Proceedings of the 3rd International Conference on Functional Materials and Civil Engineering
DOI: 10.54254/2755-2721/2025.25770

Predictive Model of Sand Body Development Type Based on Graph Convolutional Neural Network Algorithm

Wenwei Min[1], Hongwen Deng[2], Yinghai Guo[1*]

[1]*School of Resources and Geosciences, China University of Mining and Technology, Xuzhou, China*
[2]*School of Energy Resources, China University of Geosciences (Beijing), Beijing, China*
**Corresponding Author. Email: guoyh@cumt.edu.cn*

Abstract. In this study, a sand body development type prediction model was constructed based on graph convolutional neural network (GCN), and through comparison experiments with traditional machine learning algorithms, such as CatBoost and random forest, it was found that GCN significantly outperforms other models (e.g., CatBoost 0.878) in terms of core metrics, such as accuracy (0.942), recall (0.942), and so on, and that its advantage stems from the fact that it can effectively capture the node-to-node The advantage stems from the effective capture of geologic correlation features between nodes. The performance of traditional integrated models is limited by the lack of explicit graph structure modeling capability, in which Adaboost (0.467) shows a serious imbalance between precision (0.707) and recall due to the sample imbalance problem. The research results confirm that graph neural networks have unique advantages in processing geologic data with spatial correlation, and the established prediction model provides a new technical means for identifying the spatial spreading pattern of sand bodies and revealing the evolutionary characteristics of sedimentary systems, which is of great theoretical value and practical significance for the prediction of oil and gas reservoirs and the analysis of sedimentary phases.

Keywords: Graphic Convolutional Neural Networks, Sand Body Development Types, Sedimentary Phase Analysis.

1. Introduction

In this study, a sand body development type prediction model was constructed based on graph convolutional neural network (GCN), and through comparison experiments with traditional machine learning algorithms, such as CatBoost and random forest, it was found that GCN significantly outperforms other models (e.g., CatBoost 0.878) in terms of core metrics, such as accuracy (0.942), recall (0.942), and so on, and that its advantage stems from the fact that it can effectively capture the node-to-node The advantage stems from the effective capture of geologic correlation features between nodes. The performance of traditional integrated models is limited by the lack of explicit graph structure modeling capability, in which Adaboost (0.467) shows a serious imbalance between precision (0.707) and recall due to the sample imbalance problem. The research results confirm that graph neural networks have unique advantages in processing geological data with spatial correlation, and the established prediction model provides a new technical means for identifying the spatial

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.25770

spreading pattern of sand bodies and revealing the evolutionary characteristics of sedimentary systems,

which is of great theoretical value and practical significance for the prediction of oil and gas reservoirs and the analysis of sedimentary phases.

The formation and distribution of fluvial sands, estuarine dams and turbidites are controlled by sedimentary dynamics and geological environment. River channel sands are mainly developed in the river environment, formed by lateral migration or vertical accretion of the river channel, and their morphology is affected by paleomorphology, hydrodynamic conditions, and the supply of material sources, and they are usually distributed in the form of strips or lenses [1]. The estuarine dam sand body is formed in the estuarine area where the river meets the sea (or lake), and the sediments carried by the river are unloaded due to the deceleration of the water flow, and are modified by the tides, waves and littoral currents, and are mostly in the form of a fan or a dolichoic shape. Turbiditic sand bodies, on the other hand, are developed in deep-water environments, where gravity-driven turbidity currents carry land-sourced detritus to basin slopes or abyssal plains for rapid accretion, and are characterized by poor sorting and abrupt changes in grain size. The core of studying these sand bodies lies in analyzing the hydrodynamic conditions, material source background and preservation mechanism in the sedimentary environment, and comprehensively determining their genesis types through the petrographic assemblage, depositional sequences, and three-dimensional spatial spreading characteristics [2].

Based on the multi-parameter fusion of source system, transportation system, depositional system and preservation system, machine learning algorithms provide a new paradigm for sand body type prediction [3]. The source system (source distance, parent rock type) determines the original composition and supply of sediment, which directly affects the development scale of sand body; the transport system (paleoslope, hydrodynamic index) controls the sediment transport path and sorting process, which determines the morphology and internal structure of the sand body; the depositional system (thickness of sand/mudstone, depositional phases) reflects the energy conditions of depositional environments and the distribution of phase zones, which is a direct basis for the identification of the type of sand body The preservation system (stratigraphic thickness, organic matter content, seismic attributes) affects the later diagenetic modification and spatial preservation of the sand body [4]. Machine learning integrates the above multi-scale parameters through nonlinear modeling, which can efficiently explore the hidden geological laws, break through the limitations of traditional single-factor analysis, and significantly improve the accuracy and generalization ability of sand body type prediction [5].

Machine learning-driven sand type prediction is of great practical significance in oil and gas exploration. On the one hand, by predicting the spatial spread of sand bodies with high accuracy, well deployment and reservoir modeling can be optimized, and exploration risks can be reduced; on the other hand, the key parameters (e.g., hydrodynamic index, paleoslope) inverted by the algorithm can provide dynamic constraints for paleogeographic reconstruction. In addition, the method can quickly process big data such as seismic attributes and logging curves to reveal depositional patterns that are difficult to be captured by traditional geostatistics, such as the hidden distribution pattern of turbiditic sands [6]. Future research needs to further integrate the simulation of geological processes with deep learning to improve the interpretability of the model, and construct a specialized algorithmic framework for different basin types to promote the leapfrog development of sedimentology from qualitative description to intelligent prediction. In this paper, we use the graphical convolutional neural network algorithm to predict the sand body development type, and

DOI: 10.54254/2755-2721/2025.25770

compare the effects of multiple machine learning algorithms, which provides a basic research model for the study of sand body development type.

2. Data set sources and data analysis

The dataset in this paper is a private dataset, which contains 304 well samples and consists of four major geologic system parameters: the source system (source distance, host rock type) reflects the sediment supply characteristics, the transport system (paleoslope, hydrodynamic index) depicts the dynamic conditions of sediment transport, the depositional system (thickness of sand/mudstone, sedimentary phase) characterizes the differences in depositional environments, and the preservation system (thickness of the stratum, organic matter content, seismic attributes) characterize the post-depositional preservation conditions. The predictive variable sand development type (channel sand/estuarine dam/turbidite sand) can be used to study the sedimentary system evolution. Selected datasets are presented as shown in Table 1.

Table 1. Partial data

Distance to source	Paleo slope	Hydrodyna mic index	Sedimentary phase code	Thickness of sand body	Thickness ofmudstone	Sand body development type
13.66	0.76	0.31	5.00	4.02	4.44	turbidite sand
17.19	1.10	0.01	5.00	4.30	6.44	turbidite sand
1.82	6.71	0.93	2.00	21.69	4.57	River sand
11.47	8.06	0.75	2.00	23.03	1.49	River sand
11.17	7.84	0.74	2.00	2.29	3.37	River sand
16.33	2.55	0.74	2.00	7.77	3.78	River sand
26.71	3.71	0.29	5.00	7.00	3.66	turbidite sand
12.53	2.41	0.50	4.00	7.80	1.97	estuary dam
41.19	2.32	0.03	5.00	0.00	4.71	turbidite sand
1.94	6.42	0.88	2.00	21.32	4.79	River sand
7.49	1.72	0.62	4.00	8.66	2.20	estuary dam

3. Method

3.1. Machine learning classification models

Decision tree makes decision through tree structure, each node splits the data based on feature threshold, and selects splitting metrics (e.g., information gain, Gini coefficient) to maximize the category purity. Recursive splitting until the termination conditions (e.g., depth limit or node purity) are satisfied, and the final leaf node gives the prediction result (classification or regression), and the core idea is hierarchical feature selection with local optimal splitting [7].

Random Forest is an integrated method based on Bagging, which improves generalization by training multiple decision trees in parallel. Each tree introduces double randomness during training: 1) samples have put-back sampling (Bootstrap); 2) some features are randomly selected during node splitting. Ultimately, the voting or averaging of aggregation results reduces the risk of overfitting of a single tree and enhances the robustness of the model [8].

DOI: 10.54254/2755-2721/2025.25770

Adaboost is a Boosting class algorithm that iteratively trains weak classifiers and adjusts sample weights. Each round increases the weight of misclassified samples, so that the subsequent classifiers pay more attention to difficult cases; ultimately, the weighted combination of all weak classifiers. The core is “error-driven learning”, which gradually corrects the model bias through the exponential loss function.

The gradient boosting tree takes the additive model as a framework, and optimizes the loss function with gradient descent. Each round of training new trees to fit the current model of the negative gradient, gradually superimposed on the model, through the function space of the gradient descent to reduce the bias. Commonly used regression trees are used as base learners and support customized loss functions [9].

CatBoost is an improvement for GBDT, which mainly optimizes the category feature processing and training efficiency. It adopts the ordering sequence to avoid information leakage in the training set and reduce overfitting; automatically transforms category variables into target statistics and enhances robustness by combining with the ranking strategy; and the symmetric tree structure improves the inference speed.

3.2. Graph convolutional neural network

Graph Convolutional Neural Network (GCN) is a deep learning model specialized for processing graph-structured data. Traditional convolutional neural networks (CNNs) are excellent in processing Euclidean data such as images and text, but their convolution operation relies on the regular grid structure of the data, and cannot be directly applied to graph data. Graph data has non-Euclidean characteristics, where nodes are connected to each other by complex edges, and the number of neighbors of each node may not be fixed (e.g., user relationships in social networks, atomic connections in molecular structures.) GCN achieves efficient extraction of node features and modeling of relationships by extending the idea of convolution to the graph domain. The core idea is to design a local information aggregation mechanism that enables each node to update its own representation by the features of its neighboring nodes, thus capturing the spatial or spectral domain features of the graph structure. The network structure of the graphical convolutional neural network is shown in Figure 1.

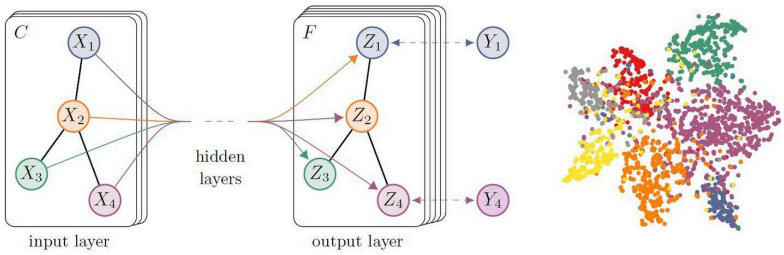

Figure 1. The network structure of the Graphical Convolutional Neural Network

The principle of GCN is mainly based on two types of methods: spectral domain and spatial domain. Spectral-domain methods map the graph signal to the frequency domain via the graph Fourier transform, and define the convolution operation using the feature decomposition of the Laplace matrix.ChebNet reduces the computational complexity by approximating the graph convolution kernel by Chebyshev polynomials; GCN is further simplified to a first-order approximation using a symmetrically normalized adjacency matrix to achieve feature propagation. Spatial domain methods, on the other hand, define convolution directly between a node and its

DOI: 10.54254/2755-2721/2025.25770

neighbors, and update the current node representation by aggregating the features of neighboring nodes. Typical models such as GraphSAGE use a sampling and aggregation strategy for feature fusion in the local neighborhood of a node, while GAT (Graph Attention Network) introduces an attention mechanism to dynamically assign weights to different neighbors [10].

4. Result

When implementing the graph neural network classification prediction experiments in Matlab R2024a, in terms of hardware and software configurations, the hardware is a workstation equipped with NVIDIA RTX 4090 GPU (24GB video memory), Intel i9-13900K CPU, and 64GB DDR5 RAM; and the software is installed with Deep Learning Toolbox and In the experimental parameter settings, a two-layer GCN network is constructed with a hidden layer dimension of 128, a learning rate of 0.01, an L2 regularization coefficient of 0.0005, a Dropout rate of 0.5, 200 iterations of Adam's optimizer, and a training test split in the ratio of 7:3.

For comparing models, Decision Tree, Random Forest, Adaboost, GBDT and CatBoost are used for prediction respectively, and for evaluation metrics, Accuracy, Recall, Precision and F1 are used, and for training set partitioning, the training and testing sets are randomly partitioned according to the ratio of 7:3.

Firstly, the comparison results of the GCN proposed in this paper with Decision Tree, Random Forest, Adaboost, GBDT and CatBoost are output, as shown in Table 2. The metrics comparison of each model is shown in Figure 2.

Table 2. Classification comparison results of different models

Model	Accuracy	Recall	Precision	F1
Decision Trees	0.818	0.818	0.803	0.808
Random Forest	0.849	0.849	0.84	0.842
Adaboost	0.467	0.467	0.707	0.49
Gradient Boosting Tree	0.836	0.836	0.824	0.825
CatBoost	0.878	0.878	0.872	0.874
GCN	0.942	0.942	0.941	0.941

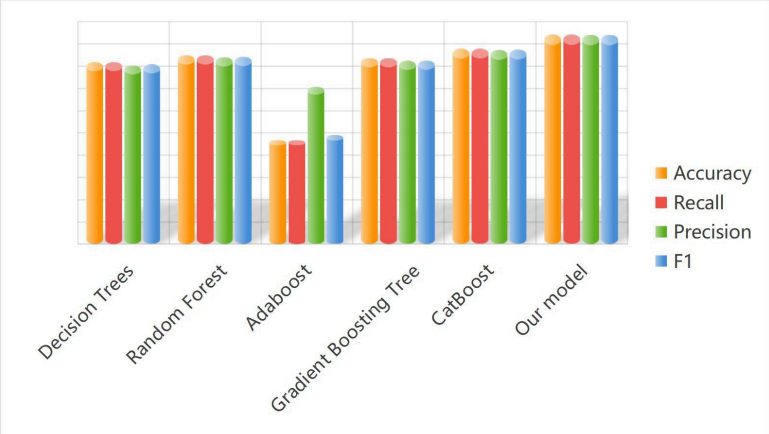

Figure 2. The metrics comparison of each model

DOI: 10.54254/2755-2721/2025.25770

The six classification models compared in the figure show that GCN significantly outperforms the other models in accuracy (0.942), recall (0.942), precision (0.941), and F1 score (0.941) for tasks that rely on inter-node associations. Traditional integration models such as CatBoost (accuracy 0.878) and Random Forest (0.849) perform next best, possibly due to the presence of nonlinear features in the data but lack of explicit graph structure support.Adaboost performs exceptionally poorly (accuracy 0.467), possibly due to sample imbalance or failure of the weak learner combinations that lead to model failure, and its high precision (0.707) with a low recall The contradiction between its high precision rate (0.707) and low recall rate suggests the risk of overfitting to a small number of classes. Overall, the excellent performance of GCN highlights the irreplaceability of graph neural networks in dealing with relational data, while the difference in the performance of traditional models reflects the importance of data characteristics and algorithmic fitness. In summary, the GCN model in this paper achieves good results in the prediction of sand body development types, and the prediction accuracy is much higher than that of the basic model, which is of great significance to the study of sand body development types.

5. Conclusion

In this paper, by comparing the performance of multiple machine learning algorithms in the prediction of sand body development type, we verify the significant advantages of graph convolutional neural network (GCN) in this field. The experimental results show that the GCN model demonstrates a prediction capability beyond traditional methods in the core metrics of accuracy (94.2%), recall (94.2%), precision (94.1%), and F1 score (94.1%), which is attributed to its network architecture that efficiently captures the characteristics of spatial associations among nodes. In contrast, integrated learning models such as CatBoost (87.8%) and Random Forest (84.9%), although capable of handling nonlinear features of the data, suffer from obvious performance limitations due to the lack of the ability to explicitly characterize the intrinsic associations of the data. Of particular note, the Adaboost model shows a significant imbalance between precision (70.7%) and recall (46.7%), revealing the algorithm's inherent flaw of being prone to overfitting a few classes in scenarios with uneven sample distribution. This study innovatively confirms the irreplaceability of graph neural networks in modeling the spatial relationships of geologic bodies, and its topological learning mechanism constructed through the adjacency matrix provides a more fitting solution for the geologic prediction task with strong spatial correlation such as sand body development. The research results not only provide a high-precision intelligent prediction tool for reservoir characterization research, but also open up a new technical path for modeling complex spatial structures in the field of geology.

References

[1] Liang, Weiyu, Longwei Qiu, and Yongqiang Yang. "Sedimentary system evolution and sandbody development characteristics of Jurassic Shaximiao Formation in the central Sichuan Basin." EGU General Assembly Conference Abstracts. 2024.

[2] Qianran, W. U., et al. "Diversity of depositional architecture and sandbody distribution of sublacustrine fans during forced regression: A case study of Paleogene Middle Sha 3 member in dongying sag, bohai bay basin, East China." Petroleum Exploration and Development 50.4 (2023): 894-908.

[3] Ye, Jianhong, et al. "Creep characteristics of calcareous coral sand in the South China Sea." Acta Geotechnica 17.11 (2022): 5133-5155.

[4] Ding, Xuanming, et al. "Experimental study on the behavior of single pile foundation under vertical cyclic load in coral sand." Ocean Engineering 280 (2023): 114672.

DOI: 10.54254/2755-2721/2025.25770

[5] Zhang, He, et al. "Whole petroleum system and hydrocarbon accumulation model in shallow and medium strata in northern Songliao Basin, NE China." Petroleum Exploration and Development 50.4 (2023): 784-797.
[6] Jianhui, Z. E. N. G., et al. "Complex gas-water contacts in tight sandstone gas reservoirs: Distribution pattern and dominant factors controlling their formation and distribution." Oil & Gas Geology 44.5 (2023): 1067-1083.
[7] Bouzid T , Grich A , Naboulsi A , et al. Optimizing Eriochrome Black T adsorption through In-Situ polymerization of Poly(aniline-co-formaldehyde) on biochar: Multivariate approach using full factorial Design, Box-Behnken, Artificial intelligence, and DFT [J]. Separation and Purification Technology, 2024, 351.
[8] Liu, Jianmin, et al. "Digital economy development and corporate bankruptcy risk: Based on the perspective of institutional isomorphism." Emerging Markets Finance and Trade 60.4 (2024): 793-807.
[9] Islam, Jahirul, et al. "Cognitive Modelling of Bankruptcy Risk: A Comparative Analysis of Machine Learning Models to Predict the Bankruptcy." 2024 12th International Symposium on Digital Forensics and Security (ISDFS). IEEE, 2024.
[10] Hu, Qian, et al. "Lateral load response of large-diameter monopiles in sand." Géotechnique 72.12 (2022): 1035-1050.