IOP Publishing Measurement Science and Technology
Meas. Sci. Technol. 31 (2020) 045101 (13pp) https://doi.org/10.1088/1361-6501/ab4a45

Fast prediction of reservoir permeability based on embedded feature selection and LightGBM using direct logging data

Kaibo Zhou[1], Yangxiang Hu[1], Hao Pan[1], Li Kong[1], Jie Liu[2,3,6], Zhen Huang[4] and Tao Chen[5]

[1] Key Laboratory of Image Information Processing and Intelligent Control of Education Ministry of China, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, People's Republic of China
[2] School of Hydropower and Information Engineering, Huazhong University of Science and Technology, Wuhan 430074, People's Republic of China
[3] Nondestructive Detection and Monitoring Technology for High Speed Transportation Facilities, Key Laboratory of Ministry of Industry and Information Technology, Nanjing 210016, People's Republic of China
[4] School of Electrical and Electronic Engineering, Wuhan Polytechnic University, Wuhan 430023, People's Republic of China
[5] China Petroleum Logging Co., Ltd, Xi'an 710077, People's Republic of China

E-mail: jie_liu@hust.edu.cn (J Liu)

Received 22 July 2019, revised 1 October 2019
Accepted for publication 2 October 2019
Published 9 January 2020

Abstract

Permeability estimation plays an important role in reservoir evaluation and hydrocarbon development, etc. Traditional physical model-based methods have problems with being time consuming and high cost. The applications of machine learning are currently becoming more and more extensive, however, there are still several limitations to previous machine learning-based permeability estimation methods, such as a limited number of samples, a requirement of prior knowledge, and some parameters needing to be calculated indirectly. In this paper, a hybrid reservoir permeability prediction approach, which is based on a certain scale of permeability dataset, embedded feature selection (EFS) and a light gradient boosting machine (LightGBM), is proposed. First, EFS is used to select features from the raw dataset. Then a LightGBM is adopted to predict the permeability. The influence of feature selection threshold, the base learners' number and dataset size on prediction results is also investigated. In addition, different feature selections and prediction models are compared. The proposed hybrid approach is also verified on other datasets. The experimental results show that the proposed approach can effectively predict the reservoir permeability based on limited direct logging data.

Keywords: permeability prediction, direct logging data, light gradient boosting machine, embedded feature selection, machine learning

(Some figures may appear in colour only in the online journal)

[6] Author to whom any correspondence should be addressed.

Nomenclature

ANN	Artificial neural network
EFS	Embedded feature selection
FFS	Filter feature selection
LightGBM	Light gradient boosting machine
MIR	Mutual information regression
RFE	Recursive feature elimination
RFR	Random forest regression
SVM	Support vector machine
WFS	Wrapper feature selection
XGBoost	Extreme gradient boosting

1. Introduction

Permeability is the characteristic of rock to allow fluids to pass through. It is a key parameter in assessing reservoir quality in petroleum engineering. If reservoir permeability can be accurately estimated, this is conducive to reservoir evaluation and production optimization, thereby decreasing production cost. However, due to the heterogeneity of reservoirs and their complex stratigraphic structure, it is a challenge to predict reservoir permeability accurately [1, 2]. Currently, conventional reservoir permeability prediction methods include core analysis, well test analysis and the empirical formula.

For example, as a common method, Darcy's law is often used to study fluid flow. It measures permeability by definition, e.g. the volume of fluid flowing through a unit section (cm^2) in unit time (s) under unit pressure (MPa). Feng *et al* studied gas multiple flow mechanisms and apparent permeability evaluation in shale reservoirs by Darcy's law [3]. Although Darcy's law is the most accurate method to directly measure the permeability of core samples, it has some disadvantages, such as being high cost, and time consuming [4, 5].

Another of the direct measurement methods, well test analysis, is a valuable method for measuring the dynamic response of a reservoir through determination of the hydraulic connectivity and effective permeability of the reservoir. Osorio *et al* used the method for geological interpretation of channelized heterolithic beds [6]. Chen *et al* used it to measure *in situ* stress, stress-dependent permeability, pore pressure and the gas-bearing system in multiple coal seams [7]. However, the method must be combined with others (such as stratigraphic comparison, log interpretation and core analysis) to comprehensively cope with more complex reservoir engineering problems.

The empirical formula is a series of nonlinear equations based on several physical models using core data. It can reasonably interpret the influences of formation parameters on reservoir permeability based on petrophysical theory. It has attracted the interest of researchers in the past few years [8–10]. The main idea is to investigate the relationship between the formation parameter and permeability through the statistical regression. Thereby, the physical model between the permeability and these parameters will be established. Several excellent ideas are to consider the influence of temperature, viscosity and liquid phase interaction. However, in addition to the parameters obtained by direct logging, the empirical formula method also requires stratigraphic parameters, such as porosity, shale content, particle size etc, obtained by combining other measurement methods. On the one hand, the physical model-based method cannot predict permeability effectively only using logging data. On the other hand, as the stratum is very complicated and difficult to describe, stratigraphic parameters are coupled and correlated. Thus, the empirical formula method cannot establish a very accurate model. In addition, a 2D scanning electron microscope image, or 3D computed tomography scanning image-based core reconstruction methods, called digital cores, also have been established for permeability studies [11, 12]. Finally, the fluid flow is simulated by the finite element analysis method to obtain flow parameters such as permeability. However, due to the high cost and low accuracy, it is not widely applied to practical production.

In recent years, intelligent computing methods have become more widely used in drug development [13, 14], medical treatments [15, 16] and other engineering fields [17–18]. At the same time, intelligent computing has also been successfully developed in the field of petroleum engineering. For example, Elkatatny *et al* used an artificial neural network (ANN) to predict bubble point pressure in oil reservoirs [19]. Ahmadi *et al* [20] and Menad *et al* [21] used a support vector machine (SVM) for enhancing oil recovery. Wang *et al* used data mining methods for pore structure prediction [22]. Meanwhile, Merembayev *et al* compared the effects of five intelligent computing methods on lithology classification [23]. Similarly, an ANN [24–26], SVM [27, 28], and other hybrid algorithms [29] have been applied to permeability prediction. For example, Saemi *et al* developed a neural network architecture optimized by genetic algorithms to predict permeability [24]. Elkatatny *et al* extracted a mathematical equation from the ANN model for permeability prediction [26]. Gu *et al* used a continuous restricted Boltzmann machine (CRBM), particle swarm optimization (PSO) and a SVM hybrid technique to predict permeability [28]. The CRBM is used to extract characteristics information from the original inputs. PSO is used to optimize the parameters of the SVM.

Compared with the statistical and classical regression methods, these intelligent computing methods can avoid complex physical models and directly establish the nonlinear relationship between input and output, but there are still some problems. For example, a certain prior knowledge is needed to select the relevant input features. Besides this, an ANN has a number of parameters that need to be adjusted, which takes significant computation time, and an SVM has no general solutions for nonlinear problems. Therefore, the current machine learning-based approaches have low prediction efficiency. Meanwhile, some of the input features of these intelligent computing methods require indirect measurement or additional calculation, such as array induced tool resistance, water saturation, neutron porosity, density porosity and reservoir type. This undoubtedly increases the prediction cost.

For the above problems, the feature selection methods can select input features without prior knowledge and improve prediction accuracy and efficiency. Among the common feature

selection methods, the embedded method inserts the feature selection step into the training process of the regression model and the selection of optimal feature subset can improve the performance of the regression model [30]. The amount of data available for permeability prediction is increasing, which will exert great influence on permeability prediction. Under such circumstance, a light gradient boosting machine (LightGBM) is applied to the permeability prediction. The LightGBM is an improved algorithm based on the gradient boosting decision tree (GBDT), which is an open-source, fast and efficient algorithm released by Microsoft Research Asia in 2016 [31]. It has been applied in the bioengineering [32] and financial industries [33]. To the best of the authors' knowledge, no one has investigated the effectiveness of LightGBM-based reservoir permeability prediction in reservoir evaluation. In this paper, a hybrid reservoir permeability prediction approach based on embedded feature selection (EFS) and the LightGBM (EFS–LightGBM) is proposed. Firstly, EFS is used for feature selection. Then LightGBM is adopted to predict the permeability of the optimal subset. Two statistical quality measurement methods, the coefficient of determination (R^2) and root mean squared error (RMSE), are used to evaluate the prediction performance of the model. The main contributions of this paper are as follows.

(1) Reservoir permeability can be predicted only using direct logging data, not indirect measurements or calculation parameters.
(2) EFS–LightGBM is first applied to select features without prior knowledge and improve accuracy and efficiency of permeability prediction.
(3) The effectiveness of the proposed hybrid reservoir permeability prediction approach is verified by the cross-hole permeability prediction in the same area.

The remaining content of this paper is organized as follows. Section 2 introduces the research background and data sources. Section 3 introduces the methodology. The experimental verification process is described in section 4. Section 5 discusses the comparison of the different methods and the validation of the proposed method on other datasets. Finally, conclusions are drawn in the final section.

2. Background

The main research content of this paper is to predict the permeability by the formation parameters obtained by logging. The heterogeneity of a reservoir's physical properties is caused due to the complex geological structures in the stratum, such as caves, faults and fractures, etc. The reliability and generalization ability of a model largely depends on the amount and type of data involved in the model training process. Therefore, a large and reliable database is needed.

The data used in this study was obtained from the logging data of well W1 in an oilfield in northwest China. The data contains 9314 data points, which have an interval of 0.125 m from 1625 m to 2790 m. The parameters of the data include: (1) 22 parameters obtained by direct measurements, such as depth (DEPTH), acoustic logging (AC), compensated neutron logging (CNL), density logging (DEN), fullbore formation microresistivity (FMIT), gamma ray (GR), and spontaneous potential (SP). (2) 6 indirectly measured parameters, such as array induced tool resistance (AT) and thorium potassium ratio (THK). (3) 25 indirectly calculated parameters, such as porosity from density (PERN), water filled porosity (PORW), and water saturation (SW). (4) Permeability (PERM). In this paper, 9297 data points were selected from the raw data to constitute the experimental dataset, containing permeability and 22 parameters obtained directly by logging.

The original reservoir data was obtained by instruments, such as double-emission and double-receiving compensation acoustic (used to obtain AC), natural gamma ray (used to obtain GR), compensation density (used to obtain DEN), formation microresistivity imager (used to obtain FMIT) and dual induction-eight lateral (used to obtain CLL8). These instruments are sub-modules of the Express and Image Logging System (EILog) developed by China Petroleum Logging Co., Ltd. The statistical descriptions of the parameters used are shown in table 1.

3. Methodology

3.1. The proposed approach

The flowchart of the proposed approach is shown in figure 1, where the input data is derived from the direct logging data obtained in section 2. First, the original data is imported to the initial model. Then the optimal feature subset is selected by EFS. Finally, the LightGBM model is used to predict permeability.

In this study, 75% of the dataset was used as the training set and the remaining proportion was the test set. In order to improve the generalization ability of the model and eliminate contingency influence in sample segmentation, four-fold cross-validation was used in all experiments. As a tree model, LightGBM is not sensitive to the range of data. In order to consider the influence of parameters on the model, the data was not normalized in this paper.

The prediction performance of the LightGBM model was evaluated by two statistical quality indicators: R^2 and RMSE. Due to the large numerical distribution intervals, the relative size of the data has a greater impact on the RMSE. Therefore, the R^2 score was considered. The two indicators are defined as follows:

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

where y_i is the real value of permeability, $\hat{y}_i$ is the predicted value of permeability, $\bar{y}$ is the mean value of permeability, and m is the amount of data. Therefore, the predictive effect is better when R^2 is closer to 1.

$$\mathrm{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2}, \tag{2}$$

Table 1. Statistical descriptions of the parameters used.

Identifier	Parameter	Symbol	Unit	Min.	Max.	Avg.	STDEV
1	Depth	DEPTH	m	1627	2789	2208	335.494
2	Acoustic logging	AC	$\mu s\ m^{-1}$	184.686	415.516	243.513	24.619
3	Azimuth of drill drift	AZIM	deg	0	359.868	110.401	112.810
4	Caliper	CAL	cm	20.865	41.134	22.674	1.078
5	*X* value of caliper	CALX	cm	20.298	41.292	22.633	1.041
6	*Y* value of caliper	CALY	cm	20.964	43.678	22.714	1.303
7	IL (induction log)-deep conductivity	CILD	$S\ m^{-1}$	3.308	8999.998	87.416	224.088
8	Conductivity of laterolog 8	CLL8	$S\ m^{-1}$	1.571	1237.462	68.498	54.226
9	Compensated neutron logging	CNL	%	6.632	88.743	23.949	7.460
10	Density logging	DEN	$g\ cm^{-3}$	1.326	2.707	2.496	0.143
11	Deviation	DEVI	deg	0.272	2.396	1.122	0.441
12	Fullbore formation microresistivity	FMIT	Ohm	0.001	18.594	0.288	0.850
13	Gamma ray	GR	API	8.011	532.692	104.087	28.474
14	Depth of magnetic marks	MMD	\	−7372.410	15 069.45	−11.513	920.662
15	MMD of CAL	MMDCAL	\	−9084.070	9999.695	−9.8138	989.508
16	MMD of DEN	MMDDEN	\	−3723.090	7555.332	9.905	344.495
17	Litho-density logging	PE	b/e	0.899	4.601	3.167	0.438
18	4 m resistivity	R4_0	Ohm	2.533	180.780	36.795	29.569
19	Mud filtrate resistivity	RMF	Ohm	0.346	0.622	0.483	0.079
20	True formation resistivity	RT	Ohm	0.111	302.253	28.775	34.379
21	Spontaneous potential	SP	mV	30.705	93.027	76.865	13.870
22	Corrected spontaneous potential	SPC	mV	−59.728	2.733	−13.630	13.824
23	Permeability	PERM	mD	0.010	57.490	1.293	3.520

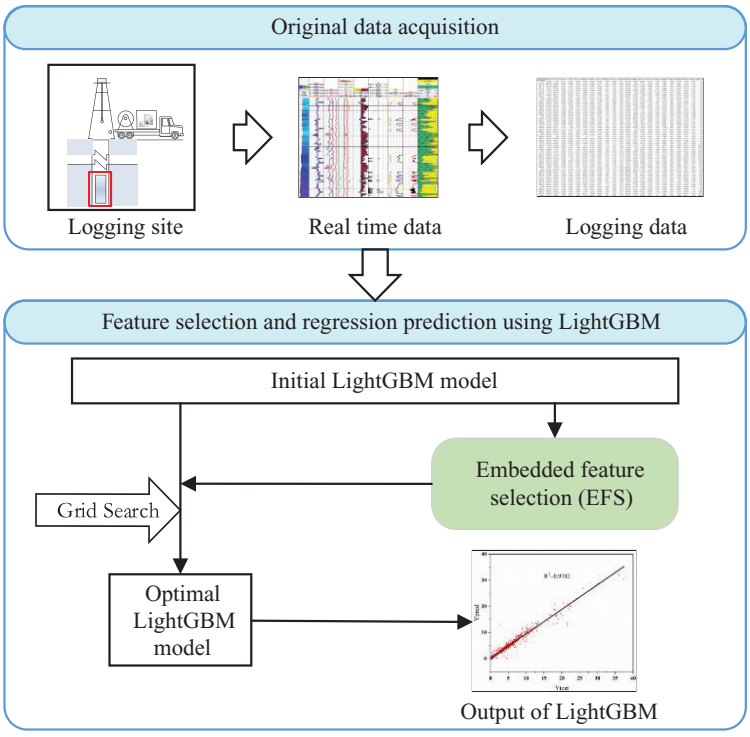

Figure 1. Flowchart of the proposed approach.

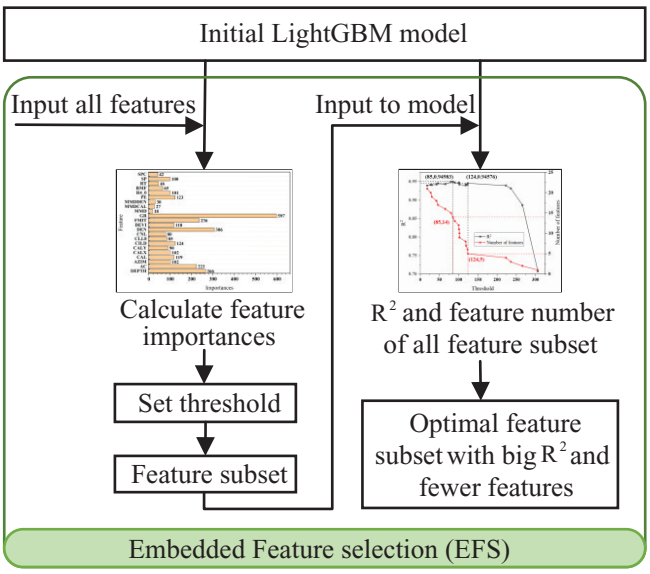

Figure 2. The flow of EFS.

where RMSE represents the weighted average of the deviation between the prediction value and the real value.

3.2. Embedded feature selection

In order to study the influence of input features on permeability prediction and obtain sufficiently regression effects with as few features as possible, it is necessary to reduce the dimension of the input features [34]. Feature selection and extraction are two common steps in feature dimension reduction. As extracted features from the latter lose the physical meaning of the original feature space and lack interpretability, the feature selection was adopted in this paper.

EFS used in this paper is to embed feature selection into the whole learning process. It allows the regression algorithm to determine features to use. That is to say, feature selection and model training are performed simultaneously.

When using the EFS, some machine learning algorithms or models are used for training. Then, the weight coefficients of each feature are obtained according to the regularization term or loss function of the model. Finally, the features are selected in order according to the weight coefficient. These weight coefficients usually represent a certain contribution or importance of a feature to the model and can be ranked. Based on the evaluation of this contribution, the most useful features for the model establishment can be selected. Moreover, considering the contribution of features to the model, the features related to correlation filtering and the features without discrimination for variance filtering will be deleted, due to lack of contribution. The flow of EFS is shown in figure 2.

3.3. LightGBM

Recently, XGBoost [35] and LightGBM [31] algorithms, which have been proposed based on the GBDT algorithm, have greatly improved prediction performance. As XGBoost adopts a pre-sorted algorithm in sorting, it is necessary to pre-sort all features according to the values. Meanwhile, the level-wise split method is used to traverse the same layer of leaves in the division of the sub-model, which leads to significant unnecessary search and split. The principle of the level-wise split method is shown in figure 3. The model therefore consumes a lot of time and space on large-scale data.

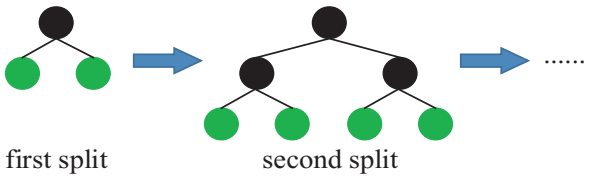

Figure 3. Principle of the level-wise split method.

To deal with these problems, a modified LightGBM is proposed as shown in figure 4. Using a histogram-based decision tree algorithm, the basic idea is to discretize continuous feature values into k integers, construct a histogram with a width of k, and then index the histogram according to the interval of the feature. Therefore, it does not need to sort according to each feature and compare the values of different features. It greatly reduces storage space and computational cost.

The choice is made to split the nodes of the weak learner leaf-wise. By controlling the depth of the tree and the minimum amount of data per leaf node, the over-fitting phenomenon is avoided. Therefore, it is not necessary to traverse the entire training data at each iteration, which leads to a small computational cost. The principle of the leaf-wise method is shown in figure 5.

The codes of the proposed EFS–LightGBM were written by Python 3.7.1 and run on the Jupyter Notebook (version 4.6.14) platform on a laptop with Core (TM) i7-8750H CPU and 16G RAM.

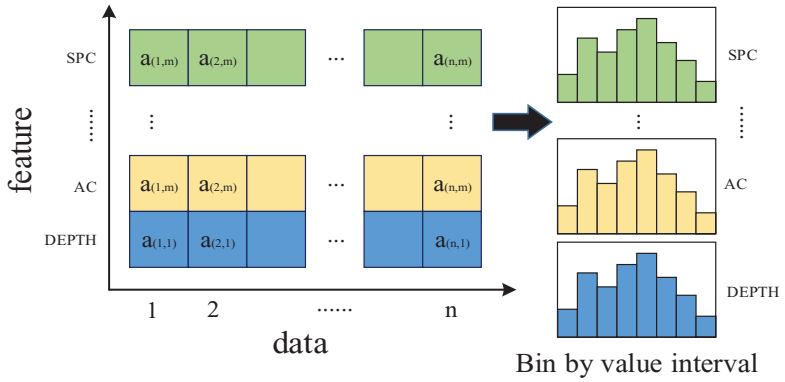

Figure 4. Principle of histogram-based decision tree algorithm.

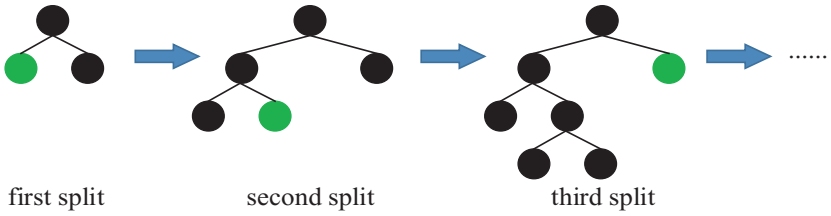

Figure 5. Principle of the leaf-wise method.

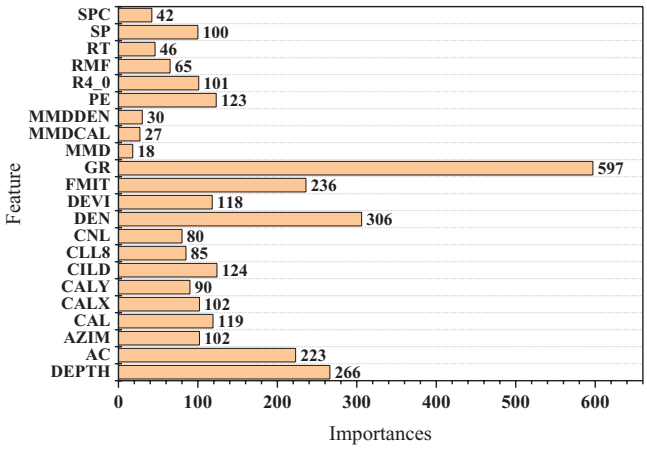

Figure 6. Contribution of each feature to the model.

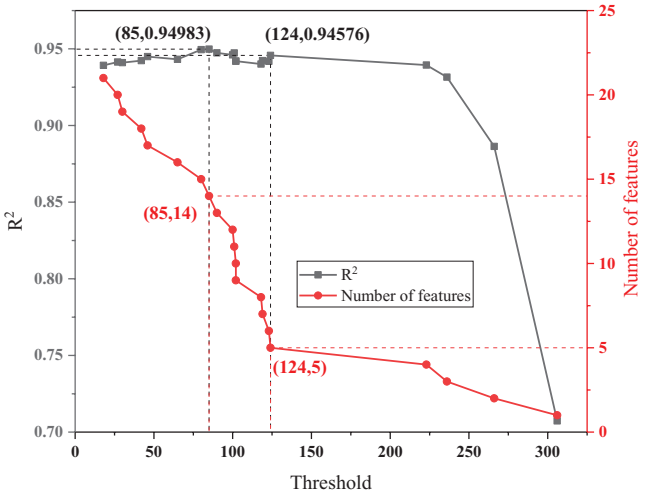

Figure 7. Relationship between the setting threshold and model performance and features number.

4. Experiments and analysis

4.1. Feature selection

In this section, the proposed EFS–LightGBM approach presented is used for feature selection and model evaluation. First, all the data are imported into the LightGBM model for training, thus the contribution value of each feature to the model (i.e., the importance of the feature) can be obtained through the model interface. The global importance of feature j is measured by the average of its importance in a single tree (base learner):

$$\hat{J}_j^2 = \frac{1}{N}\sum_{n=1}^{N}\hat{J}_j^2(T_n), \tag{3}$$

where N is the number of trees (base learners). The importance of feature j in a single tree is as follows:

$$\hat{J}_j^2(T) = \sum_{t=1}^{L-1}\hat{i}_t^2 1(v_t = j), \tag{4}$$

where L is the number of leaf nodes of the tree, $L-1$ is the number of non-leaf nodes of the tree, v_t is the feature associated with node t, and $\hat{i}_t^2$ is the squared loss reduction value after node t splitting.

Second, EFS is used to select the feature with the contribution value of the feature as the threshold. Therefore, the features whose contribution value are lower than the set threshold can be eliminated. Finally, the selected feature subset is used to import the LightGBM model to obtain the R^2 score of the model. The contribution of each feature to the model is shown in figure 6.

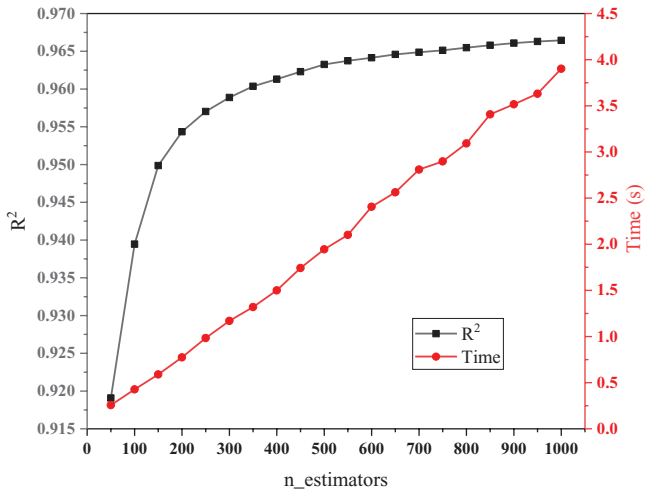

Figure 8. The effect of n_estimators on the model R^2 score and time consumption.

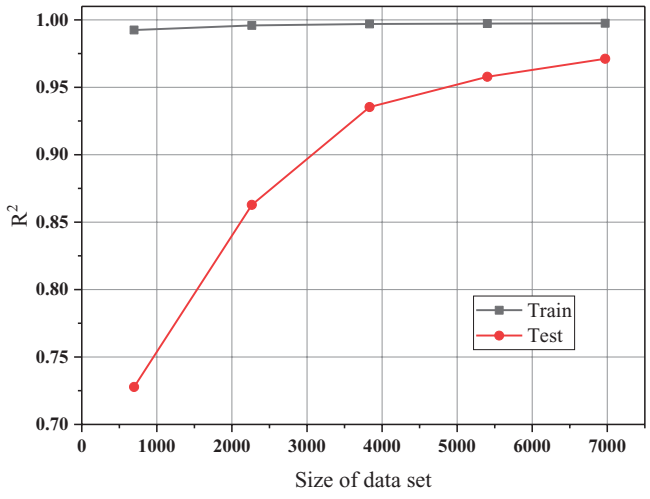

Figure 9. Effect of dataset size on R^2 score.

After sorting the contribution of the feature, the influence of the contribution threshold on feature selection and permeability prediction should be studied. The relationship between the setting threshold and model performance and feature number are shown in figure 7. When the threshold is 85, the R^2 score and feature number are 0.9498 and 14, respectively. The R^2 score is the highest at this time. When the threshold is 124, the R^2 score and feature number are 0.9457 and 5. Although the R^2 score is not the highest at this time, the feature number is reduced from 14 to 5, obtaining a similar R^2 score. Thus, the model performance can be improved by subsequent hyperparameters adjustment. After comprehensive consideration of the R^2 score and feature number, the threshold of this paper was set to 124, and the five features including DEPTH, AC, DEN, FMIT and GR were selected.

It can be seen from figure 7 that the R^2 score is lower when more features are input, and prior knowledge is not required to select useful features. Some of the features can also be regarded as noise. This proves that the EFS used in this paper has, to some extent, played a role in denoising.

Table 2. The hyperparameters of LightGBM.

Hyperparameter	Symbol	Value
Number of boosting iterations	n_estimators	500
Maximum depth of a tree	max_depth	15
Maximum number of leaves in one tree	num_leaves	16
Subsample ratio of columns when constructing each tree	colsample_bytree	1
L1 regularization term on weights	reg_alpha	2.1
L2 regularization term on weights	reg_lambda	0.01
Boosting learning rate	learning_rate	0.2

Table 3. The performance of the optimized model on the selected features.

Train		Test		Time (s)	Number of features
R^2	RMSE	R^2	RMSE		
0.9974	0.1787	**0.9712**	**0.5959**	1.37	**5**

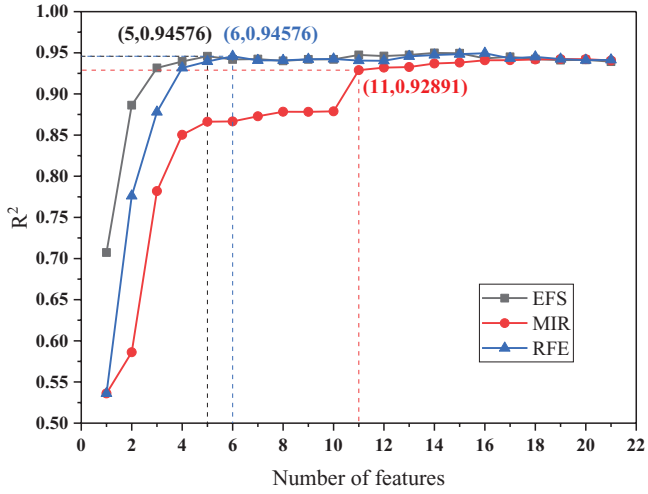

Figure 10. The relationship between the number of features selected by the three methods and the R^2 score.

4.2. Number of base learners

In the above, five selected features are input into the LightGBM model. Then, the hyperparameters are adjusted to further improve the performance of the model. Among all the hyperparameters, the number of base learners (denoted as n_estimators) is the most important, which directly determines the complexity and performance of the model.

The effect of base learner number on the R^2 score and time consumption are shown in figure 8. The increase of n_estimators leads to the increment of model complexity as well as the logarithmically increment of the R^2 (e.g. the increment is smaller and smaller), while the time consumption of the model increases linearly. Considering the real-time performance of logging evaluation, it is not necessary to select a larger n_estimators value for a higher R^2 score. To weigh the R^2 score and time consumption, n_estimators was set to 500, where the R^2 score is greater than 0.96 and the increment is less than 0.001.

Table 4. The features selected by the three methods.

	1	2	3	4	5	6	7	8	9	10	11
	DEPTH	AC	AZIM	CAL	CALX	CALY	CILD	CLL8	CNL	DEN	DEVI
MIR	√	\	√	\	\	\	√	\	\	√	\
RFE	√	√	\	\	\	\	√	\	\	√	\
EFS	√	√	\	\	\	\	\	\	\	√	\
	12	13	14	15	16	17	18	19	20	21	22
	FMIT	GR	MMD	MMDCAL	MMDDEN	PE	R4_0	RMF	RT	SP	SPC
MIR	√	√	\	\	\	√	\	√	√	√	√
RFE	√	√	\	\	\	\	\	\	\	\	\
EFS	√	√	\	\	\	\	\	\	\	\	\

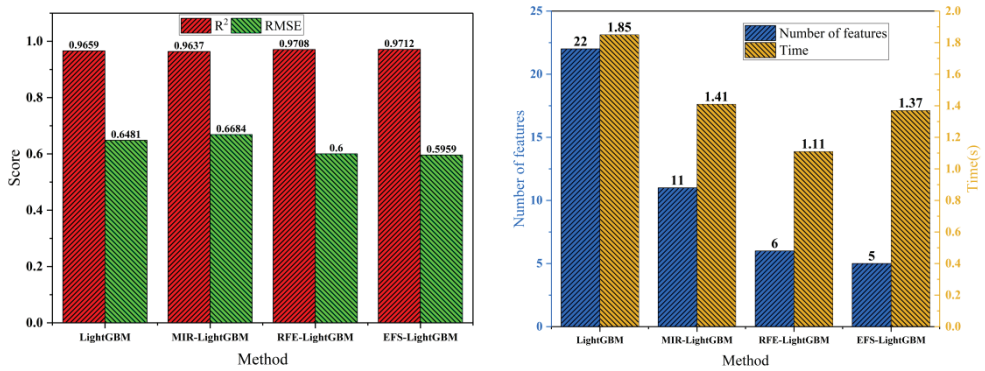

Figure 11. The performance of the four models: (a) the R^2 score and RMSE of the four scenarios; (b) the feature number and time of the four scenarios.

4.3. Dataset size

In order to explore the effect of dataset size on permeability prediction, part of the original dataset is randomly selected as a data subset. Then the model is trained under different data subsets to obtain the R^2 score of the corresponding training set and test set. The effect of the dataset size on R^2 score is shown in figure 9. The model performs well on the training set, and as the sample size increases, the performance on the test set is better. Meanwhile, improved performance can be obtained under the current overall sample size. Although increasing the sample size can improve the performance of the model to a certain extent, there is no need to obtain more data for a small increase.

As a common model optimization method, grid search is adopted to select the optimal hyperparameter by traversing within a certain range with a fixed step size. Based on the work above, grid search is used to optimize other hyperparameters of the model, thus best permeability prediction performance with maximize R^2 and minimize RMSE is achieved. That is, the performance of the model is optimal under this set of hyperparameters. Final model hyperparameters are listed in table 2. The performance of the optimized model on the selected features is shown in table 3.

5. Discussion

To further highlight the performance of the proposed method, the following three aspects are discussed in this section: different feature selection methods, different regression models, and performance on other datasets.

5.1. Comparison with other feature selection methods

Commonly used feature selection methods include filter feature selection (FFS) and wrapper feature selection (WFS) [36]. The evaluation function of FFS is independent and has no relationship with the regression algorithm. It directly evaluates the relationship between the features and then removes the data with low correlation and high redundancy according to the evaluation results. WFS requires subsequent regression calculations, improving the calculation cost.

In this section, the mutual information regression (MIR) in FFS and the recursive feature elimination (RFE) in WFS are compared with proposed EFS. The threshold of MIR is the mutual information quantity. As this paper focuses on the selected features and the predictive effects of the model, no specific mutual information values are given. The threshold of EFS is the contribution of the feature to the model (e.g. the importance of the feature as shown in figures 6 and 7). The threshold of RFE is the number of selected features (from 22 to 1). Although these thresholds are different, they can be converted into the relationship between the feature number and the R^2 score as shown in figure 10. The optimal feature numbers selected by MIR and RFE are 11 and 6, respectively. The selected features by these methods are shown in table 4. It can be seen that all these methods select DEPTH, DEN, FMIT and

Table 5. Detailed information for each indicator.

	Train		Test			
	R^2	RMSE	R^2	RMSE	Time (s)	Number of features
LightGBM	**0.9978**	**0.1645**	0.9659	0.6481	1.85	22
MIR–LightGBM	0.9970	0.1924	0.9637	0.6684	1.41	11
RFE–LightGBM	0.9976	0.1704	0.9708	0.6000	**1.11**	6
EFS–LightGBM	0.9974	0.1787	**0.9712**	**0.5959**	1.37	**5**

Table 6. Optimal hyperparameters of the corresponding model.

Hyperparameter	Symbol	LightGBM	MIR–LightGBM	RFE–LightGBM
Number of boosting iterations	n_estimators	500	500	500
Maximum depth of a tree	max_depth	13	6	6
Maximum number of leaves in one tree	num_leaves	15	16	29
Subsample ratio of columns when constructing each tree	colsample_bytree	1	0.64	1
L1 regularization term on weights	reg_alpha	0.75	1.65	2.46
L2 regularization term on weights	reg_lambda	0	0	0
Boosting learning rate	learning_rate	0.1	0.19	0.32

Table 7. Main hyperparameters for RFR.

Hyperparameter	Symbol	RFR	MIR–RFR	RFE–RFR	EFS–RFR
The number of trees in the forest	n_estimators	160	103	100	240
The maximum depth of the tree	max_depth	22	28	23	22
The minimum number of samples required to be at a leaf node	min_samples_leaf	1	1	1	1
The minimum number of samples required to split an internal node	min_samples_split	2	2	2	2
The number of features for the best split	max_features	12	12	3	3

Table 8. Main hyperparameters for XGBoost.

Hyperparameter	Symbol	XGBoost	MIR–XGBoost	RFE–XGBoost	EFS–XGBoost
Number of boosted trees to fit	n_estimators	200	200	200	200
Subsample ratio of the training instance	subsample	0.536 84	0.736 84	0.386 84	0.378 94
Minimum sum of instance weight (hessian) needed in a child	min_child_weight	1	1	1	1
Maximum depth of a tree	max_depth	9	8	8	11
Minimum loss reduction required to make a further partition on a leaf node of the tree	gamma	0.03	0	0.07	0
Subsample ratio of columns when constructing each tree	colsample_bytree	0.82	0.82	1	1
L1 regularization term on weights	reg_alpha	0.74	0.83	2.1	0
L2 regularization term on weights	reg_lambda	0.71	1	0.15	1.93
Boosting learning rate	learning_rate	0.1	0.1	0.1	0.1

GR. This means that these selected features are very important for reservoir permeability prediction. Furthermore, features such as AC and CILD also contribute to the permeability prediction.

Finally, the R^2 score, RMSE and the training time of the four schemes (LightGBM, MIR–LightGBM, RFE–LightGBM, and EFS–LightGBM) on the training and test set are obtained. The R^2 score and RMSE in the four schemes are shown in figure 11(a). It can be observed that the model with EFS has the highest score on the test set with an R^2 score and RMSE of 0.9712 and 0.5959, respectively. The feature number and time consumption in the four schemes are shown in figure 11(b). This shows that EFS has the least number of selected features and the training time of the model with RFE is the least, at 1.11 s. Detailed information of each indicator is given in table 5. The optimal hyperparameters of the corresponding model are listed in table 6. Although the hyperparameters of different algorithms are different, the optimal results of each algorithm are compared.

In fact, for the training time, the models with three feature selection methods have little difference. In this study, the main consideration is obtaining a better prediction effect with

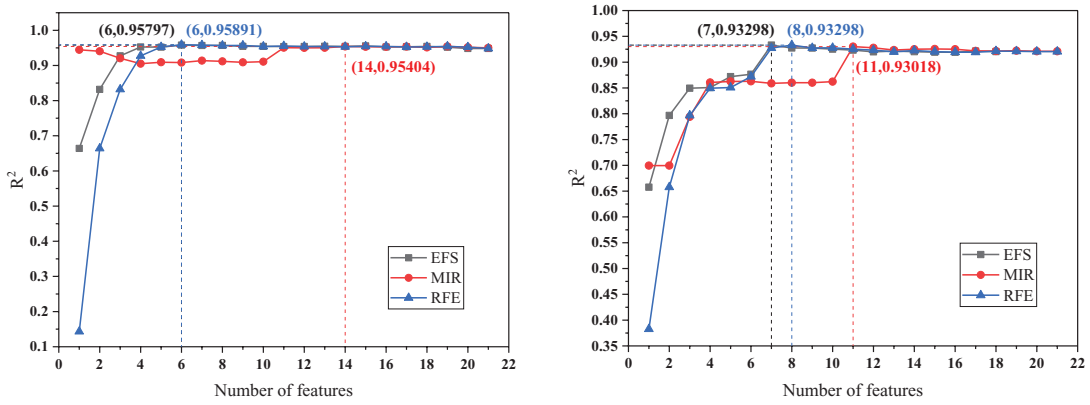

Figure 12. Comparison of different feature selection methods: (a) comparison based on RFR; (b) comparison based on XGBoost.

Table 9. R^2 and RMSE of RFR and XGBoost combined with different feature selection.

	Train		Test			
	R^2	RMSE	R^2	RMSE	Time (s)	Number of features
RFR	0.9935	0.2827	0.9564	0.7340	15.99	22
MIR–RFR	0.9938	0.2765	0.9577	0.7236	9.80	14
RFE–RFR	**0.9946**	**0.2575**	**0.9668**	**0.6384**	**3.14**	**6**
EFS–RFR	0.9941	0.2678	0.9615	0.6879	7.37	**6**
XGBoost	**0.9997**	**0.0588**	0.9731	0.5756	18.74	22
MIR–XGBoost	0.9996	0.0642	0.9754	0.5503	10.41	11
RFE–XGBoost	0.9985	0.1342	0.9762	0.5414	**8.39**	8
EFS–XGBoost	0.9994	0.0850	**0.9771**	**0.5313**	10.81	**7**

Table 10. Statistical descriptions of the features of wells #2 and #3.

Well	Feature	Min.	Max.	Avg.	STDEV
2#	DEPTH	1440	2652.25	2046.125	350.000
	AC	187.485	429.447	244.304	25.360
	DEN	1.258	2.734	2.471	0.152
	FMIT	0.001	12.788	0.366	0.986
	GR	10.397	795.188	90.312	26.330
	PERM	0	55.736	2.550	6.429
3#	DEPTH	1235	2365	1800	326.257
	AC	187.352	512.971	248.909	25.164
	DEN	1.275	2.716	2.522	0.151
	FMIT	0.001	4.342	0.265	0.445
	GR	7.38	200.877	80.774	21.463
	PERM	0.01	54.391	2.429	5.339

a small number of features. Therefore, the performance of the EFS used in this paper is superior.

5.2. Comparison with other regression models

The performance of the regression model is the most critical factor for prediction accuracy. Therefore, selection of the regression model is also very important. Two other tree-based regression models are discussed in this section. One is the random forest regression (RFR) and the other is extreme gradient boosting (XGBoost). They also combine the three feature selection methods above. The main hyperparameters for RFR are listed in table 7. The main hyperparameters for XGBoost are listed in table 8.

Table 11. R^2 and RMSE for wells #2 and #3.

	Train		Test		
Well	R^2	RMSE	R^2	RMSE	Time (s)
2#	0.9982	0.2859	0.9775	0.9581	1.34
3#	0.9988	0.1806	0.9829	0.6849	1.35

The relationship between the feature number and the R^2 score of the three feature selection methods are shown in figure 12. The R^2 score, RMSE, feature number and training time of the corresponding hybrid model are given in table 9. It can be seen that the combination with RFE is the best when using RFR. For this, R^2 is 0.9668, RMSE is 0.6384, the feature number is 6 and the training time is 3.14 s. When using XGBoost, the combination with EFS is the best. For this, R^2 is 0.9771, RMSE is 0.5313, feature number is 7 and the training time is 10.81 s. Among them, the R^2 score curve of the MIR–RFR method drops at first and then rises. The main reason for this is that the initially selected features have already contained a large amount of information. With the increasing number of features, the weight of the initially selected features is reduced, resulting in a decrease in overall performance. When feature number is increasing to a certain amount, the selected features already have a certain amount of information, so the performance of the model improves. Moreover, it

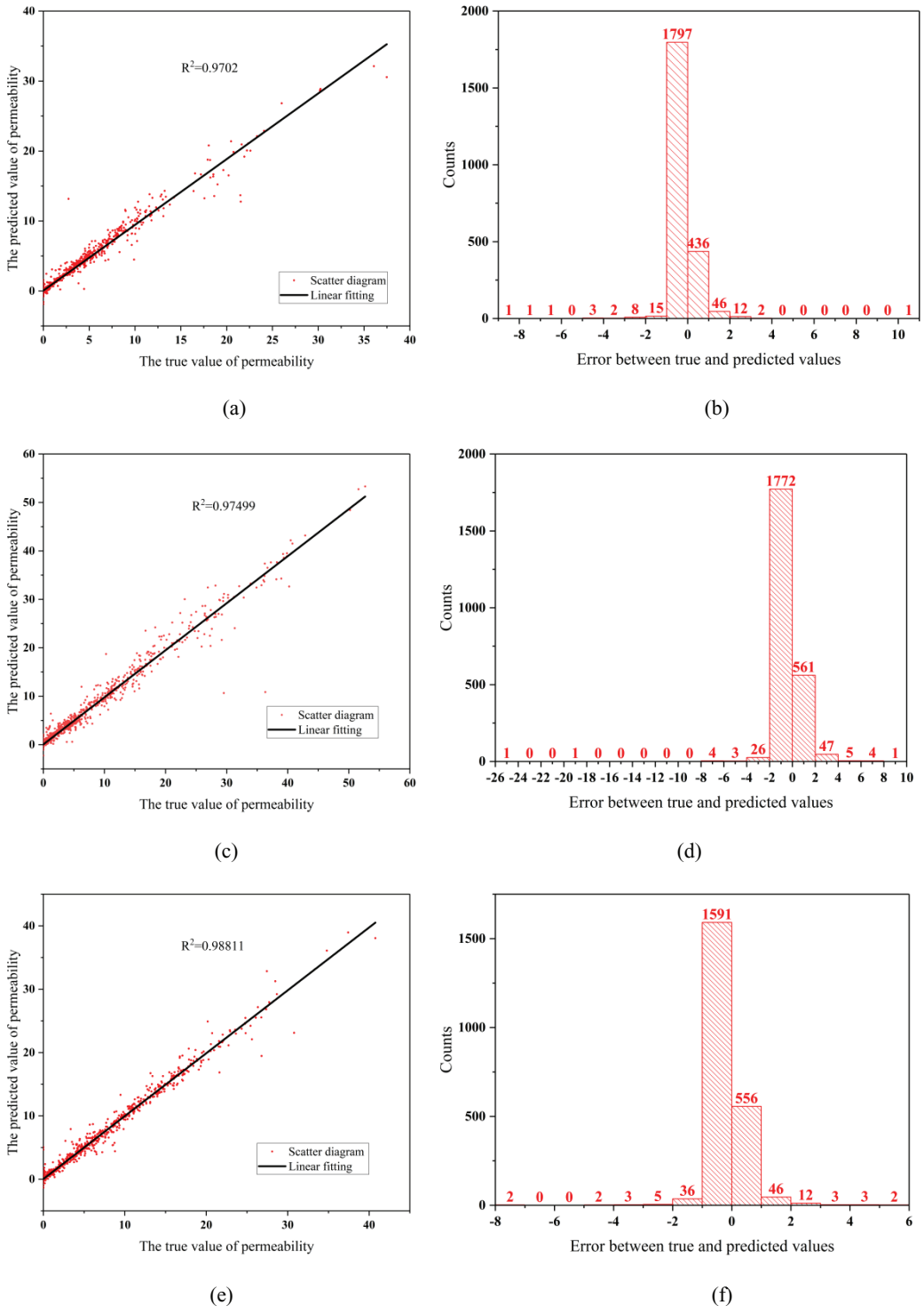

Figure 13. Regression diagram between test data and predictive data of three datasets: (a) regression diagram of well #1; (b) residual distribution of well 1#; (c) regression diagram of well #2; (d) residual distribution of well #2; (e) regression diagram of well #3; (f) residual distribution of well #3.

can be known that DEPTH and RMF are two main features. The mutual information (range (0–1)) between them and the permeability are 0.903 and 0.730, respectively.

The proposed hybrid EFS–LightGBM method in this paper has an R^2 of 0.9712, RMSE of 0.5959, feature number of five and training time of 1.37 s. Although the R^2 and RMSE scores are not the best, the difference is very small. They are only 0.0059 and 0.0646 different respectively from the optimal value, but its advantages in feature number are very obvious and the training time of the model is also superior under the current dataset.

5.3. Verification with other practical datasets

Considering the generalization ability of the proposed model, the EFS–LightGBM was used to predict the reservoir

permeability of another two wells, 2# and 3#, in the same area. The input of the model is the same as the five selected features for well 1#. The statistical descriptions of features for wells 2# and 3# are given in table 10.

The prediction results for these two wells are given in table 11. It can be seen that the R^2 score and RMSE of the prediction result for well #2 are 0.9775 and 0.9581 respectively. For well #3, they are 0.9829 and 0.6849. The relationship between the permeability prediction values and the actual values of the three well test sets including the statistical distribution of the residuals are shown in figure 13. It is obvious that the predicted values are very close to the true values. Meanwhile, the regression residuals are also concentrated in a small interval. It can therefore be seen that the established model for well #1 in this paper also has a good prediction effect on wells #2 and #3. This proves that the proposed model can be effectively applied to the permeability prediction of this area or other similar strata.

6. Conclusions

In this paper, a hybrid reservoir permeability prediction method based on EFS–LightGBM using direct logging data is proposed. In order to improve the efficiency and accuracy of prediction, EFS is used for feature selection and then the LightGBM algorithm is applied for permeability prediction. Furthermore, the influence of feature-threshold selection, base learners' number and dataset size on the prediction results are also studied. The experimental results show that the proposed method has a satisfied prediction result with a small number of features. Its R^2, RMSE and time are 0.9712, 0.5959 and 1.37 s, respectively. In addition, different feature selections and regression models are investigated. Finally, other datasets from another two wells are used to verify the proposed method. The results show that the proposed method has excellent prediction ability, the minimum feature number, less time consumption and good generalization ability. It is proved that the reservoir permeability prediction based on direct logging data is effective with a certain sample size. It can be applied to on-site reservoir inversion and evaluation.

Feature selection has a great influence on the prediction results, which can also continue to be improved in future work. In practice, the reservoir permeability prediction of other wells requires a large amount of data for training. The question of how to further improve the generalization ability of the model under insufficient data will be investigated in future work.

Acknowledgments

Financial support from the National Natural Science Foundation of China (No. 61873101), the National Science and Technology Major Project (No. 2017ZX05019-001), the Fundamental Research Funds for the Central Universities (No. 2019kfyXJJS137) and the Changzhou Key Laboratory of high technology (No. CM20183004) is acknowledged.

ORCID iDs

Kaibo Zhou https://orcid.org/0000-0003-0055-3193
Hao Pan https://orcid.org/0000-0001-9324-0545
Jie Liu https://orcid.org/0000-0002-0750-1030

References

[1] Li K 2008 A new method for calculating two-phase relative permeability from resistivity data in porous media *Transp. Porous Media* **74** 21–33
[2] Mohebbi A and Kaydani H 2015 Permeability estimation in petroleum reservoir by meta-heuristics: an overview *Artificial Intelligent Approaches in Petroleum Geosciences* pp 269–85
[3] Feng X *et al* 2019 Gas multiple flow mechanisms and apparent permeability evaluation in shale reservoirs *Sustainability* **11** 2114
[4] Winardhi C W, Maulana F I and Latief F D E 2016 Permeability estimation of porous rock by means of fluid flow simulation and digital image analysis *IOP Conf. Ser.* **29** 12005
[5] Yang S *et al* 2018 A new fracture permeability model of CBM reservoir with high-dip angle in the southern Junggar Basin, NW China *Energy Explor. Exploit.* **37** 125–43
[6] Osorio R *et al* 2017 Geological interpretation of channelized heterolithic beds through well test analysis *J. Pet. Sci. Eng.* **158** 516–28
[7] Chen S *et al* 2018 *In situ* stress, stress-dependent permeability, pore pressure and gas-bearing system in multiple coal seams in the Panguan area, western Guizhou, China *J. Nat. Gas Sci. Eng.* **49** 110–22
[8] Ngo V T, Lu V D and Le V M 2018 A comparison of permeability prediction methods using core analysis data for sandstone and carbonate reservoirs *Geomech. Geophys. Geo-Energy Geo-Resourc.* **4** 129–39
[9] Singh H and Cai J 2019 A feature-based stochastic permeability of shale: part 1—validation and two-phase permeability in a Utica shale sample *Transp. Porous Media* **126** 527–60
[10] Feng Q *et al* 2019 Apparent permeability model for shale oil with multiple mechanisms *J. Pet. Sci. Eng.* **175** 814–27
[11] Ramandi H L, Mostaghimi P and Armstrong R T 2017 Digital rock analysis for accurate prediction of fractured media permeability *J. Hydrol.* **554** 817–26
[12] Naraghi M E, Javadpour F and Ko L T 2018 An object-based shale permeability model: non-Darcy gas flow, sorption, and surface diffusion effects *Transp. Porous Media* **125** 23–39
[13] Zhang Y *et al* 2019 A combined drug discovery strategy based on machine learning and molecular docking *Chem. Biol. Drug Des.* **93** 685–99
[14] Yao L *et al* 2019 Prediction of antiepileptic drug treatment outcomes of patients with newly diagnosed epilepsy by machine learning *Epilepsy Behav.* **96** 92–7
[15] Tao R *et al* 2019 Magnetocardiography-based ischemic heart disease detection and localization using machine learning methods *IEEE Trans. Biomed. Eng.* **66** 1658–67
[16] Wang X *et al* 2019 Prediction of the 1-year risk of incident lung cancer: prospective study using electronic health records from the state of Maine *J. Med. Internet Res.* **21** e13260
[17] Hu Z *et al* 2019 Data-driven fault diagnosis method based on compressed sensing and improved multi-scale network *IEEE Trans. Ind. Electron.* 1–10

[18] Liu J *et al* 2018 An integrated multi-sensor fusion-based deep feature learning approach for rotating machinery diagnosis *Meas. Sci. Technol.* **29** 55103
[19] Elkatatny S and Mahmoud M 2018 Development of a new correlation for bubble point pressure in oil reservoirs using artificial intelligent technique *Arab. J. Sci. Eng.* **43** 2491–500
[20] Ahmadi M A, Zendehboudi S and James L A 2018 Developing a robust proxy model of CO_2 injection: coupling Box–Behnken design and a connectionist method *Fuel* **215** 904–14
[21] Menad N A *et al* 2019 Modeling temperature-based oil-water relative permeability by integrating advanced intelligent models with grey wolf optimization: application to thermal enhanced oil recovery processes *Fuel* **242** 649–63
[22] Wang X *et al* 2018 Improved pore structure prediction based on MICP with a data mining and machine learning system approach in Mesozoic strata of Gaoqing field, Jiyang depression *J. Pet. Sci. Eng.* **171** 362–93
[23] Merembayev T, Yunussov R and Yedilkhan A 2018 Machine learning algorithms for classification geology data from well logging (IEEE) *14th Int. Conf. Electron. Comput. & Comput.* pp 206–12
[24] Saemi M, Ahmadi M and Varjani A Y 2007 Design of neural networks using genetic algorithm for the permeability estimation of the reservoir *J. Pet. Sci. Eng.* **59** 97–105
[25] Lim J 2005 Reservoir properties determination using fuzzy logic and neural networks from well data in offshore Korea *J. Pet. Sci. Eng.* **49** 182–92
[26] Elkatatny S *et al* 2018 New insights into the prediction of heterogeneous carbonate reservoir permeability from well logs using artificial intelligence network *Neural Comput. Appl.* **30** 2673–83
[27] Gholami R, Shahraki A R and Jamali Paghaleh M 2012 Prediction of hydrocarbon reservoirs permeability using support vector machine *Math. Probl. Eng.* **2012** 1–18
[28] Gu Y, Bao Z and Cui G 2018 Permeability prediction using hybrid techniques of continuous restricted Boltzmann machine, particle swarm optimization and support vector regression *J. Nat. Gas Sci. Eng.* **59** 97–115
[29] Ahmadi M *et al* 2014 Connectionist model predicts the porosity and permeability of petroleum reservoirs by means of petro-physical logs: application of artificial intelligence *J. Pet. Sci. Eng.* **123** 183–200
[30] Maldonado S and López J 2018 Dealing with high-dimensional class-imbalanced datasets: embedded feature selection for SVM classification *Appl. Soft Comput.* **67** 94–105
[31] Ke G *et al* 2017 LightGBM: a highly efficient gradient boosting decision tree *Adv. Neural Inf. Process. Syst.* **30** 3146–54
[32] Zhan Z, You Z-H, Li L-P, Zhou Y and Yi H-C 2018 Accurate prediction of ncRNA-protein interactions from the integration of sequence and evolutionary information *Frontiers Genet.* **9** 458
[33] Ma X *et al* 2018 Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning *Electron. Commerce Res. Appl.* **31** 24–39
[34] Venkatesh B and Anuradha J 2019 A review of feature selection and its methods *Cybern. Inf. Technol.* **19** 3–26
[35] Chen T and Guestrin C 2016 XGBoost: a scalable tree boosting system *Proc. of the 22nd ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.* pp 785–94
[36] Hammami M *et al* 2019 A Multi-objective hybrid filter-wrapper evolutionary approach for feature selection *Memetic Comput.* **11** 193–208