energies

MDPI

Article

Re-Evaluation of Oil Bearing for Wells with Long Production Histories in Low Permeability Reservoirs Using Data-Driven Models

Yongchao Xue *, Chong Cao *, Qingshuang Jin and Qianyu Wang

State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum, Beijing 102249, China

* Correspondence: xyc75@cup.edu.cn (Y.X.); caochong17_rsc@163.com (C.C.)

Abstract: The re-evaluation of oil-bearing wells enables finding potential oil-bearing areas and estimating the results of well logging. The re-evaluation of oil bearing is one of the key procedures for guiding the development of lower production wells with long-term production histories. However, there are many limitations to traditional oil-bearing assessment due to low resolution and excessive reliance on geological expert experience, which may lead to inaccurate and uncertain predictions. Based on information gain, three data-driven models were established in this paper to re-evaluate the oil bearing of long-term production wells. The results indicated that the RF model performed best with an accuracy of 95.07%, while the prediction capability of the neural network model was the worst, with only 79.8% accuracy. Moreover, an integrated model was explored to improve model accuracy. Compared with the neural network, support vector machine, and random forest models, the accuracy of the fusion model was improved by 20.9%, 8.5%, and 1.4%, which indicated that the integrated model assisted in enhancing the accuracy of oil-bearing prediction. Combined with the long-term production characteristics of oil wells in the actual oil field, the potential target sweet spot was found, providing theoretical guidance for the effective development of lower production wells in the late period of oilfield development.

Keywords: formation evaluation; oil bearing; data-driven models; long production history; low permeability reservoirs

check for updates

Citation: Xue, Y.; Cao, C.; Jin, Q.; Wang, Q. Re-Evaluation of Oil Bearing for Wells with Long Production Histories in Low Permeability Reservoirs Using Data-Driven Models. *Energies* **2023**, *16*, 677. https://doi.org/10.3390/en16020677

Academic Editor: Vladimir Alvarado

Received: 8 October 2022
Revised: 30 November 2022
Accepted: 26 December 2022
Published: 6 January 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Oil-bearing evaluation is of great significance for the identification of potential oil-bearing areas and the exploration of remaining sweet spots in the secondary description of reservoirs, which is vital to further develop and enhance oil recovery for wells with long-term production histories [1–4].

There are numerous reports on the prediction of oil bearing in previous studies. NMR (nuclear magnetic resonance) is frequently employed to predict the oil bearing of shale, such as the division of different hydrocarbon-containing components in shale [5]. The authors of Ref. [6] applied nuclear magnetic resonance to study the Chang 7 shale in the Yan'an area. The distribution of portable oil can be forecasted based on differences in the relaxation mechanisms of different hydrogen-containing components in the rock measured by NMR. However, the NMR method is precious and only routinely implemented for the examination of oil-bearing cores. The capture range of NMR is too short for the whole oil reservoir.

In addition, lithofacies can also be used for oil-bearing prediction. In general, the constraints of lithofacies can only identify a fixed oil-bearing range, and it is extremely difficult to accurately determine hydrocarbon bearing. The authors of Ref. [7] studied the oil-bearing properties of different lithofacies based on the extraction of cores, fluorescent thin sections, rock pyrolysis, total organic carbon (TOC), and nuclear magnetic resonance

(NMR) testing. The results confirmed that the oil-bearing and physical properties of shale are better than those of shell limestone. The authors of Ref. [8] studied the physical properties of lacustrine deep-water turbidite reservoirs, revealing that the petrophysical properties seem to have a significant impact on hydrocarbon accumulation in turbidite sandstones. The authors of Ref. [9] used core, geological analysis, production, and logging data to characterize the lithology, oil-water layer, and oil layer thickness of the Chang 7 tight sandstone reservoir. The reservoir identification standard, the normalized difference curve, is superimposed on the resistivity curve to accurately determine the thickness of tight and fine sandstone reservoirs. However, this method is mainly used to calculate the thickness of the oil layer, and requires a considerable amount of data. In addition, the authors of Ref. [10] utilized the thermal terahertz analysis (TTA) method to detect the oil-bearing characteristics of desert reservoirs. However, this method is only applicable to macro level oil-bearing predictions.

The previous studies indicate that seismic data is frequently used to predict oil bearing in the plane [11–13]. However, the resolution of seismic data is too coarse to make an accurate oil-bearing prediction.

Well logging data is also often utilized for identification of hydrocarbon-bearing intervals, prediction of oil bearing, and the discovery of remaining oil [14–16]. These processes mostly rely on expert experience since the nonlinear relationship between the complex variables and multi-type data in the actual oil field is difficult to capture. Fortunately, a large amount of oil field data (i.e., geological information, oil testing, well logging) makes it possible to identify oil bearing [17–22].

In recent years, data-driven models have been widely adopted in the oil industry, including production prediction, injection-production parameter optimization, deposits with high paraffin content analysis, fracturing parameter optimization, recovery factor forecast, etc. [23–28].

A small number of scholars have conducted research on the application of machine learning in oil bearing. The authors of Ref. [29] evaluated a machine learning algorithm suitable for borehole geomagnetism to predict the remaining oil saturation in the vertical direction. Six data-driven models were established and tested. The results showed that the AdaBoost-random forest model performed best, with a prediction accuracy of 87%. However, this method requires corresponding bore-ground electromagnetic method data. The authors of Ref. [30] investigated the impact of physical properties (e.g., thickness, permeability, porosity, net to gross ratio, etc.) on oil saturation. Ten data-driven methods were applied, including random forest, lasso, gradient boosting, ridge regression, Adaboost, elastic net regression, support vector machine, linear regression, multilayer perceptron, and polynomial regression. This study revealed that NWIV (adjacent water injection variation) was the variable with the greatest influence on OSV (oil saturation variation). However, this was a prediction of dynamic oil bearing and could not be used for the identification and prediction of static oil bearing. The authors of Ref. [31] estimated gas hydrate saturation by machine learning (ML) algorithms, including ridge regression, decision tree, k-nearest neighbor, reduced order models, neural network, etc. The results showed that the k-nearest neighbor model performs best with an accuracy of 82.96%.

In this paper, logging data with high resolution was adopted to make oil-bearing predictions for wells with low production rates during long-term production periods. The purpose of this paper was to combine various types of oil field data and evaluate the oil bearing of a reservoir with data-driven models to provide theoretical guidance for the EOR of oil wells with low production rates under long-term production periods.

2. Data Preparation

2.1. Geology of the Target Block

The target area of this study was the Chang 8 ultra-low permeability reservoir in Changqing oil field, Ordos Basin, China. The main rock type of the reservoir is siltstone (80%). The porosity and permeability of siltstone are relatively small, which also leads to

the lower porosity and permeability of this area. The average thickness of the oil layer is 14.6 m, the average porosity is 9.97%, the average permeability is 0.56×10^{-3} μm^2, and the original formation pressure is 15.8 MPa. A total of 287 oil-producing wells and 94 water injection wells were put into production in December 2006. The flow rate of the wellhead for this block is 740 t/day, the oil rate is 576 t/day, and the average oil rate of a single well is 2.0 t/day with 22.1% water cut.

The development wells in this area were put into production around July 2005, and generally have a production history of approximately 10 years. The average daily oil rate per well dropped from 2.8 to 0.6 tons/day. As shown in Figure 1, the average monthly oil production of a single well in the target block has been decreasing since it was put into production. Figure 2 shows the current oil saturation field map of the target block. It can be seen from the figure that there is still a large amount of remaining oil in the target block that needs to be re-evaluated and excavated.

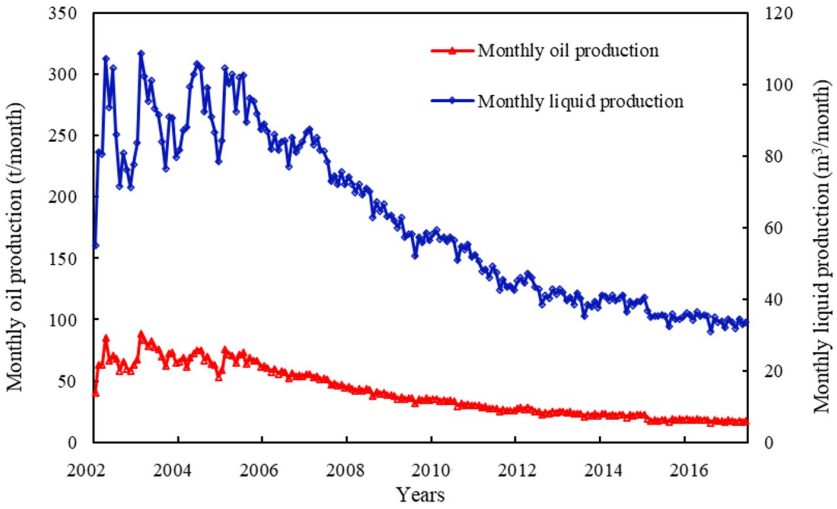


Figure 1. Average monthly oil production.

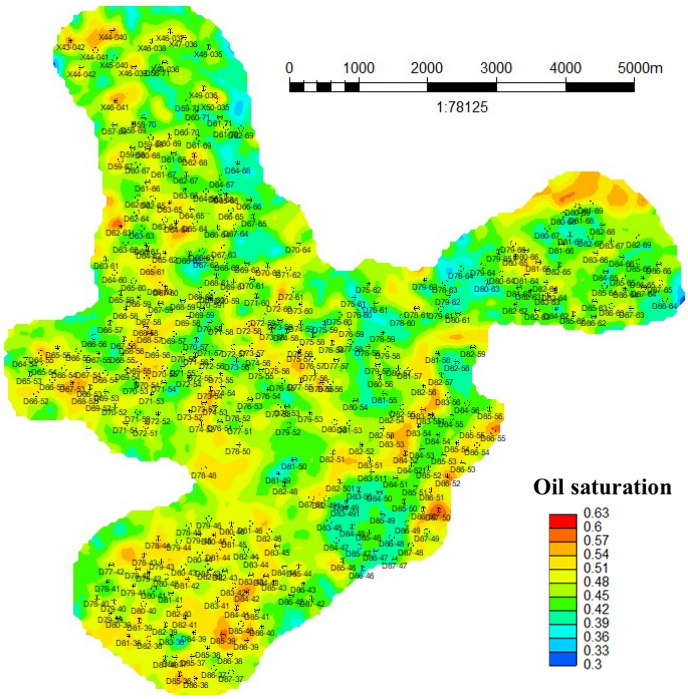


Figure 2. Initial field map of oil saturation.

In the late stage of oil field development, the remaining oil is of vital significance to improve the production and recovery of a single well. Therefore, quickly re-evaluating the oil bearing of wells with long production histories is a necessary and urgent task that can provide guidance for the development of the oil field in the later period.

The well logging data for oil-bearing prediction is mainly dependent on the rich experience of geological engineers, and it is often miscellaneous and tedious work. The interpretation of results given by different engineers may vary widely. Therefore, in order to improve the oil-bearing prediction of oil wells, it is necessary to fully exploit the characteristics of logging data. In addition, there is a complex nonlinear relationship between oil-bearing capacity and logging data, which is extremely challenging for general models or equations to describe. Machine learning models are strong adaptable to nonlinear relationships and can extract the hidden features and relationships behind the disorganized oil field data.

2.2. Oil Field Dataset

The oil field data collected from a block in Changqing oil field, Ordos Basin, China included logging, perforation, and production data. There are more than 400 wells in the target block, including 361 old wells. There were 8 variables in the raw data, such as spontaneous potential (SP), natural gamma (GR), acoustic transit time (AC), resistivity (RT), porosity (POR), permeability (PERM), water saturation (SW), and perforation. The data labels for oil bearing were 1 for an oil-bearing layer and 0 for an interlayer. The distribution and statistical information of the oil field data are shown in Figure 3 and Table 1, respectively.

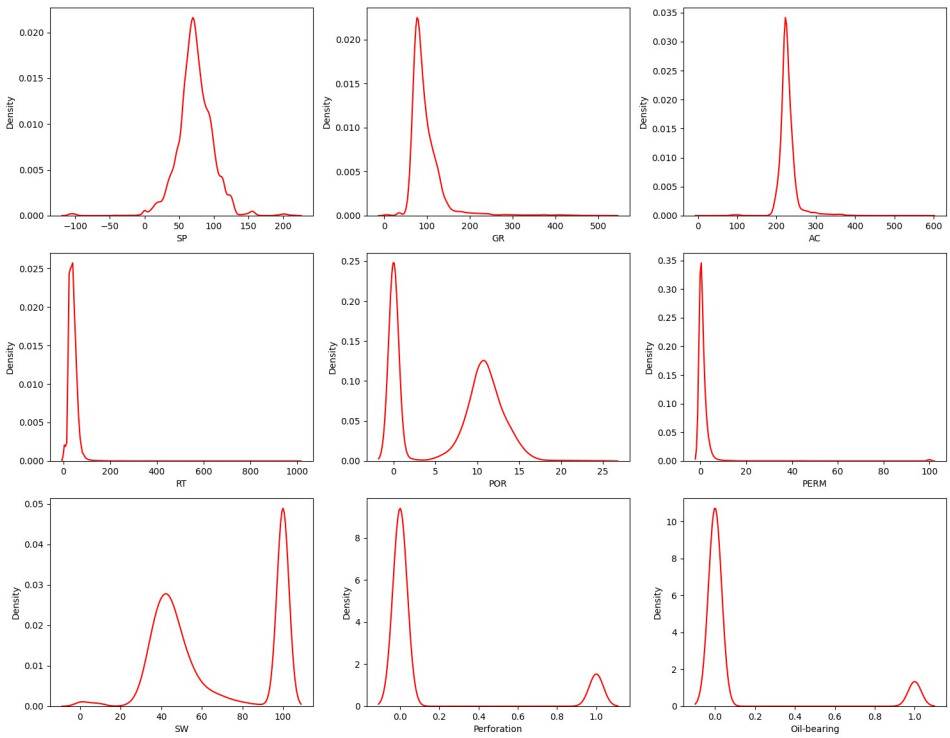


Figure 3. The distribution of data.

Table 1. Logging data characteristics and distribution.

Statistics	DEPTH (m)	SP (mV)	GR (API)	AC (μs/m)	RT (Ω·m)
mean	2031.6	73	97.1	228.4	40.1
std	61.3	26	41.6	24.4	19.9
min	1800	−111.1	0.1	1	0
25%	1999.9	59.5	75.3	217.9	28
50%	2040	72.1	86.8	225.1	38
75%	2072.5	87.7	107	234.5	49
max	2169.9	218.1	532.5	594.1	1011
Statistics	POR (%)	PERM (mD)	SW (%)	Perforation	Oil-bearing
mean	6.9	1.5	65	0.1	0.1
std	5.6	7.1	28.2	0.3	0.3
min	0	0	0	0	0
25%	0	0	41.7	0	0
50%	9	0	52	0	0
75%	11	1.1	100	0	0
max	25	100	100	1	1

2.3. Data Processing Procedures

In this paper, the oil field data including well logging, perforation, oil testing, and production data collected from 361 vertical wells in Changqing oil field, Ordos Basin, China. Before model training, it is necessary to process the raw data and convert it to available, easily identified data points. In general, the detailed procedures of oil-bearing prediction can be divided into: oil field data collection, data processing, feature selection, data splitting, model evaluation, and prediction, as shown in Figure 4. Among them, data processing, which includes filling the missing records, noise reduction, perforation coding, and oil-bearing labeling, is the crucial step for predictive capability.

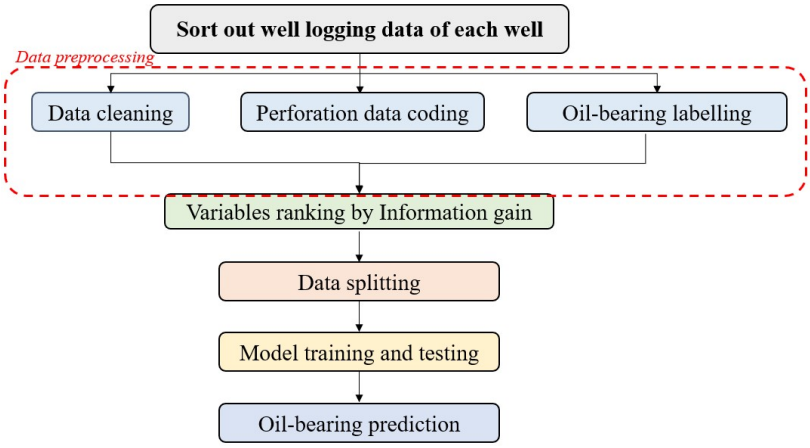


Figure 4. Procedures of oil-bearing prediction.

(1) **Sort out oil field data.** The well logging, perforation, production, and test data collected from the oil field were classified and sorted out. Each column in a row was carefully checked. If more than 50% of data were missing, this record was directly deleted. If only individual values were missing in one row, the neighboring average was used to fill the row.

(2) **Perforation data coding.** For the oil-content checking problem in this paper, the well perforation data belonged to the classification data and needed to be coded before training. In this paper, the one-coding technique was used to process the perforation data and convert the classified data into numerical data. One-coding is also known as one-bit effective encoding. The state registers are adopted to encode the states,

which have their own independent register bit with only one bit. Each variable might become m binary values with m possible values. One-coding can deal with classified data and expand features to a certain extent with only 0 and 1 in the vertical space, as shown in Figure 5.

(3) **Oil-bearing labeling.** The oil content was measured using oil testing data. If there was oil production, this item was marked as 1. Otherwise, it was marked as 0.

(4) **Variable ranking.** Multiple variables provide a large amount of information for feature extraction; however, the correlation between multiple variables also increases the complexity of the problem. In this paper, information gain was introduced to analyze the correlation between variables and rank them. Then, the performance of the well was predicted using the preferred variables.

(5) **Training and testing data splitting.** The processed well logging data and marked oil production data together constitute the oil-content evaluation dataset. Before training, it was necessary to divide the training and testing datasets. If the proportion of the training dataset is too small, the training model will be over-fitted and the generalization performance will be poor. If the proportion is too large, the test dataset will be too small and the reliability of the test effect will be reduced. Thus, the data were randomly split into 90% training and 10% testing datasets.

(6) **Model training.** After splitting the data, data-driven models, such as neural network, support vector machine, and random forest models, were trained on the training dataset. The model parameters were compared and adjusted to optimize the prediction accuracy of the model.

(7) **Oil-bearing prediction.** The testing dataset was adopted to predict oil bearing, an integrated model was explored to improve model accuracy, and finally, applications in the actual oil field were achieved.

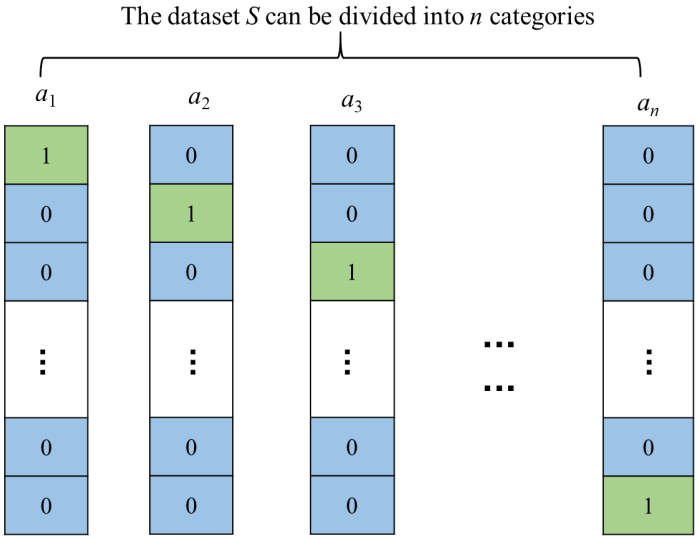


Figure 5. Schematic diagram of one-hot encoding.

2.4. Information Gain

"Information entropy" is the most commonly used indicator to measure the purity of a sample set. Assuming that in the training dataset D, $|D|$ is the sample capacity, that is, the number of samples or elements in D. There are K classes C_k where $|C_k|$ is the number of samples of C_k, and the sum of $|C_k|$ is $|D|$, $k = 1, 2, \ldots, K$. Thus, the information entropy of dataset D can be calculated using the following formula [32]:

$$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|} \quad (1)$$

According to each feature A (i.e., per logging curve in this paper), D (oil bearing in this paper) is divided into n subsets $D_1, D_2, \ldots, D_n$, where D_i represents the dataset in D corresponding to the ith value of feature A. $|D_i|$ is the number of samples of D_i, the sum of $|D_i|$ is $|D|$, $i = 1, 2, \ldots, n$, the set of samples belonging to C_k in D_i is recorded as D_{ik} (i.e., the intersection), and $|D_{ik}|$ is the number of samples of D_{ik}. The conditional information entropy of selected A can be calculated as follows:

$$H(D|A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \sum_{k=1}^{K} \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|} \tag{2}$$

Then, the information gain of feature A to dataset D is:

$$g(D, A) = H(D) - H(D|A) \tag{3}$$

Generally speaking, the greater the information gain, the greater the "purity improvement" obtained using feature A for partitioning. It is noted that most of the data in this study were metric variables. However, the above method is used to handle categorical variables. A new method was provided for metric variables based on the information gain. The processing steps of the new method are as follows:

Given a training dataset D and a metric variable a, it is assumed that a appears with n different values on D. First, these values are sorted from smallest to largest and denoted as $\{a^1, a^2, a^3, \ldots \ldots, a^n\}$. Based on the division point t, D is divided into subsets D_{t-} and D_{t+}, where D_{t-} is the sample with values no greater than t on attribute a. The results are the same whatever value t took between a^i and a^{i+1} to divide D. Thus, for the metric variable a, the set of candidate division points containing $n - 1$ elements are examined.

$$T_a = \left\{ \frac{a^i + a^{i+1}}{2} \middle| 1 \le i \le n - 1 \right\} \tag{4}$$

That is, the median of the interval $[a^i, a^{i+1})$ is taken as the candidate division point. Thus, metric variables are handled in the same way as categorical variables, and we selected the optimal division point for the division of the sample set using the following formula.

$$Gain(D, a) = \max_{t \in T_a} Gain(D, a, t) = \max_{t \in T_a} \left(Ent(D) - \sum_{\lambda \in \{-,+\}} \frac{|D_t^\lambda|}{|D|} Ent(D_t^\lambda) \right) \tag{5}$$

where Gain(D, a, t) is the information gain of the sample set D after dichotomizing based on the division point t. When dividing, the division point is chosen that gives the largest Gain(D, a, t).

3. Principles of Data-Driven Models

3.1. Support Vector Machine

SVM (support vector machine) is a common classification and regression machine learning method. In Figure 6, dots and circles refer to different results divided by support vector machine. Red dots and red circles emphasize that these dots and circles are crucial support vectors for classification. In the field of machine learning, it is a supervised learning model, which is one of the most robust and generalizing methods of all the well-known data mining algorithms. It is usually used for pattern recognition.

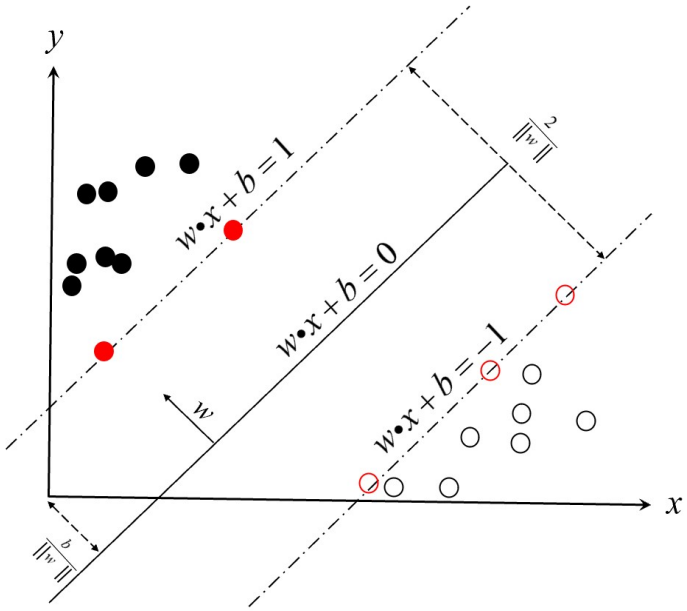


Figure 6. Schematic diagram of SVM.

Suppose the training sample set is $S = \{(x_1,y_1), (x_2,y_2), (x_3, y_3) \ldots , (x_n,y_n)\}$ where n represents the number of training samples. The support vector machine learns the training samples and establishes functions describing the relationship between input variables and output, as follows:

$$f(x_i) = \omega^T \phi(x_i) + b \tag{6}$$

where $\phi(x_i)$ represents the mapping function, ω represents the weight vector, and b represents the threshold vector.

In practical cases, it is difficult for the data to be linearly separated and it is difficult to judge whether the seemingly linearly separable results are caused by overfitting. To solve this problem, the concept of "soft margins " was introduced. Soft margins allow some samples not to satisfy the constraints, achieving linear separability of most samples. In order to measure the tolerance of the soft margins to the number of samples and parameters that do not meet the constraints, a penalty term is introduced, and the penalty term coefficient is set as the penalty factor C (C > 0). The larger the value of C, the greater the tolerance of the model to sample errors. When C is infinite, all samples will be forced to satisfy the constraints.

By introducing the kernel function to replace product operation $\phi(x_i)^T \phi(x)$, the prediction decision function of the support vector machine can be obtained as follows:

$$f(x) = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) k(x, x_i) + b \tag{7}$$

where $k(x, x_i)$ is the kernel function, given by:

$$k(x, x_i) = \exp(-\|x_i - x\|/2\sigma^2) \tag{8}$$

Table 2 shows the kernel functions commonly used in support vector machine model.

Table 2. Kernel functions for SVM.

Kernel Function	Expression	Parameter
Linear kernel	$\kappa\left(x_i, x_j\right) = x_i^T x_j$	
Polynomial kernel	$\kappa\left(x_i, x_j\right) = \left(x_i^T x_j\right)^d$	$d \geq 1$, d is the polynomial degree
Gaussian kernel	$\kappa\left(x_i, x_j\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right)$	$\sigma > 0$, σ is the Gaussian kernel bandwidth
Laplace kernel	$\kappa\left(x_i, x_j\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|}{2\sigma^2}\right)$	$\sigma > 0$
Sigmoid kernel	$\kappa\left(x_i, x_j\right) = \tanh\left(\beta\left(x_i^T x_j + \theta\right)\right)$	tanh is the hyperbolic tangent, $\beta > 0, \theta < 0$

When the selected kernel function is a Gaussian kernel (also known as a radial basis, RBF function), the parameter γ mainly defines the influence of a single sample on the entire classification hyperplane. When γ is relatively small, the effect of a single sample on the entire classification hyperplane is relatively small, and it is not easily selected as a support vector. Conversely, when γ is relatively large, a single sample has a greater impact on the entire classification hyperplane, and it is more easily selected as a support vector, or the entire model will have more support vectors.

3.2. Artificial Neural Network

An artificial neural network consists of forward propagation of the signal and back propagation of the error. During forward propagation, input samples are fed from the input layer, processed by the hidden layer, and then transmitted to the output layer. If the predicted and target values of the output layer are not met, the error will be passed to the hidden layer through back propagation to each layer of neurons, which can further modify the weight of each neuron. Moreover, the weight adjustment process of forward propagation and error back propagation is carried out repeatedly, and this process continues until the error of the network output is reduced to an acceptable level.

As shown in Figure 7, for the specific training sample $X_i = (x_1, x_2, \ldots, x_n)$, the output vector of the hidden layer is $Y_i = (z_1, z_2, \ldots, z_p)$, the output value of the output layer is $\boldsymbol{O_i}$, and the expected output is $\boldsymbol{y_i}$. The weight matrix from the input layer to the hidden layer is represented by $V = (V_1, V_2, \ldots, V_p)$, where the column vector V_j is the weight vector corresponding to the jth neuron in the hidden layer, from the hidden layer to the output. The weight vector between layers is represented by $W = (w_1, w_2, \ldots, w_p)$, where w_k is the weight of the kth neuron in the hidden layer corresponding to the output layer neuron. Table 3 shows the equations of common activation functions.

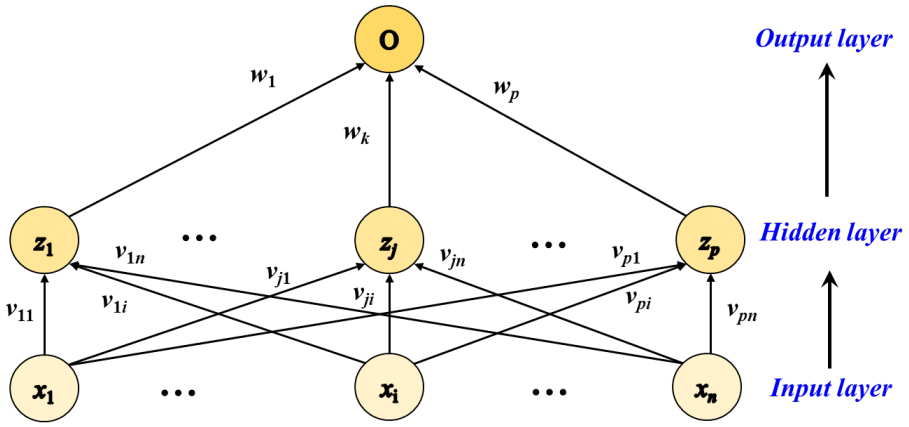


Figure 7. Schematic diagram of neural network.

Table 3. Activation functions.

Activation Function	Equation
Purelin activation function	$f(x) = x$
Sigmoid activation function	$f(x) = \sigma(x) = \frac{1}{1+e^{-x}}$
Tanh activation function	$f(x) = \tanh(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$
ReLU activation function	$f(x) = \begin{cases} 0 \; for \; x < 0 \\ x \; for \; x \geq 0 \end{cases}$

In this paper, the gradient descent method was used to adjust the weight value. The adjustment of the weights was proportional to the error of gradient descent by introducing the learning rate. For a three-layer neural network, the number of neurons in the input and output layers are dependent on the specific problem. If the number of neurons in the hidden layer is too small, the prediction and generalization capabilities of the network will be reduced. If the number of neurons in the hidden layer is too large, long-term non-convergence of the network training occurs and the fault tolerance performance decreases. If there are an excessive number of hidden layer neurons, convergence of network training is difficult. To the best of the author's knowledge, there is no consensus on the number of neurons in the hidden layer. In this paper, as few hidden layer neurons were selected as possible under the condition of satisfying the error requirements.

3.3. Random Forest

Random forest is a machine learning algorithm based on decision trees, as shown in Figure 8. There are three common algorithms used in decision trees, namely ID3, C4.5, and CART. CART is classified according to the Gini index. The CART algorithm adopts a bisection recursive segmentation method based on the Gini index, and thus the dataset can be divided into two branches with the smallest Gini index. This paper adopted the random forest algorithm based on CART. The expression of the Gini index is as follows:

$$Gini(p) = \sum_{i=1}^{k} p_i(1 - p_i) = 1 - \sum_{i=1}^{k} {p_i}^2 \tag{9}$$

where k is the number of categories and the probability of each category appearing is P_i.

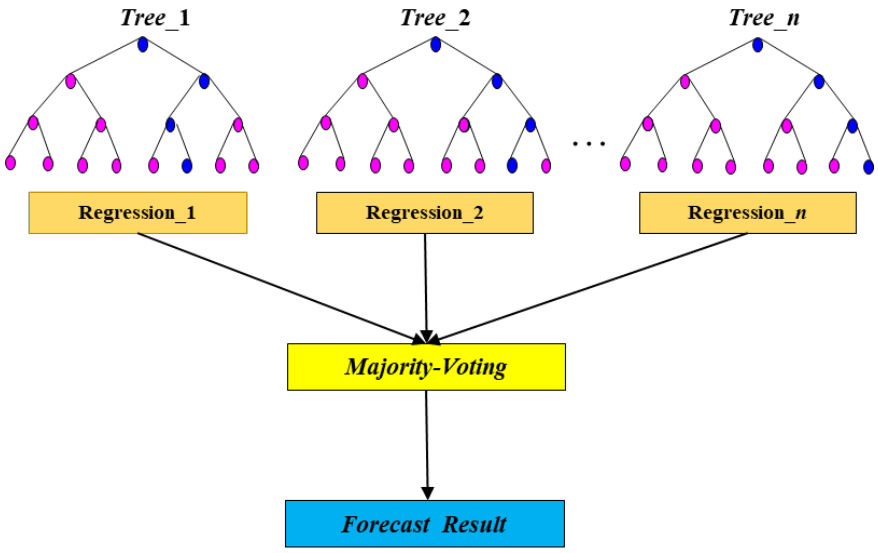


Figure 8. Schematic diagram of random forest.

Dataset S can be split into two sub-sample sets, S_1 and S_2, according to certain conditions. Thus, the Gini split index of dataset $\boldsymbol{S}$ is shown in Equation (10):

$$Gini(S) = \frac{m_1}{m} Gini(S_1) + \frac{m_2}{m} Gini(S_2) \tag{10}$$

where m is the number of datasets S; m_1 is the number of datasets S_1; and m_2 is the number of datasets S_2.

Random forest is an ensemble learning algorithm proposed by Breiman, which is an ensemble classifier involving multiple decision trees. Each decision tree in the random forest is obtained based on Bootstrap sampling, and then the category with the majority votes is selected as the final classification result with the combination of multiple decision trees. The construction process of random forest is as follows:

Assuming that in the original training data set D, n is the number of samples, M is the total number of features, and the number of decision trees required to be constructed is K.

(1) Extract the training subset. n samples were randomly selected and repeated n times from the original training dataset D, and the others formed the out-of-bagging dataset.
(2) Build a decision tree. Firstly, m feature subsets ($m < M$) were randomly selected from M features, and then the optimal splitting points were selected by relevant criteria and divided into sub-nodes. Meanwhile, the training dataset was divided into corresponding nodes, and finally, all nodes were obtained.
(3) Random forest generation. Step (2) was repeated until $\boldsymbol{k}$ decision trees were established to form a random forest $\{t_i, i = 1, 2, \ldots, k\}$.
(4) Using k decision trees in the random forest, the predicted results $\{t_1(x), t_2(x), \ldots, t_k(x)\}$ were obtained.
(5) The mode of the predicted result of each decision tree was taken as the final prediction result of each sample.

$$T(x) = \text{argmax} \sum_{i=1}^{k} I(t_i(x) = y) \tag{11}$$

4. Results and Discussion

4.1. Variables Ranking

Feature selection is a fundamental problem in feature engineering. Its goal is to obtain the optimal feature subset. Feature selection can eliminate irrelevant or redundant features, so as to reduce the number of features, improve the accuracy of the model, and reduce the running time.

Feature selection was performed by the method of information gain in this study. Logging data was taken as a feature and the oil-bearing property was taken as labels (oil-bearing marked 1, oil-free marked 0) in the process of feature optimization. The information gain of each logging curve to the oil-bearing result was calculated, and the logging curve feature with large information gain was selected as the final training feature. The information gain of each logging curve on the oil bearing results is shown in the figure below.

It can be seen from Figure 9 that the information gains of the *SP* (spontaneous potential), *GR* (natural gamma), and *AC* (acoustic transit time) logging curves were large, higher than 0.45, while those of other logging curves were small, lower than 0.2. In order to simplify the model and improve the prediction efficiency, the *SP*, *GR*, and *AC* logging curves were selected as the final feature optimization results.

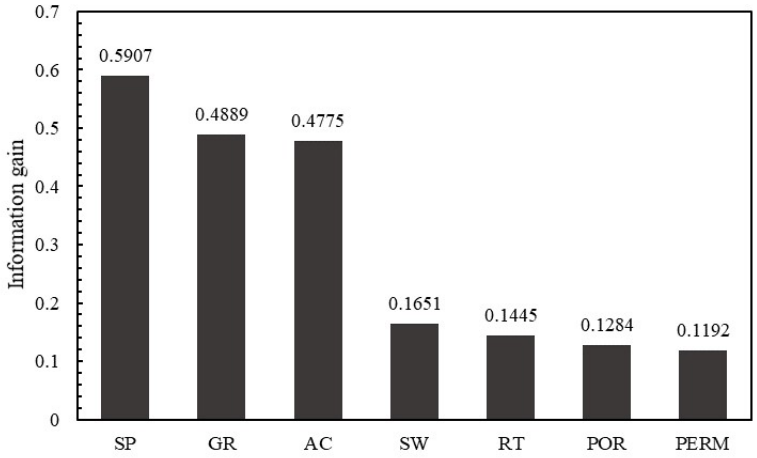


Figure 9. Information gain of each variable.

4.2. Oil-Bearing Prediction

Based on the PCA analysis results, this paper used the *SP*, *AC*, and *GR* logging curves as input features, and well depth was adopted as the index to establish the data-driven models for evaluating oil-bearing wells with long production histories. For the binary classification problem, there are many evaluation indicators. In order to make a comparative analysis of traditional well logging results, the accuracy of prediction was introduced in this paper, as shown in Equation (12). Accuracy referred to the ratio of the total number of accurate prediction results to the total number of samples in the binary classified problem.

$$Accuracy = \frac{TP + TN}{N} \tag{12}$$

where *TP* is the number of true positive samples; *TN* is the number true negative samples; *N* is the total number of samples; *Accuracy* is the accuracy of prediction.

4.2.1. Oil-Bearing Prediction Based on SVM Model

In SVM models, it has been reported that the kernel function and penalty factor have a significant impact on the prediction effect, which needs to be optimized. Four kernel functions (linear kernel, sigmoid kernel, poly kernel, Gaussian kernel) were compared in the oil-bearing analysis under different penalty factors, as shown in Figure 10.

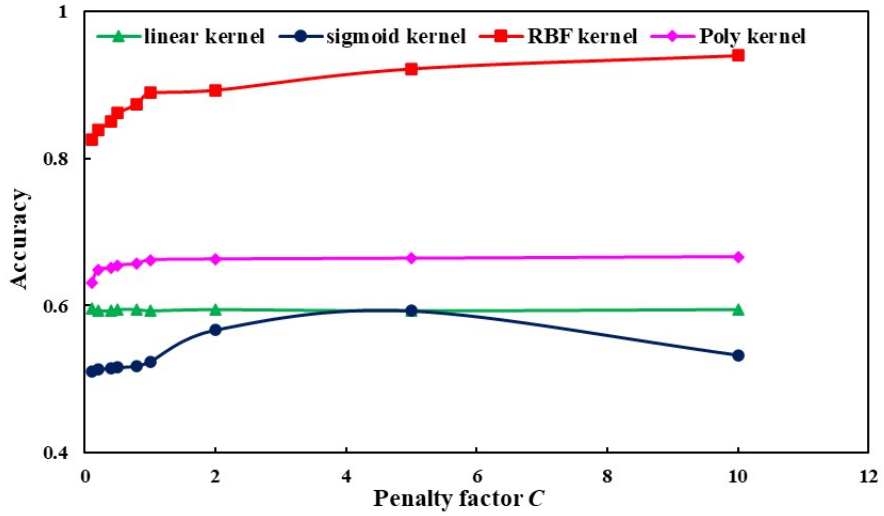


Figure 10. Comparison of the prediction results of different kernel functions of the SVM model.

As shown in Figure 10, it can be seen from the comparative analysis of the four kernel functions that the Gaussian kernel function (RBF kernel) had the best model prediction capability, with a prediction accuracy is as high as 97.3% when the penalty factor was 5. This paper further tested the oil-bearing prediction capability of the Gaussian kernel function under different penalty factors, and the results are shown in Table 4. It can be seen from the table that a better prediction effect of the Gaussian kernel function was achieved with a larger penalty factor. When the penalty factor exceeded 1, further increasing the penalty factor had little significance for the prediction of oil bearing.

Table 4. Gaussian kernel prediction results of SVM model.

Penalty Factor C	0.1	0.2	0.4	0.5	0.8	1	2	5	10
Accuracy	0.825	0.838	0.851	0.862	0.874	0.889	0.893	0.922	0.94

4.2.2. Oil-Bearing Prediction Based on Artificial Neural Network

The effect of the number of hidden layers and the learning rate on oil bearing was analyzed in this paper. As shown in Table 5, the number of neurons in the hidden layer (12 and 32) was compared and the accuracy of these two network configurations was calculated. It was found that more hidden layers had little effect on the prediction results; thus, the 12-hidden layer neural network was selected to evaluate and predict the oil content in this paper.

Table 5. Accuracy of two types of network models.

Random Split (Training: Testing)	Accuracy	Number of Hidden Layers
9:1	0.7978	12
9:1	0.7980	32

In the parameter update process of the neural network, too large a learning rate may cause the parameters to move back and forth on both sides of the optimal value, and too small a learning rate will greatly reduce the optimization speed. Therefore, the dynamic learning rate concept was introduced in this paper, as shown in Equation (13). For the training process of a neural network, a larger learning rate can quickly obtain the optimal solution at the beginning of training, and then the learning rate is gradually reduced as the number of iterations increases, so that the model is more stable in the later stage of training.

$$l_r = l_{r0} \times D^{\frac{k}{s}} \tag{13}$$

where l_r is the learning rate for the current step, l_{r0} is the initial learning rate, D is the decrease rate, k represents the training time, and s is the control coefficient for learning rate decrease speed.

In this paper, the oil content prediction capability under a dynamic learning rate was analyzed, as shown in Table 6. The predictive results indicated that the artificial neural network performed best with nearly 80% accuracy.

Table 6. Learning rate and decrease rate results.

Random Split (Training: Testing)	Accuracy	Learning Rate	Decay Rate
9:1	0.7980	0.1	0.98

4.2.3. Oil-Bearing Prediction Based on Random Forest

Random forest (RF) can be constructed using several decision tree learners. In this paper, the three variables *AC*, *SP*, and *GR* were used as the input to evaluate the oil content

using the RF model. Random forest models with different configurations were attempted to compare the oil-bearing prediction capability.

It can be seen from Figure 11 that under the same number of decision trees, the prediction accuracy of the model on the testing dataset increased and then decreased with increasing tree depth. As for 101 decision trees, the accuracy of the RF model with the maximum depth of 10 increased by 2.92% and 1.84% respectively compared to RF models with maximum depths of 8 and 12. In addition, the prediction accuracy of the random forest model increased by 0.65% with an increase in the number of decision trees from 61 to 101. As shown in Table 7, 95.07% prediction accuracy on the testing dataset was achieved with 101 decision trees and maximum depth of 10.

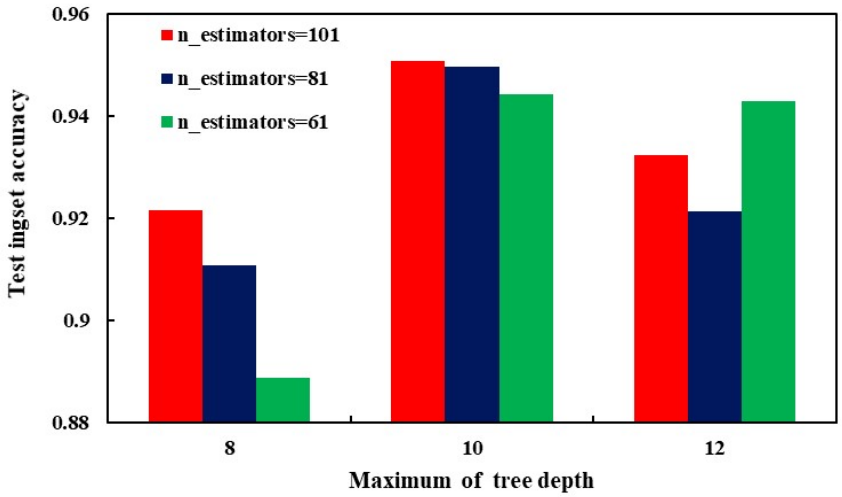


Figure 11. The performance of random forest models with different parameters on the testing dataset.

Table 7. Random forest model.

CV	n Estimators	Max Depth	Training Set Accuracy	Test Set Accuracy
10	101	8	0.9468	0.9215
		10	0.9999	0.9507
		12	0.9871	0.9323
	81	8	0.9519	0.9108
		10	0.9999	0.9496
		12	0.9858	0.9214
	61	8	0.9514	0.8888
		10	0.9999	0.9442
		12	0.9896	0.9428

4.2.4. Integration of Data-Driven Models

Generally speaking, the single data-driven model has limitations, especially when oil field data is of poor quality with incomplete records. In recent years, the combination of surrogate models has been explored. In essence, bagging integrates the sub-model in accordance with the voting method. The results obtained by sampling modeling determine the final voting results through majority voting, as shown in Figure 12. The three machine learning models mentioned above were fused using the bagging method to evaluate the oil-bearing performance. In this paper, 7446 data pairs were obtained by converting the well logging data of 47 wells into well interval data with a spacing of 0.125 m. Based on the interpretation of well logging data, the oil-bearing predictions of the ANN, SVM, and RF models and the integrated model are shown in Table 8. It can be concluded that the RF and fusion models had better prediction capabilities. As shown in Figure 13 and Table 8, the fusion model performed best among these four data-driven models, with 96.5% accuracy.

Unfortunately, the neural network model's prediction capability was the worst, with only 79.8% accuracy. Compared with the neural network, SVM, and random forest models, the accuracy of the fusion model was improved by 20.9%, 8.5%, and 1.4%, which indicated that the integrated model assisted in improving the accuracy of oil-bearing prediction.

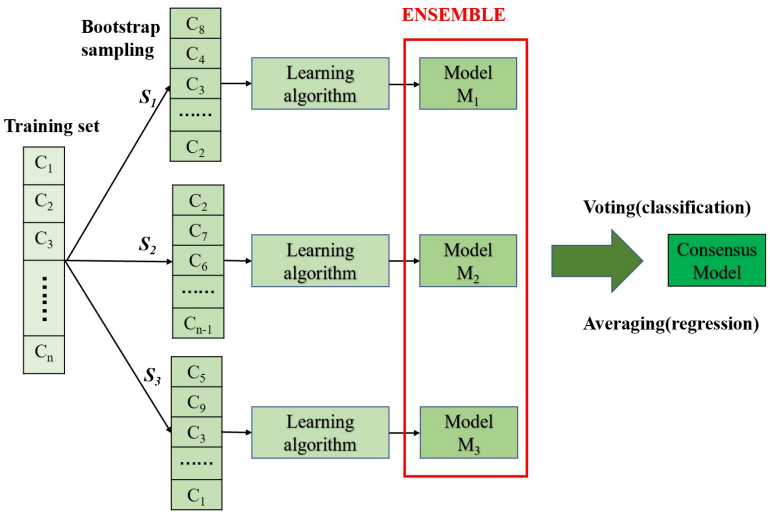


Figure 12. Schematic diagram of bagging.

Table 8. Forecasted results of traditional and data-driven methods.

	Oil Bearing Layers	Interlayers	Accuracy (%)
Test data	1048	6398	100
Well log Interpretation	1118	6328	78.91
Neural Network	1121	6325	79.80
SVM	1098	6348	88.90
Random Forest	1064	6382	95.17
Integrated model	1053	6393	96.50

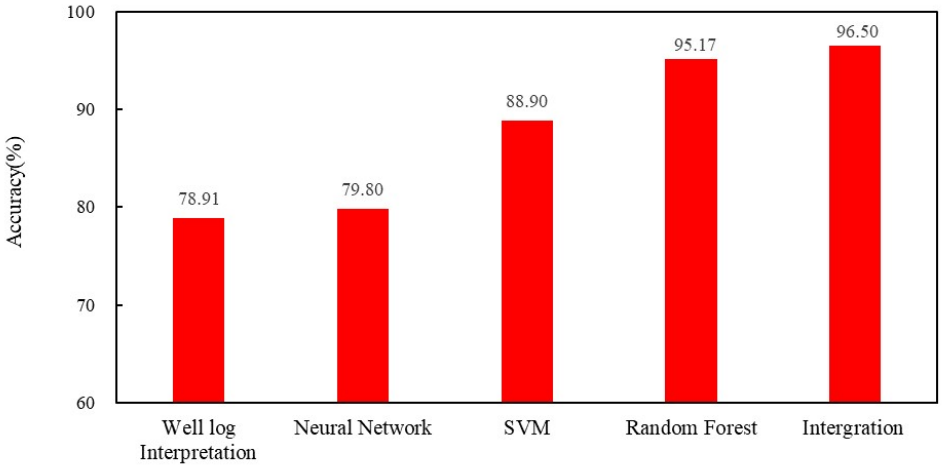


Figure 13. Comparative analysis of data-driven models and well logging method.

4.3. Field Application

4.3.1. Continuous Oil-Bearing Areas (Spot Areas)

The typical wells with long-term production in Changqing oil field, Ordos Basin, China were selected in this paper. As shown in Figure 14, the initial production of these wells was relatively high. After 3 years, the production rate dropped sharply and remained relatively low. However, interpretation of the initial well logging results of this block

showed that its physical properties were relatively good (with 64.23% oil saturation), as shown in Table 9. Therefore, it is reasonable to re-evaluate the oil content of the intervals of wells by integrating the logging data.

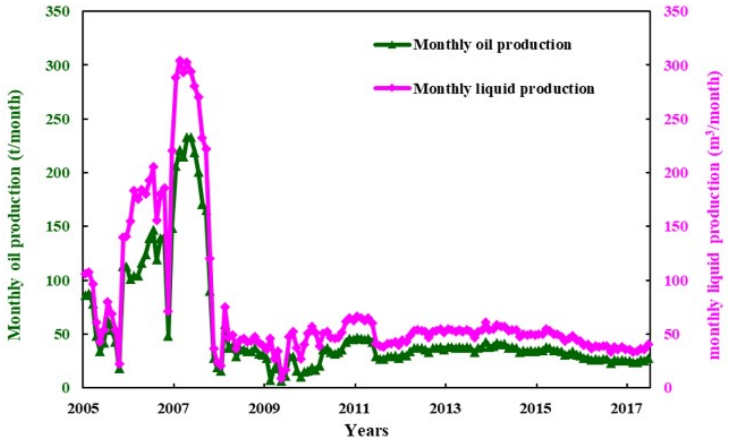


Figure 14. Long-term production rate of a typical well.

Table 9. Physical property results.

Layer Thickness/m	Porosity %	Permeability md	Oil Saturation%	Resistance Ω·m	Conclusion of Explanation	Oil Test	First Birth (First Year Average)
7.3	10.22	1.04	64.23	69.4	Oil-water same layer	18.62 (O) + 37.5 (W)	2.8 (O) + 3.78 (L)

Figure 15 shows the interpretation results of two logging interpretations. The predicted results of the integrated model proposed in this paper for these long-term production wells can be seen in Table 10, which indicated that continuous oil-bearing areas could be found in 2045–2055 m, 2060–2065 m, and 2070–2085 m. The remaining areas had sporadic oil or were continuous oil-free areas. Figure 16 shows the prediction result of oil bearing. The vertical axis is depth, and the blue dots on the x- axis at 0 represents no oil content and at 1 represents oil content.

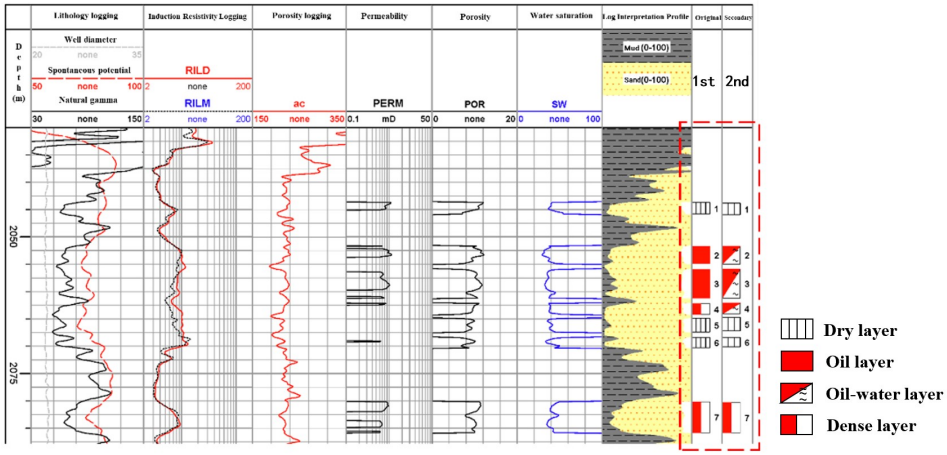


Figure 15. Results of primary and secondary interpretations of wells.

Table 10. Prediction of oil-bearing layers.

Depth (m)	Predicted	Depth (m)	Predicted	Depth (m)	Predicted	Depth (m)	Predicted
2043.75	1	2050.75	1	2057.75	0	2064.75	1
2043.88	1	2050.88	1	2057.88	0	2064.88	1
2044.00	0	2051.00	1	2058.00	0	2065.00	0
2044.13	0	2051.13	1	2058.13	0	2065.13	0
2044.25	0	2051.25	1	2058.25	1	2065.25	0
2048.75	1	2055.75	1	2062.75	1	2069.75	1
2048.88	1	2055.88	1	2062.88	1	2069.88	1
2049.00	1	2056.00	1	2063.00	1	2070.00	1
2049.13	1	2056.13	0	2063.13	1	2070.13	1
2049.25	1	2056.25	0	2063.25	1	2070.25	1

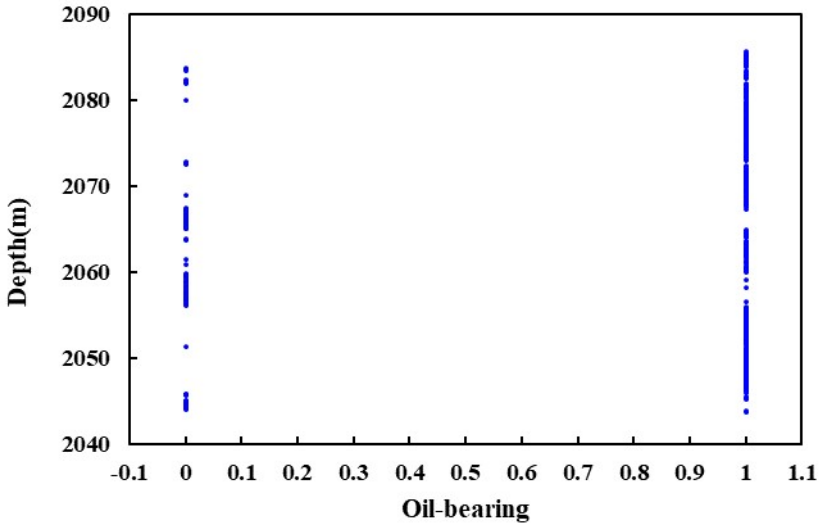


Figure 16. Oil bearing layer prediction.

The predicted results of the remaining layers are shown in the Table 10.

The predicted results of the model were consistent with the results of the first and second logging in most of the intervals.

4.3.2. Prediction of Continuous Interlayers

It is well-known that the interlayer is oil-free, and the predicted results indicated an interlayer if there was less oil content. In addition to the prediction of oil-bearing intervals, this model can also be used for applications related to the identification of interlayers. The oil-bearing prediction was applied the another producing well in the Changqing oil field, Ordos Basin, China. The main purpose of this section focuses on the identification of interlayers.

Figure 17 shows the logging interpretation results of the interlayer. From the oil content prediction results, as shown in Figure 18, interlayers may occur around 1640, 1652, and 1660 m. Similarly, interpretation of the secondary logging results indicated the presence of several interlayers in the 1640–1680 m interval, which validated the predicted results. Table 11 shows some of the predicted results. The different color in the table footer is to distinguish the oil-bearing layer from the interlayer. Figure 18 shows the forecast results of interlayers, where the vertical axis is depth, 0 refers to interlayers, and 1 to oil layers. As can be seen from the figure, there are interlayers around 1640 m,1652 m and 1660 m.

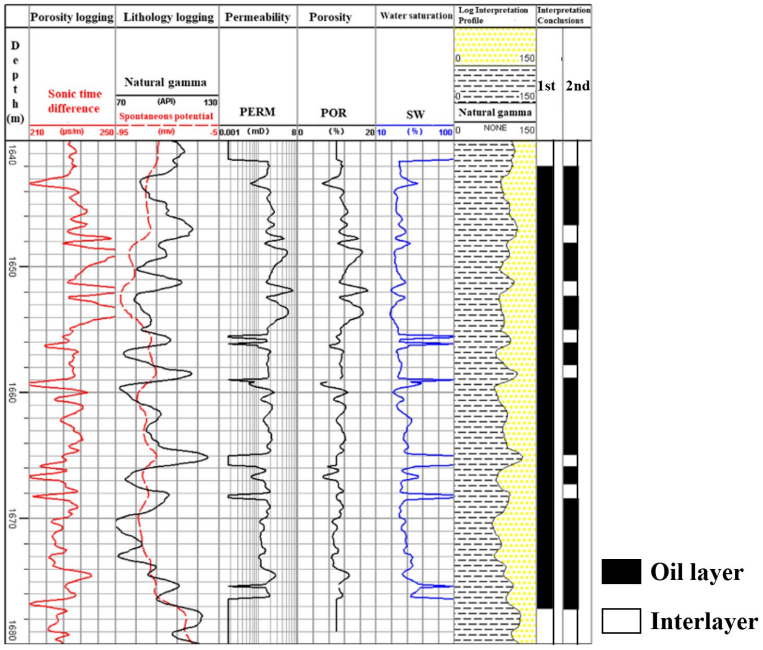


Figure 17. Well loggings of this well.

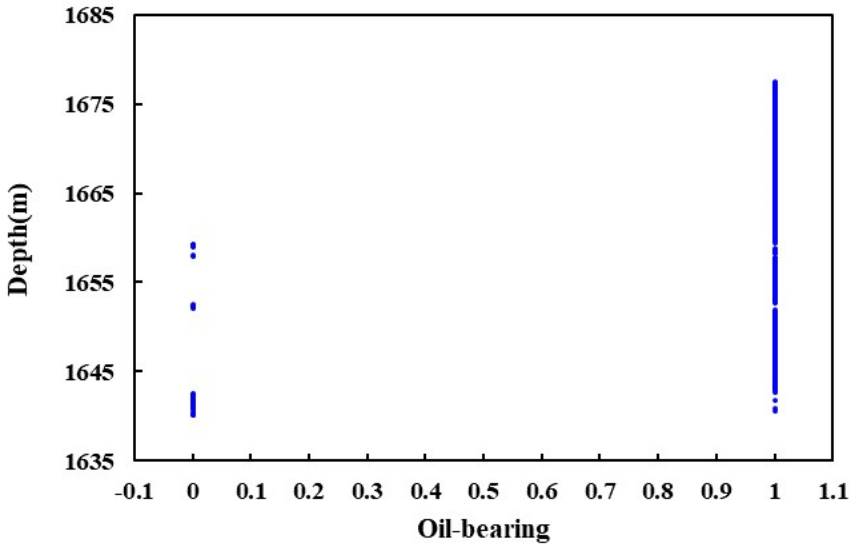


Figure 18. Interlayer predictions of this well.

Table 11. Prediction of interlayers.

Depth (m)	Predicted	Depth (m)	Predicted	Depth (m)	Predicted	Depth (m)	Predicted
1640.08	0	1651.58	1	1655.08	1	1658.58	1
1640.21	0	1651.70	1	1655.20	1	1658.70	1
1640.33	0	1651.83	1	1655.33	1	1658.82	1
1640.46	0	1651.95	1	1655.45	1	1658.95	0
1640.58	1	1652.08	0	1655.58	1	1659.07	0
1640.71	0	1652.20	0	1655.70	1	1659.20	0
1640.83	0	1652.33	0	1655.83	1	1659.32	0
1640.96	1	1652.45	0	1655.95	1	1659.45	1
1641.08	0	1652.58	0	1656.08	1	1659.57	1
1641.21	0	1652.70	1	1656.20	1	1659.70	1
1641.33	0	1652.83	1	1656.33	1	1659.82	1

5. Limitations and Future Work

In this work, the oil bearing of wells with long production periods was re-evaluated using data-driven models. This is a preliminary effort utilizing machine learning models for oil-bearing performance assessment, which combined multiple logging data and quickly determined whether oil was present. This work qualitatively analyzed the oil bearing of wells with long production histories; however, it could not accurately predict oil saturation. Moreover, the applicability of this work was limited by the number of wells and inadequate geological reservoir data. In the near future, we will further collect data from as many wells as possible to predict oil bearing more accurately and quantitatively assess oil saturation using higher quality geological reservoir data.

6. Conclusions

In this paper, three data-driven models were established to re-evaluate the oil bearing of long-term production wells, and an integrated model to improve model accuracy was explored. Combined with the long-term production characteristics of oil wells in the actual oil field, the oil-bearing property was evaluated and the potential sweet spot was found. The conclusions of this paper can be summarized as follows.

(1) In this paper, a simple and feasible procedure for oil-bearing evaluation was explored, which included oil field data collection, perforation data coding, oil-bearing labeling, variable ranking, training and testing data splitting, model training, and oil-bearing prediction
(2) Three data-driven models, including support vector machine, random forest, and artificial neural networks models, were adopted and optimized to predict oil bearing. The results indicated that the RF model performed best with an accuracy of 95.07%, while the neural network model performed the worst with only 79.8% accuracy.
(3) The ensemble learning model was explored to improve the predictive accuracy. Compared with the ANN, SVM, and RF models, the accuracy of this fusion model was improved by 20.9%, 8.5%, and 1.4%, which indicated that the integrated model assisted in improving the accuracy of oil-bearing prediction, which can be utilized to re-evaluate the oil bearing of old wells with long production histories.
(4) The ensemble learning model was applied in the re-evaluation of oil bearing for old wells with long production histories. The potential target sweet spot was found, which might provide theoretical guidance for the effective development of lower production wells in the later period.

Author Contributions: Conceptualization, supervision, and funding acquisition, Y.X.; Methodology, writing—original draft preparation and Visualization, C.C.; Software, Validation, Investigation, Q.J.; Data curation and Formal analysis Q.W. All authors have read and agreed to the published version of the manuscript.

Funding: The authors gratefully acknowledge support by the Strategic Cooperation Technology Projects of China National Petroleum Corporation (CNPC) and China University of Petroleum, Beijing (CUPB) (No. ZLZX2020-02-04).

Data Availability Statement: No new data were created in this study. Data sharing is not applicable to this article.

Acknowledgments: All authors thank Zhenhua Rui, Hua Tian and Longjun Wang for their contributions to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huang, Z.K.; Hao, Y.Q.; Li, S.J.; Wo, Y.J.; Sun, D.S.; Li, M.W.; Chen, J.P. Oil-bearing potential, mobility evaluation and significance of shale oil in Chang 7 shale system in the Ordos Basin: A case study of well H317. *Geol. China* **2020**, *47*, 210–219.
2. Li, J.; Lu, S.; Xue, H.; Xie, L.; Zhang, P. Quantitative evaluation on the elastic property of oil-bearing mudstone/shale from a Chinese continental basin. *Energy Explor. Exploit.* **2015**, *33*, 851–868. [CrossRef]

3. Galkin, S.V.; Martyushev, D.A.; Osovetsky, B.M.; Kazymov, K.P.; Song, H. Evaluation of void space of complicated potentially oil-bearing carbonate formation using X-ray tomography and electron microscopy methods. *Energy Rep.* **2022**, *8*, 6245–6257. [CrossRef]
4. Feng, M.; Wang, X.; Du, Y.; Meng, W.; Tian, T.; Chao, J.; Wang, J.; Liang, Z.; Xu, Y.; Xiao, W. Organic geochemical characteristics of shale in the lower sub-member of the third member of Paleogene Shahejie Formation (Es3L) in Zhanhua Sag, Bohai Bay Basin, eastern China: Significance for the shale oil-bearing evaluation and sedimentary environment. *Arab. J. Geosci.* **2022**, *15*, 375. [CrossRef]
5. Li, J.; Huang, W.; Lu, S.; Wang, M.; Chen, G.; Tian, W.; Guo, Z. Nuclear magnetic resonance T_1–T_2 map division method for hydrogen-bearing components in continental shale. *Energy Fuels* **2018**, *32*, 9043–9054. [CrossRef]
6. Gao, C.; Chen, Y.; Yin, J.; Liang, Q.; Hao, S.; Zhang, L.; Zhao, Q.; Sun, J.; Xu, J. Experimental study on reservoir characteristics and oil-bearing properties of Chang 7 lacustrine oil shale in Yan'an area, China. *Open Geosci.* **2022**, *14*, 234–251. [CrossRef]
7. Zhang, R.; Tang, L. Oil-bearing evaluation of different lithofacies in Da'anzhai Member, Central Sichuan Basin: Implications for shale oil development. *Pet. Sci. Technol.* **2022**, *44*, 1–15. [CrossRef]
8. Li, Q.; Jiang, Z.; Liu, K.; Zhang, C.; You, X. Factors controlling reservoir properties and hydrocarbon accumulation of lacustrine deep-water turbidites in the Huimin Depression, Bohai Bay Basin, East China. *Mar. Pet. Geol.* **2014**, *57*, 327–344. [CrossRef]
9. Feng, C.; Yang, H.; Pu, R.; Wang, Y.; Wang, D.; Liang, X.; Zhang, M.; Huang, Y.; Fei, S. Lithology and Oil-Bearing Properties of Tight Sandstone Reservoirs: Chang 7 Member of Upper Triassic Yanchang Formation, Southwestern Ordos Basin, China. *Geosci. J.* **2017**, *21*, 201–211. [CrossRef]
10. Chen, R.; Zhan, H.; Li, Y.; Miao, X.; Yue, W.; Zhao, K. Thermal terahertz analysis (TTA) for detecting the oil bearing features in a desert reservoir. *J. Pet. Sci. Eng.* **2021**, *197*, 107966. [CrossRef]
11. Gu, Z.; Xia, T.; Liu, C.; Chen, J. Application of a new attribute and enhancement technology in oil and gas detection. In *First International Meeting for Applied Geoscience & Energy*; Society of Exploration Geophysicists: Houston, TX, USA, 2021; pp. 1161–1165.
12. Li, F.; Guo, X.; Deng, Y.; Zhou, L.; Zhao, X.; Baoli, Y. Applications of low-frequency information in hydrocarbon detection: Two cases in Junggar and Tuha Basins. In Proceedings of the SEG International Exposition and Annual Meeting, Dallas, TX, USA, 16–21 October 2016; OnePetro: Richardson, TX, USA, 2016.
13. Song, X.; Zhang, F.; Sang, K.; Yang, J. Application of direct inversion of reservoir parameters in sweet spot reservoir prediction: A case study of Baisan formation in Junggar Basin. In Proceedings of the SEG International Exposition and Annual Meeting, San Antonio, TX, USA, 15–20 September 2019; OnePetro: Richardson, TX, USA, 2019.
14. Roueché, J.N.; Karacan, C.Ö. Zone identification and oil saturation prediction in a waterflooded field: Residual oil zone, East Seminole Field, Texas, USA, Permian Basin. In Proceedings of the SPE Improved Oil Recovery Conference, Tulsa, OK, USA, 14–18 April 2018; OnePetro: Richardson, TX, USA, 2018.
15. Hidayat, H.; Supriady, S.; Anwar, T.; Ariestya, G.; Giriansyah, B. Revealing Hidden Potentials to Maintain Production by Using Integrated Petrophysical Interpretation Techniques in Mahakam Mature Fields. In Proceedings of the SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition, Bali, Indonesia, 29–31 October 2029; OnePetro: Richardson, TX, USA, 2020.
16. Liang, X.; Zou, C.-C.; Mao, Z.-Q.; Shi, Y.-J.; Li, G.-R.; Gao, H.-P.; Xie, X.-H. The Correlation Analysis Method and Its Application in Hydrocarbon-Bearing Formation Identification in Tight Sandstone Reservoirs. In Proceedings of the SPE Unconventional Resources Conference and Exhibition-Asia Pacific, Brisbane, Australia, 11–13 November 2013; OnePetro: Richardson, TX, USA, 2013.
17. Qigui, C.; Shaobin, G.; Haihong, W.; Chengyu, W.; Xiaowei, L. Comprehensive reservoir evaluation of Chang-6 oil-bearing layers in Midwest Ordos Basin. *Pet. Geol. Exp.* **2010**, *32*, 415–419.
18. Huang, S. Entropy model of gray systems and its application in oil-bearing area evaluation. *J. Geomech.* **2006**, *12*, 77–83.
19. Dan, W.; Cheng, Q.; Niu, X. Integrated evaluation of low permeability reservoirs of Chang 4+ 5-Chang 8 Formations of main oil-bearing blocks in Ordos Basin. *J. Oil Gas Technol.* **2011**, *33*, 48–53.
20. Freeman, D.W.; Fenn, C.J. An evaluation of various logging methods for the determination of remaining oil saturation in a mixed salinity environment. In Proceedings of the Middle East Oil Show, Bahrain, Bahrain, 11–14 March 1989; OnePetro: Richardson, TX, USA, 1989.
21. Khatchikian, A. Log evaluation of oil-bearing igneous rocks. In Proceedings of the SPWLA 23rd Annual Logging Symposium, Corpus Christi, TX, USA, 6–9 July 1982; OnePetro: Richardson, TX, USA, 1982.
22. Cao, C.; Jia, P.; Cheng, L.; Jin, Q.; Qi, S. A review on application of data-driven models in hydrocarbon production forecast. *J. Pet. Sci. Eng.* **2022**, *212*, 110296. [CrossRef]
23. Ilyushin, Y.V. Development of a Process Control System for the Production of High-Paraffin Oil. *Energies* **2022**, *15*, 6462. [CrossRef]
24. Duplyakov, V.M.; Morozov, A.D.; Popkov, D.O.; Shel, E.V.; Vainshtein, A.L.; Burnaev, E.V.; Osiptsov, A.A.; Paderin, G.V. Data-driven model for hydraulic fracturing design optimization. Part II: Inverse problem. *J. Pet. Sci. Eng.* **2022**, *208*, 109303. [CrossRef]
25. Makhotin, I.; Orlov, D.; Koroteev, D. Machine Learning to Rate and Predict the Efficiency of Waterflooding for Oil Production. *Energies* **2022**, *15*, 1199. [CrossRef]
26. Cao, C.; Cheng, L.; Jia, Z.; Jia, P.; Zhang, X.; Xue, Y. An Integrated Model Combining Complex Fracture Networks and Time-Varying Data Modeling Techniques for Production Performance Analysis in Tight Reservoirs. In Proceedings of the SPE EuropEC—Europe Energy Conference Featured at the 83rd EAGE Annual Conference & Exhibition, Madrid, Spain, 6–9 June 2022.

27. Cao, C.; Cheng, L.; Zhang, X.; Jia, P.; Shi, J. Seepage proxy model and production forecast method based on multivariate and small sample. *Chin. J. Theor. Appl. Mech.* **2021**, *53*, 2345–2354.
28. Martirosyan, A.V.; Ilyushin, Y.V. Modeling of the Natural Objects' Temperature Field Distribution Using a Supercomputer. *Informatics* **2022**, *9*, 62. [CrossRef]
29. Guo, Q.; Zhuang, T.; Li, Z.; He, S. Prediction of reservoir saturation field in high water cut stage by bore-ground electromagnetic method based on machine learning. *J. Pet. Sci. Eng.* **2021**, *204*, 108678. [CrossRef]
30. Huang, R.; Wei, C.; Yang, J.; Xu, X.; Li, B.; Wu, S.; Xiong, L. Quantitative analysis of the main controlling factors of oil satu-ration variation. *Geofluids* **2021**, *2021*, 6515846.
31. Singh, H.; Seol, Y.; Myshakin, E.M. Prediction of gas hydrate saturation using machine learning and optimal set of well-logs. *Comput. Geosci.* **2021**, *25*, 267–283. [CrossRef]
32. Peirolo, R. Information gain as a score for probabilistic forecasts. *Meteorol. Appl.* **2011**, *1*, 9–17. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.