



Article

# Tight Oil Well Productivity Prediction Model Based on Neural Network

Yuhang Jin \* , Kangliang Guo, Xinchun Gao and Qiangyu Li 

School of Geosciences, Yangtze University, Wuhan 430100, China; jpickguo@sina.com (K.G.); 201971270@yangtzeu.edu.cn (X.G.); liqiangyu2000@163.com (Q.L.)

\* Correspondence: 2022710423@yangtzeu.edu.cn

**Abstract:** Productivity prediction has always been an important part of reservoir development, and tight reservoirs need accurate and efficient productivity prediction models. Due to the complexity of the tight oil reservoir, the data obtained by the detection instrument need to extract data features at a deeper level. Using the Pearson correlation coefficient and partial correlation coefficient to analyze the main control of productivity factors, eight characteristic parameters of volume coefficient, water saturation, density, effective thickness, skin factor, shale content, porosity, and effective permeability were obtained, and the specific oil production index was used as the target parameter. Two sample structures of pure static parameters and dynamic and static parameters (shale content, effective permeability, porosity, water saturation, and density as dynamic parameters, volume coefficient, skin factor, and effective thickness as static parameters) were created, and corresponding model structures (BP (Backpropagation), neural network model, and LSTM-BP (Long Short-Term Memory Backpropagation) neural network model) were designed to compare the prediction effects of models under different sample structures. The mean absolute error, root mean square error, mean relative percentage error, and coefficient of determination were used to evaluate the model results. The LSTM-BP neural network was used to predict the production capacity of the test set. The results showed that the average absolute error was 0.07, the root mean square error was 0.10, the average absolute percentage error was 21%, and the coefficient of determination was 0.97. Using wells in the WZ area for testing, the LSTM-BP model's predictions are evenly distributed on both sides of the 45° line, separating the predicted values from actual values, with errors from the line being relatively small. In contrast, the BP model and analytical method are unable to achieve such an even distribution around the line. Experiments show that the LSTM-BP neural network model can effectively extract dynamic parameter features and has a stronger generalization ability.

**Keywords:** tight oil reservoirs; production capacity prediction; neural networks



**Citation:** Jin, Y.; Guo, K.; Gao, X.; Li, Q. Tight Oil Well Productivity Prediction Model Based on Neural Network. *Processes* **2024**, *12*, 2088. <https://doi.org/10.3390/pr12102088>

Academic Editors: Qingbang Meng, Bin Liang and Zhan Meng

Received: 6 August 2024

Revised: 18 September 2024

Accepted: 21 September 2024

Published: 26 September 2024



**Copyright:** © 2024 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

The tight reservoir in the WZ block is located in the Beibu Gulf sea area. The reservoir has the characteristics of strong heterogeneity, complex pore structure, and poor pore connectivity. The porosity is low, usually between 5% and 20%; the permeability is low, distributed between 0.04 and 140 md, most of which is within 10 md. At the same time, the water saturation of tight reservoirs is high, usually between 30% and 90%.

Machine learning originated in the 1950s [1,2] and was not applied to oil and gas production until the 1980s [3–5]. In the 21st century, with the gradual expansion of machine learning algorithms and the improvement in computational power, machine learning has begun to make significant strides in various fields. Jani D.B. (2017) used an ANN (Artificial Neural Network) model to accurately predict the performance of solid desiccant cooling systems [6]. Nasirzadeh F. (2020) employed an ANN model combined with the PI method to provide a novel approach for predicting labor productivity [7]. Soroush Ahmadi (2024) utilized the response surface method to optimize the corrosion inhibition performance of

2-mercaptobenzothiazole (2-MBT) for carbon steel in 1 M HCl [8]. Machine learning is becoming increasingly prevalent in oil and gas production [9–11]. Due to the complexity of low-permeability tight reservoirs, productivity is affected by many factors. In order to obtain more accurate productivity prediction values, determining the main control factors can reduce the complexity of the model and accelerate the training speed of the model. Li X. (2013) used the three parameters of permeability, porosity, and first closure pressure as input parameters of the model to predict productivity [12]. Cao Q. (2016) used thickness, average porosity, average clay content, and density as the inputs for the model, but the prediction accuracy on the test set was low [13]. The data features mapped by the normal parameters cannot fully represent the capacity data and finally affect the results of capacity prediction. With the improvement in the feature extraction ability of the model, researchers began to add fracturing information to the model parameters. Researchers, such as Alimkhanov Rustam (2014), Yue Ming (2024), Wu Lei (2023), and Qin Ji (2022), added the information generated by fracturing to the model [14–17]. Alimkhanov Rustam (2014) [17] used the geological information (total thickness and net thickness, porosity, permeability coefficient, oil saturation coefficient, net-to-gross ratio, reservoir interval, macroscopic heterogeneity, transmission) before and after fracturing in the Povkh oilfield as parameters to predict the fracturing effect and then optimize the fracturing parameters. Wu Lei (2023) [14] and Qin Ji (2022) [15] used the fracture half-length and the number of fractures as part of the parameters to generate a model that can be used to optimize the fracturing parameters.

In recent years, with the widespread promotion of machine learning, some researchers have used data mining methods to predict production capacity [18–21]. Dong (2022) [19] used the regression tree model and combined the Spearman correlation coefficient and recursive feature elimination algorithm to rank the importance of influencing factors and predict the initial production capacity. Hui G. (2021) [21] uses four methods (linear regression, neural network, regression tree, and decision tree) to predict the natural gas production capacity in the Fox Creek area. Regression tree and decision tree methods have better prediction accuracy. With improvements in development needs, a single model can no longer meet the actual production needs, so researchers began to study the composite model [22,23]. Liu Jie (2023) [22] used the KNN-BP (K-Nearest Neighbors Backpropagation) neural network model to predict the productivity of tight sandstone gas reservoirs in the SM block. Compared with the single network model (BP neural network) and other algorithms (support vector machine, random forest, linear regression), the prediction accuracy of the composite neural network model is higher. Fargalla and Mandella Ali M. (2024) [23] proposed a new model called TimeNet. The model combines a convolutional neural network, a bidirectional gate cycle control unit, an attention mechanism, and Time2 Vec, which can not only capture complex nonlinear time information but also extract formation spatial characteristics. It has a very good effect on the productivity prediction of the Fenchuganj conventional sandstone gas field and the Marcellus shale gas field. For different problems, the selection of the model will also have a focus. Machine learning models are divided into unsupervised and supervised models. Unsupervised models, such as the K-means (K-means Clustering) model, are suitable for clustering analysis. Li Yuanzheng (2022) used the K-means model to classify reservoirs based on pore structure parameters [24]. Supervised models include support vector machines, long short-term memory neural models, deep neural models, etc. Based on the long-term and short-term memory neural models, Fu (2023) predicted the production capacity with time-series parameters, which were used to predict daily production capacity data [25].

Considering the special reservoir physical conditions in the WZ area and the large gap between the specific oil production index of some wells, there may be a linear relationship between the main control factors of productivity and the specific oil production index, while the neural network model is just good at dealing with nonlinear problems. Furthermore, conventional neural network models often use static parameter samples for training and prediction, meaning that parameters with dynamic changes are simply averaged or replaced with the median within their change range to treat them as static parameters.

However, reservoirs in the WZ region are classified as tight reservoirs, and their reservoir properties change dramatically with depth, making a static approach inappropriate. Pearson correlation analysis and bias analysis are used to identify the main controlling factors for productivity. Subsequently, dynamic parameter samples are constructed from the main controlling factors with depth variations in the original data. These dynamic parameters are then processed using an LSTM (Long Short-Term Memory) network to convert them into static parameters. The remaining controlling factors are combined with the converted static parameters to serve as the input for a BP network, which is then used to predict productivity. The LSTM network excels at extracting features from dynamic data and performs better than simply averaging or using the median of the change range for parameters with dynamic variations. Meanwhile, the BP network is well suited for handling nonlinear problems and is convenient to set up, with a short training cycle.

## 2. Production Capacity Prediction Model Establishment Process

(1) Problem Description. In view of the special productivity conditions in this area, the specific productivity index under multi-mechanism control is transformed into the specific productivity index affected by fixed-characteristic parameters (logging parameters, crude oil parameters, formation parameters, construction parameters), and a set of mathematical models from multi-characteristic parameters to target parameters is formed. This helps to simplify the description of practical problems and facilitate the establishment of capacity prediction models quickly and efficiently.

(2) Data Preparation. The neural network model belongs to supervised learning, and the sample data set must be involved to update the model parameters. Before training, the sample set needs to be constructed from the original data according to the initially selected feature parameters. The sample set will be divided into a training set, a validation set, and a test set. The training set is used for the forward propagation of the model to form a computational network; the validation set is used to update the parameters of the computing network; and the test set is used to evaluate the model prediction effect after each round of parameter updates.

(3) Establish and optimize the model. Because there are more logging data and physical property data in the original data, other geological parameter data are limited. Considering the inconsistency in the amount of data, the data set is established in two ways: a single sample is taken as a static parameter by the average value of each influencing parameter; the logging and physical property influence parameters of continuous depth are taken as part of dynamic parameters; and other parameters are taken as static parameters. Two network models, the BP neural network and the LSTM-BP neural network, are established correspondingly. The model optimization is divided into hyperparameters and parameter optimization. The hyperparameters are manually set before the model is established, and the parameters are optimized by the loss function and the optimization function in the model function calculation based on the training set and the verification set. The selection of hyperparameters and parameters determines the prediction effect of the model.

(4) Model evaluation. The test set is input into the optimized model, and the predicted value is output. Four prediction evaluation indices were selected: mean absolute error (*MAE*), root mean square error (*RMSE*), mean absolute percentage error (*MAPE*), and coefficient of determination ( $R^2$ ). These were used to evaluate the prediction effect of the model. Through four evaluation indices, the model with the best prediction effect is selected.

(5) Model application. The actual production application is carried out, the verification samples are collected from the original data according to the main influencing factors of capacity, and the samples are input into the model for capacity prediction.

### 3. Data Feature Preprocessing

#### 3.1. Data Reduction

After collecting the original data of the study area, the influencing factors related to tight oil wells are preliminarily screened, mainly including logging parameters (natural gamma, acoustic time difference, density, deep resistivity, shale content), formation parameters (effective permeability, porosity, volume coefficient, skin factor, effective thickness, water saturation), crude oil parameters (oil relative density, crude oil viscosity), and construction parameters (oil well radius). Some parameters are not dimensionalized for calculation (Table 1).

**Table 1.** Factors affecting production capacity.

Parameter Name	Minimum Value	Maximum Value
volumetric coefficient	0.7	1.6
water saturation	0.3	0.6
density/(g·cc <sup>-1</sup> )	2.4	2.6
deep resistivity/ohmm	7.5	49
effective thickness/m	2.92	38.2
Oil well radius/m	0.08	0.16
oil viscosity/(Mpa·s <sup>-1</sup> )	0.3	3.62
Relative density of oil	0.81	0.88
gamma ray/api	49.5	164.9
skin factor	−2	5.5
acoustic interval transit time/(us·ft <sup>-1</sup> )	69.2	96.6
argillaceous content	0.051	0.199
porosity	0.1 138	0.2 136
effective permeability/md	0.05	140

#### 3.2. Analysis of Main Control Parameters of Production Capacity

Before establishing the two network models, it is necessary to consider whether the characteristic parameters of the data set are correlated with the predictors and whether the characteristic parameters are intersected. Feature selection can not only reduce the number of feature parameters, accelerate the training time of the model, and reduce the possibility of overfitting but also improve the generalization ability of the model so that it can maintain a certain robustness in practical applications and reduce the prediction effect on the test set. The actual application of the prediction effect is not good.

The data set structure of the two models is different. The data set combined with dynamic parameters and static parameters is essentially used to extract the internal mathematical information of dynamic parameters. It is the same as the static parameter data set in purpose, and the dynamic and static parameter data sets make it difficult to analyze the main control factors. Therefore, the data set composed of all static parameters is used to analyze the main control parameters. Peel correlation analysis and partial correlation analysis are jointly involved in feature selection. The purpose is to remove the parameters with weak correlation with the target parameters and retain the feature parameters with weak correlation between the feature parameters and strong correlation with the target parameters. In this way, the difficulty of model training is reduced, and the speed of model fitting is accelerated.

First, Pearson correlation analysis was used. Pearson correlation analysis is a statistical method used to measure the degree of linear correlation between two variables. It is based on the Pearson correlation coefficient, which is usually represented by the symbol ' $r$ '. The Pearson correlation coefficient is between  $-1$  and  $1$ . It reveals the strength of relationships, which helps to understand which factors have the greatest impact on productivity. Pearson correlation analysis is a statistical method used to measure the degree of linear correlation between two variables. On one hand, using the Pearson correlation coefficient can identify key factors that have a significant linear correlation with productivity. On the other hand,

it can also identify other nonlinear factors that do not exhibit linear correlation, providing a basis for the use of neural networks.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{Y})^2}} \quad (1)$$

Equation (1):  $r$  is the correlation coefficient;  $n$  is the number of samples;  $x_i$  and  $y_i$  are two characteristic parameter values;  $\bar{X}$ ,  $\bar{Y}$  is the mean value of two characteristic parameters.

Through the Pearson correlation analysis (Figure 1), it can be found that the volume coefficient, water saturation, acoustic time difference, shale content, porosity, and effective permeability have a high correlation with the specific oil production index. In the figure, the squares closer to blue represent a stronger negative correlation, while those closer to red indicate a stronger positive correlation. The correlation coefficient between the specific productivity index and oil well radius, crude oil density, and oil relative density is low. It is worth noting that the specific productivity index has a high correlation with the shale content, but the correlation with the relative density of oil is very low. At the same time, there is a certain correlation between the shale content and the relative density of oil, indicating that there is a nonlinear correlation between the relative density of oil and the specific productivity index (the absolute value of the correlation coefficient is close to 1, indicating that the higher the linear correlation between the two, the closer the absolute value is to 0, indicating that there is a nonlinear correlation between the two). Therefore, the use of the neural network structure is more helpful to fit nonlinear problems.

Partial correlation measures the relationship between two variables while controlling for the influence of other variables. It helps us control for confounding factors, allowing us to more accurately assess the independent effect of a specific factor on productivity. Through partial correlation analysis, it can be found (Figure 2) that the contribution of the oil well radius, crude oil viscosity, oil relative density, and natural gamma contrast oil production index is small. At the same time, water saturation, deep resistivity, acoustic time difference, and shale content are in the same contribution interval. In the selection of parameters, the parameters with a small contribution and high correlation with other parameters can be eliminated; the parameters with the same contribution can be used to distinguish the correlation of the parameters with other parameters and the parameters that can contain other unselected parameters and will not conflict with the existing parameters can be selected.

Pearson correlation coefficients help to initially identify factors related to productivity and the strength of their relationships. Partial correlation, on the other hand, provides a more refined analysis by controlling for the influence of other variables, helping to determine the independent effect of a specific factor. Combined with Pearson correlation analysis and partial correlation analysis, these four parameters are eliminated because the contribution of the oil well radius, crude oil viscosity, oil relative density, and natural gamma is too small. Density and deep resistivity, as well as acoustic time difference and clay content, have a relatively equal contribution. Density has a significant correlation with water saturation and volume coefficient and also has a certain correlation with unselected natural gamma. The deep resistivity has a great correlation with the effective thickness and volume coefficient and has a certain correlation with the density, and the deep resistivity is eliminated. Acoustic time difference and mud content are the same, and the acoustic time difference is eliminated. The final training sample parameters are volume coefficient, water saturation, density, effective thickness, skin factor, shale content, porosity, and effective permeability. The specific oil production index is used as the target parameter.

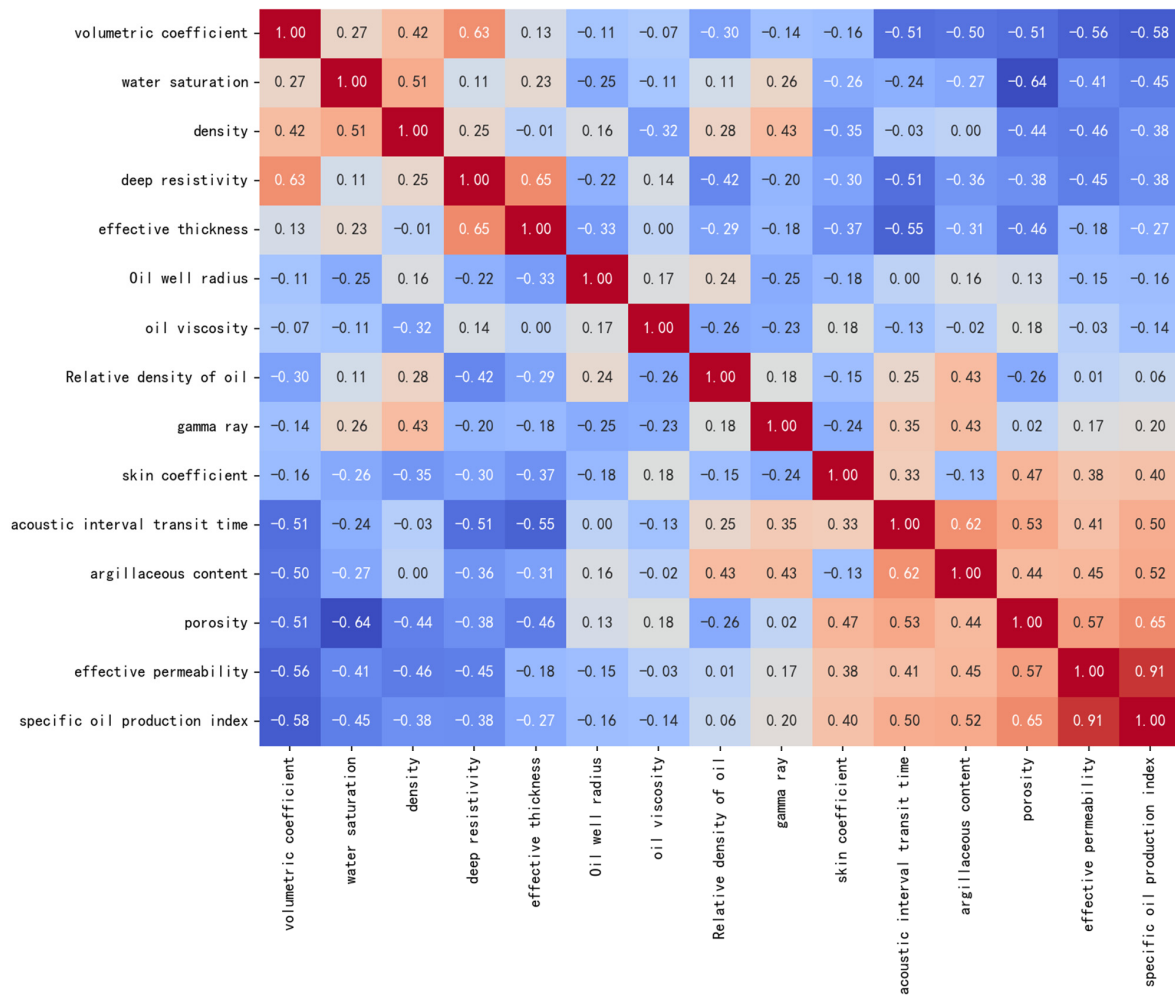


Figure 1. Heat map of main control parameters.

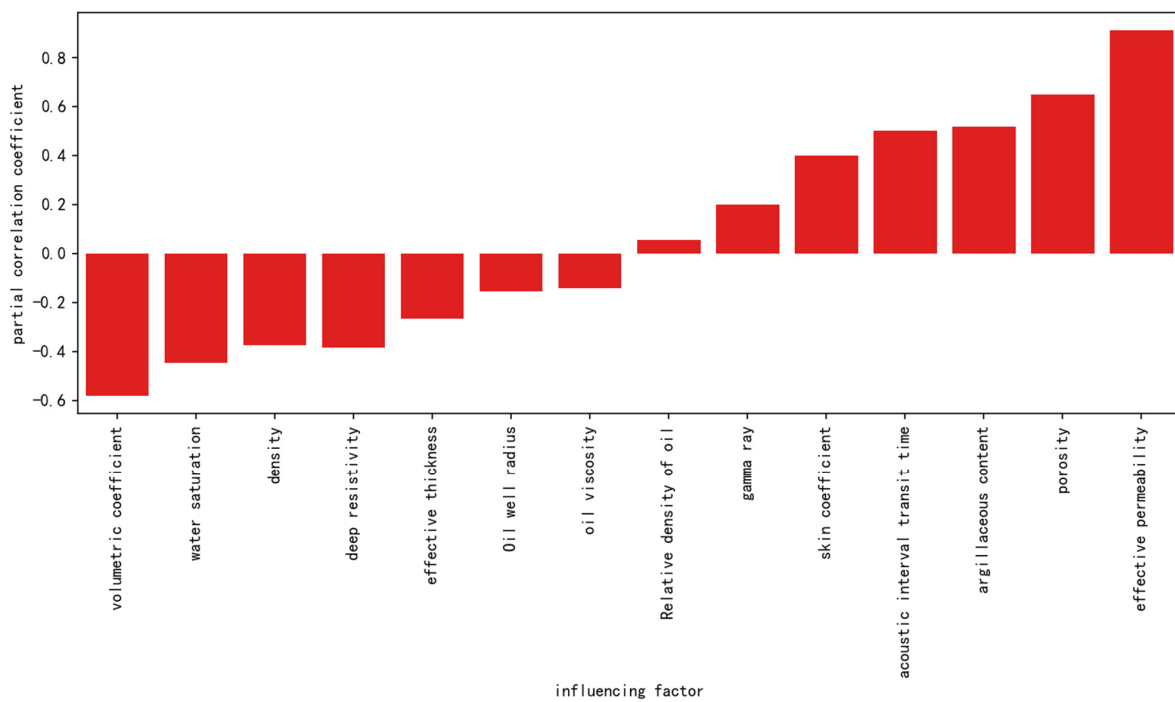


Figure 2. Partial correlation analysis.

### 3.3. Data Processing

Considering the specific application scenarios of the problem, the data set is divided into three parts: training set, verification set, and test set. The training set and the validation set jointly optimize the model parameters; that is, they establish the functional relationship between the characteristic parameters and the specific oil production index. The test set itself does not participate in the training process. Its role is to evaluate the optimized model prediction effect and generalization ability after the model establishes a complete prediction function, so that the model has the ability to process unknown data. The prediction effect of the model evaluated using the unused validation set can improve the authenticity and applicability of the model, and the prediction accuracy obtained by using the test set is more reliable.

In the raw data, it is inevitable that some parameters may be missing or anomalous. To address this issue, the nearest-neighbor algorithm is used. In machine learning, the data characteristics of similar parameters are also similar. The basic process is as follows: identify the missing or anomalous parts of the sample parameters; find other samples with the same missing or anomalous values to form a data set; calculate the Euclidean distance between these samples; select the top-three samples with the smallest Euclidean distances as similar samples for the sample to be repaired; and calculate the average value of the parameters of these similar samples to use as the replacement value.

In the original data, because the characteristic parameters have their own physical dimensions and the numerical distribution range of each other is inconsistent, it is bound to cause the model to be difficult to fit in the model fitting process. In order to make the characteristic parameters comparable, the unified dimension method is used to express the characteristic parameters with percentages in decimal numbers, and then all the characteristic parameters are normalized. Data normalization can convert data of different variables into the same range, so that the value of each sample datum falls between 0 and 1. For a certain feature  $x$  in the same sample set, its normalization formula is:

$$x' = \frac{(x - X_{min})}{X_{max} - X_{min}} \quad (2)$$

Equation (2):  $x'$  is the normalized data of characteristic parameters;  $x$  is the original data of characteristic parameters;  $x_{min}$  is the minimum value of characteristic parameters;  $x_{max}$  is the maximum value of characteristic parameters. In the whole process of model training, it is normalized according to different sample sets. At the same time, in the evaluation model, the model prediction results belong to the normalized results, and the anti-normalization operation is needed to obtain the real prediction data.

## 4. Model Establishment and Evaluation

### 4.1. Model Evaluation Indicators

The mean absolute error (MAE) is the mean of the absolute error between the predicted value and the actual value. It is a linear index, which means that all individual differences have equal weight on the average. It can clearly indicate the gap between the predicted value and the actual value, without considering the direction of the error, is not sensitive to outliers, and the calculation is simpler.

$$\eta_{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3)$$

Equation (3):  $\eta_{MAE}$  is the mean absolute error,  $m^3 \cdot d^{-1} \cdot Mpa^{-1}$ ;  $y_i$  is the real value of specific oil production index,  $m^3 \cdot d^{-1} \cdot Mpa^{-1}$ ;  $\hat{y}_i$  is the predicted value of the specific oil production index,  $m^3 \cdot d^{-1} \cdot Mpa^{-1}$ .

The root mean square error (RMSE) is the square root of the mean square of the difference between the predicted value and the actual observed value. It takes into account the direction of the difference between the predicted value and the actual value, so it can

better reflect the accuracy of the model. It has good properties in mathematics, such as differentiability, so that it can be easily used in optimization algorithms. The smaller its value, the better the prediction effect of the model. The expression is:

$$\eta_{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4)$$

Equation (4):  $\eta_{RMSE}$  is the root mean square error.

The mean absolute percentage error ( $MAPE$ ) represents the percentage of the mean prediction error relative to the actual value, so it is easier to understand and can be used to compare the prediction performance of different time-series data sets because it is the percentage relative to the actual value. Compared with other evaluation criteria, it is more intuitive and relatively measured. The smaller its value is, the better its expression is:

$$\eta_{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (5)$$

Equation (5):  $\eta_{MAPE}$  is the mean absolute percentage error.

The coefficient of determination ( $R^2$ ) represents the proportion of the variance in the dependent variable that can be explained by the independent variable, ranging from 0 to 1. Without the influence of dimension, the coefficients of determination between different data sets can be directly compared. The closer to 1, the stronger the model's ability to explain the data, and vice versa.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (6)$$

Equation (6):  $R^2$  is the coefficient of determination;  $\bar{y}_i$  is the true average value of the specific oil production index,  $\text{m}^3 \cdot \text{d}^{-1} \cdot \text{Mpa}^{-1}$ .

#### 4.2. Model Design and Evaluation

A BP neural network is usually composed of an input layer, a hidden layer, and an output layer. The input layer accepts data input from outside, and each input node corresponds to a feature or attribute of the input data. The hidden layer is located between the input layer and the output layer and can have multiple layers. Each layer contains multiple neurons (nodes), and each neuron receives input from the previous layer of neurons and generates output for the next layer (Figure 3). The BP neural networks are characterized by a simple network structure and a short training time. They can achieve good training results with fully static parameter samples. However, in productivity prediction problems, there is a process of converting dynamic parameters to static ones, and using the average or median of dynamic values can inevitably lead to a loss of feature data. The BP model is a conventional network model in neural networks, and using the BP neural network as a control group is of significant importance.

The BP neural structure uses sample data with full static parameters (Figure 4). Using the root mean square error as the loss function to calculate the loss value makes it easier for gradient descent methods to update the weights. The Backpropagation algorithm is used to change the model weight and bias until it converges to a more reasonable range. Finally, the test set is used to evaluate the model.

Considering the large span of each layer, the selected parameter value is the average value of the continuous value of the layer. Due to the large span of the layer and the uneven distribution of crude oil in the layer, the selection of the average value will cause a certain error. At the same time, each layer will be divided into smaller layers, and the number is not uniform. In order to extract this layer information more accurately, the LSTM long-term and short-term memory networks are selected.



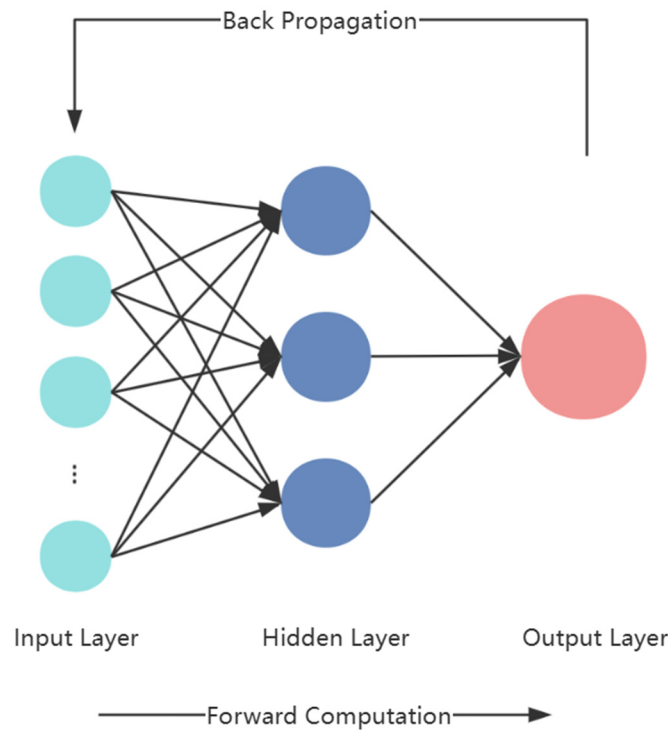


Figure 3. BP (Backpropagation) Neural Network structure.

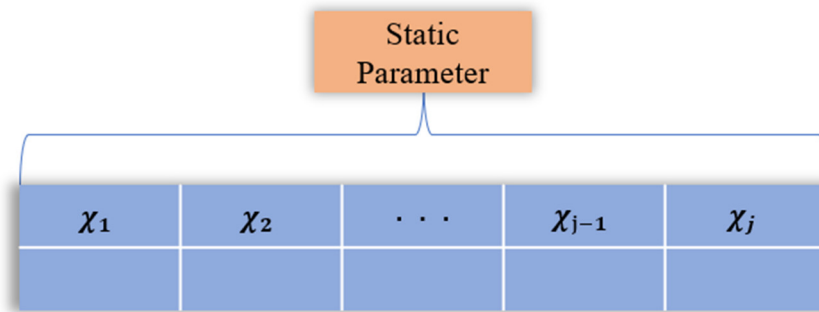


Figure 4. Sample structure of BP (Backpropagation) neural networks.

According to the actual raw data sources, shale content, effective permeability, porosity, water saturation, and density change with depth, which are dynamic parameters. The volume coefficient, skin factor, and effective thickness are derived from logging data and belong to static data. The LSTM-BP neural structure uses a sample structure that combines dynamic and static parameters (Figure 5).

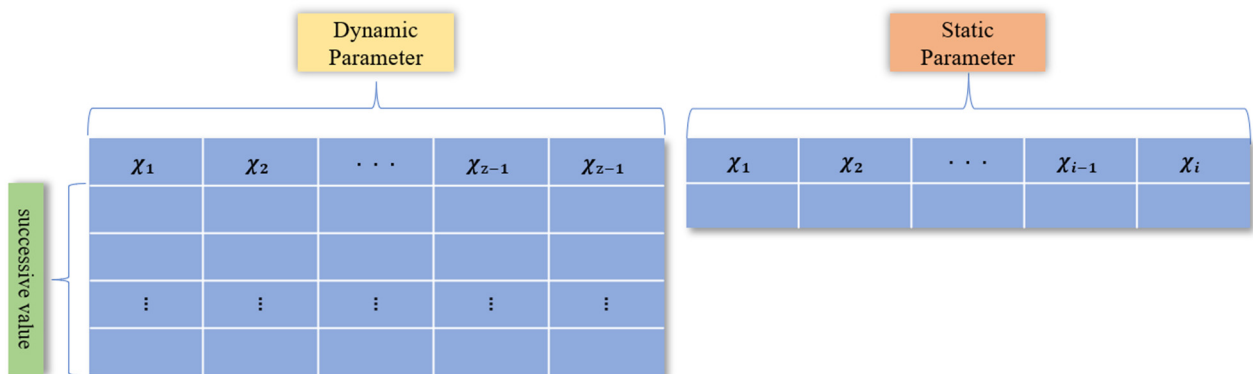
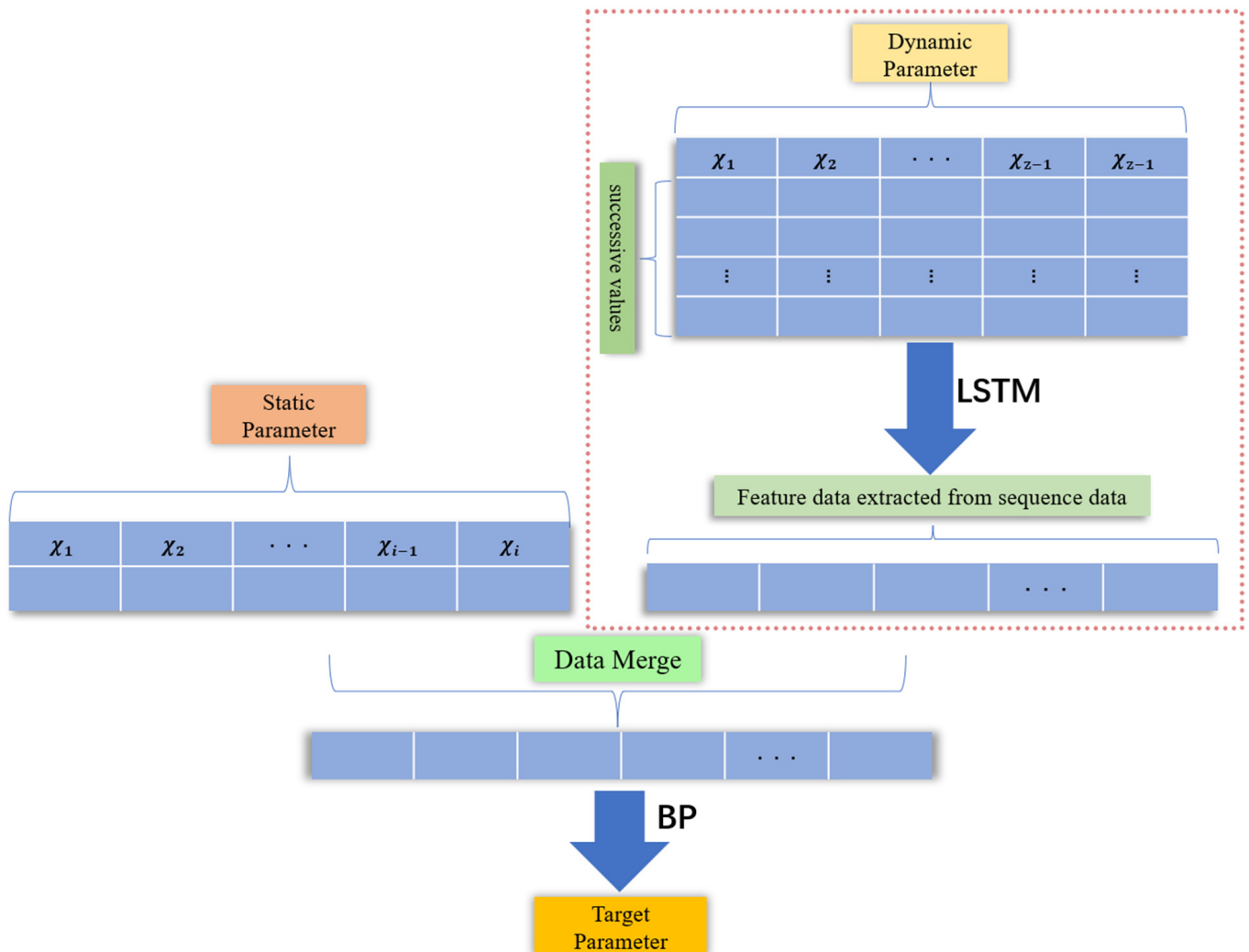


Figure 5. Sample structure of LSTM-BP neural networks.

In the LSTM-BP neural network, the LSTM network structure can process sequence data well and capture long-term dependencies in sequence data (Figure 6). In the overall structure, LSTM is used to capture the feature relationship in the dynamic parameters and output the feature data. The static parameters and the characteristic data of the LSTM output are merged as the input of the BP neural network, and, finally, the target parameters are output.



**Figure 6.** Sample structure of BP neural networks.

Create the LSTM-BP neural network structure and use random initialization for the initial model weights. Input the training set samples into the model to output predicted values, use the root mean square error as the loss function to calculate the loss value, and optimize the weights using gradient descent, reloading them into the model. After each round of weight updates, calculate and record the test set loss (Figure 7). Once all training iterations are complete, select the best weights. When using the model for prediction, training set involvement and multiple rounds of training are not required. The LSTM-BP model not only extracts data features from dynamic changes but also combines the efficient nonlinear data fitting capabilities of BP neural networks. This model provides better prediction results for productivity forecasting problems.

Verification results for a small sample (Table 2): For oil wells with actual production indices in the range of  $0.01\text{--}0.1 \text{ m}^3 \cdot \text{d}^{-1} \cdot \text{MPa}^{-1}$ , the relative prediction error of the LSTM-BP model is similar to that of the BP model. However, when the production index exceeds  $0.1 \text{ m}^3 \cdot \text{d}^{-1} \cdot \text{MPa}^{-1}$ , the LSTM-BP model outperforms the BP model in terms of

prediction accuracy. This is because the sample data set contains more samples in the  $0.01\text{--}0.1 \text{ m}^3\cdot\text{d}^{-1}\cdot\text{MPa}^{-1}$  range. During training, both models achieve good results. In the range of  $>0.1 \text{ m}^3\cdot\text{d}^{-1}\cdot\text{MPa}^{-1}$ , however, the limited number of samples cannot meet the training needs of the BP model, whereas the LSTM-BP model can provide accurate predictions even with fewer samples by extracting more precise features. For production indices in the  $0.01\text{--}0.1 \text{ m}^3\cdot\text{d}^{-1}\cdot\text{MPa}^{-1}$  range, the magnitude of the production index is too small; any change in the main control parameters greatly affects the final prediction value, making both models somewhat inadequate for wells with such small production index magnitudes. Nevertheless, the LSTM-BP model demonstrates superior generalizability.

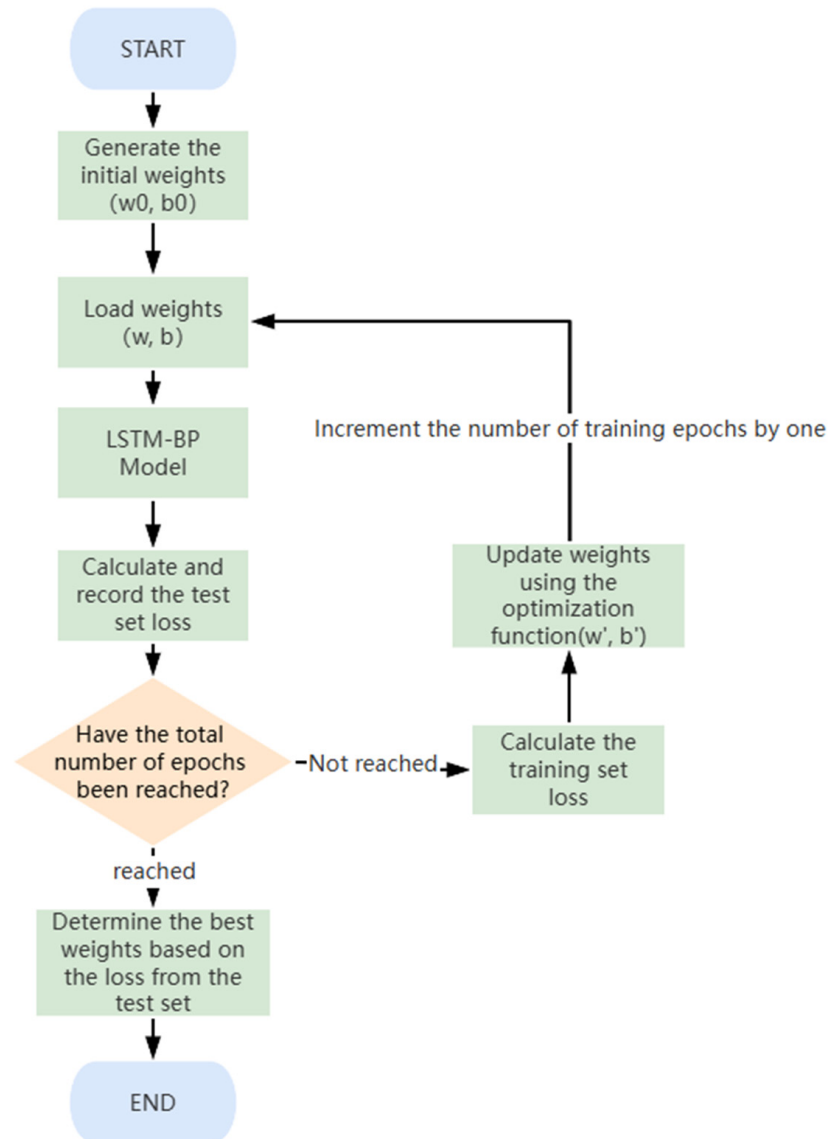


Figure 7. LSTM-BP model prediction flowchart.

Table 2. Comparison of validation results.

Well Name	Actual Oil Production Index ( $\text{m}^3\cdot\text{d}^{-1}\cdot\text{MPa}^{-1}$ )	LSTM-BP Predicted Value ( $\text{m}^3\cdot\text{d}^{-1}\cdot\text{MPa}^{-1}$ )	LSTM-BP Relative Error (%)	BP Predicted Value ( $\text{m}^3\cdot\text{d}^{-1}\cdot\text{MPa}^{-1}$ )	BP Relative Error (%)
WZ-1	0.01	0.0078	22.00	0.0073	27.00
WZ-2	0.03	0.0235	21.67	0.0213	29.00
WZ-3	0.67	0.4925	26.49	0.4574	31.73
WZ-4	0.59	0.4394	25.51	0.3978	32.57

The total number of samples in the data set is 43, with 35 samples used for training the model and 8 samples used for testing. After constructing the BP neural network and the LSTM-BP neural network models, the LSTM-BP neural network achieved an average absolute error of 0.0765, a root mean square error of 0.10, an average absolute percentage error of 21.18%, and a coefficient of determination ( $R^2$ ) of 0.97 on the test set (Table 3). The average absolute error of the BP model differs by 0.05 from that of the LSTM-BP model, but the average absolute percentage error differs by as much as 10%. This indicates that while the BP model can accurately predict some test samples, it fails to predict others accurately, and the poor prediction performance affects the overall prediction accuracy. Additionally, the average absolute percentage error of the BP model exceeds 30%, which is not suitable for practical applications. This demonstrates that the LSTM-BP model, by extracting dynamic parameter data features, provides a more effective capability for production prediction.

**Table 3.** Indicators for model evaluation.

Algorithm	Sample Set	$\eta_{MAE}$	$\eta_{RMSE}$	$\eta_{MAPE}$	$R^2$
BP neural network	Testing set	0.12	0.22	31.85%	0.88
LSTM-BP neural network	Testing set	0.07	0.10	21.18%	0.97

## 5. Applicable Analysis

Conventional production prediction methods commonly use analytical approaches, employing Darcy's law combined with the thickness of the feature segment to obtain the production rate for the entire test segment.

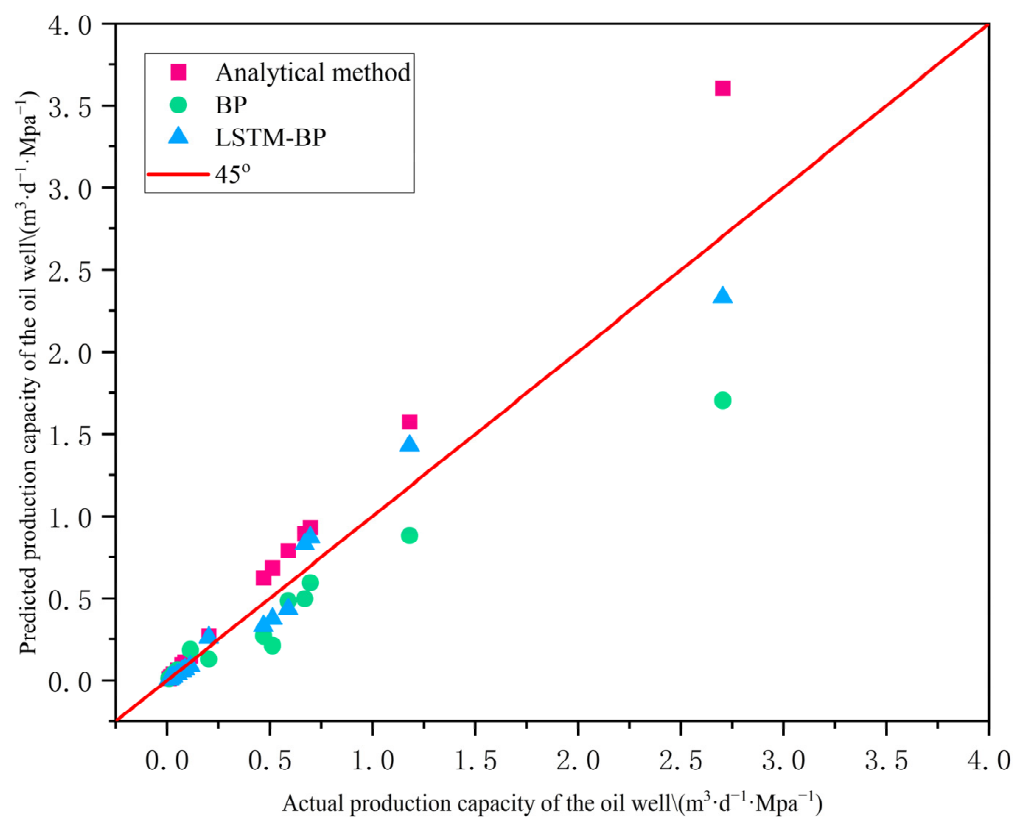
$$Q = \frac{0.543 \sum_i K_i h_i (P_e - P_{wf})}{u_o B_o \left( \ln \frac{0.472 r_e}{r_w} + S \right)} \quad (7)$$

In the formula,  $p_e$  is the formation pressure (MPa);  $p_{wf}$  is the formation flowing pressure (MPa);  $u_o$  is the viscosity of the oil (mPa·s);  $B_o$  is the oil formation volume factor;  $K_i$  is the oil-phase permeability of the layer segment represented by test point  $i$  (mD);  $h_i$  is the thickness of the layer segment represented by test point  $i$  (m);  $r_e$  and  $r_w$  are the detection radius of the test segment and the wellbore radius, respectively (m);  $S$  is the skin factor of the test segment.

Select 10 samples from the training set and 8 samples from the test set to form the prediction samples. Apply the LSTM-BP neural network to these samples for production prediction: Firstly, the original data are normalized through feature engineering. The characteristic parameters of the sample are divided into dynamic parameters and static parameters, and the dynamic parameters will be partially processed by the LSTM neural network. After that, the static parameters are merged with the data processed by the LSTM part; the combined data are sent to the BP neural network for capacity prediction; and, finally, the prediction results are shown. Use three methods—an analytical approach, BP model, and LSTM-BP model—to predict the production rates of these 18 samples and perform a comparative analysis (Figure 8).

After the model is established, the generalization ability of the model is an important indicator. Generalization means that the trained model should not only fit the known samples but also achieve more effective predictions in the unknown samples. The comparative analysis of the scatter plots of the true value and the predicted value can effectively reflect the generalization ability of the model. Through the scatter plot of the true value and the predicted value, the analytical method is constrained by the limitations of the idealized model and cannot fully capture the data characteristics of the entire sample. The analytical method is affected in the production range below  $0.25 \text{ m}^3 \cdot \text{d}^{-1} \cdot \text{MPa}^{-1}$ , where the model learns the data characteristics of this range but fails to learn from the entire sample data set. Consequently, predictions in subsequent ranges are also based on the idealized model, lead-

ing to the model's predicted values not being evenly distributed around the 45° reference line. The predicted value of the BP neural network cannot be evenly distributed near the reference line. This shows that although the BP neural network has a better determination coefficient and lower relative error in the test set, it cannot predict effectively in the face of unknown samples, and its generalization ability is poor. The LSTM-BP neural network has higher prediction accuracy in the test set, and its scatter points are evenly distributed near the reference line and can be divided into upper and lower parts by the reference line. This shows that the LSTM-BP neural network can achieve more accurate prediction in the face of known and unknown samples, and its generalization ability is strong. Due to the small number of oil well samples with a production index greater than 1 m<sup>3</sup>/d/MPa in the original data, all three models are affected in their training, leading to poor prediction performance for oil wells in this range. However, the LSTM-BP model, due to its ability to extract dynamic parameter features, mitigates the impact of uneven sample sizes to some extent.



**Figure 8.** Scattered points distribution of oil well productivity.

The LSTM-BP model achieves better prediction results in the WZ region. The LSTM-BP model is designed to extract dynamic changing parameters, making it particularly effective for production prediction where dynamic parameters are prevalent. It performs well not only for specific reservoirs but also for conventional reservoirs and deep offshore reservoirs. However, specific regional training samples are required for re-training in particular regions.

Due to the limited number of data samples in the WZ region, a more complex weight optimization function could not be used during model training. If the environment changes and more data samples are available, the model's performance can be further enhanced, and more effective optimization functions can be employed for weight updates. Future research will focus on selecting optimization functions and expanding the data set to optimize the model's network structure.

## 6. Conclusions

(1) Based on the main controlling factors of oil well productivity in the WZ area, samples composed of pure static parameters and dynamic and static parameters were used, and the corresponding neural network model was established to realize the oil well productivity prediction in the WZ area.

(2) The Pearson correlation coefficient and partial correlation coefficient were used to determine that the main controlling factors for production capacity in the WZ region are formation volume factor, water saturation, density, effective thickness, skin factor, mud content, porosity, and effective permeability.

(3) By comparing the model prediction effects under the two samples, the LSTM-BP neural network under dynamic and static parameters has a better prediction effect on the test set, and its coefficient of determination is 0.97. At the same time, the LSTM-BP neural network model also shows a good prediction effect on the reserved 18 oil wells, and its generalization ability is strong. Finally, the LSTM-BP neural network was selected to predict the oil well productivity of the tight reservoir in the WZ area.

(4) The LSTM-BP model outperforms both the analytical method and the BP model in terms of how its predicted values align with the 45° reference line. The LSTM-BP model's predictions are not only evenly distributed on both sides of the reference line but also exhibit smaller errors relative to the reference line. This results in a significant improvement in the production prediction accuracy.

**Author Contributions:** K.G. was the original data provider and provided experimental ideas. Y.J., X.G. and Q.L. performed data collection and data processing. Y.J. realized the model construction and the writing of the first draft of this article. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author.

**Acknowledgments:** Thank you to all participants for the technical or data support provided.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Pospichal, J.; Kvasnika, V. 70th Anniversary of Publication: Warren Mcculloch & Walter Pitts—A Logical Calculus of the Ideas Immanent in Nervous Activity. *Adv. Intell. Syst. Comput.* **2015**, *316*, 1–10.
2. Ali, J.K. Neural Networks: A New Tool for the Petroleum Industry? In Proceedings of the SPE European Petroleum Computer Conference, Aberdeen, UK, 15–17 March 1994.
3. Wang, H.; Wang, M.; Chen, S.; Hui, G.; Pang, Y. A Novel Governing Equation for Shale Gas Production Prediction Via Physics-Informed Neural Networks. *Expert Syst. Appl.* **2024**, *248*, 123387. [[CrossRef](#)]
4. Al-Kaabi, A.; Lee, W.J. Using Artificial Neural Networks to Identity the Well Test Interpretation Model. In Proceedings of the Proceedings—Petroleum Computer Conference, Denver, CO, USA, 25–28 June 1990.
5. Accarain, P.; Desbrandes, R. Neuro-Computing Helps Pore Pressure Determination. *Pet. Eng. Int.* **1993**, *65*, 39–42.
6. Jani, D.B.; Mishra, M.; Sahoo, P.K. Application of artificial neural network for predicting performance of solid desiccant cooling systems—A review. *Renew. Sustain. Energy Rev.* **2017**, *80*, 352–366. [[CrossRef](#)]
7. Nasirzadeh, F.; Kabir, H.D.; Akbari, M.; Khosravi, A.; Nahavandi, S.; Carmichael, D.G. ANN-based prediction intervals to forecast labour productivity. *Eng. Constr. Archit. Manag.* **2020**, *27*, 2335–2351. [[CrossRef](#)]
8. Ahmadi, S.; Khormali, A. Optimization of the Corrosion Inhibition Performance of 2-Mercaptobenzothiazole for Carbon Steel in Hcl Media Using Response Surface Methodology. *Fuel* **2024**, *357*, 129783. [[CrossRef](#)]
9. Yang, Y.; Zhao, H.; Yuan, S. Research on Oil and Gas Well Productivity Prediction Model Based on Gwo-Svm Algorithm. *Energy Environ. Prot.* **2024**, *46*, 178–183.
10. Dai, Q.; Zhang, L.; Zhang, K.; Chen, G.; Ma, X.; Wu, D.; Cao, C.; Yao, J. Horizontal Well Location Optimization Method Based on Machine Learning Agent Model. In Proceedings of the Twelfth National Conference on Fluid Mechanics, Xi'an, China, 22–25 September 2022.
11. Song, H.; Du, S.; Yang, J.; Wang, M.; Zhao, Y.; Zhang, J.; Zhu, J. Forecasting and Influencing Factor Analysis of Coalbed Methane Productivity Utilizing Intelligent Algorithms. *Chin. J. Eng.* **2024**, *46*, 614–626.

12. Li, X.; Chan, C.W.; Nguyen, H.H. Application of the Neural Decision Tree Approach for Prediction of Petroleum Production. *J. Pet. Sci. Eng.* **2013**, *104*, 11–16. [[CrossRef](#)]
13. Cao, Q.; Banerjee, R.; Gupta, S.; Li, J.; Zhou, W.; Jeyachandra, B. Data Driven Production Forecasting Using Machine Learning. In Proceedings of the SPE Argentina Exploration and Production of Unconventional Resources Symposium, Buenos Aires, Argentina, 1–3 June 2016.
14. Wu, L. Optimization Design of Fracturing Parameters for Tight Oil Reservoirs Based on Numerical Simulation and Machine Learning. Master's Thesis, Xi'an Shiyou University, Xi'an, China, 2023.
15. Qin, J. Shale Oil and Gas Productivity Prediction and Fracturing Optimization Based on Deep Learning. Master's Thesis, China University of Petroleum, Beijing, China, 2022.
16. Yue, M.; Dai, Q.; Liao, H.; Liu, Y.; Fan, L.; Song, T. Prediction of Orf for Optimized CO<sub>2</sub> Flooding in Fractured Tight Oil Reservoirs Via Machine Learning. *Energies* **2024**, *17*, 1303. [[CrossRef](#)]
17. Alimkhanov, R.; Samoylova, I. Application of Data Mining Tools for Analysis and Prediction of Hydraulic Fracturing Efficiency for the Bv8 Reservoir of the Povkh Oil Field. In Proceedings of the SPE Russian Oil and Gas Exploration & Production Technical Conference and Exhibition, Moscow, Russia, 14–16 October 2014.
18. Zheng, Y.; Liu, B.; Zhang, X.; Xue, Y.; Song, F.; Jiang, T. Productivity Prediction Method of Ultra-Low Permeability Reservoir Based on Data-Driven and Geological Law Fusion-Taking Yuan 284 Ultra-Low Permeability Reservoir as an Example. *Pet. Geol. Eng.* **2022**, *36*, 75–81.
19. Dong, Y.; Song, L.; Zhang, Y.; Qiu, L.; Yu, Y.; Lu, C. Initial Productivity Prediction Method of Offshore Oil Wells Based on Physical Constraint Data Mining Algorithm. *Pet. Geol. Recovery Factor* **2022**, *29*, 137–144.
20. Purbey, R.; Parijat, H.; Agarwal, D.; Mitra, D.; Agarwal, R.; Pandey, R.K.; Dahiya, A.K. Machine Learning and Data Mining Assisted Petroleum Reservoir Engineering: A Comprehensive Review. *Int. J. Oil Gas Coal Technol.* **2022**, *30*, 359–387. [[CrossRef](#)]
21. Hui, G.; Chen, S.; He, Y.; Wang, H.; Gu, F. Machine Learning-Based Production Forecast for Shale Gas in Unconventional Reservoirs Via Integration of Geological and Operational Factors. *J. Nat. Gas Sci. Eng.* **2021**, *94*, 104045. [[CrossRef](#)]
22. Liu, J.; Tian, L.; Liu, S.; Li, N.; Zhang, J.; Ping, X.; Ma, X.; Zhou, J.; Zhang, N. Productivity Prediction Model of Tight Gas Wells Based on Composite Machine Algorithm-Taking Sm Block in Ordos Basin as an Example. *Daqing Pet. Geol. Dev.* **2024**, *43*, 69–78. [[CrossRef](#)]
23. Fargalla, M.A.M.; Yan, W.; Deng, J.; Wu, T.; Kiyangi, W.; Li, G.; Zhang, W. Timenet: Time2vec Attention-Based Cnn-Bigru Neural Network for Predicting Production in Shale and Sandstone Gas Reservoirs. *Energy* **2024**, *290*, 130184. [[CrossRef](#)]
24. Li, Y. Research on Reservoir Pore Structure Evaluation and Reservoir Classification Prediction Method Based on Deep Learning. Master's Thesis, China University of Petroleum, Beijing, China, 2020.
25. Fu, H.; Fang, Q.; Du, Y. Tight Gas Reservoir Productivity Prediction Based on Arima-Rts and Lstm. *Petrochem. Technol.* **2023**, *30*, 120–122.

**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.