Fuel 337 (2023) 127191

Contents lists available at ScienceDirect

Fuel

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

Full Length Article

Identification of coal structures by semi-supervised learning based on limited labeled logging data

Jinxiong Shi [a,b,c], Xiangyuan Zhao [d], Lianbo Zeng [c,*], Yunzhao Zhang [e], Shaoqun Dong [c]

[a] Key Laboratory of Exploration Technologies for Oil and Gas Resources of the Ministry of Education, Yangtze University, Wuhan 430100, China
[b] School of Geosciences, Yangtze University, Wuhan 430100, 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, China

ARTICLE INFO

Keywords:
Coal structure
Well logging data
Limited labeled-sample
Semi-supervised learning
Laplacian support vector machine

ABSTRACT

Coal structure is a critical parameter in coalbed methane (CBM) development due to its significant impacts on methane enrichment, fluid flow and hydraulic fracturing. Traditional statistical analysis and data-driven machine learning methods for coal structure identification are highly dependent on the labeled logging data and have potential limitations when labeled logging data is limited. To address this issue, this paper proposed a semi-supervised learning method based on Laplacian support vector machine (LapSVM) to identify coal structure by using few labeled logging data. By mining the structure information from abundant unlabeled data, LapSVM can improve the model performance and alleviate the over-reliance on labeled data. To evaluate and verify the effectiveness and reliability of the proposed LapSVM method in coal structure identification, datasets collected from 32 CBM wells in the southern Qinshui Basin, China, are utilized in this study. The particle swarm optimization (PSO) is adopted for parameter optimization of LapSVM models. For the LapSVM model, the addition of unlabeled data is conducive to enhance model accuracy, and unavoidably increases the computational cost at the same time. The comparison of training, testing and blind-well test results between the LapSVM and standard support vector machine (SVM) models indicates that the LapSVM outperforms traditional SVM and possesses higher accuracy and generalization in coal structure identification. It has been demonstrated that the LapSVM can be a reliable tool for coal structure identification when limited labeled logging data is available.

1. Introduction

Coal structure refers to the damage degree of in situ coals after underwent tectonic movements, and is a critical parameter reflecting the characteristics of coal reservoirs [1–3]. Initially, coal structure was proposed from the perspective of gas control in underground coal mines [4,5], and was classified as the undeformed coal and deformed coal. The undeformed coal (i.e., primary structural coal) is defined as coals with well-preserved original structures [6–8], including the bedding texture, lineation texture and primary cleat. By contrary, the deformed coal has characteristics of deformation and metamorphism, in which original structures are damaged under the action of tectonic stress [9,10]. The ductile and brittle deformations (e.g., crumple structure, schistose bedding, and exogenous fracture) are common in deformed coals. With the increase of deformation degree, deformed coals are typically classified as cataclastic, granulated and mylonitized coals [10–12]. Besides, some researches divided deformed coals into three series of deformations and ten classes [13,14]. Due to distinct differences of petrological, physical and mechanical properties, coal structure has significant influences on the adsorption, desorption, gas-bearing and percolation abilities of coalbed methane (CBM) reservoirs [15–18], and plays pivotal roles in the process of drilling, completion and hydraulic fracturing [19–22]. Exploitation practices of CBM indicate that the spatial distribution of coal structures possesses strong heterogeneity, which greatly restricts the CBM development [23–25]. Therefore, the accurate identification of coal structure can provide a better understanding for the variation of reservoir characteristics, and is significant to the effective development of CBM resources.

There are many different approaches for identifying coal structure, including the direct and indirect methods. The underground mine

[*] Corresponding author.
E-mail address: lbzeng@sina.com (L. Zeng).

https://doi.org/10.1016/j.fuel.2022.127191
Received 18 October 2022; Received in revised form 28 November 2022; Accepted 15 December 2022
Available online 22 December 2022

observation and core description are the most direct methods, and have relatively high identification accuracy [10]. However, both methods are difficult to implement in many situations hindered by the complex condition of underground coal mines and low recovery rate of coal cores. Indirect identification methods of coal structure mainly depend on the geophysical data including seismic interpretation and well logging analysis. Seismic data interpretation is an effective technique for predicting regional distribution of coal structures [3], but it is generally inefficient to obtain vertical distribution due to the limitation of data quality and resolution. The geophysical logging data, with advantages of good continuity, high vertical resolution and convenient data acquisition, has been widely applied in many fields of CBM development, such as the gas content estimation [26], physical and mechanical properties evaluation [10,27–29], coal macrolithotype prediction [30,31] and coal structure identification [10–12,32,33].

The key of coal structure identification through geophysical logging data is to establish the corresponding relationship between well logging parameters and coal structures [11,33]. To address this issue, numerous mathematical methods have been introduced to the quantitative identification of coal structure. Traditional methods mainly include the characteristic curve [12], empirical formula [15], and cross plot analysis [31] techniques. In practical application, all these methods depend heavily on the experience and knowledge of analysts, and commonly have shortcomings of high subjectivity, low efficiency and poor practicability. In addition, relationships between the logging data and coal structure are highly complex and generally exhibit nonlinear features. Most of these methods are ineffective in establishing the precise and universal mathematical relationship between logging parameter and coal structure limited by their weak data mining ability. Recently, with the rapid development of artificial intelligence, several machine learning methods have been used to develop coal structure identification models, such as the cluster analysis (CA) [10], linear discriminant analysis (LDA) [11], principal component analysis (PCA) [32], and kernel Fisher discriminant analysis (KFD) [33] algorithms. Compared with traditional mathematical methods, machine learning algorithms possess stronger ability in extracting hidden information from original data and dealing with complex nonlinear problems, which greatly improves the identification accuracy. In addition, machine learning methods are highly applicable benefiting from their learning autonomy, computing efficiency and adjustmental flexibility.

Currently, common machine learning methods for coal structure identification include the unsupervised learning and supervised learning algorithms [34–36]. Major differences between these two algorithms are reflected in the utilization of class-label information for modeling. Unsupervised learning algorithms partition unlabeled data (i.e., logging data without coal structure label) into subgroups based on similarities and differences of data structure features. Despite unsupervised learning algorithms have proven to be useful in areas without observed samples, it is difficult to evaluate the effectiveness of model performance since no specific output is provided. In contrast, supervised learning method utilizes labeled data (i.e., logging data with coal structure label) for identification model training. Compared with unsupervised learning, supervised learning algorithms usually have better performance because the labeled data can provide more accurate and reliable guidance for classification [37]. In addition, it is noteworthy that supervised learning algorithms are data-driven methods and require sufficient labeled samples for modeling to guarantee models can obtain strong generalization ability. However, labeled logging data are typically limited due to the high cost and time-consuming of manual labeling. Most supervised learning methods are unable to excavate enough classification information to develop a powerful identification model when there are too few labeled samples. Meanwhile, there are massive unlabeled logging data that containing abundant valuable information, which are not used for model development by supervised learning [38]. In this context, supervised learning methods cannot sufficiently exert their effects in coal structure identification. The limited labeled-sample problem is unavoidable in coal structure identification and should be given serious consideration.

Semi-supervised learning algorithm, a branch of machine learning, takes advantage of learning strategies from unsupervised learning and supervised learning methods [39]. In the modeling process, the semi-supervised learning utilizes both the feature information of labeled data and data structure of unlabeled data to construct classifiers [40,41]. Thus, semi-supervised learning methods overcome the shortcoming of unsupervised learning and supervised learning methods that completely relying on unlabeled or labeled data, and show strong ability in handling few labeled-sample recognition and classification issues [34]. A variety of semi-supervised learning algorithms have been proposed and can be roughly classified as the self-training, co-training, generative model, and graph-based model [42,43]. The Laplacian support vector machine (LapSVM) is a typical graph-based methods in semi-supervised learning [40]. It is an extension of traditional support vector machine (SVM), and introduces the structure information of unlabeled data to standard SVM in form of Laplacian manifold regularization to obtain more accurate classification. By combining classification strategies of SVM (kernel-based method) and semi-supervised learning, LapSVM generally outperforms other semi-supervised methods in highly nonlinear and few labeled-sample tasks [44–47]. In recent years, LapSVM has been gradually applied in petroleum geology, such as the lithology identification [35] and fracture predication based on well logging data [37]. The LapSVM provides a good opportunity for coal structure identification with limited logging data, whereas few studies have been performed to investigate its application.

In this paper, a semi-supervised coal structure identification method based on LapSVM was proposed to address the limited labeled-sample problem and improve the identification accuracy. Core samples and logging data collected from the Anze CBM field in the southern Qinshui Basin, China, were used for the development of LapSVM model. In modeling processes, the particle swarm optimization (PSO) technique was employed to derive optimal parameters of LapSVM model. To illustrate the significance of unlabeled data, the LapSVM model was trained with different sizes of labeled data and unlabeled data. Then, predicted results of the LapSVM-based model were compared with these of traditional SVM model to demonstrate the performance improvement of the proposed LapSVM method. In addition, blind-well tests were also performed to verify the reliability and generalization of the built LapSVM model. Finally, a sensitivity analysis was conducted to evaluate the variable importance for LapSVM model.

2. Data acquisition and analysis

2.1. Geological background

In this study, core samples and well logging data used for coal structure identification were collected from the Permian Shanxi Formation of the Anze CBM field in the southern Qinshui Basin, China (Fig. 1). The Anze block has experienced multistage tectonic movements after the coal-bearing strata deposited and now is manifested as a SE plunging syncline with an average dip of formations less than 5° [48]. A series of small normal faults with NNE-SSW and NE-SW strikes and gentle folds with axial oriented NNE-SSW are widespread in this region [48,49], which have significant impacts on coal deformation and coal structure [3,22]. In the study area, the Permian Shanxi Formation is one of the main coal-bearing strata [49,50], in which the No.3 coal seam is the major layer for CBM production (Fig. 1). The No.3 coal seam is developed in interdistributary bay environments of lower delta plain, and is mainly composed of vitrinite, inertinite and some minerals. Macrolithotypes of No.3 coal seam are dominated by semi-bright and bright coals, and coal ranks are semi-anthracite to anthracite with the maximum vitrinite reflectance ($R_{o,max}$) ranging from 1.9 % to 2.7 % [49]. The current burial depth of the No.3 coal seam is approximately 600 m to 1300 m, and gradually increases from northwest toward

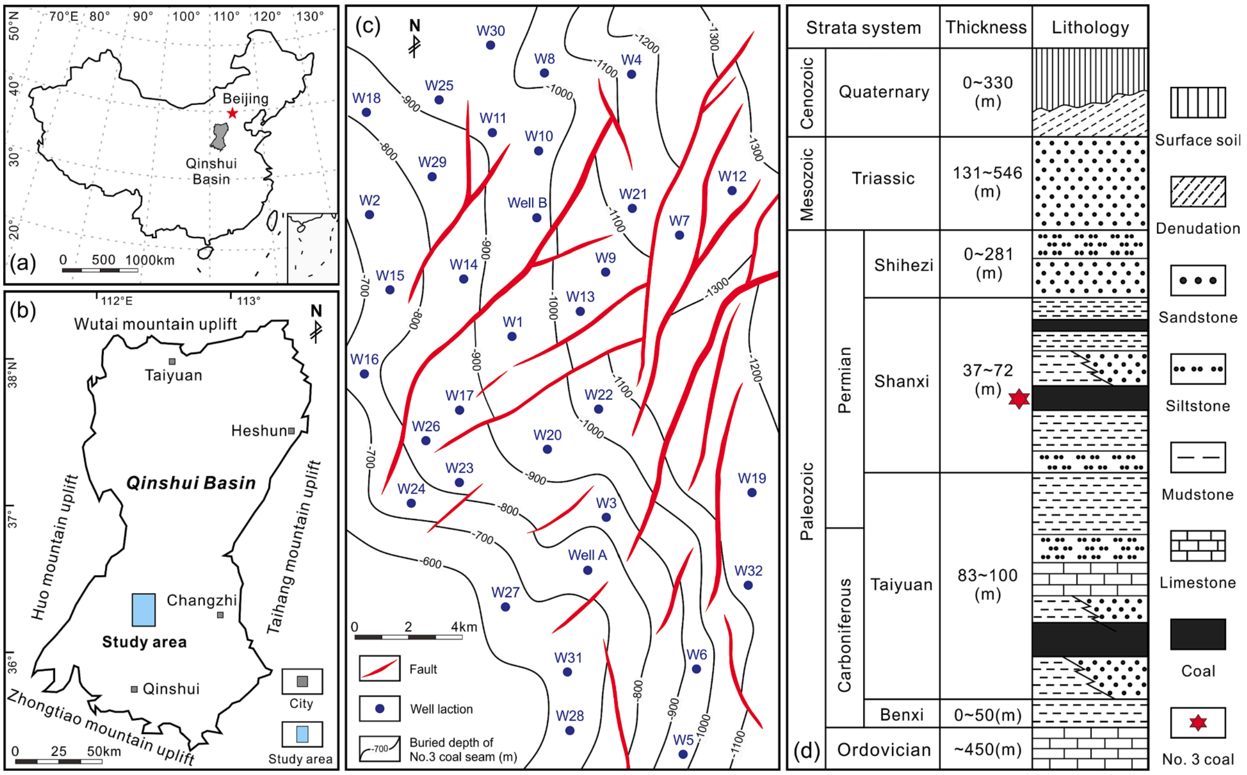


Fig. 1. Locations of (a) the Qinshui Basin in China and (b) the Anze CBM field in southern Qinshui Basin. (c) Main geologic structure, burial depth of No. 3 coal seam and well distribution of the Anze block. (d) Stratigraphic characteristics of the southern Qinshui Basin.

southeast of the study area (Fig. 1). Drilling data indicates that the thickness of No.3 coal seam ranges from 2.36 to 6.82 m, with an average of 5.26 m.

2.2. Types and characteristics of coal structure

Due to the inferior mechanical strength of coal seam, the recovery rate of coal cores is generally low in No. 3 coal seam of the Anze block. Totals of 52 coal cores with lengths of 16.3 m were obtained from 5 CBM wells in the study area. The length of each sample is approximately between 0.20 m and 0.45 m.

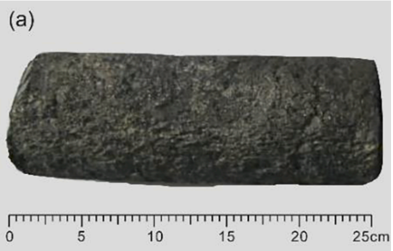
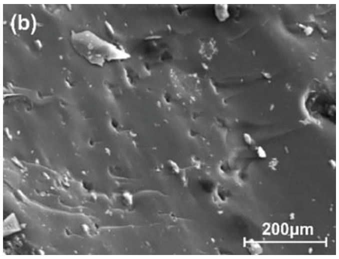
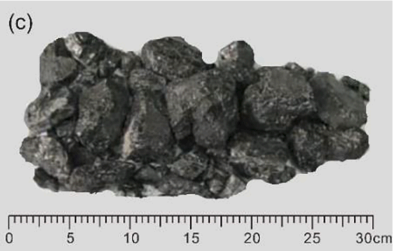
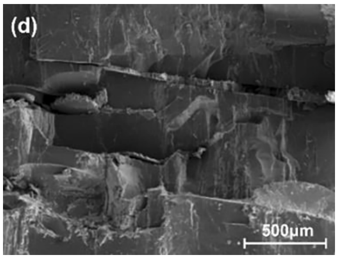
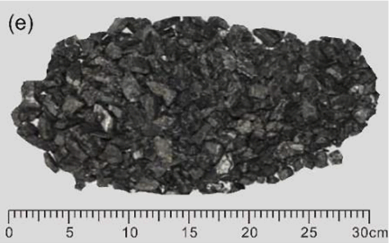
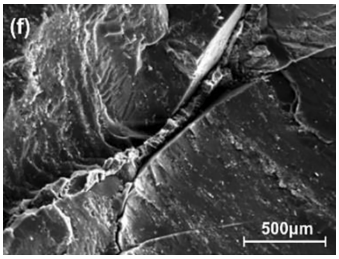
Coal structure of the No.3 coal seam in the Anze block can be classified into the undeformed coal, cataclastic coal and granulated coal following Chinese National Standard GB/T 30050-2013, whereas the mylonitized coal is not found. Macroscopic and microscopic observation shows that core samples with different coal structures have significant discrepancy in the macrolithotype distinguishability, bedding integrity, degree of fragmentation, development degree of fracture and crumple, and rock mechanics strength (Table 1). Undeformed coals commonly possess intact block structures (Table 1a) and coal macrolithotypes are easy to distinguish. Original bedding textures, cleats and organic pores are well preserved, while tectonic fractures and crumple structures are rarely seen (Table 1b). These coals are firm and can hardly be crushed by hand. Cataclastic coals present angular block structures (Table 1c) with relatively visible boundaries between different macrolithotypes. Primary sedimentary beddings and cleats are damaged to some extent (Table 1d), but can still be distinguished. These coals are usually cut by different groups of cleats and tectonic fractures, and can be broken into cm-scale fragments by hand. Granulated coals show granular or lumpy structures (Table 1e), and it is difficult to distinguish coal macrolithotypes. The sedimentary textures have been severely damaged and are indiscernible, and tectonic fractures (Table 1f) and crumple structures are well developed. Granulated coals are easily crushed into mm-scale fragments and even powders by hand.

2.3. Logging responses of coal structure

Well logging data used for modeling were collected from 32 CBM wells drilled in the study area (Fig. 1c), including 5 coring wells (aforementioned in Section 2.2) and 27 non-coring wells. Logging types of each well consist of the natural gamma ray (GR), caliper (CAL), interval transit time (AC), compensation density (DEN), and deep lateral resistivity (LLD) logs. A total of 1575 datasets were recorded with a sampling interval of 0.10 m. For a semi-supervised learning algorithm, the learning effect of the trained model depends largely on the comprehensiveness of input variables. Hence, all five types of well logs were used as input data for modeling to guarantee more information can be learned by semi-supervised models. In addition, due to various interfering factors during the logging and data collection process, outliers and unrealistic values inevitably exist within well logs, which will greatly affect the accuracy of model training. Hence, to ensure the comprehensiveness and validity of logging data, the data cleaning and quality evaluation of raw dataset are necessary. In this study, the Three-sigma Criterion was employed to deal with these abnormal data in well logs. Based on the standard deviation (SD) of individual well logs, all logging values outside the range of ± 3.0 SD were designated as outliers and excluded from raw data. Through the preprocessing, totals of 1530 datasets were finally obtained and the statistical properties of the preprocessed datasets were presented in Fig. 2. It can be seen that all well logs have wide a range of logging values and roughly follow a normal distribution, which are available for modeling.

After the cleaning of abnormal data, coal samples of 5 coring wells were depth-matched to well logs according to core description reports. On this basis, a total of 158 labeled datasets were obtained, including 50 undeformed coal, 55 cataclastic coal and 53 granulated coal datasets. The information of each dataset comprises the logging values of five well logs (the GR, CAL, AC, DEN, and LLD logs) and corresponding labels of coal structure. For instance, totals of 6 coal cores comprising 27 labeled datasets were collected from well 1. The vertical profile of coal

Table 1
Types and characteristics of coal structures in the No.3 coal seam of southern Qinshui Basin.

Coal structure	Core photos	SEM photos	Characteristic description
Undeformed coal			(1) Macrolithotypes are easy to distinguish;(2) Intact large-scale block structure;(3) Original banding textures and cleats are distinctly;(4) Tectonic fractures are undeveloped;(5) Hard to crush by hand.
Cataclastic coal			(1) Macrolithotypes are relatively clear;(2) Angular block structure;(3) Original banding structures can be traced intermittently;(4) Tectonic fracture can be observed;(5) Can be crushed into cm-scale fragments.
Granulated coal			(1) Macrolithotypes are indistinguishable;(2) Granular or lumpy structures;(3) Tectonic fractures are well-developed;(4) Easy to be crushed into mm-scale fragments and powders.

Note: (a) Undeformed coal with block structure, core observation, 1089.70 m; (b) Well-preserved organic pores in undeformed coal, SEM image, 1102.65 m; (c) Cataclastic coal with angular block structure, core observation, 1136.23 m; (d) Original cleats are damaged in cataclastic coal, SEM image, 1057.69 m; (e) Granulated coal with granular structure, core observation, 1135.32 m; (f) Two tectonic fractures in granulated coal, SEM image, 1178.30 m. SEM = scanning electron microscope.

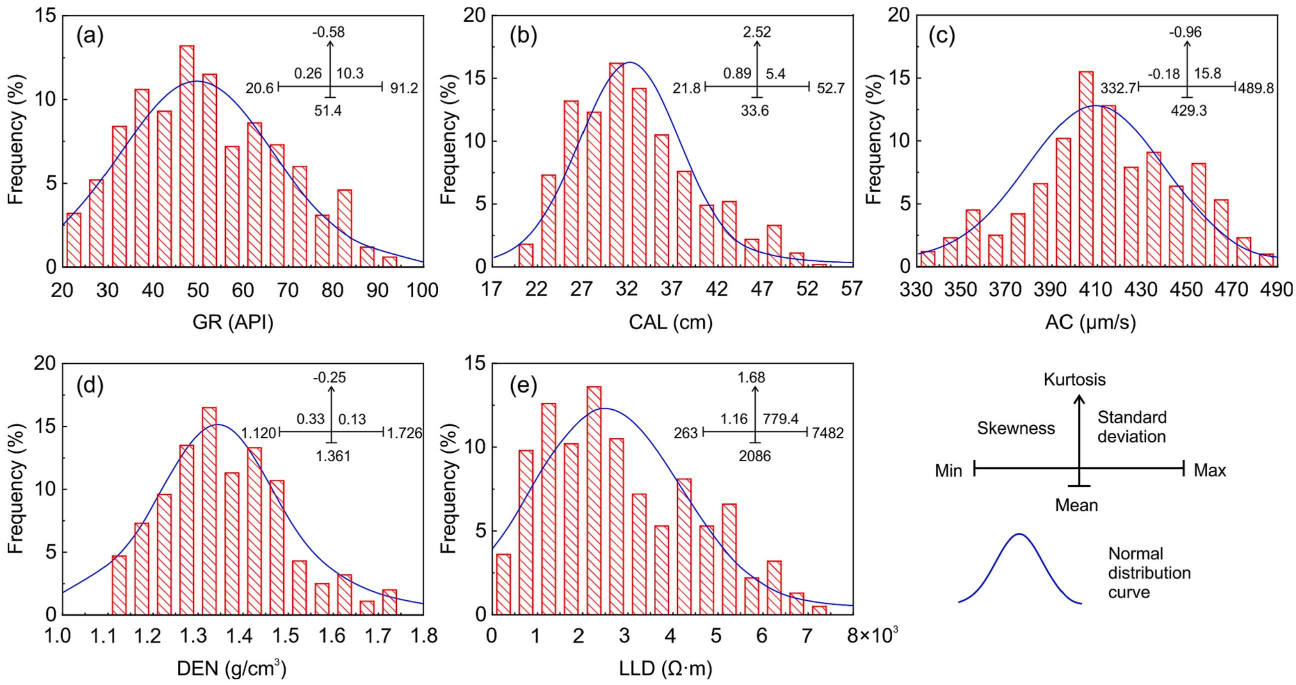


Fig. 2. Frequency distribution and statistical properties of the (a) GR, (b) CAL, (c) AC, (d) DEN, and (e) LLD logs. Min = minimum; Max = maximum.

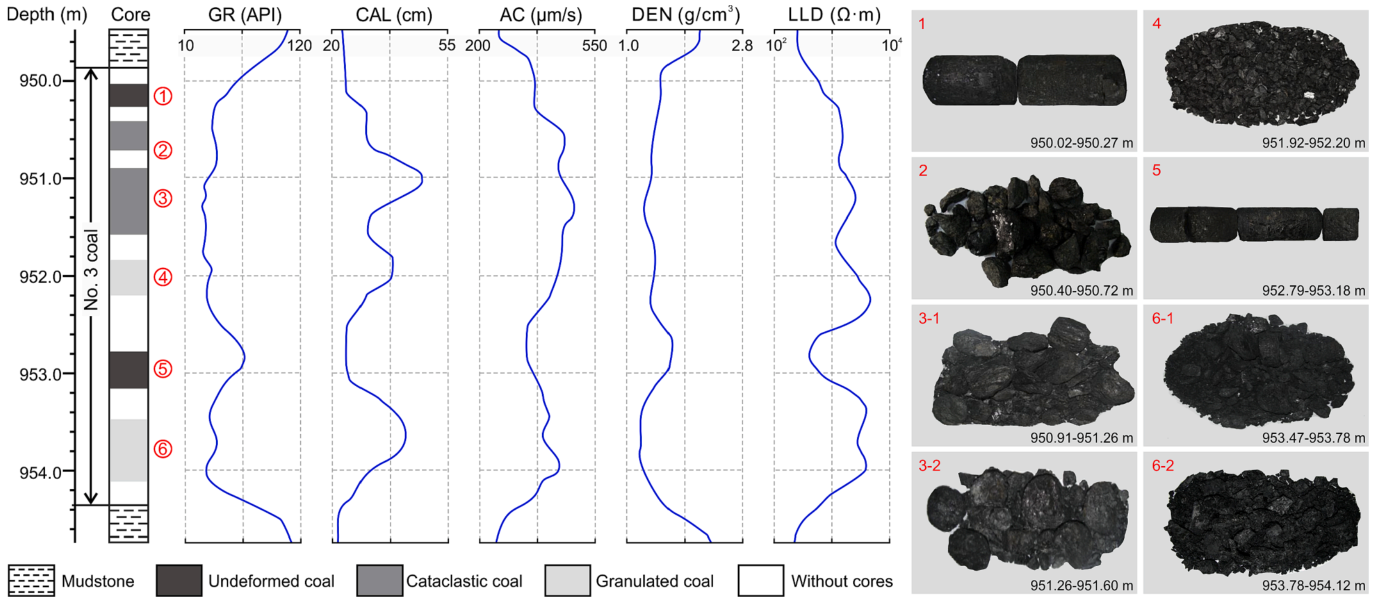


Fig. 3. Coal structure distribution and well logging response of No.3 coal seam of well 1 in the study area.

structures is undeformed, cataclastic, cataclastic, granulated, undeformed, and cataclastic coals from the top to bottom (Fig. 3). It can be observed that logging responses differ from different coal structures and deformed coals commonly have low values of GR and DEN, and high values of CAL, AC and LLD. To more intuitively reflect the relationship between the logging data and coal structures, value ranges of all well logs for each coal structure are presented in Fig. 4, where the diagonal line are boxplots of well logging data and others are scatter cross-plots of pairwise variables. From the perspective of average value, the GR and DEN logs gradually decrease from the undeformed coal to granulated coal, while the CAL, AC and LLD logs present a gradual increasing trend. These logging response characteristics of coal structures basically consist with previous research results [11–13]. Nevertheless, it can be found that large numbers of data points belonging to different coal structures possesses similar logging responses, and there are certain degrees of value overlaps among different coal structures. These overlaps make it difficult to determine the accurate decision boundary for classification and will cause multiple interpretation problems in coal structure identification. This suggests the attempt to classify a multivariate and nonlinear dataset by traditional statistical analysis based on correlation coefficient is infeasible, particularly when there only are a small number of labeled logging data.

3. Methodology

3.1. Laplacian support vector machine

Laplacian support vector machine (LapSVM) is a graph-based semi-supervised classification method proposed by Belkin et al. [40]. It is built on the standard SVM framework and incorporates a manifold regularizer to SVM model. LapSVM inherits the advantage of traditional SVM and overcomes the insufficient training problem when the labeled data is limited. To better illustrate the principle of LapSVM, we will briefly introduce the SVM first.

Support vector machine (SVM) is a kernel-based supervised learning algorithm proposed by Cortes and Vapnik [51]. The basic idea of SVM is to project the input labeled data from original space into the high-dimensional feature space by nonlinear mapping, and then construct an optimal separating hyperplane with maximum classification interval [52,53]. Thus, the nonlinear relationship of original data is indirectly transformed into a linear one. Given a training dataset of l labeled samples $T = \{x_i, y_i\}_{i=1,2,\cdots,l}$, where $x_i \in R^N$ is the feature vector of input data, y_i is the class label of input data, and N is the dimension of x_i. The goal of the SVM is to find the optimal separating hyperplane f in feature space, which can be derived by solving the following optimization problem:

$$f^* = \underset{f\in H_k}{\operatorname{argmin}}\frac{1}{l}\sum_{i=1}^{l}V(x_i, y_i, f) + \gamma_H\|f\|_H^2 \tag{1}$$

In this formula, the first term $V(x_i, y_i, f)$ is the hinge loss function measuring the error between predicted label and true label, and can be expressed as:

$$V(x_i, y_i, f) = max\{0, 1 - y_i f(x_i)\} \tag{2}$$

In Eq. (2), $f(x) = w\cdot\phi(x) + b$ is the decision function, where w is a weight vector, $\phi(x)$ is the nonlinear mapping, and b is a bias term. According to the Representer Theorem [54], w can be expressed as $w = \sum_{i=1}^{l}\alpha_i\phi(x_i)$, where $\alpha_i = \{\alpha_1, \alpha_2, \cdots, \alpha_l\}$ is the Lagrange multiplier. In addition, the kernel function $k(x_i, x_j)$, defined as the dot product operation of $\langle\phi(x_i), \phi(x_j)\rangle$, is utilized to reflect the nonlinear mapping $\phi(x)$ implicitly. Thus, decision function $f(x)$ can be further written as:

$$f(x) = \sum_{i=1}^{l}\alpha_i k(x_i, x) + b \tag{3}$$

The common kernel function includes the linear kernel, sigmoid kernel, radial basis function (RBF) kernel and polynomial kernel. Compared with other kernels, the RBF kernel has better performance in handling nonlinear problems since it has fewer parameters and smaller complexity [55,56]. Hence, the RBF kernel $k(x_i, x) = exp(-g\|x_i - x\|^2)$ is selected as the kernel function in this work, where $g = 1/2\sigma^2$, and σ is the width parameter of RBF.

The second term $\|f\|_H^2$ is the squared norm of function f in the reproducing kernel Hilbert space (RKHS), and is used to preserve the smoothness of solution in RKHS. γ_H is the corresponding weight that controls the functional complexity of decision function f in RKHS. The $\|f\|_H^2$ is written as:

$$\|f\|_H^2 = \|w\|^2 = w^T w = (\varphi\alpha)^T(\varphi\alpha) = \alpha^T k\alpha \tag{4}$$

The optimization problem in Eq (1) can be transformed into the following constrained optimization problem by introducing the slack

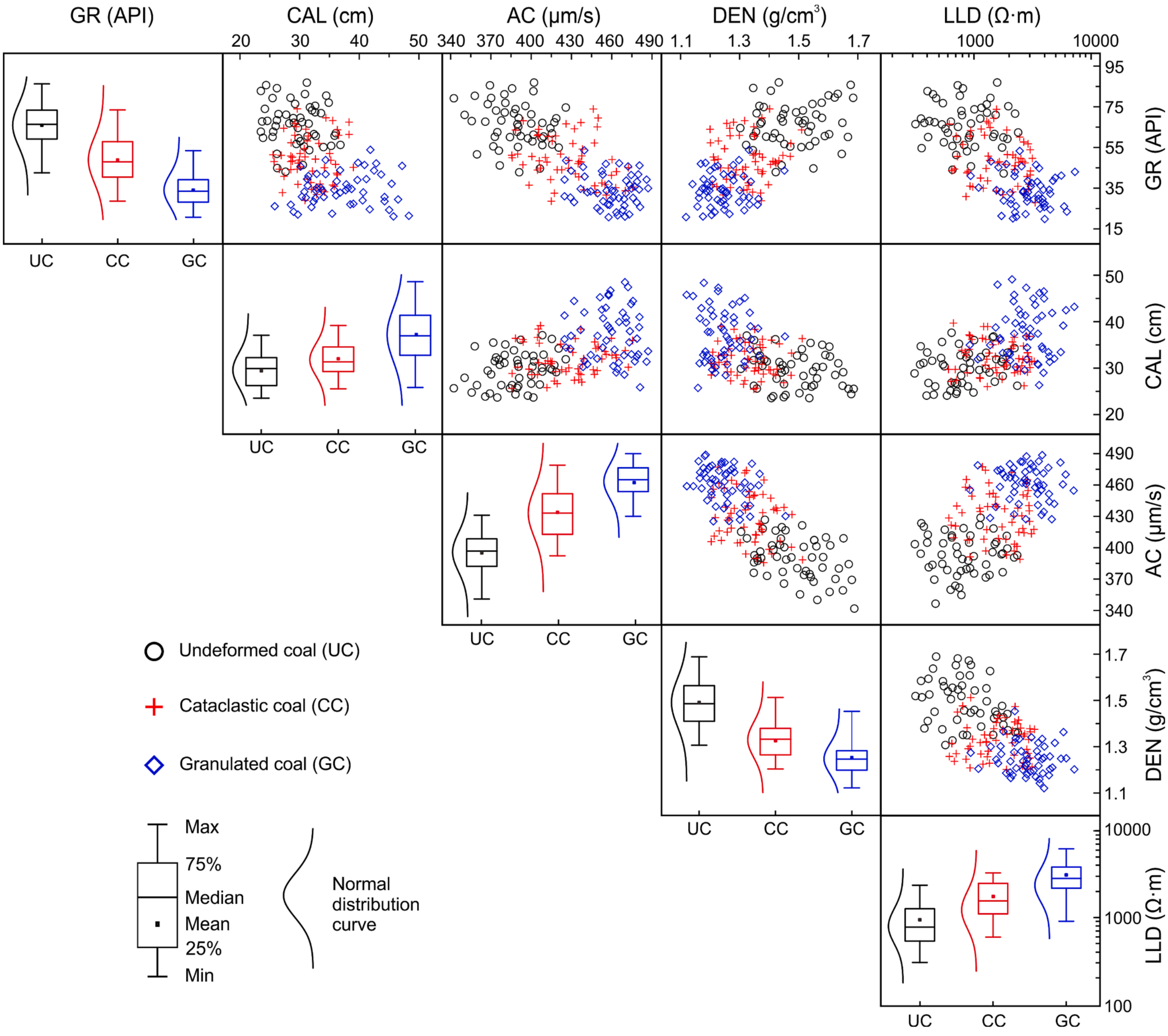


Fig. 4. Pairwise scatterplots and boxplot of different well logs by coal structures. The diagonal line are boxplots of logging data and others are scatter cross-plots of pairwise variables against coal structures.

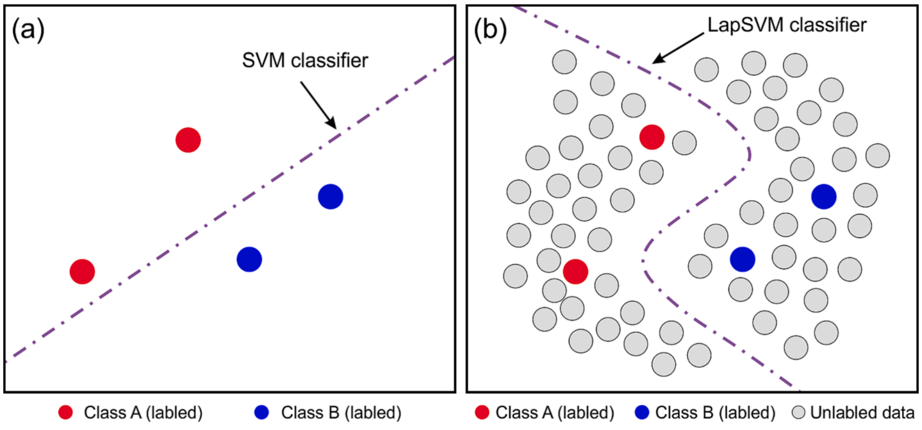


Fig. 5. Schematic diagram of the classification strategy of SVM and LapSVM algorithms.

variable $\xi_i = \{\xi_1, \xi_2, \cdots, \xi_l\}$:

$$f^* = \underset{\alpha\in R^l, \xi_i\in R}{\operatorname{argmin}}\frac{1}{l}\sum_{i=1}^{l}\xi_i + \gamma_H\alpha^T K\alpha$$

$$s.t.\begin{cases} y_i\left(\sum_{j=1}^{l}\alpha_i k(x_i, x) + b\right) \geq 1 - \xi_i \\ \xi_i \geq 0, i = 1, 2, \cdots l \end{cases} \tag{5}$$

By solving the optimization problem in Eq (5), the decision function can be obtained and its value can be calculated from samples in the low-dimensional space to realize nonlinear classification. Therefore, SVM processes unique advantages in handling high dimensional and nonlinear problems. However, its performance will be greatly restricted when labeled samples for training are insufficient.

Different from the SVM that are trained only with labeled data, LapSVM takes the additional information contained in unlabeled data into account to reveal more about data structure (Fig. 5). In the LapSVM algorithm, it specifies a standard SVM as the learner core and introduces a graph-based regularizer to constrain the SVM's classification function be smooth along the graph. To effectively utilize the unlabeled data, there two basic assumptions about data distribution in LapSVM algorithm, namely, the cluster assumption and manifold assumption [40]. Specifically, cluster assumption states that samples in the same cluster tend to have the same class label, and manifold assumption states that the intrinsic data structure of high-dimensional data can be roughly represented by a low-dimensional manifold.

Given the additional dataset of u unlabeled samples $U = \{x_i\}_{i=l+1,l+2,\cdots,l+u}$, and normally $u \gg l$. Based on the standard SVM framework (Eq. (1)), a manifold regular term is added to LapSVM to achieve semi-supervised classification. In LapSVM, the decision function f can be achieved by minimizing:

$$f^* = \underset{f\in H_K}{\operatorname{argmin}}\frac{1}{l}\sum_{i=1}^{l}V(x_i, y_i, f) + \gamma_H\|f\|_H^2 + \gamma_M\|f\|_M^2 \tag{6}$$

where $\|f\|_M^2$ is the manifold regular term related to inherent data structure of all samples and is used to keep decision function smooth along the intrinsic manifold [44]. γ_M is the corresponding regularization parameter that controls the intrinsic geometric complexity of marginal distribution. The $\|f\|_M^2$ is expressed as:

$$\|f\|_M^2 = \frac{1}{(l+u)^2}\sum_{i,j=1}^{l+u}W_{ij}\left(f(x_i) - f(x_j)\right)^2 = \frac{f^T Lf}{(l+u)^2} \tag{7}$$

where W_{ij} is the edge weight in data adjacency graph, $W = [W_{ij}]_{n\times n}$ is the edge weight matrix, and $n = l + u$. L is the graph Laplacian matrix, D is a diagonal matrix with diagonal elements $D_{ii} = \sum_{j=1}^{l+u}W_{ij}$, and $L = D - W$. Decision function $f = \sum_{i=1}^{l+u}\alpha_i k(x_i, x) = K\alpha$, and $K_{i,j} = k(x_i, x_j), i, j = 1, 2, \cdots, i + u$.

Then, by introducing slack variables ξ_i, the unconstrained optimization problem of Eq. (7) can be written as [40]:

$$f^* = \underset{\alpha\in R^{l+u}, \xi_i\in R}{\operatorname{argmin}}\frac{1}{l}\sum_{i=1}^{l}\xi_i + \gamma_H\alpha^T K\alpha + \gamma_M\frac{\alpha^T KLK\alpha}{(l+u)^2}$$

$$s.t.\begin{cases} y_i\left(\sum_{j=1}^{l+u}\alpha_i k(x_i, x) + b\right) \geq 1 - \xi_i \\ \xi_i \geq 0, i = 1, 2, \cdots l \end{cases} \tag{8}$$

Finally, the objective function Eq. (8) can be further simplified to a Lagrangian function:

$$L(\alpha, \beta) = \frac{1}{2}\alpha^T(2\gamma_H K + 2\gamma_M KLK)\alpha - \alpha^T KJ^T Y\beta + \sum_{i=1}^{l}\beta_i \tag{9}$$

where $J = [I, 0]$ is a matrix of $l \times (l+u)$, I is an $l \times l$ identity matrix, $Y = diag\left(y_1 y_2, \cdots, y_l\right)$, and β_i is the Lagrange multiplier.

The coefficient α is computed as:

$$\alpha = \left(2\gamma_H I + \frac{2\gamma_M}{(l+u)^2}LK\right)^{-1} J^T Y\beta \tag{10}$$

Thus, the primal problem in Eq. (6) is reduced to an optimization task of finding optimal coefficient vector α. After solving Eq. (10), the decision function f can be obtained and used to identify coal structure.

3.2. Particle swarm optimization

Parameter optimization is a key issue for improving the overall performance of classification models. To illustrate the influence of hyperparameters, namely, the regularization parameters γ_H, γ_M, and RBF kernel parameter g, on the LapSVM model, the particle swarm optimization algorithm was employed in this work for hyperparameter optimization.

Particle swarm optimization (PSO) algorithm is a population-based search technique proposed by Kennedy and Eberhart [57], which originates from research on the socially-coordinated behavior of animal swarms. The PSO comprises several particles initialized randomly in the search space with their own position and velocity. The particles represent the potential solution of the extremum optimization problem, and are used to compute the global optimum for fitness function. Given a D-dimensional search space, the population $X = \{X_1, X_2, \cdots, X_N\}$ is the combination of N particles, $X_i = \{x_{i1}, x_{i2}, \cdots, x_{iD}\}$ and $V_i = \{v_{i1}, v_{i2}, \cdots, v_{iD}\}$ are the position and corresponding velocity of the i-th particle in the search space. During iterations, the position and velocity of each particle are updated according to the distance to its personal best position and distance to global best position. The update formulas are described as follows:

$$V_i(t+1) = \omega V_i(t) + c_1 r_1(pbest_i - x_i(t)) + c_2 r_2(gbest - x_i(t)) \tag{11}$$

$$x_i(t+1) = v_i(t+1) + x_i(t) \tag{12}$$

where $pbest_i$ is the personal best position searched by i-th particle, $gbest$ is the best-so-far position searched by entire population. c_1 and c_2 are acceleration coefficients, and t is the number of current iterations. r_1 and r_2 are independent random numbers. ω is the inertia weight, and $\omega(t) = \omega_{\max} - (\omega_{\max} - \omega_{\min})t/t_{\max}$, where $t_{\max}$ is the maximum number of iterations.

During the parameter searching process by PSO, the position of particle is continuously updated by changing velocity, and finally converges at a global optimum in search space. In this work, the particle's position is the vector of γ_H, γ_M and g, and is denominated as $P\ (\gamma_H, \gamma_M, g)$. The average accuracy of 5-fold cross-validation is set as the fitness function to assess the performance of optimization process. The PSO algorithm procedure terminates when a minimum error threshold or the maximum iteration is achieved.

3.3. Workflow of coal structure identification

To improve the accuracy of coal structure identification under the situation of limited labeled logging data is available, this paper proposed a LapSVM-based coal structure identification method based on the semi-supervised learning strategy. Datasets used for modeling consist of 158 labeled data and 1372 unlabeled data, of which the latter is over 8 times more than the former. Input variables include the GR, CAL, AC, DEN, and LLD logs. Besides, other eight sets of data with different labeled-

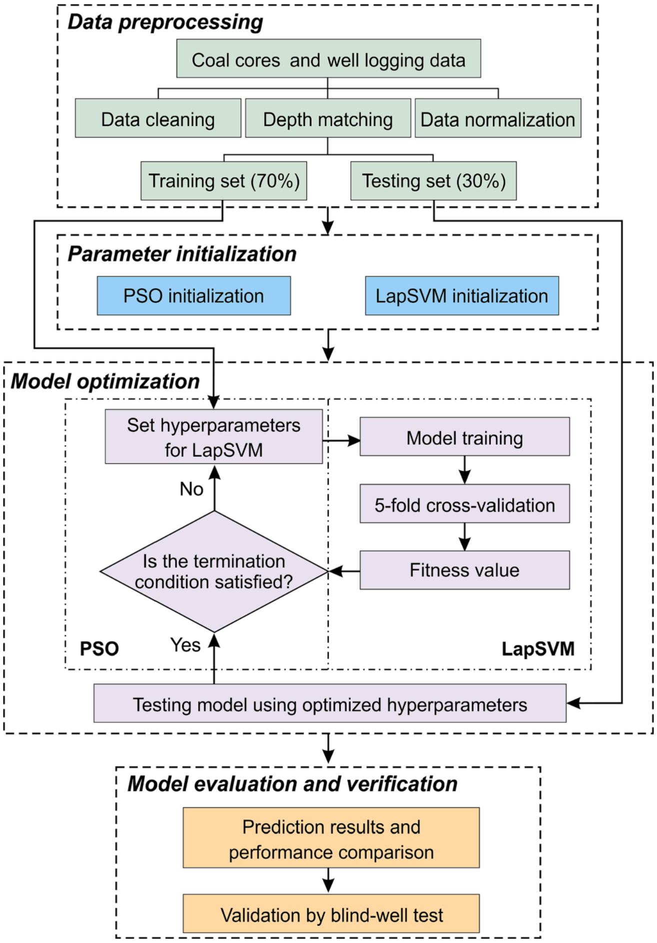

Data preprocessing
Coal cores and well logging data
Data cleaning
Depth matching
Data normalization
Training set (70%)
Testing set (30%)
Parameter initialization
PSO initialization
LapSVM initialization
Model optimization
Set hyperparameters for LapSVM
Model training
No
5-fold cross-validation
Is the termination condition satisfied?
Fitness value
PSO
Yes
LapSVM
Testing model using optimized hyperparameters
Model evaluation and verification
Prediction results and performance comparison
Validation by blind-well test

Fig. 6. Workflow of coal structure identification by LapSVM and PSO.

unlabeled data ratios are designed to investigate the impact of unlabeled data on model performance. The development, optimization and validation of the LapSVM model were conducted in the Matlab 2016b environment running Windows system with a 2.4 GHz CPU. The overall flowchart of the LapSVM-based coal structure identification method is shown in Fig. 6. The specific workflow is as follows:

Step 1. Data preprocessing. Input datasets are first standardized to eliminate the order-of-magnitude differences between well logs and reduce the computational complexity. The standardization function $x_{nor} = (x - x_{min})/(x_{max} - x_{min})$ is adopted to normalize each well logs into the range of [0,1], where x_{nor}, x_{min}, x_{max} are the normalized, minimum and maximum values, respectively. Then, input datasets are randomly portioned into the training set (~70 %) and testing set (~30 %) by a module *randperm* in Matlab to avoid model over-fitting.

Step 2. Parameter initialization. To reconcile the model performance and operational efficiency, the parameters of PSO are set as $t_{max} = 100$, $N = 20$, and parameters of LapSVM are set as $\gamma_H \in [0.01, 5]$, $\gamma_M \in [0.01, 6]$, $g \in [0.01, 20]$.

Step 3. Model construction and optimization. The training data are firstly assigned into the initial LapSVM model for training. In the training process, the PSO randomly assigns hyperparameters to LapSVM model, and the performance of this set of hyperparameter is estimated by the average accuracy of 5-fold cross-validation. The optimum hyperparameter is obtained when the termination condition is satisfied. Then, the trained model is applied to testing datasets to verify its classification performance.

Step 4. Model evaluation and verification. The performance of LapSVM model is compared with traditional SVM model for coal structure identification to demonstrate the superiority of the proposed identification method. Further, the built LapSVM model is applied to blind-well tests to validate its stability and generalization ability.

3.4. Performance evaluation indicators

For multi-classification problems, the confusion matrix and its associated evaluation parameters provide an effective approach to appraise the model performance. Commonly used evaluation indicators include the accuracy, precision, recall, and F1-score [58], which assess the model performance from different perspectives. Specifically, accuracy is the ratio of correctly classified samples and total samples, and measures the overall classification accuracy. Precision refers to the ratio of correctly classified samples and samples classified into this class, and reflects the accuracy of a specific classification result. Recall is the ratio of correctly classified samples and total samples in this class, and reflects the classification accuracy of a specific class. F1-score is the harmonic mean of precision and recall, and is known as a balanced indicator. The higher of indicator values (close to 1) imply that the model has better classification performance. The expression of these evaluation indicators is defined as follows in Table 2.

4. Results and discussion

4.1. Determination of optimal parameters

The previously defined hyperparameters γ_H, γ_M, and g have significant impacts on the classification accuracy and generalization ability of LapSVM model. Hereinto, the regularization parameter γ_H controls the smoothness and complexity of separating hyperplane in ambient spaces. When the value of γ_H is small, the hyperplane complexity is strong. The classification accuracy of training samples is high at this time, whereas the accuracy of testing samples is quite low, indicating the model is over-fitting (Fig. 7a). With the increase of γ_H, the complexity of hyperplane will reduce and the testing accuracy increases sharply. When γ_H increasing to a certain value, the hyperplane will be too smooth to capture classification features, and the under-fitting is prone to occur (Fig. 7a). The regularization parameter γ_M determines the contribute of Laplacian-based regularizer to separating hyperplane in the intrinsic space. As the γ_M increases, the more manifold structure information of data will be considered for constructing the hyperplane. When the γ_M is too large, LapSVM forces a solution that is smoother relative to the intrinsic geometry and ignores the labeled data, so that its classification effect will be similar to the unsupervised learning (Fig. 7b). In particular, when the γ_M reduces to zero, the LapSVM model will not impose manifold constraints during training and degenerate into a standard SVM. The kernel parameter g controls the geometric feature of RBF kernel and further determine the action scope of kernel function. The smaller the g is, the wider the action scope of kernel function and the easier the model is to under-fitting. On the contrary, the over-fitting will occur when the value of g is too large (Fig. 7c). Therefore, the choose of

Table 2
Definitions of performance evaluation indicators used in this work.

Evaluation indicators	Mathematical expression
Accuracy	$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$
Precision	$Precision = \frac{TP}{TP + FP}$
Recall	$Recall = \frac{TP}{TP + FN}$
F1-score	$F1 - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$

Note: TP = True Positives; FP = False Positives; TN = True Negatives; FN = False Negatives. TP, FP, TN, FN represent the numbers of positive samples correctly predicted, negative samples incorrectly predicted as positive samples, positive samples incorrectly predicted as negative samples, and negative samples correctly predicted, respectively.

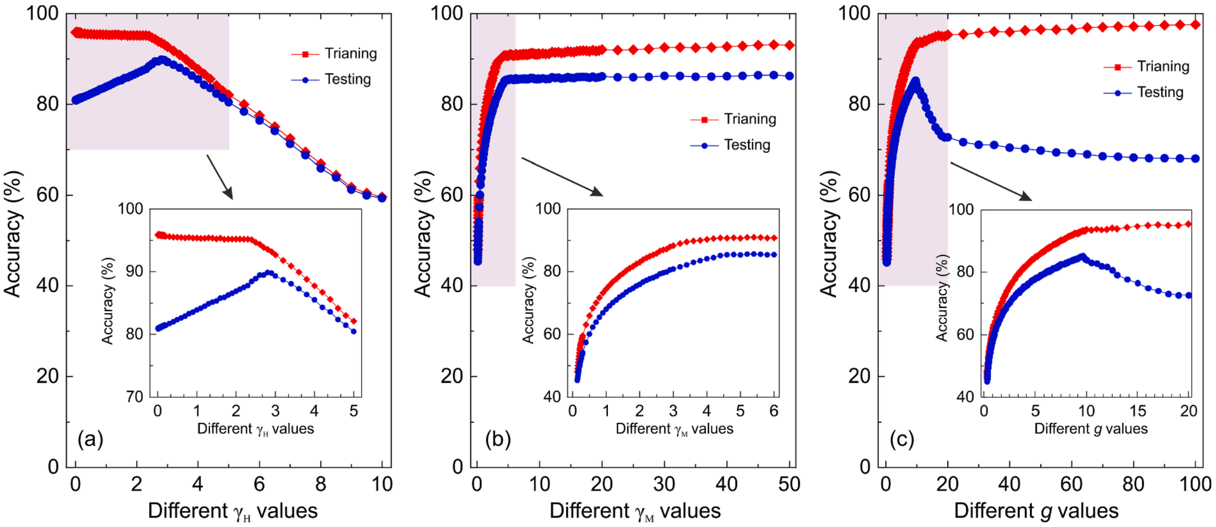


Fig. 7. Training and testing accuracies of LapSVM models with different values of hyperparameters. (a) Default γ_M, g and different γ_H values; (b) Default γ_H, g and different γ_M values; (c) Default γ_H, γ_M and different g values.

appropriate hyperparameters for the LapSVM model is critical to achieve an optimal classifier.

It is noteworthy that the effect of hyperparameters γ_H, γ_M, and g on LapSVM is not independent, and they collectively balance the accuracy and generalization of models. Meanwhile, these hyperparameters commonly have a wide range of values, and the parameter optimization process requires large amounts of computing resources. Accounting the model performance and operational efficiency, searching ranges of hyperparameters in LapSVM were set as $\gamma_H \in [0.01, 5]$, $\gamma_M \in [0.01, 6]$, and $g \in [0.01, 20]$ according to numerous test results (Fig. 7). The PSO algorithm was adopted for parameter optimization of the LapSVM model and the searching process of optimal hyperparameters was presented in

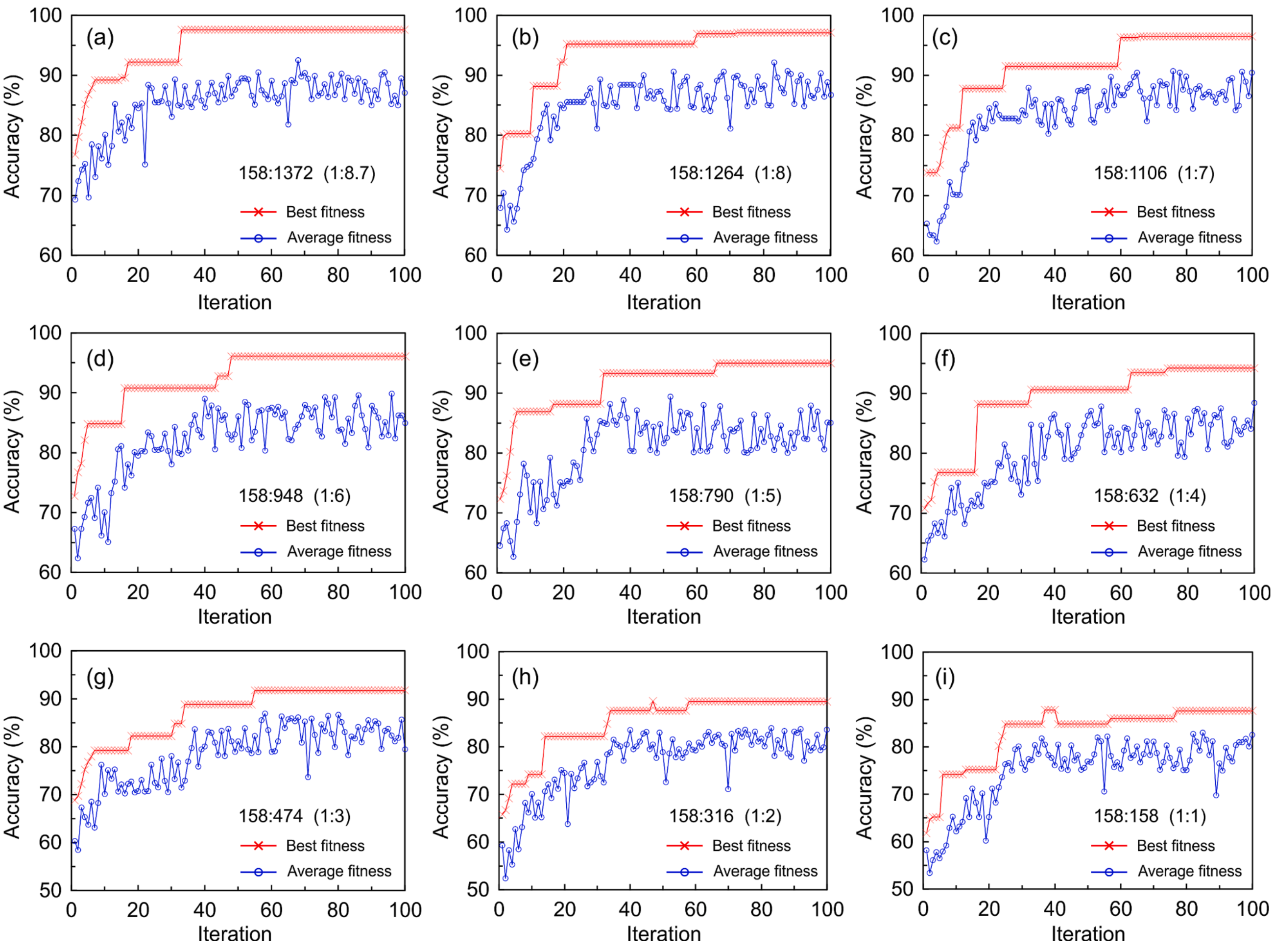


Fig. 8. Searching process of optimal hyperparameters in LapSVM models with different ratios of labeled-unlabeled data by PSO algorithm.

Fig. 8a. It can be found that the fitness value (average accuracy of 5-fold cross-validation) increases correspondingly with the iteration process and finally converges to a stable value, which indicates that the parameter search is featured with optimal results. By the optimization through PSO, optimal hyperparameters of the LapSVM model were obtained as $[\gamma_H, \gamma_M, g] = [0.89, 3.06, 7.65]$.

4.2. Impact of unlabeled data on LapSVM

By utilizing large amounts of unlabeled data, LapSVM mines the structure information of data to enhance the model performance and alleviates the over-reliance on labeled data. Hence, it is necessary to illustrate the effect of unlabeled data on LapSVM model performance. As shown in Table 3, eight sets of data with different labeled-unlabeled data ratios were designed, of which the ratios varied from 1:1 to 1:8. The minimum ratio was set as 1:1 because it cannot be called a few labeled-sample issue when the amount of unlabeled data was less than that of labeled data. It should be noted that the labeled datasets in each set of data are the same (i.e., 158 labeled samples mentioned in Section 2.3), and the unlabeled datasets are randomly sampled from 1372 unlabeled data. Under the same running environment, each set of data was trained and tested twenty times individually to investigate how the unlabeled data impacted the performance of LapSVM in terms of the classification accuracy, computation time and memory consumption. The hyper-parameters of each model were optimized by PSO (Fig. 8b-i).

Results of independent experiments on eight sets of data are presented in Fig. 9. It can be seen that the classification accuracy become higher with the unlabeled data increase (Fig. 9a). Average accuracies of LapSVM models increase rapidly from 84.6 % to 95.5 % before the labeled-unlabeled data ratio reaches 1:6, and the improvement of model accuracy is insignificant as the amount of unlabeled data continues to increase. In addition, it can be also found that the computation time (Fig. 9b) and memory consumption (Fig. 9c) increase sharply with the increase of unlabeled data. Average computation times increase approximately 10 times from 0.45 s to 4.31 s, and the average memory consumption increases about 200 times from 22.0 kb to 4336.5 kb. Thus, the addition of more unlabeled data to modeling process can greatly improve the classification accuracy of LapSVM, since they provide more useful structure information of data for classification. Simultaneously, it also leads to significant increase of computational costs, although these are acceptable for this work. To further enhance the performance of LapSVM, it is needed to reduce the computational cost under the premise of high model accuracy.

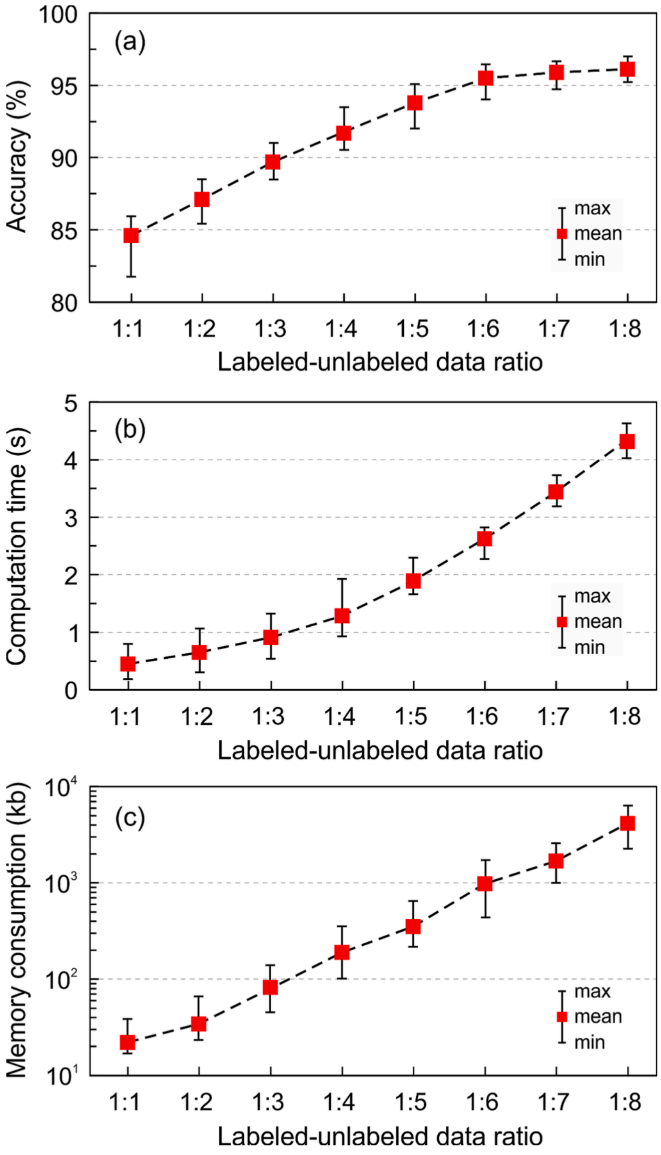


Fig. 9. Performance comparison of (a) classification accuracy, (b) computation time and (c) memory consumption for LapSVM models with different ratios of labeled and unlabeled data. Each ratio of dataset is trained and tested twenty times individually under the same running environment.

4.3. Performance comparison and validation

To demonstrate the performance improvement of the proposed LapSVM method, training and testing results of the LapSVM model were compared with these of a traditional SVM model. The RBF kernel was also selected as the kernel function for SVM. As it belongs to the supervised learning algorithm, SVM utilizes only the labeled data for model development. In the modeling process, the 70 % of 158 labeled data was utilized for training and the other 30 % was used for testing. In addition, the cross validation and grid searching methods were applied for the parameter optimization of SVM. The optimal parameter can be determined when the SVM model obtains the highest accuracy. In the end, the grid searching and cross validation found $[C, g] = [5.33, 6.25]$ as optimal parameters of the SVM model (Fig. 10).

Confusion matrices of training and testing results for SVM and LapSVM models are displayed in Fig. 11. In the confusion matrix, the diagonal positions represent the correct classification in each class and the others are misclassification. The classification result shows that training and testing accuracies of the SVM model are 90.9 % and 68.8 %, and these of the LapSVM model are 96.4 % and 91.7 % (Table 4). Form these results, it can be found that both the SVM and LapSVM perform well during model training as their classification accuracies of training data are greater than 90 %. Meanwhile, the performance of LapSVM model is significantly better than that of SVM in the testing process, with the testing accuracy increased by 22.9 %. This is mainly due to the SVM method is based on the supervised learning and is data-driven, and abundant labeled data are required for training to keep the model performance. Thus, SVM model cannot be completely trained and is prone to overfitting when the training data is limited. Different from SVM,

Table 3
Numbers of labeled and unlabeled data in each dataset.

	Labeled data	Unlabeled data	Labeled-unlabeled data ratio
Dataset 1	158	158	1:1
Dataset 2	158	316	1:2
Dataset 3	158	474	1:3
Dataset 4	158	632	1:4
Dataset 5	158	790	1:5
Dataset 6	158	948	1:6
Dataset 7	158	1106	1:7
Dataset 8	158	1264	1:8

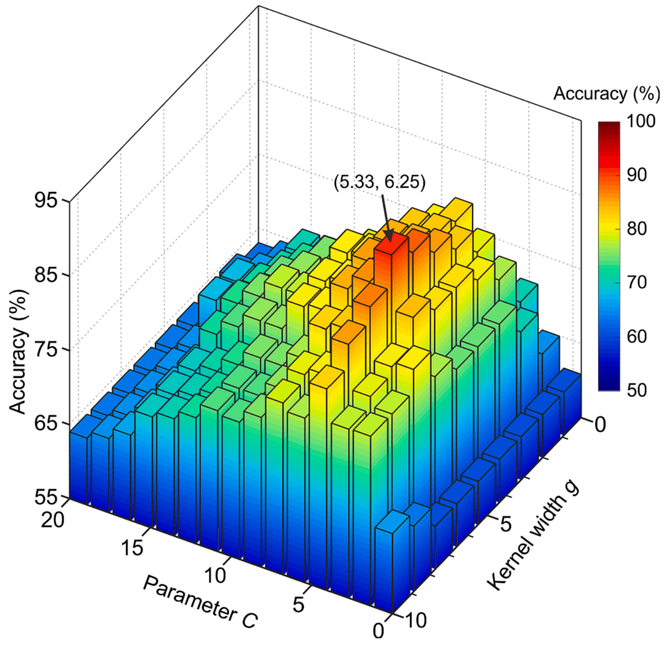


Fig. 10. Determination of optimal hyperparameters in the SVM model by grid searching and cross validation.

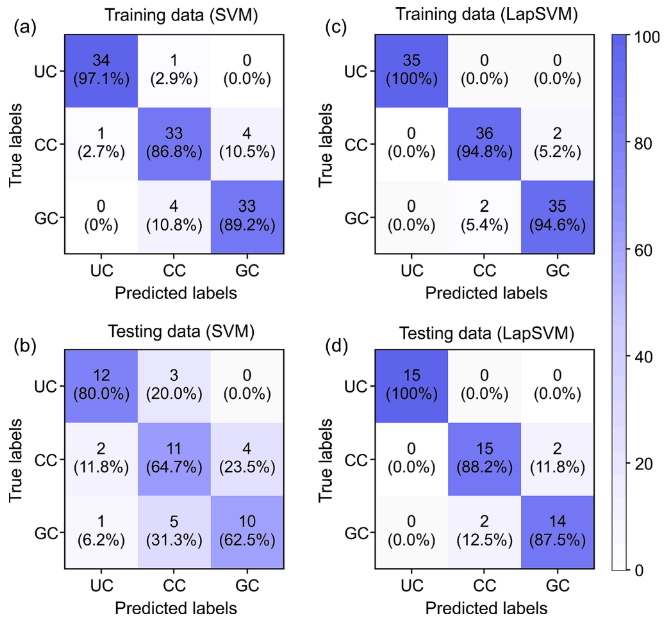


Fig. 11. Confusion matrices of training and testing results for SVM and LapSVM models. The diagonal positions represent the correct classification in each class and others are misclassification. UC = undeformed coal; CC = cataclastic coal; GC = granulated coal.

Table 4
Accuracy of SVM and LapSVM models based on training, testing and total data.

	Acc_training	Acc_testing	Acc_total
SVM	90.9 %	68.8 %	84.2 %
LapSVM	96.4 %	91.7 %	94.9 %

LapSVM constructs the decision boundary by considering the intrinsic features of both labeled and unlabeled data, thereby it relieves the dependence on labeled data. Hence, the LapSVM model can achieve high classification accuracy in both training and testing processes, and the small difference in accuracy (4.7 %) indicates that LapSVM possesses stronger stability.

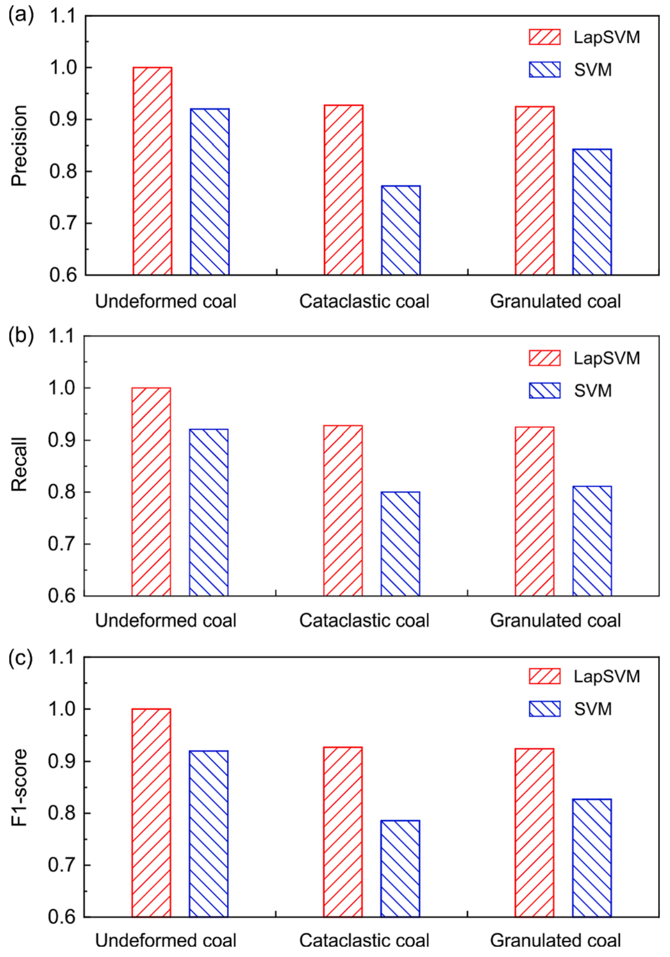


Fig. 12. Performance comparison by evaluation indicators (a) precision, (b) recall, and (c) F1-score of SVM and LapSVM models. The precision, recall and F1-score are calculated based on training and testing results.

Furthermore, other evaluation indicators precision, recall and F1-score were calculated according to confusion matrices for the performance comparison of SVM and LapSVM models. Calculation results of these evaluation indicators for each model are presented in Fig. 12. It is observed that the precision, recall and F1-score of undeformed coals for the two models are all more than 0.92, and the value differences between models are very small. It manifests that both the SVM and LapSVM models have good abilities to identify undeformed coals. Meanwhile, the LapSVM model shows better performance than standard SVM model in the classification of cataclastic and granulated coals. The values of precision, recall and F1-score for LapSVM model are over 0.90, which are 0.09–0.16 higher than these of SVM model. The accurate prediction of cataclastic coal and granulated coal is critical for CBM development, because the former is usually high-quality reservoirs and are most favorable for CBM production, while the latter is adverse intervals due to its poor hydraulic fracturing effects [11,30]. In terms of statistical metrics (i.e., accuracy, precision, recall and F1-score), the proposed LapSVM method outperforms standard SVM method over 10 % in coal structure identification and has its own advantages in the prediction of cataclastic and granulated coals.

To further validate and compare the generalization ability, the built SVM and LapSVM models were applied to a new set of data from two blind-wells in the study area, of which all 93 labeled datasets (40 datasets of well A and 53 datasets of well B) were not used in modelling

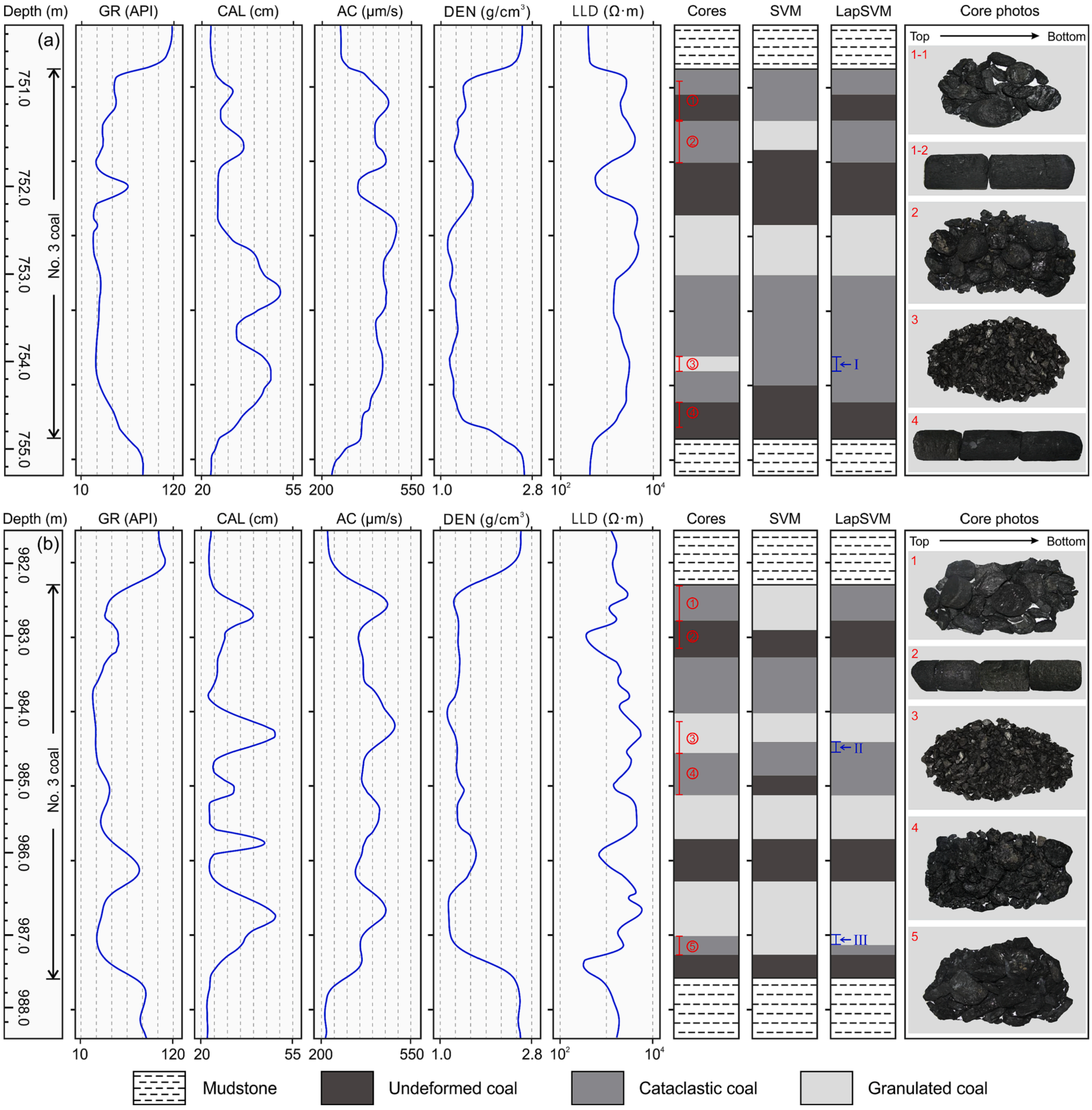


Fig. 13. Coal structure identification of blind-test well A and well B by the core observation, SVM and LapSVM methods.

process. According to core observations, coal structures of well A are cataclastic, undeformed, cataclastic, undeformed, granulated, cataclastic, granulated, cataclastic coal, and undeformed coals from top to bottom, and those of well B are cataclastic, undeformed, cataclastic, granulated, cataclastic, granulated, undeformed, granulated, cataclastic, and undeformed coals. Coal structure identification results of SVM and LapSVM models are displayed in Fig. 13. It can be found that identification results of the LapSVM model are highly consistent with

Table 5
Confusion matrices of blind-test wells by SVM and LapSVM.

		Predicted (SVM)			Predicted (LapSVM)		
		Undeformed coal	Cataclastic coal	Granulated coal	Undeformed coal	Cataclastic coal	Granulated coal
Actual (well A)	Undeformed coal	9 (81.8 %)	2 (18.2 %)	0 (0 %)	11 (100 %)	0 (0 %)	0 (0 %)
	Cataclastic coal	3 (15.8 %)	13 (68.4 %)	3 (15.8 %)	0 (0 %)	19 (100 %)	0 (0 %)
	Granulated coal	1 (10.0 %)	2 (20.0 %)	7 (70.0 %)	0 (0 %)	2 (20.0 %)	8 (80.0 %)
Actual (well B)	Undeformed coal	13 (92.9 %)	0 (0 %)	1 (7.1 %)	14 (100 %)	0 (0 %)	0 (0 %)
	Cataclastic coal	2 (10.0 %)	10 (50.0 %)	8 (40.0 %)	0 (0 %)	19 (95.0 %)	1 (5.0 %)
	Granulated coal	0 (0 %)	2 (10.5 %)	17 (89.5 %)	0 (0 %)	2 (10.5 %)	17 (89.5 %)

these of core observations except for a few depths. The identification accuracy of LapSVM for blind-test data is 94.6 % (classification accuracies of well A and well B well A are 95.0 % and 94.3 %, respectively, Table 5) and is quite close to the total accuracy of modeling process (94.9 %). By contrast, identification results of the SVM model are not ideal. The identification accuracy of SVM is only 74.2 % (classification accuracies of well A and well B well A are72.5 % and 75.4 %, respectively, Table 5), and is much lower than that of modeling data (84.2 %). These results further confirm the stability and generalization of the built LapSVM model.

In blind-well tests of well A and well B, misclassifications of coal structure by LapSVM method mainly occurred at the interface of different coal structures. For instance, granulated coals at depth of 753.96–754.13 m (interval I in well A, Fig. 13a) and 984.10–984.21 m (interval II in well B, Fig. 13b) were misclassified into cataclastic coals, and cataclastic coals at depth of 987.00–987.17 m (interval III in well B, Fig. 13b) were misclassified as granulated coals. In the research of Chen et al. [11], the authors attributed these misclassifications to the “boundary effects”, which was mainly caused by the low resolution of logging data. In the interface of coal structure transformation, logging values may be disturbed or diminished due to the low logging resolution and are not the true logging response of coal structures. In this context, the misclassification of coal structure cannot be imputed to the identification ability of LapSVM model. The successful application of LapSVM in new wells indicates that it is a reliable method to identify coal structure when limited labeled logging data is presented.

4.4. Sensitive analysis

The validity and reliability of the proposed LapSVM method have been verified by above performance evaluation and blind-well test. In the meantime, it is worth noting that the LapSVM model operates as a “black box” in the coal structure identification due to the difficulty in visualizing the classification process, thereby it remains unclear the effect of input variables on identification results. The establishment and validation of LapSVM model should not be the end of coal structure identification. The sensitive analysis of input variables can provide a way to unlock the “black box” of model and make the machine learning method more understandable in geological views, which is conducive to the application and promotion of methods.

Shapley additive explanations (SHAP) is an effective interpretative tool of machine learning models and is adopted in this work for sensitive analysis. Based on cooperative game theory, SHAP estimates the importance of input variables by considering the contribution margin when adding a variable to model [59]. The variable importance is represented by SHAP value ψ_i, which is defined as:

$$\psi_i = \sum_{S\subseteq N\{x_i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}[V(S\cup\{x_i\})-V(S)] \tag{13}$$

where N is the set of all input variables, S is the subset of N, $V(S\cup\{x_i\})$ and $V(S)$ are model outputs that trained with and without input variables S, respectively. SHAP value of the input variable i is determined by the average value of all possible permutations of variable set.

Sensitive analysis results of five logging variables for LapSVM model based on SHAP are shown in Fig. 14. The abscissa is the SHAP value that represents the contribution of a certain logging variable to model outputs. The ordinate is the types of logging variables that are arranged in order of importance degree from top to bottom, namely GR, AC, DEN, LLD and CAL. The higher the SHAP value, the greater the contribution of logging variable. From the perspective of geophysical interpretation, different contributions of well logs reflect the distinct sensitivities of logging responses. Specifically, the GR log measures the radioactivity of coal seams and is sensitive to the variation of radioactive elements density in coals. The DEN and AC can precisely detect the diversity of bulk density in coals caused by different development degree of pores and fractures. The LLD log measures the resistivity of coals that are associated with gas content. In general, gas content increases with the deformation degree since high pore/fracture space provide more surface areas for gas adsorption. However, numerous practices of CBM development indicated that gas content of coal reservoirs was also controlled by complex storage conditions, such as the thickness of coal seam and sealing capacity of roof and floor. The CAL is the record of borehole size, and it is susceptible because of the weak mechanical strength of coals. According to above sensitive analysis and geophysical interpretation, it can be concluded that the GR, DEN and AC logs have largest contributions in the identification process of LapSVM, followed by the LLD log, while the CAL log contribute the least.

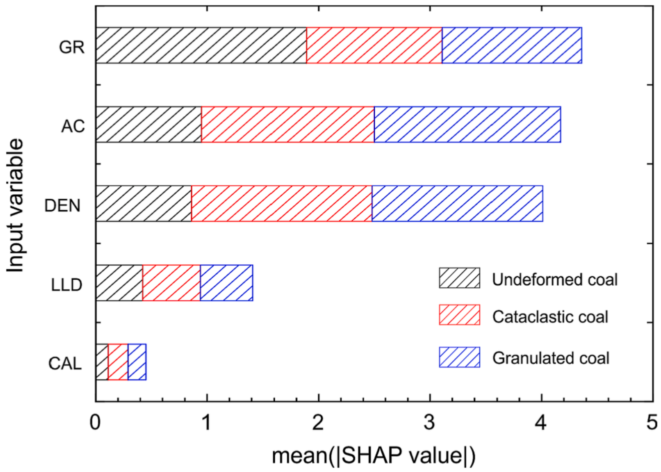


Fig. 14. SHAP values of each input variable used in LapSVM model.

4.5. Future work

The effectiveness and superiority of the proposed LapSVM method have been demonstrated through above analysis results. Nonetheless, there are still some issues should be considered in future work to further improve the LapSVM model. First, the computational cost of LapSVM is high. In this work, the computation time and memory consumption of model are acceptable as the scale of data for modeling is relatively small. However, computational costs will predictably be striking when massive data are added. Thus, one of future work is toward reducing computational cost for improving the model performance. Second, the proposed LapSVM method currently is unable to handle the data imbalance problem. To ensure the performance of classification model, data used for model training should be balanced, such as the dataset used in this study, where sample numbers of each coal structure are approximately the same. However, data imbalance is ubiquitous in practice and always aggravated by data acquisition. Normally, classification model trained with imbalanced data is biased towards majority classes, resulting in underprediction of minority ones [60]. Hence, the imbalanced data can dramatically reduce the model performance. The prototype generation (PG) algorithm and synthetic minority over-sampling technique (SMOTE) are potential solutions. Furthermore, the LapSVM method proposed in this work for coal structure identification is developed based on high-rank coal. However, coal ranks have significant diversities from region to region due to the variation of depositional environment and coalification condition, which may create difficulties to the promotion of the proposed LapSVM method in different coals. The model optimization based on actual coal properties and logging data is the critical issue when the LapSVM-based method is adopted to other regions. The PSO used in this work may be not the most appropriate method for model optimizing in other coal reservoirs, and more parameter optimization methods, such as the whale optimization algorithm (WOA) and grey wolf optimizers (GWO), will be explored in future work.

5. Conclusions

The limited labeled-sample problem poses a great challenge to the accurate identification of coal structure. In this study, a semi-supervised learning strategy-based LapSVM method was proposed to identify coal structure under limited labeled logging data. The LapSVM-based method can improve the identification performance of model by mining structure information of abundant unlabeled data, and alleviate the reliance on labeled data. Datasets collected from 32 CBM wells in the southern Qinshui Basin, China, were applied to evaluate the performance of proposed LapSVM method. Experiments with different ratios of labeled and unlabeled data illustrated that the addition of unlabeled data was conducive to improve model accuracy, but increased the computational cost simultaneously. The modeling and blind-well test results validated the outstanding validity and generalization of the proposed LapSVM method. The LapSVM model outperformed the standard SVM model with more than 10 % increase in accuracy, precision, recall and F1-score. Particularly, LapSVM had better performs in the identification of cataclastic and granulated coals. The sensitive analysis based on SHAP revealed that the GR, DEN and AC logs had greatest contributions in identification process of LapSVM. To extend the application of the proposed LapSVM identification method, the computational cost, data imbalance and parameter optimization issues will be explored in future work.

CRediT authorship contribution statement

Jinxiong Shi: Conceptualization, Methodology, Investigation, Writing – original draft, Software, Visualization. **Xiangyuan Zhao:** Methodology, Investigation, Formal analysis, Project administration, Supervision. **Lianbo Zeng:** Methodology, Writing – review & editing, Validation, Supervision. **Yunzhao Zhang:** Formal analysis, Investigation, Visualization, Supervision. **Shaoqun Dong:** Methodology, Software, Validation, Visualization.

Declaration of Competing Interest

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

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 41472123). Special thanks to the PetroChina Huabei Oilfield Company for providing core samples and for Professor Renfang Pan (Yangtze University) for his support for this paper. The authors are also grateful to editors and anonymous reviewers for their constructive suggestions.

References

[1] Yu S, Bo J, Ming Li, Chenliang H, Shaochun Xu. A review on pore-fractures in tectonically deformed coals. Fuel 2020;278:118248.

[2] Hou QL, Li HJ, Fan JJ, Ju YW, Wang TK, Li XS, et al. Structure and coalbed methane occurrence in tectonically deformed coals. Sci China Earth Sci 2012;55 (11):1755–63.

[3] Cao L, Yao Y, Liu D, Yang Y, Wang Y, Cai Y. Application of seismic curvature attributes in the delineation of coal texture and deformation in Zhengzhuang field, southern Qinshui Basin. AAPG Bull 2020;104(5):1143–66.

[4] Farmer IW, Pooley FD. A hypothesis to explain the occurrence of outbursts in coal, based on a study of west Wales outburst coal. Int J Rock Mech Min Sci 1967;4(2): 189–93.

[5] Shepherd J, Rixon LK, Griffiths L. Outbursts and geological structures in coal mines: a review. Int J Rock Mech Min Sci 1981;18(4):267–83.

[6] Ju YW, Jiang B, Hou QL, Tan YJ, Wang GL, Xiao WJ. Behavior and mechanism of the adsorption/desorption of tectonically deformed coals. China Sci Bull 2010;54: 88–94.

[7] Li W, Jiang B, Moore TA, Wang G, Liu JG, Song Y. Characterization of the chemical structure of tectonically deformed coals. Energy Fuel 2017;31(7):6977–85.

[8] Xu H, Tang DZ, Zhao JL, Li S, Tao S. A new laboratory method for accurate measurement of the methane diffusion coefficient and its influencing factors in the coal matrix. Fuel 2015;158(2):239–47.

[9] Frodsham K, Gayer RA. The impact of tectonic deformation upon coal seams in the South Wales coalfield, UK. Int J Coal Geol 1999;38(3-4):297–332.

[10] Fu XH, Qin Y, Wang GX, Rudolph V. Evaluation of coal structure and permeability with the aid of geophysical logging technology. Fuel 2009;88(11):2278–85.

[11] Chen S, Liu P, Tang D, Tao S, Zhang T. Identification of thin-layer coal texture using geophysical logging data: investigation by wavelet transform and linear discrimination analysis. Int J Coal Geol 2021;239:103727.

[12] Teng J, Yao YB, Liu DM, Cai YD. Evaluation of coal texture distributions in the southern Qinshui basin, North China: Investigation by a multiple geophysical logging method. Int J Coal Geol 2015;140:9–22.

[13] Ju Y, Li X. New research progress on the ultrastructure of tectonically deformed coals. Progr Nat Sci 2009;19(11):1455–66.

[14] Yao HF, Kang ZQ, Li W. Deformation and reservoir properties of tectonically deformed coals. Petrol Explor Dev 2014;41(4):460–7.

[15] Hou HH, Shao LY, Guo SQ, Li Z, Zhang ZJ, Yao ML, et al. Evaluation and genetic analysis of coal structures in deep Jiaozuo Coalfield, northern China: Investigation by geophysical logging data. Fuel 2017;209:552–66.

[16] Guo X, Huan X, Huan HH. Structural characteristics of deformed coals with different deformation degrees and their effects on gas adsorption. Energy Fuel 2017;31(12):13374–81.

[17] Liu DM, Jia QF, Cai YD, Gao CJ, Qiu F, Zhao Z, et al. A new insight into coalbed methane occurrence and accumulation in the Qinshui Basin. China Gondwana Res 2022;111:280–97.

[18] Song Yu, Jiang Bo, Li M, Hou C, Mathews JP. Macromolecular transformations for tectonically-deformed high volatile bituminous via HRTEM and XRD analyses. Fuel 2020;263:116756.

[19] Li JQ, Liu DM, Yao YB, Cai YD, Qiu YK. Evaluation of the reservoir permeability of anthracite coals by geophysical logging data. Int J Coal Geol 2011;87(2):121–7.

[20] Pan JN, Hou QL, Ju YW, Bai HL, Zhao YQ. Coalbed methane sorption related to coal deformation structures at different temperatures and pressures. Fuel 2012; 102:760–5.

[21] Gharagheizi F, Ilani-Kashkouli P, Mohammadi AH. Estimation of lower flammability limit temperature of chemical compounds using a corresponding state method. Fuel 2013;103:899–904.

[22] Wang Y, Liu D, Cai Y, Yao Y, Pan Z. Constraining coalbed methane reservoir petrophysical and mechanical properties through a new coal structure index in the southern Qinshui Basin, northern China: implications for hydraulic fracturing. AAPG Bull 2020;104(8):1817–42.

[23] Meng ZP, Zhang JC, Wang R. In-situ stress, pore pressure and stress-dependent permeability in the Southern Qinshui Basin. Int J Rock Mech Min Sci 2011;48(1): 122–31.

[24] Wang YJ, Liu DM, Cai YD, Yao YB, Zhou YF. Evaluation of structured coal evolution and distribution by geophysical logging methods in the Gujiao Block, northwest Qinshui basin, China. J Nat Gas Sci Eng 2018;51:210–22.

[25] Ghosh S, Chatterjee R, Paul S, Shanker P. Designing of plug-in for estimation of coal proximate parameters using statistical analysis and coal seam correlation. Fuel 2014;134(9):63–73.

[26] Fu XH, Qin Y, Wang GX, Rudolph V. Evaluation of gas content of coalbed methane reservoirs with the aid of geophysical logging technology. Fuel 2009;88(11): 2269–77.

[27] Özgen Karacan C. Elastic and shear moduli of coal measure rocks derived from basic well logs using fractal statistics and radial basis functions. Int J Rock Mech Min Sci 2009;46(8):1281–95.

[28] Oyler DC, Mark C, Molinda GM. In situ estimation of rock strength using logging. Int J Coal Geol 2010;83:484–90.

[29] Ghosh S, Chatterjee R, Shanker P. Estimation of ash, moisture content and detection of coal lithofacies from well logs using regression and artificial neural network modelling. Fuel 2016;177:279–87.

[30] Roslin A, Esterle JS. Electrofacies analysis using high-resolution wireline geophysical data as a proxy for inertinite-rich coal distribution in Late Permian Coal Seams, Bowen Basin. Int J Coal Geol 2015;152(3):10–8.

[31] Xu H, Tang DZ, Mathews JP, Zhao JL, Li BY, Tao S, et al. Evaluation of coal macrolithotypes distribution by geophysical logging data in the Hancheng Block, eastern margin, Ordos Basin. China Int J Coal Geol 2016;165:265–77.

[32] Ren PF, Xu H, Tang DZ, Li YK, Sun CH, Tao S, et al. The identification of coal texture in different rank coal reservoirs by using geophysical logging data in northwest Guizhou, China: Investigation by principal component analysis. Fuel 2018;230:258–65.

[33] Shi J, Zeng L, Dong S, Wang J, Zhang Y. Identification of coal structures using geophysical logging data in Qinshui Basin, China: Investigation by kernel Fisher discriminant analysis. Int J Coal Geol 2020;217:103314.

[34] Dong S, Zeng L, Lyu W, Xu C, Liu J, Mao Z, et al. Fracture identification by semi-supervised learning using conventional logs in tight sandstones of Ordos Basin. China J Nat Gas Sci Eng 2020;76:103131.

[35] Lan X, Zou C, Kang Z, Wu X. Log facies identification in carbonate reservoirs using multiclass semi-supervised learning strategy. Fuel 2021;302:121145.

[36] Liu S, Yang Y, Yu L, Zhu F, Cao Yu, Liu X, et al. Predicting gas production by supercritical water gasification of coal using machine learning. Fuel 2022;329:125478.

[37] Dong S, Zeng L, Du X, He J, Sun F. 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 Petrol Sci Eng 2022;210:110081.

[38] Li Z, Kang Yu, Feng D, Wang X-M, Lv W, Chang Ji, et al. Semi-supervised learning for lithology identification using laplacian support vector machine. J Pet Sci Eng 2020;195:107510.

[39] Melacci S, Belkin M. Laplacian support vector machines trained in the primal. J Mach Learn Res 2011;12(Mar):1149–84.

[40] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res 2006;7(Nov):2399–434.

[41] Gomez-Chova L, Camps-Valls G, Munoz-Mari J, Calpe J. Semisupervised image classification with Laplacian support vector machines. IEEE Geosci Remote Sens Lett 2008;5(3):336–40.

[42] Halder A, Ghosh S, Ghosh A. Aggregation pheromone metaphor for semi-supervised classification. Pattern Recogn 2013;46(8):2239–48.

[43] Chen W-J, Shao Y-H, Deng N-Y, Feng Z-L. Laplacian least squares twin support vector machine for semi-supervised classification. Neurocomputing 2014;145:465–76.

[44] Ding SF, Zhu ZB, Zhang XK. An overview on semi-supervised support vector machine. Neural Comput Appl 2017;28(5):969–78.

[45] Gong TL, Chen H, Xu C. Learning performance of LapSVM based on Markov Subsampling. Neurocomputing 2021;432:10–20.

[46] Qi ZQ, Tian YJ, Shi Y. Successive overrelaxation for laplacian support vector machine. IEEE Trans Neural Netw Learn Syst 2015;26(4):674–83.

[47] Tan M, Bai Y, Zhang H, Li G, Wei X, Wang A. Fluid typing in tight sandstone from wireline logs using classification committee machine. Fuel 2020;271:117601.

[48] Zhao XZ, Liu SQ, Sang SX, Pan ZJ, Zhao WX, Yang YH, et al. Characteristics and generation mechanisms of coal fines in coalbed methane wells in the southern Qinshui Basin. China J Nat Gas Sci Eng 2016;34:849–63.

[49] Chen S, Tao S, Tian W, Tang D, Zhang B, Liu P. Hydrogeological control on the accumulation and production of coalbed methane in the Anze Block, southern Qinshui Basin, China. J Petrol Sci Eng 2021;198:108138.

[50] Shi J, Zeng L, Zhao X, Zhang Y, Wang J. Characteristics of natural fractures in the upper Paleozoic coal bearing strata in the southern Qinshui Basin, China: Implications for coalbed methane (CBM) development. Mar Petrol Geol 2020;113:104152.

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

[52] Zhong Z, Carr TR. Application of mixed kernels function (MKF) based support vector regression model (SVR) for CO2-reservoir oil minimum miscibility pressure prediction. Fuel 2016;184:590–603.

[53] Lei CK, Deng J, Cao K, Xiao Y, Ma L, Wang WF, et al. A comparison of random forest and support vector machine approaches to predict coal spontaneous combustion in gob. Fuel 2019;239:297–311.

[54] Schölkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond. London: Massachusetts Institute of Technology; 2002. p. 626.

[55] Hotta K. View independent face detection based on horizontal rectangular features and accuracy improvement using combination kernel of various sizes. Pattern Recogn 2009;42(3):437–44.

[56] Zhong Z, Liu SY, Kazemi M, Carra TR. Dew point pressure prediction based on mixed-kernels-function support vector machine in gas-condensate reservoir. Fuel 2018;232:600–9.

[57] Kennedy J, Eberhart R. Particle Swarm Optimization. In: IEEE international joint conference on neural networks, Perth, Australia, 1995, p. 1942–1948.

[58] Fawcett T. An introduction to roc analysis. Pattern Recogn Lett 2005;27(8):861–74.

[59] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 2017, p. 4765–4774.

[60] Yuan XH, Xie LJ, Abouelenien M. A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data. Pattern Recogn 2018;77:160–72.