Oil Shale, 2025, Vol. 42, No. 1, pp. 79–114
doi: https://doi.org/10.3176/oil.2025.1.04

ISSN 0208-189X
© 2025 Estonian Academy Publishers

Intelligent evaluation method for identifying favorable shale oil areas based on improved stacked sparse autoencoder

Rui Xu[a,b], Tie Yan[a]*, Shihui Sun[b], Jingyu Qu[a], Zhaokai Hou[a]

[a] Sanya Offshore Oil & Gas Research Institute, Northeast Petroleum University, Sanya 572025, China

[b] School of Petroleum Engineering, Northeast Petroleum University, Daqing 163318, China

Received 31 May 2024, accepted 10 January 2025, available online 14 January 2025

Abstract. *Identifying and evaluating favorable areas is crucial for shale oil exploration and development, well-location deployment, and fracturing design. Traditional machine learning methods struggle to accurately extract the characteristics of favorable shale oil areas with limited labeled data, affecting accuracy and generalization. This study proposes an intelligent method for identifying favorable shale oil areas under semi-supervised learning (SSAE-plus) to identify and evaluate favorable shale oil areas of the Qingshankou Formation in the Songliao Basin. The experimental results show that this method can effectively overcome the favorable area identification model's reliance on labeled data and can adaptively extract the characteristics of favorable shale oil areas without supervision. The accuracy of model identification is as high as 98.82%. Compared with other methods, the SSAE-plus yields higher accuracy and efficiency, while being more stable and generalizable. The SSAE-plus achieved over 95% accuracy in identifying favorable shale oil areas across six datasets. It has broad application prospects in identifying and evaluating favorable areas, and provides valuable theoretical insights for shale oil development and exploration well layout.*

Keywords: *shale oil, favorable area, autoencoder, batch normalization, semi-supervised learning.*

1. Introduction

Optimal selection of favorable areas is essential for efficient development of shale oil, as it reflects differences between potential reservoirs, supports well-location deployment, guides perforation cluster choices, and plays a key role in maximizing the return on capital investment [1–4]. With the continuous

* Corresponding author, sygcxytyb@163.com

development of unconventional oil and gas resources, increasing attention has been paid to the identification of favorable areas, along with higher requirements for identification accuracy.

The identification of favorable areas generally requires comprehensive research on geological, geochemical, and geophysical data combined with petroleum geological theory. Experts then qualitatively evaluate the distribution range and resource reserves of favorable areas based on experience and statistical standards [5–8]. Common methods include drawing cross-plots of favorable oil and gas accumulation conditions, superimposing attribute distribution maps of favorable areas, and conducting multiphase zone matching analysis. Chopra et al. [9] used the joint method to draw the cross-plot of logging and seismic data to identify favorable areas for oil and gas resources exploration and development. Yang et al. [10] proposed a multiphase zone matching analysis method, which superimposes and matches favorable sedimentary, structural, and reservoir facies zones, and evaluates favorable areas based on the degree of matching. Ismail et al. [11] identified favorable areas using color transformation overlay analysis and superimposition based on differences in geological attributes. Yao et al. [12] applied a dynamic uncertain causality diagram, integrating expert knowledge to evaluate the best shale gas exploration points.

However, due to the complexity of geological conditions, data uncertainty, and inconsistency in expert knowledge along with its wider application scope, the shortcomings of traditional favorable area identification methods that rely on human experiences, such as insufficient prediction reliability and low efficiency, are becoming increasingly obvious. Therefore, researchers are now exploring advanced data analysis methods for a breakthrough in favorable area identification technology.

With advancements in computer technology, the identification of favorable areas has exceeded the scope of ordinary geology, moving from qualitative evaluation to quantitative research. Researchers combine mathematical methods with expert knowledge and use computers to evaluate favorable areas quantitatively [13–16]. Guan et al. [17] used an analytic hierarchy process to establish a hierarchical structure model for evaluating and optimizing favorable shale gas areas, while Riahi et al. [18] applied fuzzy logic and TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) decision-making methods to prioritize potentially favorable areas and determine the best candidate areas for the next exploration stage. Akbar et al. [19] used K-mean clustering to evaluate favorable areas based on water saturation, brittleness index, and total organic carbon (TOC) content. Zhou et al. [20] summarized the key evaluation parameters of favorable shale gas areas and quantitatively evaluated their spatial distribution by multiple linear regression analysis. Based on high-pressure mercury injection experiments, organic geochemistry analysis, and scanning electron microscope, combined with grey correlation analysis, clustering, and the Kriging model, Niu et al. [21] proposed a new method for the multi-scale evaluation of shale reservoirs.

Data analysis technology is bound to promote the study of favorable area identification. However, predicting favorable areas is an integrated process controlled by multiple factors, and underground geological conditions are complex. Therefore, using a single quantitative research method inevitably has shortcomings. Reducing the risk of favorable area interpretation and the influence of human subjective factors to make the identification of favorable areas more timely and accurate has become an active research focus.

In recent years, the vigorous development of machine learning methods has provided new ideas for identifying favorable areas. Many scholars now apply machine learning for intelligent automated interpretation and analysis of these areas [22–26]. Tahmasebi et al. [27] combined neural networks with genetic algorithms to identify favorable and low-producing areas in shale formation. Hauge et al. [28] discussed some popular machine learning methods for identifying favorable areas, including logistic regression, *k*-nearest neighbor (KNN), support vector machine (SVM), and random forest (RF), comparing and analyzing the advantages and disadvantages of each method. Raef et al. [29] established an artificial neural network reservoir classification model based on seismic attributes such as instantaneous frequency, energy, and bandwidth, which can effectively identify favorable reservoir areas. Otchere et al. [30] applied supervised machine learning in favorable reservoir area prediction by comparing and analyzing artificial neural network and support vector machine models. Tang et al. [31] introduced a method for automatically determining favorable areas using the gradient-boosting decision tree model. Huo et al. [32] constructed a new hybrid learning model based on temporal convolutional networks and long short-term memory networks for predicting key parameters of favorable reservoir areas. Qin and Xu [33] designed a one-dimensional convolutional neural network to predict favorable shale gas areas.

Although machine learning methods have made significant achievements in identifying favorable areas, most belong to “shallow learning” and often rely on supervised training modes. Therefore, substantial expert knowledge is necessary to pre-design training labels and extract sample features. This process is time-consuming and labor-intensive, and subjective human calibration errors can increase prediction error rates. The model’s learning ability is limited, making it difficult to fully explore the deep features of data.

This paper introduces a new method for identifying favorable areas based on a stacked sparse autoencoder. This method adopts a semi-supervised learning mode, so that the model can fully learn data features when the number of labeled samples is small. Incorporating a deep learning network enhances the model’s feature expression ability, reduces computational complexity, and improves both the accuracy and efficiency of favorable area identification.

2. Methods

2.1. Autoencoder

The autoencoder (AE) [34, 35] is a three-layer feedforward neural network composed of input, hidden, and output layers, as shown in Figure 1. The AE can extract deep data features in unsupervised learning mode and reproduce the original data as much as possible by reconstructing the error function, aiding feature extraction and dimensionality reduction. The AE's learning process includes two stages: encoding and decoding. The encoding stage is used for feature extraction and transformation of input layer information, while the decoding stage is used for the reverse reconstruction of transformed features to achieve maximum restoration of the input layer information.

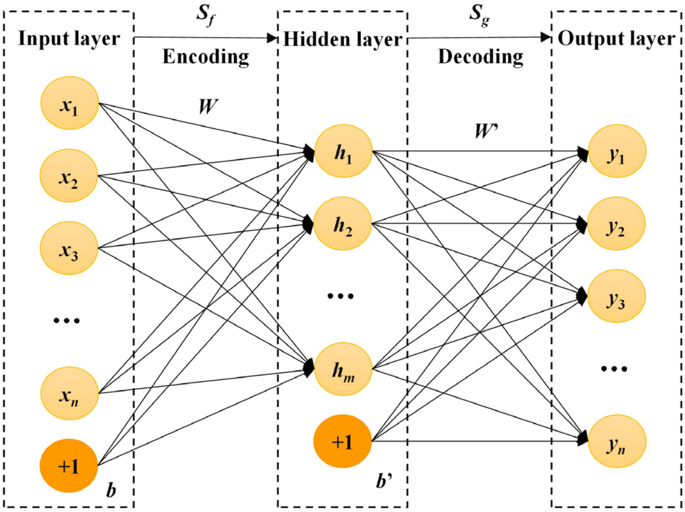


Fig. 1. Autoencoder structure.

The AE is applied to a dataset $X = [x_1, x_2, \ldots, x_n]$, containing n samples. In the encoding stage, the input layer data X is mapped to the hidden layer using the encoding function S_f, resulting in the characteristic expression $H = [h_1, h_2, \ldots, h_m]$ of the input data, as shown in formula (1):

$$H = S_f\left(WX + b\right). \tag{1}$$

In the decoding stage, the hidden layer data H is mapped to the output layer using the decoding function S_g, resulting in the reconstructed expression $Y = [y_1, y_2, \ldots, y_n]$ of the input data, as shown in formula (2):

$$Y = S_g\left(W^{'}H + b^{'}\right). \tag{2}$$

In formulas (1) and (2), S_f and S_g are nonlinear activation functions, the Sigmoid function is typically selected as the activation function, expressed as $S(x) = [1 + e(-x)]^{-1}$, W and W' are the weight matrices for encoding and decoding, respectively, and b and b' are the offset vectors in encoding and decoding, respectively.

In general, the output Y of the AE is not completely equal to the input X; rather, X is reproduced as much as possible under certain conditions. Therefore, the training goal of the AE is to minimize the reconstruction error, represented by the loss function J_{AE}. The mathematical expression for J_{AE} is as follows:

$$J_{\mathrm{AE}} = \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\left\|x_i - y_i\right\|^2\right)\right] + \frac{\lambda}{2}\sum W^2, \tag{3}$$

where x_i and y_i represent the input and output of the i-th sample from the AE, respectively, and λ is the L_2 regularization coefficient.

The first item in formula (3) is the mean squared sum of input and output sample errors, representing the reconstruction error for the entire dataset. The second item is the regularization weight penalty, which aims to prevent overfitting by restraining the weight. When the optimal parameter set $[W, b]$ is found through multiple iterations to minimize the loss function J_{AE}, the output Y of the AE is considered a reconstruction of the input X. At this point, the hidden layer output H represents the characteristic expression of the input X.

2.2. Sparse autoencoder

To make the AE fully mine feature information from original data and enhance feature expression ability, it usually uses more hidden layer nodes, but in this case, it is easy to overfit during training. To alleviate overfitting, sparsity constraints are added to the hidden layer of the AE, forcing only a few hidden layer neurons to be activated, thus forming a sparse autoencoder (SAE) [36, 37]. These sparsity constraints are implemented by controlling the average activation amount of hidden layer neurons. Assuming $h_j(x_i)$ is the output of the j-th hidden layer neuron corresponding to the i-th input sample, the average activation amount of the j-th hidden layer neuron ρ_j is given by:

$$\rho_j = \frac{1}{n}\sum_{i=1}^{n} h_j\left(x_i\right). \tag{4}$$

To limit the activation degree of neurons in the whole hidden layer, a sparsity parameter ρ is set to make the average activation amount ρ_j approach ρ, typically a small value close to 0. To ensure that ρ_j does not deviate from ρ, a sparse constraint term needs to be added to the AE's loss function J_{AE}, resulting in the SAE loss function J_{SAE}, which is expressed as follows:

$$J_{\mathrm{SAE}} = J_{\mathrm{AE}} + \beta\sum_{j=1}^{m}\mathrm{KL}\left(\rho\middle\|\rho_j\right), \tag{5}$$

$$\mathrm{KL}\left(\rho \| \rho_j\right) = \rho \ln \frac{\rho}{\rho_j} + (1-\rho) \ln \frac{1-\rho}{1-\rho_j}, \tag{6}$$

where β is the weight of the control sparse constraint term, m is the number of neurons in the hidden layer, and KL is the Kullback–Leibler divergence, which measures the difference between two probability distributions in the same event space.

The second term in formula (5) is a sparsity constraint for the hidden layer, so that ρ_j of the hidden layer approaches the set sparsity parameter ρ. To minimize the loss function J_{SAE}, the weight matrix W and the offset vector b are optimized and updated by a gradient descent algorithm, reducing the error function as follows:

$$W = W - \varepsilon \frac{\partial}{\partial W} J_{\mathrm{SEA}}, \tag{7}$$

$$b = b - \varepsilon \frac{\partial}{\partial b} J_{\mathrm{SAE}}, \tag{8}$$

where ε is the learning rate, and $\frac{\partial}{\partial b}$ represents the partial derivative for the bias parameter b.

2.3. Stacked sparse autoencoder

A single SAE can only learn the surface features of data, and its learning ability is limited. Therefore, based on single SAEs, this paper stacks several SAE networks to form the stacked sparse autoencoder (SSAE) [38, 39]. Since the ultimate goal is to classify and identify favorable areas, the SSAE's decoding layer is discarded, and a classifier for identifying favorable areas is added after the last hidden layer. Figure 2 shows the SSAE network structure, comprising three SAEs.

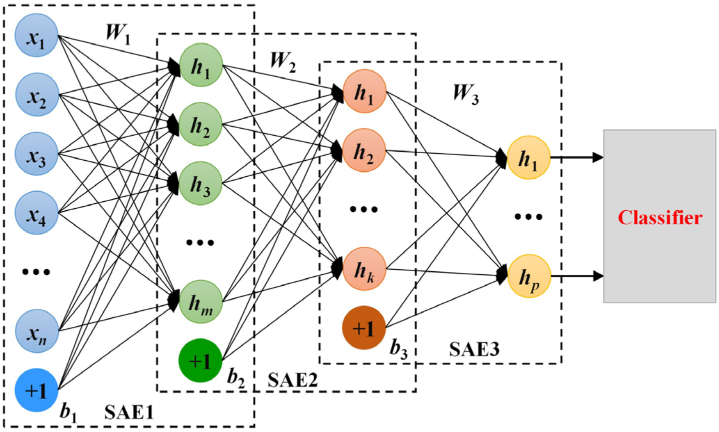


Fig. 2. Stacked sparse autoencoder structure.

Compared with the SAE, the SSAE has a deeper network structure, which allows it to learn more abstract features of original data and thus enhances its feature extraction ability. The SSAE adopts an unsupervised training method layer by layer. In this process, original data are used as input, and the hidden layer output of the previous layer of the SAE is used as the input of the next layer. After layer-by-layer training, the depth feature expression is obtained and input into the appropriate classifier for classification and identification. The whole network becomes a deep learning model with feature extraction and data classification functions. The layer-by-layer unsupervised training method considerably reduces the search of parameter space as well as avoids manually adding sample labels, improves work efficiency, and lessens reliance on subjective experience, making it more suitable for intelligent identification of favorable areas.

3. Improved stacked sparse autoencoder

While the SSAE has a strong feature expression ability, it faces problems typical of deep learning networks, such as gradient dispersion, overfitting, and slow convergence. Therefore, this study proposes improvements to the SSAE by optimizing its feature extraction process and classifier. This chapter explains the improvements in detail, introducing a batch normalization (BN) strategy, a Softmax classifier, and a semi-supervised learning mode.

3.1. Batch normalization

In training the SSAE network, updating network parameters at each layer changes data distribution. With each layer's operation, this distribution shifts further. The upper network needs constant adjustment to adapt to these evolving input data distributions. Additionally, as the network processes layer by layer, data distribution constantly approaches the upper and lower limits of the activation function, which leads to the gradual saturation of neurons and the loss of learning ability, which seriously affects convergence speed and model performance. Therefore, BN [40, 41] is introduced in this paper to solve learning gradient disappearance and slower convergence caused by input data distribution drift. The principle of the BN algorithm is shown in Figure 3.

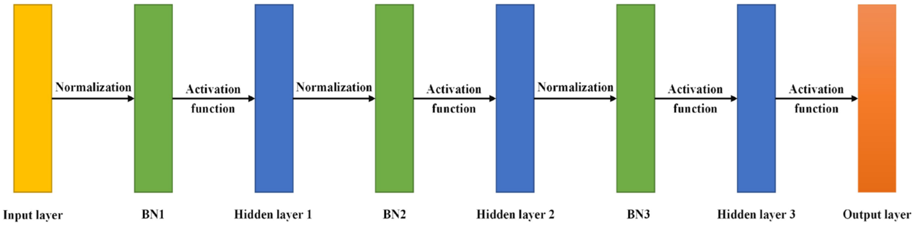


Fig. 3. Principle of batch normalization algorithm.

The BN strategy standardizes the output data of each layer of the network. The neurons are transformed (before the activation function) from an arbitrary normal distribution to a standard normal distribution with a mean value of 0 and variance of 1. The algorithm shifts data distribution away from the activation function's upper and lower limits, helping input data remain in the activation function's sensitive area, preventing gradient disappearance, and accelerating model convergence. Assuming that the input dataset of any intermediate hidden layer of the SSAE network is $h_j = [h_1, h_2, \ldots, h_m]$, the BN process is as follows:

$$\mu = \frac{1}{m}\sum_{j=1}^{m} h_j , \tag{9}$$

$$\sigma = \frac{1}{m}\sum_{j=1}^{m}\left(h_j - \mu\right)^2 , \tag{10}$$

$$h_j^{'} = \frac{h_j - \mu}{\sqrt{\sigma^2 + \varsigma}} , \tag{11}$$

where μ and σ are the mean and variance of input data, respectively, h_j' is the input data after BN, and ς is a constant set to avoid the failure of formula (11) when $\sigma = 0$.

Since BN often reduces the network's feature expression ability, parameters α and γ are introduced to offset the weakening of the network's learning ability by BN. Scaling and translation are performed based on formula (12) to restore the original feature distribution and enhance the network's expression ability:

$$y_i = \alpha h_j^{'} + \gamma . \tag{12}$$

3.2. Softmax classifier

The Softmax classifier is often used in multi-classification, mapping the outputs of multiple neurons to the [0–1] interval, and the final output result is the probability of belonging to each category. In this paper, the Softmax classifier's characteristics and the SSAE network's overall structure are comprehensively considered, and the Softmax classifier is selected as the classifier of the SSAE network for two reasons:

(1) The Softmax classifier can amplify differences between categories (the Matthew effect), making probability differences more significant and the probability of generating the maximum value closer to 1, so that the output result has better interpretability.

(2) During training, the SSAE network needs to calculate the parameter gradient based on the error and update the parameters based on the parameter gradient. Combining SSAE with the Softmax classifier simplifies parameter gradient derivation and accelerates network training.

The principle of the Softmax classifier algorithm is shown in formula (13). Assuming a_i is the output of the i-th neuron in the last hidden layer of the SSAE network with a total of k neurons, then:

$$\text{Softmax}\left(a_i\right) = \frac{e^{a_i}}{\sum_{i=i}^{k} e^{a_i}}. \tag{13}$$

3.3. Semi-supervised learning

Since most data obtained from the underground working environment are unlabeled, there is a lack of labeled data for effectively training the favorable area identification model. Manual labeling requires extracting data to ground, identify, and calibrate these data based on expert knowledge, which is time-consuming and laborious. To enable real-time intelligent identification and evaluation of favorable areas during drilling, the adaptability of model application scenarios should be considered. Therefore, this paper introduces a semi-supervised learning model that can use a small amount of labeled data and a large amount of unlabeled data for model training. This learning mode reduces reliance on manual expertise, saves time and cost caused by large-scale data labeling, and effectively identifies favorable areas.

The specific method is as follows:

1. The model is pre-trained using unlabeled data. The model automatically learns from a large set of unlabeled data to extract the deep features of the favorable area, contained in the original data, and the initial parameters of the SSAE model and the Softmax classifier are obtained.

2. Using a small amount of labeled data for supervision, the loss function of the whole network is constructed, and the network's parameters are adjusted and optimized by the gradient descent method. After iterative training, a deep neural network model with feature extraction and classification functions is achieved.

The SSAE model in the semi-supervised learning mode extracts features related to the target, ensuring correlation between the features and the output, and can better mine the correlation information between similar data and the differences between various types of data. The feature extraction ability is improved, and the classification results will be more accurate.

4. SSAE-plus intelligent identification method

By combining the BN strategy and the Softmax classifier to improve the traditional SSAE model, this paper proposes an intelligent method for identifying favorable areas in a semi-supervised learning mode (SSAE-plus). The implementation of this method is divided into three stages: data preprocessing, feature extraction, and classification identification. The specific step flow is shown in Figure 4.

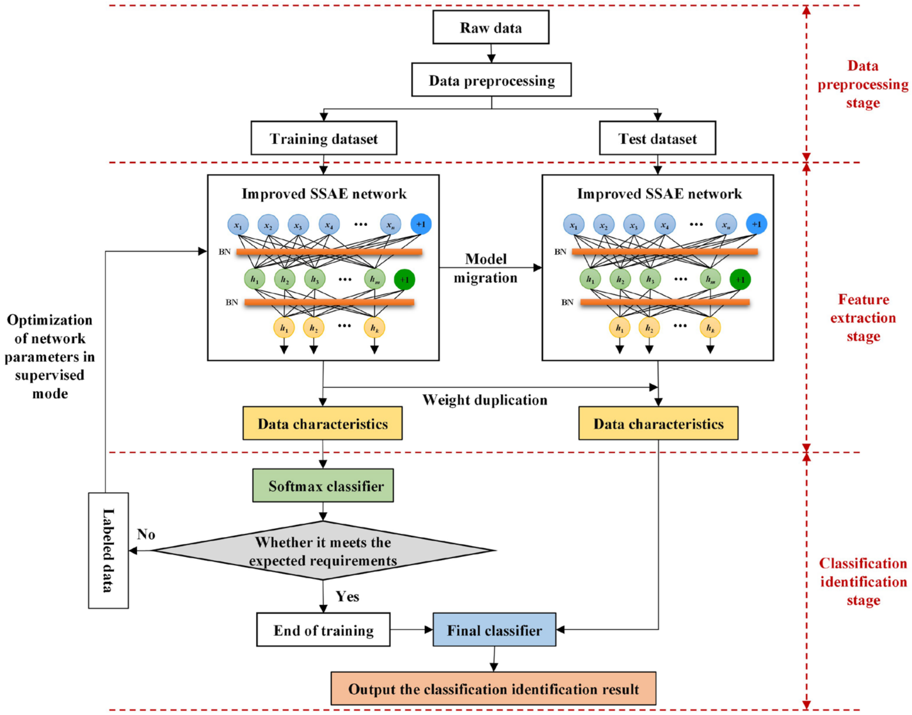


Fig. 4. Step flow of the SSAE-plus intelligent identification method.

Data preprocessing stage

To avoid calculation problems caused by differences in numerical dimensions between data feature attributes, this paper uses a linear function conversion to normalize the original data. This conversion is shown in formula (14):

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{14}$$

where x is the original sample data, x^* is the normalized sample data, and $x_{\max}$ and $x_{\min}$ are the maximum and minimum values in the original sample data, respectively.

Feature extraction stage

First, the unlabeled data is used as the input of the SSAE network, with the weight matrix W and the offset vector b randomly initialized. The number of hidden layers, the number of nodes per layer, the number of training iterations, the sparsity parameter ρ, and the learning rate ε are set. Then, unsupervised training of the SSAE network is performed layer by layer, with BN of the input data before each iteration based on formula (11), and the weights and offset values are updated based on formulas (7) and (8). The loss function is constantly calculated until the error is within the set threshold, at which point the training ends and the feature vector is output.

Classification identification stage
First, a small amount of labeled data is used as input, with the weight matrix *W* and the offset vector *b* obtained from the feature extraction stage. Then, the network is trained again under the supervision of the Softmax classifier, optimizing the overall parameters to make the whole network optimal and obtain the final classifier. Finally, the sample data of the test dataset are input into the trained network, and the favorable area identification result is obtained.

5. Application of the SSAE-plus intelligent method for identifying favorable shale oil areas of the Qingshankou Formation in the Songliao Basin

5.1. Dataset, experimental environment, and evaluation indicators

5.1.1. Dataset

To verify the application effectiveness of the SSAE-plus intelligent identification method, the study focuses on the Qingshankou Formation in the Fuyu-Changchunling area of the Songliao Basin, northeastern China. The SSAE-plus model is applied to identify and evaluate favorable shale oil areas in this formation. The Songliao Basin is a large continental sedimentary basin, with the Fuyu-Changchunling area located in the central depression, where oil shale is mainly developed in the Upper Cretaceous Qingshankou Formation. The shale deposits in the study area are thick and widely distributed, with a lithology primarily consisting of semi-deep to deep lake facies shale. Locally, extremely thin interlayers of siltstone, dolomite, and shell limestone can be seen. Based on core, thin section, and logging data, the TOC content in this area is 0.8–3.4%, with an average of 2.4%. The types of organic matter are mainly kerogen I and II. Effective porosity is 1.8–11.0%, with an average of 4.7%. Permeability is (0.0014–0.3300) × 10^3 μm^2, with an average of 0.11 × 10^3 μm^2. The study area features typical lacustrine organic-rich shale with high exploration potential.

Based on the characteristics of shale oil exploration and development, and drawing on authoritative research results and expert knowledge in oil reservoir engineering [42–44], the study comprehensively considered various factors, such as geological structure, source rock quality, reservoir physical properties, and engineering conditions, and selected ten parameters for intelligent identification and evaluation of favorable shale oil areas [45, 46]. These parameters are listed in Table 1.

Table 1. Evaluation parameters of favorable shale oil areas

Evaluation parameter	Abbreviation	Evaluation function
Shale thickness, m	δ	Only when the shale reaches a certain thickness and has a continuous distribution scale can it provide sufficient oil source conditions and storage space for the formation of shale oil.
Natural gamma ray, API	GR	Shale oil reservoir has high shale content, and the natural gamma value is positively correlated with shale content.
Compensated neutron log, %	CNL	Reflects the hydrogen content of the rock formation: the higher the hydrogen index in the oil layer, the higher the compensated neutron value.
Density, g/cm^3	DEN	Oil shale is rich in organic matter, which has lower density compared to clay minerals. When organic matter replaces the rock skeleton, the density of oil shale decreases.
Acoustic time difference, µs/ft	AC	The acoustic time difference of organic matter is greater than that of rock skeleton. When the formation contains organic matter, such as oil and gas, the AC of the formation increases with higher organic matter content.
Deep lateral resistivity log, Ω·m	LLD	Oil shale is rich in non-conductive organic matter, which reduces its conductivity, so resistivity in shale oil reservoirs shows high values.
Total organic carbon, %	TOC	Represents the ability of shale reservoirs to store and produce hydrocarbons. Higher TOC content indicates greater oil production potential.
Pyrolytic parameter S$_1$, mg/g	S$_1$	Often used to reflect the oil content in shale oil during the resource potential evaluation process.
Effective porosity, %	EPOR	Porosity influences storage and transmission capacity, and oil and gas production in shale reservoirs. The porosity of a shale oil reservoir is low.
Brittleness index, %	BI	Impacts fracturing effectiveness. A higher brittleness index facilitates fracture network formation, which is beneficial to the storage and migration of shale oil.

The logging data in the study area are complete, with the above ten evaluation parameters of favorable shale oil areas being directly or indirectly obtainable from the logging data. Shale thickness, natural gamma, compensated neutron, lithologic density, acoustic time difference, deep lateral resistivity, and effective porosity are all obtained directly from logging. Effective porosity is measured using nuclear magnetic resonance logging, which calculates porosity by assessing the relaxation time (T2 distribution) of fluids in the formation.

BI is calculated using the Poisson's ratio and Young's modulus method, based on shear wave and longitudinal wave time differences, measured by density logging data and array acoustic logging. The calculation is as follows:

$$BI = \frac{\Delta E + \Delta \nu}{2}, \tag{15}$$

$$\Delta E = \frac{E - E_{\min}}{E_{\max} - E_{\min}} \times 100\%, \tag{16}$$

$$\Delta \nu = \frac{\nu_{\max} - \nu}{\nu_{\max} - \nu_{\min}} \times 100\%, \tag{17}$$

$$E = \frac{DEN}{\Delta t_s^{\,2}} \times \frac{3\Delta t_s^{\,2} - 4\Delta t_p^{\,2}}{\Delta t_s^{\,2} - \Delta t_p^{\,2}} \times 10^6, \tag{18}$$

$$\nu = \frac{\Delta t_s^{\,2} - 2\Delta t_p^{\,2}}{2\left(\Delta t_s^{\,2} - \Delta t_p^{\,2}\right)}, \tag{19}$$

where *BI* is brittleness index (%), *E* is Young's modulus (GPa), *v* is Poisson's ratio (dimensionless), subscripts $_{\min}$ and $_{\max}$ represent the minimum and maximum values of this parameter in a certain stratum section, respectively, Δt_p is the longitudinal wave time difference (μs/m), and Δt_s is the shear wave time difference (μs/m).

The TOC was calculated using the Δlog*R* method proposed by Passey [47]. The calculation is as follows:

$$\Delta \log R = \lg\left(LLD / LLD_{baseline}\right) + 0.02\left(AC - AC_{baseline}\right), \tag{20}$$

$$\mathrm{TOC} = \left(\Delta \log R\right) \times 10^{\left(2.297 - 0.1688 LOM\right)}, \tag{21}$$

where LLD$_{baseline}$ is the baseline value of resistivity logging determined after overlaying the lateral resistivity and the acoustic time difference logging curves (Ω·m), AC$_{baseline}$ is the baseline value of acoustic time difference logging determined after overlaying the lateral resistivity and the acoustic time difference logging curves (μs/ft), and *LOM* is a maturity evaluation index, which can be obtained by referring to relevant charts.

TOC is calculated using the following regression:

$$S_1 = 0.6614 \times TOC + 1.5269 \tag{22}$$

where S_1 is the rock pyrolysis parameter (mg/g).

Referring to shale oil resource evaluation criteria [48–51] and considering the actual production performance of effective wells in the study area, the favorable shale oil areas of the Qingshankou Formation are divided into three classes: class I, class II, and unfavorable. Class I is identified as a shale oil-rich zone, with high daily production meeting industrial scale, making it the most preferred area for exploration. Class II has low daily oil production and poor capacity, suitable as a development target only after a further development of technology. Unfavorable areas show no productivity, thus being unsuitable for exploitation.

This study assigns wellhead production data to various depth intervals using tracer concentration detection data for both oil and water phases at different depths. Subsequently, production parameters for each depth interval are calculated. By analyzing the variation patterns of these parameters, quantitative evaluation criteria for different types of favorable areas are established. The daily oil production of class I favorable areas exceeds 0.2 m^3/d, with oil content over 6.0% and per-meter daily production above 0.004 m^3/d. Class II favorable areas produce 0.1–0.2 m^3/d, with oil content between 0.4–3.2%. Unfavorable areas produce below 0.1 m^3/d, with less than 0.4% oil content and per-meter daily production below 0.002 m^3/d. Based on this evaluation standard, the favorable areas of wells Y1 and Y2 are identified and evaluated, with comprehensive results shown in Figures 5 and 6, respectively.

In actual production, the most intuitive measure for defining favorable oil shale areas is production. The division of favorable areas based on production parameters is obtained by mathematical statistical analysis of actual output in the study area. Since different regions face distinct geological conditions, engineering challenges, and development objectives, the definition of favorable shale oil areas should be adapted to local conditions. Using production parameters as the definition basis is straightforward as well as highly flexible, making it convenient for the popularization and application of the method. Specifically, other shale oil blocks can directly classify favorable shale oil areas based on the actual output demand in the region, and there is no need to conduct tedious statistical analysis of the variations in logging parameters or geological data.

In this study, favorable shale oil areas are classified based on production parameters, with results shown in Figures 5 and 6. These data serve two purposes: they verify the accuracy of automatic and intelligent identification of favorable shale oil areas by the SSAE-plus model, and provide valuable labeled data to training the model. Notably, while production data are usually obtained after drilling is completed, logging data are accessible in real time during drilling. To identify and evaluate favorable shale oil areas in real time,

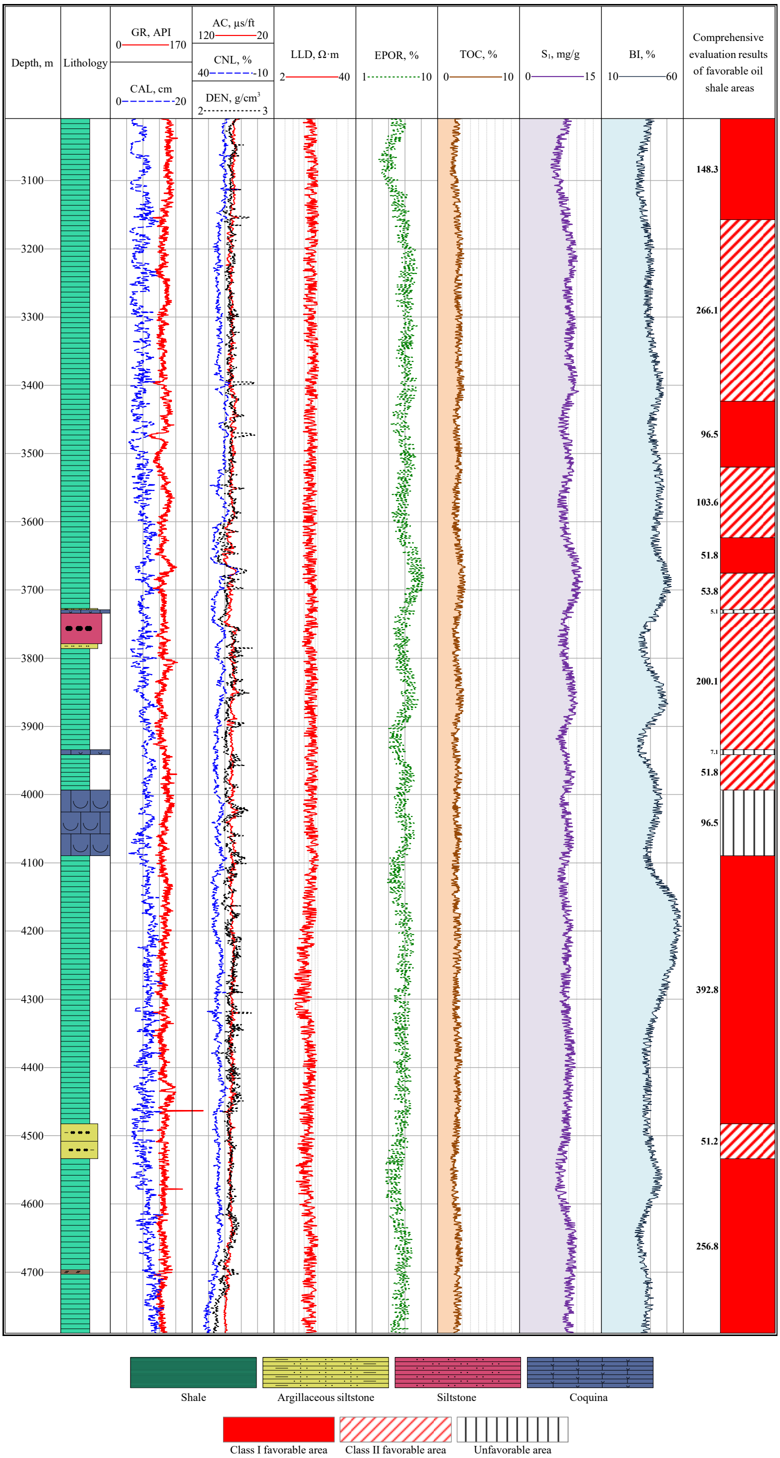


Fig. 5. Comprehensive evaluation results of favorable shale oil areas in well Y1.

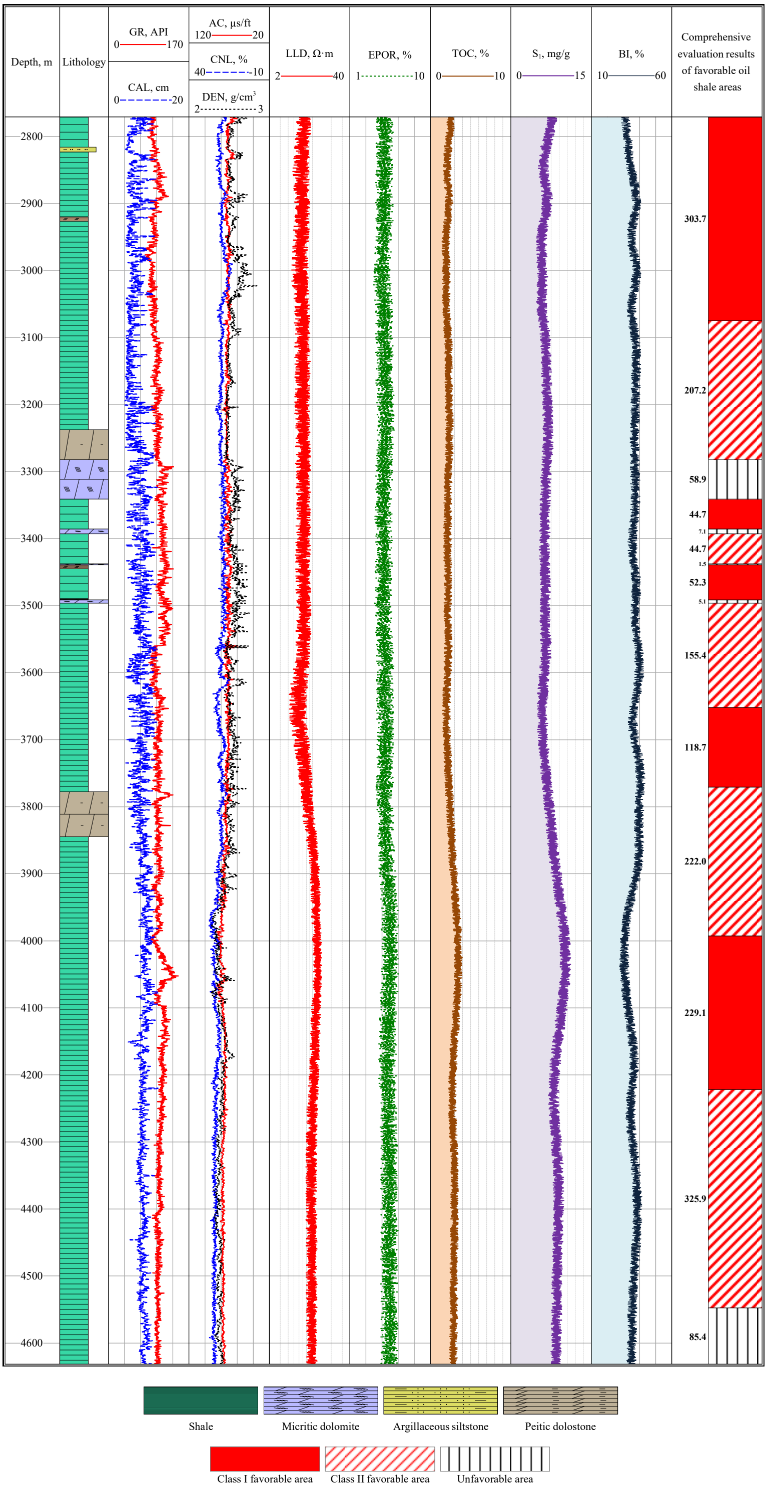


Fig. 6. Comprehensive evaluation results of favorable shale oil areas in well Y2.

so as to guide the drilling process and the formulation of development strategy dynamically, logging data are selected as evaluation parameters. The SSAE-plus model enables to identify favorable areas quickly in real time, providing more accurate guidance for drilling as well as development.

A total of 1,236 samples from well Y1 were used as the training dataset, with an equal distribution of 412 samples across class I favorable, class II favorable, and unfavorable areas. In the training dataset, 80% of the data are unlabeled, while 20% are labeled. For the testing dataset, 309 samples were selected from well Y2, containing 103 samples per classification, and all test data are unlabeled.

5.1.2. Experimental environment

This article uses Google's deep learning framework TensorFlow for experimentation. The programming language is Python. The computer hardware configuration is Intel Core i5-6500 CPU@3.20GHz, with 16 GB of memory and a 64-bit Win10 operating system.

5.1.3. Evaluation indicators

To measure the performance of the SSAE-plus intelligent identification method, accuracy rate (*ACC*), *F*1 value, and the area under the curve (*AUC*) value were selected as the evaluation indexes of the model. The identification results of favorable areas are divided into the following four situations based on real class and model prediction class:

1. True positive (*TP*) – the actual sample belongs to a certain class, and the model correctly predicts it as such.

2. False positive (*FP*) – the actual sample does not belong to a certain class, but the model incorrectly predicts it as that class.

3. True negative (*TN*) – the actual sample is not of a certain class, and the model correctly predicts it as not belonging to that class.

4. False negative (*FN*) – the actual sample belongs to a certain class, but the model incorrectly predicts it as not being in that class.

(1) *ACC* reflects the model's accuracy in predicting the correct classifications across all samples, and the larger the *ACC* value, the better the model performance. *ACC* is calculated as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}. \tag{23}$$

(2) *F*1 value measures the model's comprehensive performance, taking into account both accuracy and recall, with values ranging between [0–1]. The greater the *F*1 value, the higher the model's comprehensive performance. *F*1 value is calculated as follows:

$$F1 = \frac{2TP}{2TP + FP + FN}. \tag{24}$$

(3) AUC value is often used to measure the model's stability, and the value is the geometric area at the lower right of the receiver operating characteristic (ROC) curve. The range of AUC values is [0.5–1]. The greater the AUC value, the better the model's stability. The horizontal axis of the ROC curve is the false positive rate (FPR), and the vertical axis is the true positive rate (TPR), which are defined as follows:

$$FPR = \frac{FP}{TN + FP}, \tag{25}$$

$$TPR = \frac{TP}{TP + FN}. \tag{26}$$

5.2. Model parameter optimization

In training the SSAE-plus intelligent model for identifying favorable areas, the setting of network parameters impact network performance. In this paper, parameters such as the number of hidden layers, the number of nodes per layer, the number of training iterations, and the sparsity parameter ρ, which significantly influences the model's effectiveness, are examined through experiments. Then, the most suitable network parameter values are determined. For the convenience of discussion, other parameter values are set uniformly: weight W and offset b are randomly initialized, learning rate $\varepsilon = 0.01$, sparse constraint term weight $\beta = 3$, and L_2 regularization coefficient $\lambda = 0.002$.

5.2.1. Number of hidden layers and nodes

The SSAE-plus model is a deep neural network based on a SSAE, and its purpose is to obtain deep features of data through layered stacking. However, the number of hidden layers and nodes per layer requires experimental verification for optimal performance. If these parameters are set too low, deep features of data cannot be obtained; if set too high, gradient dispersion will occur, leading to model overfitting, increasing the network's learning burden, and prolonging training time. Therefore, this experiment studies different parameter combinations of hidden layers and nodes in each layer to determine the ideal setup, with results presented in Table 2.

As seen in Table 2, the SSAE-plus model's performance improves as the number of hidden layers increases, enhancing its ACC, $F1$, and AUC values, which is an advantage of deep learning compared to shallow learning. By adding sparse coding layers to enable multi-level abstractions and effective feature extraction from input data, the higher-level features learned by the model capture the core characteristics of the data more accurately, improving the model's classification ability. The SSAE-plus model performs best when the number of hidden layers is three. When this number continues to increase, the network complexity increases too. At this point, model training appears overfitting, the ACC, $F1$, and AUC values no longer increase or even decrease, and training time prolongs significantly. Therefore, considering the best model identification accuracy and the shortest training time, this paper establishes a

Table 2. Performance of the SSAE-plus model with different hidden layers and nodes

No. of hidden layers	No. of nodes per layer	*ACC*, %	*F*1	*AUC*	Training time, s
1	5	65.46	0.530	0.819	17.5
1	8	71.50	0.661	0.836	18.2
1	11	77.34	0.708	0.873	18.9
2	8, 5	80.91	0.793	0.923	19.7
2	11, 5	87.45	0.804	0.944	20.5
2	11, 8	90.11	0.830	0.967	21.0
3	8, 5, 3	94.27	0.893	0.975	24.1
3	11, 5, 3	96.31	0.932	0.980	26.9
3	11, 8, 5	98.82	0.975	0.993	27.3
4	11, 8, 5, 3	98.82	0.969	0.985	35.6
4	13, 8, 5, 3	97.71	0.964	0.985	36.4
4	13, 11, 8, 5	97.25	0.962	0.978	37.8

deep neural network structure with three hidden layers. Through enumeration, the optimal configuration for these layers is identified as having 11, 8, and 5 nodes, respectively. With this setting, the SSAE-plus model can be guaranteed to have good identification accuracy and relatively fast training speed.

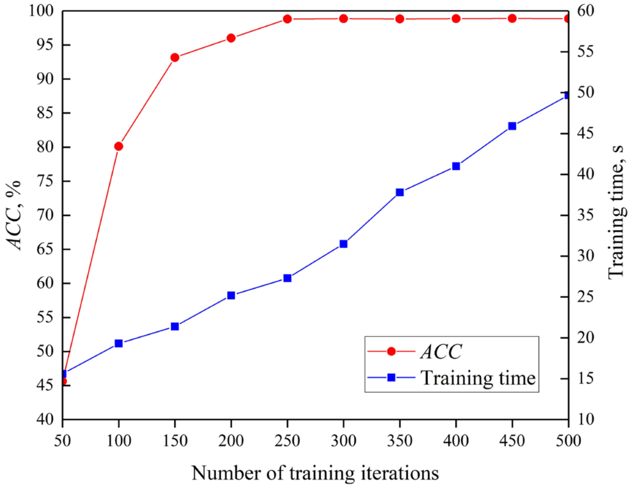


Fig. 7. Identification results using the SSAE-plus model under varying numbers of training iterations.

5.2.2. Number of training iterations

Based on the above network structure, the influence of the varying number of training iterations on the identification effect of the SSAE-plus model is studied. The experimental results are presented in Figure 7.

Figure 7 shows the identification accuracy of favorable areas and model training time for the SSAE-plus model under varying numbers of training iterations. As evident from the figure, when the number of training iterations is fewer than 250, both model identification accuracy and training time show a rapid growth trend with more iterations. However, beyond 250 iterations, the model approaches convergence, with accuracy leveling off while training time prolongs significantly. Therefore, considering both identification accuracy and training time, the number of training iterations for the SSAE-plus model is set to 250.

5.2.3. Sparsity parameter

The sparsity parameter ρ determines the integrity of data features carried or acquired by hidden layer nodes, and its optimal value requires experimental verification. In this paper, the optimal value of ρ is determined through comparative experiments, with results presented in Figure 8.

As shown in Figure 8, when ρ is lower than 0.04, increasing ρ gradually improves the identification accuracy of favorable areas. When ρ is greater than 0.04, accuracy is significantly reduced. Accuracy is the highest when ρ is equal to 0.04. Experimental results suggest that when ρ exceeds a certain threshold, the number of nodes suppressed by the hidden layer is too high, leading to

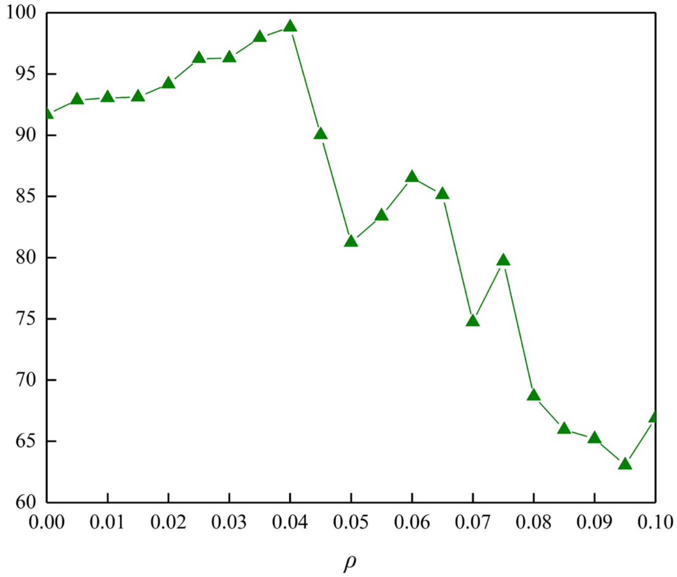


Fig. 8. Identification results of favorable areas of the SSAE-plus model under different sparsity parameters ρ.

incomplete data features, significantly decreasing feature extraction accuracy, and affecting favorable area identification. Therefore, setting ρ to 0.04 enables the SSAE-plus model to better extract deep data features, resulting in the best identification effect of favorable areas.

5.3. Model performance evaluation

This section evaluates the performance of the SSAE-plus model when all network parameters are set optimally.

5.3.1. Performance analysis of batch normalization strategy

Table 3 lists the identification results of favorable areas before and after using the BN strategy. Experimental results indicate that after the BN strategy is used to process the input data of each layer, the performance indexes of the SSAE-plus model are significantly higher: *ACC* improves by 7.08%, *F*1 value increases from 0.926 to 0.975, and *AUC* rises by 0.076. Moreover, training speed is significantly accelerated, reducing training time by nearly half. These results verify the effectiveness of the BN strategy, which successfully suppresses the gradual accumulation of parameter changes from the bottom network to the upper network, decouples network layers, and prevents the gradient from disappearing. Thus, it improves the identification accuracy of favorable areas and significantly accelerates the learning speed of the SSAE-plus model.

Table 3. Identification results of favorable areas using the SSAE-plus model before and after batch normalization strategy

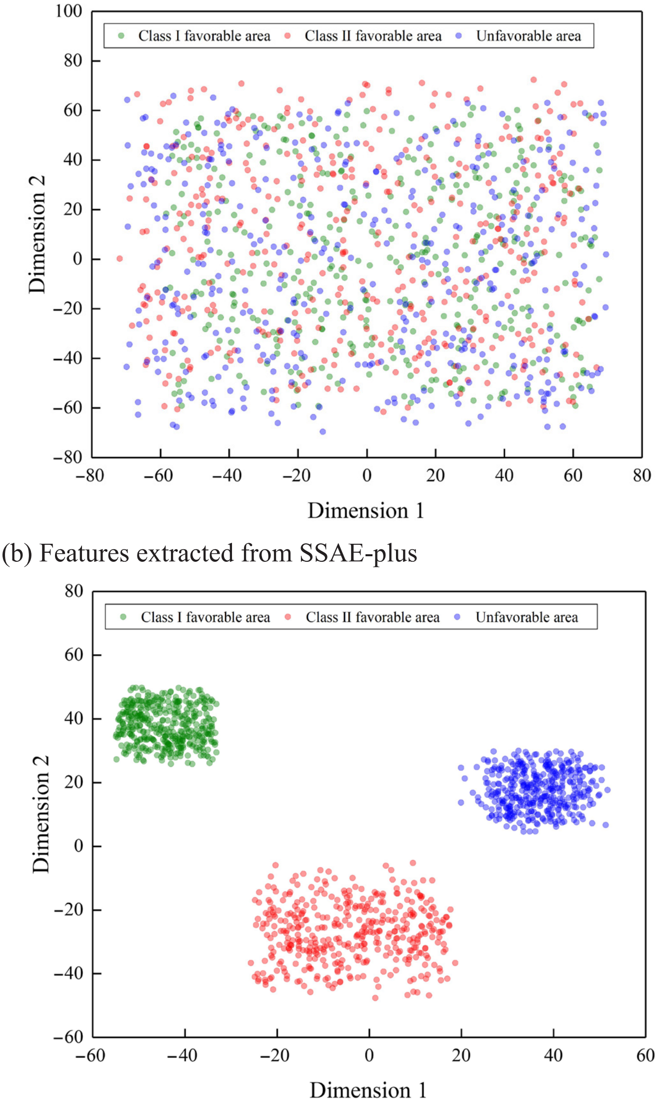
SSAE-plus model	*ACC*, %	*F*1	*AUC*	Training time, s
Before batch normalization strategy	91.74	0.926	0.917	52.8
After batch normalization strategy	98.82	0.975	0.993	27.3

5.3.2. Performance analysis of feature extraction of the SSAE-plus model

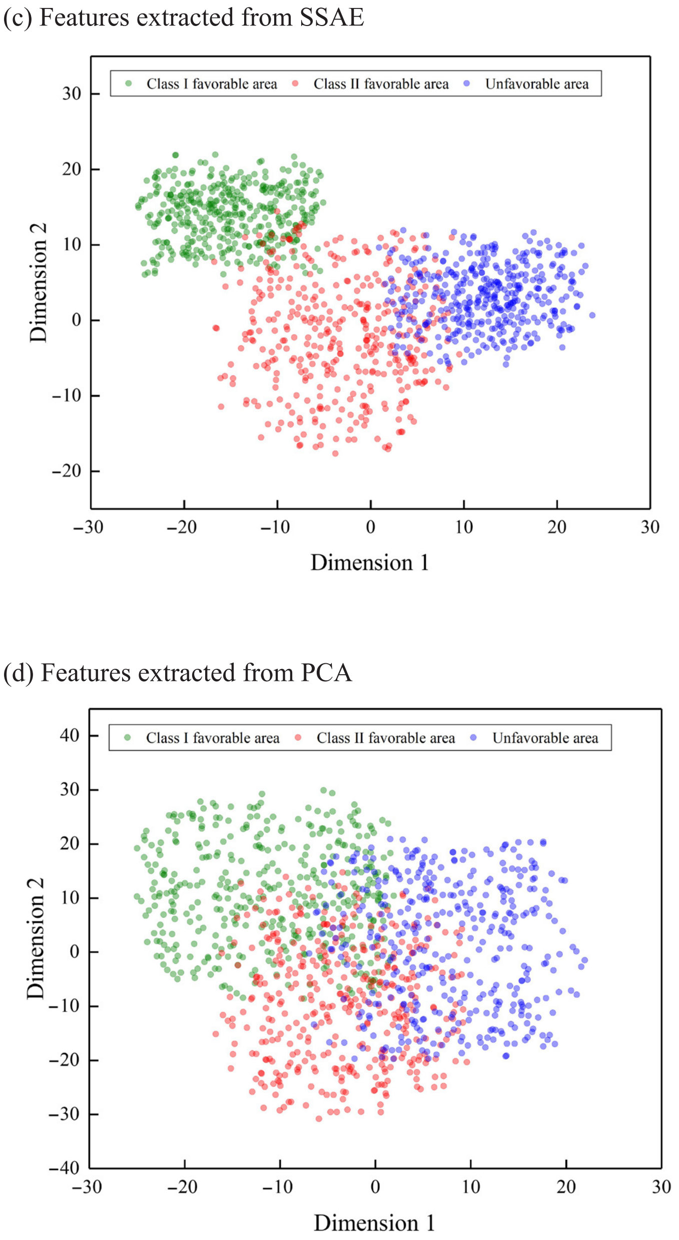
To verify the SSAE-plus model's excellent deep feature extraction performance, this paper compares the feature extraction capabilities of the SSAE-plus, SSAE, and principal component analysis (PCA) on the premise of the same samples. The t-distributed stochastic neighbor embedding (t-SNE) method is used to visualize the features extracted from the three models in two dimensions, with the results shown in Figure 9.

As evident in Figure 9, feature extraction of the original data by all three models – SSAE-plus, SSAE, and PCA – demonstrates regional inter-class separability and intra-class aggregation for all favorable areas, indicating their ability to extract features. However, the SSAE-plus model outperforms the others by achieving concentration of the same classes of favorable areas in a particular area and clear boundaries between different types of favorable areas. In contrast, the features extracted from the SSAE and PCA models overlap in different degrees, making it impossible to distinguish the three favorable areas completely. Experimental results show that the SSAE-plus model has a better feature extraction ability, transforming original space features into a new high-dimensional space, effectively extracting the key features of various favorable areas, and simplifying the learning process.

(a) Original data features



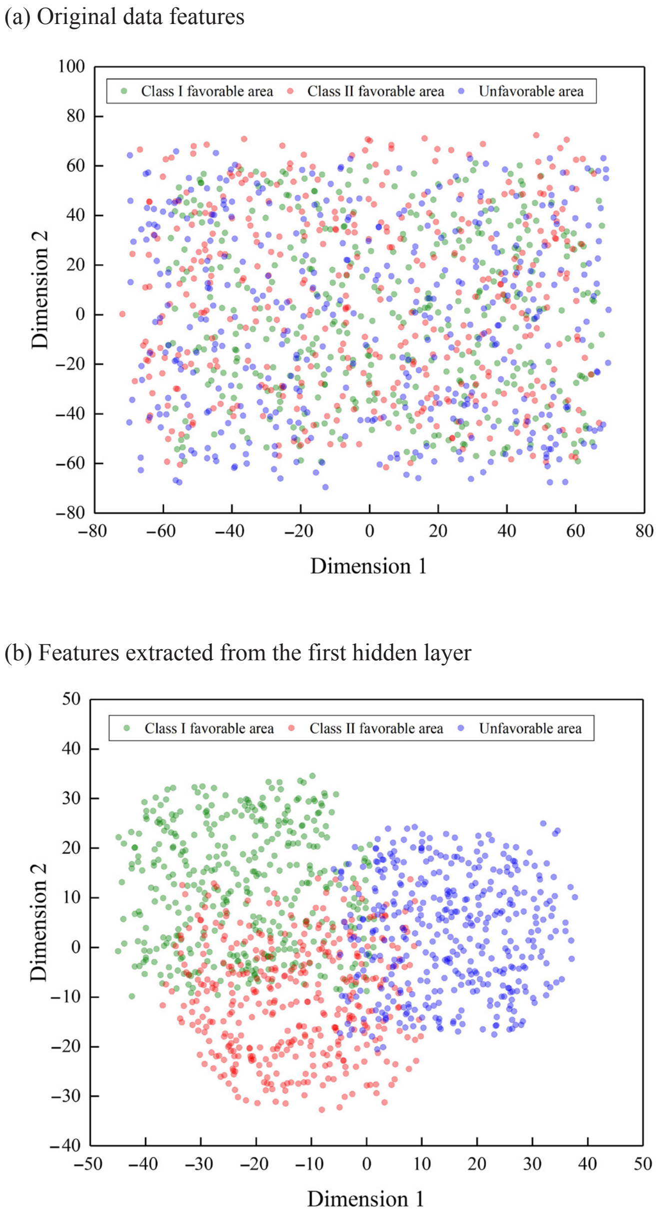
(b) Features extracted from SSAE-plus



(c) Features extracted from SSAE

(d) Features extracted from PCA

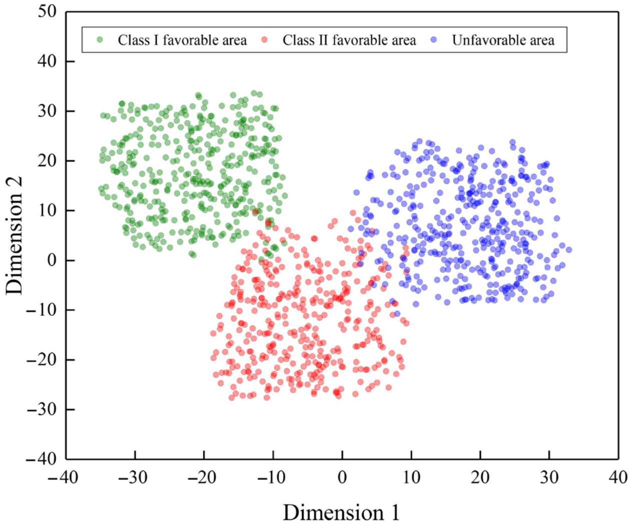
Fig. 9. Visualization of dimension reduction results for features extracted from different models.



(a) Original data features

(b) Features extracted from the first hidden layer

(c) Features extracted from the second hidden layer



(d) Features extracted from the third hidden layer

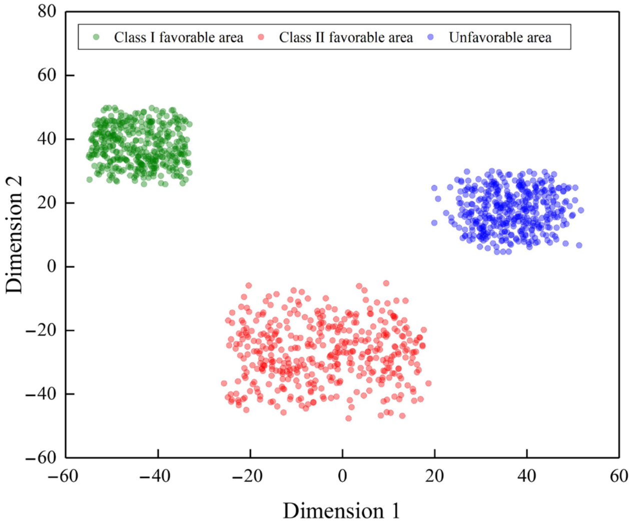


Fig. 10. Visualization of dimension reduction results for features extracted layer by layer from the SSAE-plus model.

To further understand the layer-by-layer feature extraction effect of the SSAE-plus model, the features extracted from each hidden layer of the network are visualized through dimensionality reduction, as shown in Figure 10. Experimental results demonstrate that every time sparse coding is carried out, the features of all three classes of favorable areas are transformed and updated, and after the last hidden layer, they have been completely distinguished. Favorable areas of the same class are gathered together, and the boundaries between different classes are obvious. Results show that the SSAE-plus model successfully extracts the high-dimensional features of favorable areas from original sample data through feature extraction and transformation of hidden layers, revealing the hidden characteristics of original data. This confirms that the model has an excellent ability to express the deep features of data.

5.3.3. Performance analysis of classifiers

To verify that there are obvious advantages in identifying favorable areas by using the Softmax classifier in the semi-supervised learning mode, this study compares it against SVM and KNN classifiers, using the stacked sparse autoencoder after introducing the BN strategy. The comparative identification results for favorable areas are shown in Table 4.

Table 4. Identification results for favorable areas under different classifiers

Classifier	*ACC*, %	*F*1	*AUC*	Training time, s
Softmax	98.82	0.975	0.993	27.3
SVM	98.86	0.954	0.926	39.4
KNN	88.70	0.874	0.891	15.1

As seen in Table 4, both the SVM and Softmax classifiers perform well in identifying favorable areas, achieving *ACC* values above 98.80%. However, the KNN classifier demonstrates the poorest performance among the three. While the SVM classifier has the highest identification accuracy, its *AUC* and *F*1 values are lower than those of the Softmax classifier, showing unstable model performance. The Softmax classifier offers high accuracy, higher *F*1 and *AUC* values, better comprehensive performance of the model, and faster convergence, making it more suitable for real-time intelligent identification of favorable shale oil areas.

5.3.4. Performance analysis of the SSAE-plus model

To further evaluate the overall performance of the SSAE-plus model in identifying favorable oil shale areas, five models – back propagation neural network (BPNN), RF, deep belief network (DBN), SAE, and SSAE – were selected for comparative experiments. The parameter settings for these models are provided in Table 5. Using identical input parameters and sample data as the SSAE-plus model, each model was tested ten times to eliminate the influence of accidental errors, and the results were averaged. Experimental results are shown in Figures 11–14.

Table 5. Model parameter settings

Model	Optimal parameter
BPNN	Training method = gradient descent, max_training iteration number = 250, training goal = 1e-5, learning rate $\varepsilon = 0.01$, max_training time = inf, min_gradient = 1e-10
RF	Max_features = 10, *n*_estimators = 50, max_depth = 100, min_samples_split = 3, min_samples_leaf = 1, max_leaf_nodes = none, min_impurity_split = 0
DBN	Number of hidden layers = 3, number of hidden layer nodes = 13, 10, 9, training iteration number = 250, learning rate $\varepsilon = 0.01$, batch size = 70, momentum = 0.9
SAE	Number of hidden layer nodes = 13, training iteration number = 250, sparsity parameter $\rho = 0.04$, sparse constraint term weight $\beta = 3$, learning rate $\varepsilon = 0.01$, L_2 regularization coefficient $\lambda = 0.002$
SSAE	Number of hidden layers = 3, number of hidden layer nodes = 11, 8, 5, training iteration number = 250, sparsity parameter $\rho = 0.04$, sparse constraint term weight $\beta = 3$, learning rate $\varepsilon = 0.01$, L_2 regularization coefficient $\lambda = 0.002$
SSAE-plus	Number of hidden layers = 3, number of hidden layer nodes = 11, 8, 5, training iteration number = 250, sparsity parameter $\rho = 0.04$, sparse constraint term weight $\beta = 3$, learning rate $\varepsilon = 0.01$, L_2 regularization coefficient $\lambda = 0.002$

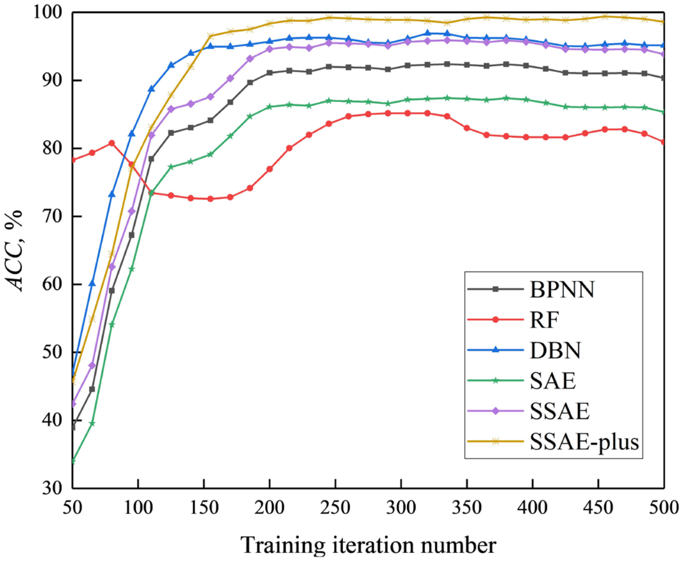


Fig. 11. Accuracy of identifying favorable areas for different models.

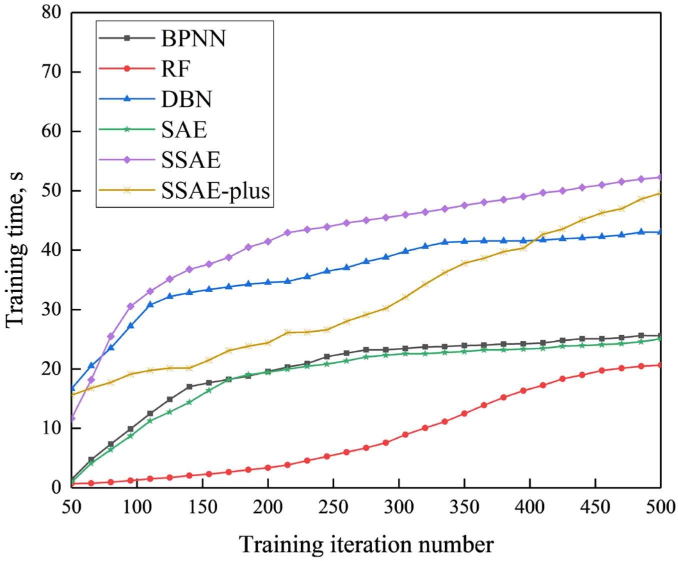


Fig. 12. Training speed of different models.

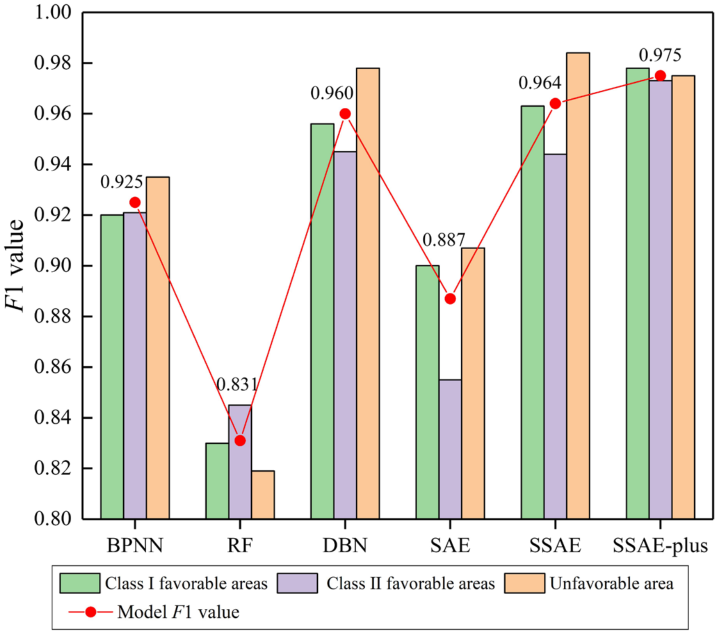


Fig. 13. *F*1 values of different models.

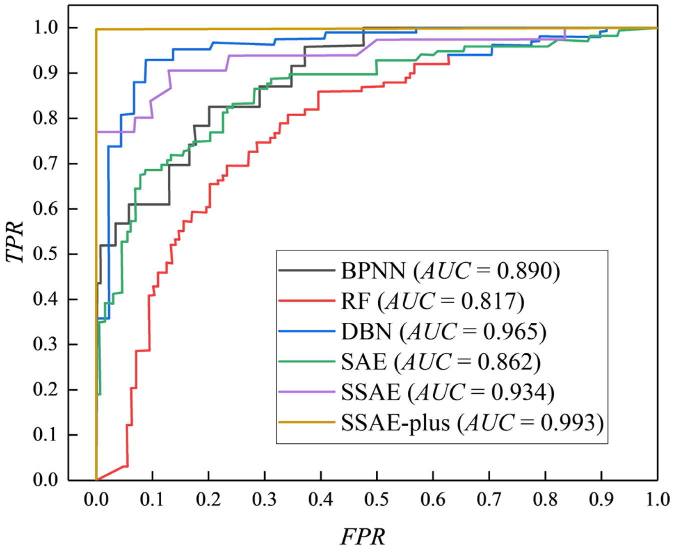


Fig. 14. ROC curves of different models.

As shown in Figure 11, the SSAE-plus, DBN, and SSAE models demonstrate high accuracy in identifying favorable areas, exceeding 90.0%, while shallow machine learning models BP, RF, and SAE offer lower accuracy. Deep learning models can learn better feature expression from the original input, thus also performing better in identifying favorable areas. Figure 12 further illustrates that the SSAE-plus model has the shortest training time among the three models with higher accuracy in identifying favorable areas, while DBN and SSAE exhibit significantly longer training durations. Therefore, SSAE-plus is considerably better than other models in terms of identification accuracy and training time.

Figure 13 shows the *F*1 values of different models. As evident, the overall *F*1 value of the SSAE-plus model is the highest, indicating its superior performance in identifying favorable areas. It outperforms the SSAE and DBN models, while RF demonstrates the poorest performance. Further analysis of the *F*1 values for different classes of favorable areas reveals that DBN and SSAE have significant advantages in identifying unfavorable areas but struggle with class I and II favorable areas. In contrast, SSAE-plus consistently performs well across all three classes of favorable areas, with *F*1 values exceeding 0.973.

As seen in the ROC curves of different models in Figure 14, the SSAE-plus model demonstrates the best performance stability and the highest *AUC* value, followed by the DBN and SSAE models. The RF model shows the weakest performance stability. SSAE-plus excels in identifying class I and II favorable areas, offering higher accuracy, shorter training time, better performance, and stronger stability, making it more suitable for engineering requirements compared to the other five models.

5.3.5. Generalizability analysis of the SSAE-plus model

This study selected six wells from the Qingshankou Formation in the Fuyu-Changchunling area of the Songliao Basin for validation data to test the generalizability of the SSAE-plus model across different datasets. Experimental results are shown in Table 6.

Table 6. Generalizability of the SSAE-plus model

Region	Well number	Sample size	*ACC*, %	*F*1	*AUC*
Qingshankou Formation in the Fuyu-Changchunling area of the Songliao Basin	GL1	1503	98.56	0.975	0.997
	GL2	826	97.83	0.961	0.983
	GL3	220	95.47	0.940	0.952
	GL4	571	96.58	0.950	0.974
	GL5	455	96.02	0.948	0.971
	GL6	1243	98.21	0.973	0.993

Based on Table 6, the SSAE-plus model offers high identification accuracy and higher *F*1 and *AUC* values across different datasets. The average identification accuracy of favorable areas in six wells is 97.11%, with *F*1 above 0.940 and an average *AUC* value of 0.978. These results show that SSAE-plus has good generalization ability and can effectively identify favorable areas in different geographical environments, indicating its potential for widespread popularization and application.

6. Conclusions

1. This paper improves the stacked sparse autoencoder (SSAE) to construct an intelligent method for identifying favorable shale oil areas in a semi-supervised learning mode (SSAE-plus). This method combines the powerful learning ability of unsupervised networks with the reliability of supervised ones, solving the problem of traditional neural networks requiring manual expertise for data labeling and feature extraction, and effectively alleviating gradient dispersion and overfitting during model training. It enables quick and accurate intelligent identification and evaluation of favorable areas.

2. The SSAE-plus model for intelligent identification of favorable areas incorporates a batch normalization strategy. Adding batch normalization layers before the activation function of each hidden layer effectively avoids the disappearance of the learning gradient during model training. The learning speed of the whole neural network is improved. SSAE-plus adopts a multi-layer stacked deep neural network structure to improve the model's ability to distinguish between different classes of favorable areas. Combined with the Softmax classifier, SSAE-plus enhances stability and is trained in a semi-supervised learning mode, which improves its generalization ability.
3. The SSAE-plus model, tested on actual field data, outperforms other machine learning methods in identifying favorable areas, with an accuracy of 98.82% and a training time of 27.3 s. The model's comprehensive performance and stability are excellent. It achieves over 95% accuracy and shows strong generalization ability. Based on reasonable and accurate evaluation results, this model significantly improves the accuracy and efficiency of identifying favorable areas. SSAE-plus is suitable for applications in well pattern deployment and fracturing design, further advancing the exploration and development of unconventional oil and gas resources such as shale oil.

Author contributions

Rui Xu contributed to the conceptualization and methodology of the study and prepared the original draft of the manuscript. Tie Yan was responsible for resources, project administration, and reviewing and editing the manuscript. Shihui Sun contributed to resources, formal analysis, software development, and data curation. Jingyu Qu was involved in methodology, formal analysis, software development, and data curation. Jinyu Feng handled investigation, software development, and data curation.

Data availability statement

Data are available on request from the authors.

Acknowledgments

This work was supported by the Hainan Province Science and Technology Special Fund's project "Research on Real-time Intelligent Sensing Technology for Closed-loop Drilling of Oil and Gas Reservoirs in Deepwater Drilling" (ZDYF2023GXJS012), the National Natural Science Foundation of China (grant No. 52174001), the National Natural Science Foundation of China (grant No. 52004064), and Heilongjiang Provincial Government and Daqing Oilfield's first batch of the scientific and technological key project "Research on the Construction Technology of Gulong Shale Oil Big Data Analysis System" (DQYT-2022-JS-750). The authors thank all reviewers for their critical and constructive comments. The publication costs of this article were partially covered by the Estonian Academy of Sciences.

References

1. Wang, E., Li, C., Feng, Y., Song, Y., Guo, T., Li, M., Chen, Z. Novel method for determining the oil moveable threshold and an innovative model for evaluating the oil content in shales. *Energy*, 2022, **239**(A), 121848. https://doi.org/10.1016/j.energy.2021.121848
2. Sun, P., Li, W., Liu, Z., Niu, D., Wu, X., Tao, L., Wang, Z., Luan, Z. Selection of favourable targets for the in-situ conversion of continental oil shale in China. *Oil Shale*, 2023, **40**(3), 177–193. https://doi.org/10.3176/oil.2023.3.01
3. Li, Y., Zhao, Q., Lyu, Q., Xue, Z., Cao, X., Liu, Z. Evaluation technology and practice of continental shale oil development in China. *Pet. Explor. Dev.*, 2022, **49**(5), 1098–1109. https://doi.org/10.1016/S1876-3804(22)60335-5
4. Mi, S., Guo, Q., Zhang, Q., Wang, J. Classification and potential of continental shale oil resources in China and resource evaluation methods and criteria. *Oil Shale*, 2023, **40**(4), 283–320. https://doi.org/10.3176/oil.2023.4.02
5. Hou, L., Zou, C., Yu, Z., Luo, X., Wu, S., Zhao, Z., Lin, S., Yang, Z., Zhang, L., Wen, D., Cui, J. Quantitative assessment of the sweet spot in marine shale oil and gas based on geology, engineering, and economics: a case study from the Eagle Ford Shale, USA. *Energy Strat. Rev.*, 2021, **38**, 100713. https://doi.org/10.1016/j.esr.2021.100713
6. Chen, B., Cai, J., Chen, X., Wu, D., Pan, Y. A review on oil shale in-situ mining technologies: opportunities and challenges. *Oil Shale*, 2024, **41**(1), 1–25. https://doi.org/10.3176/oil.2024.1.01
7. Freedman, R., Rose, D., Sun, B., Brown, R. L., Malizia, T. Novel method for evaluating shale-gas and shale-tight-oil reservoirs using advanced well-log data. *SPE Reserv. Eval. Eng.*, 2019, **22**(01), 282–301. https://doi.org/10.2118/181480-PA
8. Yan, T., Xu, R., Sun, S.-H., Hou, Z.-K., Feng, J.-Y. A real-time intelligent lithology identification method based on a dynamic felling strategy weighted random forest algorithm. *Pet. Sci.*, 2023, **21**(2), 1135–1148. https://doi.org/10.1016/j.petsci.2023.09.011
9. Chopra, S., Sharma, R., Marfurt, K. Workflows for shale gas reservoir characterization. In: *75th EAGE Conference & Exhibition incorporating SPE EUROPEC 2013*, June 10–13, 2013, London, UK. European Association of Geoscientists & Engineers, 2013, 348-00764. https://doi.org/10.3997/2214-4609.20130215

10. Yang, S., Zhao, X., Zhang, X., Zhou, L., Liu, H. Prediction of “sweet spots” of tight sandstone gas reservoirs in Jishen 1 area, Turpan-Hami Basin. *J. Xi'an Shiyou Univ. Nat. Sci. Ed.*, 2014, **29**(05), 1–8. https://doi.org/10.3969/j.issn.1673-064X.2014.05.002
11. Ismail, A., Raza, A., Gholami, R., Rezaee, R. Reservoir characterization for sweet spot detection using color transformation overlay scheme. *J. Pet. Explor. Prod. Technol.*, 2020, **10**, 2313–2334. https://doi.org/10.1007/s13202-020-00913-5
12. Yao, Q., Yang, B., Zhang, Q. Dynamic uncertain causality graph applied to the intelligent evaluation of a shale-gas sweet spot. *Energies*, 2021, **14**(17), 5228–5247. https://doi.org/10.3390/en14175228
13. Misra, S., Li, H., He, J. Index construction, dimensionality reduction, and clustering techniques for the identification of flow units in shale formations suitable for enhanced oil recovery using light-hydrocarbon injection. In: *Machine Learning for Subsurface Characterization*. Gulf Professional Publishing, 2019, 157–181. http://dx.doi.org/10.1016/B978-0-12-817736-5.00002-8
14. Li, X., Chen, K., Li, P., Li, J., Geng, H., Li, B., Li, X., Wang, H., Zang, L., Wei, Y., Zhao, R. A new evaluation method of shale oil sweet spots in Chinese Lacustrine Basin and its application. *Energies*, 2021, **14**(17), 5519–5533. https://doi.org/10.3390/en14175519
15. Liu, S., Liu, Y., Zhang, X., Guo, W., Kang, L., Yu, R., Sun, Y. Geological and engineering integrated shale gas sweet spots evaluation based on fuzzy comprehensive evaluation method: a case study of Z shale gas field HB block. *Energies*, 2022, **15**(2), 602–621. https://doi.org/10.3390/en15020602
16. Kwilosz, T., Filar, B., Miziołek, M. Use of cluster analysis to group organic shale gas rocks by hydrocarbon generation zones. *Energies*, 2022, **15**(4), 1464–1477. https://doi.org/10.3390/en15041464
17. Guan, Q., Dong, D., Wang, Y., Huang, J., Wang, S. AHP application to shale gas exploration areas assessment in Sichuan Basin. *Bull. Geol. Sci. Technol.*, 2015, **34**(5), 91–97.
18. Riahi, S., Fathianpour, N., Tabatabaei, S. H. Presenting a mapping method based on fuzzy logic and TOPSIS multi criteria decision-making methods to detect promising porphyry copper mineralization areas in the east of the Sarcheshmeh copper metallogenic district. *J. Econ. Geol.*, 2017, **9**(2), 357–374. https://doi.org/10.22067/econg.v9i2.45829
19. Akbar, M. N. A., Nugraha, S. T. K-mean cluster analysis for better determining the sweet spot intervals of the unconventional organic-rich shale: a case study. *Contemp. Trends Geosci.*, 2018, **7**(2), 200–213. https://doi.org/10.2478/ctg-2018-0014
20. Zhou, Y., Zhao, A., Yu, Q., Zhang, D., Zhang, Q., Lei, Z. A new method for evaluating favorable shale gas exploration areas based on multi-linear regression analysis: a case study of marine shales of Wufeng–Longmaxi Formations, Upper Yangtze Region. *Sediment. Geol. Tethyan Geol.*, 2021, **41**(3), 387–397. https://doi.org/10.19826/j.cnki.1009-3850.2021.05001

21. Niu, D., Li, Y., Zhang, Y., Sun, P., Wu, H., Fu, H., Wang, Z. Multi-scale classification and evaluation of shale reservoirs and 'sweet spot' prediction of the second and third members of the Qingshankou Formation in the Songliao Basin based on machine learning. *J. Pet. Sci. Eng.*, 2022, **216**, 110678. https://doi.org/10.1016/j.petrol.2022.110678
22. Amosu, A., Imsalem, M., Sun, Y. Effective machine learning identification of TOC-rich zones in the Eagle Ford Shale. *J. Appl. Geophys.*, 2021, **188**, 104311. https://doi.org/10.1016/j.jappgeo.2021.104311
23. Kalantari-Dahaghi, A. Machine learning applications in unconventional shale gas systems. In: *Unconventional Shale Gas Development* (Moghanloo, R. G., ed.). Gulf Professional Publishing, 2022, 433-443. https://doi.org/10.1016/B978-0-323-90185-7.00015-7
24. Zhan, L., Hu, J., Wang, S., Wang, K., Guo, B., Yang, X. Machine learning-based estimated ultimate recovery prediction and sweet spot evaluation of shale oil. *Int. J. Oil, Gas Coal Technol.*, 2022, **30**(1), 1–17. https://doi.org/10.1504/IJOGCT.2022.122089
25. Hui, G., Chen, Z., Wang, Y., Zhang, D., Gu, F. An integrated machine learning-based approach to identifying controlling factors of unconventional shale productivity. *Energy*, 2023, **266**, 126512. https://doi.org/10.1016/j.energy.2022.126512
26. Qian, K., He, Z., Liu, X., Chen, Y. Intelligent prediction and integral analysis of shale oil and gas sweet spots. *Pet. Sci.*, 2018, **15**, 744–755. https://doi.org/10.1007/s12182-018-0261-y
27. Tahmasebi, P., Javadpour, F., Sahimi, M. Data mining and machine learning for identifying sweet spots in shale reservoirs. *Expert Syst. Appl.*, 2017, **88**, 435–447. https://doi.org/10.1016/j.eswa.2017.07.015
28. Hauge, V. L., Hermansen, G. H. Machine learning methods for sweet spot detection: a case study. In: *Geostatistics Valencia 2016* (Gómez-Hernández, J. J., Rodrigo-Ilarri, J., Rodrigo-Clavero, M. E., Cassiraga, E., Vargas-Guzmán, J. A., eds). Springer Cham, 2017, 573–588. https://doi.org/10.1007/978-3-319-46819-8_38
29. Raef, A. E., Totten, M. W., Linares, A., Kamari, A. Lithofacies control on reservoir quality of the Viola Limestone in southwest Kansas and unsupervised machine learning approach of seismic attributes facies-classification. *Pure Appl. Geophys.*, 2019, **176**, 4297–4308. https://doi.org/10.1007/s00024-019-02205-4
30. Otchere, D. A., Ganat, T. O. A., 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. Pet. Sci. Eng.*, 2021, **200**, 108182. https://doi.org/10.1016/j.petrol.2020.108182
31. Tang, J., Fan, B., Xiao, L., Tian, S., Zhang, F., Zhang, L., Weitz, D. A new ensemble machine-learning framework for searching sweet spots in shale reservoirs. *SPE J.*, 2021, **26**(01), 482–497. https://doi.org/10.2118/204224-PA
32. Huo, F., Chen, Y., Ren, W., Dong, H., Yu, T., Zhang, J. Prediction of reservoir key parameters in 'sweet spot' on the basis of particle swarm optimization to TCN-

LSTM network. *J. Pet. Sci. Eng.*, 2022, **214**, 110544. https://doi.org/10.1016/j.petrol.2022.110544

33. Qin, Z., Xu, T. Shale gas geological “sweet spot” parameter prediction method and its application based on convolutional neural network. *Sci. Rep.*, 2022, **12**(1), 1–15. https://doi.org/10.1038/s41598-022-19711-6
34. Boyadzhiev, T., Dimitrova, S., Tsvetanov, S. Comparison of auto-encoder training algorithms. In: *Human Interaction, Emerging Technologies and Future Systems V*: *Proceedings of the 5th International Virtual Conference on Human Interaction and Emerging Technologies, IHIET 2021*, August 27–29, 2021 and *the 6th IHIET: Future Systems (IHIET-FS 2021)*, October 28–30, 2021, France. Springer Cham, 2022, 698–704. https://doi.org/10.1007/978-3-030-85540-6_88
35. Thapa, K., Seo, Y., Yang, S.-H., Kim, K. Semi-supervised adversarial auto-encoder to expedite human activity recognition. *Sensors*, 2023, **23**(2), 683. https://doi.org/10.3390/s23020683
36. Li, R., Wang, X., Quan, W., Song, Y., Lei, L. Robust and structural sparsity auto-encoder with L21-norm minimization. *Neurocomputing*, 2021, **425**, 71–81. https://doi.org/10.1016/j.neucom.2020.02.051
37. Wang, H., Sun, J., Gu, X., Song, W. A novel multi-scale and sparsity auto-encoder for classification. *Int. J. Mach. Learn. Cybern.*, 2022, **13**, 3909–3925. https://doi.org/10.1007/s13042-022-01632-5
38. Chen, Y., Chen, Y., Feng, X., Yang, X., Zhang, J., Qiu, Z., He, Y. Variety identification of orchids using Fourier transform infrared spectroscopy combined with stacked sparse auto-encoder. *Molecules*, 2019, **24**(13), 2506. https://doi.org/10.3390/molecules24132506
39. Zhang, Y., Zhao, C., Chen, M., Yuan, M. Integrating stacked sparse auto-encoder into matrix factorization for rating prediction. *IEEE Access*, 2021, **9**, 17641–17648. https://doi.org/10.1109/ACCESS.2021.3053291
40. Garbin, C., Zhu, X., Marques, O. Dropout *vs.* batch normalization: an empirical study of their impact to deep learning. *Multimed. Tools Appl.*, 2020, **79**, 12777–12815. https://doi.org/10.1007/s11042-019-08453-9
41. Awais, M., Iqbal, M. T. B., Bae, S.-H. Revisiting internal covariate shift for batch normalization. *IEEE Trans. Neural Netw. Learn. Syst.*, 2021, **32**(11), 5082–5092. https://doi.org/10.1109/TNNLS.2020.3026784
42. Xue, H., Du, T., Yan, H. The movable resource volume evaluation of Qingshankou Formation shale oil in northern Songliao Basin. *Adv. Mater. Res.*, 2014, **848**, 246–250. https://doi.org/10.4028/www.scientific.net/AMR.848.246
43. Hu, F., Liu, Z., Meng, Q., Song, Q., Xie, W. Characteristics and comprehensive utilization of oil shale of the Upper Cretaceous Qingshankou Formation in the southern Songliao Basin, NE China. *Oil Shale*, 2017, **34**(4), 312–335. https://doi.org/10.3176/oil.2017.4.02
44. Sun, L., Cui, B., Zhu, R., Wang, R., Feng, Z., Li, B., Zhang, J., Gao, B., Wang, Q., Zeng, H., Liao, Y., Jiang, H. Shale oil enrichment evaluation and production law in Gulong Sag, Songliao Basin, NE China. *Pet. Explor. Dev.*, 2023, **50**(3), 505–519. https://doi.org/10.1016/S1876-3804(23)60406-9

45. Han, X., Feng, F., Zhang, X., Cao, J., Zhang, J., Suo, Y., Yan, Y., Yan, M. An unequal fracturing stage spacing optimization model for hydraulic fracturing that considers cementing interface integrity. *Pet. Sci.*, 2023, **20**(4), 2165–2186. https://doi.org/10.1016/j.petsci.2023.05.010
46. Li, S., Chen, Z., Li, W., Yan, T., Bi, F., Tong, Y. An FE simulation of the fracture characteristics of blunt rock indenter under static and harmonic dynamic loadings using cohesive elements. *Rock Mech. Rock Eng.*, 2023, **56**, 2935–2947. https://doi.org/10.1007/s00603-022-03214-x
47. Passey, Q. R., Creaney, S., Kulla, J. B., Moretti, F. J., Stroud, J. D. A practical model for organic richness from porosity and resistivity logs. *AAPG Bull.*, 1990, **74**(12), 1777–1794. https://doi.org/10.1306/0c9b25c9-1710-11d7-8645000102c1865d
48. Sui, Y., Cao, G., Guo, T., Li, Z., Bai, Y., Li, D., Zhang, Z. Development of gelled acid system in high-temperature carbonate reservoirs. *J. Pet. Sci. Eng.*, 2022, **216**, 110836. https://doi.org/10.1016/j.petrol.2022.110836
49. Sun, P., Liu, Z., Gratzer, R., Xu, Y., Liu, R., Li, B., Meng, Q., Xu, J. Oil yield and bulk geochemical parameters of oil shales from the Songliao and Huadian Basins, China: a grade classification approach. *Oil Shale*, 2013, **30**(3), 402–418. https://doi.org/10.3176/oil.2013.3.03
50. Zhang, J., Xu, X., Bai, J., Liu, W., Chen, S., Liu, C., Li, Y. Enrichment and exploration of deep lacustrine shale oil in the first member of Cretaceous Qingshankou Formation, southern Songliao Basin, NE China. *Pet. Explor. Dev.*, 2020, **47**(4), 683–698. https://doi.org/10.1016/S1876-3804(20)60085-4
51. Sun, L., Liu, H., He, W., Li, G., Zhang, S., Zhu, R., Jin, X., Meng, S., Jiang, H. An analysis of major scientific problems and research paths of Gulong shale oil in Daqing Oilfield, NE China. *Pet. Explor. Dev.*, 2021, **48**(3), 527–540. https://doi.org/10.1016/S1876-3804(21)60043-5