WEP Warwick Evans Publishing

International Journal of Natural Resources and Environmental Studies

ISSN: 3006-2012 (Print), ISSN: 3006-0834 (Online) | Volume 3, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijnres.v3n1.17
Journal homepage: https://wepub.org/index.php/IJNRES/index

Reservoir Permeability Prediction Based on Machine Learning

Qi Wang[1, 2]

[1]School of Earth Sciences and Engineering, Xi'an Shiyou University, Xi'an, China
[2]Shanxi Key Laboratory of Petroleum Accumulation Geology, Xi'an Shiyou University, Xi'an, China

ABSTRACT

This paper discusses the challenges and solutions of carbonate reservoir permeability prediction. It is difficult to predict the permeability of carbonate reservoir because of its complex pore structure and heterogeneity. Although NMR logging is useful for characterizing pore structure, its high cost limits its wide application. Therefore, in this paper, machine learning technology and petrophysical logging data are used to compare the permeability prediction models. By constructing and testing permeability prediction models based on artificial neural network (ANN), decision tree (DT) and support vector machine (SVM), the advantages and limitations of various methods are analyzed, which provides new tools and methods for oil and gas exploration and development.

KEYWORDS

Machine learning; Permeability prediction; Carbonate reservoir.

1. INTRODUCTION

Carbonate rocks account for 20% of the total sedimentary rock area in the world, and the oil and gas reservoirs contained therein account for a large proportion of the world's total oil and gas reserves, which has important exploration and research value. Carbonate reservoirs are generally affected by lithology, structure and karst process. Diagenesis is complicated, and the reservoir space formed mainly includes dissolution pores, caves and fractures[1]. Due to the strong heterogeneity, irregularity and multi-scale pore structure in carbonate rock, the characteristics of reservoir seepage are very different and the development is difficult. Therefore, it is of great significance to the exploration and development of oil and gas fields whether the microscopic pore structure of such reservoirs can be accurately and comprehensively characterized.

Permeability is a key physical parameter controlling fluid flow in reservoirs, especially in carbonate reservoirs, but it cannot be measured directly by conventional logging. Core measurements are available, but are limited and expensive. Production and pressure accumulation data can be used for permeability estimates, but only for specific formations. Nuclear magnetic resonance (NMR) logging is used to characterize pore structure properties and help calculate permeability more accurately[2-4], but it is also not commonly used due to its high cost. Predicting permeability from porosity and conventional logging is widely studied and used in practice, but it is not easy to obtain good results with high precision. Traditional wireline logging, such as gamma ray logging, acoustic logging, and resistivity logging, can only measure the radioactive, acoustic, and resistivity properties of subsurface formations[5]. These characteristics are not directly related to the pore structure characteristics. Permeability prediction based on porosity and traditional wireline logging depends on more accurate petrophysical or mathematical models[6]. To improve the accuracy of permeability predictions, different approaches have been proposed, including many physics-based petrophysical models and data-driven machine learning. Machine learning models based on permeability predictions are based

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

on measurement data rather than petrophysics[7]. Machine learning can automatically fit nonlinear relationships between outputs and inputs without prior knowledge. In this paper, different models for predicting reservoir permeability are compared by combining machine learning concepts with petrophysical measurements. The machine learning method was built and tested with samples of precisely recorded data from the X reservoir. In addition, other intelligent methods are compared, including methods such as artificial neural networks, XGBoost, and support vector machines.

2. THEORETICAL METHOD ANALYSIS OF MACHINE LEARNING

2.1. Artificial neural network

The design of an artificial neural network (ANN) is inspired by the structure of the human brain and consists of simple, small processing units - artificial neurons or nodes - that work together to perform calculations. This unique structure makes ANN an effective mathematical tool for a variety of goals, including but not limited to function estimation and pattern recognition. The structure of an ANN usually consists of three main layers: input layer, hidden layer, and output layer[8]. The number of neurons in the input layer is equal to the input variable, while the number of neurons in the output layer is matched to the output variable. In between these two layers, there is at least one hidden layer that acts as the computing engine of the network, processing signals and assisting in making predictions. Bias and weight, as the main parameters in ANN, determine the degree of freedom of the network and the relationship between the interconnected neurons in the layer, respectively. The optimization of these parameters is a key step in the ANN training process.

In ANN, each node except the input node determines its output through a nonlinear activation transfer function, and this output is passed as the input to the next node, and so on layer by layer, until a reliable solution to the original problem is reached. Throughout the training process, ANN learns from the existing experimental data to estimate the unknown data and optimizes its internal bias and weight so that the target value of the output layer is as close as possible to the corresponding experimental data.

In petroleum engineering applications, multi-layer perceptrons (MLPS) are a common type of artificial neural network that is trained by backpropagation (BP) methods. Compared with traditional data processing methods, MLP network has significant advantages, its training time is greatly reduced, and it can efficiently process and extract relevant information. In practice, measured permeability/porosity data are usually divided into two groups: a training set and a test set. The training set is used to adjust and optimize the bias and weights of the network, while the test set is used to verify the performance of the network after training. Through continuous iteration and optimization, ANN is able to build a complex relationship model between inputs and outputs to support decision making in petroleum engineering.

2.2. Decision tree

The decision tree (DT) is a common machine learning method that is often used to form classifiers and predictive models. As the name suggests, decision trees make decisions based on tree structures. A decision tree consists of a root node, several internal nodes and several leaf nodes[9]. The root node contains the complete set of samples. Each internal node corresponds to an attribute test, the leaf node corresponds to the decision result, and the path from the root node to each leaf node corresponds to a decision test sequence.

The decision tree usually adopts a top-down design method. Each iteration cycle, a feature attribute is selected for bifurcation until it can no longer be bifurcated. Therefore, in the process of building decision trees, it is very important to select the best bifurcation feature attributes (which can not only

classify quickly, but also minimize the size of decision trees). This "optimality" can be measured by purity, which is mainly measured by information gain, information gain rate, Gini index and so on.

The advantages of decision tree in classification and regression prediction are as follows: first, the complexity of model calculation is not high; Second, they are not sensitive to missing data; Third, compared with the traditional way, it is closer to the way people make decisions. Its disadvantages are: first, it is difficult to predict the continuity field; Second, it is easy to overfit.

2.3. Support vector machine

Support Vector Machine (SVM) is a machine learning method based on VC dimension theory and structural risk minimization theory. When the sample size is small, SVM can find a balance between model complexity and learning ability to obtain good generalization ability. When solving linear regression problems, it minimizes the sum of distances from all sample points to the optimal hyperplane[10]. SVM solves the regression problem by mapping the sample to a linearly separable high-dimensional space using kernel functions. Support vector machine regression (SVR) uses the kernel function K (x, x') to transform a nonlinear regression problem in a lower n-dimensional feature space into a linear regression problem in a higher P-dimensional feature space: RN→RP. Kernel functions determine the architecture of the SVR model. Polynomial functions and radial basis functions are widely used as nonlinear kernels:

$$K_{poly}(x, x') = (ax \cdot x' + c)^d \tag{1}$$

$$K_{rbf}(x, x') = -a||x - x'||^2 \tag{2}$$

Where x, x' belong to RN, and a, c, and d represent constant parameters. We take the polynomial function Kpoly with c=0 and d=3 as the core of the SVR model for permeability prediction. SVR attempts to find a regression plane in a high-dimensional feature space so that all samples are closest to that plane. To achieve SVR for permeability prediction, we also need to find the best penalty parameters to control the balance between useful information and noise to prevent the model from overfitting. Based on parameter experiments, this paper selects cubic polynomial kernel and penalty parameter 2 as the SVR model for permeability prediction.

2.4. Random forest

Random Forest (RF) based on decision tree is one of the classical integration methods. It can be used for both regression and classification. Decision tree is a non-parametric machine learning method. It learns simple rules to divide the entire data set into several sub-data sets with similar characteristics. A split subdata set is called a branch. Each branch generates the same prediction. RF consists of many unrelated decision trees in a random manner, each of which learns and predicts independently. The final prediction averages the predictions of each decision tree to make better decisions. The complexity of random forest is controlled by the number of decision trees, the minimum number of segmentation samples and the maximum depth of decision trees[11]. These parameters should be chosen carefully to prevent over-fitting of the RF model. For regression problem, the prediction result of random forest is the average of the output of all regression decision trees. Its definition is as follows:

$$\mathrm{f}(\mathrm{x_i}) = \frac{1}{\mathrm{N_{tree}}} \sum_{\mathrm{i=1}}^{\mathrm{N_{tree}}} \mathrm{h_i}(\mathrm{x_i}) \tag{3}$$

Through parametric experiments, RF for permeability prediction utilizes 300 decision trees as independent regressors. The minimum number of split samples for each decision tree is 4. The maximum depth of each decision tree is 6.

2.5. K nearest neighbor

K-nearest Neighbor (KNN for short) algorithm is a supervised machine learning algorithm with mature theory and simple structure. Its idea is: in a feature space, if most of the K Nearest samples near a sample belong to a certain category, the sample also belongs to that category. For regression problems, the common decision-making methods include simple average method and weighted average method[13,14]. Simple average method is to use common arithmetic average algorithm for k nearest neighbor target values. The weighted average method is the weighted average of k nearest neighbor target number values considering the distance difference. In practical application, the method of cross-validation is usually used to select the optimal k value. There are many distance measurement methods in the KNN model, such as Euclidean distance, Manhattan distance, Chebyshev distance, etc., that is, for the n-dimensional eigenvector xi,xj has:

$$L_{p(x_i,x_j)} = (\sum_{l=1}^{n} |x_i^{(l)} - x_j^{(l)}|^p)^{\frac{1}{p}} \tag{4}$$

In the above formula, when p=1, is Manhattan distance; When p=2, it is Euclidean distance; When p goes to infinity, it is the Chebyshev distance.

The advantages of K-nearest neighbor regression in classification or regression prediction are as follows: first, it is easy to implement and does not require parameter estimation or training; Second, the training time complexity of the model is low. Third, it is suitable for the case of relatively large sample size. Its disadvantages are: first, when the number of features is very large, the calculation amount will be large; Second, when the sample is unbalanced, the prediction accuracy of rare category is low. Third, the method uses lazy learning method, and the prediction speed is relatively slow.

3. PERMEABILITY PREDICTION MODEL BASED ON MACHINE LEARNING

Permeability prediction can be viewed as a supervised regression problem with one output and several input features. Given n input features x ={x1, x2,... , xn} and output permeability y, the machine learning model f fits the permeability regression problem y = f(X). X and y represent m records of X(dimension n) and y(dimension 1). The prediction pattern from input characteristics to permeability is contained in the machine learning model f. We summarize the following steps for building a machine learning model for permeability prediction:

(a) Data analysis and preprocessing. Training data sets are the key to building reliable data-driven machine learning models. Input feature x should be correctly selected and normalized. The statistical characteristics of the data set also help in the subsequent model training.

(b) Model training. Machine learning models are trained on training data sets. The training process can be regarded as the minimization of the error between the predicted permeability y and the measured permeability y, that is, the loss function loss(y,$\hat{y}$). Different machine learning models are trained separately on the same training data set.

(c) Model evaluation. A trained machine learning model is tested on a test data set to evaluate its performance. Based on their performance, the best permeability prediction model can be determined.

In order to explore the accuracy of different machine learning methods for permeability prediction, we built a variety of machine learning models for permeability prediction. Train and test different machine learning models separately. Machine learning model performance is evaluated on training and test datasets.

3.1. Data preparation and preprocessing

To model the permeability in this study, data preprocessing was first performed to detect and remove outlier data and select more features that affect the permeability.

Including irrelevant features in the model input data reduces accuracy and leads to the generation of models based on these features, which can affect the application of such poorly developed models. The criterion for feature selection is the numerical value of the feature weight evaluated by the feature selection algorithm. First, the feature selection algorithm should calculate the weights for each feature, and then compare the weights for all features to the thresholds. Finally, the feature whose weight is higher than the threshold value is selected for modeling, and the rest of the features are discarded[15,16]. In this paper, Pearson Pearson correlation coefficient, also known as Pearson product difference correlation coefficient, is a way to measure vector similarity. The output ranges from -1 to +1, where 0 means no correlation, negative values mean negative correlation, and positive values mean positive correlation. This method is suitable for judging the correlation between two continuous columns of data, and Pearson correlation coefficient is more suitable for high dimensions. Pearson correlation coefficient is calculated as follows:

$$\text{Pearson} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-y)^2}} \tag{5}$$

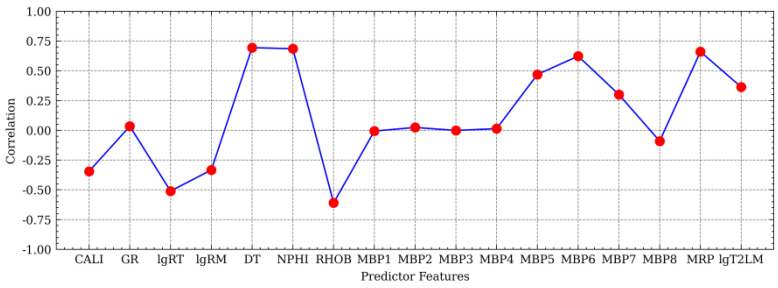


Figure 1. The pearson correlation coefficient calculates the data correlation

The permeability prediction model is established by using the porosity and permeability measured by the core and the well logging data of the xxx carbonate gas reservoir. The xxx reservoir is located in xxxx. The xxx formation in this area is composed of xxxx. Due to the influence of sedimentary, diagenetic and structural factors, the pore and pore structure characteristics of xxx are highly heterogeneous. Permeability is controlled by porosity and pore structure. Pore structure is an important factor affecting permeability. The porosity and pore structure characteristics are related to various rock and fluid properties. Rock minerals, pores and fluid composition can be fully measured with geophysical wireline logging tools. In addition to the porosity measured by the core, gamma ray (GR), acoustic wave (AC), density (DEN), compensated neutron log (CNL), deep resistivity (RD) and shallow resistivity (RS) can also be used as input features for permeability prediction to obtain more information about pore structure characteristics.

The range of input eigenvalues varies widely, for example, AC ranges from 62.1 μs/ft to 82.2 μs/ft and DEN ranges from 2.24 g/cm3 to 2.66 g/cm3. For machine learning modeling, each input feature is normalized to [0,1] with the following formula:

$$x_{id} = \frac{x_{i-}x_{i\,min}}{x_{i\,max}-x_{i\,min}} \tag{6}$$

Where, xi represents the original eigenvalue, xid represents the normalized dimensionless eigenvalue, and ximax and ximin represent the maximum and minimum values. RS and RD are normalized after logarithmic transformation. The permeability is also logarithmically transformed to be used as an output for the machine learning model.

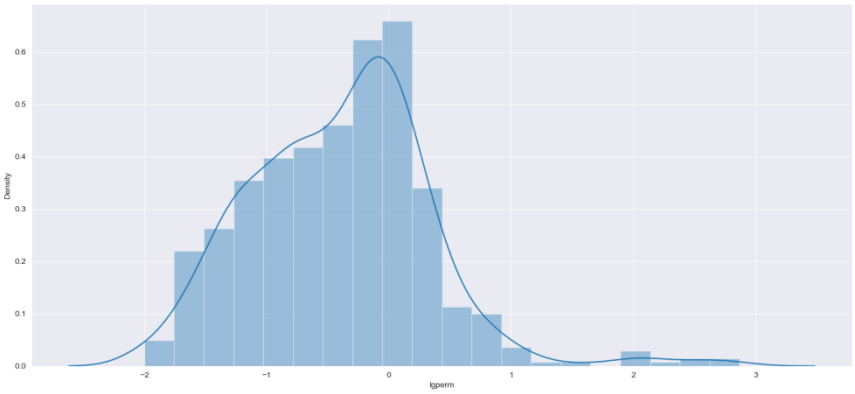
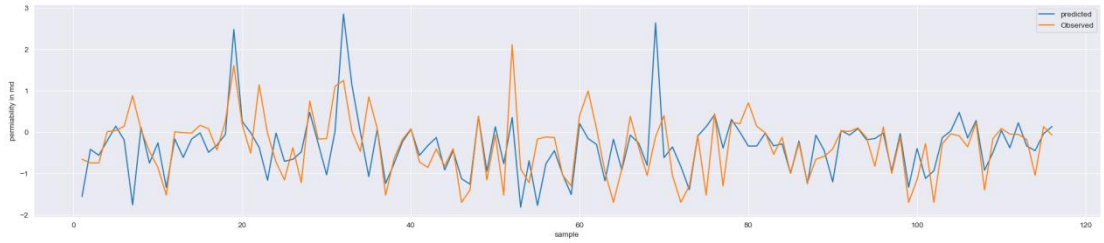


Figure 2. Statistical sample data distribution, data preprocessing

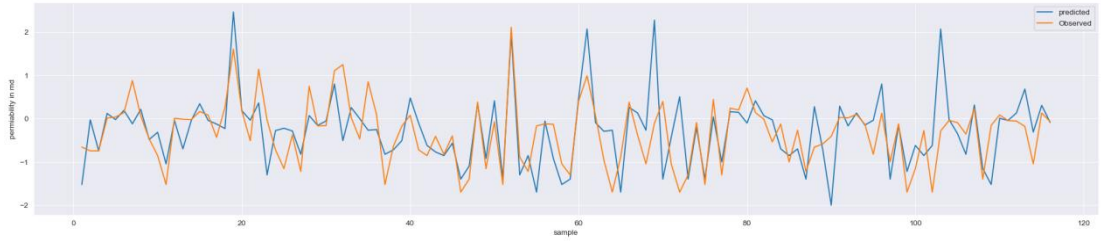
3.2. Model training and evaluation

The entire dataset of 300 samples from X Well is divided into a training dataset and a test dataset. Because of the high heterogeneity between Wells and the lack of data, the data sets are not divided by well, but randomly. 30% of the entire data set is randomly selected for model testing and the rest for model training. The machine learning model is trained on the training data set, while the test data set is only used for the final performance evaluation of the trained machine learning model. Different machine learning models are trained and tested separately on the same training and test data sets.

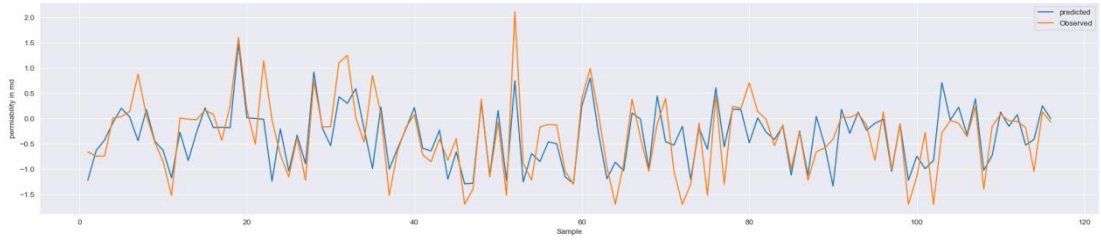
The data collected in this study include petrophysical logs (depth, diameter, gamma, resistance, neutron porosity, density, etc.). The correlation coefficient was calculated by pearson, and the eigenvalues were selected. At last, seven eigenvalues of "Depth", "GR", "CALI", "lgRM", "MBP1", "MBP2", "MBP3", "MBP4", "MBP8" and "lgperm" were selected as the input training data after data preprocessing. 30% of the entire data set is randomly selected for model testing and the rest for model training. Artificial neural network (ANN), decision tree (Decesion), support vector machine (SVM), Randomforest, K-nearest neighbor (KNN), reinforced dynamics(R_RID), eXtreme Gradient Boosting (XGB) and VOTE machine learning algorithms were used to train the model successively. The measured values and predicted values of rock permeability of various machine learning models are shown in Figure 3.


predicted
Observed
permeability in md
sample

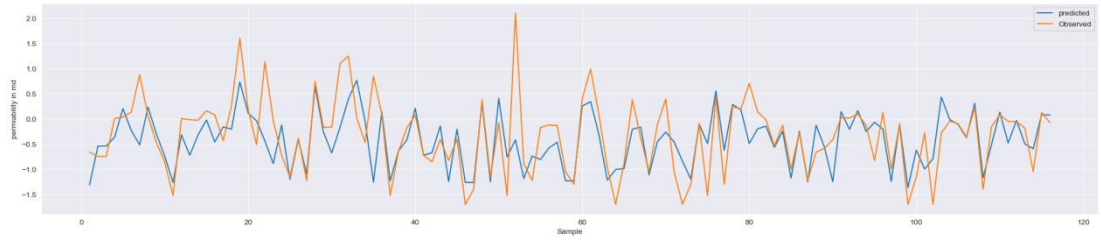
(a)


predicted
Observed
permeability in md
sample

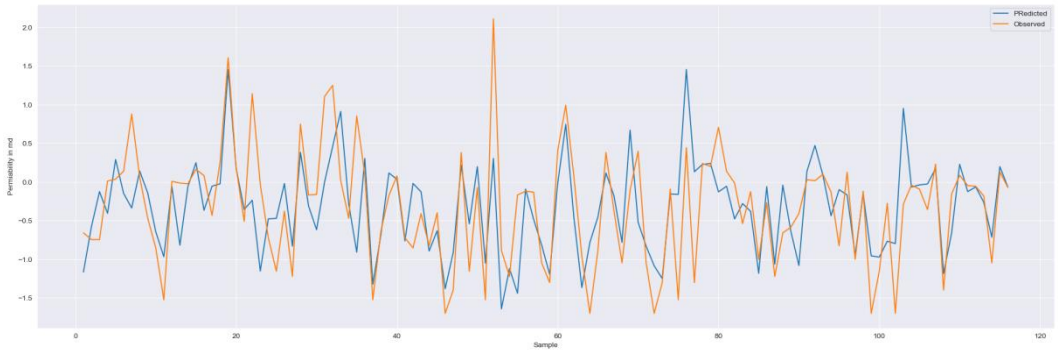
(b)


predicted
Observed
permeability in md
Sample

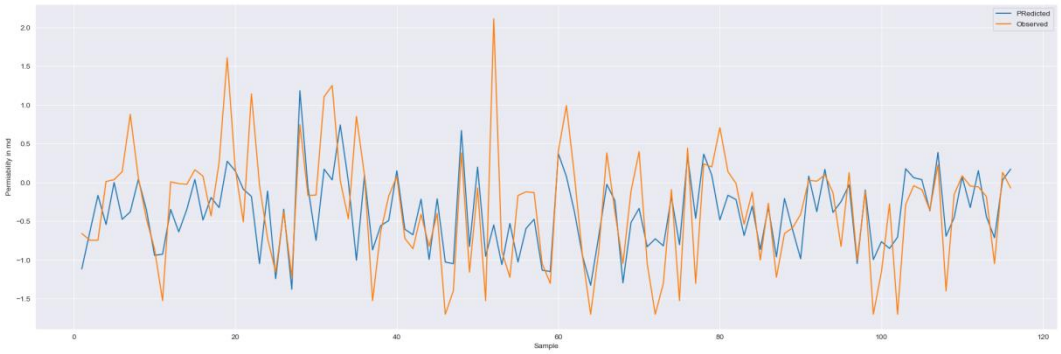
(c)


predicted
Observed
permeability in md
Sample

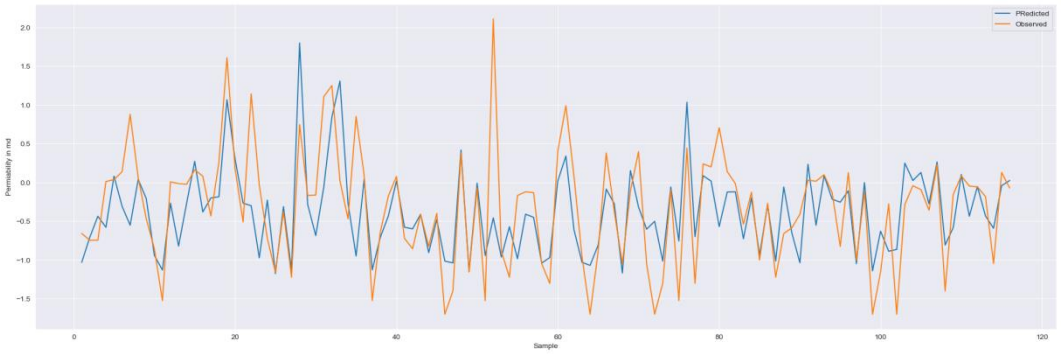
(d)


PRedicted
Observed
Permeability in md
Sample

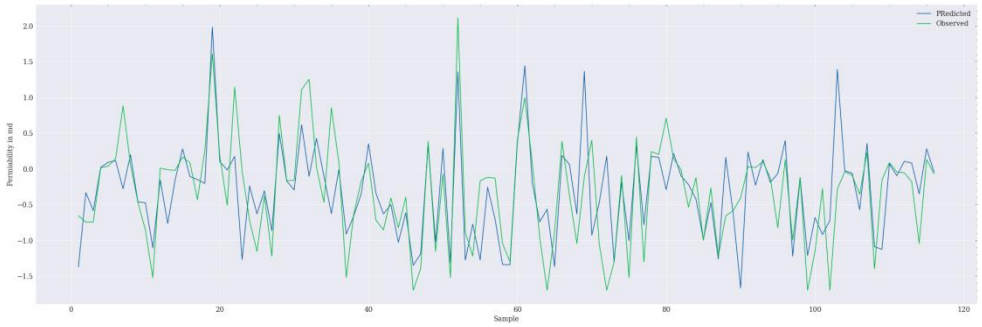
(e)



(f)



(g)



(h)

Figure 3. Comparison of different machine learning predictions: (a)Comparison of ANN predictions with raw data; (b) Comparison of Decesion predictions with raw data; (c) Comparison of SVM predictions with raw data; (d) Comparison of Random forest predictions with raw data; (e)Comparison of KNN predictions with raw data; (f) Comparison of R_RID predictions with raw data; (g) Comparison of XGB predictions with raw data; (h) Comparison of VOTE machine learning algorithms predictions with raw data.

We used determination coefficient (R2) and mean square error (MSE) to measure the performance of permeability prediction. R2 is defined as the proportion of the permeability variance that can be predicted by input features:

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{7}$$

Where yi is the true permeability, the predicted permeability, and the average value of the true permeability. The permeability used here is a logarithmic transformation. R2 is a statistical criterion used to show the variance ratio of a dependent variable described by one or more independent variables in a regression model. If R2=1, then the permeability of the rock is accurately predicted by the selected independent variable without generating errors. The higher the value, the lower the prediction error. The result evaluation of each machine learning model is shown in Figure 4. The error of artificial neural network algorithm (ANN) is the smallest, and the difference of random forest algorithm is larger.

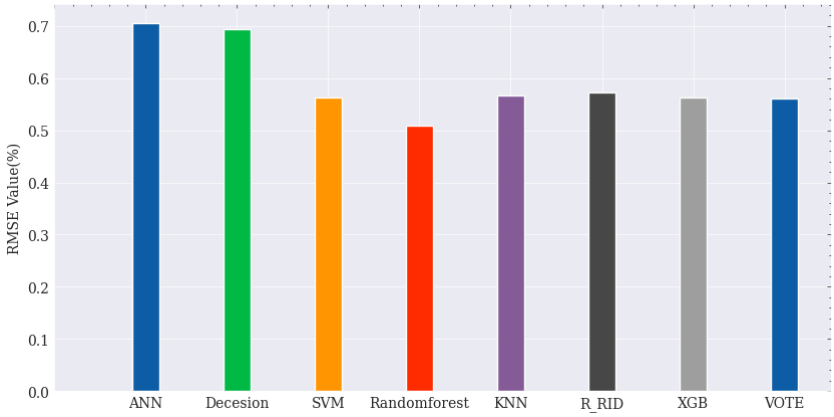


Figure 4. R2 values for different machine learning

4. SUMMARY

Numerous attempts have been made to provide an accurate and robust method for determining the permeability/porosity values of oil reservoirs. ANN, decision tree and other artificial neural networks are used to develop prediction models as robust and effective intelligent methods. In addition, precise actual porosity and permeability data from Field X were used to build the predictive model. Precise actual porosity and permeability data for the X field to determine the effectiveness and accuracy of prediction tools. Based on the results of this work, the following main inferences can be drawn:

(1) Comparisons between machine learning models show that the hybrid method can accurately predict the petrophysical properties of the reservoir.

(2) The data samples used in this paper were extracted from X oilfield. Therefore, the ability of this approach to be applied in different regions may vary depending on location, reservoir type, depth, heterogeneity, and other reservoir parameters.

REFERENCES

[1] Wang Hua, Zhang Yushun. Current situation and prospect of artificial intelligence processing and interpretation of logging data [J]. Well Logging Technology, 2021,45 (4): 345-356.

[2] YANG Y, SUN J, LI H, et al. ADMM-CSNet: a deep learning approach for image compressive sensing [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 42 (3): 521-538.

[3] LAM F, LI Y, PENG X. Constrained magnetic resonance spectroscopic imaging by learning nonlinear low dimensional models [J]. IEEE Transactions on Medical Imaging, 2019, 39 (3): 545-555.

[4] Zhang Cymbals, Zhu Jun, Su Hang. Towards the third generation of Artificial Intelligence [J]. Science in China: Information Science, 2020,50 (9) : 1281-1302.

[5] LUO S H, XIAO L Z, JIN Y, et al. A machine learning framework for low-field NMR data processing [J]. Petroleum Science, 2022, 19 (2): 581-593.

[6] HAMADA G M, ELSHAFEI M A. Neural network prediction of porosity and permeability of heterogeneous gas sand reservoirs [C]//SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition, Al-Khobar, Saudi Arabia, 2009.

[7] ZHOU X, SU G, WANG L, et al. The inversion of 2D NMR relaxometry data using L1 regularization [J]. Journal of Magnetic Resonance, 2017, 275: 46-54.

[8] SONG Y Q, KAUSIK R. NMR application in unconven tional shale reservoirs-a new porous media research frontier [J]. Progress in Nuclear Magnetic Resonance Spectroscopy, 2019, 112: 17-33.

[9] XIE R, WU Y, LIU K, et al. De-noising methods for NMR logging echo signals based on wavelet transform [J]. Journal of Geophysics and Engineering, 2014, 11 (3): 035003.

[10] GU M, XIE R, XIAO L. A novel method for NMR data denoising based on discrete cosine transform and variable length windows [J]. Journal of Petroleum Science and Engineering, 2021, 207: 108852.

[11] GE X, FAN Y, LI J, et al. Noise reduction of nuclear magnetic resonance (NMR) transversal data using improved wavelet transform and exponentially weighted moving average (EWMA) [J]. Journal of Magnetic Resonance, 2015, 251: 71-83.

[12] Xiao Lizhi. The integration and interpretability of machine learning data-driven and mechanism model [J]. Geophysical Prospecting for Petroleum, 2022, 61 (2): 205-212.

[13] LUO S, XIAO L, ZONG F, et al. Inside-out azimuth ally selective NMR tool using array coil and capacitive decoupling [J]. Journal of Magnetic Resonance, 2020, 315: 106735.

[14] GUO J, XIE R, XIAO L, et al. Nuclear magnetic resonance T1–T 2 inversion with double objective functions [J]. Journal of Magnetic Resonance, 2019, 308: 106562.

[15] PARASRAM T, DAOUD R, XIAO D. T2 analysis using artificial neural networks [J]. Journal of Magnetic Resonance, 2021, 325: 106930.

[16] Zhang Zhe, Liao Guangzhi, Xiao Lizhi, et al. Prediction of NMR T2 spectrum based on Machine Learning algorithm [J]. Science Technology and Engineering, 2023,23 (17): 7282-7292.