Marine and Petroleum Geology 150 (2023) 106168

Contents lists available at ScienceDirect

Marine and Petroleum Geology

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

Lithology identification using graph neural network in continental shale oil reservoirs: A case study in Mahu Sag, Junggar Basin, Western China

Guoqing Lu [a,b], Lianbo Zeng [a,b,d,*], Shaoqun Dong [a,c,**], Liliang Huang [e], Guoping Liu [d], Mehdi Ostadhassan [f], Wenjun He [e], Xiaoyu Du [d], Chengpeng Bao [a,c]

[a] State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum, Beijing, 102249, China
[b] College of Geoscience, China University of Petroleum, Beijing, 102249, China
[c] College of Science, China University of Petroleum, Beijing, 102249, China
[d] Institute of Energy, Peking University, Beijing, 100871, China
[e] PetroChina Xinjiang Oilfield Company, Xinjiang, Kalamay, 834000, China
[f] Department of Petroleum Engineering, University of North Dakota, Grand Forks, ND, 58202, United States

ARTICLE INFO

Keywords:
GraphSAGE
Lithology identification
Graph construction
Conventional well logs
Continental shale oil reservoirs

ABSTRACT

The continental shale oil reservoir of Fengcheng Formation in the northern slope area of Mahu Sag, Junggar Basin, Western China is very heterogeneous in lithology. Thus, the complex response characteristics of conventional logging and limited core availability in the study area has led to major challenges in lithology identification. Therefore, to resolve lithology identification by well logs in continental shale oil reservoirs, a graph neural network (GNN) method named GraphSAGE is used to train the lithology identification model based on a constructed graph, which connects the samples with adjacent depth and similar log response features on operator intention. The identification process is divided into two parts: first, based on the formation depth sequence and affinity propagation clustering method, the vertical distribution of the stratum and nodes logging curve similarity information are integrated into the graph structure, which structurally represents the conventional logging curves as graph instead of well logs as input data; Second, the nodes of the constructed graph are classified by GraphSAGE, which naturally supports combination generalization and improves sample complexity accompanied by strong relational inductive bias. To examine the effectiveness of GraphSAGE for lithology identification, a conventional log dataset labelled by direct core observations from two separate wells in Muhu Sag are used. The identification results showed that the accuracy of GraphSAGE for the lithologies exceeds 90% of the testing data, especially for transitional lithology such as dolomitic mudstone, silty mudstone and tuffaceous fine sandstone. Compared with the commonly used machine learning methods such as SVM, RF and XGBoost, GraphSAGE was more accurate in lithology identification, matching core observations. Collectively, this reflects the superiority of graph neural network in conventional logging lithology identification and effective means provided for lithology identification of continental shale oil reservoir in the Mahu Sag.

1. Introduction

With recent increase in oil and gas exploration and development from unconventional resources, shale plays in particular, they have become the topic of research from different perspectives (Ghosh et al., 2018; Gong et al., 2021; Sohail et al., 2022; Wang et al., 2022a). To date, majority of this research has been conducted on marine shale while continental shale has also shown to have high exploration and development potential, specifically in China (De Silva et al., 2015; Soeder, 2018). In 2016, PetroChina estimated that the technically recoverable resources of lacustrine shale oil is around 105 $\times$ 10^9 barrels and is mainly distributed in the Paleogene of Bohai Bay Basin, Cretaceous of Songliao Basin, Jurassic of Sichuan Basin, Triassic of Ordos Basin, Permian of Junggar basin, Paleogene-Neogene in Qaidam Basin, Permian in Santanghu Basin and Paleogene in Jianghan Basin (Du et al., 2019; Jin et al., 2021). Unlike the oil from marine shale, oil that is produced from lacustrine shale is generally heavier, has higher viscosity and poor liquidity (Jin et al., 2021).

* Corresponding author. State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum, Beijing, 102249, China.
** Corresponding author. State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum, Beijing, 102249, China.
E-mail addresses: lbzeng@sina.com (L. Zeng), dshaoqun@163.com (S. Dong).

https://doi.org/10.1016/j.marpetgeo.2023.106168
Received 26 October 2022; Received in revised form 19 January 2023; Accepted 4 February 2023
Available online 6 February 2023

Based on what was said above, the Permian Fengcheng Formation in the Mahu Sag in the northwest margin of Junggar basin contains medium-high maturity alkaline high-quality source rocks, which belongs to an alkaline lake basin shale oil system with generation and expulsion of hydrocarbons in multiple stages (Cao et al., 2015; Zhi et al., 2021). Furthermore, geological investigations of the Fengcheng Formation have delineated a favorable shale oil exploration zone with an area of >2300 km^2 in the central part of the Mahu Sag (Tang et al., 2021). Therefore, in order to expand shale oil exploration efforts, Well MY1 was drilled at a relatively a shallow depth in the continental slope of the Mahu Sag in 2008. The peak daily oil production of oil in 2020 from this well was 318 barrels, proving a prolific prospect (Tang et al., 2021). The continental shale oil reservoir of the Fengcheng Formation shows a mixed lithologies, such as dolomitic, clastic and volcanic rocks, etc. Due to various sources of sedimentation. Variations in lithologies have led to differences in reservoir quality, petrophysical properties, fracture density and oil content (Tang et al., 2021; Zhi et al., 2021). For example, in comparison to mudstone, which is generally common in marine shale, sandstone that is abundant in continental shale with larger intergranular pore will cause a higher porosity and permeability in the reservoir. Furthermore, regional or local tectonic activities in continental shale, would cause rocks with higher content of brittle minerals, such as dolomite and dolomitic mudstone to crack and form natural fractures that would increase fluid flow and hydrocarbon migration. This being said, clarifying the lithological characteristics of Fengcheng Formation in the Mahu Sag is of great significance for future exploration and development of continental shale oil.

To accurately characterize lithology in the reservoir, core observation and description would be necessary. However, due to limited number of cores and short intervals of cored wells, it is impossible to systematically describe the lithologies of the entire target formation of key wells in the study area. In this regard, conventional logging data that has high vertical resolution, good continuity and convenient acquisition operation can become useful (Xie et al., 2018; Ameur-Zaimeche et al., 2020). Moreover, its well understood that lithology would have a direct or indirect influence on the well logs which makes using conventional logging data highly beneficial (Dong et al., 2022). However, lithology identification from conventional logging data can become challenging when we have a mixture of minerals which is the case in the Fengcheng Formation, i.e., dolomitic mudstone, silty mudstone and tuffaceous fine sandstone. Besides, logging data can also get affected by other geological and drilling factors such as clay minerals, saturation of fluids, contamination of the formation with drilling fluids, etc., which makes lithology identification and classification a complicated nonlinear problem (Dong et al., 2022). Thus, to overcome the above issue, machine learning (ML) and artificial intelligence (AI) that has been applied in classification, regression and clustering of complex nonlinear problems in different disciplines can become useful (Tang and White, 2008; Delavar, 2022). In recent years, using ML methods to identify lithology has become a main direction in logging technology (Avseth and Mukerji, 2002; Tang, 2008; Dong et al., 2016; Xie et al., 2018; Abbas and Al-Mudhafar, 2021; Yuan et al., 2022). These machine learning methods can help geologists improve the efficiency and accuracy of lithology identification that can include fisher discriminant analysis (Dong et al., 2016), k-nearest neighbors clustering, KNN (Wang et al., 2018); support vector machine, SVM (Sebtosheikh et al., 2015; Al-Mudhafar, 2017a; Li et al., 2020; Dong et al., 2022); back propagation neural network, BPNN (Rogers et al., 1992; Wang et al., 2021); random forest, RF (Sebtosheikh et al., 2015; Ao et al., 2019); gradient boost decision tree, GBDT (Dev and Eden, 2019; Du et al., 2019; Yu et al., 2021); extreme gradient boosting, XGBoost (Al-Mudhafar, 2020; Han et al., 2021); Probabilistic neural networks, PNN (Specht, 1990; Al-Mudhafar, 2017b); convolutional neural network, CNN (Huang et al., 2021; Wang et al., 2022b); recurrent neural network, RNN (Li et al., 2021a; Tian et al., 2021; Zeng et al., 2022); generative adversarial network GAN (Li et al., 2019).

Since the core principle and operating characteristics of these machine learning could vary, it is necessary to select the most appropriate technique to build lithology identification models that can have a better output in a certain study area. PSRF which combines the mean-shift algorithm and the RF algorithm has been used for borehole lithology discrimination in a prototype well (Ao et al., 2019). It has been successfully applied to identify lithology in areas of shale-sand or glutenite formations, with higher flexibility and better results than other mainstream algorithms such as: LDA, ANN, SVM, DT, RF, PSANN, PSSVM (Ao et al., 2019). Similarly, BPNN is trained via backpropagating errors is a multilayer feedforward neural network algorithm and uses the gradient descent method to calculate the minimum value of the objective function (Rumelhart et al., 1986) showed an excellent classification and multi-dimensional function mapping ability to solve the lithology identification problem of conventional logging in Tarangaole uranium deposit (Rogers et al., 1992; Sun et al., 2022). XGBoost is an implementation of gradient boosting that supports both classification and regression predictive modeling with speed execution, model performance, and scalability (Chen and Guestrin, 2016; Al-Mudhafar and Wood, 2022). An integrated workflow using boosting machine learning (ML) algorithms for lithofacies classification was conducted for the carbonate reservoir of the large Majnoon oil and gas field (Iraq) (Al-Mudhafar et al., 2022). These results indicate that the XGBoost algorithm provided more accurate lithofacies classification than the other three boosting algorithms with respect to the predictions generated for both the entire dataset and testing subset (Al-Mudhafar et al., 2022). LapSVM is an improved semi-supervised classification method based on SVM. It inherits SVM's capability of handling nonlinear problems and improves the accuracy of the prediction by taking advantage of unlabelled data (Dong et al., 2020). LapSVM performed significantly better than SVM for lithology identification in Jiyang depression, Bohai Bay Basin (Li et al., 2020). Feature-Depth Smoothness based Semi-Supervised Weighted Extreme Learning Machine (FD-S2WELM) is a modified ELM method (Li et al., 2021b). Upon considering the unbalanced distribution of subsurface lithologies, the weight matrix is introduced to alleviate the class imbalance problem. Both the feature similarity and depth similarity matrix are utilized to constitute the graph Laplacians, and thus increasing the safety of the introduction of smoothness regularization (Li et al., 2021b). ISSCM is a novel semi-supervised learning algorithm developed based on the decision tree, the interpretability of which is highly beneficial to solve risk-aware problems (Li et al., 2021c). Both smoothness in the feature space and depth is utilized to generate pseudo-labels for the unlabelled data by using label propagation (Li et al., 2021c).

Sedimentation is a continuous process which creates a smooth transition of lithologies between two adjacent layers while layers with similar logging response characteristics also have similar lithology. Therefore, in the process of lithology identification, the samples relationship which is established according to the adjacent layers and logging curve similarity should be used as part of input data for model training. Moreover, the vast majority of the studies listed above solely used feature extraction from logging curves as input data for model training rather than considering such relationship which could contain sedimentation process and similarity of curve response through graph construction based on operator intention. Thus, Graph Neural Network (GNN) can solve the problems mentioned above and has been used to identify the lithology of target layers in several petroleum-bearing basins in China with high identification accuracy (Yuan et al., 2022). In this paper, a form of GNN known as GraphSAGE can become useful which has the following advantages over the commonly used machine learning methods: (1) GraphSAGE is a framework for inductive representation learning on large graphs (Hamilton et al., 2017). The input data of the algorithm is the processed graph which belongs to a data structure, and contains not only the logging curve values, but also the relationship connections between the responses. Therefore, model training process would be a combination of the curve features based on the relationship between depth locations with the logging value. (2)

GraphSAGE belongs to a semi-supervised classification method. This means the aggregation and update function in GraphSAGE can be employed repeatedly at each node and edge, which enables GraphSAGE to automatically support a combinatorial of generalization in the training process and reason in the system that was never been observed previously (Battaglia et al., 2018). Therefore, GraphSAGE is better suited to solving lithology identification problems in the wells with limited cores and training data. (3) Since the final output can be constructed and adjusted according to the training task and operator intention, higher learning efficiency in lithology identification would be expected in the training step. Compared with previous work about lithology identification based on GNN, the superiority of this identification process is to consider both depth and feature similarity in the process of graph construction and analyze the relationship between the identification accuracy and the thickness of single lithologic layer.

Considering the advantages of this model, in this paper, the GraphSAGE is used to identify lithology of the Fengcheng Formation in the Mahu Sag, and the results are compared with commonly used machine learning methods such as SVM and RF. We explained model building step by step, to ultimately predict the logging response curve which can provide a theoretical geological basis for the exploration and development of shale oil reservoirs in Fengcheng Formation.

2. Geological settings

Mahu sag, located in the northwest edge of Junggar Basin, is a secondary tectonic unit (Fig. 1a). From north to south, Wuxia Fault Zone, Kebai Fault Zone and Zhongguai Uplift are located in the west of the sag. From north to south, Quartz Beach Uplift, Yingxi Sag, Sangequan Uplift, Xiayan Uplift and Dabasong Uplift are located in the east of the sag (Zhi et al., 2021; Wang et al., 2022a) (Fig. 1b). Affected by the strong collision and compression movement of the Western Junggar Ocean towards the Kazakhstan Plate, especially in the Late Carboniferous to Early Permian, the collision between the Junggar Block and the Kazakhstan Plate intensified, forming a large nappe structure on the northwest edge of the basin, which is the most important formation period of the Mahu Sag (Lei et al., 2017). The deposition period of Fengcheng Formation is the development period of the Western foreland basin system, forming the most important set of source rocks in the basin (Zhi et al., 2021). At present, Baili oil areas of Kebai-Wuxia and west slope of the Mahu are well-known large oil areas in the world, with total proved reserves of 17.9 × 10^8t (Tang et al., 2019, 2021; Wang et al., 2022a).

The Fengcheng Formation in the Mahu Sag is a multi-source mixed fine-grained sedimentary formation formed in a deep to semi-deep alkali lake (Zhi et al., 2019). There are endogenous chemical deposits caused by arid and hot evaporation environment, volcanic materials provided by peripheral volcanic activities during the development of foreland basin, and the terrigenous clastic supply of fan delta formed by the denudation of the nappe on the western edge (Zhi et al., 2021). The Fengcheng Formation is divided from bottom to top into P_1f_1, P_1f_2 and P_1f_3, with obvious differences in sediments at different depths. P_1f_1 has a high content of volcanic material, with tuffaceous fine sandstone, basalt, ignimbrite in the lower part and organic mudstone and dolomitic rocks in the lacustrine period in the upper part (Fig. 1c). P_1f_2 was deposited in a strong evaporation environment, with high water salinity, limited exogenous input (Zhu et al., 2017). The lithology composition is mainly dolomite, dolomitic mudstone, mudstone and siltstone (Fig. 1c). In P_1f_3 deposition period, with the increase of exogenous input, the salinity decreased, and the sediment was similar to the top of the P_1f_1 (Zhi et al., 2021). The lithology composition is mainly mudstone, silty mudstone and dolomitic mudstone (Fig. 1c).

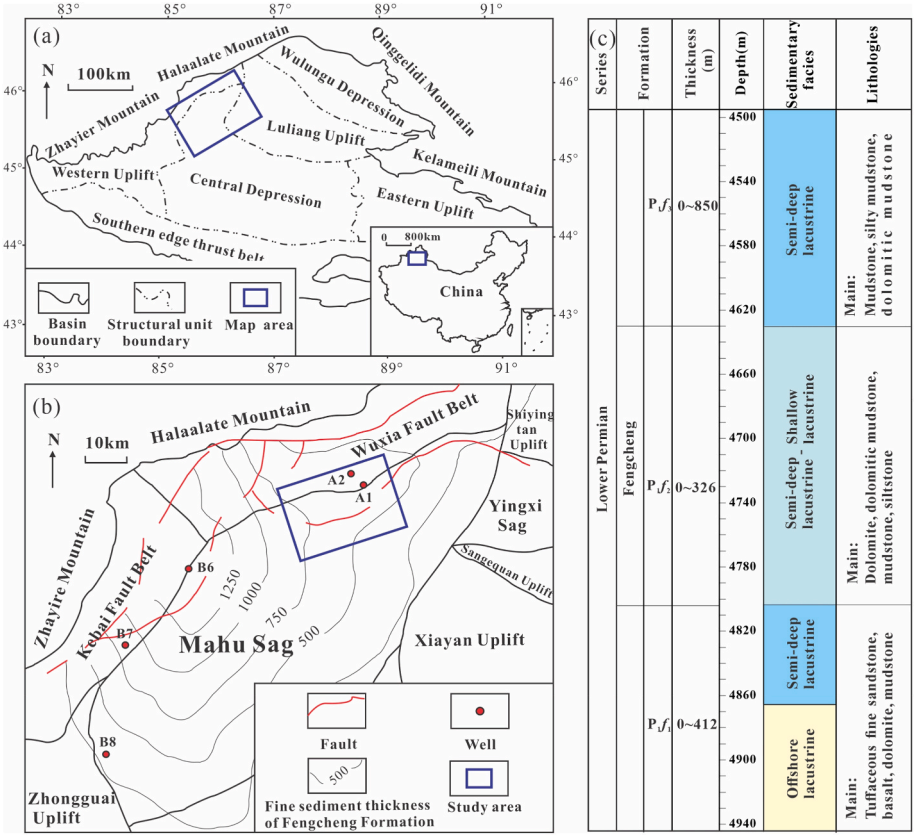

Fig. 1. Geologicl mp of Mahu Sag of the Junggar Basin (modified from Zhi et al., 2021). (a) Location of the Mahu Sag and structural units of Junggar Basin. (b) Location of the study area and the structural unit of the Mahu Sag. (c) Stratigraphic column of the Fengcheng Formation.

3. Principle of GraphSAGE for lithology identification

3.1. Basic idea of the lithology identification method based on GraphSAGE

In graph theory, a graph is a structure amounting to a set of objects in which some pairs of the objects are in some sense "related". The objects correspond to mathematical abstractions called nodes and each of the related pairs of nodes is called an edge (West, 2001). A graph is recorded as:

$$G = (V, E) \tag{1}$$

In the formula, V is the node representing the research object, and E is the edge representing the specific relationship between the objects. Graphs are divided into directed and undirected graphs. The number of edges connecting a node in the graph is called the degree of the node. The number of directed edges starting from the node is called the out-degree of the node, and the number of directed edges ending at the node is called the in-degree of the node. The degree of a node is the sum of the in-degree and out-degrees. In certain cases, each node has a feature vector, and the feature vectors of all nodes constitute the feature matrix of the graph. Therefore, the graph can also be divided into graph structure and feature matrix (Fig. 2).

Graphs have a wide range of applications in chemistry, communication networks, social networks, etc. The relationships between nodes in these systems are frequently deterministic (Zhou et al., 2020). For example, the atoms in a molecule serve as nodes, and the chemical bonds between atoms serve as edges. In the communication network, people serve as nodes, and the connections between people represents the edges. The chemical bonds between atoms and the connection between people exist objectively, and the corresponding graph structure is also determined. In conventional logging for lithology identification, if the sampling depth, the interval between two data points is 0.125 m, is taken as node, there wont be any clear relationship between the nodes, such as what we consider in chemical bonds between atoms or connections between people. Therefore, it is crucial to establish a graph structure suitable for lithology identification.

The method of constructing graph used in this paper is mainly based on the following two points: (1) The structure of the graph should be based on the general sedimentation of the stratum and should have certain geological significance. (2) Nodes with the same lithology should have connections.

Based on these two points, the process of graph construction is comprised of two steps: First, the nodes are sorted based on depth sequence, and the pairs of adjacent depth nodes are connected by edges (Fig. 3b). Due to the continuous process of sedimentation, each node at a certain depth has lithological similarity with the spatial adjacent nodes. Therefore, not only the constructed graph reflects the vertical spatial distribution of the actual stratum, but also the nodes with the same lithologic label at adjacent depths would be connected. Second, cluster the nodes with non-adjacent depth using affinity propagation (AP) algorithm, to calculate the Euclidean distance between the nodes in each group based on the normalized logging curves. To do so, we select the appropriate quantile by determining the relationship between the nodes according to the Euclidean distance from small to large (Fig. 3b). The process forms a connection between these nodes with similar logging curve characteristics but not adjacent in depth while these connected nodes could have the same lithology labels.

The training process of lithology identification is as follows: after inputting the created graph to the model, the first GraphSAGE layer is used to construct the embedding matrix H_1 integrating both the nodes' initial feature vector which is well logs standardized and the information about the local graph structure that surrounds them. Then, the embedding matrix H_1 is activated by the Leky RELU function and used as the input data of the next GraphSAGE layer for the second information integration to obtain the embedding matrix H_2. In the process of updating nodes representation after n times, the final classification result is obtained by calculating the embedding matrix H_n with Softmax layer (Fig. 3c).

3.2. Mathematical algorithms in the lithology identification method

3.2.1. Affinity propagation

Affinity propagation (AP) is a clustering method based on the concept of "message passing" between data points (Frey et al., 2007). The basic concept of this method is to regard all samples as nodes in the network, and then calculate the clustering center of each sample through the message passing of each edge in the network (Frey et al., 2007; Ortiz-Bejar et al., 2022; Sajid et al., 2022). Both AP and k-means classify groups by defining distances. However, since AP used the idea of message passing, it could get a better total square error than K-means, and the clustering ability is also greatly improved, especially suitable for clustering of high dimensional and multi types data. In the process of AP clustering, there are two kinds of message passing between nodes,

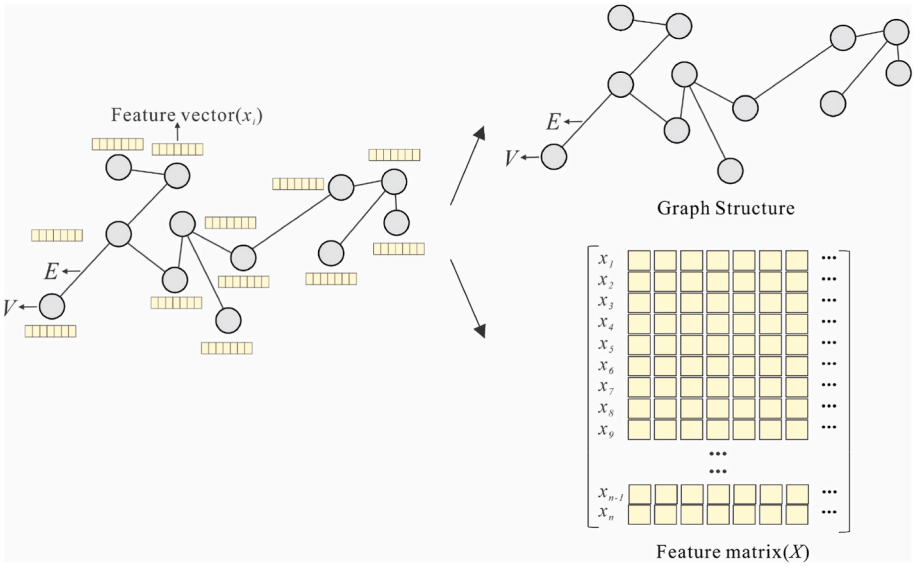

Fig. 2. Composition of graph.

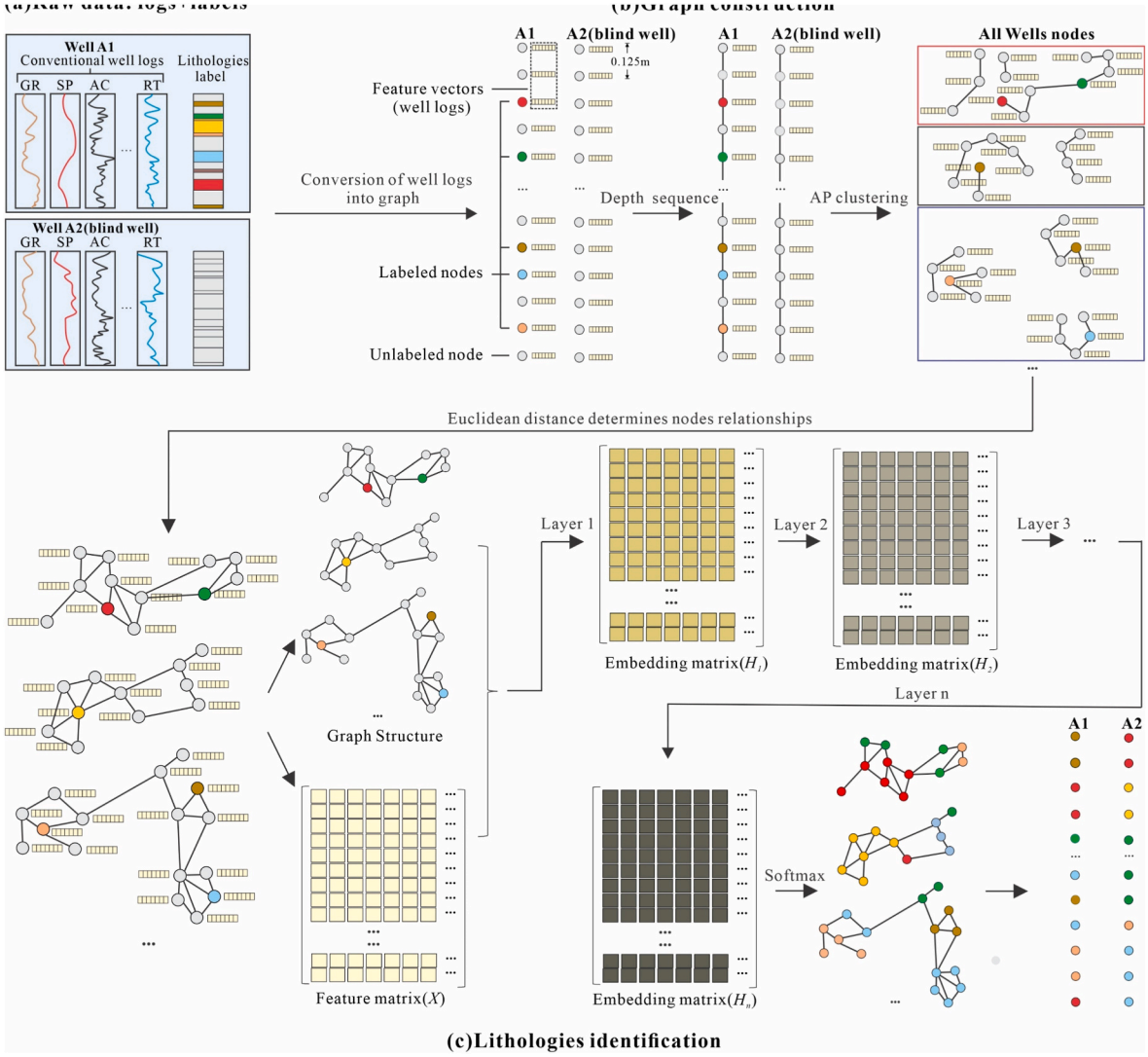

Fig. 3. The process of the GraphSAGE for lithology identification.

responsibility and availability. Through multiple iterations, the responsibility and availability of nodes are constantly being updated until multiple exemplars are generated, and the remaining nodes are assigned to the corresponding groups.

AP has the following advantages: (1) There is no need to specify the number of clusters. It makes the prior experience become an unnecessary condition and expands the scope of application. (2) It is insensitive to the initial value and does not need to select the random initial value. (3) It has low total square error.

3.2.2. GraphSAGE

The essence of lithology identification is to classify the nodes in the graph. GraphSAGE is a batch training algorithm for graph convolution (Hamilton et al., 2017). It samples the neighboring nodes of each root node, extends the adjacent nodes of the root nodes by K steps, and updates the representation of the root nodes by aggregating the hidden nodes representations hierarchically from the k-hop neighbor nodes to the root nodes. This process is called node embedding (Hamilton et al., 2017). The basic idea of node embedding is to extract the high-dimensional information of the neighbors of the root nodes into dense vector embeddings using dimension reduction method (Hamilton et al., 2017). Then these node embeddings are fed to the downstream machine learning system, and help complete the tasks of node classification, clustering and link prediction. The specific embedding process could be expressed by the following formulas:

$$h^k_{\mathcal{N}(v)} = \mathrm{AGGREGATE}_k\left(\left\{h^{k-1}_u, \forall u \in \mathcal{N}(v)\right\}\right) \tag{2}$$

$$h^k_v = \sigma\left(\mathbf{W}^k \cdot \mathrm{CONCAT}\left(h^{k-1}_v, h^k_{\mathcal{N}(v)}\right)\right) \tag{3}$$

Formula (2) represents the information aggregation of neighboring nodes. Formula (3) represents the root nodes information update representation and $h^k_{\mathcal{N}(v)}$ represents aggregated vectors. h^{k-1}_u represents a neighbor node of a root node. $\mathcal{N}(v)$ represents the neighbor nodes of root nodes, h^{k-1}_v is a root node's current representation. $\mathbf{W}^k$ represents the weight matrix of the k-th layer. σ is the activation function. CONCAT (·)is concatenation function. AGGREGATE (·)is the aggregating function. Aggregating function includes element-wise mean, long short-term memory (LSTM) and pooling. This paper adopts element-wise mean.

The above formulas contains the concept of message passing divided into aggregation and update of nodes feature vectors. Take lithologic identification as an example to explain these two processes (Fig. 4). Aggregation of feature vectors is that each node collects other feature vectors from its neighbors via a permutation equivariant function (Hamilton et al., 2017; Chen et al., 2020). Assuming that node "0″ is a root node, nodes "1″ and "2″ are the one-hop neighbors of node "0″ in Fig. 4. The process of the first layer training is separated into two steps. The first step is that the feature vectors x_1, x_2 corresponding to the "1″

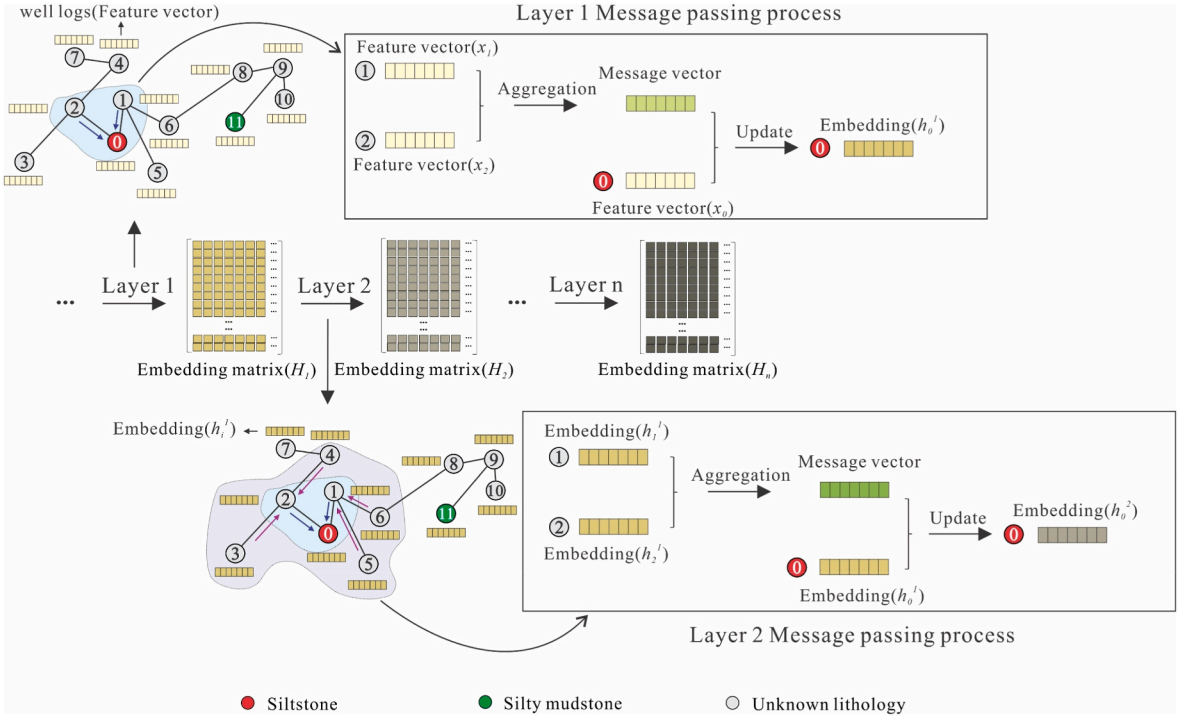

Fig. 4. Message passing process in GraphSAGE algorithm.

and "2″ nodes aggregated through the element-wise mean aggregation function to obtain the Message vector. The second step is the update function, where the root node "0″ combines the Message vector with its own feature vector to construct a new vector h_0^1. The embeddings obtained after the feature update of all nodes are combined to the embedding matrix H_1 (Fig. 4). After being activated by the activation function, the embedded matrix H_1 is input into the second layer. The training process is similar to the first layer. The embeddings h_1^1, h_2^1 corresponding to the two-hop neighbor nodes "1″ and "2″ are aggregated to obtain Message vector, which is combined with the h_0^1 to update representation of the node "0". Because h_1^1 and h_2^1 contain the information of nodes "5″, "6″ and "3″, "4″ respectively, h_0^2 contains not only the information of one-hop neighbors "1″, "2″, but also two-hop neighbors "5″, "6″, "3″, "4" (Fig. 4). Setting n layers is equivalent to aggregating the information of n-hop neighbor nodes.

Because GraphSAGE belongs to semi-supervised learning, it works on the following two conditions: all data (the labelled and the unlabelled data) are sampled from the same marginal distribution and two closer samples are more likely to fall in the same class. The latter condition is well known as the smoothness (or manifold) assumption (Chang et al., 2020; Liu et al., 2020c; Lv et al., 2020; Wu et al., 2021).

According to the message passage concept and assumptions mentioned above, the training results of the two nodes connected in the graph affect each other and tend to be divided into the same lithology. The logging data that need to be predicted also need to be involved in the construction of the graph.

3.3. Evaluation metrics of lithology identification model

In order to evaluate the validity of the model, it is necessary to establish a unified evaluation index in order to evaluate the lithology identification ability of the model. For the multi-classification problem of lithology identification, the confusion matrix is used to evaluate the results of model identification, including TP (true positive), FP (false positive), FN (false negative) and TN (true negative), which respectively represent the number of positive samples correctly identified, the number of negative samples incorrectly identified as positive samples, the number of positive samples incorrectly identified as negative samples and the number of negative samples correctly identified (Table 1) (Bressan et al., 2020). Taking mudstone identification as an example, mudstone is a positive sample, and these four parameters respectively represent the identification of mudstone samples as mudstone, non-mudstone samples as mudstone, mudstone samples as non-mudstone, and non-mudstone samples as the corresponding correct lithology.

Accuracy (Ac) is defined as the ratio of the number of correctly classified samples to the total number of samples (Bressan et al., 2020; Dong et al., 2022):

$$Ac = (TP + TN) / (TP + TN + FP + FN) \quad (4)$$

For the lithology identification model, the accuracy can only represent the overall classification ability of the classifier for all samples, but cannot reflect the classification ability of the classifier for each lithology.

Precision (Pr) defines the ratio of the number of correctly identified positive samples in a class of samples to the total number of correctly identified positive samples and negative samples incorrectly identified as positive samples (Bressan et al., 2020; Dong et al., 2022):

$$P_r = TP / (TP + FP) \quad (5)$$

The higher the accuracy, the less samples of other lithologies are misidentified as mudstones.

Recall (Re) defines the ratio of the number of correctly identified positive samples in a class of samples to the total number of correctly identified positive samples and positive samples incorrectly identified as negative samples (Bressan et al., 2020; Dong et al., 2022):

Table 1
Confusion matrix for multi-classification data.

	Results of identification	
	Positive type	Negative class
positive type	*TP*	*FN*
Negative class	*FP*	*TN*

$$R_e = TP / (TP + FN) \quad (6)$$

The higher the recall, the less samples of mudstone are misidentified as other lithologies.

3.4. Workflow of lithology identification

According to the graph structure construction process and GraphSAGE principle, the workflow of lithology identification using GraphSAGE is summarized (Fig. 5). First of all, the conventional well logs marked by the available cores are divided the training data and testing data randomly according to the ratio of 7:3 after standardization. In the second place, based on the depth sequence and AP clustering method, the graph structure is constructed, and the standardized logging curve is used as the feature matrix. Finally, GraphSAGE is used to train the constructed graph. After 5000 iterations, the trained lithology identification model is applied to the testing data and the blind well to examine the reliability of the model.

4. Lithology identification in continental shale reservoirs of Junggar Basin, China

4.1. Conventional well logs responses of lithologies

Based on the observation and description results of the core and thin sections, the lithologies of Fengcheng Formation in the study area are divided into 10 different types including: basalt (Fig. 6a, Fig. 6m and n), ignimbrite (Fig. 6b), tuffaceous fine sandstone (Fig. 6c), sedimentary tuff (Fig. 6d), dolomite (Fig. 6f, k, Fig. 6l), dolomitic mudstone (Fig. 6e), mudstone (Fig. 6h), silty mudstone (Fig. 6o and p), siltstone (Fig. 6g) and silicalite (Fig. 6i and j). The content of brittle minerals in dolomite, dolomitic mudstone and siltstone of Fengcheng Formation is high, and natural fractures are widely developed under multi-stage structural deformation (Fig. 6e and f, Fig. 6g). As the matrix porosity and permeability of shale oil reservoir are generally low, the development of natural fractures can effectively improve the reservoir quality and oil production capacity (Zeng, 2010; Zeng et al., 2016; Liu et al., 2020a; Liu et al., 2020b). The reservoirs of these lithologies with high content of brittle minerals and relatively developed structural fractures could be used as favorable reservoirs in the study area.

Because the well log response characteristics of various lithologies

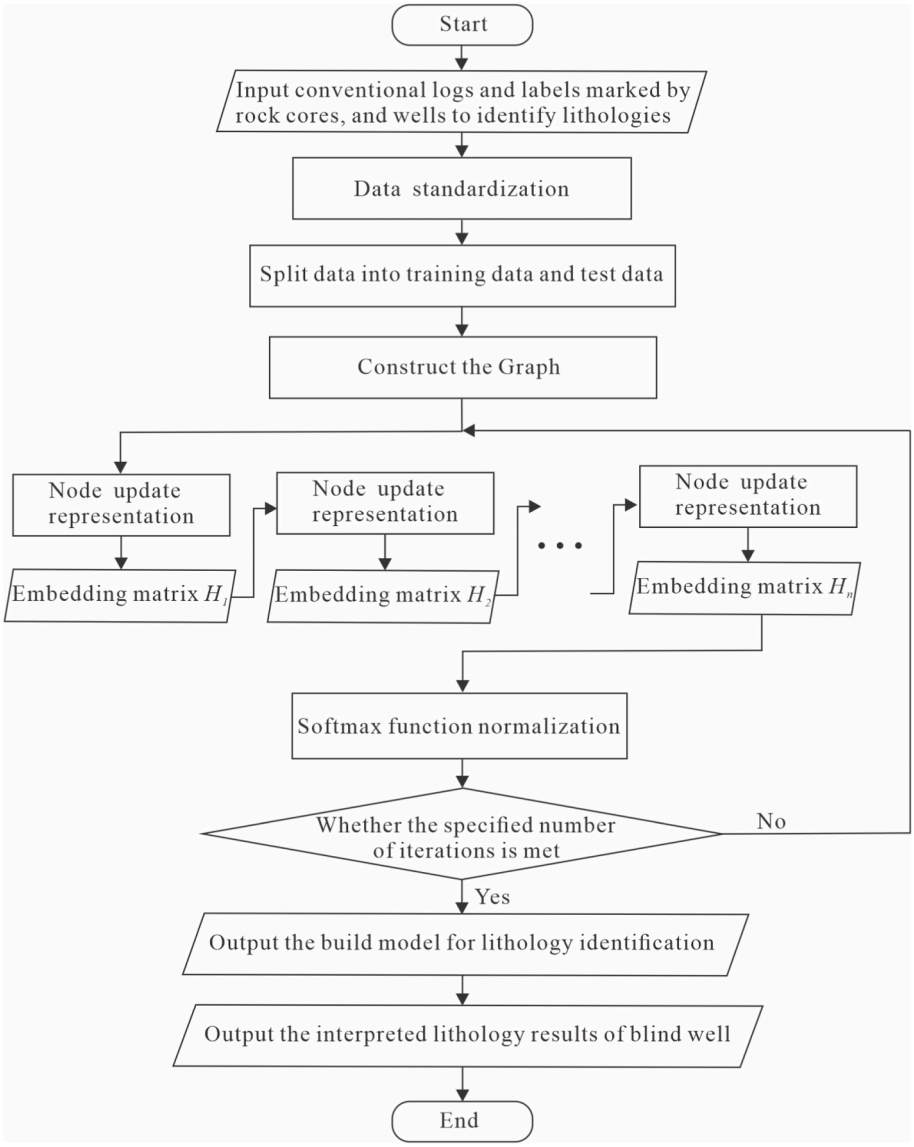

Fig. 5. Schematic flow chart of lithology identification based on the GraphSAGE algorithm.

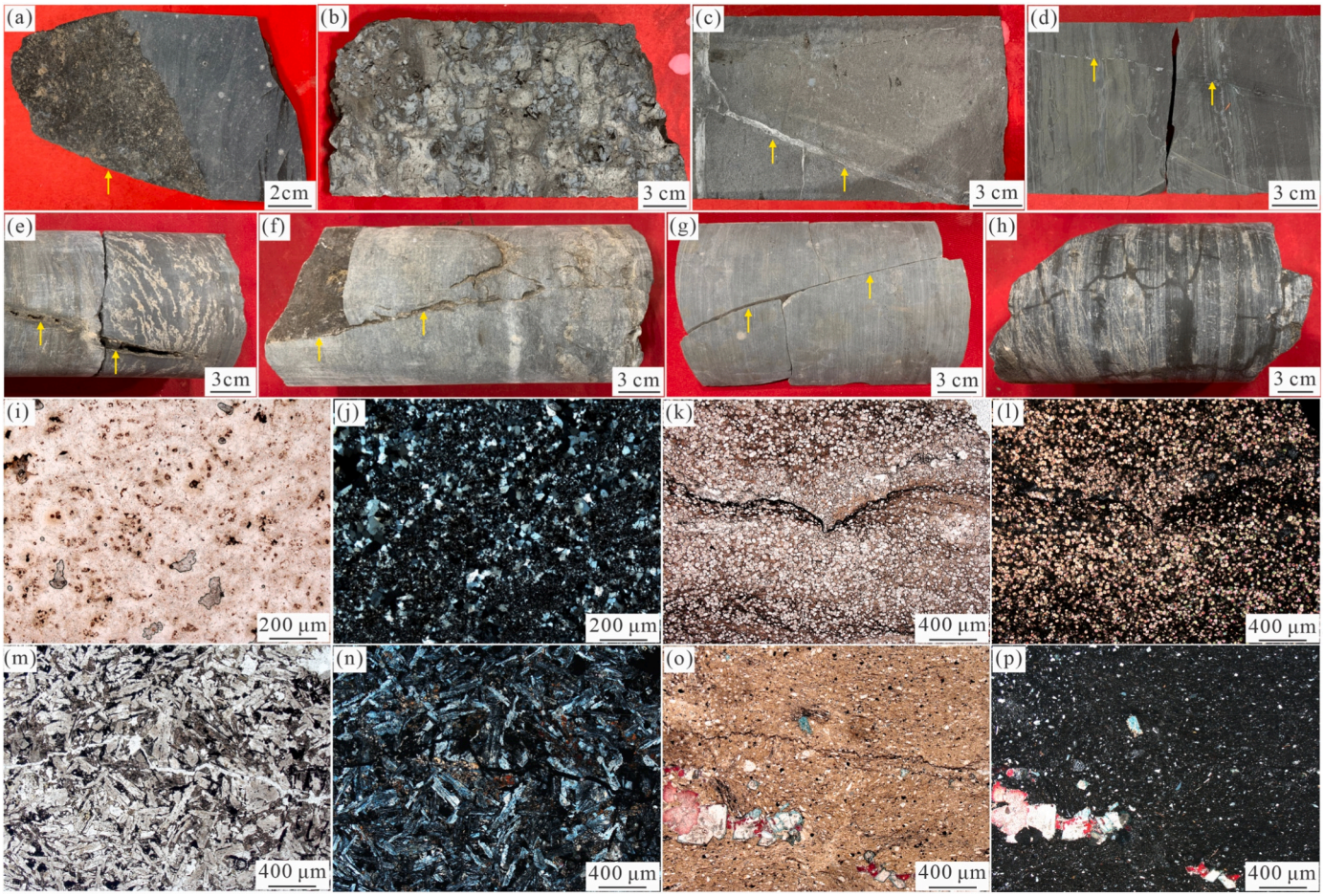

Fig. 6. Photos of the core and thin sections. (a) Well A1, 4931.55 m, Basalt. (b) Well A1, 4901.34 m, Ignimbrite. (c) Well A1, 4891.87 m, Tuffaceous fine sandstone. (d) Well A1, 4876.07 m, Sedimentary tuff. (e) Well A2, 4575 m, Dolomitic mudstone. (f) Well A2, 4558.8 m, Dolomite. (g) Well A1, 4854.7 m, Siltstone. (h) Well A2, 4524 m, Mudstone. (i) Well A1, 4787.7 m, plane-polarized light, Silicalite. (j) Well A1, 4787.7 m, orthogonal-polarized light, Silicalite. (k) Well A1, 4789.7 m, plane-polarized light, Dolomite. (l) Well A1, 4789.7 m, orthogonal-polarized light, Dolomite. (m) Well A1, 4868.76 m, plane-polarized light, Basalt. (n) Well A1, 4868.76 m, orthogonal-polarized light, Basalt. (o) Well A1, 4799.18 m, plane-polarized light, Silty mudstone. (p) Well A1, 4799.18 m, orthogonal-polarized light, Silty mudstone.

are different due to their specific physical properties, theoretically, the amplitude change of multiple logging curves can separate different lithologies (Dong et al., 2022). For example, GR is a record of radioactivity formed by radioactive elements (uranium, thorium, and potassium) (Dong et al., 2020). The higher GR value of mudstone is due to the adsorption of radioactive elements on the charged surface of clay minerals, which can be used to distinguish mudstone from other lithologies. AC records the time required for an acoustic wave to transverse a given length of the formation (Dong et al., 2020). Typically, the reference AC values for dolomite, limestone, and sandstone are 43.5, 47.6, and 55.0–51.5 μs/ft (Dong et al., 2022). AC can be used to distinguish sandstone from other lithologies. DEN reflects the bulk density of formation based on the physical phenomena of gamma ray scattering and absorption (Dong et al., 2020). The DEN of sandstone is lower than limestone and dolomite, and the DEN of mudstone varies with the degree of compaction of the mudstone (Dong et al., 2022). CAL can indirectly distinguish lithologies by measuring the size and shape of the borehole (Dong et al., 2020). Brittle mudstone increases CAL due to caving while high permeability sandstone decreases CAL values due to the mud cakes.

In practice, the response characteristics of logging curves are not only affected by lithologies, but also by geological and drilling operations such as clay minerals, pore fluids, and slurry pressure, making the lithology response more complex. The cross plot of well logs in the study area, which is composed of 100 samples randomly selected from each lithology, is shown in Fig. 7. The probability density curves of different lithologies are on the right and above side of the cross plot. The overlap and separation of curve peaks reflect the sensitivity of well logs for different lithologies (Dong et al., 2022). The AC range of Fengcheng Formation in the study area is 49.52–82.93 μs/ft, with an average of 59.55 μs/ft. CNL ranges from 0.015 to 0.38 m^3/m^3, with an average of 0.12 m^3/m^3 (Table 2). As displayed in Fig. 7a, there are four peaks in the CNL probability density curve, which can distinguish basalt, ignimbrite and tuffaceous fine sandstone from other lithologies. There are two peaks in the AC probability density curve, and the AC value of ignimbrite and tuffaceous fine sandstone is higher than that of other lithologies. The CAL range in the study area is 8.34–10.2 inch, with an average of 8.54 inch. DEN ranges from 2.31 to 2.81 g/cm^3, with an average of 2.63 g/cm^3 (Table 2). According to Fig. 7b, DEN probability density curve has four peaks, which can distinguish basalt, ignimbrite, tuff fine sandstone and other lithologies; however, there are considerable overlaps between the CAL values of each lithology, which cannot be well distinguished. The RT range in the study area is 1.15–6257.48 Ω m, with an average of 180.74 Ω m; RI range is 1.16–3262.9 Ω m, with an average of 143.25 Ω m. According to Fig. 7c, there are three clear peaks in the probability density curves of RT and RI, representing basalt, ignimbrite and tuffaceous fine sandstone respectively, but the overlapping area of other lithologies is still large. GR in the study area ranges from 37.74 to 209.09 API, with an average of 101.54 API; RXO is distributed in 0.14–348.51 Ω m, with an average of 47.98 Ω m (Table 2).

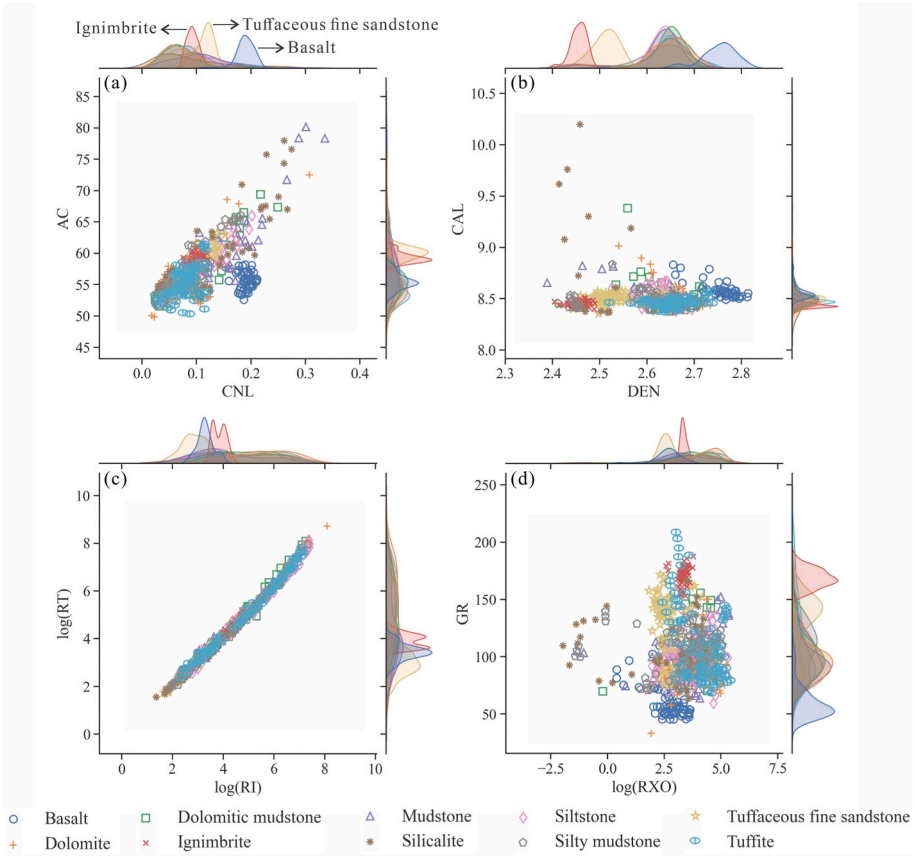

Fig. 7. Cross plots of well logs.

Table 2
Statistics of conventional well logging data in the study area.

Variables	GR	SP	CAL	AC	DEN	CNL	RI	RT	RXO
Units	API	mV	inch	us/ft	g/cm^3	m^3/m^3	Ω.m	Ω.m	Ω.m
Mean	101.54	−54.45	8.54	59.55	2.63	0.12	143.25	180.74	47.98
Deviation	27.13	13.14	0.29	6.29	0.06	0.07	251.65	385.60	45.59
Variance	735.89	172.71	0.04	39.62	0.003	0.005	63325.81	148687.3	2078.09
Minimum	37.74	−77.84	8.34	49.52	2.31	0.015	1.16	1.15	0.14
Maximum	209.09	−15.09	10.2	82.93	2.81	0.38	3262.9	6257.48	348.51
Skewness	0.73	0.54	2.26	1.37	−1.21	1.22	3.61	5.12	1.43
Kurtosis	0.42	−0.32	6.2	1.06	2.52	0.73	19.66	41.66	1.93

According to Fig. 7d, GR and RXO can distinguish between ignimbrite and basalt but there are considerable overlaps between other lithologies. Generally speaking, by cross plotting well logging curves, only three to four types of lithologies can generally be recognized in the study area. As a result of considerable overlap between other lithologies, lithology identification in the study area is a complex problem of nonlinear classification nature.

Based on the analysis of cross plot of well logs, a total of 9 logging curves including AC, DEN, CNL, GR, RT, RI, RXO, SP and CAL are selected as the input for building lithology identification models.

4.2. Lithology identification by GraphSAGE

The coring section of Well A1 is done from depth 4577–4930.375 m, which covers P_1f_3, P_1f_2 and P_1f_1. There are 2746 depth sampling points in total that are measured with the well logs, with an interval of 0.125 m between adjacent sampling points. The training and testing process of the model adopts the Random Subsampling where dataset is sampled and split into two parts: training for modeling and testing for prediction (Wang et al., 2014; Al-Mudhafar, 2016). 70% of these data from well A1 are randomly selected as the training data, and the remaining 30% are used as the testing data check the reliability of the model. To further examine the accuracy of the lithology identification models, another well, Well A2, was used as the blind test, and its data was not used for the model training. All the data are standardized, and the lithology description results of cores and thin sections are used as labels. Dolomite, dolomitic mudstone, mudstone, silicalite, siltstone, silty mudstone, sedimentary tuff, ignimbrite, tuffaceous fine sandstone and basalt are represented by 0, 1, 2, 3, 4, 5, 6, 7, 8 and 9, respectively.

Quantile refers to the value points that divide the probability distribution range of a random variable into several equal parts, commonly used are median, quartile, percentile, etc. In the process of graph construction, after AP clustering of the conventional logging data, the relationship between the two nodes is determined according to the Euclidean distance of the nodes in each class. When the Euclidean distance between the nodes is less than the set quantile, the connection is established. In this paper, the quantile is set to be 3%, and the constructed graph of lithology identification model is shown in Fig. 8. The layout of graph visualization is obtained by Fruchterman-Reingold algorithm (Fruchterman et al., 1991; Gajdoš et al., 2016) which is based

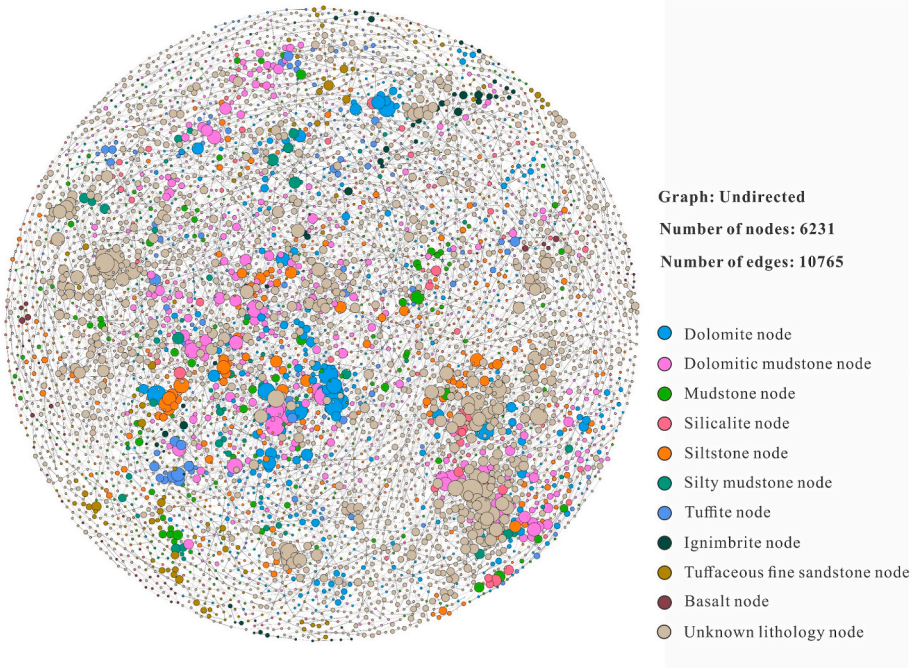

Fig. 8. The input graph visualization of lithology identification model.

on particle physics theory. This algorithm simulates the nodes in the graph as atoms and calculates the positional relationship between nodes by simulating the force field between the atoms. The input graph contains 6231 nodes in total, 2746 nodes in Well A1 and 3485 nodes in Well A2. Moreover, the graph is comprised of 10,765 undirected edges, 6229 edges of the pairs of adjacent depth nodes and 4536 edges of Euclidean distance less than the set quantile of the pairs except adjacent depth nodes after AP clustering (Fig. 8). The nodes in the graph represent different lithologies and the diameter of the nodes reflect the degree of the nodes. The connections of the paired nodes would represent the edge. The larger the degree of each node will result in larger diameter and higher number of edges as shown in Fig. 8.

The statistical data of the degree of each node in the constructed graph is shown in Fig. 9. The node ID in the graph ranges from 0 to 2745 corresponding to the order of Well A1 nodes from shallow to deep. The node ID ranges from 2745 to 6230 corresponding to the order of blind Well A2 nodes from shallow to deep. The graph shows that the minimum degree of the node is 2, and the maximum degree is 14, indicating that the node has at least 2 edges and at most 14 edges (Fig. 9a). Since the constructed graph data is an undirected graph, the in-degree of all nodes is equal to the out-degree. As the counts of degree increases, the number of corresponding nodes decreases (Fig. 9b).

Other model hyper-parameters are set as follows: the number of layers is 4, the neighbor nodes are fully sampled and he learning rate is 0.0005. The Aggregator is selected as mean aggregator and the activation function is selected as Leaky RELU. Finally, the loss function is selected as cross-entropy (Table 3) and the number of iterations is 5000 times.

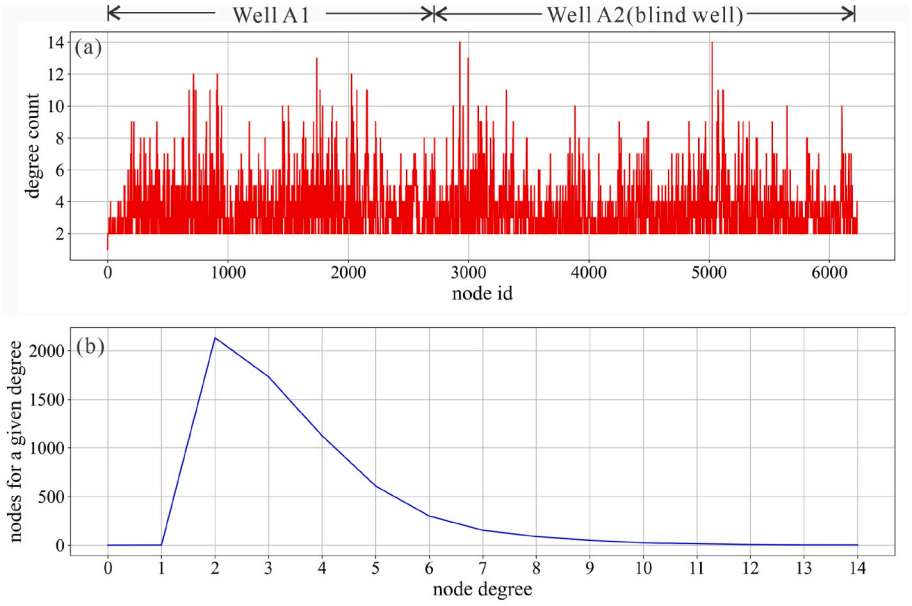

Fig. 9. Node degree statistics of the graph.

Table 3
Model hyper-parameters setting.

Hyper-parameters	Quantile	Number of Layers	Sample sizes	Aggregator	Loss function	Activation function	Learning rate
Parameter setting	3%	4	Full sampling	Mean aggregator	Cross- entropy	Leaky RELU	0.0005

The code program is developed in Python, and the computer hardware used is AMD 5900 × 12-core 24-thread CPU, 64G memory, and NVIDIA RTX3090 24G video memory GPU.

4.3. Comparison of lithology identification between GraphSAGE and other methods

Using the trained model to identify the lithologies of the testing data, the confusion matrix as shown in Fig. 10a is obtained. The red area of the confusion matrix represents the number of wrongly identified samples and the ratio of the number of wrongly identified samples to the total number of samples in the testing data. The green area represents the number of correctly identified samples and the ratio of the number of correctly identified samples to the total number of samples in the testing data. The bottom row colored by light gray represents the precision (P_r) for each lithology. The right-most column colored by light gray represents the recall (R_e) for each lithology. The dark gray box in bottom right represents the accuracy (A_c) of all existing lithologies. According to Fig. 10a, the accuracy of the testing data is 90.41%. Because the logging curve response characteristics of basalt ("9″) and ignimbrite ("7″) are notanly different from other lithologies, the precision and recall of these two lithologies are 100%. The precision and recall of tuffaceous fine

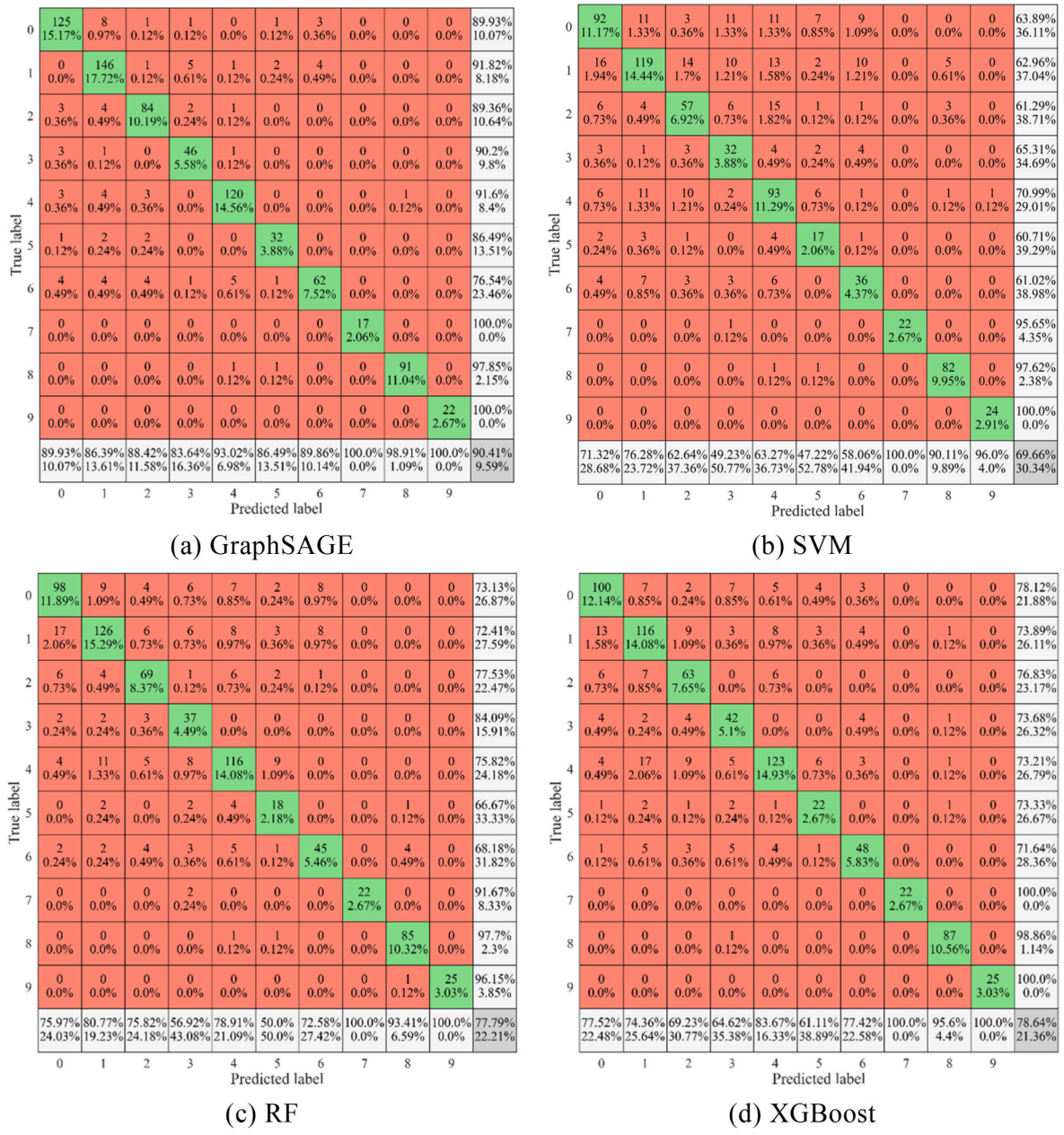

Fig. 10. Confusion matrix of four methods testing data of Well A1 for lithology identification. (Dolomite, dolomitic mudstone, mudstone, silicalite, siltstone, silty mudstone, sedimentary tuff, ignimbrite, tuffaceous fine sandstone and basalt are represented by 0, 1, 2, 3, 4, 5, 6, 7, 8 and 9.)

sandstone ("8″) are more than 98%. Moreover, the precision of silicalite ("3″) is 96%, and the recall is only 81.36%, indicating that few samples of other lithologies are wrongly identified as siliceous rock, but there are relatively many samples of siliceous rock wrongly identified as other lithologies, especially the largest number of samples wrongly identified as dolomitic mudstone. The precision and recall of dolomite ("0″), dolomitic mudstone ("1″), mudstone ("2″), siltstone ("4″), silty mudstone ("5″), sedimentary tuff ("6″) are generally more than 85%.

For comparison with GraphSAGE method, SVM, RF and XGBoost are selected as three commonly used machine learning methods for lithology identification. In the process of model training, these three methods use the grid search method to optimize the hyper-parameters (Table 4). The accuracies of SVM, RF and XGBoost are 69.66%, 77.79% and 78.64% (Fig. 10), which are lower than the identification results using GraphSAGE. However, the recall and precision of ignimbrite ("7″), tuffaceous fine sandstone ("8″) and basalt ("9″) using the three methods are high, basically reaching more than 90%, with a small gap with GraphSAGE. It shows that the logging response characteristics of these three lithologies are significantly different with other lithologies. In addition, the overlapping area of other lithologies in Fig. 7 is large, so it is difficult to identify the exact lithology. The precision and recall of SVM for these lithologies are found generally less than 77%, and the precision of silicalite ("3″) and silty mudstone ("5″) are only 49.23% and 47.22%, inferring that there are a large number of samples that were other lithologies but misidentified as these two types of lithologies. RF and XGBoost are mainly less than 80% in precision and recall, and less than 60% in identification precision of silicalite ("3″) and silty mudstone ("5″). The precision and recall of GraphSAGE are generally more than 85% for these lithologies that are difficult to identify correctly. For silicalite ("3″) and silty mudstone ("5″), the precision of GraphSAGE is 83.64% and 86.49%, and the recall is 90.2% and 86.49%.

In order to further test the developed four models, Well A2 is selected as the blind test which was not included in models training. The cored sections of Well A2 are 4517.12–4529.14 m and 4541.88–4576.76 m, belonging to P_1f_3, which lacks sedimentary tuff, tuffaceous fine sandstone, ignimbrite and basalt. The lithology identification models based on GraphSAGE, SVM, RF and XGB are used to identify the cored sections of Well A2. The identification accuracy is 75.2%, 48.11%, 55.61% and 57.6% respectively (Fig. 11). The interpretation of Well A2 is shown in Fig. 12. Ten intervals are used from this well to illustrate the identification results as follows: (1) GraphSAGE identify one part of dolomite between 4518.65 and 4519.28 m, another part is incorrectly identified. XGBoost, RF and SVM are entirely identified lithologies incorrectly. (2) The 4519.28–4520.52 m interval is correctly identified as silicalite by GraphSAGE, XGBoost, RF and misidentified by SVM. (3) Because the 4529.16–4541.78 m interval is not cored, its exact lithology is unknown but knowing that basalt is only developed in P_1f_1, results by XGBoost in 4532.03–4532.15 m and 4533.03–4533.15 m, RF in 4532.03–4532.15 m and 4532.90–4533.15 m, SVM in 4532.90–4533.15 m and 4535.15–4535.40 m cannot be accepted. GraphSAGE did not misidentify basalt in this section. (4) The 4542.15–4542.78 m interval is correctly identified as silicalite by these four methods. (5) In the 4543.65–4544.65 m interval the identification results obtained by GraphSAGE are consistent with the cores being silty mudstone but XGBoost, RF and SVM misclassified this interval as dolomite, tuffaceous fine sandstone, silicalite and siltstone. (6) Section 4547.40-4549.15 m, shows that the identification results of GraphSAGE are consistent with the cores that is observed to be silicalite however, XGBoost, RF and SVM misidentified this section as mudstone, siltstone and dolomitic mudstone. (7) In the 4558.53–4559.78 m section, GraphSAGE correctly identified dolomite to be present only between 4559.24 and 4559.78 m. XGBoost correctly recognized dolomite only between 4559.03 and 4559.15 m, RF only between 4559.03-4559.15 m and 4559.40–4559.53 m while SVM completely misidentified the section as other lithologies. (8) In 4567.78–4569.53 m, the identification results obtained by GraphSAGE is totally consistent with the cores while XGBoost, RF and SVM only accurately classified one part of dolomite in section. (9) In the 4569.53–4570.40 m interval, the identification results of GraphSAGE and SVM are exactly conforming the cores while XGBoost and RF incorrectly identified other lithologies. (10) Finally, in the 4572.15–4572.65 m section, the identification results of SVM are completely consistent with core observation results while GraphSAGE, XGBoost and RF accurately classified one part of dolomitic mudstone in the section.

In general, the identification results obtained from the GraphSAGE have the highest accuracy, recall, precision in the Well A1 testing data and matches well with the core observations in the blind test from the data of Well A2. The results obtained from GraphSAGE for dolomite and silicalite, in particular are more accurate than other three AI methods. Collectively, XGBoost has slightly higher accuracy than RF and SVM has the lowest accuracy of these four methods in the Well A1 testing dataset. The identification results obtained by XGBoost, RF and SVM all have a low matching with cores observations in the blind test data from Well A2, which reveals poor capability and general performance of these three other AI methods for lithology identification compared to the GraphSAGE.

Table 4
Search range and optimal setting of hyper-parameters of SVM, RF and XGBoost.

Method	hyper-parameters	Search range	optimal setting
SVM	Kernel function	RBF, Polynomial, Sigmoid	RBF
	Gamma	0–2	1
	C	0–25	13
RF	Max depth	0–40	25
	Min samples leaf	0–20	1
	Min samples split	0–10	2
XGBoost	Max depth	0–50	21
	Gamma	0–10	0
	Alpha	0.03–2	0.03
	Learning rate	0.001–0.5	0.15

5. Discussions

5.1. Influence of graph construction method on identification results

It was understood that GraphSAGE has better lithology identification ability than frequently-used ensemble learning and kernel methods. In this section, the reasons for the above results are analyzed in depth, and the influence of graph construction on the identification results is further analyzed. In this regard, relational inductive bias which refers to prioritizing a certain solution over other solutions through artificial preference at the beginning of model training is believed to play a significant role in our approach. This can be included both in the underlying data distribution assumptions and in the model design. Weak relational inductive bias corresponds to the low learning efficiency of data, which means that a large amount of data would be required to train models. Graph is a representation that supports arbitrary relational structure, hence, computations over graphs would also provide strong relational inductive bias, which exceeds the bias that is provided by full connection layers, convolution layers and recurrent layers (Battaglia et al., 2018).

Based on the vertical stratigraphic distribution of the Fengcheng Formation and the characteristics of the message passing process of GraphSAGE method, this paper constructs the lithology identification graph structure, which is equivalent to integrating these two preferences into the graph structure at the beginning of model training. However, the input data of SVM, RF and XGBoost is raw logging data and do not contain such structure features, which leads to lower learning efficiency and accuracy in lithology identification compared to GraphSAGE. In order to confirm that graph structure is the key factor affecting the results of lithology identification, three graph construction methods are designed based on the same training and testing data, which are not the

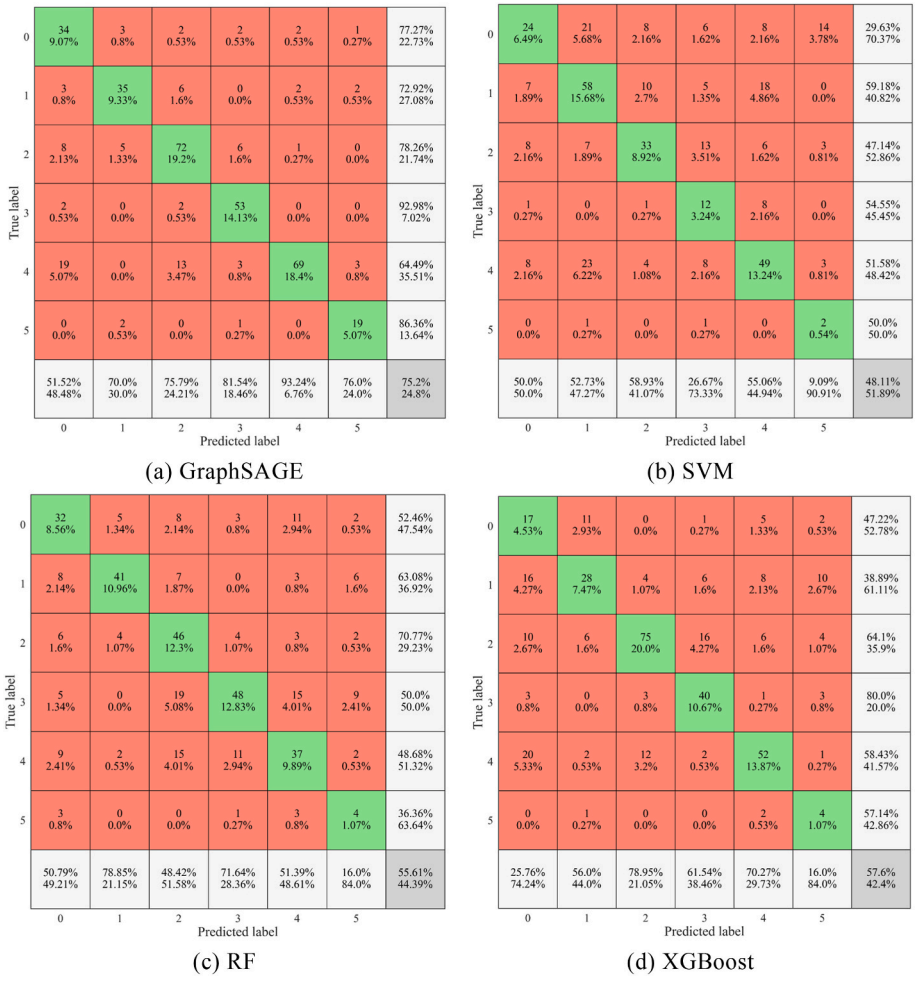

Fig. 11. Confusion matrix of four methods of blind Well A2 cored section for lithology identification. (Dolomite, dolomitic mudstone, mudstone, silicalite, siltstone and silty mudstone are represented by 0, 1, 2, 3, 4, 5.)

constructing edges, constructing edges based on depth sequences, and constructing edges based on depth sequences and clustering algorithm. To ensure that other test conditions are consistent, the classifiers used 4 layers, with learning rate of 0.0005, Leaky RELU, cross-entropy, and 5000 iterations to find out the influence of graph structure on GraphSAGE lithology identification ability. The first graph structure is that the raw logging curve is directly used as the input data, and the edges between nodes are not constructed (Fig. 13a). By comparison, it can be seen that GraphSAGE only inputs the raw data, and the testing data accuracy can reach 80.46% (Fig. 14). Although the identification accuracy is 10% lower than that of the input constructed graph structure, it is still higher than the SVM, RF and XGBoost identification accuracy. The second method is to construct graph according to the depth sequence, which fully considers the geological factors and can reflect the vertical spatial distribution of the Fengcheng Formation (Fig. 13b). The accuracy of the testing data of the model after training is 91.26% (Fig. 14). The third method is introduced in Section 3.1. After constructing edges according to the depth sequence, multiple clusters are obtained by clustering method (Fig. 13c). In each cluster, a connection is established between two nodes whose Euclidean distance obtained by calculating the feature vectors of nodes is less than the set quantile. The accuracy of the testing data after training is 90.41% (Fig. 14), which is slightly lower than the second graph structure.

Because of the randomly selected training and testing data, there is a high degree of similarity between the training and testing data. In order to further test the impact of the three graph structures on the model identification and generalization ability, Well A2 is selected as the blind test. The well logs of A2 are used for the graph construction but not for the model training. The accuracy of lithology identification in the cored section of the this well with three types of graph structure are 62.47%, 70.95%, and 75.2% (Fig. 14). The identification accuracy of the first graph structure without constructing edges is the lowest among the three. The accuracy of the third graph structure with edges construction based on the depth sequence and clustering method is much higher than the second graph structure with edges construction by the depth sequence after training. The result shows that the accuracy of these two forms of graph structures is similar in the randomly divided training and testing data, but the model generalization ability trained by the third graph structure is significantly better than that of the second (Fig. 14).

The above results show that the lithology identification model based on the third graph construction method (Sequence + Clustering) has good classification and generalization ability. The priority of the graph construction is the selection of clustering method and quantile. Four common clustering methods, K-means, Gaussian mixture, Agglomerative clustering and Affinity propagation, are selected for comparative test. The calculation results show that the model based on Affinity propagation clustering algorithm shows higher accuracy than other clustering algorithms in the testing data of Well A1 and blind well data (Fig. 15). After obtaining multiple clusters according to Affinity propagation clustering algorithm, a connection is established between two nodes whose Euclidean distance obtained by calculating the feature vectors of nodes is less than the set quantile in each cluster. According to Fig. 16, when the quantile is equal to 0, it is equivalent to the second graph construction method (Sequence). The testing data accuracy is

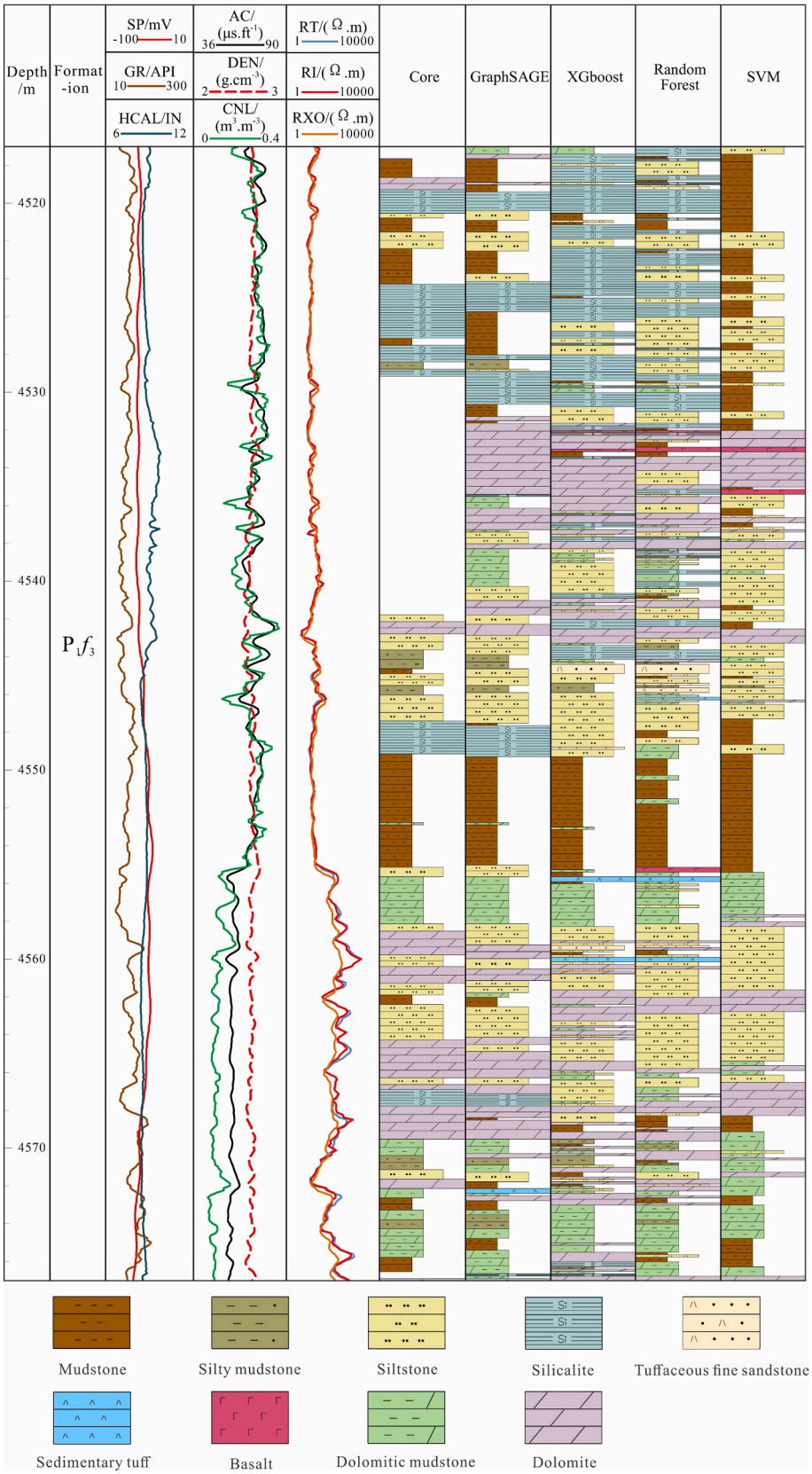

Fig. 12. The blind well identification results obtained by the models based on four different AI algorithms, compared with core description and observations.

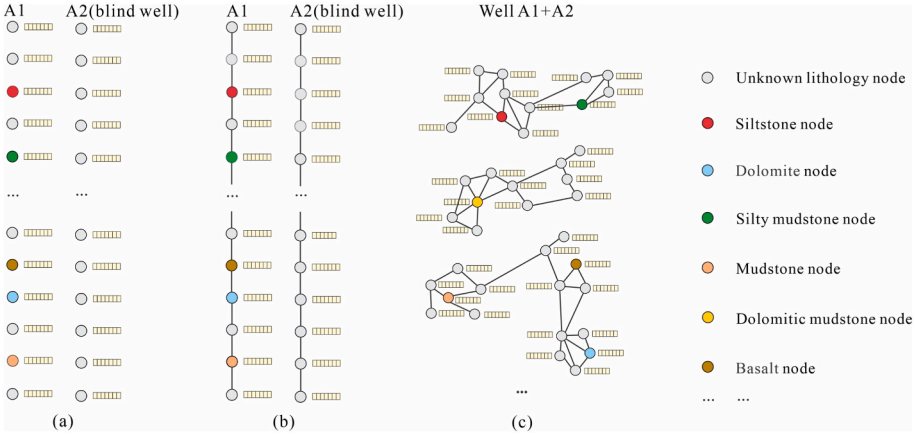

Fig. 13. Three graph structures. (a) The first graph structure, no edges. (b) The second graph structure, depth sequence. (c) The third graph structure, depth sequence + clustering.

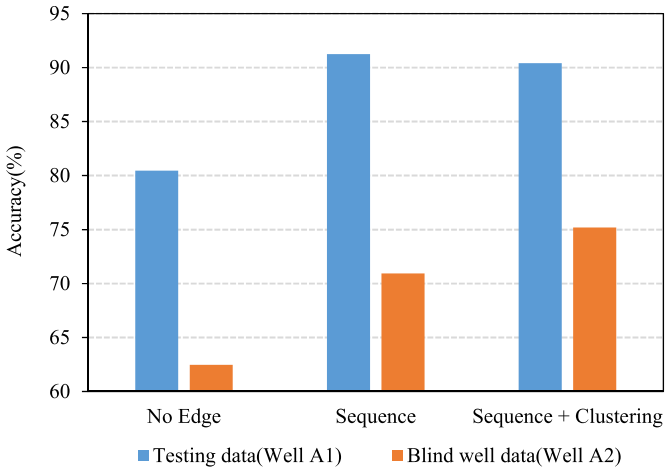

Fig. 14. Influence of graph construction method on lithology identification results in testing data of A1 and the blind well date of A2.

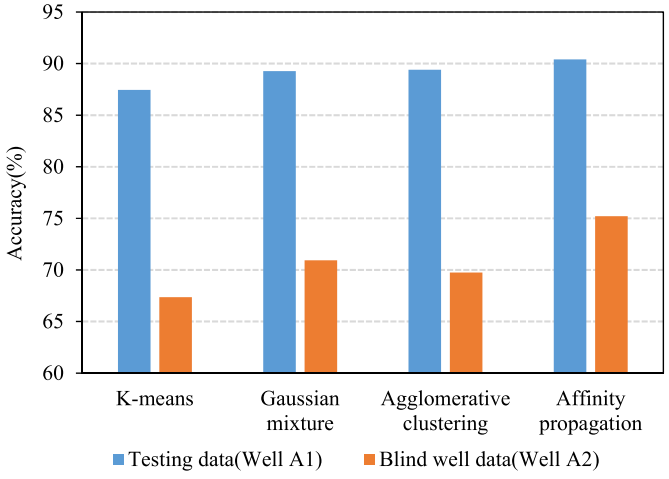

Fig. 15. Influence of clustering method on lithology identification results in testing data of A1 and the blind well date of A2.

high but the generalization ability is weak. With the increase of the quantile, the testing data accuracy decreases slightly, but the blind well data identification accuracy continues to improve. When the quantile is greater than 3, the accuracy of the testing data and blind well data shows

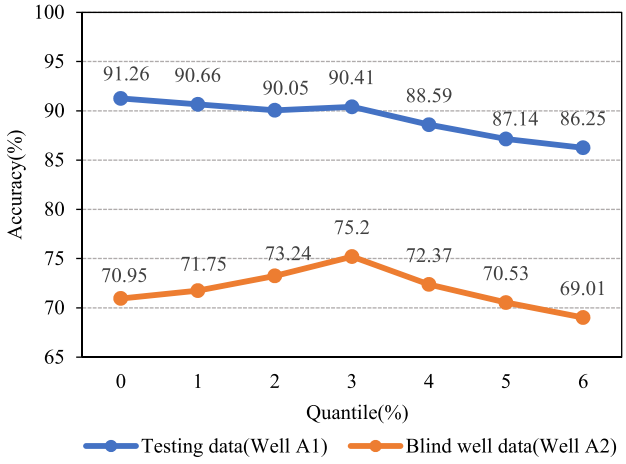

Fig. 16. Influence of quantile on lithology identification results of testing data of A1 and the blind well date of A2.

a downward trend, indicating that the increase of the number of edges between non-adjacent nodes will improve the generalization ability of the model, but the excessive number will also lead to more wrong connections and reduce the reliability of the model.

5.2. *Over-smoothing problem*

In addition to the graph structure, the layers composition in GraphSAGE also has an important impact on the lithology identification results. According to Section 3.2, the GraphSAGE message passing framework, collecting feature vectors from neighboring nodes and updating the representation of the root node, is the process of nodes information aggregation and updating. This mechanism can take advantage of the information conveyed in the graph structure, and also make the nodes connected by edges have similar feature representation. With the continuous increase of the model depth, the scope of information aggregation gradually expands. This process, which is the problem of over-smoothing, significantly will improve the probability of the root node aggregating the information of different label nodes, resulting in the decline of the classification ability of the model (Oono et al., 2020; Liu et al., 2022).

Before discussing whether the increase of the number of layers in the lithology identification model will lead to the occurrence of excessive smoothing, it is necessary to define the aggregator of GraphSAGE. In the

case of randomly dividing the training and the testing data, Mean, Pooling and LSTM aggregator are used for testing. The results of testing data show that when LSTM aggregator is selected, the runtime of the model is 17 times longer than the other two aggregators, but the identification accuracy is the lowest of the three. When the Mean aggregator is selected, the model training takes the shortest time and the identification accuracy exceeds 90% (Table 5). Under the condition that aggregator and other hyperparameters are consistent, when the number of layers is 2, the accuracy of the testing data is only 41.38%, signifying that the model is under-fitting. When the number of layers is 3, the accuracy of the testing data is 86.65%. With the continuous increase of the number of layers, the accuracy of the testing data becomes stable at about 90%, and there is no over-smoothing issue in the identification model. Considering Well A2 as the blind test, its incorporated in the construction of the graph but not in the model training. When the number of layers is 2, the identification accuracy of the blind test in this well is 27.47%. When the number of layers is 3, the accuracy is 66.13% and with 4 layers, the accuracy is improved to 75.2%. As the number of layers continues to increase, the accuracy shows a gradually downward trend. When the number of layers is 8, the identification accuracy is only 61.12%, and the model is over-smoothing in the blind well (Fig. 17).

The reason for the above results can be due to two points: the first is the division scheme of training and testing data. In well A1, as the data is randomly divided to the training and the testing data, most of the neighboring nodes around the root node of the testing data, with adjacent depth and similar logging curve characteristics, would be categorized in the same class as the root node, with known labels. Therefore, the probability of correctly classifying the nodes of testing data is high, and the accuracy is less affected as the number of layers increases. In well A2 which is used as a blind test, the neighboring nodes with known labels account for a relatively small proportion. As the aggregation depth increases, the classification of root nodes is subject to more interference information, resulting in a decrease in accuracy.

The second is the influence of the thickness of a single lithological layer. Because the vertical distribution of the stratum is integrated in the graph structure, the increase of the number of layers could represent the expansion of the information aggregation range in the actual vertical space. For example, the 2 layers network essentially aggregates the node information within the range of 0.5 m, and the nodes within the range tend to be divided into the same category. With each additional layer of model depth, the aggregation range is expanded by 0.25 m. When the depth of the model is 8 layers, the information of nodes within the true range of 2.5 m is aggregated, which is equivalent to that all nodes within this range tend to be classified into the same type of lithology (Fig. 18). Based on the thickness of a single lithologic layer statistics in the cored section of Well A1 and A2, the thickness of single lithologic layer in Well A1 is mainly 1–2 m, and the thickness of single lithologic layer in Well A2 is generally 0–1 m (Fig. 19). Herein, thickness distribution interval of single lithological layer in Well A1 corresponds to the model depth of 4–8 layers, and the thickness distribution interval of single lithological layer in Well A2 corresponds to the model depth of 2–4 layers. Therefore, for Well A1, there is no decline in the accuracy of testing data. Moreover, when Well A2 is used as a blind test and the model depth exceeds 4 layers, the model will aggregate too much information of different lithologies, resulting in the reduction of classification ability and decline in identification accuracy.

Table 5
Comparison of identification accuracy and runtime of different aggregators.

	Mean aggregator	Pooling aggregator	LSTM aggregator
Accuracy (%)	90.41	88.47	84.34
Runtime(s)	73.34	74.63	1263.75

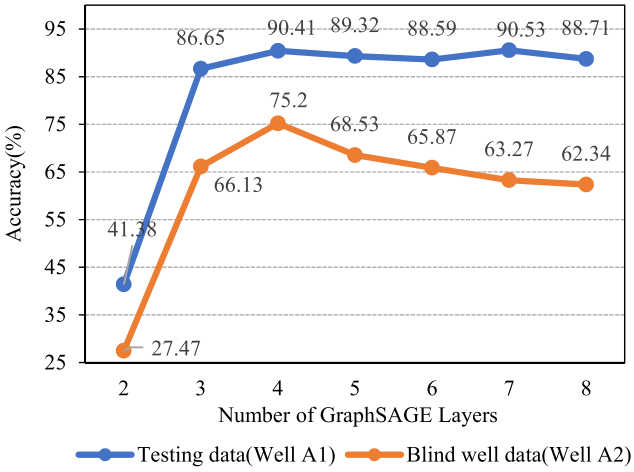

Fig. 17. Influence of model depth on lithology identification results of testing data of A1 and the blind well date of A2.

6. Conclusions

The lithology of shale oil reservoirs varies a lot and is difficult to identify. In this paper, the GraphSAGE method is used to train the model of the constructed graph, which connect datapoints from adjacent depth and similar log response features based on operator intention. This method was found to significantly improve the accuracy of lithology identification process. Based on the results the following conclusions can be made.

(1) By focusing on the problems of lithology variations, complex logging response characteristics, and difficulty in characterizing continental shale oil reservoirs, the vertical distribution of the stratum and node logging curve similarity information are integrated into the graph structure through personal preference at the beginning of model training. This process led to improvement in lithology identification accuracy in continental shale oil reservoir via conventional logging data by using GraphSAGE to classify the nodes of the graph.
(2) The training and testing process of the lithology identification model adopted Random Subsampling. The application in the continental shale oil reservoir of Fengcheng Formation in the Mahu Sag of Junggar Basin showed that the accuracy of GraphSAGE (90.41%) in lithology identification is clearly higher than the commonly used machine learning methods such as SVM (69.66%), RF (77.79%) and XGBoost (78.64%), which reflects the superiority of graph neural network in conventional logging lithology identification.
(3) The graph structure is the key factor affecting the lithology identification results. The accuracy of the testing data, graph constructed based on the depth sequence and the combination of the depth sequence and AP clustering, is much higher when raw logging data would be the single input. Furthermore, the graph based on the combination of the depth sequence and AP clustering exhibited a higher accuracy and generalization in the blind test in a second well for lithology identification.
(4) The number of layers of neural network also has an important impact on the result of lithology identification. GraphSAGE has the over-smoothing problem when the model depth exceeds 4 layers in the process of lithology identification of a blind well. The reason for over-smoothing is not only affected by how training and testing data are divided, but also affected by the thickness of single lithology layer.

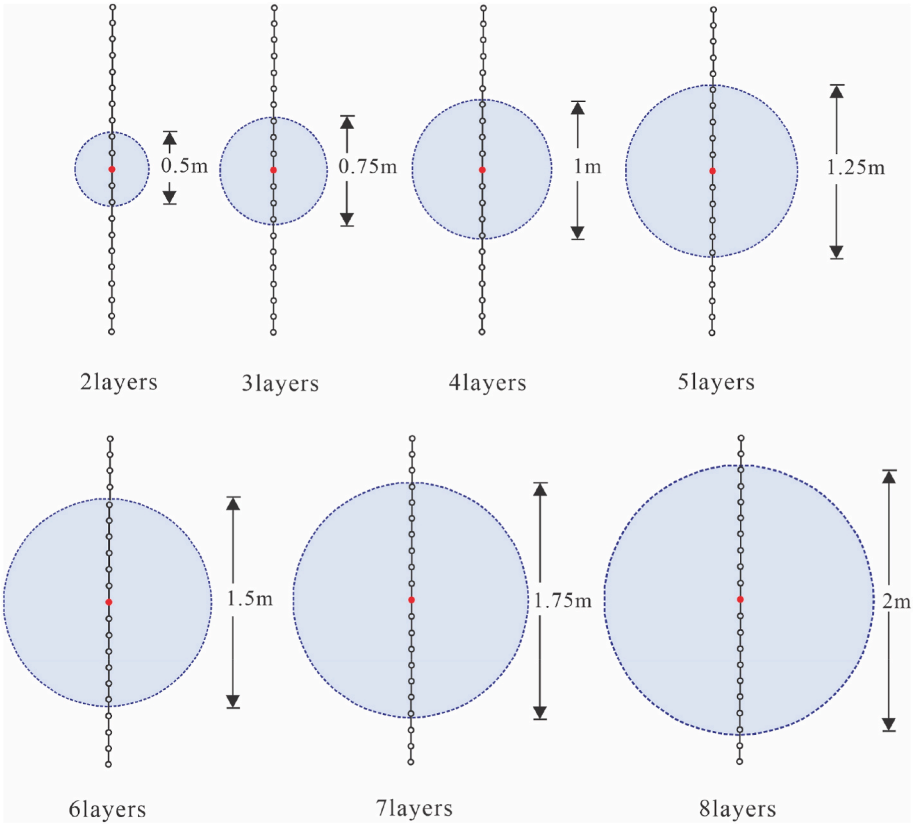

Fig. 18. Comparison between the number of layers and the aggregation range.

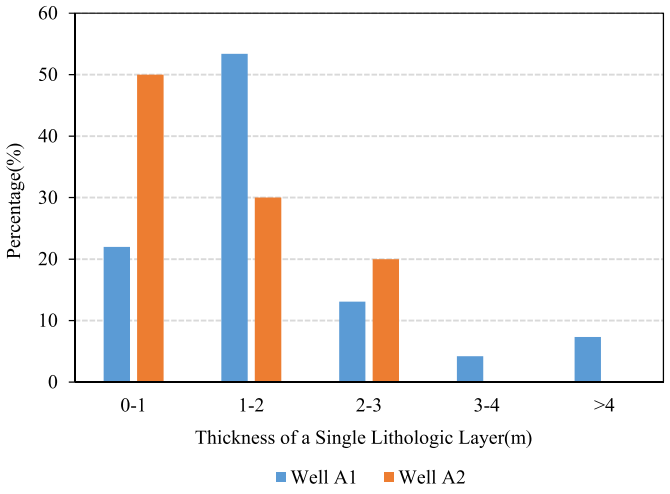

Fig. 19. Statistical results of thickness of single lithologic layer in Well A1 and Well A2 based on the cores.

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 data that has been used is confidential.

Acknowledgments

This study was financially supported by the National Natural Science Foundation of China (Grant No. 42090020, U1663203).

References

Abbas, M.A., Al-Mudhafar, W.J., 2021. Lithofacies classification of carbonate reservoirs using advanced machine learning: a case study from a southern Iraqi oil field. In: Offshore Technology Conference. https://doi.org/10.4043/31114-MS. Virtual and Houston, Texas, USA.

Al-Mudhafar, W.J., 2016. Incorporation of bootstrapping and cross-validation for efficient multivariate facies and Petrophysical modeling. In: SPE Low Perm Symposium. OnePetro. https://doi.org/10.2118/180277-MS.

Al-Mudhafar, W.J., 2017a. Integrating kernel support vector machines for efficient rock facies classification in the main pay of Zubair formation in South Rumaila oil field, Iraq. Modeling Earth Systems and Environment 3 (1), 1–8. https://doi.org/10.1007/s40808-017-0277-0.

Al-Mudhafar, W.J., 2017b. Integrating well log interpretations for lithofacies classification and permeability modeling through advanced machine learning algorithms. J. Pet. Explor. Prod. Technol. 7 (4), 1023–1033. https://doi.org/10.1007/s13202-017-0360-0.

Al-Mudhafar, W.J., 2020. Integrating machine learning and data analytics for geostatistical characterization of clastic reservoirs. J. Petrol. Sci. Eng. 195, 107837 https://doi.org/10.1016/j.petrol.2020.107837.

Al-Mudhafar, W.J., Wood, D.A., 2022. Tree-based ensemble algorithms for lithofacies classification and permeability prediction in heterogeneous carbonate reservoirs. In: Offshore Technology Conference. OnePetro. https://doi.org/10.4043/31780-MS.

Al-Mudhafar, W.J., Abbas, M.A., Wood, D.A., 2022. Performance evaluation of boosting machine learning algorithms for lithofacies classification in heterogeneous carbonate reservoirs. Mar. Petrol. Geol. 145, 105886 https://doi.org/10.1016/j.marpetgeo.2022.105886.

Ameur-Zaimeche, O., Zeddouri, A., Heddam, S., Kechiched, R., 2020. Lithofacies prediction in non-cored wells from the Sif Fatima oil field (Berkine basin, southern Algeria): a comparative study of multilayer perceptron neural network and cluster analysis-based approaches. J. Afr. Earth Sci. 166, 103826 https://doi.org/10.1016/j.jafrearsci.2020.103826.

Ao, Y., Li, H.Q., Zhu, L.P., Ali, S., Yang, Z.G., 2019. Logging lithology discrimination in the prototype similarity space with random forest. Ieee Geosci Remote S 5 (16), 687–691. https://doi.org/10.1109/LGRS.2018.2882123.

Avseth, P., Mukerji, T., 2002. Seismic lithofacies classification from well logs using statistical rock physics. Petrophysics-The SPWLA Journal of Formation Evaluation and Reservoir Description 43 (2).
Battaglia, P.W., Hamrick, J.B., Bapst, V., et al., 2018. Relational Inductive Biases, Deep Learning, and Graph Networks. https://doi.org/10.48550/arXiv.1806.01261 arXiv preprint arXiv: 1806.01261.
Bressan, T.S., de Souza, M.K., Girelli, T.J., Junior, F.C., 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.
Cao, J., Lei, D.W., Li, Y.W., Tang, Y., Abulimit, Chang, Q.S., Wang, T.T., 2015. Ancient high-quality alkaline lacustrine source rocks discovered in the lower permian Fengcheng Formation, Junggar Basin. Acta Pet. Sin. 36 (7), 781–790. https://doi.org/10.7623/syxb201507002 (in Chinese with English abstract).
Chang, J., Kang, Y., Li, Z.R., Zheng, W.X., Lv, W.J., Feng, D.Y., 2020. Cross-domain lithology identification using active learning and source reweighting. Geosci. Rem. Sens. Lett. IEEE. https://doi.org/10.1109/LGRS.2020.3041960.
Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939785.
Chen, D.L., Lin, Y.K., Li, W., Li, P., Zhou, J., Sun, X., 2020. Measuring and Relieving the Over-smoothing Problem for Graph Neural Networks from the Topological View, pp. 3438–3445. https://doi.org/10.1609/aaai.v34i04.5747.
De Silva, P.N.K., Simons, S.J.R., Stevens, P., Philip, L.M., 2015. A comparison of north American shale plays with emerging non-marine shale plays in Australia. Mar. Petrol. Geol. 67, 16–29. https://doi.org/10.1016/j.marpetgeo.2015.04.011.
Delavar, M.R., 2022. Hybrid machine learning approaches for classification and detection of fractures in carbonate reservoir. J. Petrol. Sci. Eng. 208, 109327 https://doi.org/10.1016/j.petrol.2021.109327.
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., Lyu, W.Y., Xu, C.S., Liu, J.J., Mao, Z., Tian, H., Sun, F.W., 2020. 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., Du, X.Y., He, J., Sun, F.T., 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.
Du, J.H., Hu, S.Y., Pang, Z.L., Lin, S.H., Hou, L.H., Zhu, R.K., 2019. The types, potentials and prospects of continental shale oil in China. China Petroleum Exploration 24 (5), 560–568. https://doi.org/10.3969/j.issn.1672—7703.2019.05.003 (in Chinese with English abstract).
Frey, B.J., Dueck, D., 2007. Clustering by passing messages between data points. Science 972 (315), 972–976. https://doi.org/10.1126/science.1136800.
Fruchterman, T.M.J., Reingold, E.M., 1991. Graph drawing by force-directed placement. Software Pract. Ex. 11 (21), 1129–1164. https://doi.org/10.1002/spe.4380211102.
Gajdoš, P., Ježowicz, T., Uher, V., Dohnálek, P., 2016. A parallel fruchterman-reingold algorithm optimized for fast visualization of large graphs and swarms of data. Swarm Evol. Comput. 26, 56–63. https://doi.org/10.1016/j.swevo.2015.07.006.
Ghosh, S., Galvis-Portilla, H.A., Klockow, C.M., Slatt, R.M., 2018. An application of outcrop analogues to understanding the origin and abundance of natural fractures in the woodford shale. J. Petrol. Sci. Eng. 164, 623–639. https://doi.org/10.1016/j.petrol.2017.11.073.
Gong, L., Wang, J., Gao, S., Fu, X.F., Liu, B., Miao, F.B., Zhou, X.P., Meng, Q.K., 2021. Characterization, controlling factors and evolution of fracture effectiveness in shale oil reservoirs. J. Petrol. Sci. Eng. 203, 108655 https://doi.org/10.1016/j.petrol.2021.108655.
Hamilton, W.L., Ying, R., Leskovec, J., 2017. Inductive Representation Learning on Large Graphs, pp. 1024–1034. https://doi.org/10.48550/arXiv.1706.02216.
Han, R.Y., Wang, Z.W., Wang, W.H., Xu, F.H., Qi, X.H., Cui, Y.T., 2021. Lithology identification of igneous rocks based on XGboost and conventional logging curves, a case study of the eastern depression of liaohe basin. J. Appl. Geophys. 195, 104480 https://doi.org/10.1016/j.jappgeo.2021.104480.
Huang, W.L., Gao, F., Liao, J.P., Chuai, X.Y., 2021. A deep learning network for estimation of seismic local slopes. Petrol. Sci. 1 (18), 92–105. https://doi.org/10.1007/s12182-020-00530-1.
Jin, Z.J., Zhu, R.K., Liang, X.P., Shen, Y.Q., 2021. Several issues worthy of attention in current lacustrine shale oil exploration and development. Petrol. Explor. Dev. 48 (6), 1276–1287. https://doi.org/10.11698/PED.2021.06.20 (in Chinese with English abstract).
Lei, D.W., Chen, G.Q., Liu, H.L., Li, X., Abulimit, Tao, K.Y., Cao, J., 2017. Study on the forming conditions and exploration fields of the Mahu giant oil (gas) province, Junggar Basin. doi:cnki:sun:dzxe.0.2017-07-013 Acta Geol. Sin. 91 (7), 1604–1619 (in Chinese with English abstract).
Li, G.H., Qiao, Y.H., Zheng, Y.F., Li, Y., Wu, W.J., 2019. Semi-supervised learning based on generative adversarial network and its applied to lithology recognition. IEEE Access 7, 67428–67437. https://doi.org/10.1109/ACCESS.2019.2918366.
Li, Z.R., Kang, Y., Feng, D.Y., Wang, X.M., Lv, W.J., Chang, J., Zheng, W.X., 2020. Semi-supervised learning for lithology identification using laplacian support vector machine. J. Petrol. Sci. Eng. 195, 107510 https://doi.org/10.1016/j.petrol.2020.107510.
Li, K.W., Xi, Y.J., Su, Z.X., Zhu, J.B., Wang, B.S., 2021a. Research on reservoir lithology prediction method based on convolutional recurrent neural network. Comput. Electr. Eng. 95, 107404 https://doi.org/10.1016/j.compeleceng.2021.107404.
Li, Z.R., Wu, Y.P., Kang, Y., Lv, W.J., Feng, D.Y., Yuan, C.H., 2021b. Feature-depth smoothness based semi-supervised weighted extreme learning machine for lithology identification. J. Nat. Gas Sci. Eng. 96, 104306 https://doi.org/10.1016/j.jngse.2021.104306.
Li, Z.R., Kang, Y., Lv, W.J., Zheng, W.X., Wang, X.M., 2021c. Interpretable semisupervised classification method under multiple smoothness assumptions with application to lithology identification. Geosci. Rem. Sens. Lett. IEEE 18 (3), 386–390. https://doi.org/10.1109/LGRS.2020.2978053.
Liu, G.P., Zeng, L.B., Sun, G.Q., Zu, K.W., Qin, L.B., Mao, Z., Ostadhassan, M., 2020a. Natural fractures in tight gas volcanic reservoirs and their influences on production in the xujiaweizi depression, Songliao Basin, China. AAPG Bull. 10 (104), 2099–2123. https://doi.org/10.1306/05122017169.
Liu, G.P., Zeng, L.B., Wang, X.J., Ostadhassan, M., Wang, Z.L., Mao, Z., Tie, Q., 2020b. Natural fractures in deep tight gas sandstone reservoirs in the thrust belt of the southern Junggar Basin, northwestern China. Interpretation 4 (8), P81–P93. https://doi.org/10.1190/INT-2020-0051.1.
Liu, H.N., Wu, Y.P., Cao, Y.C., Lv, W.J., Han, H.W., Li, Z.R., Chang, J., 2020c. Well logging based lithology identification model establishment under data drift: a transfer learning method. Sensors 20 (13), 3643. https://doi.org/10.3390/s20133643.
Liu, X., Sun, D., Wei, W., 2022. Alleviating the over-smoothing of graph neural computing by a data augmentation strategy with entropy preservation. Pattern Recogn. 132, 108951 https://doi.org/10.1016/j.patcog.2022.108951.
Lv, W.J., Kang, Y., Zheng, W.X., Wu, Y.P., Li, Z.R., 2020. Feature-temporal semi-supervised extreme learning machine for robotic terrain classification. IEEE Trans. Circuits Syst. II 67 (12), 3567–3571. https://doi.org/10.1109/TCSII.2020.2990661.
Oono, K., Suzuki, T., 2020. Graph Neural Networks Exponentially Lose Expressive Power for Node Classification. https://doi.org/10.48550/arXiv.1905.10947.
Ortiz-Bejar, J., Paternina, M.R.A., Zamora-Mendez, A., Lugnani, L., Tellez, E., 2022. Power system coherency assessment by the affinity propagation algorithm and distance correlation. Sustainable Energy, Grids and Networks 30, 100658. https://doi.org/10.1016/j.segan.2022.100658.
Rogers, S.J., Fang, J.H., Karr, C.L., Stanley, D.A., 1992. Determination of lithology from well logs using a neural network. AAPGBull 76, 731–739. https://doi.org/10.1306/BDFF88BC-1718-11D7-8645000102C1865D.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 303–310. https://doi.org/10.1038/323533a0.
Sajid, S., Chouinard, L., Carino, N., 2022. Condition assessment of concrete plates using impulse-response test with affinity propagation and homoscedasticity. Mech. Syst. Signal Process. 178, 109289 https://doi.org/10.1016/j.ymssp.2022.109289.
Sebtosheikh, M.A., Salehi, A., 2015. Lithology prediction by support vector classifiers using inverted seismic attributes data and petrophysical logs as a new approach and investigation of training data set size effect on its performance in a heterogeneous carbonate reservoir. J. Petrol. Sci. Eng. 134, 143–149. https://doi.org/10.1016/j.petrol.2015.08.001.
Soeder, D.J., 2018. The successful development of gas and oil resources from shales in north America. J. Petrol. Sci. Eng. 163, 399–420. https://doi.org/10.1016/j.petrol.2017.12.084.
Sohail, G.M., Radwan, A.E., Mahmoud, M., 2022. A review of Pakistani shales for shale gas exploration and comparison to north American shale plays. Energy Rep. 8, 6423–6442. https://doi.org/10.1016/j.egyr.2022.04.074.
Specht, D.F., 1990. Probabilistic neural networks. Neural Network. 3 (1), 109–118. https://doi.org/10.1016/0893-6080(90)90049-Q.
Sun, Y.Q., Chen, J.P., Yan, P.B., Zhong, J., Sun, Y.X., Jin, X.Y., 2022. Lithology identification of uranium-bearing sand bodies using logging data based on a BP neural network. Minerals-Basel 5 (12), 546. https://doi.org/10.3390/min12050546.
Tang, H., 2008. Improved carbonate reservoir facies classification using artificial neural network method. In: Canadian International Petroleum Conference. OnePetro. https://doi.org/10.2118/2008-122.
Tang, H., White, C.D., 2008. Multivariate statistical log log-facies classification on a shallow marine reservoir. J. Petrol. Sci. Eng. 61 (2–4), 88–93. https://doi.org/10.1016/j.petrol.2008.05.004.
Tang, Y., Guo, W.J., Wang, X.T., Bao, H.J., Wu, H.S., 2019. A new breakthrough in exploration of large conglomerate oil province in Mahu sag and its implications. Xinjing Pet. Geol. 40 (2), 127–137. https://doi.org/10.7657/XJPG20190201 (in Chinese with English abstract).
Tang, Y., Cao, J., He, W.J., Guo, X.G., Zhao, K.B., Li, W.W., 2021. Discovery of shale oil in alkaline lacustrine basins: the late paleozoic Fengcheng Formation, Mahu sag, Junggar Basin, China. Petrol. Sci. 5 (18), 1281–1293. https://doi.org/10.1016/j.petsci.2021.04.001.
Tian, M., Omre, H., Xu, H.M., 2021. Inversion of well logs into lithology classes accounting for spatial dependencies by using hidden markov models and recurrent neural networks. J. Petrol. Sci. Eng. 196, 107598 https://doi.org/10.1016/j.petrol.2020.107598.
Wang, G., Ju, Y., Carr, T.R., Li, C., Cheng, G., 2014. Application of artificial intelligence on black shale lithofacies prediction in marcellus shale, appalachian basin. Unconventional Resources Technology Conference. https://doi.org/10.15530/URTEC-2014-1935021.
Wang, X.D., Yang, S.C., Zhao, Y.F., Wang, Y., 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, Q.S., Zhang, X.J., Tang, B., Ma, Y.J., Xing, J.S., Liu, L.F., 2021. Lithology identification technology using BP neural network based on XRF. Acta Geophys. 6 (69), 2231–2240. https://doi.org/10.1007/s11600-021-00665-8.

Wang, X.J., Jin, Z.J., Chen, G., Peng, M., Huang, L.L., Wang, Z.L., Zeng, L.B., Lu, G.Q., Du, X.Y., Liu, G.P., 2022a. Multi-scale natural fracture prediction in continental shale oil reservoirs: a case study of the Fengcheng Formation in the Mahu sag, Junggar Basin, China. Front. Earth Sci. 10, 1–15. https://doi.org/10.3389/feart.2022.929467.

Wang, X.P., Zuo, R.G., Wang, Z.Y., 2022b. Lithological mapping using a convolutional neural network based on stream sediment geochemical survey data. Nat. Resour. Res. 31 (5), 2397–2412. https://doi.org/10.1007/s11053-022-10096-x.

West, D.B., 2001. Introduction to Graph Theory. Prentice hall, Upper Saddle River.

Wu, Y.P., Yang, Y.X., Lv, W.J., Chang, J., Li, Z.R., Feng, D.Y., Xu, T., Li, J., 2021. Robust unilateral alignment for subsurface lithofacies classification. IEEE Trans. Geosci. Rem. Sens. 60, 1–13. https://doi.org/10.1109/TGRS.2021.3070050.

Xie, Y.X., Zhu, C.Y., Zhou, W., Li, Z.D., Liu, X., Tu, M., 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.

Yu, Z.C., Wang, Z.Z., Zeng, F.C., Song, P., Baffour, B.A., Wang, P., Wang, W.F., Li, L., 2021. Volcanic lithology identification based on parameter-optimized GBDT algorithm: a case study in the jilin oilfield, Songliao Basin, ne China. J. Appl. Geophys. 194, 104443 https://doi.org/10.1016/j.jappgeo.2021.104443.

Yuan, C.H., Wu, Y.P., Li, Z.R., Zhou, H.S., Chen, S.B., Kang, Y., 2022. Lithology identification by adaptive feature aggregation under scarce labels. J. Petrol. Sci. Eng. 215, 110540 https://doi.org/10.1016/j.petrol.2022.110540.

Zeng, L.B., 2010. Microfracturing in the upper triassic Sichuan Basin tight-gas sandstones: tectonic, overpressure, and diagenetic origins. AAPG Bull. 12 (94), 1811–1825. https://doi.org/10.1306/06301009191.

Zeng, L.B., Lyu, W.Y., Li, J., Zhu, L.F., Weng, J.Q., Yue, F., Zu, K.W., 2016. Natural fractures and their influence on shale gas enrichment in Sichuan Basin, China. J. Nat. Gas Sci. Eng. 30, 1–9. https://doi.org/10.1016/j.jngse.2015.11.048.

Zeng, L.L., Ren, W.J., Shan, L.Q., Huo, F.C., Meng, F.Y., 2022. Lithology spatial distribution prediction based on recurrent neural network with kriging technology. J. Petrol. Sci. Eng. 214, 110538 https://doi.org/10.1016/j.petrol.2022.110538.

Zhi, D.M., Tang, Y., Yang, Z.F., Guo, X.G., Zheng, M.L., Wan, M., Huang, L.L., 2019. Geological characteristics and accumulation mechanism of continental shale oil in Jimusaer sag, Junggar Basin. Oil Gas Geol. 40 (3), 524–534. https://doi.org/10.11743/ogg20190308 (in Chinese with English abstract).

Zhi, D.M., Tang, Y., He, W.J., Guo, X.G., Zheng, M.L., Huang, L.L., 2021. Orderly coexistence and accumulation models of conventional and unconventional hydrocarbons in lower permian Fengcheng Formation, Mahu sag, Junggar Basin. Petrol. Explor. Dev. 48 (1), 38–51. https://doi.org/10.11698/PED.2021.01.04 (in Chinese with English abstract).

Zhou, J., Cui, G.Q., Hu, S.D., Zhang, Z.Y., Yang, C., Liu, Z.Y., Wang, L.F., Li, C.C., Sun, M. S., 2020. Graph neural networks: a review of methods and applications. AI Open 1, 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001.

Zhu, S.F., Qin, Y., Liu, X., Wei, C.J., Zhu, X.M., Zhang, W., 2017. Origin of dolomitic rocks in the lower permian Fengcheng Formation, Junggar Basin, China: evidence from petrology and geochemistry. Min. Pet. 2 (111), 267–282. https://doi.org/10.1007/s00710-016-0467-x.