Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 12, No. 1, 2024

Research on Lithology Identification based on Machine Learning

Kunkun Li

School of Xi'an Shiyou University, Xi'an 710065, China

Abstract: Machine learning has great potential in lithology identification. Through supervised learning, unsupervised learning, semi-supervised learning, deep learning and other methods, features can be automatically extracted from complex seismic data and logging data to achieve efficient and accurate lithology classification. These methods not only improve the accuracy and efficiency of lithology identification, but also reduce the workload of geologists, allowing them to focus on higher-level analysis and decision making. Despite significant progress, machine learning still faces many challenges in lithology identification. First, data quality and quantity limitations remain a major problem, especially in certain areas where obtaining high-quality seismic and logging data is difficult. Second, the training and interpretability of complex models also need to be addressed, especially because the "black box" nature of deep learning models makes it difficult for geologists to understand their internal mechanisms and predictions. In addition, the generalization ability and overfitting of models, as well as the need for real-time data processing in practical applications, are also urgent challenges to be solved. To address these challenges, this paper proposes several solutions and future directions. Data enhancement and synthesis techniques can extend existing data sets and improve the robustness and accuracy of models. The development of interpretative models and visualization tools helps geologists understand and trust the decision-making process of models. Multi-source data fusion technology can effectively use multi-source heterogeneous data such as seismic data, well logging data and geological map to improve model performance. Online learning and transfer learning technologies can update models in real time, improve the adaptability and generalization ability of models, and develop more accurate and interpretable models.

Keywords: Machine Learning; Lithology Identification; Deep Learning.

1. Introduction

Lithology identification is an important task in geology and petroleum exploration, which determines the types of underground rocks by analyzing geological data [1]. Traditional lithology identification methods rely on the expertise and experience of geologists, combining geological maps, rock samples, seismic data and well logging data. Although these methods are effective in some cases, they often have problems such as strong subjectivity and low efficiency because they rely on human judgment [2]. With the development of information technology, data-driven machine learning has shown great potential in lithology identification.

Machine learning is a technology that uses computers to automatically learn from data and make predictions. Its powerful data processing and pattern recognition capabilities make it widely used in various fields [3]. Especially in the field of lithology identification, machine learning methods can process a large amount of geological data, automatically extract features and classify them, and significantly improve the accuracy and efficiency of lithology identification [4]. In recent years, with the rise of deep learning technology, convolutional neural network (CNN) [5] and other models have achieved remarkable results in lithology identification. These methods can not only process structured data, but also unstructured data such as seismic images, which further expands the application range of lithology identification.

In summary, through systematic review and analysis of existing literature, this paper aims to provide a comprehensive reference for researchers and engineers, help them understand the current research progress of lithology identification based on machine learning, and provide guidance for future research and application. It is hoped that this paper can promote the further application and development of machine learning technology in lithology identification, and make contributions to geology, petroleum exploration and other related fields.

2. Summary of Existing Research Results

Fang Dazhi[6]et al. used SVM to classify logging data in the lithology identification study of an oilfield. Their research shows that SVM is excellent at distinguishing between different types of rocks, especially when dealing with high-dimensional data. By optimizing kernel function and parameters, SVM model realizes high precision lithology classification, and the identification accuracy reaches more than 95%. Wang Qi [7] et al. used random forest model to conduct integrated learning of multiple groups of logging data in the study of lithology identification of complex carbonate rocks. Studies have shown that random forests are excellent at handling noise and outliers in data and have high robustness. Finally, the identification accuracy of the model reaches more than 90%, which significantly improves the reliability of lithology identification. Chen Weicheng [8] uses artificial neural network to identify lithology of logging data of sandstone type uranium mine, and computates the number of hidden layer neurons through two ways. GridSearchCV is used to optimize the number of hidden layer neurons and the learning rate. By identifying the actual data, the model can reach convergence within 200 iterations. The results of confusion matrix show that the recognition accuracy of permeable sandstone is 80%, and the recognition accuracy of impermeable siltstone and muddy siltstone is 60%. Zhou Yuankai [9] et al. discussed the application of deep neural network model in lithology classification of uranium logging

interpretation in the Nalinggou area, alleviated the impact of class imbalance on classification results by using models with different structures, analyzed the hierarchical structure and training process of the model, and explained the internal mechanism and decision logic of the model in a more comprehensive way. The results show that the recognition accuracy of long and short time memory network is higher than 80% while maintaining high training efficiency, and the accuracy of 8-layer fully connected network is higher than 90%. Gao Jiatian [10] et al., using the well logging data of several Wells in the Songliao Basin to conduct model research, proposed a lithology identification method based on PSO-BP. The lithology identification method based on PSO-BP is realized through data preprocessing of log source data, constructing network identification model, optimizing lithology identification model and evaluating the output result of the model. After repeated tests, the results show that the average accuracy of lithology identification using PSO-BP method can reach 92.2%, which provides reliable support for reservoir prediction. Zhao Ranlei [11] et al. used principal component analysis to screen out four characteristic logging curves sensitive to volcanic lithology identification as input, and used XGBoost algorithm to build a lithology identification model for volcanic lithology identification. After the lithology identification results were given by the model, the results showed that the accuracy rate of XGBoost algorithm in blind well section identification reached 96.13%. Chen Ganghua [12] et al., aiming at the characteristics of longitudinal sequential logging data, constructed a bidirectional short-short-memory neural network (BiLSTM) lithography identification model, adopted random forest method to select features of conventional logging data and other parameters, and trained BiLSTM model with selected parameters as input variables. The results show that the lithology identification accuracy of the BiLSTM model is 0.86, which proves that the BILSTM model is suitable for the lithology identification of beach bar sand reservoir.

The application of machine learning in lithology identification has achieved remarkable results, and each method has its advantages and disadvantages. Support vector machines (SVM) and random forests (RF) perform well in handling high and noisy data. Convolutional neural network (CNN) has advantages in extracting complex features and processing 3D data. (XGBoost) has significant advantages in handling large-scale data and improving predictive performance; Long short-term memory network (LSTM) is outstanding in capturing the features of time series.

Overall, the machine learning method has demonstrated its strong adaptability and accuracy in lithology identification. In the future, by combining multiple machine learning methods and interdisciplinary cooperation to further improve the performance and interpretation of the model, it will promote the development of lithology identification technology, and provide more powerful tools and methods for geological research and resource exploration.

3. Classification of Machine Learning Methods

In lithology identification, machine learning methods can be divided into supervised learning, unsupervised learning, semi-supervised learning and deep learning according to their learning styles and application scenarios. Each of these approaches has advantages and disadvantages, and is suitable for different data types and task requirements. Each method and its application to lithology identification are described in detail below.

3.1. Supervised Learning

Supervised learning is a method of training a model with labeled data so that it can predict the output based on the input data. In lithology identification, supervised learning is often used to learn from existing lithology label data in order to classify unlabeled data.

(1) Support Vector Machine (SVM) [13] : SVM is a supervised learning algorithm for classification and regression, which separates different categories of data by finding the optimal hyperplane. In lithology identification, SVM can be used to distinguish between different types of rocks.

(2) Decision Tree [14] : Decision tree conducts classification or regression by building a tree model, in which each node represents a feature and each branch represents the possible value of the feature. In lithology identification, decision tree can intuitively show the influence of features on classification results.

(3) Random Forest [15] : Random forest is an integrated learning method composed of multiple decision trees, which improves classification accuracy and robustness through voting mechanism. In lithology identification, random forest can process high dimensional data and noise data, and the effect is remarkable.

(4) Neural Network [16] : Neural network is a model composed of multiple neurons, which can learn and predict by simulating biological neural network. In lithology identification, neural network is especially suitable for processing complex nonlinear data.

3.2. Unsupervised Learning

Unsupervised learning is used to process unlabeled data and classify or cluster by mining the internal structure of the data. In lithology identification, unsupervised learning is often used to discover patterns and regularities in data.

(1) K-means Clustering [17] : K-means is a clustering algorithm that maximizes the similarity of the data within the cluster by dividing the data into K clusters. In lithology identification, K-means can be used for preliminary classification to identify potential lithology types in the data.

(2) Principal component Analysis (PCA) [18] : PCA is a dimensionality reduction technique, which maps high-dimensional data to low-dimensional space through linear transformation and retains the main features of the data. In lithology identification, PCA can be used for data dimensionality reduction and computational complexity reduction.

3.3. Deep Learning

Deep learning is a branch of machine learning that learns and makes predictions by building multi-layered neural networks, and is particularly suited for working with complex, high-dimensional data. In lithology identification, deep learning methods perform well, which can automatically extract features and accurately classify them.

(1) Convolutional Neural Network (CNN) [19] : CNN is a kind of neural network specially used to process image data, and extracts local features of images through convolutional layer and pooling layer. In lithology identification, CNN can be used to analyze seismic image data and automatically

identify the formation characteristics of different lithology.

(2) Recurrent neural network (RNN) [20] : RNN is a kind of neural network used to process sequence data, memorizing and processing time series information through cyclic structure. In lithology identification, RNNS can be used to analyze log data and capture formation characteristics that change with depth.

(3) Generative adversarial Network (GAN) [21] : GAN is a deep learning model that learns against each other through generator and discriminator, which is used to generate realistic data samples. In lithology identification, GAN can be used for data enhancement to generate synthetic data to make up for the problem of insufficient data.

4. Methods Evaluation Indicators [22]

In machine learning-based lithology identification, it is very important to select suitable models and methods and evaluate them effectively. Different machine learning methods have advantages and disadvantages, and are suitable for different data types and task requirements. The following will describe in detail the criteria for comparison and evaluation of methods, common evaluation indicators and case studies in practical applications. When selecting and evaluating machine learning methods, there are several criteria that need to be considered:

When dealing with binary classification tasks, the confusion matrix is drawn as follows:

$$N = \begin{bmatrix} N_{TP} & N_{FP} \\ N_{FN} & N_{TN} \end{bmatrix}$$

Where: N_{TP} is the quantity classified as positive in the positive sample; $N_{FN}$is the number classified as negative in the positive sample; $N_{FP}$is the number of negative samples classified as positive; N_{TN} is the amount of the negative sample that is classified as negative.

(1) Accuracy refers to the proportion of the model that is correctly classified in all samples, and the specific expression is shown in Formula 1.

$$Acc = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (1)$$

(2) Accuracy rate refers to the proportion of all samples predicted by the model to be positive, which is actually positive. For specific expressions, see Equation 2.

$$Pre = \frac{TP}{(TP+FP)} \quad (2)$$

(3) Recall rate refers to the proportion of samples that the model successfully predicts to be positive among all samples that are actually positive, and the specific expression is shown in Equation 3.

$$Re = \frac{TP}{(TP+FN)} \quad (3)$$

(4) The F1 value synthesizes the performance of precision rate and recall rate, and its value range is [0,1]. The larger the value, the better the model performance and stronger generalization ability. For specific expressions, see Equation 4.

$$F_1 = 2\frac{Pre \times Re}{Pre+Re} \quad (4)$$

(5) ROC Curve and Area Under the Curve (Receiver Operating Characteristic Curve and Area Under the Curve) : The ROC curve evaluates the model performance by describing the true positive rate (TPR) and false positive rate (FPR) under different thresholds, while the AUC is the area under the ROC curve, and the closer to 1, the better the model performance.

(6) Confusion Matrix: The classification performance of the model is analyzed in detail by recording the matching between the predicted results and the actual results. The confusion matrix includes TP, TN (True Negative), FP, and FN.

(7)Computational Complexity: The time and resource consumption of model training and prediction, which is particularly important when dealing with large-scale data.

(8) Robustness: The resistance of the model to noisy data and outliers, which measures the stability of the model in practical applications.

(9) Generalization Ability: The performance of the model when dealing with unseen data, reflecting whether the model overfits the training data.

5. Main Challenges and Future Directions

Although machine learning has made remarkable progress in lithology identification, there are still many challenges. Addressing these challenges and exploring future directions will further improve the accuracy and usefulness of lithology identification. The main current challenges and future directions are discussed in detail below.

5.1. Main Challenges

(1) Data quality and quantity:

1) Data scarcity: In some areas, obtaining high-quality seismic and logging data can be very difficult, and data scarcity can limit model training and application.

2) Noise and outliers: Geological data often contain noise and outliers, and these inaccurate data will affect the performance of the model, requiring effective data cleaning and preprocessing methods.

3) Multi-source heterogeneous data fusion: seismic data, logging data and petrophysical experiment data come from different sources and in various formats, so how to effectively merge and utilize these heterogeneous data is a challenge.

(2) Model complexity and interpretability:

1) Training of complex models: When complex models such as deep learning process large-scale data, the training time is long and the computing resource consumption is large, which requires the support of high-performance computing platforms.

2) Black-box nature of models: Complex models, especially deep learning models, are often considered "black boxes" that have difficulty explaining their internal mechanisms and predictions. This is particularly important in lithology identification, as geologists need to understand the decision-making process of the model.

(3) Generalization ability and overfitting:

1) Overfitting problem: The model performs well on training data, but poorly on test data, indicating that the model may be overfitting. Effective regularization techniques and model validation methods are needed to improve the generalization ability of the model.

2) Diversity and heterogeneity: the geological conditions of different regions vary greatly, and the model needs to be able to handle diversity and heterogeneity to ensure applicability in different geological contexts.

(4) Real-time processing and application:

1) Real-time data processing: In practical applications, real-time processing and analysis of data is an important requirement, especially in oil and gas exploration and development, which requires immediate decision making and

response.

2) Deployment and maintenance: Deploying machine learning models into actual production environments and conducting ongoing monitoring and maintenance to ensure model stability and performance is a real challenge.

5.2. Future Development Direction

(1) Data enhancement and synthesis:

1) Data enhancement technology: Through data enhancement technology, such as random transformation, noise injection, etc., to expand the existing data set and improve the robustness of the model.

2) Generate adversarial network (GAN): Use GAN to generate synthetic data to make up for the problem of insufficient data and provide more samples for model training.

(2) Model interpretability and visualization:

1) Interpretative models: Develop models with high interpretative properties that enable geologists to understand and trust the decision-making process of the model. For example, interpretable Artificial Intelligence (XAI) technology is used to provide transparency in model decisions.

2) Visualization tools: Design intuitive visualization tools that combine model predictions with geological data to help geologists better understand and apply the results.

(3) Multi-source data fusion:

1) Cross-domain data fusion: Develop effective methods and algorithms to integrate multi-source heterogeneous data such as seismic data, logging data and geological maps to improve the accuracy and generalization ability of the model.

2) Data assimilation technology: Data assimilation technology is used to combine physical models and observed data to improve data integrity and consistency.

(4) Online learning and transfer learning:

1) Online learning: Develop online learning algorithms to enable the model to update and adapt to new data in real time, and improve the model's immediate response ability.

2) Transfer learning: Through transfer learning technology, the model trained in one region is applied to other regions to improve the adaptability and generalization ability of the model.

By addressing current challenges and exploring future directions, the application of machine learning in lithology identification will continue to improve, providing more powerful tools and methods for geological research and oil exploration.

6. Conclusion

(1) Machine learning has shown great potential and advantages in lithology identification. By utilizing a variety of machine learning methods, geologists can automatically extract features from complex seismic and logging data to achieve efficient and accurate lithology classification. These methods include supervised learning, unsupervised learning, and deep learning, each of which is suitable for different data types and task requirements, providing diversified solutions for lithology identification. Machine learning not only improves the accuracy and efficiency of lithology identification, but also reduces the workload of geologists, allowing them to focus on higher-level analysis and decision making.

(2)Although machine learning has made significant progress in lithology identification, there are still many challenges, such as limitations in data quality and quantity, problems with model complexity and interpretability, generalization capabilities and risks of overfitting, and the need for real-time processing and application. Addressing these challenges requires a combination of technologies such as data enhancement, model interpretability, multi-source data fusion, online learning, and transfer learning to develop more accurate and interpretable models.

(3) In the future, the application of machine learning in lithology identification will continue to expand and deepen. Data enhancement and synthesis technologies, interpretive models and visualization tools, multi-source data fusion, online learning and transfer learning will be the main development directions in the future. These directions will not only help overcome current challenges, but will also drive innovative applications of machine learning in lithology identification. With the continuous progress of technology and the accumulation of practical experience, machine learning will certainly play a more critical role in lithology identification, providing powerful and efficient tools for geological research and petroleum exploration, and helping the efficient development and utilization of resources.

References

[1] Wu Quande, Ma Zhizhong, Guo Keyi, et al. Intelligent lithology identification method of tunnel surrounding rock based on machine learning [J/OL]. Roadbed Engineering,1-6[2024-07-22].

[2] CHENG Guojian, Guo Wenhui, Fan Pengzhao. Rock image classification based on convolutional neural networks [J]. Journal of Xi 'an Shiyou University (Natural Science Edition),2017,32(04):116-122.

[3] Jiang Li, Zhang Zhimo, Wang Qiwei, et al. Comparative study on lithology classification of petroleum logging data based on different machine learning models [J]. Geophysical and Geochemical Exploration, 2024,48(02):489-497.

[4] Wang Xinling, Zhu Xinyi, Zhang Hongbing, et al. Lithology identification method of LWD based on random tree embedding [J]. Journal of Jilin University (Earth Science Edition), 2024,54(02):701-708.

[5] WANG Jiao, WANG Chenbai, Tan Zhenkun, et al. High order radial vortex beam superposition OAM pattern recognition method based on convolutional neural networks [J/OL]. Science in China: Physics, Mechanics and Astronomy, 1-8 [2024-07-22].

[6] Fang Dazhi, Ma Wejun, Yan Xu, et al. Lithology identification based on wavelet noise reduction and artificial intelligence [J]. Well Logging Technology,2023,47(04):438-446. (in Chinese)

[7] Wang Qi, Yang Tianwei, Liu Yongzhen, et al. Lithology identification of complex carbonate rocks based on random forest algorithm [J]. Chinese Journal of Engineering Geophysics,2020,17(05):550-558.

[8] Chen Weizheng. Application of neural network to lithology identification in sandstone type uranium mine logging [J]. Heilongjiang Science, 2024,15(12):8-12.

[9] Zhou Yuankai, Liu Hu. Research on lithology identification of logging based on deep learning method [J]. Uranium Geology,2024,40(02):336-345.

[10] Gao Ya Tian, Yang Junguo. Research on lithology identification method based on PSO-BP [J]. Computer and Digital Engineering, 2024,52(04):1119-1124. (in Chinese)

[11] Zhao Ranlei, Yang Liushuan, Xu Xiao, et al. Lithology identification of volcanic rocks based on XGBoost algorithm [J/OL]. Progress in Geophysics,1-12[2024-07-22].

[12] Chen Ganghua, Zhang Yuxia, Wang Jun, et al. Application of bidirectional long and short time memory neural network in reservoir lithology identification of beach bar sand [J]. Well Logging Technology,2023,47(03):319-325.

[13] Tong Rongchao. Application of machine learning in lithology intelligent identification [J]. Chemical Minerals and Processing, 2022, 51(08):43-47+54.

[14] Wang Xinling, Zhu Xinyi, Zhang Hongbing, et al. Based on random tree embedded while drilling logging lithology recognition method [J]. Journal of jilin university (earth sciences), 2024, (02): 701-708.

[15] Huang An, CAI Wenyuan, Wei Xinlu, et al. Lithology identification of volcanic logging based on improved random forest [J]. Science Technology and Engineering, 2023,23 (09): 3696-3704.

[16] Dong Wenhao, Zhang Huai. Lithology identification of cuttings based on transfer learning [J]. Journal of University of Chinese Academy of Sciences,2023, 40 (06): 743-750.

[17] Gao Chuqiao, Zhan Wang, Zhao Bin, et al. Lithology identification of igneous rocks based on hierarchical decomposition, principal component and Gaussian mixture clustering [J/OL]. Journal of Yangtze University (Natural Science Edition),1-12[2024-07-22].

[18] Bi Wenyi, Zhang Jinyang. Dense based on principal component analysis of glutenite reservoir lithology recognition method [J]. Science and technology, the wind, 2021, (7): 106-107.

[19] Sun Junyang, Fu Yunlai, Lv Jing, et al. Research on counting method of sea cucumber seedlings based on improved YOLOv7 model [J/OL]. Computer Technology and Development, 1-7 [2024-07-22].

[20] Guo Bin, Liu Zhao, Zhu Mingang. Research on embankment settlement prediction method based on EMD-RNN [J/OL]. Water Resources and Hydropower Letters,1-9[2024-07-22].

[21] Liu Y, Dan B, Yi C C, et al. Research on gear fault classification method based on improved SAGGAN model [J/OL]. Mechanical and Electrical Engineering,1-11[2024-07-22].

[22] Tian Tian, Cheng Zhiyou, Ju Wei, et al. Study on small sample classification of tea diseases based on SimAM-ConvNeXt-FL [J]. Transactions of the Chinese Society for Agricultural Machinery, 2024,55(03):275-281.