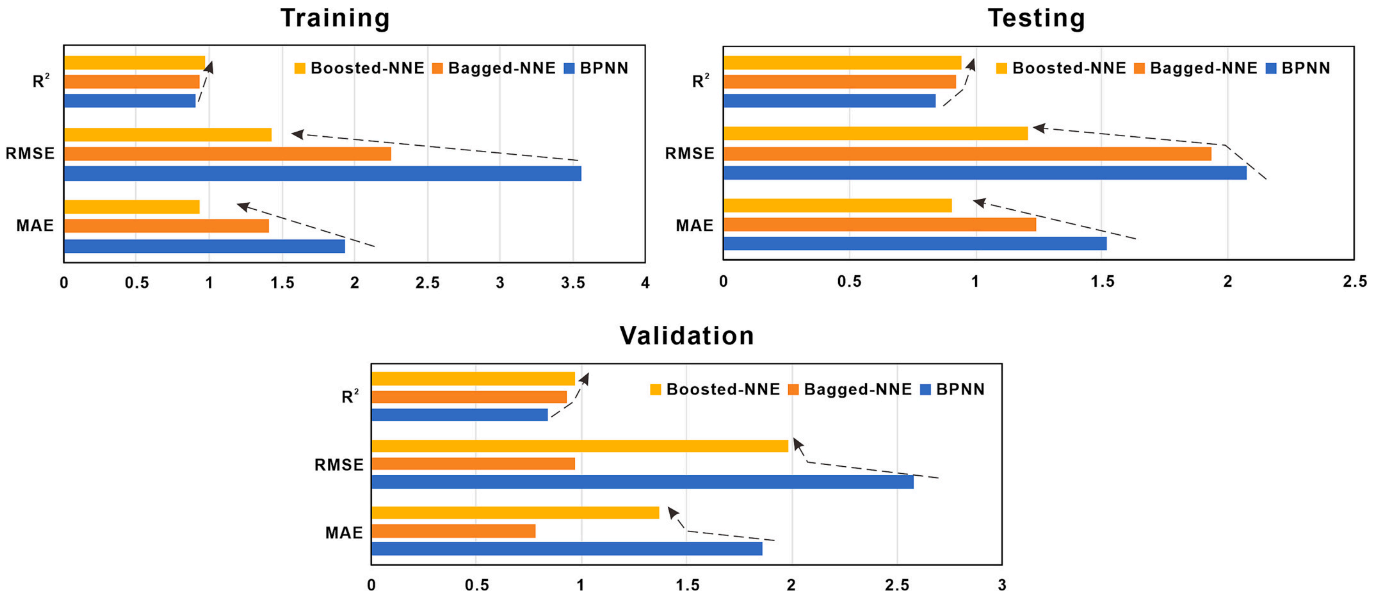


**Fig. 18.** Performance comparison of BPNN, Bagged-NNE and Boosted-NNE models in different execution stages.

technique, and the weights and thresholds in the network were updated by repeated training of each BPNN. When the model error reached the accuracy requirement or the maximum training times, the training process was ended. Since the HCH is a continuous variable, the average combination strategy was selected to integrate the prediction results of the 10 BPNN models, thereby obtaining the final comprehensive prediction output. In this way, a Bagged-NNE model with 10 BPNN models as components was built.

#### 4.3.3. *Boosted-NNE model*

For comparison with the BPNN and Boosted-NNE models, the components of the Boosted-NNE model still choose the BPNN model with a 6 × 10 × 1 structure, and the number of components is set to 10. Similarly, 70%, 15% and 15% of the datasets were randomly selected as training, testing and verification samples, respectively. The maximum training times, minimum gradient and learning rate of the model were set to 1000, $1 \times 10^{-6}$ and 0.001, respectively. After obtaining the first BPNN model by the boosting ensemble algorithm, the training set of each newly added component was adjusted according to the error signal of the previous BPNN model. When the 10 component models were generated and trained, the average combination technique was applied to integrate all the model results and obtain the final output. Consequently, a Boosted-NNE model using 10 BPNN models as components was constructed.

#### 4.3.4. *Performance of the BPNN, bagged-NNE and boosted-NNE models*

The mean absolute error (MAE), root mean square error (RMSE) and R-squared ($R^2$) were chosen as the evaluation indicators. Table 6 shows the performance of three different models in predicting SHA. For the BPNN model, the $R^2$, MAE and RMSE values in the training, testing and validation sets yielded significantly higher errors than those of the other two ensemble models, proving the limitations of a single BPNN model in SHA prediction. For the Bagged-NNE model, the measured MAE, RMSE and $R^2$ values were all superior to the BPNN during the training, testing and validation stages, verifying the improvement of the bagging ensemble algorithm on the performance of a single BPNN. On the Boosted-NNE training set, the $R^2$ was 0.9724, the MAE was 0.939 m and the RMSE was 1.426 m. The testing result was satisfied, with an $R^2$ of 0.9423, an MAE of 0.906 m and an RMSE of 1.206 m, while the validation results for the $R^2$ was 0.9714, the MAE was 1.373 m, and the RMSE was 1.981 m.

As shown in Fig. 18, the model performance after bagging and boosting ensemble has significantly improved in the training, testing and validation sets, and the Boosted-NNE model is the most excellent. In addition, by combining all the true and predicted values of each model (Figs. 19–21), the $R^2$ values of the BPNN, Bagged-NNE, and Boosted-NNE models were measured as 0.8915, 0.9333, and 0.9689, respectively. These results demonstrate that the Boosted-NNE model achieved the best performance in predicting SHA and the corresponding HCH. As shown in Fig. 22, the single-well cumulative HCH predicted by the Boosted-NNE model matches well with the actual HCH of the K gasfield in the Xihu Depression. Wells with high HCH values exhibit a wide range of planar hydrocarbon accumulations, while wells with low HCH values tend to have small-scale hydrocarbon accumulations, such as wells W13 and W14. In addition, the actual HCH of different sand layers in each single well has a good matching relationship with the predicted HCH, which confirms the accuracy and effectiveness of the Boosted-NNE model.

### 4.4. *SHA prediction result*

The best-performing Boosted-NNE model was selected to predict the SHA of the L Block in the southeast of the K gasfield (Fig. 11c), which covers 4 drilling wells. The L Block developed northeast-trending normal faults that combined with adjacent sandstone reservoirs to constitute fault-bounded traps. The LQT Fault is the main vertical conduit connecting the $E_2p_3$ source rock and the overlying shallowly buried sandstone reservoir. Previous studies have confirmed that this fault contributes to long-range vertical hydrocarbon migration [11,12]. To predict the SHA volume of sandstone reservoirs within the fault-bounded trap, the hydrocarbon expulsion rate of the $E_2p_3$ source rock at the target location was determined by numerical simulation using Petro-mod software. The source-fault contact thickness and vertical charging distance were measured by seismic interpretation. The shale gouge ratio and normal stress in the fault zone were calculated by Eq. (2) and Eq. (3). The thickness, porosity and permeability of sandstone reservoirs are characterized mainly based on logging interpretation. The fluid pressure at different burial depths was determined by the equivalent depth method based on the sonic time curve. The changes in hydrocarbon viscosity and density under different temperature and pressure conditions were obtained by PVT experiments. All the above quantitative parameters were standardized and converted into six

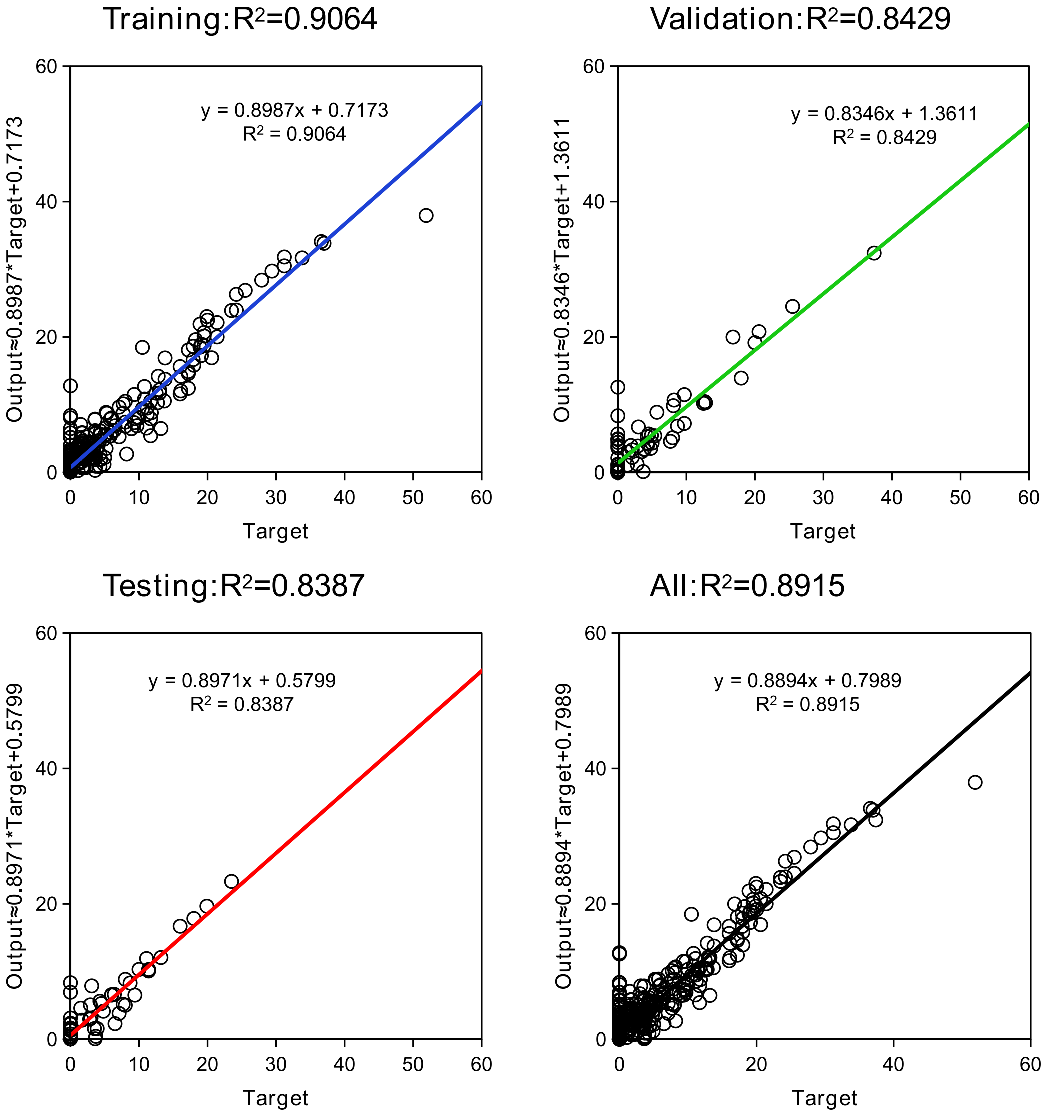


**Fig. 19.** Correlation coefficient of hydrocarbon column height using the BPNN prediction model.

principal components ($F_1$, $F_2$, $F_3$, $F_4$, $F_5$, and $F_6$), which were utilized as input variables to predict the SHA. In this way, the Boosted-NNE model can support the output of the HCH in each sandstone reservoir within the fault-bounded trap. As shown in Fig. 23, shallow hydrocarbons in the L Block tend to accumulate in the footwall due to lateral fault sealing. Consequently, the HCH predicted by the Boosted-NNE model was highly consistent with the discovered hydrocarbon accumulation in the L Block, with an $R^2$ of 0.9553 between the predicted and true values (Fig. 24). Thus, the Boosted-NNE model achieves high accuracy in practical applications, which verifies the effectiveness of the proposed method in predicting SHA.

To further verify the practicability of this method in determining predrilling hydrocarbon volumes, the HCH results obtained by the Boosted-NNE model were used to calculate geological reserves and compared with the existing evaluations. Well W1 was drilled continuously to 4865.77 m in the $E_2p_3$ member, where hydrocarbons were found in the overlying sandstone reservoirs in the $E_2p_{1+2}$ and $E_3h_1$ members. The structural trap capable of capturing hydrocarbons is a fault-nose trap with an assigned TT value of 0.66. Fluid inclusion observation and homogenization temperature analysis demonstrate that large-scale hydrocarbon charging in well W1 mainly lasted from 9.5 Ma to 3.5 Ma, and the corresponding hydrocarbon expulsion rate ranged from 0.25 to 0.60 Mtons/km^2 (Fig. 25). Fig. 26 presents the variations in the remaining 9 main geological controlling factors with burial depth. Gas is present in the entire sections of the $E_2p_{1+2}$ and $E_3h_1$ members, but the gas-bearing thickness of different sandstone reservoirs varies widely with burial depth. According to the predicted HCH results, there are nine potential exploration target intervals with rich gas. The horizontal gas-bearing area can be estimated by using the HCH and corresponding trap geometry characteristics. On this basis, hydrocarbon volumes of different sandstone reservoirs can be estimated using the volumetric method of gas reservoir geological reserves calculation (Eq. (24)) [111].

$$G = \frac{0.01^* A_g{}^* h^* \varnothing^* S_{gi}}{B_{gi}} \tag{24}$$

where $G$ is the geological reserves of gas reservoirs, $10^8$ m$^3$; $A_g$ is the gas-bearing area, km$^2$; $h$ is the effective thickness, m; $\varnothing$ is the effective porosity, %; $S_{gi}$ is the gas saturation, %; and $B_{gi}$ is the natural gas volume factor.

The results are shown in Table 7. The calculated accumulative gas geological reserve of the nine target sandstone reservoirs is 39.62 × $10^8$ m$^3$, which is highly close to the 41.24 × $10^8$ m$^3$ evaluated by the SINOPEC Shanghai Offshore Oil & Gas Company. Similarly, combined with the HCH values predicted from wells W2, W3 and W4 (Table 8), the cumulative geological reserve of the four effective traps in the L Block

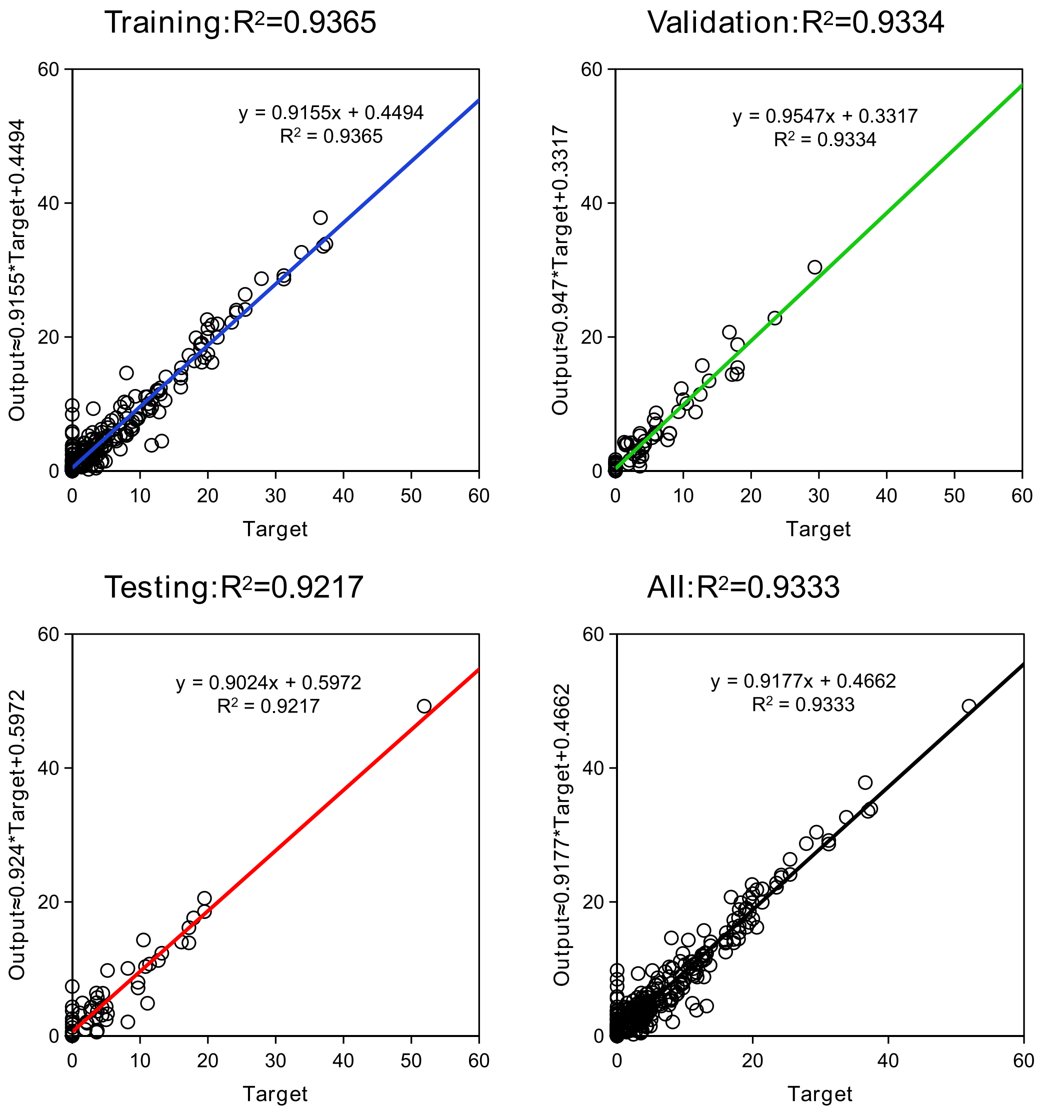


**Fig. 20.** Correlation coefficient of hydrocarbon column height using the Bagged-NNE prediction model.

was found to be 69.58 × $10^8$ $m^3$, a result that is close to the oil company's estimate of 71.66 × $10^8$ $m^3$. This indicates that the new method proposed in this paper can be successfully applied to hydrocarbon column height prediction and corresponding predrilling volume assessment.

## 5. Discussion

### 5.1. ML model comparison

In terms of SHA prediction in this study area, does the ENN algorithm (especially Boosted-NNE) outperform other well-known ML algorithms? For comparison, four well-known ML algorithms capable of regression prediction and a Linear regression (LR) model were selected to train the pre-processed data set in this investigation. Specifically, these ML algorithms include Gaussian Process Regression (GPR) (for the deeper understanding, refer to Rasmussen and Williams [107]), Classification and Regression Trees (CART) (for the deeper understanding, refer to Krzywinski and Altman [108]), K-Nearest Neighbor (KNN) (for the deeper understanding, refer to Shakhnarovich et al., [109]) and SVM (for the deeper understanding, refer to Cortes and Vapnik [110]). Meanwhile, the MAE, RMSE and $R^2$ were also selected as quantitative metrics to evaluate the performance of different ML models. The performance evaluation results of six SHA prediction models are significantly different. As shown in Fig. 27, the following sequence can be recognized: Boosted-NNE model > GPR model > CART model > KNN model > SVM model > LR model. On the whole, the Boosted-NNE model outperforms other machine learning models. Although the GPR model has the most similar MAE value to the Boosted-NNE model, it exhibits a larger MASE value and a smaller $R^2$ value, suggesting a decreasing model accuracy. In addition, although the CART model presents the smallest MAE value of all ML models, both MASE value and $R^2$ value demonstrate that its accuracy is significantly lower than Boosted-NNE and GPR models. Therefore, the evaluation system integrating multiple metrics can reveal the model performance more accurately than relying only on a single indicator, especially MAE value. From a comprehensive perspective, both the KNN and SVM models performed worse than the three ML models mentioned above, but they are significantly superior to the LR model. Thus, compared with traditional linear regression methods, ML methods that can implement nonlinear operations can achieve complex SHA predictions more accurately, among which Boosted-NNE model performs best in this investigation.

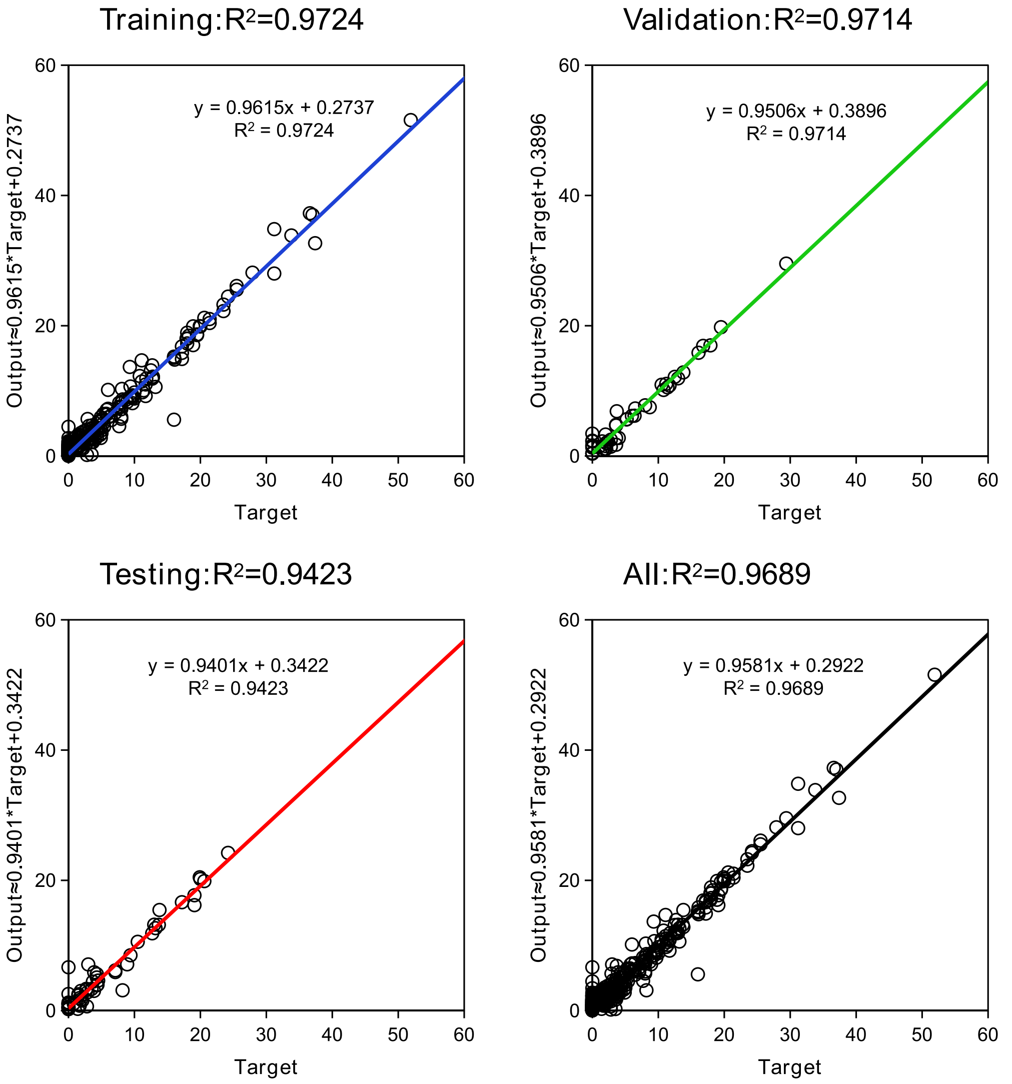


**Fig. 21.** Correlation coefficient of hydrocarbon column height using the Boosted-NNE prediction model.

## 5.2. *Variable importance on the model performance*

### 5.2.1. *Effect of principal component variables*

The importance analysis of the variables affecting the HCH is of great significance for predrilling volume estimation and future SHA exploration. In this study, the importance scores of six principal components were calculated using the feature importance method for the BPNN, Bagged-NNE and Boosted-NNE models. As shown in Fig. 28, all three models have the same 4 most important variables. $F_1$ is the most valuable principal component for predicting the accumulated hydrocarbon column heights, which primarily reflect changes in HER, NS, POR and SDA features. $F_2$ is the second most important principal component for model prediction, reflecting more variations in VCD and FDA characteristics. $F_5$ and $F_3$ are the third and fourth most important principal components contributing to the model prediction results, respectively. In the BPNN and Bagged-NNE models, the $F_6$ and $F_4$ principal components have similar and the least contributions to the predicted hydrocarbon column height. However, in the Boosted-NNE model, the importance of $F_4$ to the model prediction is higher than that of $F_6$ with the least contribution.

### 5.2.2. *Effect of single-factor control variables*

To determine the contribution of a single geological factor to the model output, a single-factor control variable analysis was performed. By keeping other influencing factors unchanged (all datasets used were the average values of the collected data), the effect of individual controlling factors on the HCH can be evaluated. In this study, the best-performing Boosted-NNE model was used to analyze the contribution of 12 main controlling factors to the predicted results, as shown in Fig. 29.

In terms of hydrocarbon supply conditions, the accumulated HCH increases with increasing hydrocarbon expulsion rate (HER) (Fig. 29a). This is consistent with the results of Pang et al. (2005) [16] and Jiang et al. (2013) [21], that is, the stronger the hydrocarbon expulsion intensity is, the better the hydrocarbon accumulation response. Nevertheless, the increased value of the HCH contributed by the HER in the prediction model was <1 m, indicating that the HER contribution to hydrocarbon accumulation is relatively limited. Comparatively, the hydrocarbon charging time (CD) presents a complex relationship with the HCH instead of a single positive or negative correlation (Fig. 29b). This may be because there are two different hydrocarbon charging mechanisms within the fault zone: short-term rapid charging driven by the “seismic pump” and long-term slow charging driven by buoyancy. When the “seismic pump” mechanism is dominant, the short-term increase in the CD contributes to the increase in the HCH value, which matches the trend of the CD value at 0–4 Ma. When the buoyancy-driven mechanism dominates, only a long-term increase in the CD value can lead to an increase in the HCH value, which is consistent with the trend

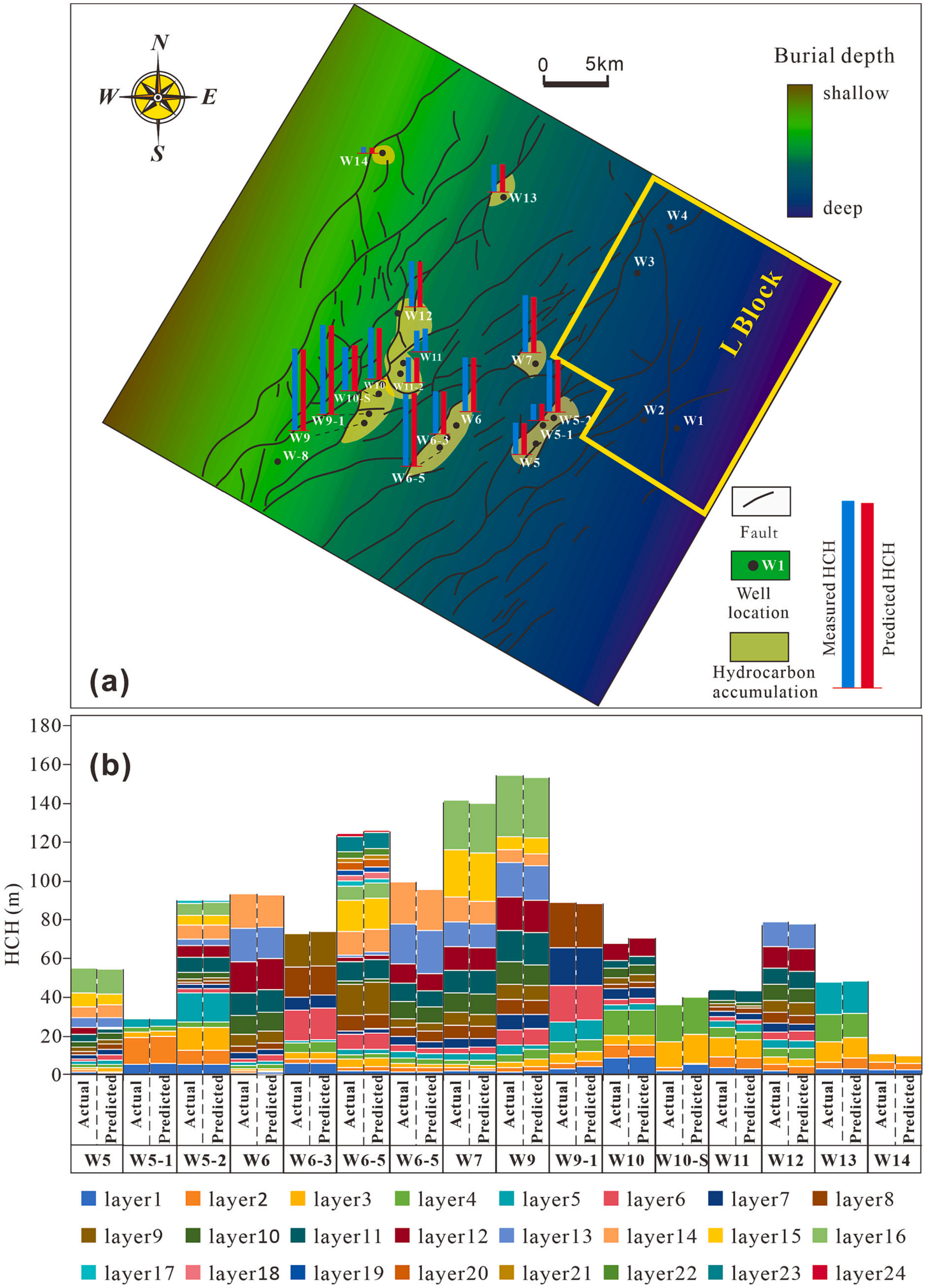


**Fig. 22.** Comparison between actual hydrocarbon accumulation and HCH predicted by the Boosted-NNE model in the K gasfield of the Xihu Depression. (a) Relationship between single-well cumulative HCH and plane hydrocarbon accumulation; and (b) actual HCH and predicted HCH of different sand layers in each single well.

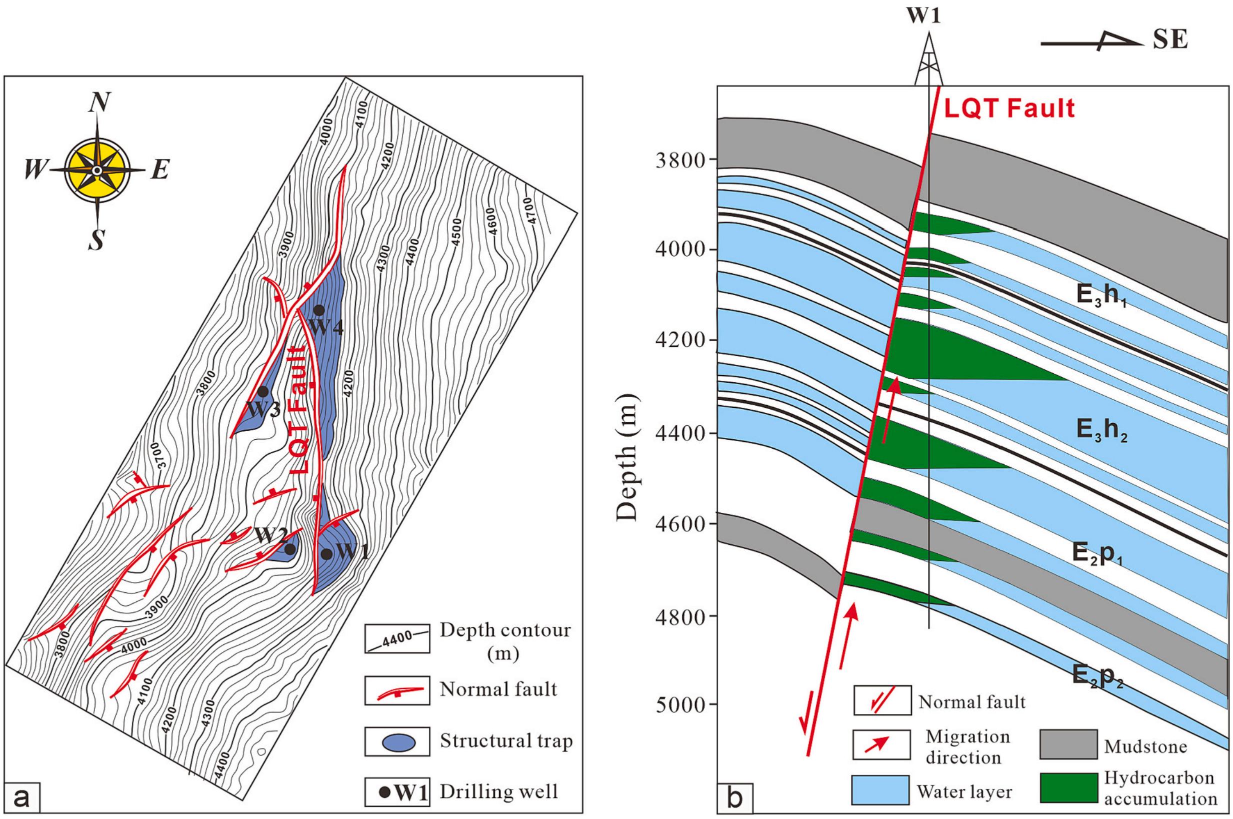


**Fig. 23.** (a) Structural contours, normal faults and trap characteristics of the Huagang Formation ($E_3h$) in the L Block. See Fig. 11c for the location. (b) Reservoir profile of well W1 in the L Block.

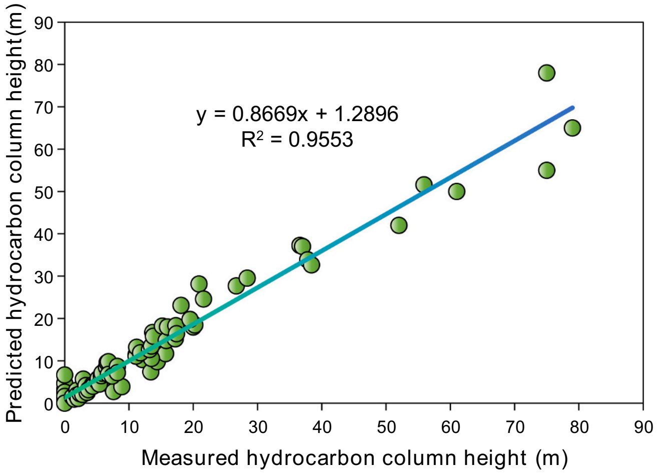


**Fig. 24.** Correlation coefficient of the measured and predicted hydrocarbon column heights in the L Block.

at 6–12 Ma. When hydrocarbon charging is in the transition period of these two mechanisms, the HCH may have a tendency to decrease because the fluid volume driven by buoyancy tends to be smaller than the fluid volume driven by the "seismic pump", such as a trend of the CD value at 4–6 Ma. Fig. 29c shows the relationship between the vertical charging distance (VCD) and HCH. When other conditions were unchanged, the VCD showed a positive correlation with the HCH in the interval 0–500 m but a negative correlation in the range 500–1500 m. Theoretically, shallow hydrocarbons tend to accumulate in the low potential energy region at the structural high position [112], which is the reason why the VCD and HCH can be positively correlated in a certain range. However, with the weakening of the hydrocarbon supply and the continuous loss of hydrocarbons during vertical migration, the accumulated HCH will gradually decrease with increasing VCD values. Therefore, it is necessary to consider whether hydrocarbon migration can be reached when analyzing long-range hydrocarbon accumulation.

With respect to fault zone features, the shale gouge ratio (SGR) is distributed in the range of 0–1, which is positively correlated with the HCH (Fig. 29d). This is consistent with the study of Yeilding et al. (1997) [24], that is, the larger the SGR value of the fault is, the larger the maximum HCH that can be sealed. Fig. 29e presents the relationship between normal stress (NS) and the HCH, exhibiting a positive correlation when NS is <120 MPa and a negative correlation when NS is larger than 120 MPa. According to reported normal stress studies [12,19,27], increasing the normal stress tends to enhance the sealing capability of the fault zone, thereby sealing more hydrocarbon column heights to prevent leakage. This is why the HCH initially showed a positive correlation with NS. Moreover, Wang et al. (2020,2021) also revealed that excessive normal stress may cause the fault zone to lose vertical transport capacity [12,19], thus making hydrocarbon vertical charging difficult. Therefore, when determining potential hydrocarbon accumulation sites, the location of NS that is too large or too small should be considered.

In terms of sandstone reservoirs, sandstone can provide seepage channels and storage space for SHA. Fig. 29f shows that the HCH value increases with increasing sandstone thickness (ST). The larger the sandstone thickness is, the more macrostorage space can be provided for large-scale SHAs. Fig. 29g presents the relationship between HCH and sandstone porosity (POR). There is a positive correlation between sandstone porosity and HCH because the larger the porosity (or throat radius) of the sandstone is, the smaller the capillary resistance [112]. As shown in Fig. 29h, there is also a positive correlation between the permeability (PERM) and HCH, but the rate of increase begins to slow down when the permeability is >500 mD. When the sandstone reservoir

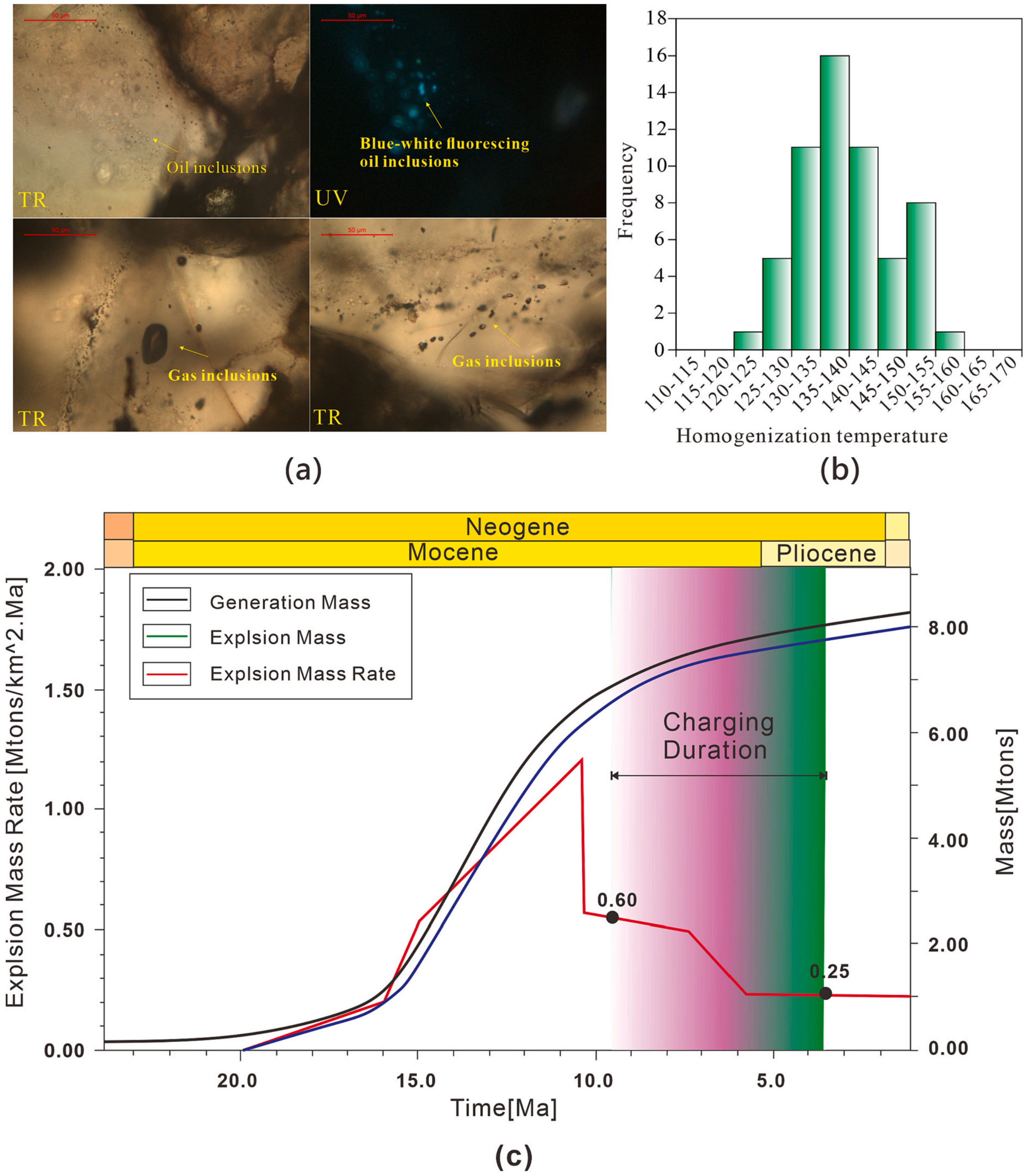


**Fig. 25.** Reservoir fluid inclusion and source rock hydrocarbon expulsion characteristics of well W1 in the L Block. (a) Oil–gas inclusion characteristics under TR plane-polarized light and UV-excited fluorescence light in the 4220–4228 m sandstone reservoir. (b) Homogenization temperature distribution of fluid inclusions in the sandstone reservoir of the Huagang formation ($E_3h$). (c) Hydrocarbon generation and expulsion characteristics of $E_2p_3$ source rock in well W1.

permeability increases from 500 mD to 2000 mD, the rock porosity only increases by <5%, which leads to a slow increase in the HCH value. Thus, when other conditions remain similar, thick sandstone reservoirs with good physical properties are often favorable sites for SHAs.

For parameters related to fault–sand intersection geometry, the relationship between the sandstone dip angle (SDA) and HCH presents an asymmetric inverse “W” shape in the SDA range of −90° to 90° (Fig. 29i). Completely vertical sandstone (SDA = -90° or 90°) has the most convenient vertical hydrocarbon migration but is not conducive to lateral hydrocarbon flow into the fault-bounded trap, so the HCH value is relatively low in this case. In addition, the contact area of the fault zone with horizontal sandstone (SDA = 0°) is often smaller than that with inclined sandstone, resulting in less hydrocarbon accumulation within the horizontal sandstone. Controlled by buoyancy, up-dip sandstone (0° < SDA < 90°) is generally superior to down-dip sandstone (−90° < SDA < 0°) in laterally diverting hydrocarbons. Therefore, the up-dip sandstone reservoir has a slightly higher maximum HCH value than that of the down-dip sandstone. Similarly, the HCH corresponding to vertical and horizontal faults is the lowest and reaches its maximum when the FDA approaches 60° (Fig. 29j). Fig. 29k shows that HCH is generally higher when the fault–sand intersection angle (FSIA) is in the range of 100° -280°. This is because sandstone reservoirs and faults (the

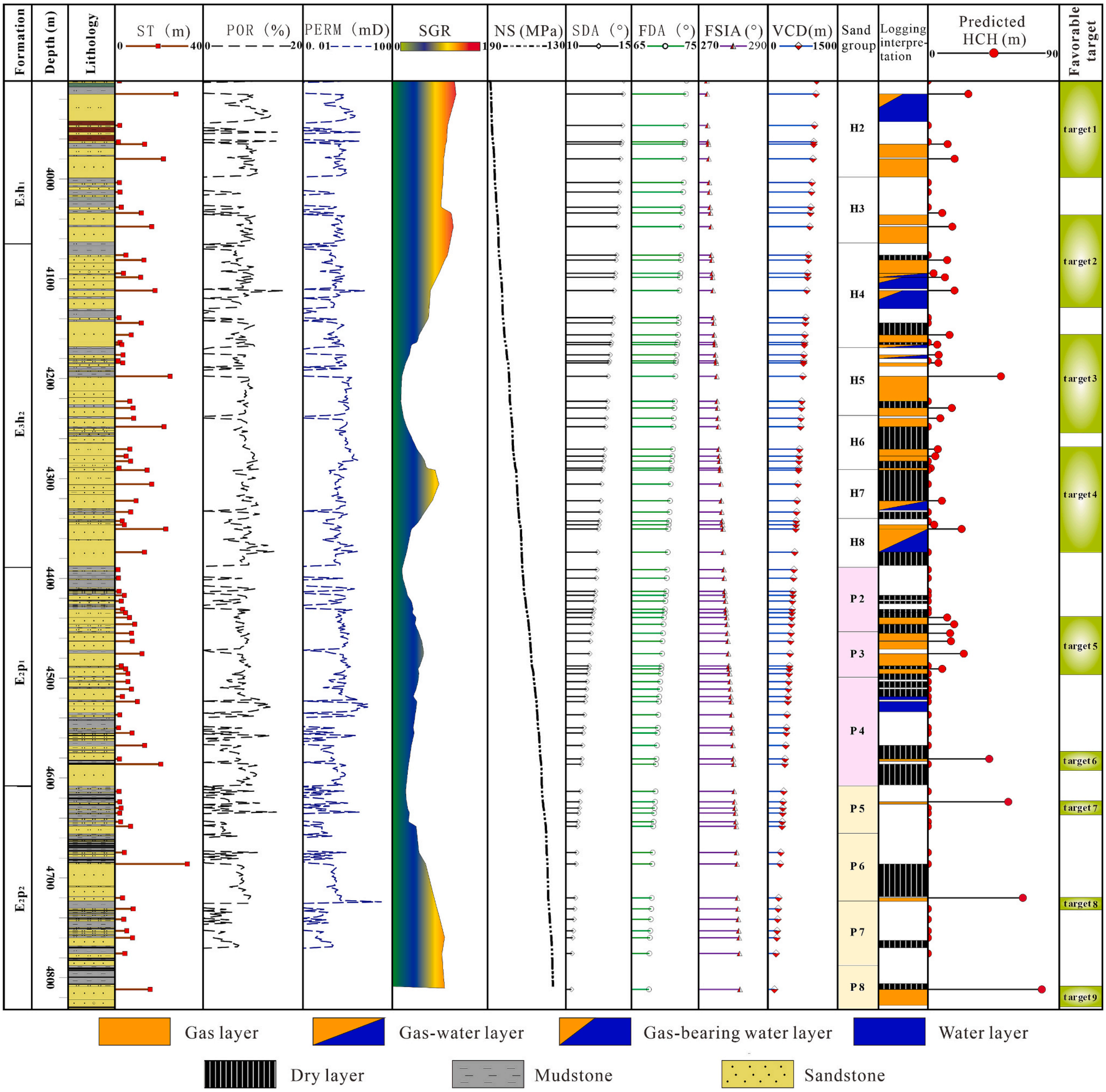


**Fig. 26.** Prediction of favorable hydrocarbon accumulation targets of the Huagang formation ($E_3h$) and Pinghu formation ($E_2p$) of well W1 in the L Block. Key: ST = sandstone thickness; POR = sandstone porosity; PERM = sandstone permeability; SGR = shale gouge ratio; NS = normal stress to fault plane; SDA = sandstone dip angle; FDA = fault dip angle; FSIA = fault–sand intersection angle; VCD = vertical charging distance; and HCH = hydrocarbon column height.

FDA is mostly close to 70°) in this study area are more likely to form fault-sealed effective traps (e.g., fault-anticlines, fault-noses, fault-blocks) under this configuration. Therefore, the optimal fault–sand spatial configuration for hydrocarbon accumulation in different basins needs to be determined according to the actual fault–sand geometric features.

Unlike other continuous variables, fault-bounded traps were assigned four discrete TT values according to their classification. When the other geological conditions remained constant, the HCH of the site without effective traps (TT = 0) was zero (Fig. 29l), while the sandstone pinch-out (TT = 0.23), fault-nose (TT = 0.66), fault-block (TT = 0.84) and fault-anticlinal (TT = 1) traps contributed 3.1 m, 9.8 m, 5.4 m and 3.5 m increases in the HCH value, respectively. These HCH predictions based on the TT variation are not completely consistent with the statistical characteristics of the hydrocarbon-bearing thickness in Table 1. This can be attributed to the fact that the HCH in a fault-bounded trap is usually greater than the hydrocarbon-bearing thickness due to the inclination of the formation. Thus, according to the HCH values, the fault-nose and fault-block traps are the two most influential trap categories for the change in SHA volume in this study area, followed by the fault-anticline and sandstone pinch-out traps.

### 5.3. *Uncertainty analysis and recommendations*

#### 5.3.1. *Uncertainty analysis*

This study proposes a new method to quantitatively predict SHAs and their corresponding HCHs, which can help reduce the risk of offshore hydrocarbon exploration. Nevertheless, quantitative evaluation of SHAs remains an inexact science, and many potential uncertainties exist in the process of model construction and geological

**Table 7**
Favorable target geological reserve calculation based on the predicted hydrocarbon column height in well W1.

| Evaluation unit | *HCH* | $A_g$ | *h* | ∅ | $S_{gi}$ | $B_{gi}$ | *G* |
|---|---|---|---|---|---|---|---|
| target1 | 70.60 | 3.50 | 14.80 | 9.40 | 42.20 | 0.00391 | 5.25 |
| target2 | 73.50 | 2.70 | 11.90 | 9.30 | 40.40 | 0.00391 | 3.09 |
| target3 | 102.20 | 4.07 | 21.10 | 9.20 | 44.50 | 0.00383 | 9.18 |
| target4 | 59.30 | 2.54 | 12.80 | 9.60 | 47.20 | 0.00383 | 3.85 |
| target5 | 96.60 | 2.76 | 24.80 | 9.10 | 49.20 | 0.00382 | 8.02 |
| target6 | 42.00 | 2.54 | 2.10 | 9.00 | 41.90 | 0.00382 | 0.53 |
| target7 | 64.00 | 3.99 | 2.20 | 9.60 | 40.30 | 0.00302 | 1.13 |
| target8 | 70.00 | 4.60 | 3.50 | 9.10 | 40.50 | 0.00302 | 1.97 |
| target9 | 83.00 | 4.15 | 13.10 | 9.10 | 40.30 | 0.00302 | 6.60 |

Note: HCH = hydrocarbon column height, m; $A_g$= gas-bearing area, km$^2$; h= effective thickness, m; ∅= effective porosity, %; $S_{gi}$= gas saturation, %; $B_{gi}$= natural gas volume factor; and G= geological reserves of gas reservoirs, $10^8$ m$^3$.

**Table 8**
Predicted and actual cumulative geological reserves of four wells in the L Block of the K gasfield.

| Well | Number of target sandstone reservoirs | Actual accumulative *HCH* | Predicted accumulative *HCH* | Actual *G* | Predicted *G* |
|---|---|---|---|---|---|
| W1 | 9 | 693.00 | 661.20 | 41.24 | 39.62 |
| W2 | 3 | 20.20 | 21.30 | 2.44 | 2.58 |
| W3 | 1 | 59.70 | 58.63 | 2.43 | 2.23 |
| W4 | 5 | 224.89 | 228.04 | 25.55 | 25.15 |

Note: HCH = hydrocarbon column height, m; and G= geological reserves of gas reservoirs, $10^8$ m$^3$.

characterization. Currently, our model can realize the two-dimensional prediction of the HCH for sandstone reservoirs in a single well. However, an unclear three-dimensional understanding of sandstone reservoirs, fault zones and fault-bounded traps may lead to biased judgments about hydrocarbon volumes and favorable targets. Although twelve major geological factors were identified as model inputs in this contribution, we believe that different geological backgrounds have different major controlling factors for SHA. In addition, uncertainties arising from the quantification of geological parameters should be considered when forecasting the SHA. The accuracy of the geological interpretation models directly affects the understanding degree of geological frameworks, such as the spatial variation of fault zones, sandstone reservoirs and traps. In the process of source rock hydrocarbon expulsion analysis, the initial parameter setting and calibration in the numerical simulation model will have a great impact on the evaluation results. Moreover, selecting the most appropriate quantitative method is very important in geological characterization because there are often multiple approaches available for a geological feature. With fluid pressure prediction as an example, quantitative work can be achieved based on either disequilibrium compaction or hydrocarbon generation pressurization mechanisms, but the most suitable method needs to be selected according to the actual geological background. Thus, all of these uncertainties should be accounted for when quantitatively predicting SHA and conducting favorable target screening.

### 5.3.2. *Future recommendation*

With the vigorous development of technology, digital and intelligent oil fields can be an inevitable stage in hydrocarbon exploration.

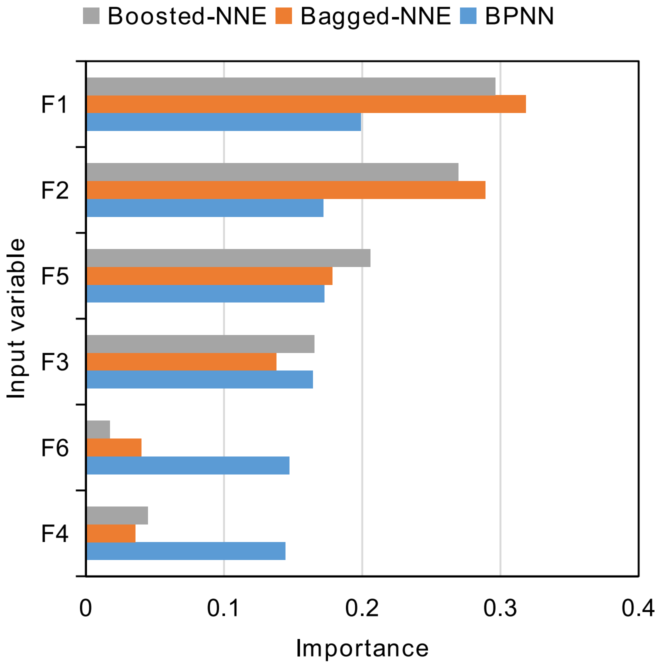


**Fig. 28.** Importance of principal component variables.

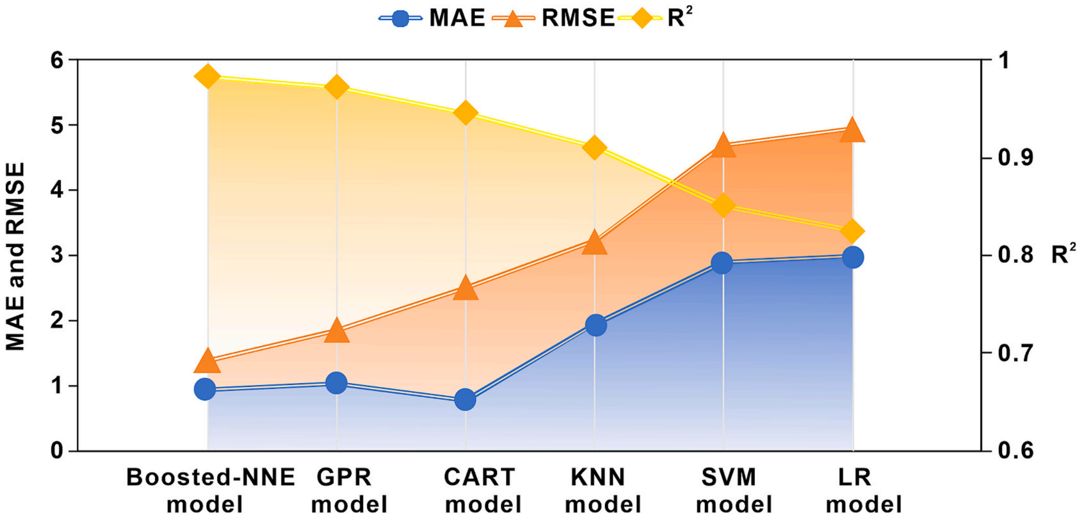


**Fig. 27.** Comparison of the performance of the Boosted-NNE model with the four well-known ML models and one LR model in predicting SHA.

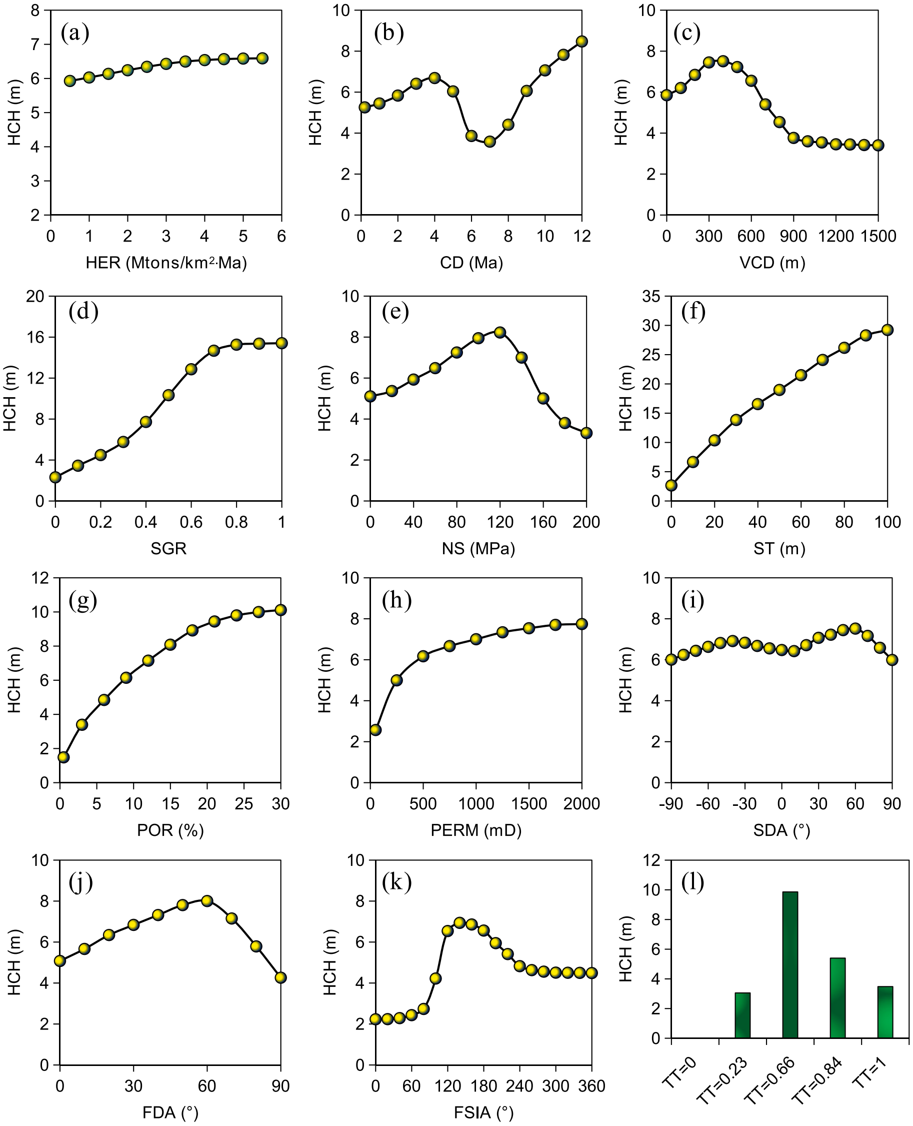


**Fig. 29.** Effect of different single factors on the hydrocarbon column height.

Although this study established an intelligent prediction model based on ensemble neural networks, we still recommend the use of more machine learning algorithms in future petroleum-related studies. This is because each machine learning algorithm has its own applicability, strengths, and weaknesses in regard to analyzing complex geological problems. Combining advanced machine learning algorithms with systematic geological evaluation can greatly improve the understanding of hydrocarbon resource evaluation and prediction. In addition, geological evaluation should be as consistent as possible with the actual conditions. As mentioned previously, there are many uncertainties in quantifying the geological factors associated with hydrocarbon accumulation. Once a large amount of unreal data is applied to model training and prediction, deviation between the model output and the actual understanding of petroleum geology tends to appear. Thus, improving the quantitative characterization of the three-dimensional features of geological bodies is of great significance to enhance model accuracy. To achieve future basin-scale applicability, the size of the original geological dataset needs to be appropriately increased because only the amount of data that reaches a certain threshold can guarantee the stability of the machine learning model. The more spatial range and geological factors covered by the original dataset, the more consistent the trained model rules are with the actual geological laws. Therefore, to improve the performance and applicability of prediction models, more diverse machine learning algorithms, more accurate quantitative evaluation, and more abundant geological datasets are the direction of our next efforts.

## 6. Conclusion

This paper proposes a new method combining the S–F–$S_d$ evaluation and ENN algorithm to realize shallow hydrocarbon accumulation prediction, which provides guidance for predrilling volume estimation and ideas for intelligent exploration technology. The conclusions can be summarized as follows:

(1) Twelve key geological parameters related to source–fault sand (source rocks, fault zones and sandstone reservoirs) are the main controlling factors determining shallow hydrocarbon accumulation. In addition, the six principal components obtained by dimensionally reducing the main controlling factors can effectively characterize the supply, migration, storage conditions and dynamic accumulation process of shallow hydrocarbons.

(2) The performance of models constructed based on different neural network algorithms varies greatly in the training, testing and verification stages. According to the MAE, RMSE and R-squared

($R^2$) values, the Boosted-NNE model is superior to the Bagged-NNE and BPNN models in characterizing the hydrocarbon column height and is preferentially recommended for predicting shallow hydrocarbon accumulation.

(3) The hydrocarbon column height predicted by the Boosted-NNE model is highly consistent with the discovered hydrocarbon accumulation in the fault-bounded traps ($R^2 = 0.9553$), which verifies the applicability of the Boosted-NNE model. Moreover, the geological reserves of the target sandstone reservoir calculated based on the forecast results are close to the existing evaluation, which can be a good complement to the current predrilling volume estimation.

(4) In terms of variable importance, $F_1$, $F_2$, $F_5$ and $F_3$ are the four top principal components contributing to the model output. For single-factor control, the hydrocarbon expulsion rate, shale gouge ratio, sandstone thickness, porosity and permeability are positively correlated with the accumulated hydrocarbon column height, while the remaining 7 main controlling factors show a more complex correlation with hydrocarbon column height rather than a purely positive or negative correlation.

Since the accuracy of the SHA prediction predominantly relies on the geological input to the model, it is strongly recommended to reduce the uncertainty caused by the quantification and screening of controlling geological factors. Furthermore, more diverse machine learning algorithms and richer geological datasets are highly recommended to improve hydrocarbon resource evaluation and advance intelligent oil–gas field technologies.

## CRediT authorship contribution statement

**Fuwei Wang:** Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Dongxia Chen:** Writing – review & editing, Supervision, Conceptualization. **Meijun Li:** Supervision, Conceptualization. **Zhangxin Chen:** Supervision, Conceptualization. **Qiaochu Wang:** Visualization, Formal analysis. **Mengya Jiang:** Software, Investigation. **Lanxi Rong:** Investigation, Formal analysis. **Yuqi Wang:** Software, Investigation. **Sha Li:** Software, Investigation. **Khawaja Hasnain Iltaf:** Investigation. **Renzeng Wanma:** Investigation. **Chen Liu:** Investigation.

## 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.

## Data availability

Data will be made available on request.

## Acknowledgments

This work was financially supported by the Science Foundation of China University of Petroleum, Beijing (No. 2462023XKBH011), the National Natural Science Foundation of China (Grant No. 41972124), the 2021 AAPG Foundation Grants-in-Aid Program (William E. & Jean Crain Named Grant), and the 2022 AAPG Foundation Grants-in-Aid Program (Grants-in-Aid General Fund Grant). We appreciate the support from the Research Institute of CNOOC Shanghai Branch and SINOPEC Shanghai Offshore Oil & Gas Company for providing the data used in this study. Sincere gratitude is extended to the handling editor, Prof. Zita Vale, for her diligent efforts in improving the quality of our manuscripts. Two anonymous reviewers provided constructive comments and are gratefully acknowledged.

## References

[1] Kvenvolden KA. Methane hydrate—a major reservoir of carbon in the shallow geosphere? Chem Geol 1988;71(1–3):41–51.

[2] Arthur MA, Cole DR. Unconventional hydrocarbon resources: prospects and problems. Elements 2014;10(4):257–64.

[3] Zou C, Zhai G, Zhang G, Wang H, Zhang G, Li J, et al. Formation, distribution, potential and prediction of global conventional and unconventional hydrocarbon resources. Petrol Explor Dev 2015;42(1):14–28.

[4] Wang H, Ma F, Tong X, Liu Z, Zhang X, Wu Z, et al. Assessment of global unconventional oil and gas resources. Petrol Explor Dev 2016;43(6):925–40.

[5] Zheng M, Li J, Wu X, Wang S, Guo Q, Yu J, et al. China's conventional and unconventional natural gas resources: potential and exploration targets. J Nat Gas Geosci 2023;3(6):295–309.

[6] Zou C, Tao S, Hou L, Zhu R, Yuan X, Zhang G, et al. Unconventional petroleum geology. 2nd ed. Beijing: Geological Publishing House; 2013.

[7] Christopher JS, Tracey JM, Cheryl AW, Geoffrey SE, Thomas MF, Phuong AL, et al. Assessment of undiscovered conventional oil and gas resources of China. U. S. Geological Survey; 2020.

[8] Deng Y, Xue Y, Yu S, Liu C. Shallow hydrocarbon migration and accumulation theory and discovery of giant oilfield group in Bohai Sea. Acta Geol Sin 2017;38 (01):1–8.

[9] Deng Y, Xu J, Sun L, Cao J, Xu L, Yu X, et al. National Science and technology major project supports CNOOC for increase of its offshore oil reserves and production. Petrol Sci Technol Forum 2021;40(3):56.

[10] Xu C, Peng J, Wu Q, Sun Z, Ye T. Vertical dominant migration channel and hydrocarbon migration in complex fault zone, Bohai Bay sag. China Petrol Explor Dev 2019;46(4):720–8.

[11] Wang F, Chen D, Wang Q, Du W, Chang S, Wang C, et al. Quantitative evaluation of caprock sealing controlled by fault activity and hydrocarbon accumulation response: K gasfield in the Xihu depression, East China Sea basin. Mar Petrol Geol 2021;134:105352.

[12] Wang F, Chen D, Du W, Zeng J, Wang Q, Tian Z, et al. Improved method for quantitative evaluation of fault vertical sealing: a case study from the eastern Pinghu Slope Belt of the Xihu depression, East China Sea Shelf Basin. Mar Petrol Geol 2021;132:105224.

[13] Chu R, Yan D, Qiu L, Wang H, Wang Q. Quantitative constraints on hydrocarbon vertical leakage: insights from underfilled fault-bound traps in the Bohai Bay basin. China Mar Petrol Geol 2023;149:106078.

[14] Tissot BP, Welte DH. Petroleum formation and occurrence. New York: Springer-Verla; 1984. p. 699.

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

[16] Pang X, Li M, Li S, Jin Z. Geochemistry of petroleum systems in the Niuzhuang south slope of Bohai Bay basin: part 3. Estimating hydrocarbon expulsion from the Shahejie formation. Org Geochem 2005;36(4):497–510.

[17] Zheng D, Pang X, Ma X, Li C, Zheng T, Zhou L. Hydrocarbon generation and expulsion characteristics of the source rocks in the third member of the upper Triassic Xujiahe formation and its effect on conventional and unconventional hydrocarbon resource potential in the Sichuan Basin. Mar Petrol Geol 2019;109: 175–92.

[18] Sibson RH, Moore JMM, Rankin AH. Seismic pumping—a hydrothermal fluid transport mechanism. J Geol Soc London 1975;131(6):653–9.

[19] Wang F, Chen D, Wang Q, Shi X, Xie G, Wang Z, et al. Evolution characteristics of transtensional faults and their impacts on hydrocarbon migration and accumulation: a case study from the Huimin depression, Bohai Bay basin, eastern China. Mar Petrol Geol 2020;120:104507.

[20] Wang F, Chen D, Wang Q, Shi X, Jiang M, Du W, et al. Quantitative evaluation of sandstone carrier transport properties and their effects on hydrocarbon migration and accumulation: a case study of the Es32 in the southern slope of Dongying depression, Bohai Bay basin. Mar Petrol Geol 2021;126:104937.

[21] Jiang F, Pang X, Guo J. Quantitative analysis model and application of the hydrocarbon distribution threshold. Acta Geol Sin English Ed 2013;87(1):232–42.

[22] Allan US. Model for hydrocarbon migration and entrapment within faulted structures. AAPG Bull 1989;73(7):803–11.

[23] Knipe RJ. Juxtaposition and seal diagrams to help analyze fault seals in hydrocarbon reservoirs. AAPG Bull 1997;81(2):187–95.

[24] Yielding G, Freeman B, Needham DT. Quantitative fault seal prediction. AAPG Bull 1997;81(6):897–917.

[25] Jaeger JC, Cook G. Fundamentals of rock mechanics. London: Chapman and Hall; 1979. p. 593.

[26] Lü Y, Li G, Wang Y, Song G. Quantitative analyses in fault sealing properties. Acta Petrol Sin 1996;17(3):39–45.

[27] Fu G, Li Y, Zhang Y, Nie H. Methods of research on fault-vertical-sealed oil & gas and its applications. Nat Gas Industry 1997;17(6):22–5.

[28] Bekele EB, Person MA, Rostron BJ, Barnes R. Modeling secondary oil migration with core-scale data: Viking formation. Alberta Basin AAPG Bull 2002;86(1): 55–74.

[29] King PR. The connectivity and conductivity of overlapping sand bodies. London: Graham & Trotman; 1990. p. 353–62.

[30] Jackson MD, Yoshida S, Muggeridge AH, Johnson HD. Three-dimensional reservoir characterization and flow simulation of heterolithic tidal sandstones. AAPG Bull 2005;89(4):507–28.

[31] Lei Y, Luo X, Song G, Zhang L, Hao X, Yang W, et al. Quantitative characterization of connectivity and conductivity of sandstone carriers during secondary

petroleum migration, applied to the third member of Eocene Shahejie formation, Dongying depression. East China Mar Petrol Geol 2014;51:268–85.

[32] Liang F. The research on shale gas enrichment pattern and the favorable area optimizing of Wufeng-longmaxi shale in middle and upper Yangtze region. Beijing, China: China University of Mining and Technology; 2018.

[33] Lewandowska-Śmierzchalska J, Tarkowski R, Uliasz-Misiak B. Screening and ranking framework for underground hydrogen storage site selection in Poland. Int J Hydrogen Energy 2018;43(9):4401–14.

[34] Liu G, Hu S, Zhao W. Oil resource abundance of petroleum plays in Chinese basins and its prediction model. Petrol Explor Dev 2006;33(6):759–61.

[35] Wang W, Pang X, Chen Z, Chen D, Zheng T, Luo B, et al. Quantitative prediction of oil and gas prospects of the Sinian-lower Paleozoic in the Sichuan Basin in Central China. Energy 2019;174:861–72.

[36] Ma K, Pang X, Pang H, Lv C, Gao T, Chen J, et al. A novel method for favorable zone prediction of conventional hydrocarbon accumulations based on RUSBoosted tree machine learning algorithm. Appl Energy 2022;326:119983.

[37] Alqahtani H, Kumar G. Machine learning for enhancing transportation security: a comprehensive analysis of electric and flying vehicle systems. Eng Appl Artif Intel 2024;129:107667.

[38] Zlobina K, Jafari M, Rolandi M, Gomez M. The role of machine learning in advancing precision medicine with feedback control. Cell Rep Phys Sci 2022;3 (11):101149.

[39] Sarwar S, Aziz G, Tiwari AK. Implication of machine learning techniques to forecast the electricity price and carbon emission: evidence from a hot region. Geoscience Frontiers 2023:101647. https://doi.org/10.1016/j.gsf.2023.101647.

[40] Kolachalama VB. Machine learning and pre-medical education. Artif Intell Med 2022;129:102313.

[41] Zhou W, Li X, Qi ZL, Zhao HH, Yi J. A shale gas production prediction model based on masked convolutional neural network. Appl Energy 2024;353(Part A): 122092.

[42] Liu Y, Zeng J, Qiao J, Yang G, Liu S, Cao W. An advanced prediction model of shale oil production profile based on source-reservoir assemblages and artificial neural networks. Appl Energy 2023;333:120604.

[43] Ren X, Hou J, Song S, Liu Y, Chen D, Wang X, et al. Lithology identification using well logs: a method by integrating artificial neural networks and sedimentary patterns. J Petrol Sci Eng 2019;182:106336.

[44] Kamenski A, Cvetković M, Močilac IK, Saftić B. Lithology prediction in the subsurface by artificial neural networks on well and 3D seismic data in clastic sediments: a stochastic approach to a deterministic method. Int J Geomathemat 2020;11:8.

[45] Liu W, Chen Z, Hu Y, Xu L. A systematic machine learning method for reservoir identification and production prediction. Petrol Sci 2023;20(1):295–308.

[46] Wang Q, Chen D, Li M, Wang F, Wang Y, Du W, et al. Application of machine learning for evaluating and predicting fault seals: a case study in the Huimin depression, Bohai Bay basin. East China Geoenergy Sci Eng 2023;228:212064.

[47] Otchere DA, Ganat AA, Gholami R, Ridha S. Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: comparative analysis of ann and svm models. J Petrol Sci Eng 2021;200:108182.

[48] Gholami R, Moradzadeh A, Maleki S, Amiri S, Hanachi J. Applications of artificial intelligence methods in prediction of permeability in hydrocarbon reservoirs. J Petrol Sci Eng 2014;122:643–56.

[49] Hosseini M, Riahi MA, Mohebian R. A Meta attribute for reservoir permeability classification using well logs and 3D seismic data with probabilistic neural network. Bollettino di Geofisica Teorica ed Applicata 2019;60(1):81–96.

[50] Ahmadi MA, Chen Z. Comparison of machine learning methods for estimating permeability and porosity of oil reservoirs via petro-physical logs. Petroleum 2019;5:271–84.

[51] Li K, Zhou G, Yang Y, Li F, Jiao Z. A novel prediction method for favorable reservoir of oil field based on grey wolf optimizer and twin support vector machine. J Petrol Sci Eng 2020;189(6):106952.

[52] Ahmadi MA, Mahmoudi B. Development of robust model to estimate gas–oil interfacial tension using least square support vector machine: experimental and modeling study. J Supercritical Fluids 2016;107:122–8.

[53] Wang Q, Chen D, Li M, Li S, Wang F, Yang Z, et al. A novel method for petroleum and natural gas resource potential evaluation and prediction by support vector machines (SVM). Appl Energy 2023;351:121836.

[54] Chen H, Wang Y, Zuo M, Zhang C, Jia N, Liu X, et al. A new prediction model of CO2 diffusion coefficient in crude oil under reservoir conditions based on BP neural network. Energy 2022;239:122286.

[55] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 1943;5:115–33.

[56] Hebb DO. The organization of behavior: a neuropsychological theory. Psychology Press; 2005. p. 17–275.

[57] Karayiannis N, Venetsanopoulos AN. Artificial neural networks: Learning algorithms, performance evaluation, and applications. Springer Science & Business Media; 1992. p. 3–217.

[58] Turan NG, Mesci B, Ozgonenel O. Artificial neural network (ANN) approach for modeling Zn (II) adsorption from leachate using a new biosorbent. Chem Eng J 2011;173(1):98–105.

[59] Giri AK, Patel RK, Mahapatra SS. Artificial neural network (ANN) approach for modelling of arsenic (III) biosorption from aqueous solution by living cells of *Bacillus cereus* biomass. Chem Eng J 2011;178:15–25.

[60] Mahmoud A, Elkatatny S, Mahmoud M, Abouelresh M, Abdulraheem A, Ali A. Determination of the total organic carbon (TOC) based on conventional well logs using artificial neural network. Int J Coal Geol 2017;179:72–80.

[61] Jiang D, Chen H, Xing J, Wang Y, Wang Z, Tuo H. A new method for dynamic predicting porosity and permeability of low permeability and tight reservoir under effective overburden pressure based on BP neural network. Geoenergy Sci Eng 2023;226:211721.

[62] Liu M, Fu X, Meng L, Du X, Zhang X, Zhang Y. Prediction of CO2 storage performance in reservoirs based on optimized neural networks. Geoenergy Sci Eng 2023;222:211428.

[63] Kotsiantis S. Combining bagging, boosting, rotation forest and random subspace methods. Artific Intellig Rev 2011;35:223–40.

[64] García-Pedrajas N, García-Osorio C, Fyfe C. Nonlinear boosting projections for ensemble construction. J Mach Learn Res 2007;8:1–33.

[65] García-Pedrajas N, Maudes-Raedo J, García-Osorio C, Rodríguez-Díez JJ. Supervised subspace projections for constructing ensembles of classifiers. Inform Sci 2012;193:1–21.

[66] Breiman L. Bagging predictors Machine learning24; 1996. p. 123–40.

[67] Freund Y, Schapire RE. Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference on Machine Learning-ICML. Morgan Kaufman; 1996. p. 148–56.

[68] Schultz RA, Fossen H. Terminology for structural discontinuities. AAPG Bull 2008;92(7):853–67.

[69] Faulkner DR, Jackson CAL, Lunn RJ, Schlische RW, Shipton ZK, Wibberley CAJ, et al. A review of recent developments concerning the structure, mechanics and fluid flow properties of fault zones. J Struct Geol 2010;32(11):1557–75.

[70] Faulkner DR, Lewis AC, Rutter EH. On the internal structure and mechanics of large strike-slip fault zones: field observations of the Carboneras fault in southeastern Spain. Tectonophysics 2003;367(3–4):235–51.

[71] Pei Y, Paton D, Knipe R, Wu K. A review of fault sealing behaviour and its evaluation in siliciclastic rocks. Earth-Sci Rev 2015;150:121–38.

[72] Childs C, Manzocchi T, Walsh J, Bonson CG, Nicol A, et al. A geometric model of fault zone and fault rock thickness variations. J Struct Geol 2009;31:117–27.

[73] Otsuki K. On the relationship between the width of shear zone and the displacement along fault. Jour Geol Soc Japan 1978;84:661–9.

[74] Robertson EG. Relationship of fault displacement to gouge and breccia thickness. Am Inst Min Eng Trans 1983;274:1426–32.

[75] Pizzati M, Balsamo F, Storti F. Displacement-dependent microstructural and petrophysical properties of deformation bands and gouges in poorly lithified sandstone deformed at shallow burial depth (Crotone Basin, Italy). J Struct Geol 2020;137:104069.

[76] Edmundson IS, Davies R, Frette LU, Mackie S, Kavli EA, Rotevatn A, et al. An empirical approach to estimating hydrocarbon column heights for improved predrill volume prediction in hydrocarbon exploration. AAPG Bull 2021;105(12): 2381–403.

[77] Jia X, An H, Fang W, Sun X, Huang X. How do correlations of crude oil prices co-move? A grey correlation-based wavelet perspective. Energy Econ 2015;49: 588–98.

[78] Jebli I, Belouadha F, Kabbaj M, Tilioua A. Prediction of solar energy guided by Pearson correlation using machine learning. Energy 2021;224:120109.

[79] Wang P, Long Z, Wang G. A hybrid prognostics approach for estimating remaining useful life of wind turbine bearings. Energy Rep 2020;6:173–82.

[80] Qian J, Wu J, Yao L, Mahmut S, Zhang Q. Comprehensive performance evaluation of wind-solar-CCHP system based on emergy analysis and multi-objective decision method. Energy 2021;230:120779.

[81] Sun Y. Using SPSS software to analyze the correlation between variables. J Xinjiang Educ Inst 2007;2:120–3.

[82] Zhang Y, Wang G, Wang X, Fan H, Shen B, Sun K. TOC estimation from logging data using principal component analysis. Energy Geosci 2023;4(4):100197.

[83] Yang H, Jin J, Hou F, He X, Hang Y. An ANN-based method for predicting Zhundong and other Chinese coal slagging potential. Fuel 2021;293:120271.

[84] Deng Y, Zhu M, Xiang D, Cheng X. An analysis for effect of cetane number on exhaust emissions from engine with the neural network. Fuel 2002;81(15): 1963–70.

[85] Mittal S, Pathak S, Dhawan H, Upadhyayula S. A machine learning approach to improve ignition properties of high-ash Indian coals by solvent extraction and coal blending. Chem Eng J 2021;413:127385.

[86] Rumelhart D, Hinton G, Williams R. Learning internal representations by error propagation. Read Cognit Sci 1988:399–421.

[87] Wu D, Zhang D, Liu S, Jin Z, Chowwanonthapunya T, Gao J, et al. Prediction of polycarbonate degradation in natural atmospheric environment of China based on BP-ANN model with screened environmental factors. Chem Eng J 2020;399: 125878.

[88] Hariharan N, Senthil V, Krishnamoorthi M, Karthic SV. Application of artificial neural network and response surface methodology for predicting and optimizing dual-fuel CI engine characteristics using hydrogen and bio fuel with water injection. Fuel 2020;270:117576.

[89] Cheng J, Wang X, Si T, Zhou F, Zhou J, Cen K. Ignition temperature and activation energy of power coal blends predicted with back-propagation neural network models. Fuel 2016;173:230–8.

[90] Zhu H, Kong D, Qian X. Shale gas production prediction method based on adaptive threshold denoising BP neural network. Sci Technol Eng 2017;17(31): 128–32.

[91] Shen H, Wang Z, Gao C. Determining the number of BP neural network hidden layer units. J Tianjin Univ Technol 2008;24(5):13.

[92] Stathakis D. How many hidden layers and nodes? Int J Remote Sens 2009;30(8): 2133–47.

[93] Souza PVC, Torres LCB, Guimaraes AJ, Araujo VS, Araujo VJS, Rezende TS. Data density-based clustering for regularized fuzzy neural networks based on nullneurons and robust activation function. Soft Comput 2019;23(23):12475–89.

[94] Warner B, Misra M. Understanding neural networks as statistical tools. Am Statistic 1996;50(4):284–93.

[95] McClelland JL, Rumelhart DE. An interactive activation model of context effects in letter perception: I. An account of basic findings. Psychol Rev 1981;88(5):375.

[96] Masnadi-Shirazi H, Vasconcelos N. Asymmetric boosting. In: Proceedings of the 24th International Conference on Machine Learning; 2007. p. 609–19.

[97] Hadavandi E, Shahrabi J, Shamshirband S. A novel Boosted-neural network ensemble for modeling multi-target regression problems. Eng Appl Artif Intel 2015;45:204–19.

[98] Nash JE, Sutcliffe JV. River flow forecasting through conceptual models part I—A discussion of principles. J Hydrol 1970;10(3):282–90.

[99] Duan M, Ye J, Wu J, Shan C, Lei C. Overpressure formation mechanism in Xihu depression of the East China Sea shelf basin. Earth Sci 2017;42(1):119–29.

[100] Zhang J, Lu Y, Krijgsman W, Liu J, Li X, Du X, et al. Source to sink transport in the Oligocene Huagang formation of the Xihu depression, East China Sea shelf basin. Mar Petrol Geol 2018;98:733–45.

[101] Ye J, Qing H, Bend SL, Gu H. Petroleum systems in the offshore Xihu Basin on the continental shelf of the East China Sea. AAPG Bull 2007;91(8):1167–88.

[102] Zhang Y, Gartrell A, Underschultz JR, Dewhurst DN. Numerical modelling of strain localisation and fluid flow during extensional fault reactivation: implications for hydrocarbon preservation. J Struct Geol 2009;31(3):315–27.

[103] Zhang Y, Ye J, Su K, Li L, Xu J, Zhang Y. The burial history and evolution of Xihu depression. Geotectonicaet Metallogenia 2009;33(2):215–23.

[104] Abbas A, Zhu H, Zeng Z, Zhou X. Sedimentary facies analysis using sequence stratigraphy and seismic sedimentology in the paleogene Pinghu formation, Xihu depression, East China Sea shelf basin. Mar Petrol Geol 2018;93:287–97.

[105] Shan C, Ye J, Cao Q, Lei C, Peng Y, Tian Y. Controlling factors for gas accumulation in Kongqueting gas field of Xihu Sag. Mar Geol Quatern Geol 2015;35(1):135–44.

[106] Su A, Chen H, Zhao J, Zhang T, Feng Y, Wang C. Natural gas washing induces condensate formation from coal measures in the Pinghu Slope Belt of the Xihu depression, East China Sea basin: insights from fluid inclusion, geochemistry, and rock gold-tube pyrolysis. Mar Petrol Geol 2020;118:104450.

[107] Rasmussen CE, Williams CKI. Gaussian processes for machine learning, 3. Printed. In: Adaptive computation and machine learning. Cambridge, Mass: MIT Press; 2008.

[108] Krzywinski M, Altman N. Classification and regression trees. Nat Methods 2017;14:757–8.

[109] Shakhnarovich G, Darrell T, Indyk P. Nearest-neighbor methods in learning and vision: theory and practice19(2). Cambridge: The MIT Press; 2006. p. 377.

[110] Cortes C, Vapnik V. Support vector networks. Machine Learn 1995;20(3):273–97.

[111] Yang T, Fan S. Calculation method of oil and natural gas reserves. Beijing: Petroleum Industry Press; 1998. p. 32–56.

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