electronics

MDPI

Article

Using a Deep Neural Network with Small Datasets to Predict the Initial Production of Tight Oil Horizontal Wells

Yuxi Yang [1,2], Chengqian Tan [1,2,*], Youyou Cheng [1,3], Xiang Luo [4] and Xiangliang Qiu [4]

[1] School of Earth Sciences and Engineering, Xi'an Shiyou University, Xi'an 710065, China; 21211020236@stumail.xsyu.edu.cn (Y.Y.)
[2] Shaanxi Key Laboratory of Petroleum Accumulation Geology, Xi'an Shiyou University, Xi'an 710065, China
[3] Laboratory of Basin Structure and Hydrocarbon Accumulation, CNPC, Beijing 100083, China
[4] School of Petroleum Engineering, Xi'an Shiyou University, Xi'an 710065, China
* Correspondence: cqtan@xsyu.edu.cn

Abstract: Due to its abundant reserves, tight oil has emerged as a significant substitute for conventional petroleum resources. It has become one of the focal points of exploration and research, and a new hot spot in global unconventional oil and gas exploration and development. This has led to a significant increase in the demand for forecasting the production capacity of tight oil horizontal wells. The deep neural network (DNN), as a mature model, has demonstrated significant advantages in many fields. However, due to the confidentiality and uniqueness of oilfield data, acquiring large datasets has become a challenge. Traditional methods using small datasets for training DNN models result in low accuracy and overfitting issues, which hinders the development of neural networks in the petroleum industry. This study aims to predict the initial production capacity of tight oil horizontal wells by using a small dataset of 650 data points through a DNN model. The research results indicate that pre-trained and fine-tuned DNNs outperform shallow neural networks, supporting vector machines, and DNN trained with traditional methods in terms of better generalization performance. Their accuracy reached 91.3%, demonstrating that it is reasonable to use a small dataset with pre-trained and fine-tuned DNN models.

Keywords: small datasets; DNN; tight oil prediction; initial production capacity; horizontal wells

check for updates

Citation: Yang, Y.; Tan, C.; Cheng, Y.; Luo, X.; Qiu, X. Using a Deep Neural Network with Small Datasets to Predict the Initial Production of Tight Oil Horizontal Wells. *Electronics* **2023**, *12*, 4570. https://doi.org/10.3390/electronics12224570

Academic Editors: Junhua Ding, Haihua Chen, Yunhe Feng and Tozammel Hossain

Received: 2 October 2023
Revised: 6 November 2023
Accepted: 7 November 2023
Published: 8 November 2023

1. Introduction

1.1. Traditional Methods

In recent decades, tight oil has become a major focus of research in the global petroleum geology field. Due to the fact that the accumulation conditions and mechanisms of tight oil are significantly different from conventional reservoirs, and that various factors, such as geology, development, and engineering, contribute to challenges in tight oil horizontal well development, including rapid production decline, low individual well productivity, and low recovery rates [1,2]. Therefore, forecasting the production capacity of tight oil horizontal wells has become a crucial basis for planning and deploying unconventional oil and gas field development. As a result, researchers worldwide have conducted extensive studies on the prediction of production capacity for tight oil horizontal wells, considering a combination of geological, development, and engineering factors. Quantitative characterization methods have been widely applied in analyzing production capacity prediction, and have achieved good results. The researchers conducted sensitivity analysis, uncertainty analysis and history matching, and predicted gas production to assess the impact of injecting carbon dioxide (CO_2) on well productivity [3]. By incorporating the Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm and introducing four new feature parameters, a quantitative characterization model for shale oil vertical and horizontal wells was developed, which enables the fitting, prediction, and parameter inversion of dynamic production curves throughout the entire life cycle [4]. However, due to the simplicity and intuitiveness of the quantitative characterization model, it only analyzes the primary reservoir characteristics that affect production capacity, while neglecting the secondary factors. Therefore, some researchers have opted to use numerical simulation methods to establish complex models as an alternative to quantitative characterization methods. By establishing an equivalent aperture numerical model, the spatial variation in fracture aperture caused by fracture deformation can be evaluated to assess its impact on well productivity. This approach enables the prediction of production capacity by considering the influence of spatial changes in fracture aperture [5]. The grey correlation analysis method is employed to analyze the geological and engineering factors. An evaluation model is established to predict the production for the first three years [6]. Numerical simulation methods can be used to assess the impact of various reservoir characteristics, such as reservoir geological shape, fluid properties, and reservoir heterogeneity, on production capacity prediction [7]. This approach addresses the limitations of simple and intuitive quantitative characterization models, which often overlook comprehensive consideration of reservoir characteristics. Indeed, the establishment of numerical simulation requires a significant amount of reservoir data, which can make the modeling process complex. Additionally, the computation time can be relatively long due to the detailed calculations involved. To address this issue, semi-analytical models have been used as an alternative. While these models overcome the complexity of numerical simulation in terms of modeling and solving difficulties, they may still have limitations in their application [8–10].

1.2. Machine Learning Methods

With the advancement of machine learning theory, various artificial intelligence algorithms have been applied in the field of petroleum engineering. By utilizing a Deep Neural Network (DNN) model, the cumulative oil production of the Bakken shale oil reservoir can be predicted [11]. Similarly, the Long Short-Term Memory (LSTM) structure can be employed to forecast the production of a carbonate reservoir in the Middle East [12]. For more sophisticated predictions, a Deep Long Short-Term Memory (DLSTM) framework has been established to enhance the accuracy of oil production forecasts [13].

Conducting researches on predicting oil well production capacity through machine learning has become a trend in the field, but rather limited contributions to the study of initial production capacity prediction for oil wells were made [14]. Estimating initial production capacity accurately in oilfield development provides vital insights for reservoir reserve assessment and resource availability. It serves as a crucial reference point for planning and executing effective oilfield development strategies [15]. Additionally, the test results of initial production capacity can aid in determining appropriate production strategies, including production rates and duration. Well-designed production strategies can maximize output while mitigating issues like reservoir pressure decline and decreased production capacity that may result from overproduction. It plays a critical role in the oilfield development process [16].

The researches of the previous scholars have shown that, in comparison to traditional numerical simulation and semi-analytical methods, machine learning has demonstrated its advantages of simplicity and higher accuracy in many fields. This trend is not only evident in theoretical studies but has also been validated in practical applications. This article aims to explore how to leverage machine learning for predicting initial production capacity in a simpler and more accurate manner.

1.3. Potential for Applying DNN with Small Dataset

While machine learning has been introduced to the field of petroleum geology in recent years, the application of DNN in petroleum geology remains limited. This is because the uniqueness of geological conditions in each oil field region prevents the use of all domestic oil field data for training. Therefore, we can only use well data from specific regions for training. However, the number of wells in a particular region is not sufficient to construct a large dataset, which means that using machine learning methods with large datasets to predict initial production capacity for tight oil horizontal wells will pose a significant challenge to the researchers. At present time, advances in machine learning have made DNN more easily trainable and have provided some useful tools, such as pre-training with restricted Boltzmann machines (RBM) and sparse autoencoders (SAE), to handle small datasets [17,18].

In the field of petroleum geology research, the potential of using DNN with small datasets is evident. Regression and classification problems that were previously addressed using traditional machine learning methods such as Shallow Neural Network (SNN), random forests, support vector machines (SVM), and others can now be tackled with DNN, leading to higher accuracy and improved generalization performance.

In this study, we employed a DNN pre-trained with SAE. There are multiple reasons for choosing this approach. Firstly, SAE, as a deep learning technique, is capable of automatically learning essential features from the data while reducing data dimensionality, which enhances the performance of predictive models. Secondly, SAE pre-training can be used to initialize the neural network's weights, optimizing them through layer-wise training for better data fitting. Additionally, with the capability to effectively handle data sparsity and noise, SAE can improve the model's robustness. This research focused on analyzing data related to tight oil horizontal wells to identify which factors influence initial production capacity. We used a DNN model pre-trained with SAE for initial production capacity prediction [17,18]. Through experimental results, we demonstrated the significant advantages of this method over traditional approaches: it showed higher prediction accuracy and better generalization performance. This not only contributes to improving the accuracy of initial production capacity prediction, but also opens up new directions for similar research in exploring the potential of machine learning in resource extraction and production forecasting.

2. Method and Workflow

The application of deep learning in the field of oil will contribute to reducing development risks and improving resource utilization efficiency in the petroleum industry. In this chapter, we introduce a model algorithm for predicting the initial production capacity of tight oil horizontal wells.

2.1. DNN Pre-Training

Pre-training involves initializing the weights and biases of the DNN with values close to the global optimum solution, which assists in bypassing the traps of local optima during subsequent fine-tuning steps [19]. Figure 1 illustrates the pre-training process, which involves initializing the DNN using SAE. The upper section displays the structure of the DNN, while the lower section depicts the structures of six autoencoders. brittleness, sand ratio, sanding intensity, fracturing fluid intensity, flowback time, flowback rate, fracture density, horizontal section length, well spacing, permeability, energy-storage coefficient, pressure difference, and total volume are used as 13 input variables for the model. Initial production capacity is the output variable of the DNN. $W^j(0)$ represents the initial weight matrix layer of the j-th layer in the DNN. Arrows indicate the direction of data transmission. An autoencoder is a special type of shallow neural network with a single hidden layer that shares the same input and output layers. Taking a DNN with 5 hidden layers as an example, structured as 13-(7-6-5-4-3)-1, the structures of the 6 autoencoders would be as follows: 13-(7)-13, 7-(6)-7, 6-(5)-6, 5-(4)-5, 4-(3)-4, 3-(1)-3. Each autoencoder has the same number of neurons as the corresponding layer in the DNN. The hidden layer of the previous autoencoder serves as the input and output layer for the next autoencoder. Each trained autoencoder provides initial weights and biases for the corresponding layer of the DNN after SAE initialization. Following SAE initialization, the DNN is trained by using the Adam algorithm, as utilized in the SNN.

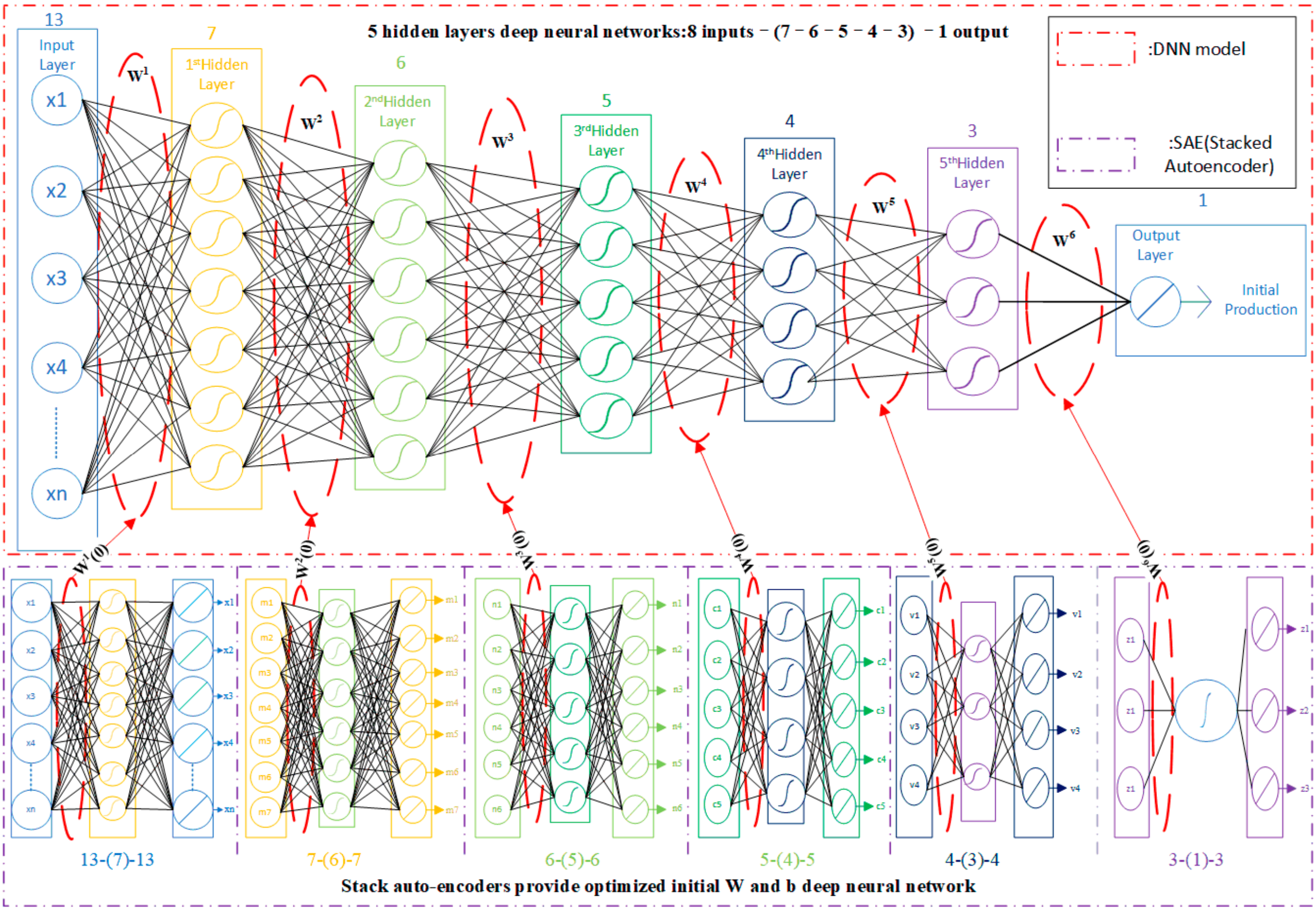


Figure 1. DNN Pre-training model.

2.2. *Improving the L2 Regularization Expression*

The performance function for a neural network, Mean Squared Error (MSE), is formulated as follows:

$$MSE = \frac{\sum_{i=1}^{n} (y_n - \hat{y}_n)^2}{n} \tag{1}$$

The L2 regularization expression [20] is:

$$\mathrm{J}(w, b) = \mathrm{MSE} + \frac{\lambda}{2n}||w||^2 + \frac{\lambda}{2n}b^2 \tag{2}$$

We improved the L2 regularization expression, and the formula is as follows:

$$\mathrm{J}(\hat{y}, w) = \lambda\mathrm{MSE} + \frac{(1-\lambda)}{2n}||w||^2 \tag{3}$$

where n represents the number of training samples, $||w||^2$ represents the squared norm, $||w||^2 = \sum_{i=1}^{n^{[l-1]}} \sum_{j=1}^{n^{[l]}} (W_{ij}^{[l]})^2$ represents the sum of squares of all weights in the matrix, b represents the bias, λ is a hyperparameter, y_n is the actual value, and $\hat{y}_n$ is the model's output value, i.e., the predicted value.

The purpose of L2 regularization is to increase the model adaptability and reduce the likelihood of overfitting. Due to the inherent limitations of small datasets, over-fitting may not be avoided while using a L2 regularization expression. Therefore, hyperparameters are applied to combine the Mean-Squared Error (MSE) and $||w||^2$ to strengthen the relationship between the MSE and $||w||^2$. Subsequently, random initialization is used to train the SNN, while optimized values are employed to initialize and fine-tune the DNN model. Training the SNN and fine-tuning the DNN involves adjusting weights and biases along the negative gradient of an improved L2 expression using the Adam algorithm to minimize the expression, simultaneously, to determine the optimal value of α, which is based on the Bayesian regularization method proposed by MacKay, in relation to reducing local minima [21]. The experimental section in the next part will demonstrate the effectiveness of this improved L2 regularization expression for small datasets.

2.3. Workflow

The workflow for the prediction of initial production capacity in tight oil horizontal wells is depicted in Figure 2. It includes the following steps: data collection, data analysis, data pre-processing, partitioning into training/testing/prediction datasets, training SVM/SNN/DNN models, testing the machine learning models, comparing the accuracy of DNN, SNN, and SVM models, selecting the model with the highest accuracy, and using the best model to predict the prediction dataset.

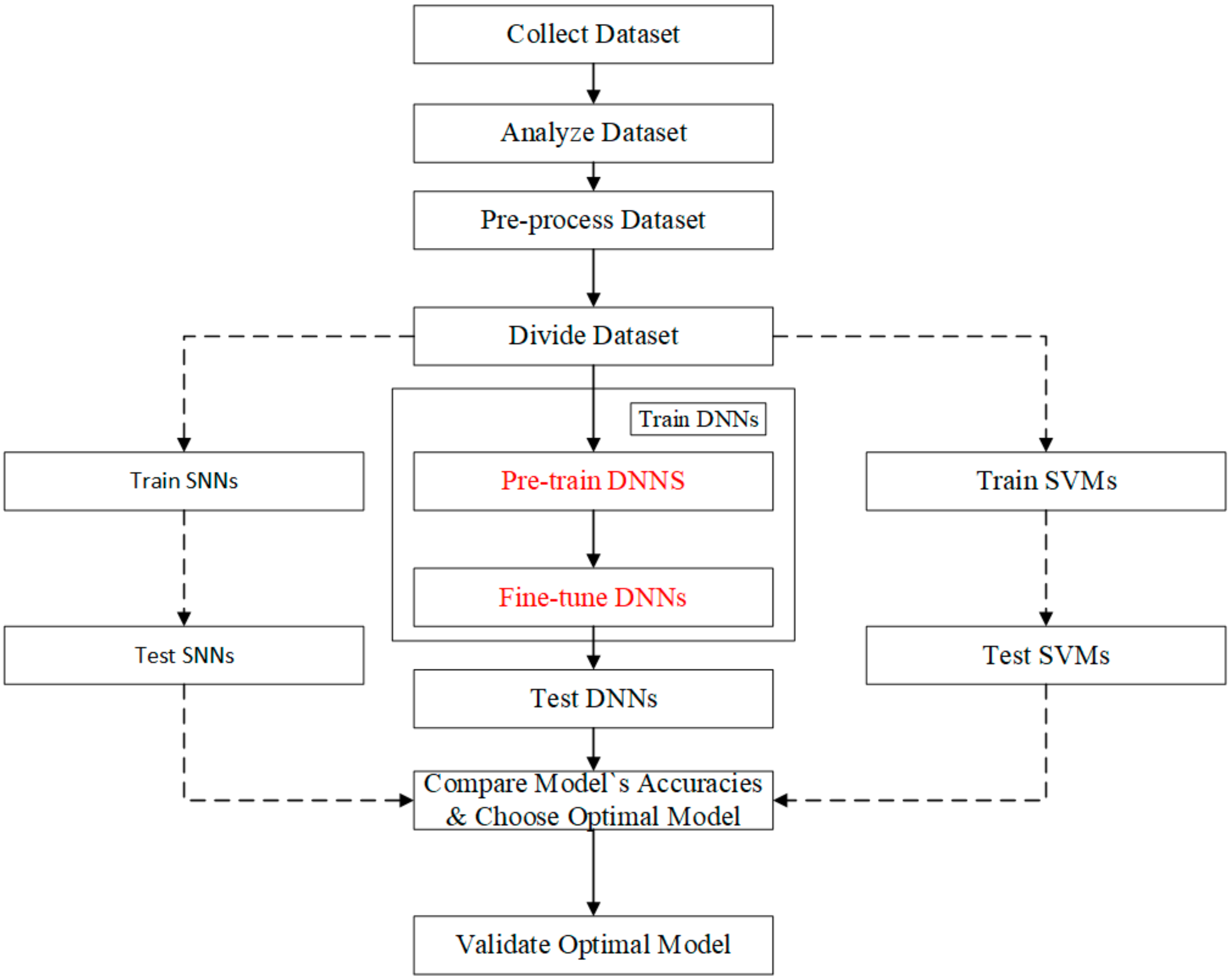


Figure 2. The workflow for predicting the initial production capacity of tight oil horizontal wells.

3. Experiment

We first analyzed the data related to the factors affecting the initial production capacity of tight oil horizontal wells, then designed an experimental model, and provided a description of the experimental equipment.

3.1. Collecting Data

This study focuses on the z183 Chang 7 oil reservoir, primarily a tight oil reservoir, which is located in the southern part of the Ordos Basin [22]. This reservoir is predominantly a tight oil reservoir, consisting of seven layers. The depositional environment is characterized by deep lake to semi-deep lake sedimentation. The average porosity is only 6.68%, mainly concentrated in the range of 2% to 10%. The permeability is concentrated between 0.05 mD and 0.2 mD [23,24]. The reservoir mainly comprises lithic feldspathic sandstone and feldspathic lithic sandstone. Overall, the z region exhibits a high quartz content and low feldspar content. Production in the z183 oil reservoir began in 2013, and up to the present time, 97 horizontal wells have been developed, resulting in a cumulative oil production of 57.8 tons. The primary development technique employed is hydraulic fracturing in horizontal wells, which has become the main approach for developing the tight oil reservoir in this region.

In the field of petroleum geology, the quantitative assessment of reservoirs is primarily analyzed through geological, developmental, and engineering factors. The evaluation and analysis of low-permeability tight oil reservoirs are conducted through fifteen geological factors, such as permeability, porosity, pore throat radius, movable fluid saturation, clay mineral content, and others [25]. The analysis and assessment of tight oil horizontal wells are made according to ten development and engineering factors, including reservoir properties, horizontal segment length, heterogeneity, fracture density, fracture connectivity, etc. [26]. Therefore, we collected data from 50 wells in the z183 oil reservoir for our experiment, and identified 13 factors that may influence the initial production capacity of tight oil horizontal wells. These 13 factors can be categorized into three major groups. Firstly, the geological factors include the energy-storage coefficient, permeability, and brittleness. Secondly, the development factors include well spacing, horizontal section length, fracture density, flowback rate, flowback time, and pressure difference. Thirdly, the engineering factors include fracturing fluid intensity, sanding intensity, sand ratio, and total capacity.

3.2. Analyzing Data

Figure 3 displays histograms of the minimum, maximum, mean, and standard deviation of the 13 variables in the dataset. Except for the flow back time, the other 12 variables have well-distributed data and are suitable for modeling.

3.3. Data Pre-Processing

As is well known, the Pearson algorithm uses correlation coefficients to assess the degree of association between variables and the target variable, as well as the strength and direction of the relationship between two variables, in order to select the most relevant features. In this study, to avoid negative impacts of certain factors on production capacity prediction, which may lead to reduced model accuracy, the correlation between each factor and initial production capacity is visualized by creating a heatmap. This allows for the selection of influencing factors that have a positive correlation with initial production capacity, aiding in the establishment of an optimal solution for the model (Figure 4).

From the chart, it is evident that the energy-storage coefficient, brittleness, well spacing, horizontal section length, fracture density, fracturing fluid intensity, sanding intensity, and total capacity have strong correlations with initial production capacity. Horizontal segment length has the closest relationship with initial production capacity, with a linear correlation of 0.4. Therefore, we selected factors with a linear relationship greater than 0 and factors with a linear relationship greater than 0.2 (taking the average of the highest linear correlation coefficients) as the dataset for model training. These two preprocessed datasets, along with the original dataset, were used as inputs for training and testing the SVM, SNN, and DNN models.

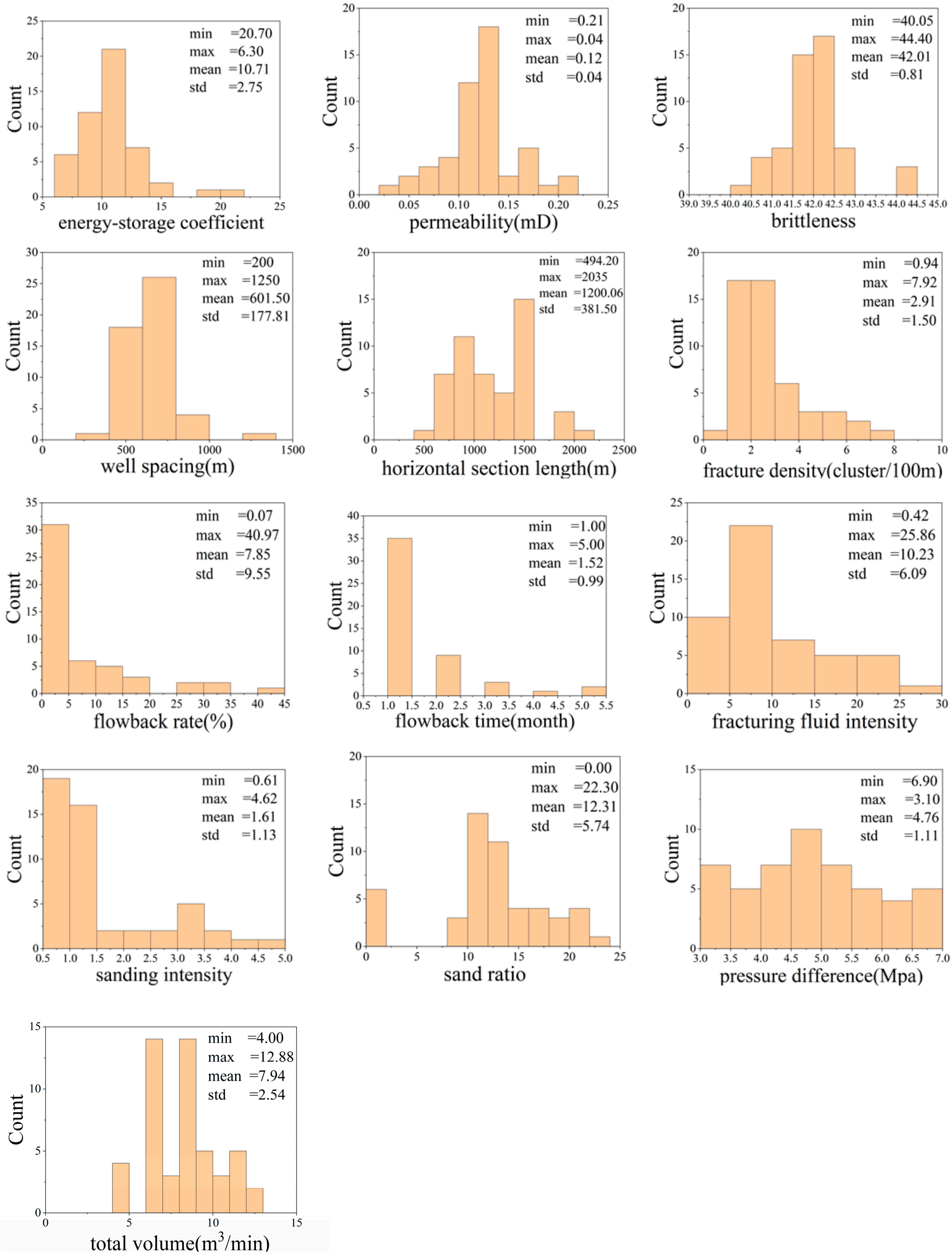
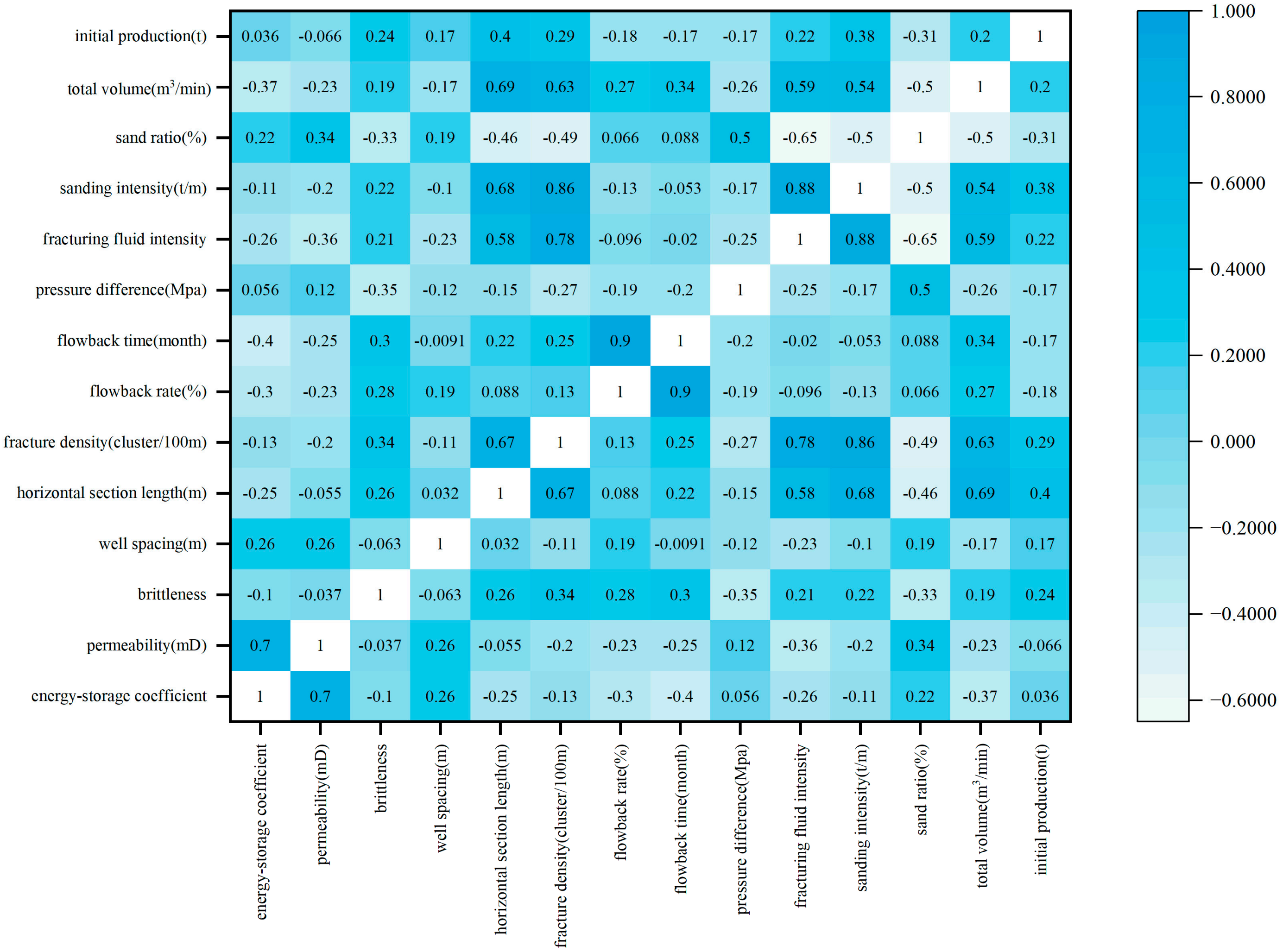


Figure 3. Histograms of the 13 variables in the dataset, along with statistical information such as mean, minimum, maximum, and standard deviation, are displayed on the histograms.



	energy-storage coefficient	permeability(mD)	brittleness	well spacing(m)	horizontal section length(m)	fracture density(cluster/100m)	flowback rate(%)	flowback time(month)	pressure difference(Mpa)	fracturing fluid intensity	sanding intensity(t/m)	sand ratio(%)	total volume(m^3/min)	initial production(t)
initial production(t)	0.036	-0.066	0.24	0.17	0.4	0.29	-0.18	-0.17	-0.17	0.22	0.38	-0.31	0.2	1
total volume(m^3/min)	-0.37	-0.23	0.19	-0.17	0.69	0.63	0.27	0.34	-0.26	0.59	0.54	-0.5	1	0.2
sand ratio(%)	0.22	0.34	-0.33	0.19	-0.46	-0.49	0.066	0.088	0.5	-0.65	-0.5	1	-0.5	-0.31
sanding intensity(t/m)	-0.11	-0.2	0.22	-0.1	0.68	0.86	-0.13	-0.053	-0.17	0.88	1	-0.5	0.54	0.38
fracturing fluid intensity	-0.26	-0.36	0.21	-0.23	0.58	0.78	-0.096	-0.02	-0.25	1	0.88	-0.65	0.59	0.22
pressure difference(Mpa)	0.056	0.12	-0.35	-0.12	-0.15	-0.27	-0.19	-0.2	1	-0.25	-0.17	0.5	-0.26	-0.17
flowback time(month)	-0.4	-0.25	0.3	-0.0091	0.22	0.25	0.9	1	-0.2	-0.02	-0.053	0.088	0.34	-0.17
flowback rate(%)	-0.3	-0.23	0.28	0.19	0.088	0.13	1	0.9	-0.19	-0.096	-0.13	0.066	0.27	-0.18
fracture density(cluster/100m)	-0.13	-0.2	0.34	-0.11	0.67	1	0.13	0.25	-0.27	0.78	0.86	-0.49	0.63	0.29
horizontal section length(m)	-0.25	-0.055	0.26	0.032	1	0.67	0.088	0.22	-0.15	0.58	0.68	-0.46	0.69	0.4
well spacing(m)	0.26	0.26	-0.063	1	0.032	-0.11	0.19	-0.0091	-0.12	-0.23	-0.1	0.19	-0.17	0.17
brittleness	-0.1	-0.037	1	-0.063	0.26	0.34	0.28	0.3	-0.35	0.21	0.22	-0.33	0.19	0.24
permeability(mD)	0.7	1	-0.037	0.26	-0.055	-0.2	-0.23	-0.25	0.12	-0.36	-0.2	0.34	-0.23	-0.066
energy-storage coefficient	1	0.7	-0.1	0.26	-0.25	-0.13	-0.3	-0.4	0.056	-0.26	-0.11	0.22	-0.37	0.036

Figure 4. A heatmap of the 13 influencing factors.

3.4. Dataset Dividing

To make effective use of the small dataset, it was divided into three groups. The first group used factors with linear relationships > 0 as input data for the model, the second group used factors with linear relationships $\geq$ 0.2 as input data for the model, and the third group used all 13 influencing factors as input data for the model. Then, data were randomly selected in a 3:1:1 ratio for training, validation, and testing the model's generalization ability. This means that three-fifths of the dataset (390 data points) was used for training, one-fifth (130 data points) for validating the model's accuracy, and the remaining one-fifth (130 data points) was used to further test the model's ability to generalize to unseen data. Holdout cross-validation was applied to reserve enough data for validating the model's performance, and an independent test set was provided for the final evaluation of the model. This test set helps ensure the generalization performance of the machine learning model, in other words, how well the model performs on new data. The reason for separating validation data and test data is to prevent the model from overfitting to the validation set, thereby providing a more accurate assessment.

3.5. Training Model

The SNN and DNN structures used in this study are similar to those shown in Figure 5. Both the SNN and DNN consist of 13 input neurons, a varying number of neurons (DNN model includes different layers of hidden neurons), and 1 output neuron for initial production capacity. The activation function for the hidden layers is tanh (x). The optimizer for the models is Adam. The SNN model has two structures: 13-(X1)-1 and 13-(X1)-(X2)-1. For the training of SNN, we specifically focused on the impact of the number of neurons in a single hidden layer on model accuracy. The number of neurons in a single hidden layer (X1) varies from 1 to 30, and for the dual hidden layers, the number of neurons in X1 ranges from 2 to 30, and the number of neurons in X2 ranges from 4 to 30. The structure of the DNN model is as follows: 13-(4-3-2)-1, 13-(9-6-3)-1, 13-(10-8-6-4)-1, 13-(5-4-3-3)-1, 13-(9-6-3-2-2)-1 and 13-(11-9-7-5-3-3)-1. For the training of DNN, our primary focus was on the impact of the number of hidden layers on model accuracy. In this experiment, we trained the model by using the six specified numbers of hidden layers (ranging from 3 layers to 6 layers), and determined the model with the highest accuracy based on the number of hidden layers. All SNN and DNN models were trained using random initialization. We also trained the DNN models using pretraining and fine-tuning, and compared their accuracy with DNN models that were not pretrained.

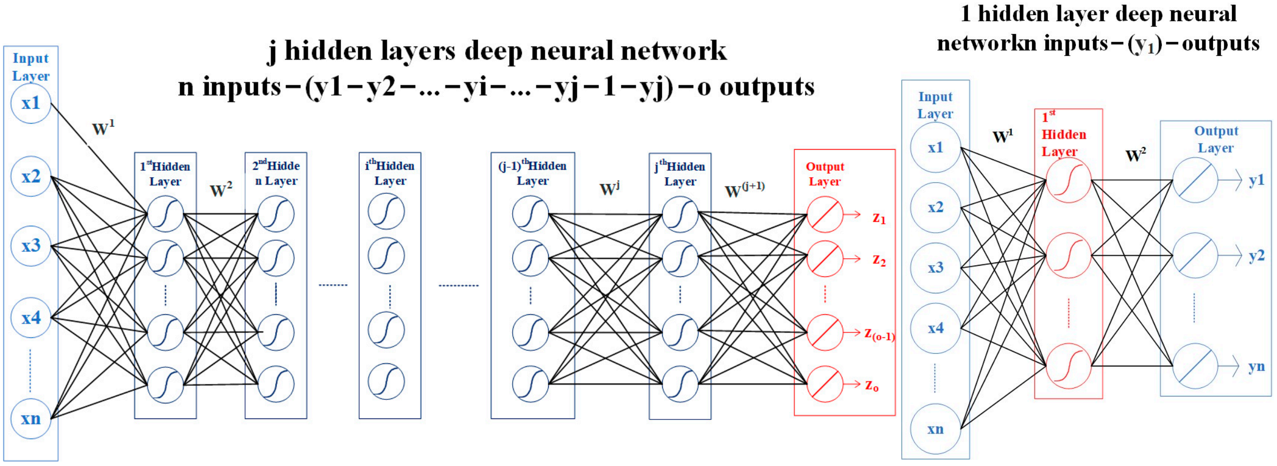


Figure 5. SNN with one hidden layer and DNN with j hidden layers.

SVM, SNN, and DNN are trained using the same learning rate. The maximum training epochs were set to 1000. The best model was selected for comparison. The DNN model with SAE pretraining was trained, tested, and predicted using the Machine Learning Toolbox in MATLAB 2023a. SVM, SNN, and non-pretrained DNN models were trained, tested, and predicted using the PyTorch framework with GPU support. All models were trained, tested, and predicted on a personal computer equipped with an i7-12700h CPU, 16 GB RAM, and an RTX 3080ti GPU.

4. Results and Discussion

4.1. Comparison of Model Accuracy

After training the DNN, SNN, and SVM models, the accuracies of these three models were compared. The results in Table 1 indicate that negatively correlated factors have an adverse impact on training accuracy, while the highest accuracy is achieved when factors with a linear correlation > 0 are used as input neurons. This suggests that factors with a linear correlation > 0 have a positive impact on the model's accuracy. The subsequent research will use the dataset with eight influencing factors (linear correlation > 0) for the model.

Figure 6a illustrates the impact of the number of neurons on the accuracy of SNN when there is only one hidden layer. As the number of neurons increases from 1 to 10, the training accuracy of the SNN rises from 0.67 to 0.955, and the validation accuracy improves from 0.45 to 0.854. However, beyond 10 neurons, accuracy begins to decline, with fluctuations observed between 11 and 30 neurons. The highest training and validation accuracies was achieved with 19 neurons, reaching 0.99 for training and 0.876 for validation (with only two decimal places in the graph). Nevertheless, the model exhibits overfitting, primarily due to the limited dataset, which is one of the key challenges in this experiment when dealing with small data. Consequently, for subsequent experiments, an SNN structure of 8-(19)-1 was selected for comparison.

Table 1. Model accuracy comparison between the original dataset and preprocessed datasets.

	Original Dataset	**Linear Correlation > 0**	**Linear Correlation ≥ 0.2**
Model	Training/Validation with input of 13 Influencing Factors	Training/Validation with input of eight Influencing Factors	Training/Validation with six Influencing Factors as Input
SVM	0.903/0.835	0.951/0.862	0.912/0.856
SNN	0.854/0.77	0.99/0.876	0.936/0.878
DNN	0.933/0.852	0.986/0.913	0.954/0.90

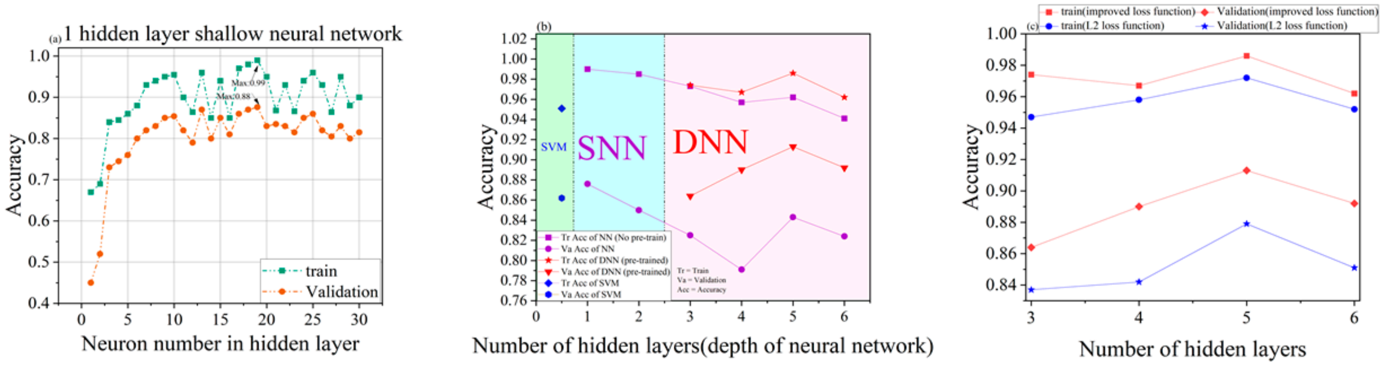


Figure 6. The impact of various factors on the accuracy of the neural network: (**a**) neurons; (**b**) hidden layers; and (**c**) loss function.

Figure 6b illustrates the training and validation accuracy of the SVM, SNN, and DNN models. When the number of hidden layers in DNN ranges from three to six, both training and validation accuracies of DNN are lower than those of SNN and SVM. However, the pre-trained and fine-tuned DNN model achieves higher training and validation accuracy compared with SNN, SVM, and non-pretrained DNN models. The 8-(7-5-4-3-3)-1 structure with five hidden layers, pre-trained and fine-tuned using SAE, exhibits the best accuracy. It achieved a training accuracy of 0.99 and a validation accuracy of 0.913, which was 0.051 higher than SVM's validation accuracy, 0.063 higher than SNN's validation accuracy, and 0.07 higher than the non-pretrained DNN's validation accuracy.

Figure 6c demonstrates the impact of using L2 regularization expression and the improved expression as the model loss function on model accuracy. It is evident that when using the L2 regularization expression as the model loss function, overfitting occurs due to the limitations of the small dataset, and the accuracy on the validation set is consistently below 90%. In contrast, when using the improved loss function, overfitting is mitigated, and the validation accuracy significantly surpasses that of the DNN models using the L2 regularization expression. This indicates that the improved L2 expression is suitable for models with small datasets.

4.2. Validation

To further validate the suitability of the pre-trained and fine-tuned DNN model for small datasets, predictions are made using the testing dataset across all models.

Figure 7 displays linear regression results for each model on the prediction dataset. The DNN model pre-trained and fine-tuned with SAE (correlation coefficient R = 0.95) outperformed the SNN model (R = 0.91) by 4% and the SVM model (R = 0.88) by 8%. This further confirms that the SAE pre-trained and fine-tuned DNN model has superior generalization compared with SNN and SVM.

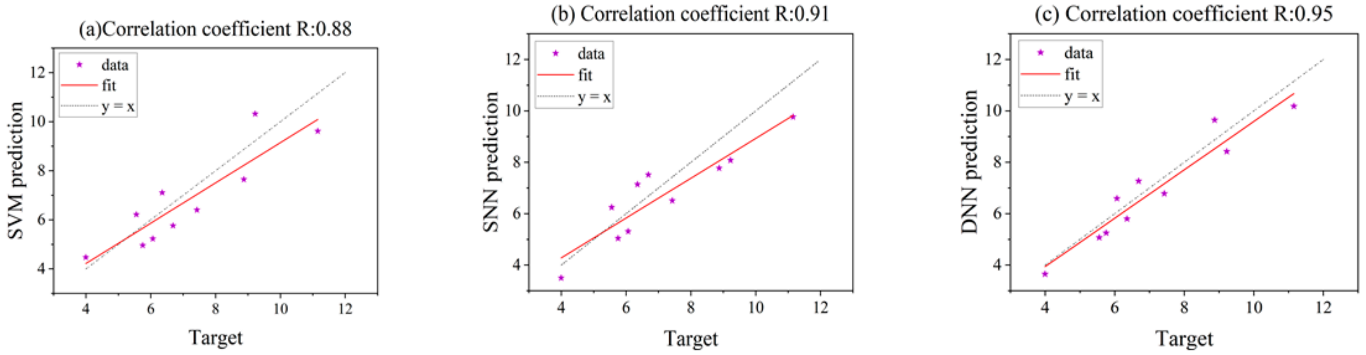


Figure 7. Different models were used to perform regression analysis on the test dataset: (**a**) SVM; (**b**) SNN; and (**c**) DNN.

4.3. Discussion

The factors influencing the production capacity of tight oil horizontal wells, ranked by their importance, are horizontal section length, sanding intensity, fracture density, brittleness, fracturing fluid intensity, total capacity, well spacing, and the energy-storage coefficient. This indicates that the factors influencing the production capacity of tight oil horizontal wells primarily lie in the development and engineering aspects, followed by geological factors. It suggests that in the same region, under similar geological conditions, development and engineering factors play a crucial role in determining the production capacity of tight oil horizontal wells. Based on these findings, these factors can be used as guidelines for the development of tight oil horizontal wells in the Z reservoir, with the aim of achieving higher production capacity.

The experiments conducted above reveal that it is possible to predict the initial production level of horizontal wells through development factors, engineering factors, and geological factors. Furthermore, it validates that using SAE pre-training and fine-tuning for the DNN model can yield good results, even with limited data. In comparison with other methods that rely on dynamic production data to forecast capacity, this approach is more convenient, requires less data, and, most importantly, can be applied to newly developed wells for prediction.

While the research model has shown promising results in predicting the initial production capacity of tight oil horizontal wells, it is evident from the heat map that the influence of the 13 factors listed in the table is somewhat limited. To further enhance the model's accuracy, it is essential to incorporate a more comprehensive set of influencing factors into the training process.

5. Conclusions

In this research, a specialized deep neural network was designed to predict the initial production capacity of tight oil horizontal wells. Initially, 13 influencing factors, including development, engineering, and geological factors, were considered as input parameters for the model. Through preliminary data preprocessing, it was validated that these 13 influencing factors could serve as input neurons for the model. Subsequently, a heat map analysis revealed that the primary factors influencing the initial production capacity of tight oil horizontal wells in the Z region were engineering and development factors. As a result, eight and six positively correlated numerical features were selected as input parameters for the model. During the training and testing processes, 520 data points were utilized. Statistical and graphical analysis of the developed model, as well as its predictions on the testing dataset, demonstrated that the DNN model, trained with SAE pre-training and fine-tuning, exhibited robustness, and could accurately predict the initial production capacity of tight oil horizontal wells (as is shown in Figure 8). Additionally, we chose the optimal model by comparing the number of hidden layers in the DNN. The results indicated that when the number of hidden layers was set to 5, the model achieved the highest accuracy in predicting the initial production capacity of tight oil horizontal wells, with an accuracy rate of 91.13%.

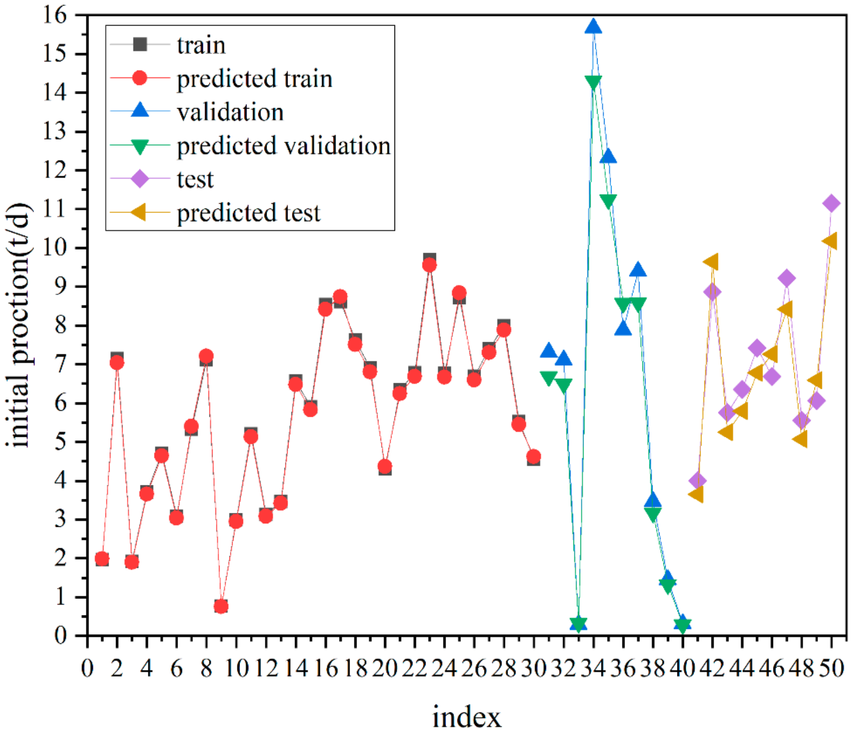


Figure 8. Indices for the training, validation, and testing datasets for initial production capacity.

Although using a large dataset with a DNN model can improve accuracy, when access to a large dataset is limited, employing a DNN model with a small dataset and specialized methods may be a reasonable choice. Due to the confidentiality and scarcity of oilfield data, small datasets are quite common in the field of petroleum engineering. For instance, in the case of initial production capacity prediction for tight oil horizontal wells in this study, the DNN model with SAE pre-training and fine-tuning proved effective.

In conclusion, the results of this study provide strong support and new insights in the field of petroleum production capacity research. By delving into such factors as development, engineering, and geology, we have successfully established a viable method for predicting the initial production capacity of horizontal wells. On the one hand, this method demonstrates not only a high level of accuracy but also practical applicability, as it requires relatively small amounts of data, making it suitable for various types of wells. On the other hand, by validating the deep neural network model by using a small dataset and specialized training methods, this method has showcased its enormous potential in petroleum production capacity research, providing a powerful tool for future research and applications.

The achievements of this study are significant not only for predicting the production capacity of existing wells but also for providing strong support for the development of new wells. It has the potential to expedite decision-making and resource optimization in the petroleum industry, reduce development risks, and improve resource utilization efficiency. Studies in the future are expected to focus on further expansion of the application scope of this method, and refining the model for enhanced accuracy to better meet the needs of oilfield development.

Hopefully, these research findings will have a positive impact on the sustainable development of the petroleum industry, providing robust support for future researches and practices.

Author Contributions: All authors contributed to the study conception and design. Material preparation, and analysis were performed by C.T., Y.C. and X.L. Data curation and Investigation, Y.Y. The first draft of the manuscript was written by Y.Y. The final manuscript was read and approved by X.L., and X.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This work is financially supported by the Scientific Research and Technology Development Project of CNPC (2021DJ3301) and Sponsored by CNPC Innovation Fund (2022DQ02-0202).

Data Availability Statement: The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

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

References

1. Cheng, Y.; Luo, X.; Hou, B.; Zhang, S.; Tan, C.; Xiao, H. Pore Structure and Permeability Characterization of Tight Sandstone Reservoirs: From a Multiscale Perspective. *Energy Fuels* **2023**, *37*, 9185–9196. [CrossRef]
2. Meng, M.; Chen, Z.; Liao, X.; Wang, J.; Shi, L. A well-testing method for parameter evaluation of multiple fractured horizontal wells with non-uniform fractures in shale oil reservoirs. *Adv. Geo-Energy Res.* **2020**, *4*, 187–198. [CrossRef]
3. Keles, C.; Tang, X.; Schlosser, C.; Louk, A.; Ripepi, N. Sensitivity and history match analysis of a carbon dioxide "huff-and-puff" injection test in a horizontal shale gas well in Tennessee. *J. Nat. Gas Sci. Eng.* **2020**, *77*, 103226. [CrossRef]
4. Chen, Y.; Zhou, Y.; Liang, C.; Xu, T.; He, Y. Quantitative Characterization Model of Shale Oil Horizontal Well Production Change. *J. Southwest Pet. Univ. (Sci. Technol. Ed.)* **2021**, *43*, 97–103. [CrossRef]
5. Kadeethum, T.; Salimzadeh, S.; Nick, H. Well productivity evaluation in deformable single-fracture media. *Geothermics* **2020**, *87*, 101839. [CrossRef]
6. Zeng, Q.; Yao, J. Production calculation of multi-cluster fractured horizontal well accounting for stress shadow effect. *Int. J. Oil Gas Coal Technol.* **2020**, *23*, 293–311. [CrossRef]
7. Dai, S.; Zhang, J.; Gao, X.; Wang, X. Analysis of Factors Affecting Productivity of Horizontal Wells Based on Grey Relational Theory. *IOP Conf. Ser. Earth Environ. Sci.* **2018**, *108*, 032048. [CrossRef]
8. Sun, R.; Hu, J.; Zhang, Y.; Li, Z. A semi-analytical model for investigating the productivity of fractured horizontal wells in tight oil reservoirs with micro-fractures. *J. Pet. Sci. Eng.* **2020**, *186*, 106781. [CrossRef]
9. Xie, Y.; He, Y.; Hu, Y.; Jiang, Y. Study on Productivity Prediction of Multi-Stage Fractured Horizontal Well in Low-Permeability Reservoir Based on Finite Element Method. *Transp. Porous Med.* **2022**, *141*, 629–648. [CrossRef]
10. Lu, Y.; Li, H.; Wang, J.; Liu, T.; Wu, K. Productivity evaluation model for multi-cluster fractured wells based on volumetric source method. *Energy Rep.* **2022**, *8*, 8467–8479. [CrossRef]
11. Wang, S.; Chen, Z.; Chen, S. Applicability of deep neural networks on production forecasting in Bakken shale reservoirs. *J. Pet. Sci. Eng.* **2019**, *179*, 112–125. [CrossRef]
12. Huang, R.; Wei, C.; Wang, B.; Yang, J.; Xu, X.; Wu, S.; Huang, S. Well performance prediction based on Long Short-Term Memory (LSTM) neural network. *J. Pet. Sci. Eng.* **2022**, *208*, 109686. [CrossRef]
13. Sagheer, A.; Kotb, M. Time series forecasting of petroleum production using deep LSTM recurrent networks. *Neurocomputing* **2019**, *323*, 203–213. [CrossRef]
14. Tang, J.; Peng, C. Predicting Method of the Initial Productivity for the Horizontal Well in Fuling Shale Gas Reservoirs. *Pet. Geol. Oilfield Dev. Daqing* **2020**, *39*, 160–167. [CrossRef]
15. Yu, Q.; Mu, Z.; Liu, P.; Hu, X.; Li, Y. A new evaluation method for determining reservoir parameters for the development of edge-water-driven oil reservoirs. *J. Pet. Sci. Eng.* **2019**, *175*, 255–265. [CrossRef]
16. Chen, L.; Liu, Z.; Ma, N.; Wang, Y. Prediction of Oilfield-Increased Production Using Adaptive Neurofuzzy Inference System with Smoothing Treatment. *Math. Probl. Eng.* **2019**, *37*, 9185–9196. [CrossRef]
17. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507. [CrossRef]
18. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
19. Feng, S.; Zhou, H.; Dong, H. Using deep neural network with small dataset to predict material defects. *Mater. Des.* **2019**, *162*, 300–310. [CrossRef]
20. Cortes, C.; Mohri, M.; Rostamizadeh, A. L2 regularization for learning kernels. *arXiv* **2012**, arXiv:1205.2653. [CrossRef]
21. MacKay, D.J.C. Bayesian Interpolation. *Neural Comput.* **1992**, *4*, 415–447. [CrossRef]
22. Yan, J.; He, X.; Zhang, S.; Feng, C.; Wang, J.; Hu, Q.; Cai, J.; Wang, M. Sensitive parameters of NMR T2 spectrum and their application to pore structure characterization and evaluation in logging profile: A case study from Chang 7 in the Yanchang Formation, Heshui area, Ordos Basin, NW China. *Mar. Pet. Geol.* **2020**, *111*, 230–239. [CrossRef]
23. Liu, X.; Liu, Y.; Lu, Z.; Zhang, T.; Yu, L.; Yi, T.; Xu, H.; Lei, Y.; Zang, Q.; Awan, R.S.; et al. Study on the characteristics and influencing factors of Chang 7 ultralow-porosity and low-permeability reservoirs in the Heshui area, Ordos Basin. *Interpretation* **2022**, *10*, T581–T593. [CrossRef]

24. Zhou, Z.; Wang, G.; Ran, Y.; Lai, J.; Cui, Y.; Zhao, X. A logging identification method of tight oil reservoir lithology and lithofacies: A case from Chang7 Member of Triassic Yanchang Formation in Heshui area, Ordos Basin, NW China. *Pet. Explor. Dev.* **2016**, *43*, 65–73. [CrossRef]
25. Fang, W.; Qin, X.; Liu, C.; Wang, R.; Pu, J.; Jiang, H.; He, W. Quantitative Evaluation of Well Performance Affected by Fracture Density and Fracture Connectivity in Fractured Tight Reservoirs. *Geofluids* **2021**, *2022*, 2805348. [CrossRef]
26. Jiang, D.-L.; Chen, H.; Xing, J.-P.; Shang, L.; Wang, Q.-H.; Sun, Y.-C.; Zhao, Y.; Cui, J.; Duncan, I. A novel method of quantitative evaluation and comprehensive classification of low permeability-tight oil reservoirs: A case study of Jidong Oilfield, China. *Pet. Sci.* **2022**, *19*, 1527–1541. [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.