Petroleum Science 21 (2024) 1632–1648

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

Identification of reservoir types in deep carbonates based on mixed-kernel machine learning using geophysical logging data

Jin-Xiong Shi [a, b, c], Xiang-Yuan Zhao [d], Lian-Bo Zeng [c, *], Yun-Zhao Zhang [e], Zheng-Ping Zhu [a, b, **], Shao-Qun Dong [c]

[a] Key Laboratory of Exploration Technologies for Oil and Gas Resources of the Ministry of Education, Yangtze University, Wuhan, 430100, Hubei, China
[b] School of Geosciences, Yangtze University, Wuhan, 430100, Hubei, China
[c] College of Geosciences, China University of Petroleum (Beijing), Beijing, 102249, China
[d] Petroleum Exploration and Production Research Institute of SINOPEC, Beijing, 100083, China
[e] School of Earth Sciences and Engineering, Xi'an Shiyou University, Xi'an 710065, Shaanxi, China

ARTICLE INFO

Article history:
Received 18 June 2022
Received in revised form
20 September 2023
Accepted 21 December 2023
Available online 30 December 2023

Edited by Jie Hao and Meng-Jiao Zhou

Keywords:
Reservoir type identification
Geophysical logging data
Kernel Fisher discriminant analysis
Mixed kernel function
Deep carbonates

ABSTRACT

Identification of reservoir types in deep carbonates has always been a great challenge due to complex logging responses caused by the heterogeneous scale and distribution of storage spaces. Traditional cross-plot analysis and empirical formula methods for identifying reservoir types using geophysical logging data have high uncertainty and low efficiency, which cannot accurately reflect the nonlinear relationship between reservoir types and logging data. Recently, the kernel Fisher discriminant analysis (KFD), a kernel-based machine learning technique, attracts attention in many fields because of its strong nonlinear processing ability. However, the overall performance of KFD model may be limited as a single kernel function cannot simultaneously extrapolate and interpolate well, especially for highly complex data cases. To address this issue, in this study, a mixed kernel Fisher discriminant analysis (MKFD) model was established and applied to identify reservoir types of the deep Sinian carbonates in central Sichuan Basin, China. The MKFD model was trained and tested with 453 datasets from 7 coring wells, utilizing GR, CAL, DEN, AC, CNL and RT logs as input variables. The particle swarm optimization (PSO) was adopted for hyper-parameter optimization of MKFD model. To evaluate the model performance, prediction results of MKFD were compared with those of basic-kernel based KFD, RF and SVM models. Subsequently, the built MKFD model was applied in a blind well test, and a variable importance analysis was conducted. The comparison and blind test results demonstrated that MKFD outperformed traditional KFD, RF and SVM in the identification of reservoir types, which provided higher accuracy and stronger generalization. The MKFD can therefore be a reliable method for identifying reservoir types of deep carbonates.

1. Introduction

Carbonate rocks possess abundant hydrocarbon resources and production around the world. In recent years, carbonate reservoirs in the deep layer of petroliferous basins have become increasingly important exploration targets worldwide (Katz and Everett, 2016; Zhu et al., 2019; Mahdaviara et al., 2020). Due to complex diagenesis and tectonism during the long burial progress, various storage spaces, such as the pores, cavities and fractures, were intricately developed within the deeply buried carbonates and formed multiple types of reservoirs (Lu et al., 2017; Tian et al., 2019; Souvik et al., 2021). Different types of carbonate reservoirs exhibit distinct storage and percolation capacities over a range of scales, resulting in significant heterogeneity of reservoir quality and discrepancy of hydrocarbon enrichment within carbonate reservoirs, which poses potential risks for effective exploitation (Matonti et al., 2015; Zhang et al., 2015). Therefore, the accurate identification of reservoir types is one of key issues for the reservoir quality evaluation and hydrocarbon rational development in deep carbonates, meanwhile it is always full of challenges.

For underground reservoirs, drilling cores and geophysical logging data are commonly used in the identification of reservoir types (Tian et al., 2019; Lan et al., 2021; Souvik et al., 2021). Core

[*] Corresponding author.
[**] Corresponding author.
E-mail addresses: lbzeng@sina.com (L.-B. Zeng), 501096@yangtzeu.edu.cn (Z.-P. Zhu).

https://doi.org/10.1016/j.petsci.2023.12.016
1995-8226/

description is the most direct and efficient approach to identify reservoir types, but it cannot be widely used due to the low coring rate and time-consuming. Image logs have high longitudinal continuity and resolution, and can provide abundant information for reservoir type identification (Tian et al., 2019; Lan et al., 2021; Zheng et al., 2021). However, these data are not always available from all wells due to the high cost. By contrast, the conventional logging data is more prevalent, economic and reliable. Conventional logging measures the physicochemical property of rocks around the wellhole by acoustical, electrical, and radioactive detections (Tokhmchi et al., 2010; Ghosh et al., 2016), which has wide applications in the recognition and prediction of lithology, physical properties, mechanical properties, and hydrocarbon-bearing properties (Khalifah et al., 2019; Lan et al., 2021; Méndez et al., 2021; Zheng et al., 2021; Dong et al., 2022a). By utilizing conventional logging data, traditional methods generally adopt the cross-plot analysis, empirical formula, and interpretation chart techniques to identify reservoir types. Most methods are greatly dependent on geologists' experience, and have characteristics of low efficiency and high subjectivity. Furthermore, these methods fail to attain anticipated accuracy due to the complex logging responses. Results from above practices suggest that the reservoir type identification based on well logs is not a simple linear classification problem.

With the rapid development of artificial intelligence in recent years, machine learning techniques have been widely applied in many fields of petroleum geology with highly promising results (Ghosh et al., 2016; Zhang et al., 2021; Liu et al., 2022). Hitherto, the supervised learning algorithm has been the hottest research focus and shown its advantages in handling complex classification and regression problems (Jordan and Mitchell, 2015; Dong et al., 2019; Liu et al., 2020a,b; Shi et al., 2023). Numerous supervised learning classifiers have been introduced to well logging interpretation, such as the support vector machine (SVM), decision tree (DT), artificial neural network (ANN), and random forest (RF). The kernel Fisher discriminant analysis (KFD), a kernel-based multi-classification tool, presents remarkable ability to capture nonlinear relationships among high-dimensional variables (Billings and Lee, 2002; Xu et al., 2004; Dong et al., 2022a). Typically, KFD maps the linearly non-separable data from original space (low-dimensional) into an implicit nonlinear feature space (high-dimensional) by using kernel function (Xu et al., 2004; Dong et al., 2016). This mapping effectively converts the nonlinear relationship into a linear separable one, and realizes classification through a hyperplane. Taking advantages of kernel functions, KFD method can extract more hidden features and potential relations form input data and improve the classification precision, with successful applications in lithology, fracture and coal structure prediction (Dong et al., 2016, 2020; Shi et al., 2020).

The performance of KFD largely depends on kernel function. Geometric shapes of different kernel functions vary greatly and determine the data distribution in feature space (Xu et al., 2004; Pilario et al., 2019). Generally, kernel functions used by KFD classifier include the local and global kernels. The local kernel presents high interpolation capability, whereas the global kernel possesses strong extrapolation ability (Brailovsky et al., 1999; Sridevi, 2018). When dealing with practical problem, KFD usually adopts only a single kernel function. The selection of kernel function theoretically is dependent on features of input data (Zhu et al., 2012; Xu et al., 2015; Chen et al., 2018), which are usually unknown. In most cases, with the same experimental data, different types of kernels are individually applied and verified to achieve the best prediction result, which is cumbersome and prone to error. Limited by inherent features of the local and global kernels, KFD model based a single kernel cannot well interpolate and extrapolate at the same time. Furthermore, complex nonlinear problems complicate the application of a single-kernel based KFD method. Recent studies indicate that the mixed-kernel function, a combination of the local and global kernels, has dramatical potentiality to improve the overall model performance (Xu et al., 2015; Pilario et al., 2019). A mixed-kernel based KFD model has both good interpolation and extrapolation abilities, and presents property superior to single-kernel based models. To date, the application of mixed-kernel based KFD method in the reservoir type identification using well logging data has not been reported.

To improve the identification accuracy, a new mixed kernel Fisher discriminant analysis (MKFD) method was proposed in this paper for identifying reservoir types in deep carbonates using geophysical logging data. In this work, datasets obtained from the Sinian carbonates of the Sichuan Basin were utilized to establish the MKFD model. By comparing modeling results of MKFD model with these of traditional KFD, RF and SVM models, the improved performance of the proposed method was evaluated and demonstrated. In addition, a blind well test was conducted to verify the validity and reliability of the built MKFD model. Furthermore, a variable importance analysis was implemented to better understand the geological implication of the MKFD result. Finally, future work was discussed to expand the application of proposed MKFD identification method in petroleum geology.

2. Geological setting and data analysis

2.1. The study area

In this work, datasets including core samples and well logging data were collected from the Sinian dolomite reservoirs of central Sichuan Basin, southwest China (Fig. 1). The Upper Sinian Dengying Formation (Fig. 1), target layer of this study, is currently the main exploitation and production layer of natural gas in this area. The present burial depth of the Dengying Formation ranges from approximately 4700–5500 m, with thickness of 260–350 m (average ~300 m). The Dengying Formation was deposited in a shallow-water carbonate platform sedimentary environment as the Upper Yangtze Platform entered into the rift valley filling stage after Nanhua glacial period. Sedimentary facies mainly include the restricted platform, platform margin, platform marginal slope and platform basin facies (Zhou et al., 2016). The dominant sediments in the interior platform and platform margin are dolomites. The lithology of Dengying Formation consists mainly of granular dolomite, crystalline dolomite and microbial dolomite (Zhou et al., 2016; Shi et al., 2022). Core plug analysis indicates that these dolomite rocks have low porosity ranging from 0.13% to 8.59% (average 3.06%), and low permeability ranging between 0.0001 and 25.3×10^{-3} μm^2 (average 0.59×10^{-3} μm^2) (Zhou et al., 2020).

2.2. Data acquisition

In the central Sichuan Basin, more than 60 wells have penetrated the Dengying Formation, which provide a certain amount of core samples and large numbers of conventional logging data. Totals of 103.5 m of core samples were collected from 8 typical wells with depths ranging 4900–5300 m. In addition, approximately 300 samples were selected from coring intervals for thin-section preparation. Thin sections were in sizes of 20 mm × 20 mm × 30 μm, and were impregnated with blue-epoxy resin. The microscopic observation was performed on a Leica DM4500 microscope. In the study area, conventional logging data include the caliper (CAL), gamma ray (GR), bulk density (DEN), acoustic velocity (AC), compensated neutron (CNL), and true formation resistivity (RT) logs. The sampling interval of well logs is 0.125 m.

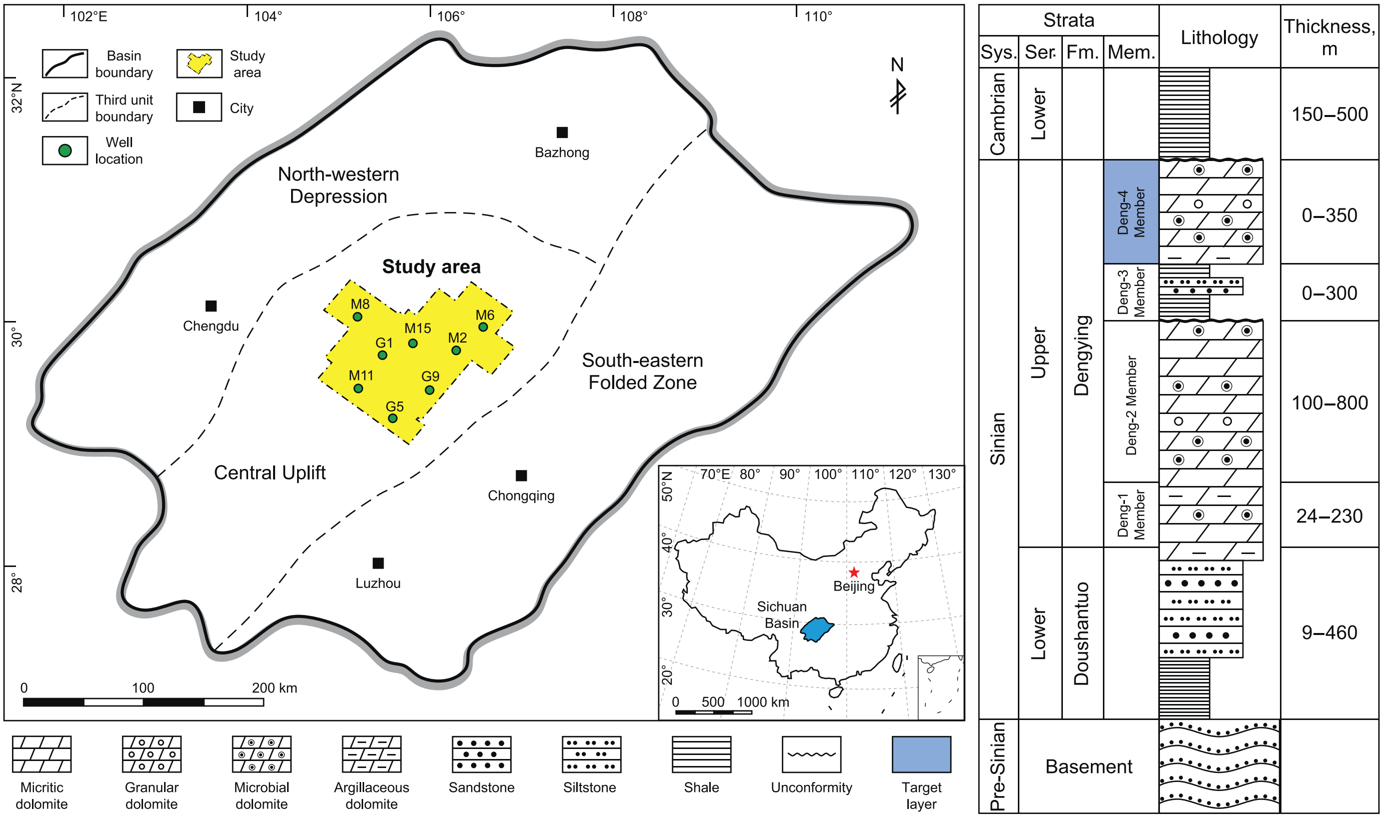


Fig. 1. Locations of the Sichuan Basin and study area, and stratigraphic characteristics of Sinian Formation.

2.3. Type and characteristic of carbonate reservoirs

Due to the compaction and cementation during long burial progress, extremely few primary pores were preserved in the current matrix of Dengying carbonate reservoirs. Simultaneously, influenced by multistage karstification and tectonism, various types of secondary pores and natural fractures were intricately developed within these carbonate rocks, which were the dominant reservoir spaces in the Dengying reservoirs. Core and thin section observations show that secondary pores mainly comprise dissolution pores (diameter ≤2.0 mm) and caves (diameter >2.0 mm) (Zhou et al., 2020), and natural fractures are dominated by tectonic fractures and solution-enlarged fractures. It is noticed that reservoir spaces refer to pores and fractures that are not full-filled by dolomite or quartz minerals in this paper. Carbonate reservoirs in the Dengying Formation can be divided to following five types according to the type and assemblage of storage spaces, and petrophysical properties of reservoirs (Table 1).

Pore reservoirs (Type I): Storage spaces in pore reservoirs are mainly dissolution pores and a few primary pores, while dissolution caves and natural fractures are rarely developed (Table 1(a)). The dissolution pores are composed mainly of the intercrystalline, intragranular, intergranular dissolution pores. These dissolution pores commonly present isolated and scattered distribution (Table 1(b)), with pore diameters ranging from 0.6 to 1.3 mm. Core plug measurements show that the porosity and permeability of these reservoirs are generally less than 3.0% and 0.01×10^{-3} μm^2, respectively.

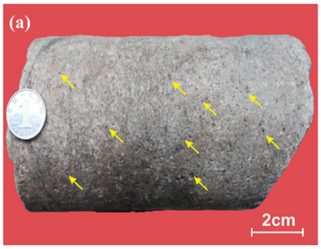
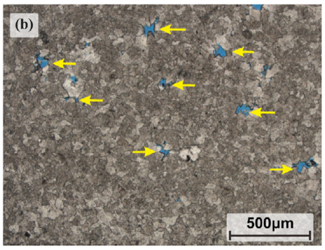
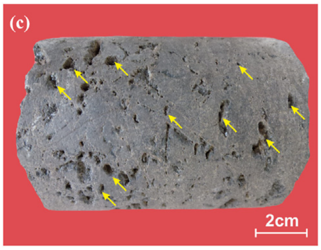
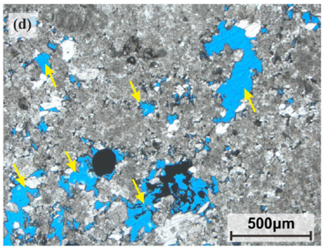
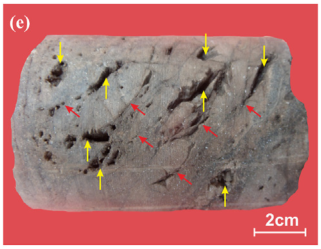
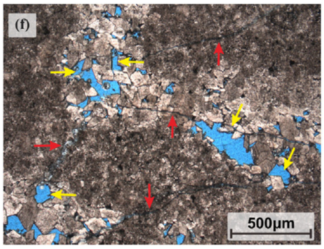
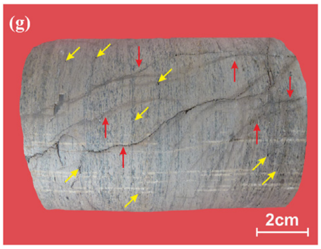
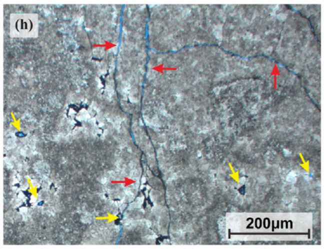
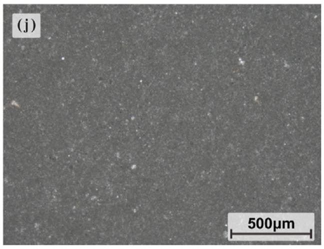
Pore-cavern reservoirs (Type II): In the pore-cavern reservoirs, dissolution caves are the dominant reservoir spaces, followed by dissolution pores (Table 1(c)), which are closely associated with the karstification. The majority of dissolution caves and pores are irregular with varied sizes, and layered or distributed along bedding planes. Commonly, these dissolution pores and caves are well developed and intensely distributed (Table 1(d)), resulting in relatively high porosity (>3.0%) and permeability (>0.1×10^{-3} μm^2).

Fractured-cavern reservoirs (Type III): Storage spaces in the fractured-cavern reservoirs consist mainly of dissolution caves, fractures and some dissolution pores (Table 1(e)). High-angle tectonic fractures are well developed and parts of them are solution-enlarged. The dissolution pores and caves are usually distributed beaded along fractures or parallel to bedding planes, forming large-scale fracture-cave systems with high connectivity (Table 1(f)). Hence, fractured-cavern reservoirs generally present high petrophysical properties, with porosity and permeability more than 3.0% and 0.5×10^{-3} μm^2, respectively.

Fractured-pore reservoirs (Type IV): Storage spaces in fractured-pore reservoirs compose dissolution pores and tectonic fractures (Table 1(g)), while caves and solution-enlarged fractures are inferior-developed. The dissolution pores are mostly scattered (Table 1(h)), and parts of them are connected by fractures. Most of fractures are tectonic ones and different groups of them usually form fracture networks (Table 1(h)). Core plug analysis indicates that these reservoirs have characteristics of low porosity (3.0%) and relatively high permeability (>0.1×10^{-3} μm^2).

Tight reservoirs (Type V): These reservoirs refer to carbonate bedrocks in the Dengying Formation, where the majority of primary pores were destroyed by strong compaction and cementation during the burial progress (Table 1(i)). Extremely few secondary pores and natural fractures are developed in these reservoirs (Table 1(g)). Hence, tight reservoirs almost have few actual storage

Table 1
Characteristics of different reservoir types in the Sinian Dengying carbonates, central Sichuan Basin.

Reservoir types	Storage spaces	Petrophysical properties	Cores[a]	Thin sections[b]
Pore reservoirs (Type I)	Dominated by dissolution pores; caves and fractures are undeveloped	Por <3.0%, Perm <0.01 × 10^{-3} μm^2		
Pore-cavern reservoirs (Type II)	Dominated by dissolution caves and some dissolution pores; fractures are undeveloped	Por >3.0%, Perm >0.1 × 10^{-3} μm^2		
Fractured-cavern reservoirs (Type III)	Dissolution caves, fractures and some dissolution pores in high connectivity	Por >3.0%, Perm >0.5 × 10^{-3} μm^2		
Fractured-pore reservoirs (Type IV)	Dissolution pores and tectonic fractures (forming fracture network)	Por <3.0%, Perm >0.1 × 10^{-3} μm^2		
Tight reservoirs (Type V)	Carbonate bedrocks; primary pores, secondary pores and natural fractures are rarely developed	Por <2.0%, Perm <0.001 × 10^{-3} μm^2		

[a] Macroscopic characteristics of different reservoir types in the Dengying carbonates. **(a)** Dissolution pores, M8, 5011.73 m; **(c)** dissolution caves are well developed, G5, 5223.56 m; **(e)** dissolution caves (yellow arrow) and fractures (red arrow) form high connectivity systems, G5, 5329.12 m; **(g)** dissolution pores (yellow arrow) and tectonic fractures (red arrow) developed, M11, 5138.89 m; **(i)** tight dolomite, G1, 5032.97 m.

[b] Microscopic characteristics of different reservoir types in the Dengying carbonates. **(b)** Interparticle and intercrystalline dissolution pores, G5, 5125.42 m; **(d)** M11, 5137.54 m; **(f)** dissolution caves and pores connected by fractures with good connectivity, M11, 5123.81 m; **(h)** Fracture network interconnecting dissolution pores, M6, 5128.38 m; **(j)** no storage spaces exist in the matrix, G9, 5188.23 m. Por = porosity; Perm = permeability.

and seepage capability. The porosity and permeability of core samples are generally less than 2.0% and 0.001 × 10^{-3} μm^2, respectively.

2.4. Logging response of carbonate reservoirs

After depth matching of core samples to well logs according to core description reports, datasets including 6 well logs (GR, CAL, DEN, AC, CNL and RT) and corresponding reservoir types was collected from 7 coring wells in the Dengying Formation. The logging data were selected from multiple sets of reservoirs at different depths to avoid sample specificity and were mainly concentrated in the middle section of each interval with different types to eliminate the interference of boundary effect. Further, to guarantee the validity, raw data was statistically evaluated to remove abnormal values in well logs. According to the three-sigma criterion, all values of well logs outside the range of ±3.0 standard deviation (SD) were considered as outliers and removed. After the pre-processing, totals of 453 datasets were obtained from raw data and were used to train and test the MKFD model, including 106 pore reservoirs, 75 pore-cavern reservoirs, 68 fractured-cavern reservoirs, 86 fractured-pore reservoirs, and 118 tight reservoirs datasets. Value ranges of well logs for each reservoir type are shown in Table 2.

As shown in Fig. 2, logging responses of different reservoir types are diverse. From the perspective of average value, the DEN and RT logs gradually decrease in proper order of tight reservoirs, pore reservoirs, fractured-pore reservoirs, pore-cavern reservoirs, and fractured-cavern reservoirs, while the GR, CAL, AC and CNL logs present an increasing trend. In deep carbonate reservoirs, the massive presence of pores and caves can reduce the rock weight per unit volume, protract the sonic transit time and supply abundant spaces for gas accumulation (Tian et al., 2019; Dong et al., 2022a), resulting in low DEN and high AC, CNL values. The development of fractures can also generate the same logging responses of DEN, AC and CNL as pores and caves since fractures provide certain storage spaces (Aghli et al., 2016; Dong et al., 2022b), whereas the degree of fracture-related logging responses is much inapparent than that of the latter two. In the meantime, well-developed open fractures are profit to the invasion of drilling fluids because they service as effective seepage paths in tight rocks (Lyu et al., 2016; Lai et al., 2022), which leads to relatively low RT values in fractured rocks. In addition, the mechanic strength of rocks will be reduced when pores, caves and fractures are abundantly existent, where there is a higher potential of hole enlargement (Tokhmchi et al., 2010). These logging response characteristics for carbonate reservoirs are basically consistent with the results of previous research (Lu et al., 2017; Feng et al., 2021; Lan et al., 2021; Zheng et al., 2021). Nonetheless, there are considerable overlaps of well-log values among different reservoir types, indicating that the feature information associated with classification is concealed by those irrelevant but stronger.

To reflect the variation of logging responses more intuitively, three-dimensional cross-plots were generated (Fig. 3). It can be found that some reservoir types possess the similar log responses when distinguished by three-dimensional cross-plot. The overlaps of well logging values are still obvious, and decision boundaries of different reservoir types are obscure. Those overlaps will cause multiple interpretation problems in the reservoir type identification. All of these indicate that there is complex nonlinear relation between reservoir types and well logs. Hence, it is crucial to employ appropriate nonlinear method to analyze multi-variate dataset and extract meaningfully geological features for identifying reservoir types.

3. Methodology

3.1. Kernel Fisher discriminant analysis

Kernel Fisher discriminant analysis (KFD) is a supervised machine learning algorithm proposed by Mika et al. (1999). The KFD is on basis of the linear discriminant analysis (LDA) and kernel method, and possesses unique advantages in handling problems with small samples and high nonlinearity (Baudat and Anouar, 2000; Chu et al., 2011). The basic idea of KFD is projecting the original input data from low-dimensional space R (usually nonlinear space) into high-dimensional feature space F (linear space) through nonlinear mapping $\varphi(\bullet)$. In the feature space F, nonlinear relations of input data are indirectly transformed into linear relations, and thereby the linear decision boundary for classification can be extracted.

For a given dataset $X=\{x_1,x_2,\dots,x_n\}$ with C classes, $\varphi(X)$ is the images of X under the map φ. Mathematically, the purpose of KFD is searching projection vectors that minimize the scatter within classes and maximize the scatter between each class in the feature space F. It can be formulated as maximizing the Fisher criterion function (Baudat and Anouar, 2000):

$$J^{\varphi}(\boldsymbol{\alpha})=\frac{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{b}}^{\varphi}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{w}}^{\varphi}\boldsymbol{\alpha}} \tag{1}$$

where $\boldsymbol{S}_{\mathrm{w}}^{\varphi}$ is the within-class scatter matrix, $\boldsymbol{S}_{\mathrm{b}}^{\varphi}$ is the between-class scatter matrix, and $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\dots,\alpha_n)^{\mathrm{T}}$ is the projection vector in F.

The optimal problem of Eq. (1) can be solved by the generalized characteristic equation:

$$\boldsymbol{S}_{\mathrm{b}}^{\varphi}\boldsymbol{\alpha}=\lambda\boldsymbol{S}_{\mathrm{w}}^{\varphi}\boldsymbol{\alpha} \tag{2}$$

where $\lambda=(\lambda_1,\lambda_2,\dots,\lambda_n)$ is the nonzero eigenvalue of $\boldsymbol{\alpha}$, representing the classification capacity of corresponding projection vector.

According to the reproducing kernel theory, the solution $\boldsymbol{\alpha}F$ must lie within the sample span in F. Hence, $\boldsymbol{\alpha}$ can be expressed as a linear combination of $\varphi(x_i)$:

$$\boldsymbol{\alpha}=\sum_{i=1}^{n}\alpha_i\varphi(x_i) \tag{3}$$

Thus, the discriminant function $f^{\varphi}(\varphi)$ in feature space can be denoted as Eq. (4):

$$f^{\varphi}(\varphi)=\alpha^{\mathrm{T}}\varphi(x)=\sum_{i=1}^{n}\alpha_i(\varphi(\boldsymbol{x}_i),\boldsymbol{x}) \tag{4}$$

Normally, the number of projection vectors m is less than the number of classes C, i.e., $m\le C-1$. To determine the number of projection vectors, the contribution rate is usually adopted. For instance, let $\boldsymbol{\alpha}_{\mathrm{opt}}=(\alpha_1,\dots,\alpha_m)^{\mathrm{T}}$ be the optimal projection vector,

Table 2
Range of well log values for different reservoir types in Dengying carbonates, central Sichuan Basin.

Reservoir types	GR, API	CAL, in	AC, μs/m	DEN, g/cm^3	CNL, %	RT, Ω·m
Pore reservoirs	3.9–10.6 (6.8)	6.03–6.45 (6.25)	43.0–48.5 (45.7)	2.704–2.858 (2.772)	1.20–4.47 (2.79)	2883–37,277 (16,329)
Pore-cavern reservoirs	4.3–12.2 (7.5)	6.11–6.47 (6.26)	43.5–50.4 (46.4)	2.668–2.849 (2.758)	1.13–5.38 (3.02)	2438–26,312 (9836)
Fractured-cavern reservoirs	4.0–12.8 (7.8)	6.07–6.57 (6.29)	44.6–49.6 (46.7)	2.643–2.838 (2.739)	1.33–5.10 (3.20)	2209–21,432 (7847)
Fractured-pore reservoirs	4.2–11.6 (7.6)	6.03–6.51 (6.27)	43.4–48.6 (45.9)	2.686–2.842 (2.763)	1.19–4.20 (2.69)	2513–24,098 (9577)
Tight reservoirs	3.2–10.7 (6.4)	6.00–6.36 (6.23)	42.6–47.4 (44.7)	2.696–2.875 (2.795)	0.79–3.92 (2.28)	2893–35,598 (18,243)

Note: 3.9–10.6 (6.8) = minimum–maximum (mean).

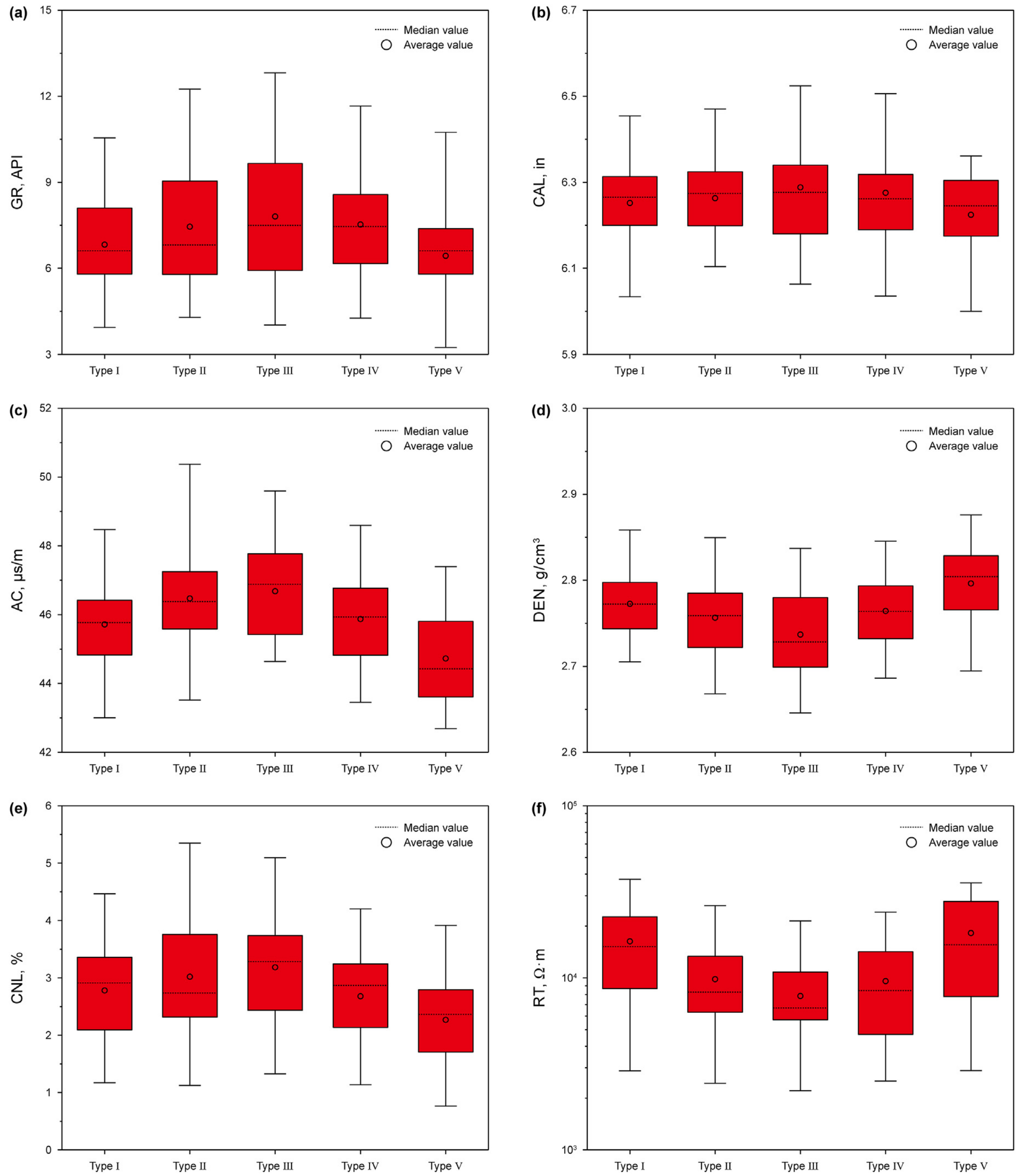


Fig. 2. Boxplots of well logs for different reservoir types. Type I–V correspond to pore reservoirs, pore-cavern reservoirs, fractured-cavern reservoirs, fractured-pore reservoirs and tight reservoirs, respectively.

and $\lambda_1, \lambda_2, \ldots, \lambda_m (\lambda_1 \geq \ldots \geq \lambda_m > 0)$ are the eigenvalue of $\alpha_1, \ldots, \alpha_m$ respectively. The number of projection vectors m can be determined based on the cumulative contribution rate $\sum_{i=1}^{m} \lambda_i / \sum_{i=1}^{n} \lambda_i \geq$ *threshold* (Lei et al., 2019). Typically, the value of *threshold* is more than 0.85, and it is set to 0.95 in this work.

3.2. Mixed kernel Fisher discriminant analysis

In the KFD, various nonlinear relationships must be considered to design a $\varphi(\cdot)$ that guarantees the accuracy of mapping process. Simultaneously, it inevitably leads to a high dimensionality of the feature space. To avoid specifying $\varphi(\cdot)$ explicitly, the kernel function $K(x_i, x_j)$ is introduced to simplify the dot-product computation of transformed data $\langle \varphi(x_i), \varphi(x_j) \rangle$ in the feature space. Thus, the need for specifying $\varphi(\cdot)$ is eliminated since the nonlinear mapping is implicitly realized by kernel function, called the kernel trick (Liu et al., 2004). Hence, the discriminant function $f^{\varphi}(\varphi)$ in the feature space F is converted to

$$f^{\varphi}(\varphi) = f(x) = \sum_{i=1}^{n} \beta_i k(\boldsymbol{x}_i, \boldsymbol{x}) \tag{5}$$

where $\boldsymbol{x}_i$ is the ith one of all input samples, β_i is coefficient vector of

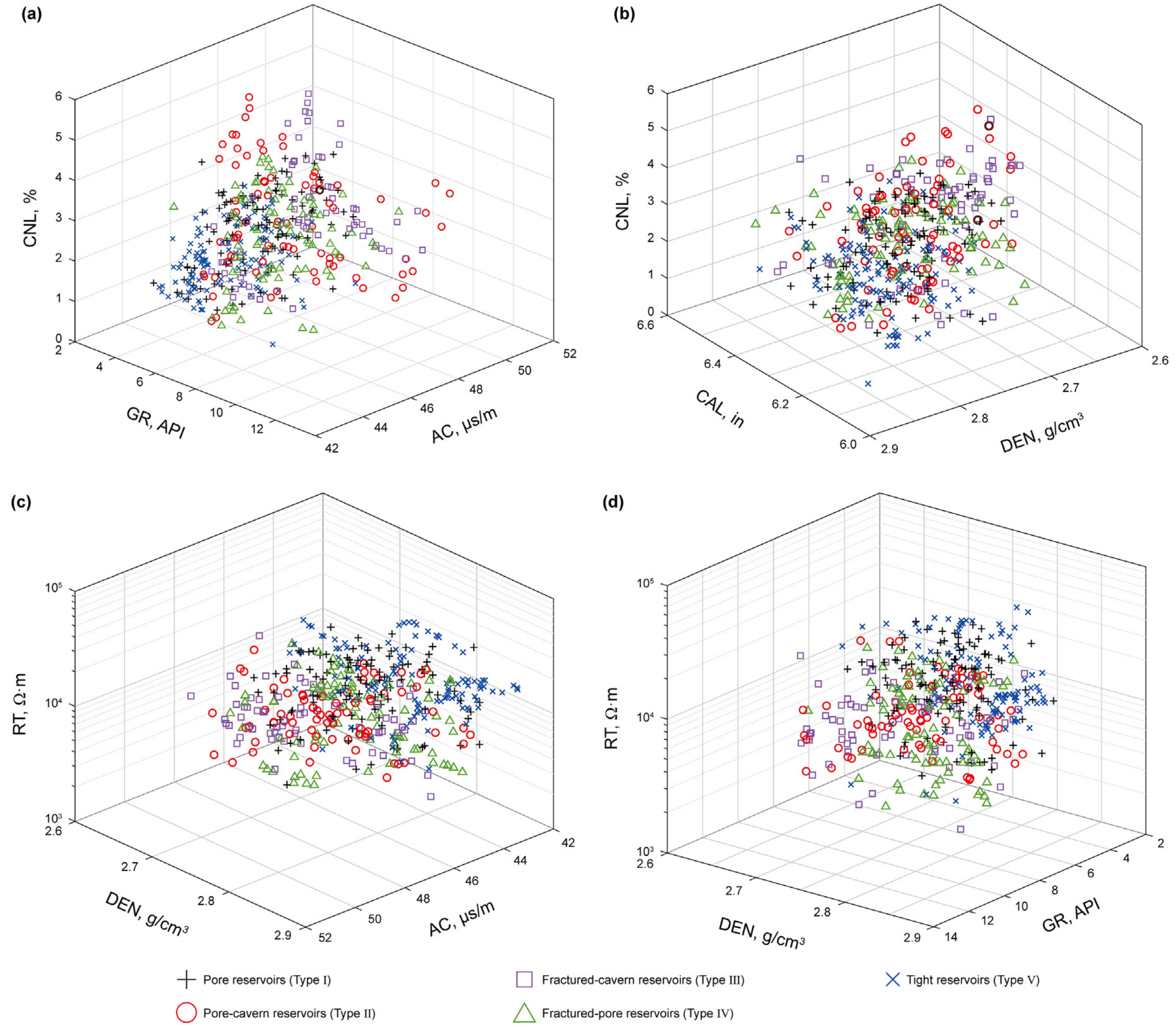


Fig. 3. Three-dimensional crossplots of well logs against different reservoir types.

the *i*th kernel, and *k* is kernel function.

The kernel function must satisfy the Mercer's condition, and common kernel functions used in classification problems include the linear, radial basis function (RBF), polynomial, and sigmoid kernel functions (Table 3). Different kernel functions have various characteristics and unique geometric properties, which determines the nonlinear processing and generalization capacity of the constructed KFD model. Therefore, the selection of suitable kernel functions is a critical important part of KFD modeling, including pre-defining the kernel type and tuning corresponding kernel parameters.

Normally, kernel function can be classified into the local kernel and global kernel (Hotta, 2009; Cheng et al., 2010; Zhu et al., 2012). The local kernel has strong interpolation capability and focuses on finding local optimal solution. By contrast, the global kernel has good extrapolation capability and aims to obtain global optimal solution (Zhu et al., 2012; Chen et al., 2018). The RBF kernel and polynomial kernel are typical local and global ones, respectively. Sample plots show the comparison of two types of kernels (Fig. 4). For the RBF kernel, data points adjacent to the test point have significant effects on kernel values (Fig. 4(a)), indicating its high interpolation capability. However, the RBF kernel only extrapolates well at large values of kernel parameter σ. In contrast, the polynomial kernel exhibits good extrapolation capability since it creates a mapping over the entire space regardless of where it was trained (Fig. 4(b)). Nevertheless, it only interpolates well when the degree of polynomial is high. Therefore, a local kernel or global kernel alone cannot provide a model with both extrapolation and

Table 3
Commonly used kernel function.

Kernel function	Mathematical expression	Kernel parameter
Linear kernel	$k(x,y) = x \cdot y$	
Polynomial kernel	$k(x,y) = (x \cdot y + 1)^d$	d
Radial basis function (RBF) kernel	$k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$	σ
Sigmoid kernel	$k(x,y) = \tanh(\gamma x \cdot y + \theta)$	γ, θ

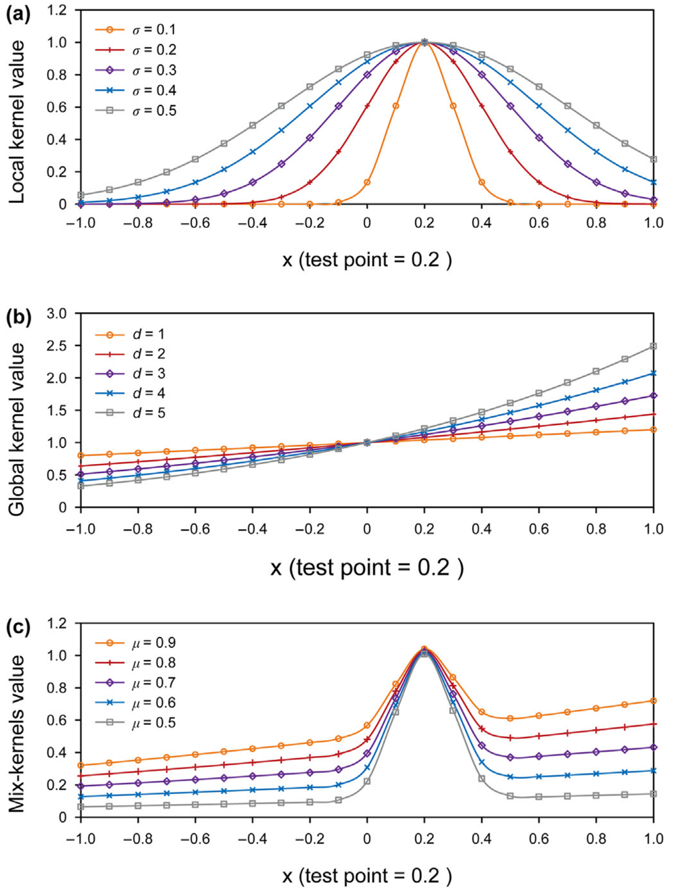


Fig. 4. Sample plots of kernel values for **(a)** local kernel (RBF kernel), **(b)** global kernel (polynomial kernel), and **(c)** mixed kernel at fixed $\sigma = 0.1$, $d = 2$. Data over $[-1, 1]$, test point $x_i = 0.2$.

interpolation properties at the same time.

Typically, to achieve strong learning and generalization property of KFD model, a kernel that possesses both strong interpolation and extrapolation capabilities is desired. Tuning an optimal composite kernel has attracted numerous researches (Cheng et al., 2010; Chen et al., 2018). The common approach is to construct a mixed kernel function with a linear combination of single kernels, which combines characteristics of each kernel. In this work, a mixture of the RBF kernel and polynomial kernel was formed as given by Eq. (6):

$$K_{\text{mix}} = \mu K_{\text{RBF}} + (1-\mu)K_{\text{poly}} \tag{6}$$

where K_{RBF} and K_{poly} are the RBF and polynomial kernel, respectively. $\mu(0,1)$ is the mixing coefficient.

Since both the K_{RBF} and K_{poly} are positive definite Mercer kernel, the linear combination K_{mix} of them also satisfies the Mercer's condition, which has been proven by numerous studies (Brailovsky et al., 1999; Pilario et al., 2019). It can be seen from Fig. 4(c) that the mixed kernel possesses not only global but also local effects. All data points that far away from and adjacent to the test point have significant influence on kernel values. Compared with the single-kernel model, the mixed-kernel model can receive interpolation and extrapolation ability simultaneously, and presents good performance of learning and generalization.

3.3. Particle swarm optimization

Hyper-parameters including the polynomial degree (d), RBF kernel width (σ), and mixing coefficient (μ) have significant impacts on the classification performance of MKFD model. In this work, the particle swarm optimization algorithm is applied to optimize hyper-parameters in the MKFD model.

Particle swarm optimization (PSO) algorithm is a population-based heuristic search technique proposed by Kennedy and Eberhart (1995). It originates from research on the socially-coordinated behavior of animal swarms. The PSO comprises several particles initialized randomly in the search space, which have their own position and velocity. The particles represent the potential solution of extremum optimization problem, and are used to compute the global optimum for fitness function. Given a T-dimensional search space, the population $X = \{X_1, X_2, \ldots, X_N\}$ is the combination of N particles. The $X_i = \{x_{i1}, x_{i2}, \ldots, x_{iT}\}$ and $V_i = \{v_{i1}, v_{i2}, \ldots, v_{iT}\}$ are the position and corresponding velocity of the ith particle in search space. During iterations, the position and velocity of each particle are updated according to the distance to its personal best position *pbest* and distance to global best position *gbest*. The update formulas are described as follows:

$$V_i(k+1) = \omega V_i(k) + c_1 r_1 (pbest_i - x_i(k)) + c_2 r_2 (gbest - x_i(t)) \tag{7}$$

$$x_i(k+1) = v_i(k+1) + x_i(k) \tag{8}$$

where $pbest_i$ is the personal best position searched by the ith particle, *gbest* is the best-so-far position searched by entire population. c_1 and c_2 are acceleration coefficients, and k is the number of current iterations. r_1 and r_2 are independent random numbers. ω is inertia weight and can be described as follows:

$$\omega(k) = \omega_{\max} - k\left(\frac{\omega_{\max} - \omega_{\min}}{k_{\max}}\right) \tag{9}$$

where $\omega(k)$ is the inertia weight after kth iteration, $\omega_{\min}$ is the final inertia weight, $\omega_{\max}$ is the initial inertia weight, and $k_{\max}$ is the preset maximum number of iterations.

During the searching process of optimal hyper-parameters by PSO, the position of particle is continuously updated by changing velocity, and finally will converge at a global optimum in search space. In this work, the particle's position is the vector of σ, d and μ, and is denominated as $P(\sigma, d, \mu)$. The average accuracy of five-fold cross-validation is set as the fitness function to evaluate the performance of optimization process. The PSO algorithm procedure is terminated when a minimum error threshold or the maximum number of iterations $k_{\max}$ is achieved.

3.4. Workflow of PSO-MKFD identification model

According to the theoretical analysis of KFD, MKFD and PSO, an identification method of reservoir types based on the hybrid PSO-MKFD model is proposed to improve the identification accuracy. For supervised machine learning methods, more effective information can be learned by model to extract key features for classification with the increase of input variable types. Hence, all available well logs in the study area, including the GR, CAL, DEN, AC, CNL, and RT logs, were selected as input variables for modeling. The modeling process in this work is carried out in the Matlab environment. The workflow of MKFD-based reservoir type identification model proceeds in the following steps (Fig. 5):

Step 1: Data collection and preprocessing. The classification standard of reservoir types is established under constraints of reservoir space type, assemblage and petrophysical properties, and then is associated with logging data to obtain the labeled dataset. To improve the calculation speed and reduce the estimation errors during modeling process, all input parameters are integrally normalized into [0,1] using the *mapminmax* function, namely $x_{\text{nor}} = (x_i - x_{\min})/(x_{\max} - x_{\min})$, where x_{nor}, x_i, x_{min} and x_{max} are the normalized, original, minimum and maximum values, respectively. To avoid model overfitting, the input dataset is randomly split into two subsets: ~75% ($n = 339$) training data and ~25% ($n = 114$) testing data (Zhong and Carr, 2016; Shi et al., 2020).

Step 2: Parameter initialization. Initialize parameters in PSO include the c_1, c_2, ω_{min}, ω_{max}, k_{max} and N. To reconcile the model performance and operational efficiency, here, $c_1 = 1.5$, $c_2 = 1.7$, $\omega_{min} = 0.4$, $\omega_{max} = 0.9$, $k_{max} = 100$ and $N = 20$ is adopted according to numerous test results.

Step 3: Model optimization. The training data are firstly assigned into the initial MKFD model for training. In this process, PSO randomly assigns hyper-parameters to model, and the performance of this set of hyper-parameters is estimated by the average accuracy of five-fold cross-validation. When the termination condition is satisfied, that is, the fitness value converges and tends to be stable, the optimal hyper-parameters are outputted. Then, the trained model is applied to the test dataset to verify the classification performance of MKFD model.

Step 4: Model evaluation. To demonstrate the improved performance, prediction results of MKFD model are compared with those of basic-kernel KFD, RF and SVM models. In addition, the built MKFD model is validated by a blind well test to validate its stability and generalization, in which the dataset is not formerly used during training and testing process. Furthermore, a variable importance analysis is conducted to better understand the geological implication of MKFD result.

3.5. Performance evaluation criteria

For multi-classification problems, the model performance is usually evaluated by confusion matrix and relevant parameters calculated based on it. The commonly used evaluation indicators include the accuracy, precision, recall, and F1-score (Table 4), which appraise the performance of model from different perspectives. Specifically, accuracy refers to the ratio of total correctly classified samples and total samples in all classes. Precision is the ratio of correctly classified samples and all samples classified into this class, which reflects the accuracy specific to this class. Recall is defined as the ratio of correctly classified samples and total samples in this class, indicating a model's comprehensiveness for classifying each class. F1-score is defined as the harmonic mean of precision and recall. The higher of indicator values (close to 1) imply that the model has better classification performance. All these indicators offer complementary information for the evaluation of model performance.

4. Results

4.1. Determination of optimal hyper-parameters

The performance of MKFD model is highly dependent on hyper-parameters, including the RBF kernel width σ, polynomial degree d, and mixing coefficient μ. The mixing coefficient μ represents the weight of two kernels and adjusts the extrapolation and interpolation capabilities of mixed kernel. For a single kernel, kernel parameters control the kernel's complexity, and further determine the data distribution in the mapping feature space. For RBF kernel, the smaller value of σ, the higher complexity of model, and over-fitting is prone to occur. On the contrary, if the σ value is too large, the model will be too constrained to capture the complexity of dataset, and under-fitting is inevitable. For polynomial kernel, large values of d will lead to high computational and learning complexity, and the model is tended to over-fitting. Hence, to balance the classification accuracy, generalization capacity and computational cost of MKFD model, search ranges of kernel parameters were set to $\sigma[0.01, 10]$ and $d[1, 6]$, respectively.

The searching process of optimal hyper-parameters in MKFD model by PSO is presented in Fig. 6(a). It can be found that the fitness value (average accuracy of five-fold cross-validation) increased accordingly with iteration process. Within the maximum iterations ($k_{\max} = 100$), the average accuracy finally converges to a stable value, indicating that the parameter search is featured with optimal results. Thus, the corresponding hyper-parameters are defined as optimum values. By PSO optimization, optimal hyper-parameters of MKFD model are determined to be $[\sigma,\ d,\ \mu] = [1.572, 2.385, 0.736]$.

4.2. Model training and testing results

Totals of 453 datasets obtained from 7 wells were utilized for MKFD modeling, and input variables include the GR, CAL, DEN, AC, CNL and RT logs. Purposefully to circumvent over-fitting in modeling, from total datasets, ~75% (339 data) was randomly sampled for model training, and the remaining ~25% (114 data) was employed for model testing. Through the training and testing of MKFD, three features were extracted for the classification of reservoir types. Fig. 7 presents the distribution characteristics of three extracted classification features of reservoir types. It can be seen that each reservoir type was well separated from others except for a few data points. In addition, data points belonging to each class were densely clustered around their own cluster centroids, and different classes had obvious decision boundaries. Compared with original inputs (Fig. 3), datasets projected by the MKFD model are more separable, indicating that the model was well trained to effectively classify reservoir types. The classification accuracies of training and testing datasets are 93.8% and 91.2%, respectively.

It should be noticed that the classification ability of each extracted feature is different. For the three extracted features in this work, eigenvalues of Feature 1, Feature 2 and Feature 3 are 6.10, 2.96 and 0.11, respectively. Therefore, classification abilities of Feature 1 (66.5%) and Feature 2 (32.3%) are much higher than that of Feature 3 (1.2%), and the total classification contribution of the first two is 98.8%. The three-dimensional cross-plots with three features extracted by the MKFD model shows that the categories were well-separated in the Feature 1 and Feature 2 (Fig. 7), indicating that they contain almost all the feature information involved in initial variables. In contrast, all categories have a wide range of Feature 3 values. Hence, although the addition of Feature 3 increases the distance between categories, it can barely improve the classification ability to MKFD model.

4.3. Performance comparison

To demonstrate the improved performance of the proposed MKFD model, two KFD models with different basic kernels (i.e., the RBF kernel and polynomial kernel), RF model and SVM model, were applied for identifying reservoir types under the premise of keeping training and testing datasets unchanged. In modeling process, the five-fold cross-validation technique was adopted in parameter optimization for RBF-KFD and Poly-KFD models, and optimal hyperparameters of RF and SVM models were obtained by PSO. Searching ranges of parameters σ in RBF-KFD, d in Poly-KFD, *ntre* (total number of trees) and *ntd* (maximum tree depth) in RF, and C

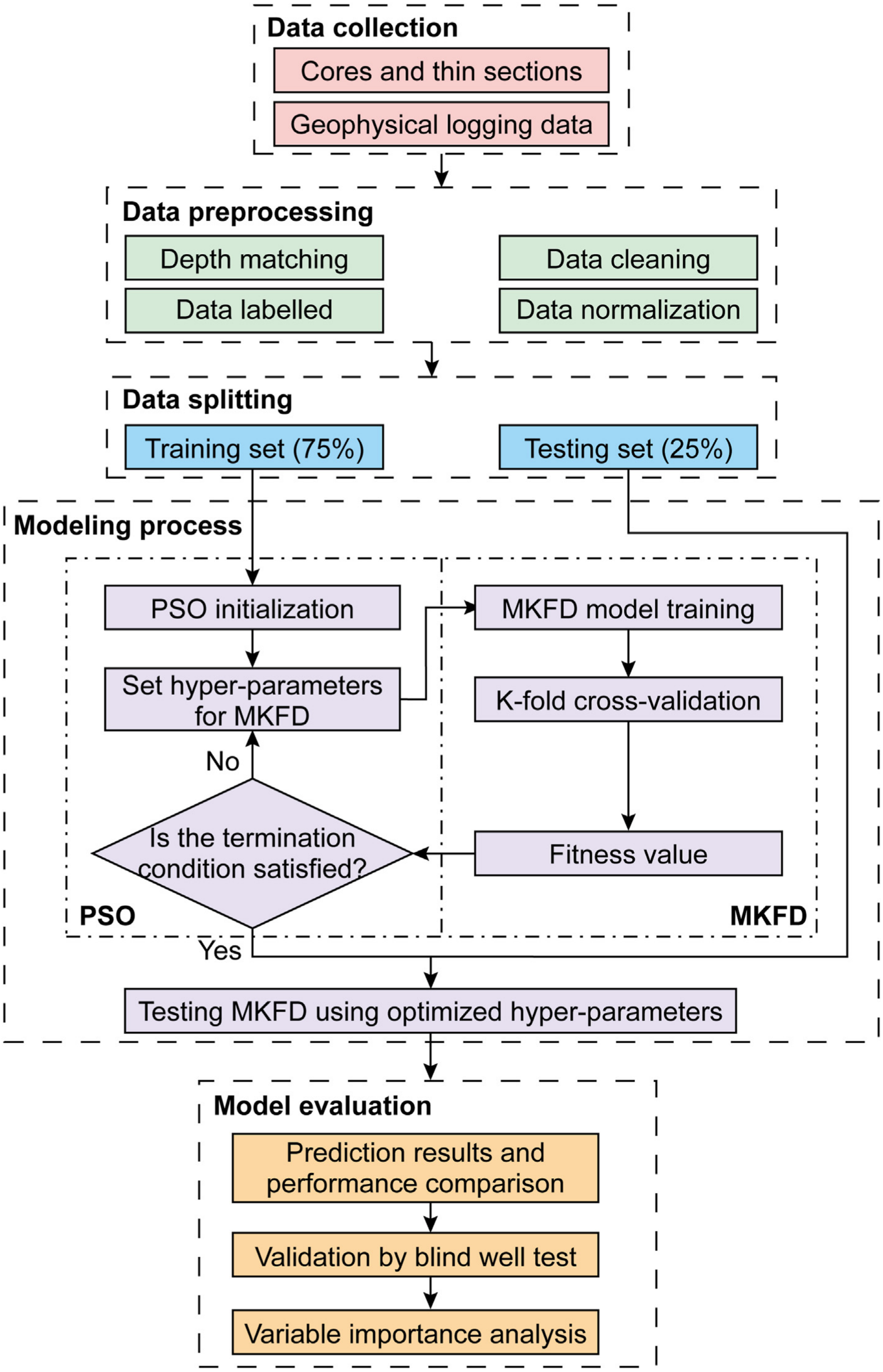


Fig. 5. Workflow of MKFD model for the identification of reservoir types.

Table 4
Definitions of performance evaluation indicators used in this work.

Evaluation indicators	Mathematical expression
Accuracy	$accuracy = \frac{TP+TN}{TP+FP+FN+TN} \times 100\%$
Precision	$precision = \frac{TP}{TP+FP} \times 100\%$
Recall	$recall = \frac{TP}{TP+FN} \times 100\%$
F1-score	$F1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall} \times 100\%$

Note: TP = true positives; FP = false positives; TN = true negatives; FN = false negatives.

(error penalty factor) and g (width parameter of RBF kernel) in SVM were set to $\sigma[0.01, 10]$, $d[1, 6]$, $ntre[10, 500]$, $ntd[1, 5]$, $C[0.01, 100]$, $g[0.01, 50]$, respectively. The optimal parameter can be determined when the model gains the highest accuracy (Fig. 6). The performance of MKFD, RBF-KFD, Poly-KFD, RF and SVM models was quantitatively compared on the basis of confusion matrix and four evaluation indicators.

Confusion matrices of training and testing results for five models are presented in Fig. 8, in which diagonal positions represent the correct classification and others represent incorrect classification. According to these confusion matrices, classification accuracies of the three models were obtained (Table 5). The training accuracies of MKFD, RBF-KFD, Poly-KFD, RF and SVM models are 93.8%, 87.6%, 83.5%, 87.0% and 83.8%, respectively. The corresponding accuracies for testing data are 91.2%, 78.9%, 70.2%, 76.3 and 73.7%, respectively. Judging from these values, the modeling performance of MKFD method outperforms, which provides more accurate prediction. In addition, the difference between the training and testing accuracies is the smallest, indicating the MKFD model has stronger ability to avoid over-fitting compared with other two KFD models.

The evaluation indicators precision, recall, and F1-score are further used to compare the model performance. The indicator values of all reservoir types by MKFD are all over 90% (90.7–95.7%), while those of two KFD models are in the range of 70.3%–89.6% (Fig. 9). Compared to RBF-KFD, Poly-KFD, RF and SVM models, the precision, recall, and F1-score values of MKFD model increase ranging from 3.1% to 22.8%, 6.6%–14.7%, and 4.8%–18.8%, respectively. Besides, it can be observed that the misclassification of MKFD mainly exists between the pore-cavern and fractured-cavern reservoirs, as well as the pore reservoirs and tight reservoirs (Fig. 8). However, the misclassification of other four models exist among each reservoir type. From the perspective of reservoir quality, the pore-cavern and fractured-cavern reservoirs are most favorable for gas production in the study area, followed by fractured-pore reservoirs, and the reservoir quality of pore reservoirs and tight reservoirs are worst. In this context, the misclassification between pore-cavern and fractured-cavern reservoirs, as well as pore reservoirs and tight reservoirs is acceptable. However, the misclassification between reservoirs with good and poor quality will misguide the geologist. Compared with traditional KFD, RF and SVM models, the MKFD has certain advantages in identifying reservoir types for deep dolomites in the Sichuan Basin.

4.4. Generalization ability validation

To validate the generalization ability, five models were applied to another set of data from two coring intervals (total 41.5 m) in the Well M15, where all 322 datasets were previously not used for modeling. Identification results based on the core analysis and MKFD, RBF-KFD, Poly-KFD, RF, SVM prediction are shown in Fig. 10. It can be found that identification results of the MKFD model are highly consistent with that of core observation except for a few depths. The identification accuracy of blind-test data based on MKFD is 92.7%, which is quite approximate the total accuracy of modeling process (93.2%). In addition, it can be seen that the misclassification phenomena are similar these in modeling data. For RBF-KFD, Poly-KFD, RF, and SVM models, although their total accuracies of validation data exceed 75.0% (82.2%, 77.6%, 81.7%, and 79.2%, respectively), differences between total accuracies in the validation and modeling are larger than that of MKFD model, indicating the low generalization abilities of these two models.

It should be noted that the misclassification could not be entirely attributed to the discriminant ability of MKFD model. The performance of supervised learning model is dependent largely on the comprehensiveness and reliability of input data. On the one hand, data with large size are beneficial for model to learn more information and extract more key features for further discriminant. The accuracy and reliability of supervised learning method will be worse when data used in modeling are insufficient. In this work, data employed in modeling process may not provide adequate feature information for MKFD, despite all the available ones have been used. On the other hand, the accuracy of well logging data was inevitably disturbed by the geo-pressure, geo-temperature and drilling fluids in the drilling process. Logging data with inapparent errors may still exist, although some abnormal values in raw data have been cleaned. Nonetheless, the successful application in blind-test well M15 indicates that the proposed MKFD method can be used as a reliable tool for identifying reservoir types in deep carbonates.

4.5. Variable importance

To investigate the impact of input variable on MKFD's outcome, well logs (i.e., 453 datasets of 7 wells) with different combinations were used as input data for modeling, in which the training and testing data ratio was set as 3:1. Each MKFD model with different well-log combinations was trained and tested twenty times randomly. Experimental results of nine MKFD models trained and tested with different well-log combinations are presented in Table 6. It can be seen that the classification accuracy, including the training, testing and total accuracies, become lower with the reduction of well-log types. Compared with MKFD built based on all types of well logs, total accuracies of MKFDs modelled by using five, four and three types of well logs decreased 2.5%–8.2%, 10.6%–16.2%, and 19.5%–36.9%, respectively. This is mainly because the quantity of well logs vitally affects the ability of MKFD to learn more generalized rules for classification. More types of well logs implies more feature information can be learned by MKFD, resulting in higher classification accuracy.

Furthermore, to better understand the geological implication of MKFD results, variable importance analysis was conducted to explain the relation between input variables and model outputs. The Shapley additive explanations (SHAP) method is applied in this study. Based on the game theory, SHAP appraises the importance of input variables by considering the marginal contribution when add a variable to model (Male et al., 2020). The result is presented in the form of SHAP value, and can be determined as

$$\Phi_i = \sum_{S \subseteq N \setminus \{x_i\}} \left(\frac{|S|!(|N| - |S| - 1)!}{|N|!} \right) [V(S \cup \{x_i\}) - v(S)] \tag{10}$$

where N is the set of all input variables, S is a subset of N, and $V(S)$ is the model output corresponding to input variables S. The models $V(S \cup \{x_i\})$ and $V(S)$ are trained with and without input variables, respectively. The SHAP value Φ_i of input i is determined by the average value of all possible permutations of variable set. The higher the SHAP value, the more important the input variable is.

The variable importance analysis results of 6 logging variables for MKFD model based on SHAP are shown in Fig. 11. It can be found that the SHAP values of well logs decreased in the order of DEN, AC, RT, CNL, CAL and GR for the proposed MKFD model. Typically, the value variation of well logs is related to the changes of rock physicochemical property around wellbore (Ghosh et al., 2016; Wang et al., 2020). As mentioned previously, types and development degree of storage spaces in Dengying dolomites are diverse, resulting in differences of porosity, permeability and gas content. The DEN and AC logs are sensitive to porosity change, and can well reflect the size of storage spaces. The RT and CNL are sensitive to the variation of gas content in reservoir spaces. However, the gas content of some high porosity reservoirs may be low because of the poor preservation condition, resulting in the similar RT and CNL logging responses of low porosity reservoirs. Due to the deeply burial depth of Dengying reservoirs, the CAL are easily interfered by formation pressure and in-situ stress in drilling process. In addition, the variation of GR values is inapparent since the extremely low clay content in dolomites. Hence, it can be concluded that DEN and AC logs are the most effective data for reservoir type identification of deep carbonates, followed by the RT and CNL, while CAL and GR logs contribute little.

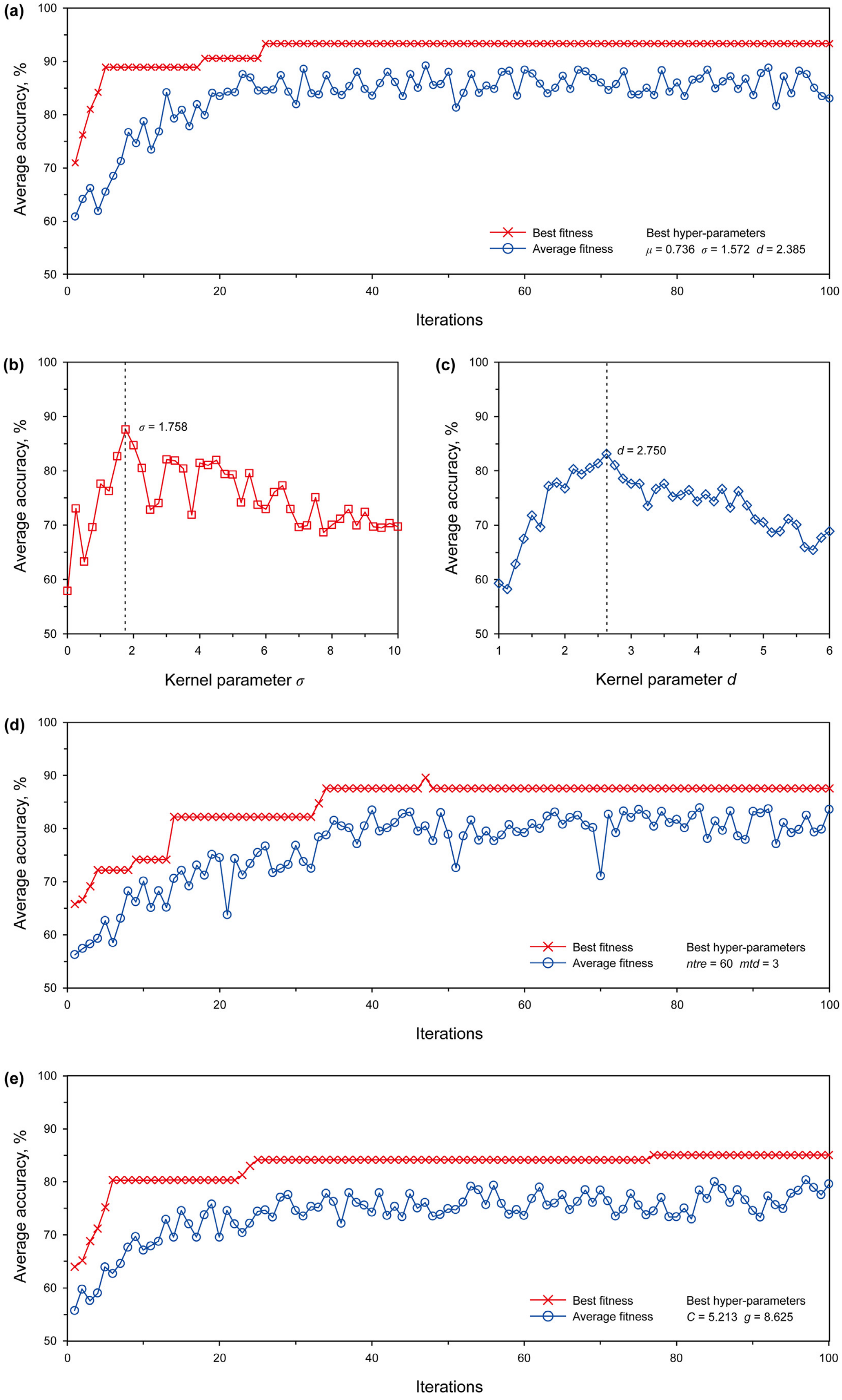


Fig. 6. The searching process of **(a)** optimal hyper-parameters in MKFD model by PSO, kernel parameters **(b)** σ in RBF-KFD model and **(c)** d in Poly-KFD model by 5-fold cross-validation, and optimal hyper-parameters in **(d)** RF and **(e)** SVM models by PSO.

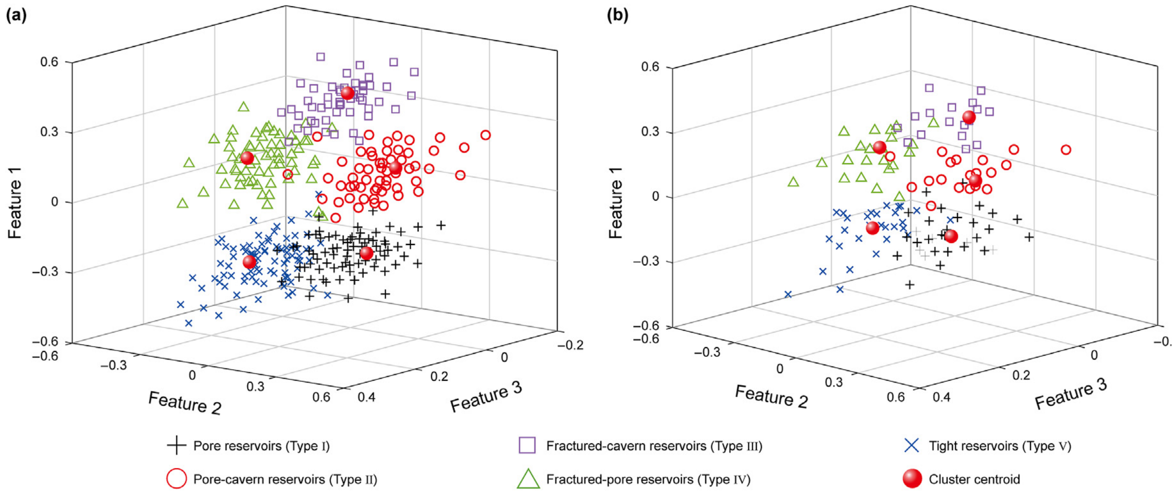


Fig. 7. Three-dimensional crossplots of classification features for reservoir types extracted by MKFD model. **(a)** Training dataset (n = 339) and **(b)** testing dataset (n = 114).

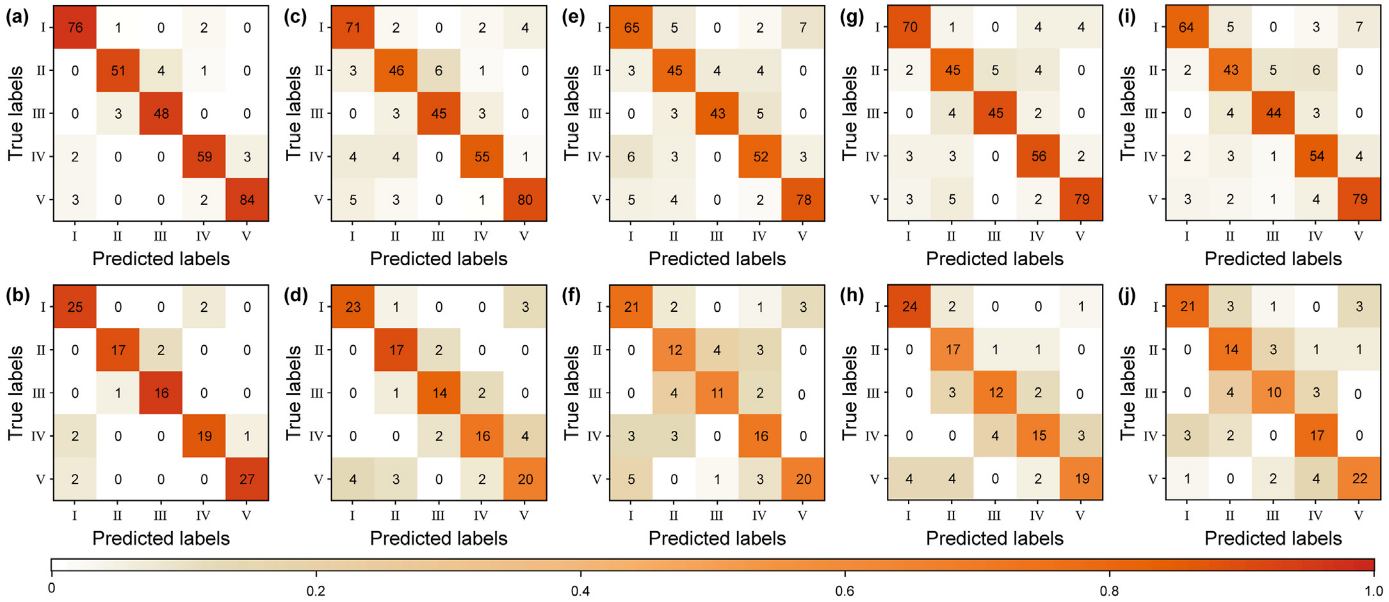


Fig. 8. Confusion matrix of different models. MKFD on **(a)** training dataset and **(b)** testing dataset; RBF-KFD model on **(c)** training dataset and **(d)** testing dataset; Poly-KFD model on **(e)** training dataset and **(f)** testing dataset; RF on **(g)** training dataset and **(h)** testing dataset; SVM on **(i)** training dataset and **(j)** testing dataset. Codes I-V correspond to pore reservoirs, pore-cavern reservoirs, fractured-cavern reservoirs, fractured-pore reservoirs and tight reservoirs, respectively.

5. Discussion

Identification of reservoir types is a fundamental work in hydrocarbon exploration and development. Traditional artificial identification and cross-plot analysis methods are heavily dependent on geologist's experience and inefficient, which cannot satisfy requirements for the research of highly heterogeneous reservoirs. By comparison, the proposed MKFD method can automatically and accurately identify reservoir types, and reduce the influence of human subjective factors. By training and optimization processes, the built MKFD model is feasible to identify reservoir types of non-cored wells, which greatly improves the work efficiency. In addition, the spatial distribution prediction of different reservoir types based on multi-well analysis has an important guiding significance for the determination of development programs. The MKFD method can also be applied to the identification of lithology, lithofacies, sedimentary microfacies and fractures for carbonate reservoirs as well as clastic reservoirs, metamorphic reservoirs and igneous reservoirs. With the exploration and development gradually turning to more complex and heterogeneous reservoirs, the MKFD identification method is increasingly valuable for reservoir prediction, characterization and valuation.

Table 5
Performance comparison of different identification models with training, testing and total data.

Model	Accuracy, %			Avd, %
	Training data	Testing data	Total data	
MKFD	93.8	91.2	93.2	2.6
RBF-KFD	87.6	78.9	85.4	8.7
Poly-KFD	83.5	70.2	80.1	13.3
RF	87.0	76.3	85.3	10.7
SVM	83.8	73.7	81.2	10.1

Note: Avd = absolute value of difference between the accuracies of training and testing data.

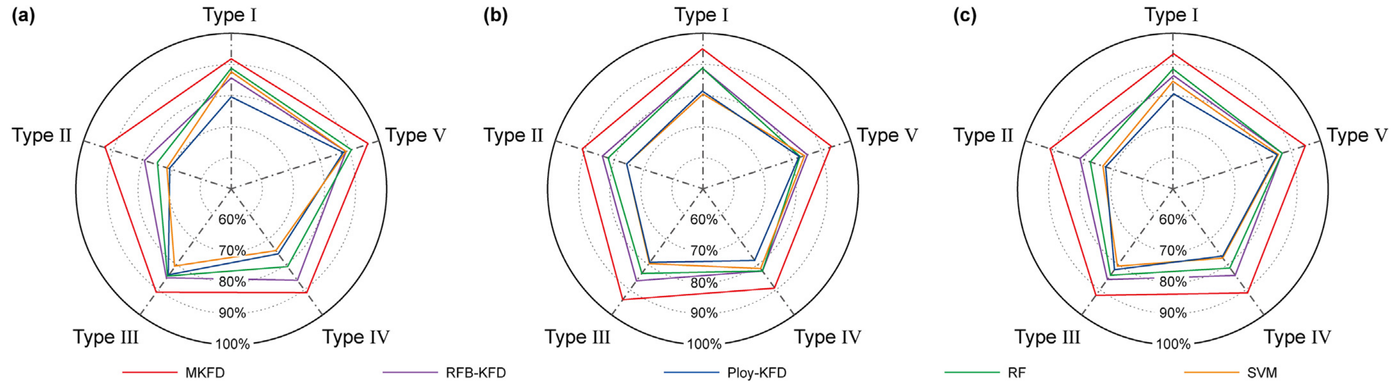


Fig. 9. Comparison of identification results by different evaluation indicators for MKFD, RBF-KFD, Poly-KFD, RF and SVM models. **(a)** Precision, **(b)** recall, and **(c)** F1-score. Type I–V correspond to pore reservoirs, pore-cavern reservoirs, fractured-cavern reservoirs, fractured-pore reservoirs and tight reservoirs, respectively.

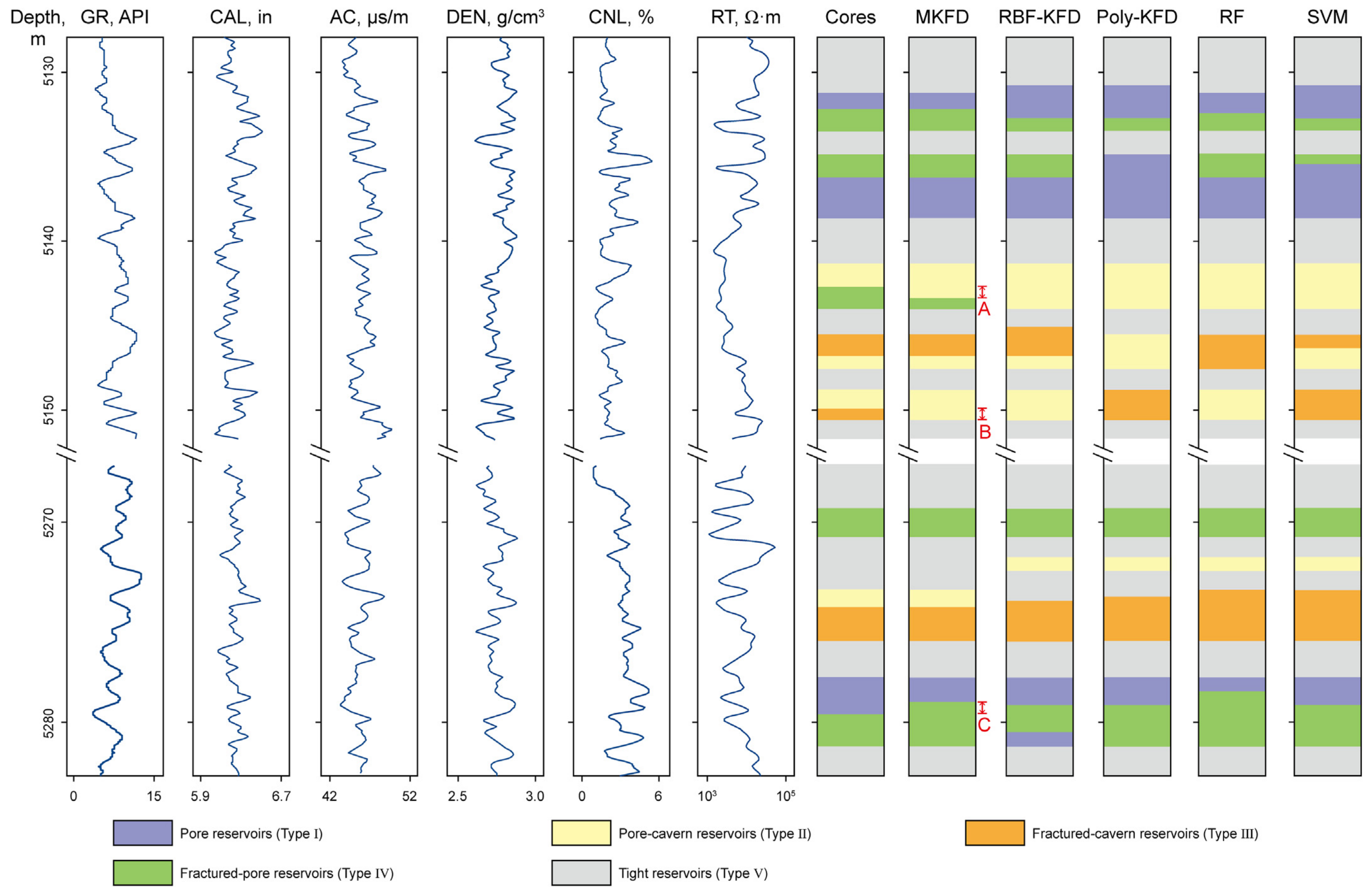


Fig. 10. Identification results of reservoir types in well M15 by the MKFD, RBF-KFD, Poly-KFD, RF and SVM methods.

Meanwhile, there are still some problems should be considered in future work to further expand its applications in petroleum geology.

First, the misclassification of blind-well tests by MKFD method mainly occurred at the interface of different reservoir types (Fig. 10). For instance, fractured-pore reservoirs (at depth of 5142.80–5143.35 m, interval A) and fractured-cavern reservoirs (at depth of 5150.10–5150.70 m, interval B) were misclassified into pore-cavern reservoirs. These misclassification phenomena have also been referred in other literatures. In the research of Chen et al. (2021), the authors attributed these misclassifications to the "boundary effects", which was mainly caused by the low resolution of logging data. In the transformation interface of reservoir types, logging values may be disturbed or diminished due to the low logging resolution and cannot represent the true logging response of reservoir types, which makes it difficult to identify reservoir types, particularly when the reservoir thickness is very thin. Therefore, the improvement of logging resolution is necessary prior to well logs are used for MKFD modeling. The wavelet transform is a potential solution to solve the logging resolution problem.

Second, well logging data used for modeling are generally balanced to ensure the classification performance of model, such as

Table 6
Performance of MKFD model trained and tested with different types of well logs.

Well log types	Accuracy, %		
	Training data	Testing data	Total data
CAL, GR, AC, CNL, RT	87.0 (85.8–87.6)	78.9 (77.2–79.8)	85.0 (83.7–85.7)
GR, DEN, AC, CNL, RT	92.0 (91.4–92.6)	86.8 (84.2–88.6)	90.7 (89.6–91.6)
CAL, GR, DEN, AC, CNL	88.2 (87.9–89.4)	80.7 (78.9–83.3)	86.3 (85.7–87.9)
DEN, AC, CNL, RT	84.7 (83.5–86.1)	76.3 (74.6–78.9)	82.6 (81.2–84.3)
CAL, GR, CNL, RT	79.9 (79.1–81.4)	68.4 (66.7–70.2)	77.0 (76.0–78.6)
CAL, GR, DEN, AC	81.7 (81.1–83.5)	70.2 (69.3–72.8)	78.8 (78.1–80.8)
DEN, AC, RT	77.6 (76.7–79.6)	62.3 (61.4–66.7)	73.7 (72.8–76.4)
CAL, GR, CNL	59.9 (58.7–61.4)	45.6 (43.0–49.1)	56.3 (54.7–58.3)
AC, RT, GR	66.4 (65.5–68.4)	51.8 (50.0–55.3)	62.7 (61.6–65.1)

Note: 6.8 (3.9–10.6) = mean (minimum–maximum).

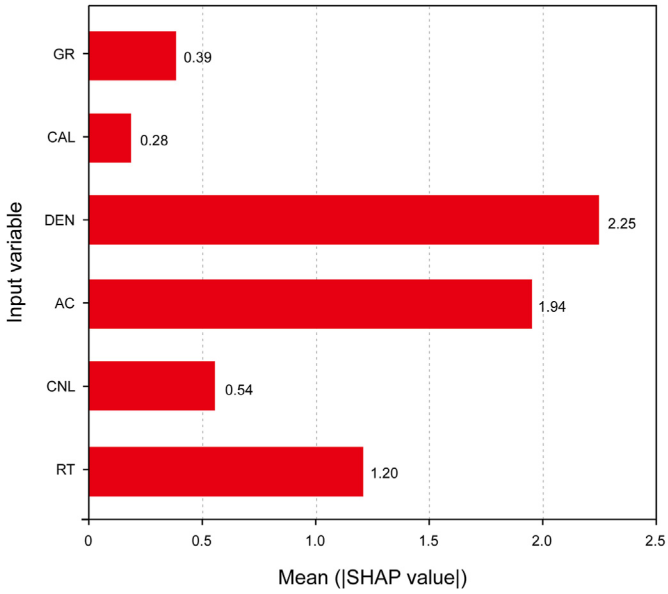


Fig. 11. The SHAP values of each input variable in MKFD model.

the dataset used in this study, in which the sample numbers of each class (i.e., reservoir types) are approximately the same. However, the imbalanced data is ubiquitous in classification problem, especially for the reservoir type identification in deep carbonates, which is associated with the strong reservoir heterogeneity and always aggravated by data acquisition. Normally, classification models trained with imbalanced data are biased towards the majority class, resulting in underprediction of the minority ones (Yuan et al., 2018). Hence, the imbalanced data can dramatically reduce the model performance. To address this problem, some effective approaches, such as the prototype generation (PG) algorithm and synthetic minority over-sampling technique (SMOTE), will be explored in future work.

Besides, MKFD is a supervised machine learning algorithm and uses only labeled logging data (with class labels) for model training. To guarantee the reliability and stability of MKFD model, sufficient labeled data are required for modeling. However, the number of labeled logging data is generally small due to limited core samples and high cost of manual labeling, which will make the model generalization ability worse. In contrast, unlabeled logging data (without class labels) are generally available in large quantities. Despite unlabeled logging data contain effective information (e.g., data distribution and geometric data structure) for classification, they cannot be utilized for model training by MKFD due to its nature of supervised learning. Recent studies have demonstrated that the combination of labeled and unlabeled data can improve the classification accuracy of supervised methods, i.e., the semi-supervised learning algorithm (Liu et al., 2020a,b; Lan et al., 2021). Different from supervised machine learning, semi-supervised learning mines additional information from abundant unlabeled data to construct appropriate models, and thus only requires a few labeled data, which can alleviate the over-reliance on labeled data. Hence, unlabeled logging data will be critical for machine learning to develop a powerful classification model when limited labeled logging data are available. In future works, the improvement of MKFD based on semi-supervised learning strategy will be considered.

6. Conclusions

Identification of reservoir types in deep carbonates using geophysical logging data is a complex nonlinear classification problem. To address this issue, this paper introduced a mixed kernel to machine learning and developed a mixed kernel Fisher discriminant analysis (MKFD) model. The proposed MKFD has well extrapolate and interpolate capabilities, and overcomes the single-kernel limitation. The MKFD method was successfully applied to reservoir type identification in deep carbonates of the Sichuan Basin, with high prediction accuracy in both modeling (93.2%) and blind test (92.7%) processes. Compared with single-kernel based KFD models, MKFD presents outstanding identification validity and generalization ability, which has over 5.0% increases in accuracy, precision, recall, and F1-score.

Reservoir type identification is a fundamental work in reservoir research, and provides basic information for reservoir quality evaluation. Well logs and machine learning are the two useful tools for reservoir type identification. The proposed MKFD method has great application potential to other petroliferous basins worldwide for identifying carbonate reservoir types. Practically, the varying reservoir conditions from region to region bring great challenges to modeling and identification. To develop a powerful identification model, the model optimization based on actual logging data is the critical issue.

The MKFD identification method greatly improve the classification accuracy and reduce the dependent on geologist's experience. Simultaneously, the MKFD method possesses high efficiency due to its automatic recognition and scalability available in non-cored wells. With the exploration and development gradually turning to more complex and heterogeneous reservoirs, the MKFD method has a wide application prospect in identification of lithology, lithofacies, sedimentary microfacies and fractures for carbonate reservoirs and other reservoirs. The proposed identification method will play more important roles in reservoir characterization, prediction and valuation. The improvement of MKFD by considering issues of low logging resolution, imbalanced data and limited labeled data will be explored to obtain better applications.

Declaration of competing interest

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

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. U21B2062) and the Natural Science Foundation of Hubei Province (No. 2023AFB307). Special thanks to the PetroChina Southwest Oil & Gas field Company for providing core samples and logging data. The authors are also very grateful to

reviewers for their constructive suggestions.

References

Aghli, G., Soleimani, B., Moussavi-Harami, R., et al., 2016. Fractured zones detection using conventional petrophysical logs by differentiation method and its correlation with image logs. J. Pet. Sci. Eng. 142, 152–162. https://doi.org/10.1016/j.petrol.2016.02.002.

Baudat, G., Anouar, F., 2000. Generalized discriminant analysis using a kernel approach. Neural Comput. 12 (10), 2385–2404. https://doi.org/10.1162/089976600300014980.

Billings, S.A., Lee, K.L., 2002. Nonlinear Fisher discriminant analysis using a minimum squared error cost function and the orthogonal least squares algorithm. Neural Networks 15 (2), 263–270. https://doi.org/10.1016/S0893-6080(01)00142-3.

Brailovsky, V.L., Barzilay, O., Shahave, R., 1999. On global, local, mixed and neighborhood kernels for support vector machines. Pattern Recogn. Lett. 20 (11–13), 1183–1190. https://doi.org/10.1016/S0167-8655(99)00086-0.

Chen, S.D., Liu, P.C., Tang, D.Z., et al., 2021. Identification of thin-layer coal texture using geophysical logging data: investigation by wavelet transform and linear discrimination analysis. Int. J. Coal Geol. 239, 103727. https://doi.org/10.1016/j.coal.2021.103727.

Chen, Y.H., Kloft, M., Yang, Y., et al., 2018. Mixed kernel based extreme learning machine for electric load forecasting. Neurocomputing 312, 90–106. https://doi.org/10.1016/j.neucom.2018.05.068.

Cheng, C.Y., Hsu, C.C., Chen, M.C., 2010. Adaptive kernel principal component analysis (KPCA) for monitoring small disturbances of nonlinear processes. Ind. Eng. Chem. Res. 49 (5), 2254–2262. https://doi.org/10.1021/ie900521b.

Chu, W.S., Chen, J.C., Lien, J.J.J., 2011. Kernel discriminant transformation for image set-based face recognition. Pattern Recogn. 44, 1567–1580. https://doi.org/10.1016/j.patcog.2011.02.011.

Dong, P., Chen, Z.M., Liao, X.W., et al., 2022a. A deep reinforcement learning (DRL) based approach for well-testing interpretation to evaluate reservoir parameters. Petrol. Sci. 19 (1), 264–278. https://doi.org/10.1016/j.petsci.2021.09.046.

Dong, S.Q., Wang, Z.Z., Zeng, L.B., 2016. Lithology identification using kernel Fisher discriminant analysis with well logs. J. Pet. Sci. Eng. 143, 95–102. https://doi.org/10.1016/j.petrol.2016.02.017.

Dong, S.Q., Zeng, L.B., Du, X.Y., et al., 2022b. Lithofacies identification in carbonate reservoirs by multiple kernel Fisher discriminant analysis using conventional well logs: a case study in A oilfield, Zagros Basin, Iraq. J. Pet. Sci. Eng. 210, 110081. https://doi.org/10.1016/j.petrol.2021.110081.

Dong, S.Q., Zeng, L.B., Liu, J.J., et al., 2020. Fracture identification in tight reservoirs by multiple kernel Fisher discriminant analysis using conventional logs. Interpretation 8 (4), 215–225. https://doi.org/10.1190/INT-2020-0048.1.

Dong, S.Q., Zeng, L.B., Lyu, W.Y., et al., 2019. Fracture identification by semi-supervised learning using conventional logs in tight sandstones of Ordos Basin, China. J. Nat. Gas Sci. Eng. 76, 103131. https://doi.org/10.1016/j.jngse.2019.103131.

Feng, Q.F., Xiao, Y.X., Hou, X.L., et al., 2021. Logging identification method of depositional facies in Sinian Dengying Formation of the Sichuan Basin. Petrol. Sci. 18 (4), 1086–1096. https://doi.org/10.1016/j.petsci.2020.10.002.

Ghosh, S., Chatterjee, R., Shanker, P., 2016. Estimation of ash, moisture content and detection of coal lithofacies from well logs using regression and artificial neural network modelling. Fuel 177, 279–287. https://doi.org/10.1016/j.fuel.2016.03.001.

Hotta, K., 2009. View independent face detection based on horizontal rectangular features and accuracy improvement using combination kernel of various sizes. Pattern Recogn. 42 (3), 437–444. https://doi.org/10.1016/j.patcog.2008.08.013.

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. Science 349 (6245), 255–260. https://doi.org/10.1126/science.aaa8415.

Katz, B.J., Everett, M.A., 2016. An overview of pre-Devonian petroleum systems: unique characteristics and elevated risks. Mar. Petrol. Geol. 73, 492–516. https://doi.org/10.1016/j.marpetgeo.2016.03.019.

Kennedy, J., Eberhart, R., 1995. Particle swarm optimization. In: IEEE International Joint Conference on Neural Networks, Perth, Australia, pp. 1942–1948.

Khalifah, H.A., Glover, P.W.J., Lorinczi, P., 2019. Permeability prediction and diagenesis in tight carbonates using machine learning techniques. Mar. Petrol. Geol. 112, 104096. https://doi.org/10.1016/j.marpetgeo.2019.104096.

Lai, J., Liu, B.C., Li, H.B., et al., 2022. Bedding parallel fractures in fine-grained sedimentary rocks: recognition, formation mechanisms, and prediction using well log. Petrol. Sci. 19 (2), 554–569. https://doi.org/10.1016/j.petsci.2021.10.017.

Lan, X.X., Zou, C.N., Kang, Z.H., et al., 2021. Log facies identification in carbonate reservoirs using multiclass semi-supervised learning strategy. Fuel 302, 121145. https://doi.org/10.1016/j.fuel.2021.121145.

Lei, C.K., Deng, J., Cao, K., et al., 2019. A comparison of random forest and support vector machine approaches to predict coal spontaneous combustion in gob. Fuel 239 (1), 297–311. https://doi.org/10.1016/j.fuel.2018.11.006.

Liu, H., Ren, Y.L., Li, X., et al., 2022. Rock thin-section analysis and identification based on artificial intelligent technique. Petrol. Sci. 19 (4), 1605–1621. https://doi.org/10.1016/j.petsci.2022.03.011.

Liu, M., Jervis, M., Li, W., et al., 2020a. Seismic facies classification using supervised convolutional neural networks and semi-supervised generative adversarial networks. Geophysics 85 (4), O47–O58. https://doi.org/10.1190/GEO2019-0627.1.

Liu, Q.S., Lu, H.Q., Ma, S.D., 2004. Improving kernel Fisher discriminant analysis for face recognition. IEEE Trans. Circ. Syst. Video Technol. 14, 42–49. https://doi.org/10.1109/ICME.2004.1394460.

Liu, X.Y., Zhou, L., Chen, X.H., et al., 2020b. Lithofacies identification using support vector machine based on local deep multi-kernel learning. Petrol. Sci. 17 (4), 954–966. https://doi.org/10.1007/s12182-020-00474-6.

Lu, X.B., Wang, Y., Tian, F., et al., 2017. New insights into the carbonate karstic fault system and reservoir formation in the Southern Tahe area of the Tarim Basin. Mar. Petrol. Geol. 86, 587–605. https://doi.org/10.1016/j.marpetgeo.2017.06.023.

Lyu, W.Y., Zeng, L.B., Liu, Z.Q., et al., 2016. Fracture responses of conventional logs in tight-oil sandstones: a case study of the Upper Triassic Yanchang Formation in southwest Ordos Basin, China. AAPG Bull. 100 (9), 1399–1417. https://doi.org/10.1306/04041615129.

Mahdaviara, M., Rostami, A., Shahbazi, K., 2020. State-of-the-art modeling permeability of the heterogeneous carbonate oil reservoirs using robust computational approaches. Fuel 268, 117389. https://doi.org/10.1016/j.fuel.2020.117389.

Male, F., Jensen, J.L., Lake, L.W., 2020. Comparison of permeability predictions on cemented sandstones with physics-based and machine learning approaches. J. Nat. Gas Sci. Eng. 77, 103244. https://doi.org/10.1016/j.jngse.2020.103244.

Matonti, C., Guglielmi, Y., Viseur, S., et al., 2015. Heterogeneities and diagenetic control on the spatial distribution of carbonate rocks acoustic properties at the outcrop scale. Tectonophysics 638, 94–111. https://doi.org/10.1016/j.tecto.2014.10.020.

Méndez, J.N., Jin, Q., Zhang, X.D., et al., 2021. Rock type prediction and 3D modeling of clastic paleokarst fillings in deeply-buried carbonates using the Democratic Neural Networks Association technique. Mar. Petrol. Geol. 127 (1), 104987. https://doi.org/10.1016/j.marpetgeo.2021.104987.

Mika, S., Rätsch, G., Weston, J., et al., 1999. Fisher discriminant analysis with kernels. Neural Network. 9, 41–48. https://doi.org/10.1109/NNSP.1999.788121.

Pilario, K.E.S., Cao, Y., Shafiee, M., 2019. Mixed kernel canonical variate dissimilarity analysis for incipient fault monitoring in nonlinear dynamic processes. Comput. Chem. Eng. 123, 143–154. https://doi.org/10.1016/j.compchemeng.2018.12.027.

Shi, J.X., Zeng, L.B., Dong, S.Q., et al., 2020. Identification of coal structures using geophysical logging data in Qinshui Basin, China: investigation by kernel Fisher discriminant analysis. Int. J. Coal Geol. 217, 103314. https://doi.org/10.1016/j.coal.2019.103314.

Shi, J.X., Zhao, X.Y., Pan, R.F., et al., 2022. Natural fractures in the deep Sinian carbonates of the central Sichuan Basin, China: implications for reservoir quality. J. Pet. Sci. Eng. 216, 110829. https://doi.org/10.1016/j.petrol.2022.110829.

Shi, J.X., Zhao, X.Y., Zeng, L.B., et al., 2023. Identification of coal structures by semi-supervised learning based on limited labeled logging data. Fuel 337, 127191. https://doi.org/10.1016/j.fuel.2022.127191.

Souvik, S., Mohamed, A., Shib, S.G., et al., 2021. Petrophysical heterogeneity of the early Cretaceous Alamein dolomite reservoir from North Razzak oil field, Egypt integrating well logs, core measurements, and machine learning approach. Fuel 306, 121698. https://doi.org/10.1016/j.fuel.2021.121698.

Sridevi, P., 2018. Identification of suitable membership and kernel function for FCM based FSVM classifier model. Cluster Comput. 6, 1–10. https://doi.org/10.1007/s10586-017-1533-9.

Tian, F., Luo, X.L., Zhang, W., 2019. Integrated geological-geophysical characterizations of deeply buried fractured-vuggy carbonate reservoirs in Ordovician strata, Tarim Basin. Mar. Petrol. Geol. 99, 292–309. https://doi.org/10.1016/j.marpetgeo.2018.10.028.

Tokhmchi, B., Memarian, H., Rezaee, M.R., 2010. Estimation of the fracture density in fractured zones using petrophysical logs. J. Pet. Sci. Eng. 72 (1–2), 206–213.

Wang, L., He, Y.M., Peng, X., et al., 2020. Pore structure characteristics of an ultra-deep carbonate gas reservoir and their effects on gas storage and percolation capacities in the Deng IV member, Gaoshiti-Moxi Area, Sichuan Basin, SW China. Mar. Petrol. Geol. 111, 44–65. https://doi.org/10.1016/j.marpetgeo.2019.08.012.

Xu, L.X., Niu, X., Xie, J., et al., 2015. A local-global mixed kernel with re- producing property. Neurocomputing 168, 190–199. https://doi.org/10.1016/j.neucom.2015.05.107.

Xu, Y., Yang, J.Y., Yang, J., 2004. A reformative kernel Fisher discriminant analysis. Pattern Recogn. 37 (6), 1299–1302. https://doi.org/10.1016/j.patcog.2003.10.006.

Yuan, X.H., Xie, L.J., Abouelenien, M., 2018. A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data. Pattern Recogn. 77, 160–172. https://doi.org/10.1016/j.patcog.2017.12.017.

Zhang, S.C., Huang, H.P., Su, J., et al., 2015. Ultra-deep liquid hydrocarbon exploration potential in cratonic region of the Tarim Basin inferred from gas condensate genesis. Fuel 160, 583–595. https://doi.org/10.1016/j.fuel.2015.08.023.

Zhang, Y.Y., Xi, K.L., Cao, Y.C., et al., 2021. The application of machine learning under supervision in identification of shale lamina combination types - a case study of Chang 7_3 sub-member organic-rich shales in the Triassic Yanchang Formation, Ordos Basin, NW China. Petrol. Sci. 18 (6), 1619–1629. https://doi.org/10.1016/j.petsci.2021.09.033.

Zheng, W.H., Tian, F., Di, Q.Y., et al., 2021. Integrated geological-geophysical characterizations of deeply buried fractured-vuggy carbonate reservoirs in Ordovician strata, Tarim Basin. Mar. Petrol. Geol. 99, 292–309. https://doi.org/10.1016/j.marpetgeo.2018.10.028.

Zhong, Z., Carr, T.R., 2016. Application of mixed kernels function (MKF) based support vector regression model (SVR) for CO_2 – reservoir oil minimum miscibility pressure prediction. Fuel 184, 590–603. https://doi.org/10.1016/j.fuel.2016.07.030.

Zhou, Y., Yang, F.L., Ji, Y.L., et al., 2020. Characteristics and controlling factors of dolomite karst reservoirs of the Sinian Dengying Formation, central Sichuan Basin, southwestern China. Precambrian Res. 343, 105708. https://doi.org/10.1016/j.precamres.2020.105708.

Zhou, Z., Wang, X.Z., Yin, G., et al., 2016. Characteristics and genesis of the (Sinian) Dengying Formation reservoir in central Sichuan, China. J. Nat. Gas Sci. Eng. 29, 311–321. https://doi.org/10.1016/j.jngse.2015.12.005.

Zhu, G.Y., Milkov, A.V., Zhang, Z.Y., et al., 2019. Formation and preservation of a giant petroleum accumulation in superdeep carbonate reservoirs in the southern Halahatang oil field area, Tarim Basin, China. AAPG Bull. 103 (7), 1703–1743. https://doi.org/10.1306/11211817132.

Zhu, X.F., Huang, Z., Shen, H.T., et al., 2012. Dimensionality reduction by mixed kernel canonical correlation analysis. Pattern Recogn. 45 (8), 3003–3016. https://doi.org/10.1016/j.patcog.2012.02.007.