2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)

Research of KPCA-SVM predicting model in hydrocarbon reservoir

Dong Xie
Fundamentals Department
Air Force Engineering University
Xi'an, China
xd040852@126.com

*Hongbo Li
Fundamentals Department
Air Force Engineering University
Xi'an, China
*lihongbo7909@163.com

Jiyang Li
Fundamentals Department
Air Force Engineering University
Xi'an, China
Lijiyang1991@163.com

Peng Bai
Fundamentals Department
Air Force Engineering University
Xi'an, China
Baipeng_kgd@163.com

***Abstract*—The progressive depletion of oil and gas resources, coupled with the scarcity of conventional reservoirs, has intensified the challenge of accurately forecasting hydrocarbon content in complex reservoirs. These reservoirs are characterized by heterogeneity, lithological variations, multiphase distribution, and intricate pore and fracture networks, making prediction particularly difficult when exploration data is limited. To tackle this issue, this study presents an innovative integration of Kernel Principal Component Analysis (KPCA) with Support Vector Machines (SVM) to develop a novel predictive model termed Kernel Principal Component Analysis Support Vector Machine (KPCA-SVM). This model capitalizes on the high precision and flexibility of KPCA in data processing and the exceptional recognition and generalization capabilities of SVM, especially with small sample sizes, to effectively address the prediction of complex hydrocarbon reservoirs under constrained conditions. Utilizing actual geological data samples, this research conducts an in-depth optimization study of key model parameters, including kernel function selection, penalty factor *C* setting, and data sample size determination, providing recommendations for their optimal selection. The findings demonstrate that the KPCA-SVM model improves predictive accuracy by over 10% compared to other existing models, establishing its significant superiority in the field of hydrocarbon reservoir prediction. This underscores not only its theoretical significance but also its practical value in real-world applications.**

Keywords—KPCA, SVM, Forecasts, Hydrocarbon

I. INTRODUCTION

With the progressive exhaustion of conventional hydrocarbon reserves, these sources are insufficient to satisfy contemporary energy requirements. Anticipating the content of intricate hydrocarbon reservoirs—marked by heterogeneity, diverse lithology, multiphase distribution, and a network of pores and fractures—emerges as a pivotal avenue for future advancements [1-4]. Conventional approaches to reservoir forecasting heavily rely on well-log data and the expertise of exploration experts [5-7]. Such techniques not only hinder exploration efficiency but also are subject to the explorers' subjective interpretations, introducing considerable unpredictability into the forecasting outcomes. The inherent variability and intricacy of complex hydrocarbon reservoirs render traditional, experience-based prediction methods increasingly unfit for reservoirs exhibiting low porosity, low permeability, and low resistivity. Consequently, there is an imperative to innovate predictive techniques to boost the precision and productivity of reservoir forecasting [8-9].

Researchers have previously introduced an array of methods predicated on linear mapping and homogeneous formation assumptions, encompassing multiple regression analysis [10], fuzzy recognition technology [11], grey system theory [12], and dynamic cluster analysis [13]. Despite achieving some level of success, these methodologies are largely tailored to the attributes of traditional hydrocarbon reservoirs [14]. They become inapplicable when confronted with complex hydrocarbon reservoirs defined by low porosity, low permeability, low resistivity, and nonlinear geological conditions. The relationship between hydrocarbon reservoirs and the collected data samples is, in fact, highly nonlinear and lacks a distinct functional link. Hence, the introduction of nonlinear information processing techniques, such as artificial neural networks (ANN) [15] and support vector machines (SVM) [16], becomes essential for achieving nonlinear predictive capabilities.

Utilizing artificial neural networks (ANN) for forecasting hydrocarbon reservoirs significantly elevates predictive accuracy [15]. Nonetheless, ANN encounters challenges when addressing nonlinear mapping issues, necessitating an ample dataset that spans the entire feature space to ensure precision. In scenarios where the training dataset is inadequate or does not encompass the complete feature space, the predictive accuracy may be compromised. Moreover, ANN grapples with intrinsic limitations that are not easily surmountable, including a propensity for local minimum traps that hinder global optimization efforts and overfitting concerns that can severely impede the model's generalization capabilities.

Support vector machines (SVM), grounded in statistical learning theory (SLT), excel in addressing challenges posed by limited data samples and nonlinear complexities, making them a popular choice across diverse domains[16]. In the context of intricate tasks like hydrocarbon reservoir forecasting, SVM offers distinct benefits over artificial neural networks (ANN), such as the implementation of structural risk minimization

979-8-3315-0976-7/25/$31.00 ©2025 IEEE

(SRM), the achievement of globally optimal solutions, and the use of kernel functions for dimensionality mapping.

This study introduces an innovative fusion of kernel principal component analysis (KPCA) [17-20] with SVM, crafting a novel predictive framework known as KPCA-SVM. Anchored in real-world geological data samples, this model harnesses KPCA for feature extraction and SVM for accurate forecasting and categorization. In the realm of hydrocarbon reservoir prediction, KPCA serves to condense data dimensions, simplifying the feature data's intricacy, while SVM capitalizes on its proficiency in small-sample learning to overcome the scarcity of data samples. This synergistic approach enhances the KPCA-SVM model's predictive precision and bolsters its resilience and flexibility in the face of complex hydrocarbon reservoirs, equipping it as a potent instrument for hydrocarbon exploration and development endeavors.

Given the pivotal role of model parameter selection in shaping predictive outcomes, this research also places a significant emphasis on parameter optimization. Through meticulous experimentation with actual data samples, this paper delves into the influence of key model parameters, including kernel functions, the penalty factor *C*, and data sample quantities, on the predictive results. It delineates an optimized parameter range for the model, which holds significant value for both theoretical insights and practical applications.

II. THEORY AND MODEL

A. Prediction Model

Unveiling the intricate nonlinear correlations between geological data samples and hydrocarbon reservoirs lies at the core of predictive modeling within this domain. As depicted in Fig.1, the KPCA-SVM model for hydrocarbon reservoir forecasting is executed through a two-tiered approach:

- Feature Extraction and Dimensionality Reduction: This initial phase employs Kernel Principal Component Analysis (KPCA) to distill essential features and condense the dimensionality of the dataset. Such a strategy not only mitigates the computational load of subsequent analytical steps but also circumvents the challenges posed by the scarcity of samples in hydrocarbon reservoir forecasting scenarios.
- Linear Prediction of Small Samples: In this subsequent phase, Support Vector Machines (SVM) are harnessed, utilizing kernel transformations to relocate the data from a lower-dimensional space to a higher-dimensional one. This maneuver effectively converts the initial nonlinear problem into a linear counterpart, facilitating more manageable data processing. Through ongoing refinement and parameter tuning of the SVM model, predictive accuracy is enhanced, thus empowering the trained model to adeptly tackle the nonlinear prediction of small-sample hydrocarbon reservoirs.

B. KPCA Processing

Kernel Principal Component Analysis (KPCA) represents an advancement over traditional Principal Component Analysis (PCA), offering a strategy for nonlinear dimensionality reduction. This technique employs a nonlinear mapping function, denoted as Φ, to project the original nonlinearly inseparable data points, x_i, into a higher-dimensional space $\Phi(x_i)$. Within this elevated-dimensional context, the data becomes linearly separable, thus enabling the application of linear methods for analysis and classification.

$\boldsymbol{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_n\}, \mathbf{x}_i \in R_d$ expresses the data sample collection that resides within a d-dimensional Euclidean space as $\Phi : R^d \rightarrow F$, the transformation process is described, with F

representing the feature space. The computation of the inner product of kernel functions for data samples within the feature space is articulated in (1):

$$K(x_i, x_j) = < \Phi(x_i) \bullet \Phi(x_j) > \tag{1}$$

Assuming the mean of the data samples in the feature space $\sum_{i=1}^{n} \Phi(\mathbf{x}_i) = 0$. The covariance matrix of the mapped data samples can be shown by (2):

$$C = \frac{1}{N} \sum_{i=i}^{N} \Phi(\mathbf{x}_i) \Phi(\mathbf{x}_i)^T \tag{2}$$

In feature space, the relationship between the eigenvalue λ and the eigenvector Q can be shown by (3):

$$\lambda Q = CQ \tag{3}$$

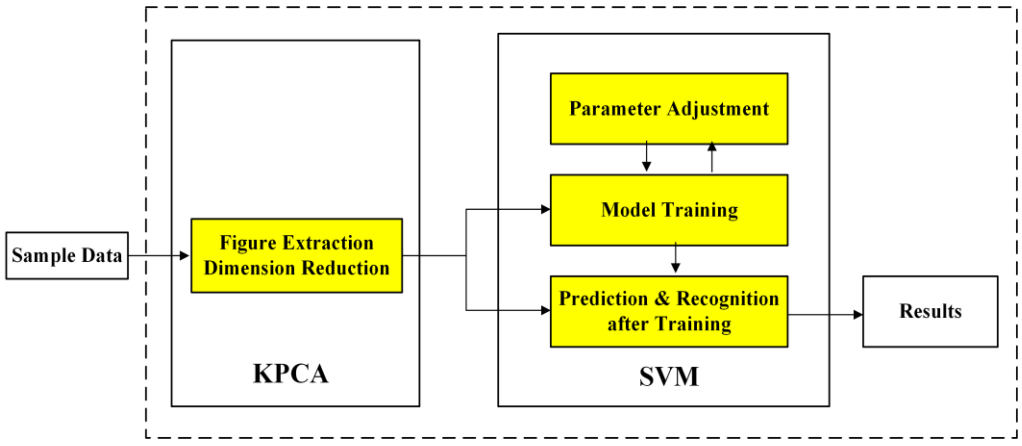

Fig. 1. KPCA-SVM prediction model

In this equation, Q is normalized eigenvector, can be expressed by (4). α_i is a coefficient of linear combination.

$$\boldsymbol{Q} = \sum_{i=}^{n} \alpha_i \Phi(\mathbf{x}_i) \tag{4}$$

Combining the definition of kernel function $K(x_i, x_j) = \left\langle \Phi(x_i) \cdot \Phi(x_j) \right\rangle$, the k-th kernel principal component of the data samples $\mathbf{x}_i$ in the original space within the feature space can be shown by (5):

$$t_k = \left\langle Q_k \cdot \Phi(\mathbf{x}) \right\rangle = \sum_{i=1}^{N} \alpha_i^k K(\mathbf{x}, \mathbf{x}_i) \tag{5}$$

For real-world applications, a criterion is established where the cumulative contribution surpasses 90%, indicating that selecting 5 to 6 principal components captures a substantial amount of the data's information. The dimensionality reduction procedure of KPCA is visually represented in Fig. 2.

C. Training and Testing of KPCA-SVM Prediction Model

Upon the application of KPCA for dimensionality reduction to the data samples, the predictive relationship between these condensed data points and the hydrocarbon reservoirs can be encapsulated through an SVM-based classification approach, as delineated below.

The data sample set is $\{(\mathbf{x}_i, \mathbf{y}_i), i = 1, ..., n\}$. where $\mathbf{x_i} \in R_d$ is the i-th data sample. $\mathbf{x}_i = (x_1, x_2, \cdots x_l)$ indicates the l-th KPCA data. And $\mathbf{y}_i = (y_1, y_2)$ is used to illustrate the corresponding hydrocarbon reservoir state, the details are shown in TABLE I.

TABLE I. hydrocarbon reservoir state of well

$\mathbf{y}_i(y_1, y_2)$	**Well State**
$\mathbf{y}_i(0, 0)$	hydrocarbon-free reservoir
$\mathbf{y}_i(0, 1)$	dry well
$\mathbf{y}_i(1, 0)$	low production
$\mathbf{y}_i(1, 1)$	high production

The predictive correlation function linking the data samples under investigation with the hydrocarbon reservoirs can be articulated through (6):

$$f(\mathbf{x}_i) = \mathbf{w} \cdot \Phi(\mathbf{x}_i) + b \tag{6}$$

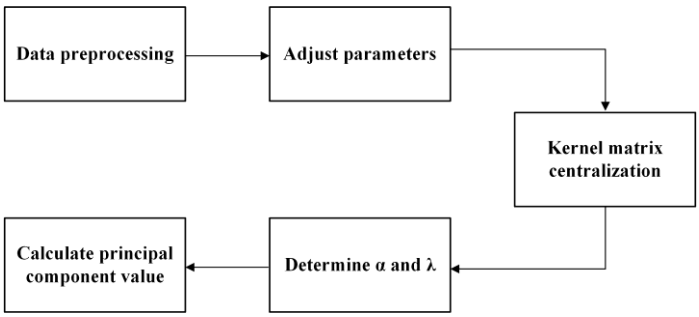

Fig. 2. The steps of KPCA Processing

In this equation, $\mathbf{w} \cdot \Phi(\mathbf{x}_i)$ represents the inner product between vectors $\mathbf{w}$ and $\Phi(\mathbf{x_i})$, where $\mathbf{w}$ is the dimension of a high-dimensional space, and b is the threshold, $b \in R$.

To solve $\mathbf{w}$ and b, relaxation variable $\xi, \xi^* \geq 0$ is introduced shown in (7).

$$\min_{\omega, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n} (\xi + \xi^*) \tag{7}$$

In this equation, C is the penalty factor. Constraints can be expressed by (8).

$$\begin{cases} \mathbf{y}_i - \mathbf{w} \cdot \Phi(\mathbf{x}_i) - b \leq \varepsilon + \xi_i \\ -\mathbf{y}_i + \mathbf{w} \cdot \Phi(\mathbf{x}_i) + b \leq \varepsilon + \xi_i^* \\ i = 1, ..., n \end{cases} \tag{8}$$

C affects the training error; the larger the value of C, the greater the penalty for samples with training errors greater than ξ. Based on Lagrange function, the kernel function perform the high-dimensional space inner product operations through operations in the original space, as shown in (9):

$$\mathbf{w} = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \Phi(\mathbf{x}_i) \tag{9}$$

The classification function of the SVM prediction model can be expressed as (10).

$$f(\mathbf{x}) = \mathbf{w} \cdot \Phi(\mathbf{x}) + b = \sum_{i=1}^{n} \left(\alpha_i - \alpha_i^*\right) K(\mathbf{x}_i, \mathbf{x}) + b \tag{10}$$

In this equation, $K(\mathbf{x}_i, \mathbf{x})$ is the Kernel function, Kernel function is a transformation function for sample data points, encompassing linear kernels, polynomial kernels, and Gaussian (RBF) kernels. When the corresponding Lagrange multiplier is non-zero, the associated sample is identified as a support vector. The structure of the SVM prediction model is shown in Fig. 3 .

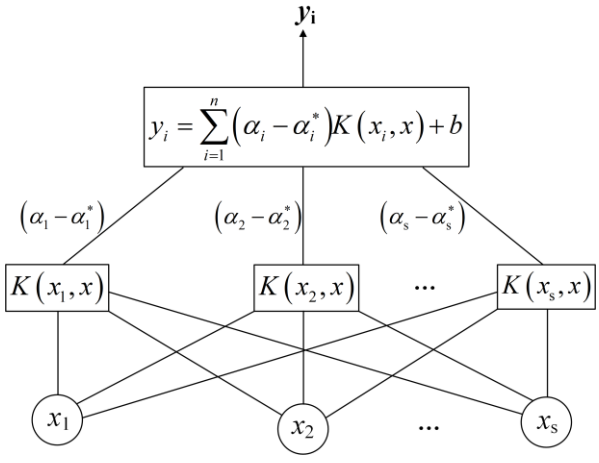

Fig. 3. Structure of SVM predicting model

As illustrated in Fig. 4, the process of developing and validating the KPCA-SVM predictive model is outlined. The initial phase involves training the model to identify an optimal parameter set, which includes decisions on the SVM algorithm, the kernel function, the penalty parameter *C*, and the error function threshold. Following this, a segment of the sample data is used for model training, which continues until the deviation between the model's predictions and actual values is within an acceptable limit. If the model's predictions do not meet the set criteria, the parameters are adjusted based on the observed discrepancies. In tandem, another subset of the sample data is employed to iteratively assess the model's performance, leading to the final refinement of the model's parameters.

The results of the Kernel Principal Component Analysis (KPCA) are depicted in Fig. 5. During the dimensionality reduction process facilitated by kernel principal components, the initial data samples are subjected to preprocessing to normalize the mean to zero. The illustration presents the count of kernel principal components, the informational load each component bears, and the aggregate informational content.

III. EXPERIMENTAL ANALYSIS

Focusing on actual seismic geological data from a specific area in Sichuan as the subject of analysis, the input samples are detailed in TABLE II. The KPCA algorithm is executed using MATLAB, and the SVM algorithm is run with the aid of the libSVM library. The data processing outcomes indicate that by incorporating six kernel principal components, a cumulative information retention rate of 95.83% is attainable, which ensures the preservation of the informational essence of the original data. To evaluate the impact of different kernel functions on recognition outcomes, samples from wells with odd numbers are assigned as training samples, while those from even-numbered wells are used as test samples. With all other parameters held constant, the recognition results for various kernel functions are presented in the accompanying table.

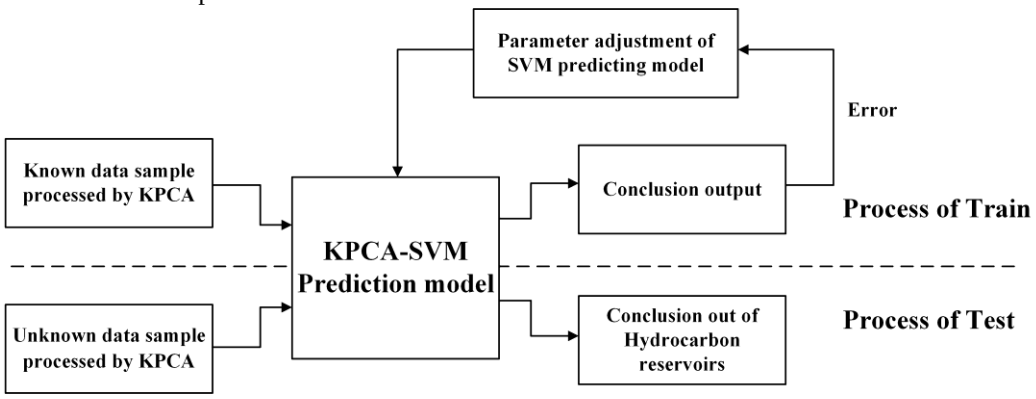

Fig. 4. The train and test flow chart of SVM analysis model

TABLE II. Sample table of seismic geological data

Num	Amp (mm)	Phase (mm)	Fre. (HZ)	Curvature (%)	Velocity (m/s)	Apparent polarity	Low velocity layer thickness	Production Remarks
1	1625	4	30	0.1	5142	1	20	Dry
2	1625	7	10	-0.07	4971	-1	30	Dry
3	1625	3	40	0.11	5142	1	40	High
4	2750	5	15	0.3	4800	-1	30	High
5	1625	5	20	0.16	5142	-1	40	High
6	1625	5	10	0.4	4971	-1	40	High
7	1625	5	15	0.15	4971	1	30	Low
8	1625	6	15	0.07	4971	-1	30	Low
9	1625	5	15	0.31	4971	1	40	Low
10	1625	5	30	0.31	4971	-1	30	High
11	2750	7	40	0.13	4971	-1	30	High
12	2750	3	35	0.1	5142	-1	30	Low
13	2750	7	15	0.24	5142	-1	30	High
14	2750	7	35	0.29	5314	1	30	Low
15	3875	6	10	0.07	4971	-1	30	High
16	2750	3	30	0.18	5142	1	20	Low
17	2750	6	15	0.1	5314	-1	20	Low

The experimental findings depicted in Fig. 6 highlight the substantial influence of the kernel function on the model's generalization capability and predictive classification performance. The selection of the kernel function significantly affects the outcomes of prediction and classification. The most favorable classification results are achieved when utilizing linear and polynomial kernel functions. The penalty factor *C*, a positive constant, determines the stringency of the penalty imposed on samples that exceed the specified error threshold. With all other parameters remaining unchanged, Fig. 7 illustrates the classification and recognition outcomes across a range of penalty factors.

Examining the outcomes depicted in Fig. 7 reveals that the impact of the penalty factor C on the analytical results wanes once it surpasses a specific threshold. This occurrence can be understood through the lens of structural risk minimization; an overly large *C* value leads to marginal empirical benefits. Therefore, *C* functions as a regulatory parameter that balances the SVM algorithm's capacity for generalization.

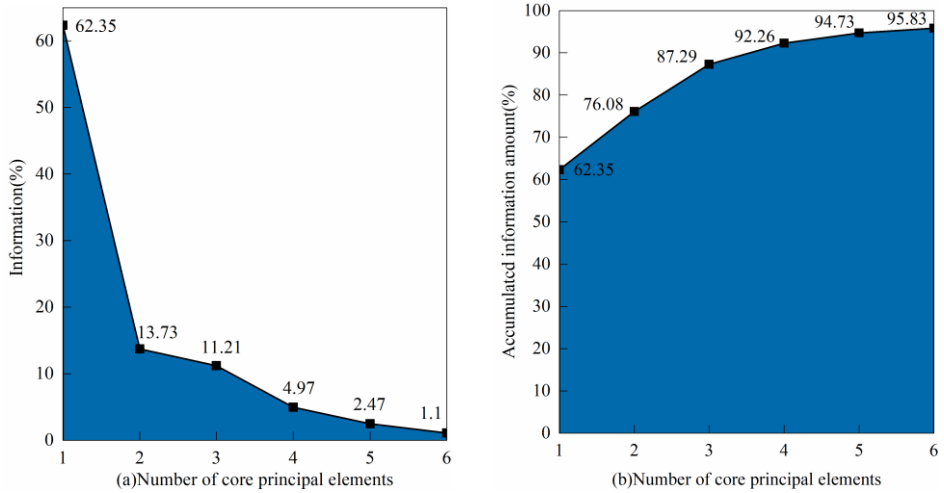

Fig. 5. Experimental results of KPCA

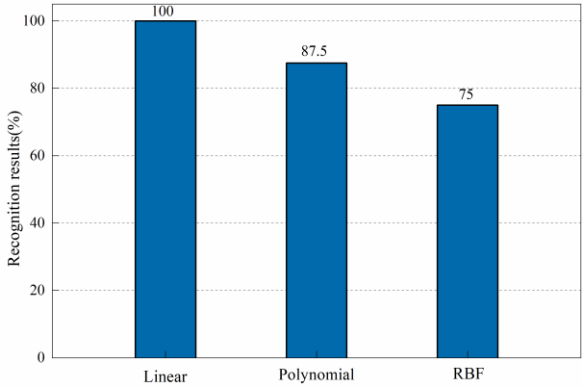

Fig. 6. Influence of different kernel functions on recognition results

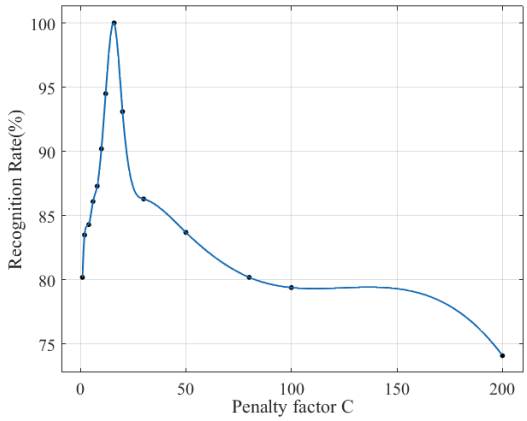

Fig. 7. Recognition rate of different penalty factors

IV. Conclusion

Support Vector Machines (SVM), an algorithm rooted in statistical learning theory, is designed to tackle the challenges of statistical learning that arise with limited data samples. It aims to attain the best generalization performance by balancing the learning capacity and the complexity of the model. In this study, the KPCA-SVM model is developed by combining the small-sample handling capabilities of SVM with the dimensionality reduction and feature extraction techniques of Kernel Principal Component Analysis (KPCA). The KPCA-SVM, which possesses strong nonlinear information extraction abilities, is suitable for the classification and prediction of hydrocarbon reservoirs and is especially fitting for predictive and classification tasks where data samples are scarce. Experiments conducted with actual data samples assess the effects of kernel functions, the penalty factor *C*, and the quantity of data samples on the predictive results, offering an optimal range for the selection of the model's key parameters. The experimental results indicate that the KPCA-SVM model outperforms other models and achieves higher classification accuracy, thus possessing substantial practical application value.

References

[1] S.G.Zhao Impact of Global Oil and Gas Industry's Changes on China's Energy Security and Corresponding Proposals during"14th FiveYear Period"[J]. Development Research, 2020(04):40-44.

[2] D.ALAIGBA, T.A.AWANI, Z.OYELEKE. Exploring the Delicate Balance in the Optimal Execution of Concurrent Gas Cap Blowdown and Oil Rim Production in a Saturated Reservoir[C]//SPE Nigeria Annual International Conference and Exhibition: August 5–7, 2024 Lagos, Nigeria. 2024:1-16.

[3] N. N. MIKHAILOV, V. A. KUZ'MIN, K. A. MOTOROVA, et al. The Effect of the Pore Space Microstructure on Hydrophobization of Oil and Gas Reservoirs[J]. Moscow University geology bulletin,2016,71(6):436-444.

[4] F.X.Zhang, X.Q. Zheng, Z.B.Li et.al. Practice of drilling optimization system in the development of unconventional oil and gas resources in China[J]. China Petroleum Exploration, 2020,25(02):96-109.

[5] HAJIREZAIE, SASSAN, WU, XINGRU, SOLTANIAN, MOHAMAD REZA, et al. Numerical simulation of mineral precipitation in hydrocarbon reservoirs and wellbores[J]. Fuel,2019,238(Feb.15):462-472.

[6] LIU L., MEHANA M., CHEN B., et al. Reduced-order models for the greenhouse gas leakage prediction from depleted hydrocarbon reservoirs using machine learning methods[J]. International Journal of Greenhouse Gas Control,2024,132.

[7] WILSON, M. J., SHALDYBIN, M. V., WILSON, L.. Clay mineralogy and unconventional hydrocarbon shale reservoirs in the USA. I. Occurrence and interpretation of mixed-layer R3 ordered illite/smectite[J]. Earth-Science Reviews: The International Geological Journal Bridging the Gap between Research Articles and Textbooks,2016,15831-50.

[8] G.Z.Liao, Y.Z.Li, L.Z.Xiao Prediction of microscopic pore structure of tight reservoirs using convolutional neural network model[J].Petroleum Science Bulletin. 2020,5(01):26-38.

[9] SYED, FAHAD IQBAL, MUTHER, TEMOOR, DAHAGHI, AMIRMASOUD KALANTARI, et al. Low-Rank Tensors Applications for Dimensionality Reduction of Complex Hydrocarbon Reservoirs[J]. 2022,244(Apr.1 Pt.A):122680.1-122680.9.

[10] BIN YAO. Analysis of the Key Factors of Pumping Well System Efficiency for Oil Field Based on Multiple Regression[J]. IOP Conference Series:Earth and Environmental Science,2021,661(1).

[11] SERGEY.GORBACHEV, VLADIMIR.SYRYAMKIN. Adaptive Neuro-Fuzzy Recognition Technology Intersecting Objects[J]. Applied Mechanics and Materials,2015,3793(1512):683-688.

[12] LEI, DAJIANG, WU, KAILI, ZHANG, LIPING, et al. Neural ordinary differential grey model and its applications[J]. Expert Systems with Application,2021,177(Sep.):114923.1-114923.7.

[13] U.M-P.JOHN, A.ADEBAYO, D.O.ANOMNEZE. Locating the Remaining Oil in a Re-Saturated Gas Cap Reservoir in a Brown Field A Case Study of a Niger Delta Reservoir[C]//SPE Nigeria Annual International Conference and Exhibition: August 5–7, 2024 Lagos, Nigeria. 2024:1-14.

[14] S S RISWATI, S IRHAM, NULL RENDY, et al. Surfactant technology for improved hydrocarbon recovery in unconventional liquid reservoirs: a systematic literature review[J]. IOP Conference Series:Earth and Environmental Science,2023,1239(1).

[15] SHOUCHUN WANG, XIUCHENG DONG, RENJIN SUN. Predicting saturates of sour vacuum gas oil using artificial neural networks and genetic algorithms[J]. Expert Systems with Application,2010,37(7).

[16] MOHAMMED A.KHAMIS, K.A.FATTAH. Estimating oii-gas ratio for volatile oil and gas condensate reservoirs: artificial neural network,support vector machines and functional network approach[J]. Journal of Petroleum Exploration and Production Technology,2019,9(1):573-582.

[17] Lahdhiri H, Taouali O. Reduced Rank KPCA based on GLRT chart for sensor fault detection in nonlinear chemical process [J].Measurement, 2020, 169:108342.

[18] NAKAYAMA, YUGO, YATA, KAZUYOSHI, AOSHIMA, MAKOTO. Clustering by principal component analysis with Gaussian kernel in high-dimension, low-sample-size settings[J]. Journal of Multivariate Analysis: An International Journal,2021,185.

[19] HEIDARY, MOHAMMAD. The use of kernel principal component analysis and discrete wavelet transform to determine the gas and oil interface[J]. Journal of geophysics and engineering,2015,12(3):386-399.

[20] ABDALLAH BASHIR MUSA. A comparison of ?1-regularizion, PCA, KPCA and ICA for dimensionality reduction in logistic regression[J]. International journal of machine learning and cybernetics,2014,5(6):861-873.