Petroleum Science 20 (2023) 733–752

Contents lists available at ScienceDirect

Petroleum Science

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

Original Paper

How to improve machine learning models for lithofacies identification by practical and novel ensemble strategy and principles

Shao-Qun Dong [a, b], Yan-Ming Sun [a, b], Tao Xu [a, b], Lian-Bo Zeng [a, c, *], Xiang-Yi Du [a, c], Xu Yang [a, b], Yu Liang [a, b]

[a] State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum, Beijing, 102249, China
[b] College of Science, China University of Petroleum, Beijing, 102249, China
[c] College of Geoscience, China University of Petroleum, Beijing, 102249, China

ARTICLE INFO

Article history:
Received 14 February 2022
Received in revised form
6 May 2022
Accepted 14 September 2022
Available online 28 September 2022

Edited by Jie Hao

Keywords:
Lithofacies identification
Machine learning
Ensemble learning strategy
Ensemble principle
Homogeneous ensemble
Heterogeneous ensemble

ABSTRACT

Typically, relationship between well logs and lithofacies is complex, which leads to low accuracy of lithofacies identification. Machine learning (ML) methods are often applied to identify lithofacies using logs labelled by rock cores. However, these methods have accuracy limits to some extent. To further improve their accuracies, practical and novel ensemble learning strategy and principles are proposed in this work, which allows geologists not familiar with ML to establish a good ML lithofacies identification model and help geologists familiar with ML further improve accuracy of lithofacies identification. The ensemble learning strategy combines ML methods as sub-classifiers to generate a comprehensive lithofacies identification model, which aims to reduce the variance errors in prediction. Each sub-classifier is trained by randomly sampled labelled data with random features. The novelty of this work lies in the ensemble principles making sub-classifiers just overfitting by algorithm parameter setting and sub-dataset sampling. The principles can help reduce the bias errors in the prediction. Two issues are discussed, videlicet (1) whether only a relatively simple single-classifier method can be as sub-classifiers and how to select proper ML methods as sub-classifiers; (2) whether different kinds of ML methods can be combined as sub-classifiers. If yes, how to determine a proper combination. In order to test the effectiveness of the ensemble strategy and principles for lithofacies identification, different kinds of machine learning algorithms are selected as sub-classifiers, including regular classifiers (LDA, NB, KNN, ID3 tree and CART), kernel method (SVM), and ensemble learning algorithms (RF, AdaBoost, XGBoost and LightGBM). In this work, the experiments used a published dataset of lithofacies from Daniudi gas field (DGF) in Ordes Basin, China. Based on a series of comparisons between ML algorithms and their corresponding ensemble models using the ensemble strategy and principles, conclusions are drawn: (1) not only decision tree but also other single-classifiers and ensemble-learning-classifiers can be used as sub-classifiers of homogeneous ensemble learning and the ensemble can improve the accuracy of the original classifiers; (2) the ensemble principles for the introduced homogeneous and heterogeneous ensemble strategy are effective in promoting ML in lithofacies identification; (3) in practice, heterogeneous ensemble is more suitable for building a more powerful lithofacies identification model, though it is complex.

Abbreviations

The abbreviations used in the paper are summarized in Table 1 below.

1. Introduction

Lithofacies identification based on logging data is an important task in petroleum exploration and development, since it is critical for reservoir characterization, calculation of reserves and 3D geological modelling (Dong et al., 2022; Ma, 2011; Martyushev and Yurikov, 2021). However, the relationships between well logs and lithofacies are usually complicated since well log responses are influenced by other factors such as porosity, oil-bearing properties,

* Corresponding author. State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum, Beijing, 102249, China.
E-mail address: lbzeng@sina.com (L.-B. Zeng).

https://doi.org/10.1016/j.petsci.2022.09.006

Table 1
Term abbreviations used in this work.

Full name	Abbr.	Full name	Abbr.
machine learning	ML	gradient boosting machines	GBMs
linear discriminant analysis	LDA	random forest	RF
Naive Bayes	NB	adaptive boosting	AdaBoost
k-nearest neighbors	KNN	gradient boosting decision tree	GBDT
iterative dichotomiser 3	ID3	eXtreme gradient boosting	XGBoost
classification and regression trees	CART	light gradient boosting machine	LightGBM
decision tree	DT	double fault	DF
support vector machine	SVM	Daniudi Gas Field	DGF
artificial neural networks	ANN	Hangjinqi Gas Field	HGF

fractures, etc. (Liu et al., 2018; Saggaf and Nebrija, 2000; Yang et al., 2019). To build lithofacies identification models, many mathematical methods have been introduced to train the prediction models utilizing logging data labelled by rock cores (Delfiner et al., 1987; Tokhmechi et al., 2009). Among them, machine learning techniques perform well in many oilfields, which can be divided into common single-classifier methods and methods based on ensemble classifiers (Tewari and Dwivedi, 2019).

Single-classifier methods are widely used in lithofacies identification, such as linear discriminant analysis (LDA) (Busch et al., 1987; Dubois et al., 2007; Li and Anderson-Sprecher, 2006), naive Bayes (NB) (Corina and Hovda, 2018; Li and Anderson-Sprecher, 2006; Moja et al., 2019), k-nearest neighbors (KNN) (Tripoppoom et al., 2019; Wang et al., 2018), decision tree (DT) (Breiman et al., 2015; Kolose et al., 2021; Li et al., 2011; Quinlan, 1986, 1996), support vector machines (SVM) (Al-Anazi and Gates, 2010; Hou et al., 2020; Liu et al., 2020; Sebtosheikh et al., 2015), artificial neural networks (ANN) (Bressan et al., 2020; Gorai et al., 2021; He et al., 2019; Lawal et al., 2021; Wang et al., 2017), and etc. The idea of LDA is to extract linear features distinguishing lithofacies according to the principle of maximizing the distance between classes and at the same time minimizing the distance within each class (Dong et al., 2016). In the Shublik Formation of the Prudhoe Bay, America, LDA obtained an accuracy of 75% in the lithofacies identification which provided support for reservoir description (Busch et al., 1987). When lithofacies become quite complex, LDA may perform poorly since linear methods are not enough to deal with nonlinear classification problems. For example, the accuracy of LDA is only about 60% in Panoma Gas Field in southwestern Kansas, America (Dubois et al., 2007). NB minimizes errors of identification based on posterior probability, which is calculated by prior probabilities of each lithofacies and prior distributions of logs against lithofacies (Corina and Hovda, 2018; Moja et al., 2019). In the Upper Tensleep Formation in Teapot Dome, Powder River Basin, Wyoming, a naive Bayes classifier was trained by petrophysical logs of seven wells. It obtained an accuracy of about 75% with a similar performance to LDA in both efficiency and consistency (Li and Anderson-Sprecher, 2006). However, naive Bayes assumes each well log is independent of any other logs, which is usually difficult to realize in practice (Ao et al., 2018). KNN is a classification method by which the data are classified based on a plurality vote of its neighbors. A new sample will be assigned to the most common class among its neighbors (Tripoppoom et al., 2019). Optimized KNN and traditional KNN methods are used to identify the lithology by well-logging data acquired by Gaoqing Oilfield in China, and the total classification accuracy is about 61% (Wang et al., 2018). DT is essentially a series of statements that are aligned through the structure of nodes and sheets and have specific if-else rules (Kolose et al., 2021). ID3 (iterative dichotomiser 3) decision tree (Quinlan, 1986), C4.5 (an extension of ID3) (Quinlan, 1996) and CART (classification and regression trees) (Breiman et al., 2015) are three representative DT methods wildly used in lithofacies identification. SVM maps original logs into a higher feature space by kernel functions and builds a hyperplane classifier to identify lithofacies (Cortes and Vapnik, 1995). For the heterogeneous carbonate reservoirs in Iran, the RBF kernel based SVM using well logs after feature selection won a lower misclassification rate (about 8%) than SVM based on polynomial and polynomial kernels (Sebtosheikh et al., 2015). For the sandstone reservoirs in the Middle East, SVM combined with feature selection based on fuzzy theory obtained a misclassification of about 10%, and outperformed LDA and probabilistic neural network (PNN) (Al-Anazi and Gates, 2010), since SVM is good at addressing high-dimensional nonlinear features in the logging data (Anifowose et al., 2015). The ANN is a bionics method imitating the biological neural network to process information and predict lithofacies (Kardani et al., 2022). For the Longmaxi-Wufeng shale reservoirs in the Fuling Gas Field of Sichuan Basin, China, the combination of ANN and hierarchical decomposition (HD) by conventional logs were used to identify lithofacies for 3-D geological modelling. The lithofacies labels of training data were determined by cores and elementary capture spectroscopy (ECS). The cross-validation accuracy of all shale lithofacies is about 85%. Nevertheless, the identification of mixed shale is relatively low (about 65%) (Wang et al., 2017).

Different from single-classifier methods, an ensemble learning method combines a set of weak classifiers (called individual classifiers or sub-classifiers (Li et al., 2013; Sun et al., 2015)) into a strong classifier to codetermine predictions of lithofacies by a joint decision-making mechanism (Tewari and Dwivedi, 2019). Two of the most well-known methods are Boosting (Schapire, 1990) and Bagging (bootstrap aggregating) (Breiman, 1996). Commonly used Boosting methods include adaptive boosting (AdaBoost) and gradient boosting machines (GBMs). GBMs include gradient boosting decision tree (GBDT), eXtreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM) and so on. The representative Bagging method is random forest (RF) (Qiao and Chang, 2021; Wang et al., 2020). For lithology identification in Daniudi Gas Field and Hangjinqi Gas Field in Ordos Basin, China, the ensemble methods (GBDT and RF) obtained lower prediction errors, compared with the single-classifier methods (e.g., SVM and ANN) (Xie et al., 2018).

The Boosting method is consisted of a series of base classifiers, which are trained and promoted sequentially. Each model tries to compensate for the weaknesses of its predecessor. The step-by-step optimization aims to upgrade weak rules to strong prediction rules (Opitz et al., 2018). AdaBoost is a Boosting algorithm that changes the deficiencies of the model by increasing the weights of misclassified data points (Freund and Schapire, 1997; Friedman et al., 2000), while GBMs determine the weights by operating on the negative partial derivatives of the loss function at each training observation (Natekin and Knoll, 2013). In a Kansas oil field, experiments on lithology identification indicated that AdaBoost can

improve the accuracy of single classifiers (SVM, C4.5 DT, CART) (Tewari and Dwivedi, 2019). For lithology identification in Daniudi Gas Field (DGF) and Hangjinqi Gas Field (HGF), Ordos Basin, China, the lithology classification by GBMs (e.g., XGBoost and LightGBM) was better than single classifiers (e.g., SVM and ANN) in terms of average precision, recall and F1-score (Dev and Eden, 2018, 2019; Xie et al., 2018).

Bagging names from the abbreviation of bootstrap aggregating (Breiman, 1996). Aggregating means a Bagging predictor aggregates sub-classifiers to generate a comprehensive lithofacies identification model through the plurality voting rule. Bootstrap means that each sub-classifier is trained by randomly sampled labelled data with random features, which is helpful for the decorrelation of the sub-classifiers. It has been proven that Bagging is particularly effective in improving the accuracy of unstable individual learners since the ensemble of unstable individual learners with large diversity (Kuncheva and Whitaker 2003) can smooth the sharp decision boundary of a classifier to reduce variance and improve accuracy (Breiman, 1996; Bühlmann and Yu, 2002; Friedman and Hall, 2006). In random forest, 'forest' means the ensemble of decision trees (Breiman, 2001). Distinguished from bagged decision tree (BDT) method in the split criterion, RF only randomly chooses k features rather than all used by BDT (Breiman, 2001; Zhang et al., 2016). RF was used to identify lithology in DGF and HGF, the precision scores for 5-fold-cross-validation on the DGF dataset and HGF dataset are above 80%, which indicates RF has a good ability in lithology recognition (Xie et al., 2018). In the case of rapid lithological classification in International Ocean Discovery Program (IODP), for cross validation, RF performed better than SVM, DT and ANN in all scenarios (Bressan et al., 2020). The sub-classifiers in Bagging can not only choose decision trees, but also choose other learners to improve their prediction capacities. On a Kansas oil field data in the United States, these single classifiers (SVM, C4.5 decision tree, CART, etc.) are ensemble using the Bagging method, and the accuracy of lithology identification can be improved (Tewari and Dwivedi, 2019).

The reviews above indicate that (1) machine learning methods are useful for lithofacies identification, but they have accuracy limits to some extent; (2) ensemble learning performs better than some single-classifier methods in many cases, and it can be used to improve the prediction capacity of single-classifier methods; (3) ensemble learning usually employs one kind of relatively simple classifiers as sub-classifiers. Based on these reviews, to improve lithofacies identification, two questions about ensemble learning are raised: (1) whether only relatively simple single-classifier method can be used as sub-classifiers, and if yes, which kind of method can be employed as sub-classifiers and how to choose proper sub-classifiers; (2) whether different kinds of machine learning methods can be combined as sub-classifiers. If yes, how to determine a proper combination.

This work introduces a simple ensemble learning strategy, which is in fact the generalization of RF. It (1) generates sub-data for sub-classifiers by random sampling with or without replacement according to a specific sub-classifier method; (2) can use either single-classifier methods or ensemble learning methods as sub-classifiers (Fig. 1). Here, the sub-classifiers in homogeneous ensemble are called base classifiers, while those in heterogeneous ensemble are called component classifiers. Base classifiers are only one kind of machine learning method while component classifiers are a combination of multiple types of machine learning methods. Based on the analyses of the two raised questions, several novel ensemble principles are proposed. In Section 2, the ensemble learning strategy, the used base classifiers and ensemble principles are presented. In Section 3, a series of experiments are implemented to analyse the two problems abovementioned and test the ensemble learning strategy and ensemble principles. The dataset is from Daniudi Gas Field in Ordos Basin, China. Based on the experiment results, the Discussions section tries to explain the two questions abovementioned and summarize principles of choosing proper sub-classifiers in the used ensemble learning strategy for improving lithofacies identification.

2. Principle of mathematical methods

2.1. The ensemble learning strategy using random selections of samples and features

The ensemble learning strategy is shown in Fig. 2, which is divided into three parts. The first part is to generate sub-datasets. The main content is to randomly sample the training dataset and perform random feature selection to generate multiple training sub-datasets. The second part is to train sub-classifiers. The multiple training sub-datasets generated in the first part are used to train the sub-classifiers. The third part is the ensemble. The multiple sub-classifiers are integrated to generate an ensemble classifier according to a voting mechanism. The testing dataset is used to evaluate the classification prediction ability of the ensemble classifier, and the parameters in sub-classifiers will be adjusted until the generated ensemble classifier meets the accuracy requirement. The final selected ensemble classifier is used for lithofacies identification.

Suppose there is a dataset $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$. $\boldsymbol{x}_i$ has m features. It is divided into a training dataset with $[N \times r_{\text{train}}]$ samples and a testing dataset with $[N \times (1 - r_{\text{train}})]$ samples (e.g., $r_{\text{train}} = 80\%$). Random sampling with or without replacement will generate n_S sub-datasets with $n_b = [[N \times r_{\text{train}}] \times r_b]$ (e.g., $r_b = 50\%$) samples for training sub-classifiers. Each sub-dataset randomly selects $n_{\text{prob}} = [m \times r_p]$ (e.g., $r_p = 80\%$) features from all m features without replacement.

2.2. Single-classifier methods will be used as sub-classifiers

As shown in Fig. 1, different single-classifier methods, including regular methods, kernel methods and artificial neural network methods, will be used in this work. The six specific methods are LDA, NB, KNN, ID3, CART and SVM displayed in Fig. 3.

(1) **LDA.** LDA is a classic supervised dimensionality reduction technique, which aims to find new projection axes maximizing separability among the known categories in the target variables (Fig. 3(a)). The mapped data meets the requirements of both maximizing distance between categories and minimizing distance within categories. The problem can be converted to a problem of solving generalized eigenvalues (Dong et al., 2016; Shi et al., 2020). The values of eigenvalues represent contributions of corresponding eigenvectors $\boldsymbol{\omega}$ to distinguishing different classes. When the projection $\boldsymbol{\omega x}$ of a new sample $\boldsymbol{x}$ is closest to the center of projected points in a class, it will be identified in this category.

(2) **NB.** The NB method utilizes a posterior probability $P(y|\boldsymbol{x})$ in Eq. (1) to determine the label y of a sample $\boldsymbol{x}$ based on the maximum posterior criterion. $P(y)$ is the prior probability of one category (e.g., $y = 1,2, \ldots$). $P(\boldsymbol{x}|y)$ is the probability distribution of $\boldsymbol{x}$ corresponding to samples in a category. Usually, $\boldsymbol{x}$ is multidimensional which makes $P(\boldsymbol{x}|y)$ difficult to determine. Hence, Naïve Bayes simplifies this problem by assuming each variable of $\boldsymbol{x}$ is independent. Subsequently, $P(\boldsymbol{x}|y)$ can be expressed by $\prod_{i=1}^{n} P(x_i|y)P(y)$, where n is the number of all categories. Because $P(\boldsymbol{x})$ is the same for each

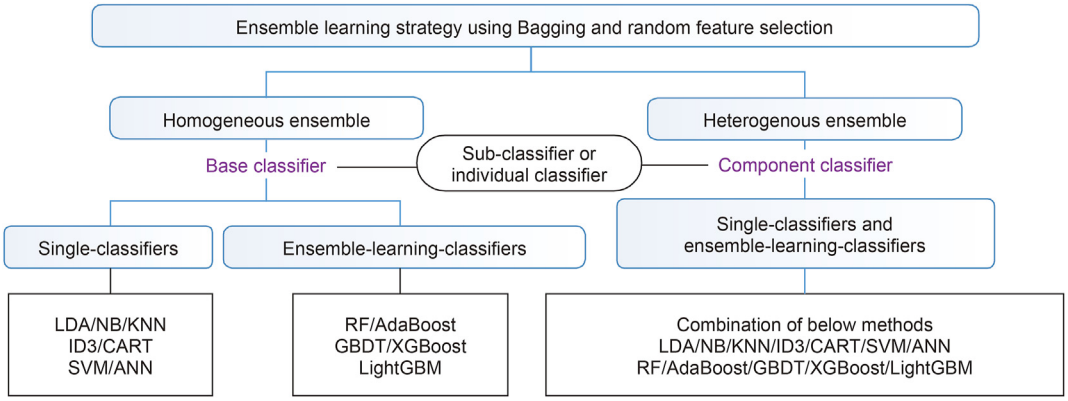

Fig. 1. Schematic diagram of experiments in this work. The sub-classifiers in homogeneous ensemble are called base classifiers, and the ones in heterogeneous ensemble are called component classifiers.

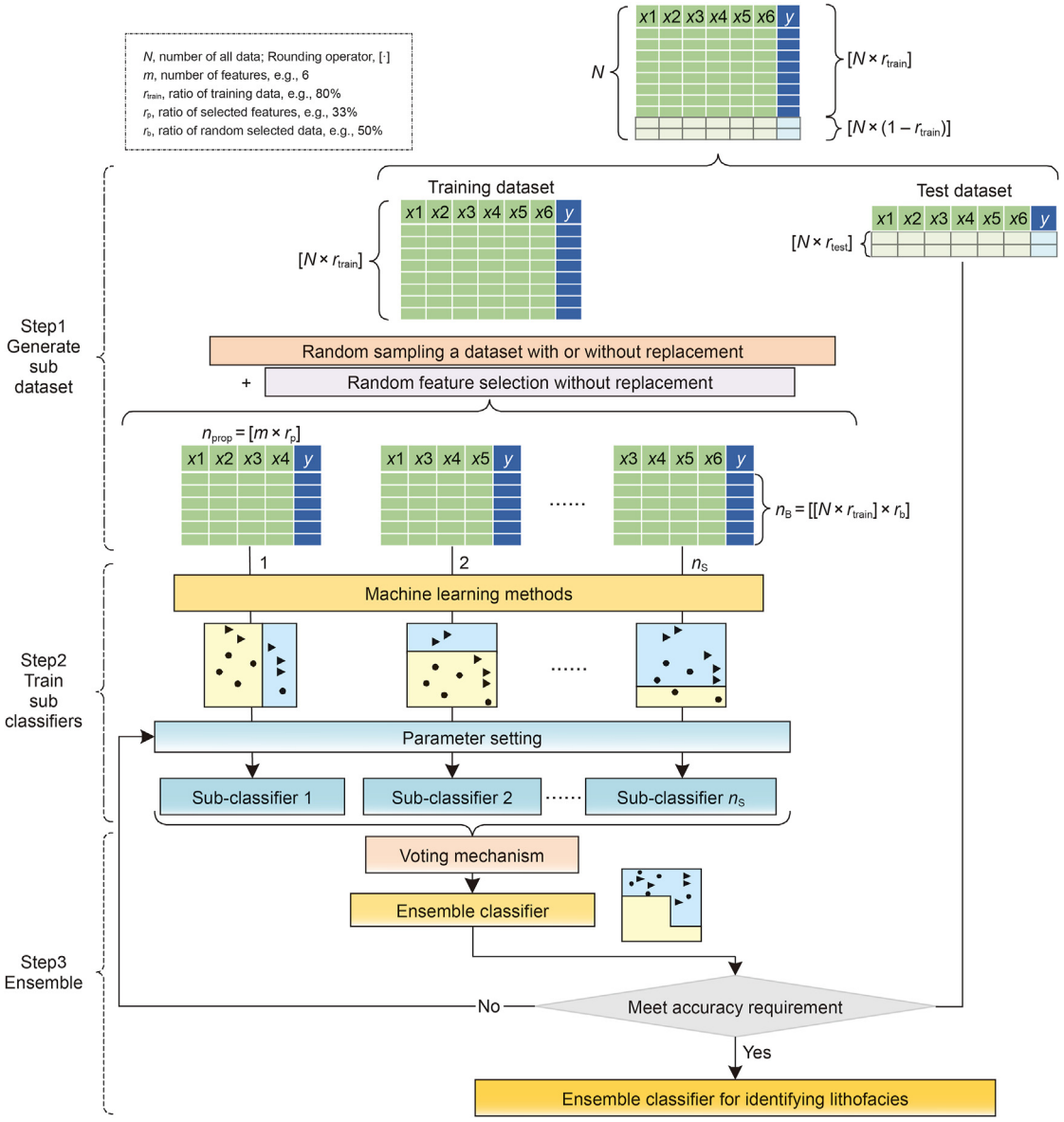

Fig. 2. Schematic diagram of the used ensemble learning strategy.

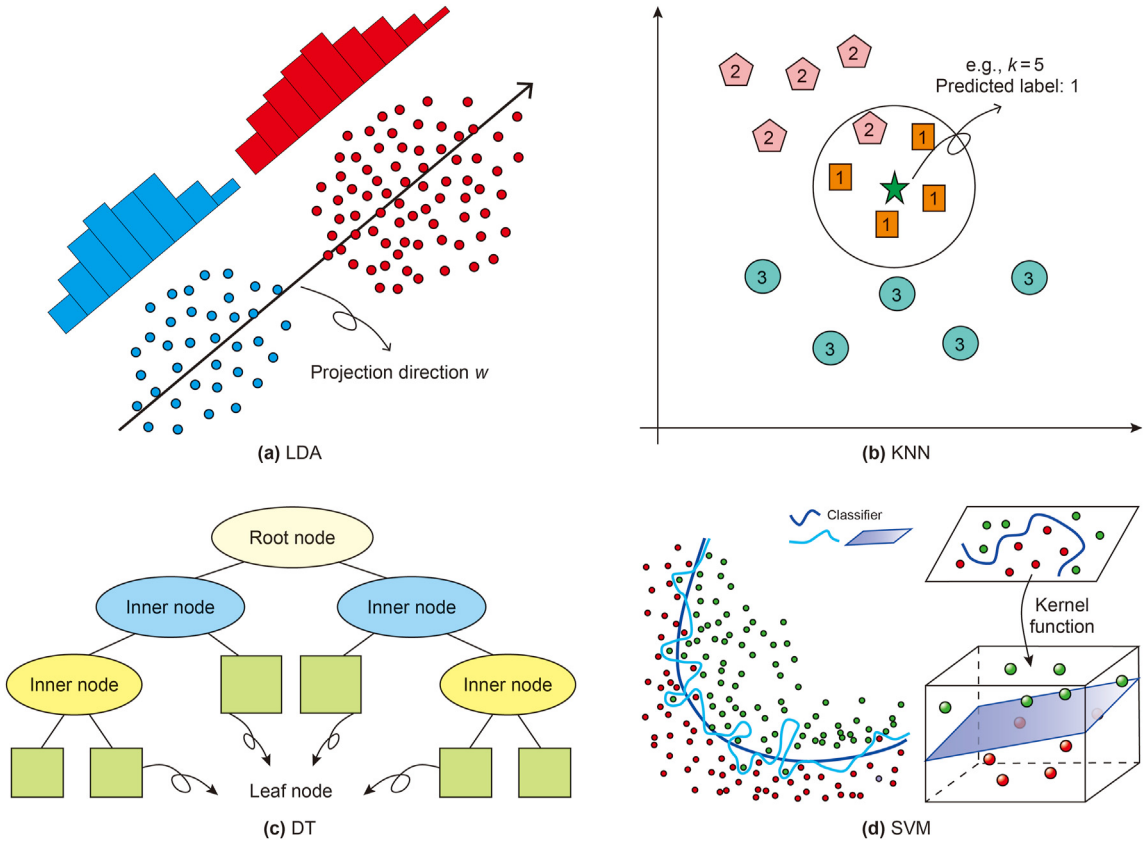

Fig. 3. Schematic diagrams of single-classifier methods.

sample, $\frac{P(\boldsymbol{x}|y)P(y)}{P(\boldsymbol{x})}$ can be substituted by $\prod_{i=1}^{n} P(x_i|y)P(y)$. After $P(x_i|y)$ and $P(y)$ are determined by the training data, the label of a new sample will be the category with the largest posterior probability.

$$P(y|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|y)P(y)}{P(\boldsymbol{x})} \propto \frac{P(\boldsymbol{x}|y)P(y)}{\prod_{i=1}^{n} P(x_i|y)P(y)} \tag{1}$$

(3) **KNN.** As shown in Fig. 3(b), a sample marked by a pentagram needs to be classified. In the K=5 neighbors, there are four "1" and one "2". According to plurality voting, this sample will be identified as "1". It should be noted that different values of K may lead to different results.

(4) **ID3.** As a decision tree algorithm, ID3 iteratively dichotomizes nodes into two or more nodes at each step (Fig. 3(c)). The division utilizes specific if-else rules based on information gain criterion. Assuming that there are K classes in a dataset D, each sample $\boldsymbol{a}$ has n features $\{a^1, a^2, \ldots, a^n\}$, the probability that the sample point belongs to the k-th class is p_k. Then the information entropy measuring the purity of the division of a dataset is defined as Eq. (2), and the information gain dividing a node will be calculated by Eq. (3). The leaf nodes in a decision tree correspond to the decision results. A depth threshold of a decision tree can be used to terminate the spliting process. Besides, the spliting process can be ended if there is only one sample in each leaf node.

$$Ent(D) = -\sum_{k=1}^{K} p_k \log_2 p_k \tag{2}$$

$$Gain(D, a) = Ent(D) - \sum_{i=1}^{n} \frac{|D^i|}{|D|} Ent\left(D^i\right) \tag{3}$$

where $\frac{|D^i|}{|D|}$ is the weight assigned to the branch node, $|D|$ represents the number of samples of data D, and $|D^i|$ indicates the number of samples whose value is a^i on attribute a.

(5) **CART.** Different from ID3, a Gini criterion is used in CART instead of information entropy. The CART tree shares the same structure with ID3 as shown in Fig. 3(c). The Gini index assessing the purity of the dataset is displayed in Eq. (4).

$$Gini(D) = \sum_{k=1}^{K} p_k(1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2 \tag{4}$$

(6) **SVM.** In SVM, a smart implicit nonlinear mapping without knowing the map function makes a nonlinear problem linearly separable as shown in Fig. 3(d). The implicit mapping is implemented through replacing inner product of two mapped vectors in the feature space by kernel function value of the original vectors. In the feature space, SVM builds an optimal classifier using support vectors by the maximum

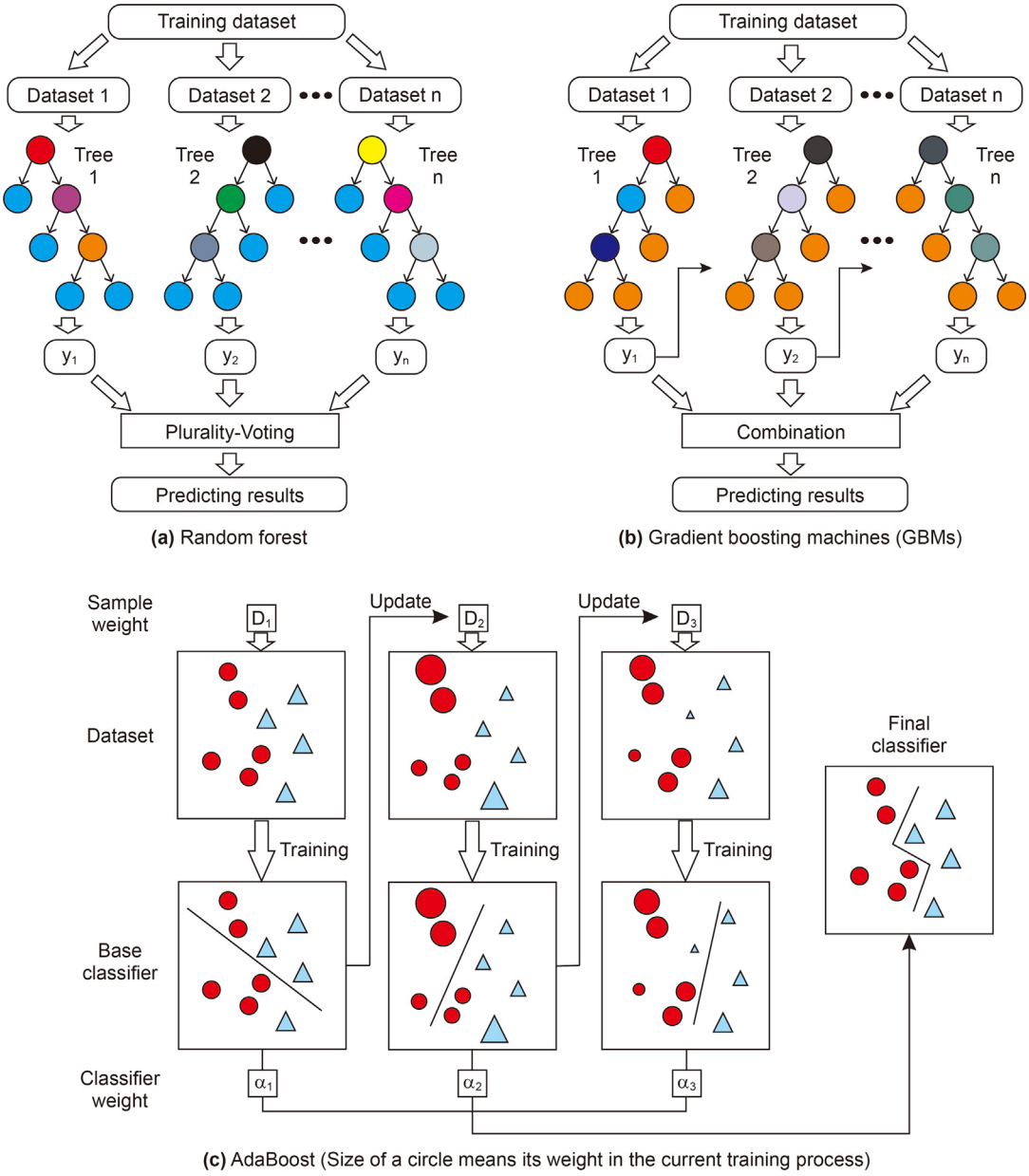

Fig. 4. Schematic diagrams of ensemble-learning-classifiers methods.

margin principle. Support vectors are data points that are closer to the hyperplane classifier.

2.3. Ensemble-learning-classifiers methods will be used as sub-classifiers

(1) **RF.** RF is a typical representative ensemble learning method combining Bagging and random feature selection. The principle of RF is shown in Fig. 4(a). The term "random" refers to generating n sub-datasets by randomly sampling with replacement and randomly removing some features of each sub-dataset. The term "forest" means that sub-classifiers are decision trees. The prediction results are based on a voting mechanism.

(2) **GBMs.** Boosting models based on decreasing gradient algorithms are termed as GBMs (Natekin and Knoll, 2013). As shown in Fig. 4(b), a GBM trains the decision tree base learners in a gradual, additive and sequential manner. The k-th base learner is an update of the $(k-1)$-th base learner, in which the weights of misclassified samples in $(k-1)$-th base learner will be increased in the k-th one. The negative gradients of pseudo-residuals between true and predicted labels against parameters in GBM will help determine a set of optimal parameters. eXtreme gradient boosting (XGBoost) (Chen and Guestrin, 2016; Liu and Wang, 2022) and light gradient boosting machine (LightGBM) (Gu et al., 2021) are recently developed tree-based scalable versions of GBMs. XGBoost applies the idea of gradient tree boosting and optimizes both the objective function and the node splitting of the tree. A regularization term is added to the objective function:

$$Obj^{(s)} = \sum_{i=1}^{n} L\left(y_i, y_i^{(s-1)} + f_s\left(x_i\right)\right) + \Omega\left(f_s\right) \tag{5}$$

where $L(\bullet,\bullet)$ is the loss function; $\hat{y}_i^{(s-1)}$ is the predicted label of sample x_i in the $(s-1)$-th iteration; $f_s(x_i)$ is the new sub-model trained in the s-th iteration; the regularization term $\Omega(f_s) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2$, T and γ are the numbers of leaf nodes and its coefficients, respectively, ω_j is the weight score of each leaf node, and λ is the weight of the leaf nodes.

To alleviate the efficiency and scalability problems of XGBoost for high-dimensional and large data, LightGBM (Gu et al., 2021) is proposed, which combines two innovative technologies, namely gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). GOSS is used to split internal nodes. EFB aims to speed up the training process without losing accuracy, especially for input with high-dimensional and sparse features. Specifically, samples with larger absolute values of gradients (i.e., $a \times 100\%$) are selected as subset A, while the remaining samples with smaller gradients are randomly chosen to form subset B (i.e., $b \times (1-a) \times 100\%$). Here, a and b are sampling ratios of data with large and small gradients, respectively. The samples about the j-th feature are split according to the variance gain $V_j(d)$ on $A \cup B$ in Eq. (6). EFB bundles these features together into a single feature bundle to achieve the purpose of dimensionality reduction.

$$V_j(d) = \frac{1}{n}\left[\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in B_l} g_i\right)^2}{n_l^j(d)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b}\sum_{x_i \in B_r} g_i\right)^2}{n_r^j(d)}\right] \tag{6}$$

where $A_l = \{x_i \in A \quad x_{ij} \le d\}$, $A_r = \{x_i \in A \quad x_{ij} > d\}$, $B_l = \{x_i \in B \quad x_{ij} \le d\}$, $B_r = \{x_i \in B \quad x_{ij} > d\}$ and g_i denotes the negative gradients of the loss function for the LightGBM outputs in each iteration.

(3) **AdaBoost.** Based on the result of the previous base classifier, the weights of each training sample will be revised in current base classifier. Misclassified samples will be setted higher weights while the weights of correctly classified ones will be reduced. The current base classifier will be updated based on the iterated samples, and many base classifiers will be obtained. The base classifiers are combined into a strong classifier after the entire training process is completed. Here, an AdaBoost model using three base classifiers is as an example displayed in Fig. 4 (c).

2.4. Ensemble principles

To construct a good ensemble model, how to select sub-classifiers and set their parameters are crucial (Yang, 2011). Typically, the settings should ensure sub-classifiers high accuracy and diversity, since the success of ensemble learning depends on the trade-off between the accuracy and diversity of sub-classifiers. Hence, ensemble criteria for proper settings are described below.

(1) **Principle for homogeneous ensemble**

DT is the base classifier of RF. The generalization ability of DT is relatively weak. DT often suffers from the issue of overfitting while RF usually works well. In RF with good prediction performance, there is a surprising phenomenon in which DTs in RF are near overfitting or just overfits. Here, an assumption about choosing and setting proper base classifiers is put forward: classifier methods that can overfit are suitable for being base classifiers.

To choose and configure base classifiers of high accuracy, based on the assumption above, an ensemble principle is proposed based on the experience of using RF, in which parameters and sub datasets should near or just make base classifiers reach over-fitting to obtain a high accuracy. In other words, we should choose base classifier that can predict training data with a high accuracy (e.g., >90%) through the operations of parameter setting and sub-data sampling.

Diversity of sub-classifiers depends on the nature of the algorithm itself (algorithm stability). In addition, it can be artificially enhanced by data sampling and feature sampling. To improve the diversity between the built sub-classifier models, the random generation of the sub-dataset in Step 1 of Fig. 2 is employed. This kind of random sampling of data and random sampling of features can be defined as data sampling disturbance and input feature disturbance. These two methods are combined and used in our ensemble strategy.

A series of experiments will be carried out in Sections 3.2-3.5 to test these proposed principles.

(2) **Principle for heterogeneous ensemble**

The combination of sub-classifiers in heterogeneous ensemble is determined in a stepwise way that is based on homogeneous ensemble models as shown in Fig. 5.

Firstly, choose a homogeneous ensemble model with the highest accuracy to become the first sub-classifier of the heterogeneous ensemble model marked as M_j ($j = 0$) which is the basis of the entire heterogeneous ensemble model. M_j represents the j-th ensemble model with accuracy A_j;

Secondly, calculate the double fault (DF) (Giacinto and Roli, 2001) of the rest sub-classifiers and generate a list $D = [h_1, h_2, \ldots, h_N]$ according to DF in decreasing order. DF focuses on the error of two classifiers on the same sample, which measures the diversity of the two classifiers. The expression of DF between i-th and j-th sub-classifiers is shown in Eq. (7) ($i \neq j$). The value range of DF is [0,1]. In the worst case, the error rate of both classifiers is 100%, in which DF equals to 1, accuracy and diversity of the classifiers are the lowest at the same time;

$$DF_{i,j} = \frac{n^{00}}{n} \tag{7}$$

where the total number of samples is n; n^{11} (n^{00}) represents the number of samples that were correctly (wrongly) classified by h_i and h_j, n^{10} represents the number of samples that were correctly classified by h_i and misclassified by h_j, and n^{01} represents the number of samples that were misclassified by h_i, h_j. The number of samples correctly classified, and they satisfy $n^{11} + n^{00} + n^{10} + n^{01} = n$.

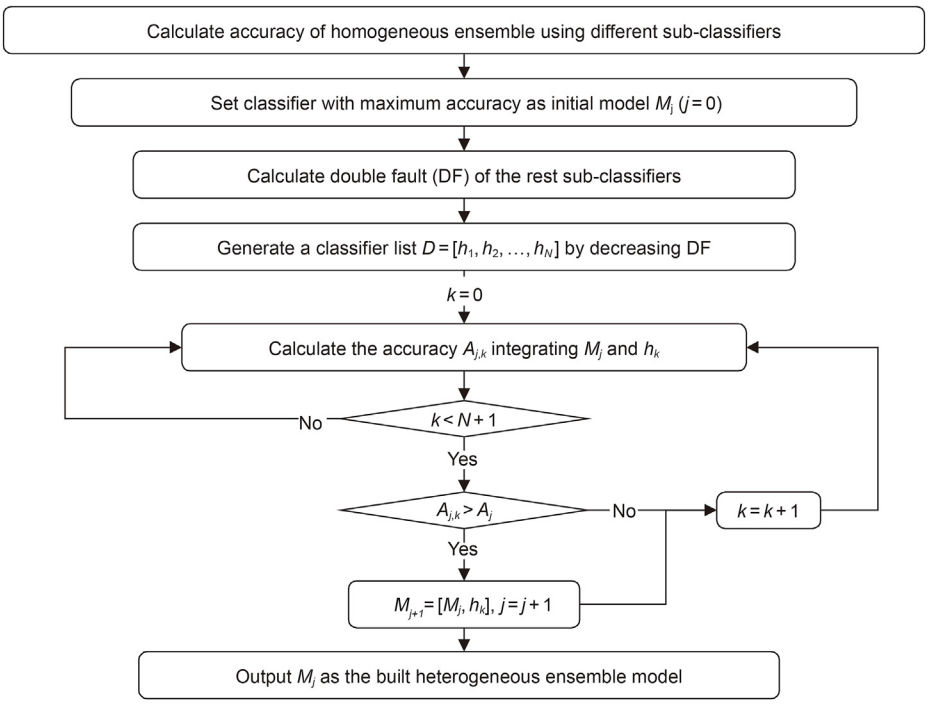

Fig. 5. Workflow of determining a proper combination of sub-classifiers in heterogeneous ensemble.

Thirdly, calculate the accuracy $A_{j,k}$ of ensemble models using M_j and h_k in a sequence of $k = 1, 2, \ldots, N$. If $A_{j,k} > A_j$, M_{j+1} becomes M_j integrated with h_k, $j = j+1$, $k = k+1$; Else, M_j remains the same.

Finally, obtain the combination of sub-classifiers in the heterogeneous ensemble model.

2.5. Metrics for evaluating performance of prediction models

Accuracy, precision, recall and F1-score are four commonly used metrics for evaluating performances of prediction models. Take a confusion matrix of four-categories in Fig. 6(a) as an example to explain these metrics. Let the i-th category as a positive class, and the rest as negative ones. The corresponding two-categories confusion matrix consists of true positive (TP), false positive (FP), false negative (FN) and true negative (TN) as shown in Fig. 6(b). Then accuracy, precision, recall and F1-score can be calculated as shown in Fig. 6(c).

Accuracy is the evaluation of the overall accuracy of the classifier. Precision is the evaluation of the accuracy of the classifier's prediction as a certain analogy. Recall is the ability of the classifier to find relevant instances in the data set. When Precision and Recall need to be considered comprehensively, F1-score is the index of the harmonic value.

2.6. Grid search method for determining optimal parameters in ensemble models

To evaluate the generalization ability of ensemble learning methods effectively, stratified k-fold cross validation (SCV) technique is applied to unbalanced dataset used as shown in the left

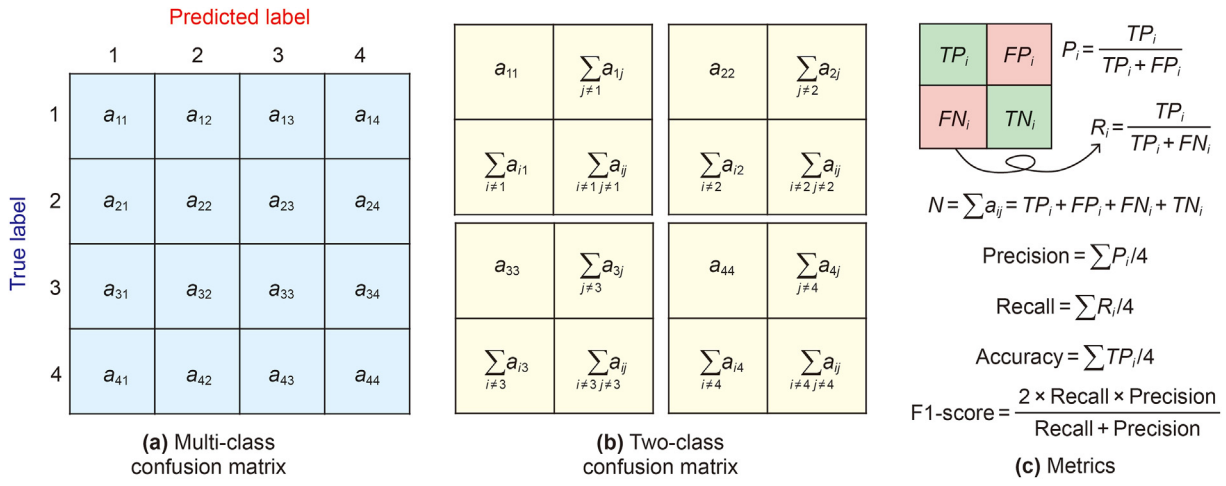

Fig. 6. The confusion matrix and evaluation metrics.

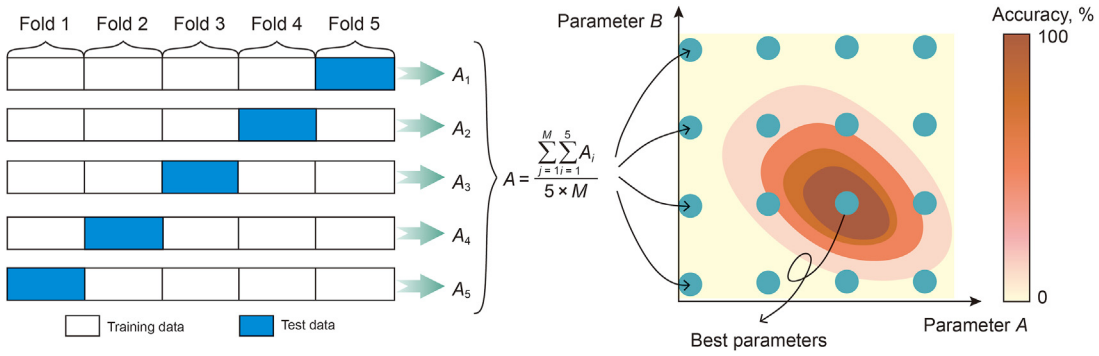

Fig. 7. Grid search and SCV.

part of Fig. 7. Distribution of each class in each fold is nearly the same.could be ensured.

All labelled data will be divided into subsets of *k* folds. *k*-1 subsets will combine into a training dataset and the rest of one fold data will be as the test data. The training dataset will be randomly sampled into subsets for each base learner to train an ensemble prediction model. Then the accuracy of the rest of one fold data will be predicted by the built model. Each fold of data will have an accuracy. The average accuracy will be as a metric to measure the performance of the proposed ensemble.

Many parameter pairs in an ensemble model are displayed in the right part of Fig. 7. Grid search technique is a method to determine optimal parameters for a prediction model. Each pair of parameters will have a further average accuracy by repeating the *k*-fold process *M* times. As shown in Fig. 7, optimal parameters will be obtained by the grid search method (Dong et al., 2020a).

3. Experiments of lithofacies identification by ensemble learning strategy and principles

In this work, all methods are programmed by Python and all calculations are conducted on an Intel (R) Core (TM) computer with 2.6 GHz CPU and 8 GB of RAM. The k-fold cross-validation, LDA, NB, KNN, SVM, DT, RF and AdaBoost are implemented in the scikit-learn library, while XGBoost and LightGBM are implemented by the XGBoost and LightGBM libraries, respectively.

3.1. Data set

The dataset used to test the ensemble learning strategy and principles is from the Daniudi Gas Field (DGF) published in the work of Xie et al., (2018). The DGF is located in the eastern portion of the Yishan Slope of Ordos Basin in China, which is one of the main gas-bearing areas of the Upper Paleozoic in the northern part of the basin (Fig. 8). The main target formations from bottom to top are the Carboniferous Taiyuan, Permian Shanxi and Xiashihezi formations. The sedimentary environment of the target formations is a fluvial-deltaic depositional environment.

This dataset has 915 samples which are consisted of well logs and the corresponding lithofacies labels obtained by rock core. For each sample, there are 7 properties, namely gamma ray log (GR), acoustic log (AC), caliper log (CAL), density log (DEN), compensated neutron log (CNL), deep investigation log (LLD) and shallow investigation log (LLS). Lithofacies are divided into eight types, including carbonate rock (CR), coal (C), pebbly sandstone (PS), coarse sandstone (CS), medium sandstone (MS), fine sandstone (FS), siltstone (S) and mudstone (M). To eliminate the influence of data range, the linearization normalization method is used to convert the data to the range of [0,1], that is $x^* = (x - x_{\min})/(x_{\max} - x_{\min})$ (Dong et al., 2020c).

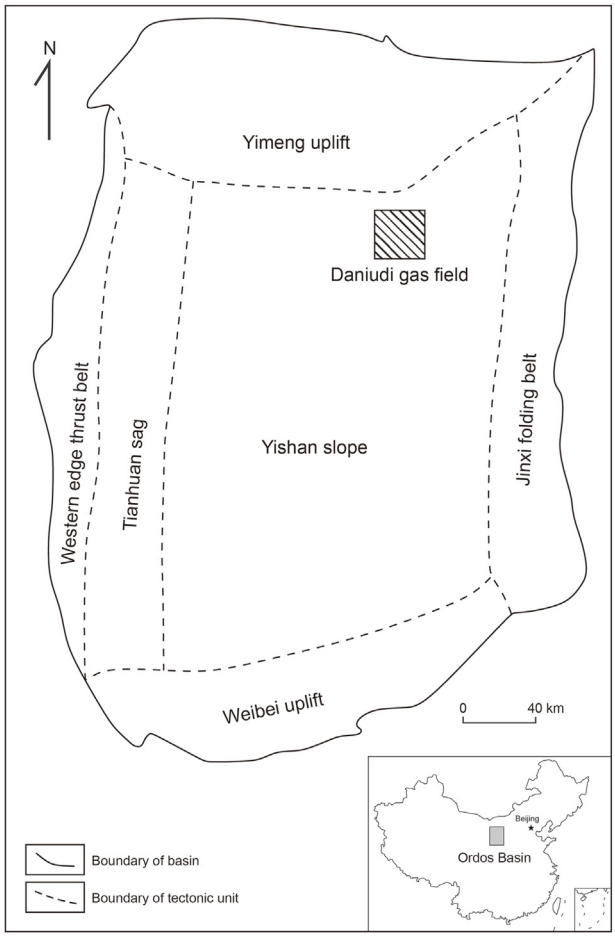

Fig. 8. Tectonic and geographic location of Daniudi gas field (by (Dong et al., 2020b), changed).

3.2. Homogeneous ensemble learning using common single-classifier methods

Six commonly used single-classifier methods are chosen as base classifiers in homogeneous ensemble learning, that is to say LDA, NB, KNN, SVM using RBF kernel, ID3 and CART. In Fig. 9, the red dotted lines show accuracies of single-classifier methods: LDA (51.1%), NB (57.7%), KNN (81.3%), SVM (83.5%), ID3 (72.5%) and CART

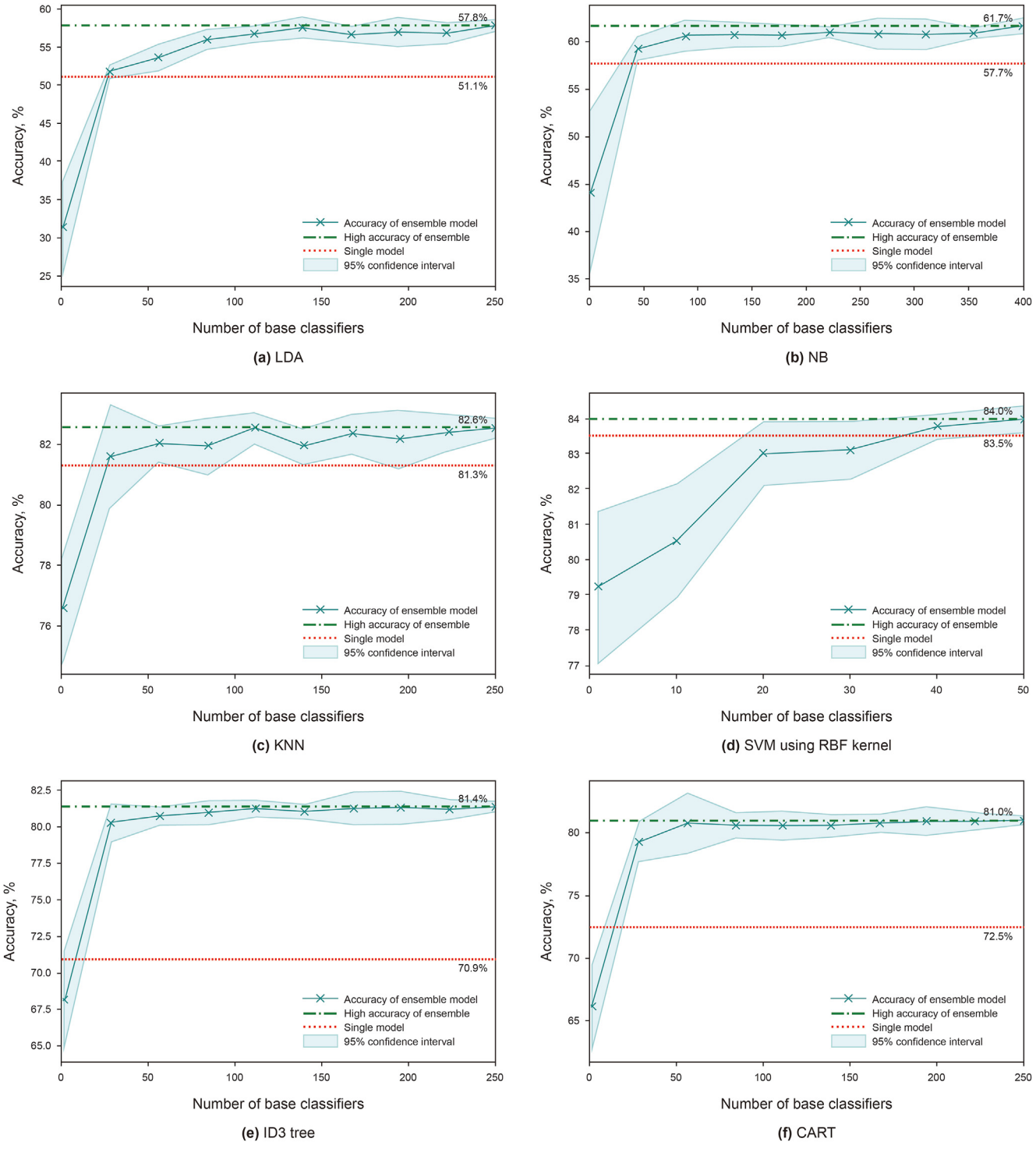

Fig. 9. Comparison of ensemble learning using regular classifiers.

(70.9%). SVM and KNN perform best, ID3 and CART perform second best, NB and LDA worst.

After homogeneous ensemble learning, the improved LDA, NB, KNN, SVM, ID3 and CART reach accuracies of 57.8%, 61.7%, 82.6%, 84.0%, 81.0%, 81.4%, respectively. They are marked by green dashed lines in Fig. 9. For the improved methods, KNN, SVM, ID3 and CART can perform best and NB and LDA worst. However, all the improved methods obtain an increase in accuracy compared with their corresponding single-classifier. The improved gaps of LDA, NB, KNN, SVM, ID3 tree and CART are 6.7%, 4.0%, 1.3%, 0.5%, 8.5%, 10.5%, respectively. The relatively two weak single-classifiers (ID3 and CART) improve significantly (>8%), the two weakest single-classifiers (LDA and NB) improve by about 5%, while the two strong single-classifiers (KNN and SVM) improve slightly (<1.5%).

As shown in Fig. 9, all the homogeneous ensemble models will gradually increase with the increase of the number of base classifiers and converge to specific upper limits of accuracy (green dashed lines). The light blue shaded area is drawn to represent the

95% confidence interval based on the accuracy obtained by five repeated experiments. Red dotted lines, green dashed lines and light blue shaded area in Section 3.3 and 3.4 have the same meanings as those expressed here.

Besides the number of base classifiers, another two key parameters need to be determined, which are data sampling ratio (r_b) and feature sampling ratio (r_p). The optimal r_b and r_p of homogeneous ensemble models corresponding to each point in Fig. 9 are obtained by the grid search method. The determination processes are displayed in Fig. 10. The accuracy corresponding to each pair of r_b and r_p are expressed by height of z-axis and colors (red on behalf of the high accuracy, and blue low). The highest accuracy point is marked by a circle. These grid search processes only select those with the highest accuracies as representatives. The optimal of r_b and r_p corresponding to LDA, NB, KNN, SVM, ID3 and CART are (0.022,1), (0.06,0.9), (3,0.8), (1.6,0.9), (2.8,0.7), and (2.8,0.8), respectively.

Accuracy, recall, precision and F1-score of different single-classifiers are shown in Fig. 11. The black error lines represent the 95% confidence interval for 5 repeated experiments. For LDA, NB, KNN, SVM, CART and ID3, relative to single models, the recall improvements of ensemble models are 4.2%, 2.5%, 0.6%, 0.8%, 9.0%, 11.0%, respectively; the precision improvements are 12.9%, 7.5%, 0.6%, 1.4%, 8.9%, 12.1%, respectively; the F1-score improvements are 5.5%, 3.1%, 0.6%, 0.7%, 9.1%, 11.6%, respectively.

Increased value and small variance of four different evaluation metrics for each classifier indicate ensemble models of single-classifiers have stronger generalization ability and stability than single-classifiers. The ensemble strategy improves single-classifiers and there are different improvements. Note there are large variance in precision and recall for LDA, which indicates that LDA tends to identify samples as certain lithologies.

3.3. Homogeneous ensemble learning using ensemble-learning base classifiers

Four representative ensemble-classifier methods are chosen as base classifiers in homogeneous ensemble learning, namely RF, AdaBoost, XGBoost and LightGBM. As displayed in Fig. 12, the accuracies of homogeneous ensemble classifier gradually rise with the increase of the number of sub-classifiers, and eventually tend to stabilize. The stable accuracies are 81.7% (RF), 83.3% (AdaBoost), 82.1% (XGBoost), 83.0% (LightGBM), respectively. The light blue shaded area and the red dashed lines have the same meaning as Fig. 9 above. Red dashed lines show accuracies of base classifiers: RF (79.8%), AdaBoost (81.3%), XGBoost (81.0%), LightGBM (82.3%). Obviously, there are improvements for homogeneous ensemble using ensemble-learning base classifiers compared to common ensemble learning methods.

The other two key parameters r_b and r_p are obtained by the grid search method similar to those in Section 3.2. The determinations are displayed in Fig. 13. The optimal r_b and r_p of RF, AdaBoost, XGBoost and LightGBM are (2.8,0.8), (3,0.8), (2.5,0.7) and (0.9,2.7), respectively. They are marked by circles.

For base-classifiers of RF, AdaBoost, XGBoost and LightGBM, the recalls of ensemble models improve 2.9%, 2.3%, 0.6%, 1.4%, respectively, compared with the original models; the precisions improve 2.3%, 1.8%, 0.8%, 1.1%, respectively; the F1-score improve 2.7%, 2.1%, 0.9%, 1.1%, respectively, as shown in Fig. 14. Even though the improvements are relatively low compared with single-classifiers, the generalization abilities of the prediction models (four common ensemble learning methods) are enhanced.

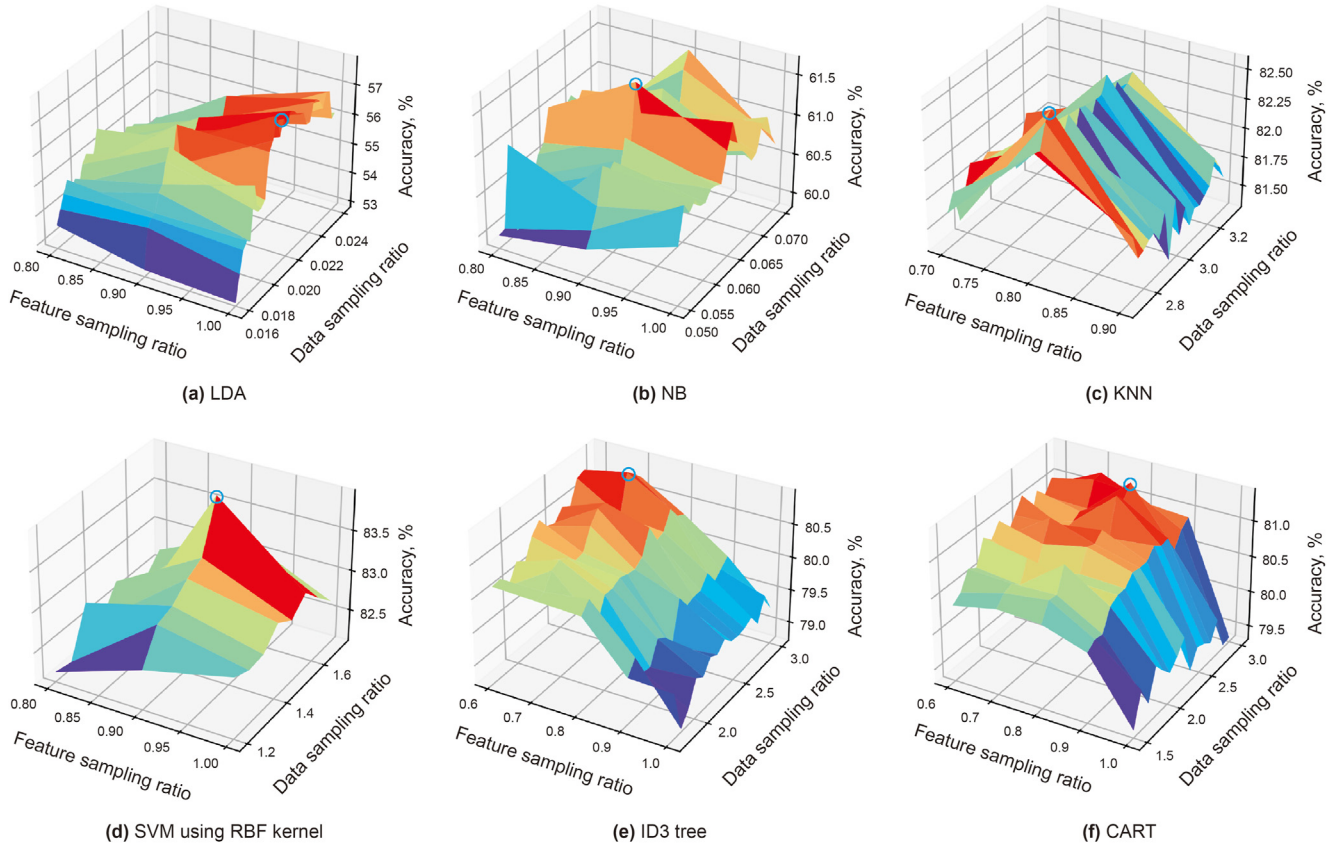

Fig. 10. Optimal parameter determination in homogeneous ensemble by different single-classifiers.

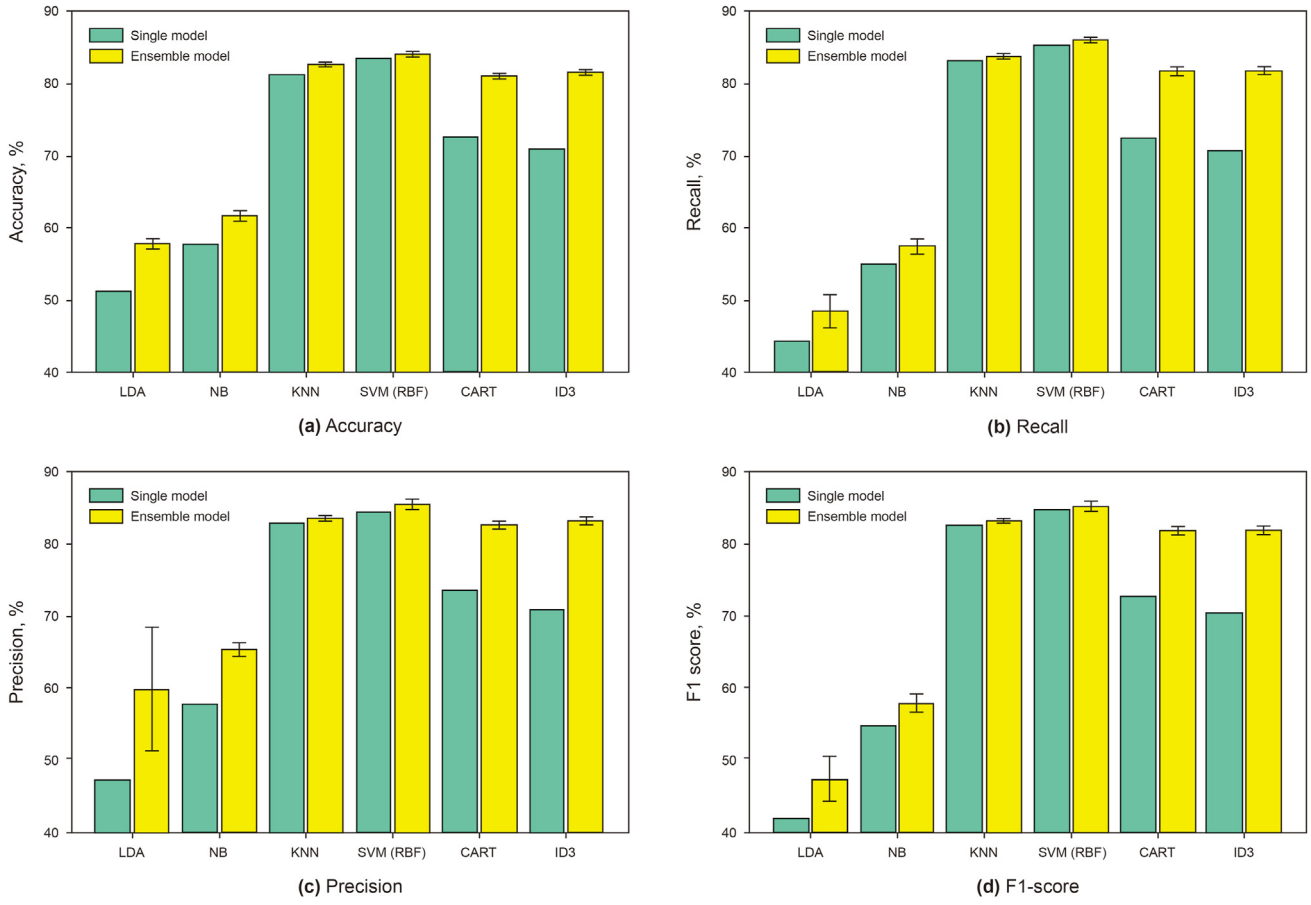

Fig. 11. Comparison of single-classifier methods and the corresponding ensemble-classifiers.

3.4. Heterogeneous ensemble learning using a mixture of component classifiers

The homogeneous models with relatively high accuracy will be as the candidate base classifiers in heterogeneous ensemble, which include SVM (84.0%), AdaBoost (83.3%), LightGBM (83.0%), KNN (82.6%), XGBoost (82.1%) and RF (81.7%). According to the workflow in Fig. 5, the building process of the heterogeneous ensemble is shown in Fig. 16. Because SVM homogeneous ensemble model has the highest accuracy and generalization ability, it is preferred as the basis of the heterogeneous ensemble process. The next base classifier will be chosen by the double fault (DF) values between this classifier and the current heterogeneous ensemble model. DF measures the diversity between this method and the current heterogeneous ensemble model. The following heterogeneous ensemble process will be:

(1) Calculate DF between SVM and other five homogeneous ensemble models, which are AdaBoost (0.1136), LightGBM (0.1202), KNN (0.1147), XGBoost (0.1191) and RF (0.1202), respectively. Hence, AdaBoost is put into heterogeneous ensemble. Due to an increase of accuracy (+0.3%), the current heterogeneous ensemble becomes [SVM + AdaBoost];
(2) Calculate DF between [SVM + AdaBoost] and other four homogeneous ensemble models, which are LightGBM (0.1303), KNN (0.1322), XGBoost (0.1338) and RF (0.1404), respectively. Therefore, LightGBM is put into heterogeneous ensemble. Due to an increase in accuracy (+0.6%), the current heterogeneous ensemble becomes [SVM + AdaBoost + LightGBM];
(3) Calculate DF between [SVM + AdaBoost + LightGBM] and other three homogeneous ensemble models, which are KNN (0.1429), XGBoost (0.1434) and RF (0.1398), respectively. Therefore, RF is put into heterogeneous ensemble. Due to a decrease in accuracy (−0.5%), the current heterogeneous ensemble model remains unchanged;
(4) Subsequently, put KNN with the second largest DF (0.1429) into [SVM + AdaBoost + LightGBM]. Due to a decrease in accuracy (−0.2%), the current heterogeneous ensemble model remains unchanged;
(5) Finally, put XGBoost into heterogeneous ensemble. Due to a decrease in accuracy (−0.1%), the current heterogeneous ensemble model remains unchanged.

Hence, the final heterogeneous ensemble is [SVM + AdaBoost + LightGBM]. There is an accuracy increase of 0.9% than the highest (SVM) homogeneous ensemble model.

Not only the evaluation metrics value of the final heterogeneous ensemble model could reach a high level, but also the variance is relatively small (Fig. 16). Results show that the model has higher generalization ability and relative stability, revealing that the proposed heterogeneous ensemble strategy is feasible.

3.5. Comparison of classification by ensemble methods

Accuracies of all experiments are shown in Fig. 17. The orange and blue bars represent average accuracy values before and after ensemble, and the red line is the best accuracy. The accuracy changes are shown in Fig. 18.

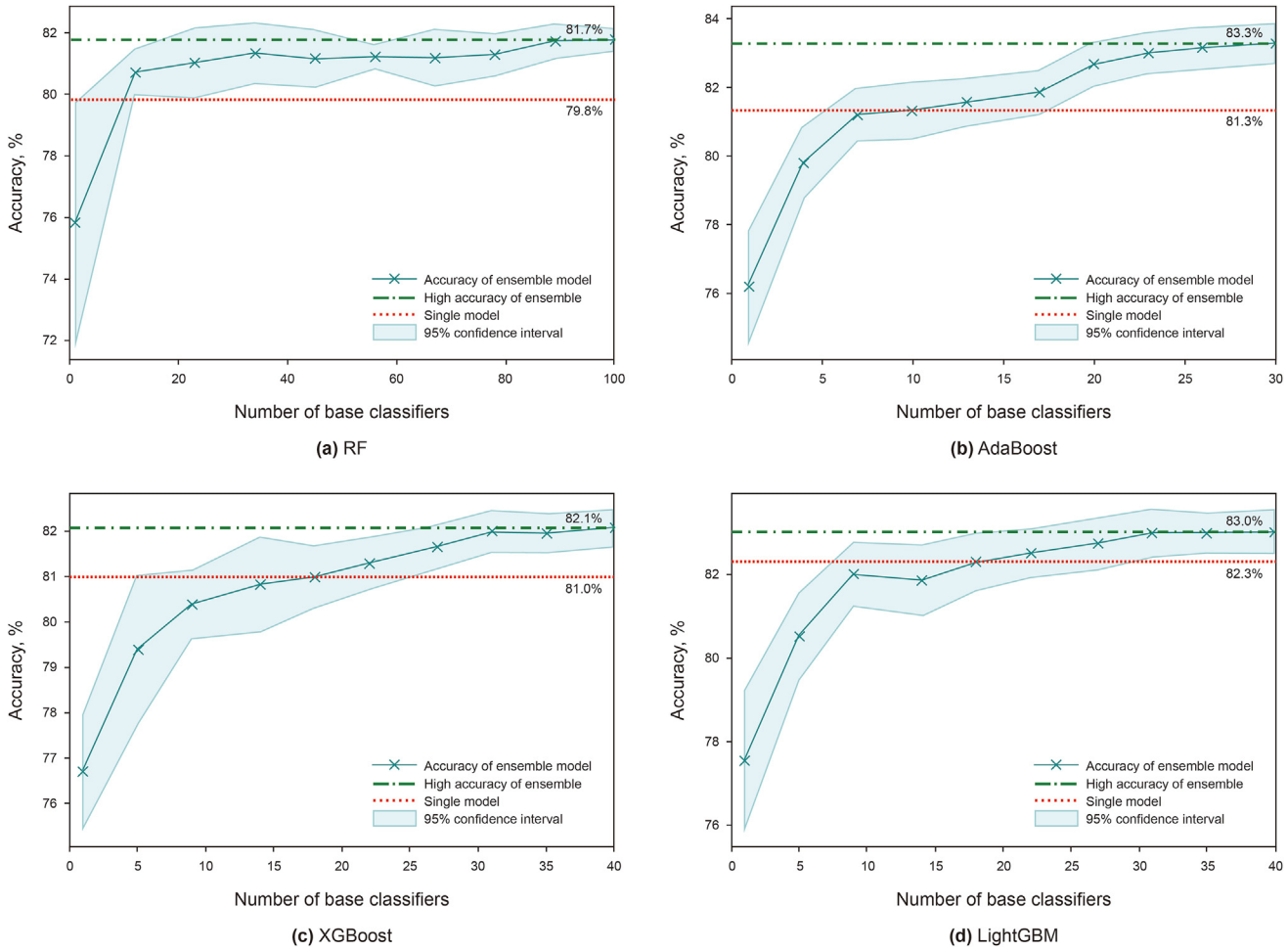

Fig. 12. Comparison of ensemble learning using ensemble base classifiers.

A single homogenous ensemble of the classifiers through the established ensemble strategy can improve the classification accuracy generally, and single classifiers as a sub-classifier have a larger increase compared with ensemble learning classifiers as sub-classifier. The model with the best generalization ability is heterogeneous ensemble model obtained by heterogeneous ensemble strategy.

4. Discussions

4.1. Which kind of method can be as sub-classifiers of ensemble models

Typically, a general idea is that ensemble learning is suitable for weak sub-classifiers. Weak classifiers refer to ones that perform only slightly better than a random classifier (Zhou, 2011). In practice, DT method is mostly used as a sub-classifier. There are rarely research focusing on other methods as sub-classifiers. This paper implements ensemble learning using not only DT (ID3 and CART) but also other single-classifiers (LDA, NB, KNN and SVM) and ensemble-learning-classifiers (RF, AdaBoost, XGBoost and LightGBM) as sub-classifiers. As shown in Figs. 9, Fig. 11, Figs. 14 and 12 in Sections 3.2 and 3.3, all the homogeneous ensemble models can perform better than the corresponding sub-classifier methods in accuracy, recall, precision and F1-score. Fig. 18 displays the change in accuracy between ensemble models and their corresponding sub-classifiers. The results indicate that (1) the improvements of DTs are highest (≥8.5%); (2) the improvements of LDA and NB are second-highest (≥4%); (3) the improvements of ensemble-learning-classifiers are third-highest (≥0.7%); (4) the improvements of KNN and SVM are relatively low (≥0.5%). The highest improvements of DTs are consistent with common sense that DTs are suitable for ensemble learning. It should be noted that the improvements of other methods mean the ensemble strategy and principles can aid to build better lithofacies identification models compared with the original machine learning methods. Hence, not only DTs but also other single-classifiers and ensemble-learning-classifiers can be sub-classifiers of homogeneous ensemble learning.

Different from homogeneous ensemble learning, not several kinds of classifiers can be combined to obtain a better prediction model as shown in Fig. 15. However, based on the results in Section 3.4, heterogeneous ensemble using a proper combination of component classifiers can obtain a better prediction model than the used homogeneous ensemble models. As shown in Fig. 17, the built heterogeneous ensemble model obtained the highest accuracy of 84.9%, and higher than the highest homogeneous ensemble model using SVM (83.9%). Therefore, the selection of component classifiers is important for heterogeneous ensemble learning.

4.2. How to select proper sub-classifiers for ensemble learning

The gap between models before and after ensemble can demonstrate the effectiveness of the ensemble learning strategy

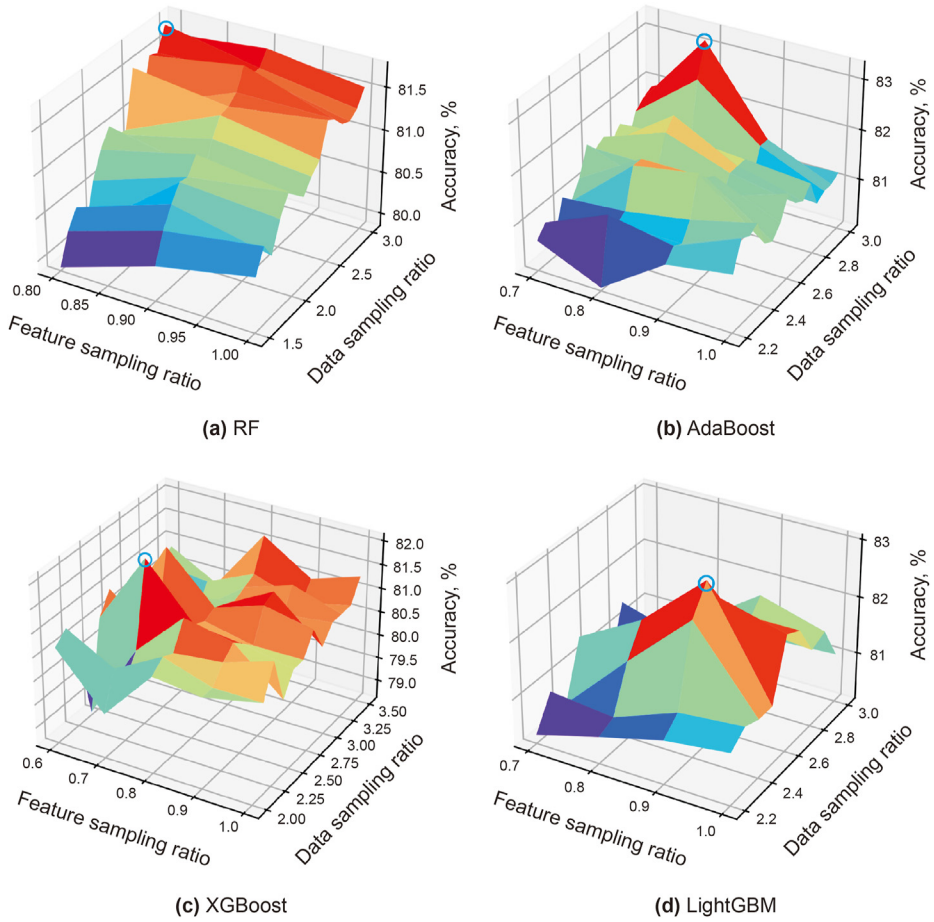

Fig. 13. Optimal parameters determination in different homogeneous ensembles.

and principles for lithofacies identification. In practice, how select proper sub-classifiers for ensemble learning should be paid more attentions since it can help build a powerful prediction model. The effect of the ensemble learning strategy and ensemble principles used in experiments shows that these methods illustrated in Section 2.4 can indeed play a role in enhancing the generalization ability. The aims of selecting sub-classifiers are to obtain high accuracy and diversity for sub-classifiers.

To obtain high accuracy in homogeneous ensemble learning, the most effective approach is to set model parameters to ensure each sub-classifier has a high accuracy (e.g., >90%) for the sub training dataset. In fact, an over-fitting state is needed for training sub-classifiers. Besides, sub-data and properties sampling can also help sub-classifier obtain high accuracy.

High diversity can be guaranteed by algorithm stability and random generation of sub-dataset. If sub-classifier is stable, such as SVM in Section 3, the built sub-classifier model will have low diversity. DTs are relatively unstable, so the corresponding ensemble models obtain obvious improvements as shown in Fig. 18. If a sub-classifier is chosen, sampling of sub-dataset includes samples and their properties will be the most effective way, which should make the sub-classifier model nearly overfitting and unstable.

It should be noted that there is a contradictory relationship between the accuracy and diversity of sub-classifiers. Therefore, the core problem of ensemble learning research always revolves around how to combine the two, to produce a sub-classifier whose two properties are effectively compatible.

To some extent, homogeneous ensemble can be the basis of building a good homogeneous ensemble model. Based on accuracy and DF, the stepwise addition and removal of sub-classifiers illustrated in Fig. 5 give a practical approach to determining an optimal combination of sub-classifiers for a homogeneous ensemble.

4.3. Just overfitting of base classifiers for the training in the ensemble process

Just overfitting is an effective approach to improving the accuracy of ensemble models. Overfitting means base classifiers can have higher accuracy for training data (>90%) regardless of that of test data. Typically, an accuracy >95% is better. If a base classifier can obtain an accuracy for training data close to 100% (>99%), it will have a good potential for the ensemble. Just overfitting means parameters in a base classifier can make it just close to overfitting. Take the construction of the SVM (RBF) ensemble model as an example. The C parameter in SVM plays a pivotal role in overfitting. For $C = 1400$, the accuracy of training data is 89%; for $C = 1500$, accuracy is 98%; for $C = 1510$, accuracy is 99.9%; for $C > 1520$, accuracy is 99.9%, too. Then the parameter C for making SVM just overfitting is about 1510. Therefore, this principle is suitable for base classifiers with parameters, such as SVM. However, for methods without adjustable parameters, such as LDA and NB, the proposed principle will not work well.

4.4. Why the ensemble strategy and principles can improve sub-classifiers

Lithofacies identification by machine learning is a classification problem of supervised learning, in which expected error E on an

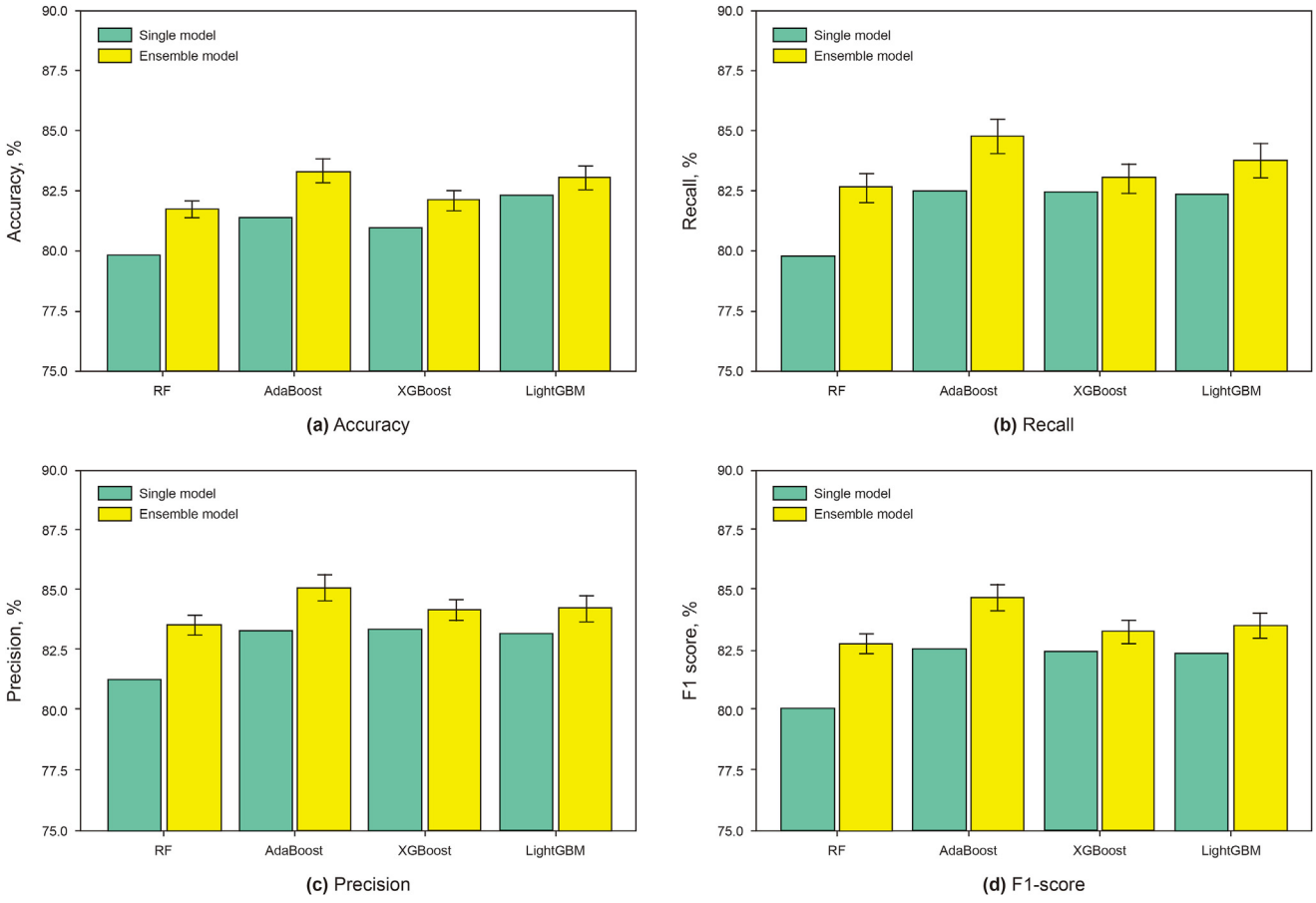

Fig. 14. Comparison of the ensemble base classifiers and the corresponding ensemble classifiers.

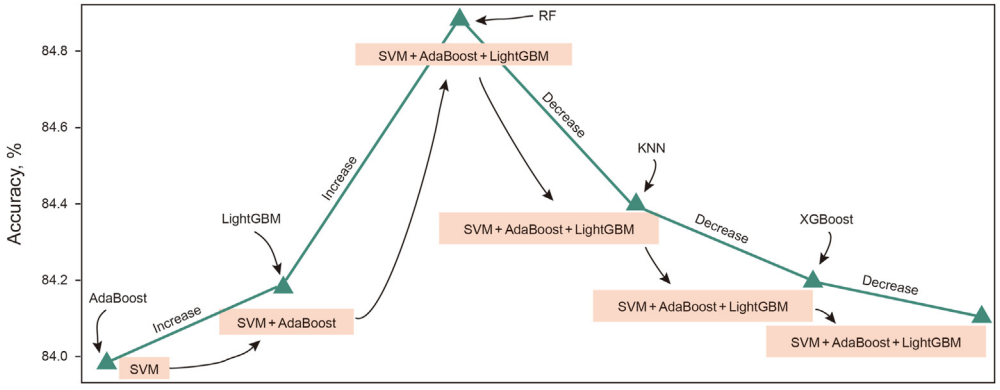

Fig. 15. Accuracy change trend of heterogeneous ensemble.

unseen sample consists of three terms, namely bias, variance (*var*), and a quantity called the base error σ^2 (irreducible) resulting from noises in the problem itself (Bauer and Kohavi, 1999). *E* is shown in Eq. (8).

$$E = bias^2 + var + \sigma^2 \tag{8}$$

where *bias* is an error from erroneous assumptions in a learning algorithm. Methods with high bias typically produce simpler models that may fail to capture important regularities (i.e. underfit) in the data; var is an error from sensitivity to small fluctuations in the training set. High-variance learning methods may be able to represent their training set well but are at risk of overfitting to noisy or unrepresentative training data.

The key difference between ensemble learning methods and other machine learning methods is that they focus on the problem of bias-variance tradeoff (Sun and Zhou, 2018). This problem is the conflict in attempting to simultaneously minimize these two sources of error that prevent supervised learning algorithms from generalizing beyond their training set. From statistical aspects, the ensemble strategy in this paper aims to reduce variances of prediction models, while the principle of just overfitting for each subclassifier aims to decrease the bias error. The reasons, that this ensemble strategy and principles can work, lie in: (1) high accuracy

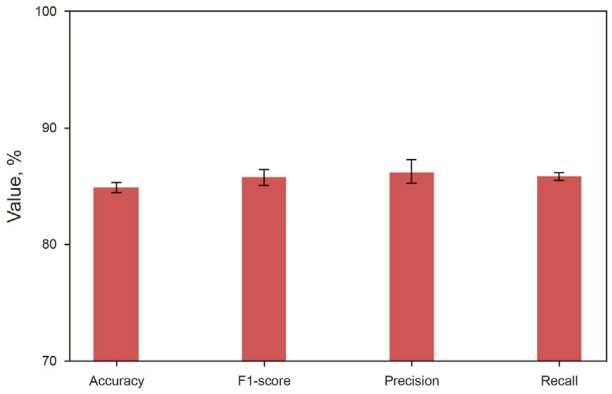

Fig. 16. Evaluation metrics of the built heterogeneous ensemble model.

of sub-classifiers can lower the upper bound of the variance of the ensemble model; (2) high diversity among sub-classifier models can reduce the lower bound of the variance (Sun and Zhou, 2018).

The following analyses try to explain the reasons why the proposed ensemble strategy and principles work for different sub-classifier methods used above mentioned.

The foundation of NB is feature independence, which ensures stable classification efficiency and capacity. Nonetheless, when there is feature dependence in the dataset, the generalization ability of the built NB model will decrease a lot. Except for GR and other features, there is a relatively high correlation (>0.65 or < -0.65) between other features in this case as exhibited in Fig. 19. The sampling module in this ensemble strategy can reduce the correlation of features in the dataset used by each sub-classifier. Besides NB is good at addressing small sample problems for each sub-data. The ensemble can not only improve the difference between the base classifiers but improves the accuracy of each base classifier so that the classification accuracy after ensemble is improved to a certain extent.

LDA is a linear feature extraction and classification method. For complex nonlinear lithofacies identification problems, the bias error of LDA will be large. In other words, the sub-classifier of LDA cannot fit the pattern of training data well. Hence, it is not the best choice for the ensemble. However, due to a decrease in variance errors, the ensemble LDA can also obtain an improvement.

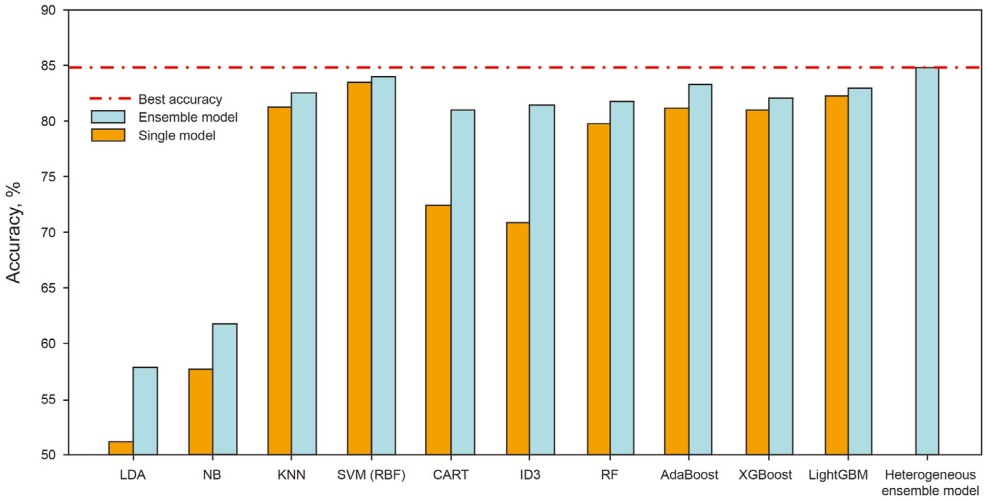

Fig. 17. Comparison of classification accuracy after adopting ensemble learning strategy.

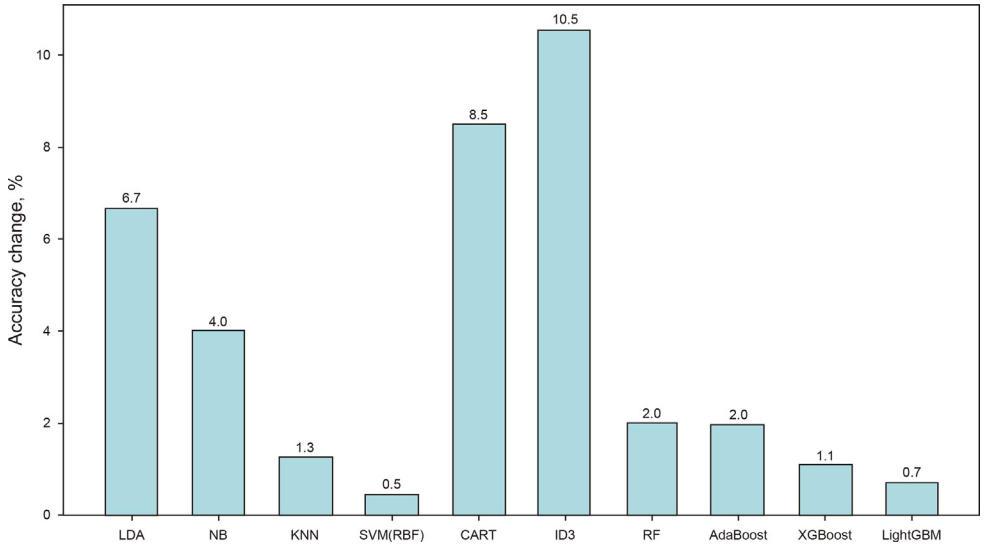

Fig. 18. Comparison of accuracy changes before and after ensemble.

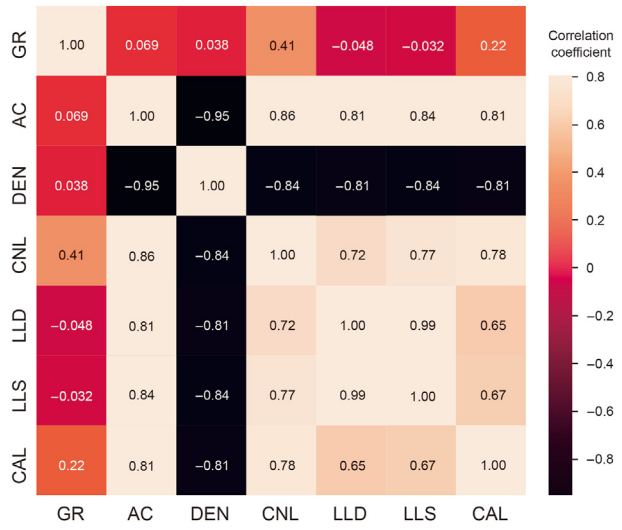

Fig. 19. Correlation matrix of seven features in the dataset.

DT is a relatively weak classifier prone to overfitting, but it can meet the requirements of the ensemble principle and is suitable for reducing bias errors. Besides, the ensemble strategy will reduce variance errors of the ensemble DT model. Therefore, the ensemble DTs achieve obvious enhancements.

KNN, SVM and ensemble classifiers all have high stability. In fact, the diversity will not be very high, so the improvement will not be quietly obvious. The general understanding is that these classifiers are not suitable for the ensemble. However, the improvement after ensemble indicates that the ensemble strategy and ensemble principles can help improve their generalization ability. The improvements benefit from the decrease of variance due to the ensemble strategy and the reduction of bias due to the ensemble principles. Even though the increases of these methods are not the best, their accuracy of lithofacies identification is high. Hence, in practice, the ensemble of these kinds of methods should be paid more attention to, which is usually ignored.

4.5. Other issues

(1) Data augmented sampling methods

The random sampling ratio in the paper can be greater than 1, which is different from the previous random sampling method of the sub-classifier. The effect of data enhancement of the sub-training set is realized. The classification accuracy of the best single RF model (single model) in the RF ensemble experiment in the above experiments is lower than the ensemble learning classifier formed in the CART ensemble experiment. This shows that this data augmentation method does achieve the purpose of further enhancing the generalization ability of the ensemble model for most classifiers.

(2) Why ANN is not used in this work

The improvement of generalization ability is important for a good ensemble, but time efficiency is also significant for practical applications. When ANN deals with supervised learning problems with high-dimensional and nonlinear classification, a large number of neurons need to be trained to obtain a machine learning model with high generalization ability. The huge number of neurons makes the time complexity of training the model extremely large (Fig. 20). The optimal parameter determination for ANN by grid searching takes over 16 hours far more than other classifiers, which seriously affects the practical application. In addition to ANN, GBDT has extremely high training time complexity due to the defects of the algorithm itself, so ANN and GBDT are not recommended in this work.

The consumed time for training optimal ensemble models one time and their corresponding base models is shown in Fig. 20(a). The bar heights represent the average time in each training of the grid searching and black error lines represent the 95% confidence intervals. For training single methods, single-classifier methods (LDA, NB, KNN, SVM, CART, ID3) uses 0.001–0.006s except ANN uses 7.93s, and ensemble-classifier methods (RF, AdaBoost, XGBoost, LightGBM, GBDT) uses 0.55–1.6s. For training ensemble models by the proposed ensemble strategy and principles, the ensemble models based on single-classifiers use 0.17–12.27s, while those based on ensemble-classifiers use 3.15–41.33s. There are increases of 6–26 times in time consuming of the proposed methods. In practice, the time to build an ensemble method by the proposed ensemble strategy and principles, and optimal parameter determination should be considered, too. If the grid searching method is used, the total number of parameter grid point repeats will be needed compared with the training one time in Fig. 20(a). For example, when the SVM using RBF kernel involving C and gamma parameters is constructed as an ensemble model with 50 base classifiers, the average time consumed is 5.344s. After the grid search method for data sampling ratio and feature sampling ratio is used in Fig. 10 (d), the total time to complete the experiment is about 13.4 min. Hence, in terms of time consumption, methods with fewer parameters will be preferred.

The time consumed by the built optimal models to predict a new sample is shown in Fig. 20(b). Single methods (LDA, NB, KNN, SVM, CART, ID3, ANN, RF, AdaBoost, XGBoost, LightGBM, GBDT) use <0.05 s. For ensemble models by the proposed methods, the used time is less than 0.2 s except for KNN (1.11 s) and SVM (0.41 s). In general, the time consumed by the proposed ensemble increases but it is acceptable.

In future work, the time of determining an optimal ensemble model still need more work. For example, use gradient-free optimization methods to determine parameters instead of grid searching.

(3) Comparison with the original paper used the same dataset

In the original paper, five machine learning methods namely, NB, SVM, ANN, RF and GBDT, are selected by the author for identification of lithofacies from the Daniudi dataset (Xie et al., 2018). Three metrics precision, recall and F1-score were employed to evaluate the built classifiers. The best prediction model is built by RF with precision (82.9%), recall (80.0%) and F1-score (80.8%). In our work, the single-classifier RF can obtain precision (81.3%), recall (79.7%) and F1-score (80.0%), which are similar to the original work. Note that the ensemble RF model can achieve precision (83.6%), recall (82.6%) and F1-score (82.7%). Besides the best homogeneous ensemble model is the ensemble SVM model with precision (86.0%), recall (86.0%) and F1-score (85.4%); the best heterogeneous ensemble model (SVM + AdaBoost + LightGBM) has precision (86.3%), recall (85.9%) and F1-score (85.7%).

In general, the ensemble learning strategy and principles can improve machine learning in lithofacies identification.

(4) Quality control of lithofacies labels of well logs in training data

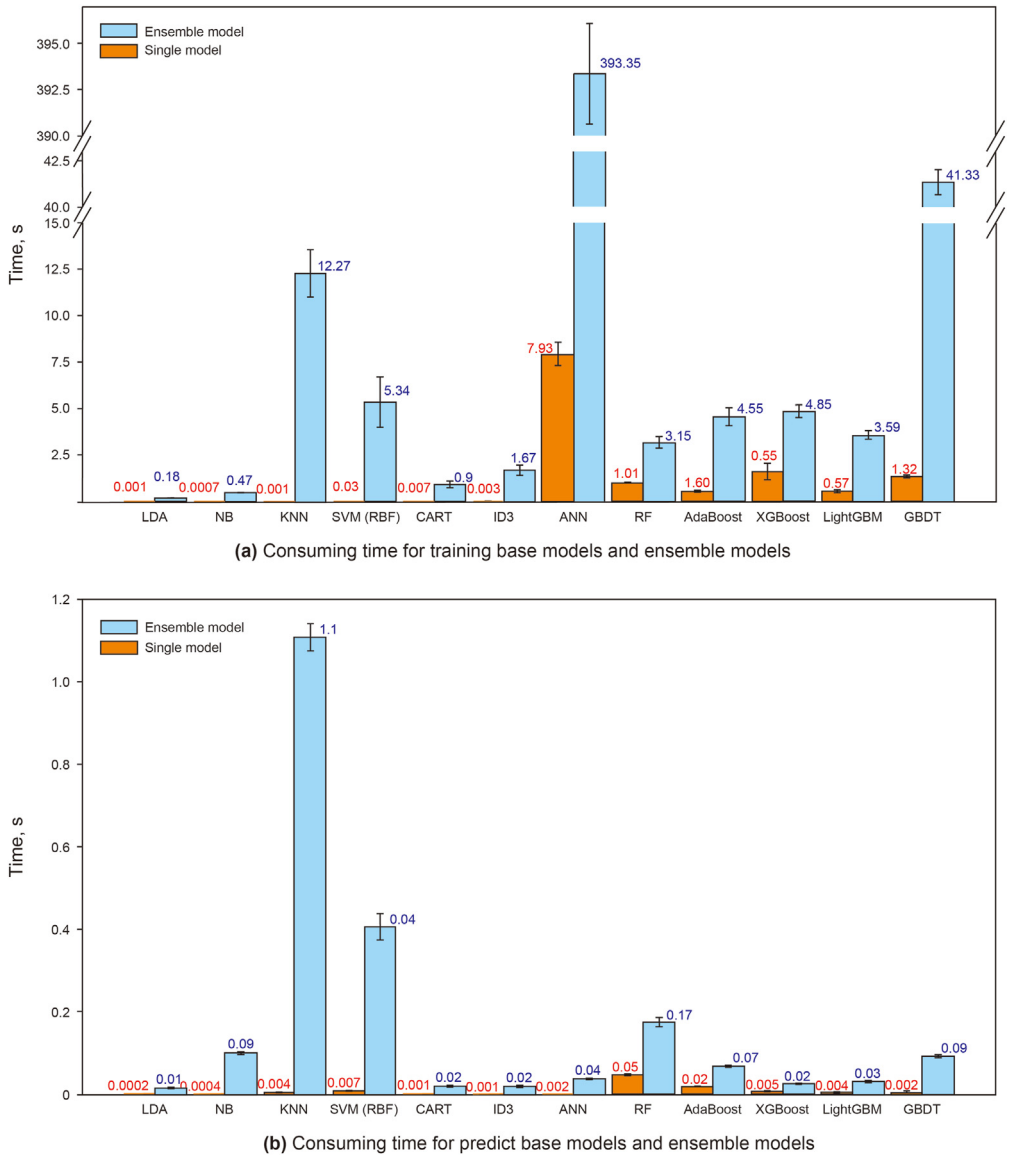

Fig. 20. Consuming time for training ensemble models training.

The interpreted labels of samples are quite important for the following lithofacies identification by machine learning methods. To build a stable and accurate model of lithofacies identification, more attention should be paid to quality control of labelled well log data. In this work, we choose a dataset from a published peer-reviewed paper to ensure the quality of labelled training data, so that we can pay more attention to the improvements of machine learning methods suitable for lithofacies identification.

(5) Built models of lithofacies identification by single methods

In this work, all prediction models of single methods are optimized by grid searching methods, which can ensure an equal comparison with the proposed method. All optimal hyperparameters for each of the single machine learning classifiers are shown in Appendix.

(6) Further works

A proper combination of heterogeneous sub-classifiers is significant for heterogeneous ensemble learning. This paper proposed an approach to determine sub-classifiers as shown in Fig. 5. However, there still needs more work to improve the efficiency of heterogeneous ensemble principles. For example, choose a better measurement instead of DF. Besides, how to more effectively improve the diversity and accuracy of sub-classifiers in heterogeneous ensemble should be paid more attentions in further works.

5. Conclusions

A simple ensemble strategy and several novel ensemble principles are proposed to aid geologists not familiar with ML to

establish a good lithofacies identification model. They can also help geologists familiar with ML further improve the accuracy of lithofacies identification.

The ensemble strategy is a generalization of the ensemble in random forest. The ensemble principles are to make sub-classifiers just overfitting by algorithm parameter setting and sub-dataset sampling. In the prediction, the ensemble strategy aims to decrease variance errors, while the ensemble principles try to reduce bias errors. The novel ensemble principles help make the ensemble strategy more practical. Based on a series of comparison experiments between original methods and the homogenous and heterogeneous ensemble methods, conclusions are listed below:

(1) All homogeneous ensemble models improve compared with the corresponding single-classifiers (e.g., LDA, NB, KNN, SVM, CART, and ID3). Among them, the ensemble ID3 obtain an accuracy increase of 10.5%. All homogeneous ensembles using ensemble-learning-classifiers (e.g., RF, AdaBoost, XGBoost, and LightGBM) also improve over their sub-classifiers, in which both the ensemble RF and ensemble AdaBoost increase by 2.0%. Single-classifiers (especially DT) are typically regarded as good candidates for ensemble learning, but these results indicate not only DT but also other single-classifiers and ensemble-learning-classifiers can be sub-classifiers of homogeneous ensemble learning;
(2) For homogeneous ensemble, the best model in this work is the ensemble SVM with an accuracy of 83.9%, which is higher than the most commonly used Random forests (ensemble ID3, 81.4%, and ensemble CART, 81.0%). This demonstrates other single-classifiers except random forests may obtain a better ensemble prediction model for lithofacies;
(3) The ensemble principles, which make sub-classifiers just overfitting by algorithm parameter setting and sub-dataset sampling, for homogeneous ensemble are proven effective for the introduced ensemble strategy;
(4) The heterogeneous ensemble model (SVM + AdaBoost + LightGBM) achieves the best accuracy of 84.9% in this work.
(5) Not all heterogeneous ensemble is better than homogeneous ones, but a proper combination of heterogeneous sub-classifiers can obtain an improvement. The proposed heterogeneous ensemble principle based on double fault (DF) is proven effective. In practice, heterogeneous ensemble is more suitable for building a more powerful lithofacies identification model, though it is complex.

There is still further work to be explored in the future, such as revisions and improvements of ensemble principles for both homogeneous and heterogeneous ensembles.

Acknowledgment

This work was financially supported by the National Natural Science Foundation of China (Grant No. 42002134), China Postdoctoral Science Foundation (Grant No. 2021T140735) and Science Foundation of China University of Petroleum, Beijing (Grant Nos. 2462020XKJS02 and 2462020YXZZ004). Special thanks to Yunxin Xie (Chengdu University of Technology, Chengdu) for opening the dataset for lithofacies identification in his work, which is used in this paper.

Appendix

Table A1
Optimal hyperparameters for each of the single machine learning classifiers

Single classifier	Hyperparameters	Optimal value
KNN	n_neighbors	1
SVM(RBF)	C	690
	gamma	30
CART	max_depth	30
ID3	max_depth	25
RF	n_estimators	500
	max_depth	30
	max_features	4
AdaBoost	base_estimator	'DecisionTreeClassifier'
	n_estimators	200
	max_depth	25
	max_features	4
	learning_rate	0.35
	algorithm	'SAMME'
XGBoost	n_estimators	400
	learning_rate	0.05
	max_depth	10
	colsample_bytree	0.6
	booster	'gbtree'
LightGBM	boosting_type	'gbdt'
	n_estimators	400
	learning_rate	0.14
	max_depth	25

References

Al-Anazi, A., Gates, I.D., 2010. On the capability of support vector machines to classify lithology from well logs. Nat. Resour. Res. 19 (2), 125–139. https://doi.org/10.1007/s11053-010-9118-9.
Anifowose, F., Labadin, J., Abdulraheem, A., 2015. Improving the prediction of petroleum reservoir characterization with a stacked generalization ensemble model of support vector machines. Appl. Soft Comput. 26, 483–496. https://doi.org/10.1016/j.asoc.2014.10.017.
Ao, Y.L., Li, H.Q., Zhu, L.P., et al., 2018. Logging lithology discrimination in the prototype similarity space with random forest. Ieee Geosci Remote S 16 (5), 687–691. https://doi.org/10.1109/lgrs.2018.2882123.
Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach. Learn. 36 (1), 105–139. https://doi.org/10.1023/A:1007515423169.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140. https://doi.org/10.1007/bf00058655.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.
Breiman, L.I., Friedman, J.H., Olshen, R.A., et al., 2015. Classification and regression trees. Ency Ecol 57 (3), 582–588. https://doi.org/10.1007/978-3-642-57292-0_10.
Bressan, T.S., de Souza, M.K., Girelli, T.J., et al., 2020. Evaluation of machine learning methods for lithology classification using geophysical data. Comput Geosci-Uk 139, 104475. https://doi.org/10.1016/j.cageo.2020.104475.
Bühlmann, P., Yu, B., 2002. Analyzing bagging. Ann. Stat. 30 (4), 927–961. https://doi.org/10.1214/aos/1031689014.
Busch, J.M., Fortney, W.G., Berry, L.N., 1987. Determination of lithology from well logs by statistical analysis. SPE Form. Eval. 2 (4), 412–418. https://doi.org/10.2118/14301-pa.
Chen, T.Q., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. https://doi.org/10.1145/2939672.
Corina, A.N., Hovda, S., 2018. Automatic lithology prediction from well logging using kernel density estimation. J. Petrol. Sci. Eng. 170, 664–674. https://doi.org/10.1016/j.petrol.2018.06.012.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

https://doi.org/10.1023/A:1022627411411.
Delfiner, P., Peyret, O., Serra, O., 1987. Automatic determination of lithology from well logs. SPE Form. Eval. 2 (3), 303–310. https://doi.org/10.2118/13290-pa.
Dev, V.A., Eden, M.R., 2018. Evaluating the boosting approach to machine learning for formation lithology classification. Comput Aided Chem Eng 44, 1465–1470. https://doi.org/10.1016/b978-0-444-64241-7.50239-1.
Dev, V.A., Eden, M.R., 2019. Formation lithology classification using scalable gradient boosted decision trees. Comput. Chem. Eng. 128, 392–404. https://doi.org/10.1016/j.compchemeng.2019.06.001.
Dong, S.Q., Wang, Z.Z., Zeng, L.B., 2016. Lithology identification using kernel Fisher discriminant analysis with well logs. J. Petrol. 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., 2022. 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. 210, 110081. https://doi.org/10.1016/j.petrol.2021.110081.
Dong, S.Q., Zeng, L.B., Liu, J.J., et al., 2020c. 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., 2020a. 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.
Dong, S.Q., Zeng, L.B., Lyu, W.Y., et al., 2020b. Fracture identification and evaluation using conventional logs in tight sandstones: a case study in the Ordos Basin, China. Energy Geosci 1, 115–123. https://doi.org/10.1016/j.engeos.2020.06.003.
Dubois, M.K., Bohling, G.C., Chakrabarti, S., 2007. Comparison of four approaches to a rock facies classification problem. Comput Geosci-Uk 33 (5), 599–617. https://doi.org/10.1016/j.cageo.2006.08.011.
Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55 (1), 119–139. https://doi.org/10.1006/jcss.1997.1504.
Friedman, J.H., Hall, P., 2006. On bagging and nonlinear estimation. J. Stat. Plann. Inference 137 (3), 669–683. https://doi.org/10.1016/j.jspi.2006.06.002.
Friedman, J.H., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann. Stat. 28 (2), 337–407. https://doi.org/10.1214/aos/1016218223.
Giacinto, G., Roli, F., 2001. Design of effective neural network ensembles for image classification purposes. Image Vis Comput. 19, 699–707. https://doi.org/10.1016/s0262-8856(01)00045-2.
Gorai, A.K., Raval, S., Patel, A.K., et al., 2021. Design and development of a machine vision system using artificial neural network-based algorithm for automated coal characterization. Int J Coal Sci Tech 8 (4), 737–755. https://doi.org/10.1007/s40789-020-00370-9.
Gu, Y.F., Zhang, D.Y., Lin, Y.B., et al., 2021. Data-driven lithology prediction for tight sandstone reservoirs based on new ensemble learning of conventional logs: a demonstration of a Yanchang member, Ordos Basin. J. Petrol. Sci. Eng. 207, 109292. https://doi.org/10.1016/j.petrol.2021.109292.
He, L., Wen, K., Wu, C.C., et al., 2019. A corroded natural gas pipeline reliability evaluation method based on multiple intelligent algorithms. Petrol Sci Bull 4 (3), 310–322. https://doi.org/10.3969/j.issn.2096-1693.2019.03.028 (in Chinese).
Hou, E.K., Wen, Q., Ye, Z.N., et al., 2020. Height prediction of water-flowing fracture zone with a genetic-algorithm support-vector-machine method. Int J Coal Sci Tech 7 (4), 740–751. https://doi.org/10.1007/s40789-020-00363-8.
Kardani, N., Bardhan, A., Samui, P., et al., 2022. Predicting the thermal conductivity of soils using integrated approach of ANN and PSO with adaptive and time-varying acceleration coefficients. Int. J. Therm. Sci. 173, 107427. https://doi.org/10.1016/j.ijthermalsci.2021.107427.
Kolose, S., Stewart, T., Hume, P., et al., 2021. Prediction of military combat clothing size using decision trees and 3D body scan data. Appl. Ergon. 95, 103435. https://doi.org/10.1016/j.apergo.2021.103435.
Kuncheva, L.I., Whitaker, C.J., 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach. Learn. 51 (2), 181–207. https://doi.org/10.1023/A:1022859003006.
Lawal, A.I., Aladejare, A.E., Onifade, M., et al., 2021. Predictions of elemental composition of coal and biomass from their proximate analyses using ANFIS, ANN and MLR. Int J Coal Sci Tech 8 (1), 124–140. https://doi.org/10.1007/s40789-020-00346-9.
Li, F.Y., Zhang, X.P., Chen, B., et al., 2013. JPEG steganalysis with high-dimensional features and Bayesian ensemble classifier. IEEE Signal Process. Lett. 20 (3), 233–236. https://doi.org/10.1109/LSP.2013.2240385.
Li, X.Y., Li, H.Q., Zhou, J.Y., et al., 2011. Identification of the quaternary low gas-saturation reservoirs in the sanhu area of the qaidam basin, China. Petrol. Sci. 8 (1), 49–54. https://doi.org/10.1007/s12182-011-0114-4.
Li, Y.M., Anderson-Sprecher, R., 2006. Facies identification from well logs: a comparison of discriminant analysis and naïve Bayes classifier. J. Petrol. Sci. Eng. 53, 149–157. https://doi.org/10.1016/j.petrol.2006.06.001.
Liu, R.C., Jiang, Y.J., Huang, N., et al., 2018. Hydraulic properties of 3D crossed rock fractures by considering anisotropic aperture distributions. Adv Geo-Energy Res. 2 (2), 113–121. https://doi.org/10.26804/ager.2018.02.01.
Liu, S.S., Wang, Z.M., 2022. Reservoir grain size profile prediction of multiple sampling points based on a machine learning method. Petrol Sci Bull 7 (1), 93–105. https://doi.org/10.3969/j.issn.2096-1693.2022.01.009 (in Chinese).
Liu, X.Y., Zhou, L., Chen, X.H., et al., 2020. 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.
Ma, Y., 2011. Lithofacies clustering using principal component analysis and neural network: applications to wireline logs. Math. Geosci. 43 (4), 401–419. https://doi.org/10.1007/s11004-011-9335-8.
Martyushev, D.A., Yurikov, A., 2021. Evaluation of opening of fractures in the Logovskoye carbonate reservoir, Perm Krai, Russia. Petrol Res 6 (2), 137–143. https://doi.org/10.1016/j.ptlrs.2020.11.002.
Moja, S.S., Asfaw, Z.G., Omre, H., 2019. Bayesian inversion in hidden markov models with varying marginal proportions. Math. Geosci. 51 (4), 463–484. https://doi.org/10.1007/s11004-018-9752-z.
Natekin, A., Knoll, A., 2013. Gradient boosting machines, a tutorial. Front. Neurorob. 7, 21. https://doi.org/10.3389/fnbot.2013.00021.
Opitz, M., Waltner, G., Possegger, H., et al., 2018. Deep metric learning with bier: boosting independent embeddings robustly. Ieee T Pattern Anal 42 (2), 276–290. https://doi.org/10.1109/tpami.2018.2848925.
Qiao, X., Chang, F., 2021. Underground location algorithm based on random forest and environmental factor compensation. Int J Coal Sci Tech 8 (5), 1108–1117. https://doi.org/10.1007/s40789-021-00418-4.
Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn. 1 (1), 81–106. https://doi.org/10.1007/bf00116251.
Quinlan, J.R., 1996. Improved use of continuous attributes in C4.5. J. Artif. Intell. Res. 4, 77–90. https://doi.org/10.1613/jair.279.
Saggaf, M.M., Nebrija, E.L., 2000. Estimation of lithologies and depositional facies from wire-line logs. AAPG Bull. 84 (10), 1633–1646. https://doi.org/10.1306/8626bf1f-173b-11d7-8645000102c1865d.
Schapire, R.E., 1990. The strength of weak learnability. Mach. Learn. 5 (2), 197–227. https://doi.org/10.1007/bf00116037.
Sebtosheikh, M.A., Motafakkerfard, R., Riahi, M., et al., 2015. Support vector machine method, a new technique for lithology prediction in an Iranian heterogeneous carbonate reservoir using petrophysical well logs. Carbonates Evaporites 30 (1), 59–68. https://doi.org/10.1007/s13146-014-0199-0.
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.
Sun, T., Zhou, Z.H., 2018. Structural diversity for decision tree ensemble learning. Front. Comput. Sci. China 12 (3), 560–570. https://doi.org/10.1007/s11704-018-7151-8.
Sun, Z.B., Song, Q.B., Zhu, X.Y., et al., 2015. A novel ensemble method for classifying imbalanced data. Pattern Recogn. 48 (5), 1623–1637. https://doi.org/10.1016/j.patcog.2014.11.014.
Tewari, S., Dwivedi, U.D., 2019. Ensemble-based big data analytics of lithofacies for automatic development of petroleum reservoirs. Comput. Ind. Eng. 128, 937–947. https://doi.org/10.1016/j.cie.2018.08.018.
Tokhmechi, B., Memarian, H., Noubari, H.A., et al., 2009. A novel approach proposed for fractured zone detection using petrophysical logs. J. Geophys. Eng. 6 (4), 365–373. https://doi.org/10.1088/1742-2132/6/4/004.
Tripoppoom, S., Ma, X., Yong, R., et al., 2019. Assisted history matching in shale gas well using multiple-proxy-based Markov chain Monte Carlo algorithm: the comparison of K-nearest neighbors and neural networks as proxy model. Fuel 262, 116563. https://doi.org/10.1016/j.fuel.2019.116563.
Wang, G.C., Ju, Y.W., Huang, C., et al., 2017. Longmaxi-Wufeng Shale lithofacies identification and 3-D modeling in the northern Fuling gas field, Sichuan Basin. J. Nat. Gas Sci. Eng. 47, 59–72. https://doi.org/10.1016/j.jngse.2017.10.003.
Wang, X.D., Yang, S.C., Zhao, Y.F., et al., 2018. Lithology identification using an optimized KNN clustering method based on entropy-weighed cosine distance in Mesozoic strata of Gaoqing field, Jiyang depression. J. Petrol. Sci. Eng. 166, 157–174. https://doi.org/10.1016/j.petrol.2018.03.034.
Wang, Y.M., Li, Y.B., Li, X.P., et al., 2020. Recent progress on ANN-based pipeline erosion predictions. Petrol Sci Bull 5 (1), 114–121. https://doi.org/10.3969/j.issn.2096-1693.2020.01.011 (in Chinese).
Xie, Y.X., Zhu, C.Y., Zhou, W., et al., 2018. Evaluation of machine learning methods for formation lithology identification: a comparison of tuning processes and model performances. J. Petrol. Sci. Eng. 160, 182–193. https://doi.org/10.1016/j.petrol.2017.10.028.
Yang, L., Ran, B., Han, Y., et al., 2019. Sedimentary environment controls on the accumulation of organic matter in the upper ordovician Wufeng–lower silurian Longmaxi mudstones in the southeastern Sichuan Basin of China. Petrol. Sci. 16 (1), 44–57. https://doi.org/10.1007/s12182-018-0283-5.
Yang, L.Y., 2011. Classifiers selection for ensemble learning based on accuracy and diversity. Procedia Eng. 15, 4266–4270. https://doi.org/10.1016/j.proeng.2011.08.800.
Zhang, Z.J., Zuo, R.G., Xiong, Y.H., 2016. A comparative study of fuzzy weights of evidence and random forests for mapping mineral prospectivity for skarn-type Fe deposits in the southwestern Fujian metallogenic belt, China. Sci. China Earth Sci. 59 (3), 556–572. https://doi.org/10.1007/s11430-015-5178-3.
Zhou, Z.H., 2011. When semi-supervised learning meets ensemble learning. Front. Electr. Electron. Eng. China 6, 6–16. https://doi.org/10.1007/s11460-011-0126-2.