

Article

Transfer Learning-Based Dynamic Production Prediction for Tight Oil: Considering Multimodal Features and Long-Term Temporal Dependencies

Qingying Lin ¹, Minghai Zhang ¹, Tiancong Mao ¹, Yunwei Kang ^{2,*}, Xingcan Li ², Xianyang Sun ², Dali Guo ² and Zixi Guo ^{3,*} 

¹ Baikouquan Oil Production Plant, PetroChina Xinjiang Oilfield Company, Karamay 834000, China

² School of Sciences, Southwest Petroleum University, Chengdu 610500, China; lxswpu@163.com (X.L.); sunxianyangswpu@163.com (X.S.); guodali@swpu.edu.cn (D.G.)

³ State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation, Southwest Petroleum University, Chengdu 610500, China

* Correspondence: kangyw_swpu@163.com (Y.K.); guozixi@vip.163.com (Z.G.)

Abstract

Tight oil horizontal wells in the Mahu block of the Junggar Basin commonly show rapid production decline and limited target-domain samples. These characteristics make accurate production prediction difficult. This work aims to address the small-sample overfitting problem of tight oil horizontal well production prediction in Mahu Sag with rapid production decline and limited measured well data. The concept of transfer learning is introduced to address the issue of insufficient target domain samples, and the Pearson correlation coefficient is utilized to select the main controlling factors for production from the production data. Next, based on the features extracted by the temporal convolutional network at different data scales, a multi-head attention mechanism is introduced to capture the dependencies across different time steps. Subsequently, an improved sparrow search algorithm is employed to optimize the hyperparameters of the bidirectional long short-term memory network. Finally, the bidirectional long short-term memory network is integrated to further extract the nonlinear features learned by the temporal convolutional network to conduct production prediction. Tailored to the exploitation conditions of tight oil horizontal wells in this block, a tight oil production prediction model based on transfer learning and the multi-head attention mechanism is proposed. Experimental results demonstrate that, compared with the standard bidirectional long short-term memory network, the proposed model's evaluation metrics show a 60.93% decrease in root mean square error, a 78.53% decrease in mean absolute percentage error, and a 43.68% increase in coefficient of determination. This verifies the effectiveness of transfer learning in solving small-sample modeling challenges, providing precise technical support for the optimization of tight oil fracturing parameters and stimulation treatments in the Mahu block. The novelty of this work lies in the integration of multi-head attention temporal convolution network, quantum sparrow optimized bidirectional long short-term memory network and cross-block transfer learning for small-sample tight oil forecasting.

Keywords: tight oil; production prediction; transfer learning; multimodal features; bidirectional long short-term memory



Academic Editor: Albert Ratner

Received: 3 June 2026

Revised: 19 June 2026

Accepted: 23 June 2026

Published: 25 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Unconventional hydrocarbon resources, including shale oil and tight oil, have become crucial for global energy supply. Their effective development relies on advanced stimulation techniques such as horizontal drilling and multi-stage hydraulic fracturing. Unlike conventional reservoirs, these formations exhibit ultralow permeability and strong heterogeneity, necessitating accurate production forecasting for economic optimization. As noted in recent studies [1], the production behavior of such reservoirs is highly sensitive to both geological and engineering parameters. In particular, tight conglomerate reservoirs, represented by the Mahu sag in the Junggar Basin, have become the main battleground for increasing reserves and production. However, under the combined influence of strong reservoir heterogeneity, low permeability, and large-scale volume fracturing technologies, the production dynamics of horizontal wells in this type of reservoir often exhibit characteristics of rapid initial decline, large fluctuations in water cut, and strong nonlinearity coupled with multiple factors. Traditional production prediction methods based on physical seepage mechanisms or empirical formulas often struggle to balance computational efficiency and prediction accuracy when dealing with such complex geological-engineering systems. With the deep integration of digital transformation in the oil and gas industry, leveraging artificial intelligence technology to mine the spatiotemporal evolution patterns behind production data and establishing high-precision production prediction models to guide the optimization of fracturing parameters [2–4] has become a research hotspot for achieving scientific development, cost reduction, and efficiency enhancement in tight oil exploitation [5]. It should be clarified that shale oil and tight oil are two distinct unconventional resources. Shale oil refers to petroleum generated and retained in organic-rich shale formations with ultralow matrix permeability, requiring extensive hydraulic fracturing to achieve economic flow. In contrast, tight oil is stored in tight sandstone or conglomerate reservoirs with better porosity but still low permeability, often exhibiting rapid production decline. Our study focuses on tight conglomerate reservoirs in the Mahu sag, which differ from typical shale plays in lithology and fracture-network behavior.

In recent years, production prediction methods for oil and gas fields have undergone a leap from traditional numerical simulation to data-driven deep learning. Wang et al. [6] and Fan et al. [7] explored the application of machine learning and ensemble learning to unconventional oil and gas production prediction at an early stage [8]. To extract nonlinear spatiotemporal features from production data, Zhang et al. [9] and Pan et al. [10] predicted shale oil production using a Convolutional Neural Network (CNN) combined with Bidirectional Long Short-Term Memory (BiLSTM) networks. Meanwhile, to overcome the blindness of manual parameter tuning, Zhu et al. [11], Li et al. [12] and Tang et al. [13] respectively employed automatic hyperparameter optimization, Particle Swarm Optimization (PSO), and Multi-strategy Sparrow Search Algorithm (MSSA) to automatically optimize the hyperparameters of neural networks. Qiu et al. [14] combined complete ensemble empirical mode decomposition with SSA-CNN-BiLSTM to handle complex short-term time series, while He et al. [15] utilized Simulated Annealing (SA) combined with Long Short-Term Memory (LSTM) for tight reservoir prediction. However, a limitation of these baseline deep learning and optimization models is that when dealing with long-sequence production data from tight oil wells, the receptive field of standard networks is restricted, making it difficult to effectively capture long-distance historical decline features and multi-scale temporal dependencies. To break through the bottleneck of sequence feature extraction, frontier research has begun to introduce attention mechanisms and complex spatiotemporal networks. Zhang et al. [16] utilized the efficient feature extraction capability of Temporal Convolutional Networks (TCN) to process the dynamic single-well time series of water-driven oilfields. Pan et al. [17] and Meng et al. [18] introduced self-attention

and cross-attention mechanisms into CNN-LSTM and BiLSTM models, assigning higher weights to key time steps. Li et al. [19], Sun et al. [20] and Kazemi et al. [21] deeply integrated Multi-Head Attention (MHA) mechanisms and Transformer architectures with CNN or U-Net to capture global internal correlations. Although complex models with attention mechanisms can improve prediction accuracy, they usually require large datasets for parameter training. For new horizontal wells in the early development stage of the Mahu block, production histories are short and target-domain samples are limited. Therefore, these models are prone to overfitting when trained only with target-domain data. To address this small-sample dilemma, some scholars have turned their attention to transfer learning. Desbordes et al. [22] pioneered research on dynamic production optimization based on transfer learning; Qian et al. [23] proposed a small-sample quality prediction method integrating deep transfer learning. Addressing the scarcity of well data, Niu et al. [24] proposed an ensemble transfer learning strategy to achieve cross-block production prediction for shale gas wells. Dong et al. [25] and Al-Alimi et al. [26] constructed a hybrid prediction framework combining stacked LSTM with transfer learning, effectively coping with short production times and small sample sizes. However, existing transfer-learning studies for production prediction still have limitations. First, most transfer models use a relatively single feature extractor. They do not fully combine the wide receptive field of TCN, the global dependency modeling of MHA, and the bidirectional memory of BiLSTM. Second, after the pre-trained model is transferred to the target domain, adaptive hyperparameter fine-tuning for scarce target-domain samples is often insufficient. This may cause the model to fall into local optima. Some scholars have also utilized the experience of other oil reservoirs to carry out production prediction [27].

However, existing transfer-learning studies still have limitations, such as single feature extraction, insufficient target domain tuning, and an easy fall into local optima. To fill these gaps, this paper proposes a novel production prediction framework that fuses transfer learning strategies with hybrid deep neural computing. The core of this method lies in breaking the limitations of traditional isolated modeling, utilizing transfer learning to map prior knowledge from mature blocks to the target block, thereby mitigating the risk of overfitting caused by small samples right from the data source. At the feature extraction level, a composite network structure containing TCN and MHA mechanisms is designed [28]. It captures local abrupt changes and long-term trends of production curves in parallel through multi-scale convolutional kernels, enhances the influence of key time steps using attention weights, and is supplemented by a bidirectional BiLSTM network to further excavate the nonlinear dependencies among the data. Meanwhile, to address the difficulty of hyperparameter optimization in deep networks [29], an improved quantum sparrow search algorithm is introduced for adaptive solving, thereby achieving a full-process intelligent prediction spanning from data transfer and feature mining to parameter closed-loop optimization [30–32].

2. Materials and Methods

2.1. Overview of the Study Area

The Mahu Oilfield is located in the Junggar Basin, Xinjiang Uyghur Autonomous Region, China. It is a large-scale tight oilfield with abundant petroleum reserves and is the world's largest tight conglomerate oilfield, boasting an exploration area of 7300 square kilometers and reserves on the scale of one billion tons. The reservoirs in this oilfield are mainly distributed in channel and sand bar structures [33]. The reservoir rock mass is composed of overlapping layers of sandy conglomerate, exhibiting continuity and tightness. Furthermore, the tight oil here is characterized by low permeability and strong heterogeneity. To enhance the production of horizontal wells, the fracturing technology of

long-interval multi-cluster, extreme limited-entry perforating, and intra-stage temporary plugging has been adopted for horizontal wells in the study area. The production of horizontal wells in the Mahu Oilfield also features a multi-stage decline, making it difficult for conventional methods [34,35], such as empirical formulas and numerical simulation, to accurately predict the horizontal well production.

2.2. Algorithm Fundamentals

2.2.1. Pearson Correlation Coefficient

The Pearson correlation coefficient is a statistic that measures the strength and direction of the linear relationship between two variables. Its calculation formula is as follows:

$$r = \frac{\sum_{i=1}^n (x_i - x^*)(y_i - y^*)}{\sqrt{\sum_{i=1}^n (x_i - x^*)^2 \sum_{i=1}^n (y_i - y^*)^2}} \quad (1)$$

where x_i and y_i are each pair of observed values in the sample data, x^* and y^* are the sample means of x and y respectively, and n is the number of samples.

2.2.2. Transfer Learning

When the target domain is constrained by severe costs and extreme data scarcity, causing traditional deep learning models to easily suffer from poor generalization performance, the introduction of transfer learning becomes an effective breakthrough approach. The core advantage of this method lies not only in its technical mechanism of cross-domain data alignment but also in the high adaptability of this mechanism to the small sample cold-start problem. It effectively utilizes the attribute that the target domain lacks superficial data but shares homologous underlying features with the source domain. Through the mapping and coordination of the feature space, the prior knowledge contained in the source domain is transferred and applied to the target domain, thereby effectively dismantling the information barriers caused by data silos.

During this process, it is first necessary to establish a distribution similarity evaluation system between the source domain and the target domain data, and knowledge transfer is realized when specific conditions are met. To effectively reduce the cross-domain data distribution discrepancy, the Maximum Mean Discrepancy (MMD) in a Reproducing Kernel Hilbert Space (RKHS) is currently widely used as a measure of distribution distance. This metric maps samples into a high-dimensional feature space via a kernel function and quantifies the distribution difference by calculating the Euclidean distance between the means of the mapped data. Its mathematical expression can be represented as:

$$MMD(X, Y) = \left\| \frac{1}{n} \sum_{i=1}^n \phi(x_i) - \frac{1}{m} \sum_{j=1}^m \phi(y_j) \right\|^2 \quad (2)$$

where x_i and y_i represent two samples, $\phi(\cdot)$ represents the mapping, and m and n represent the domains of the samples, respectively. The smaller the MMD between the two samples, the more identical their distributions are.

2.2.3. Bidirectional Long Short-Term Memory Network

When processing complex sequence data with deep contextual correlations, traditional LSTM, built upon a unidirectional time-delay mechanism, is highly prone to falling into information myopia. That is, it can only rely on historical experience and cannot utilize subsequent sequences to reverse correct the current state, causing severe context

fragmentation. The core value of BiLSTM originates exactly from the high adaptability of its network architecture to this pain point: it breaks the non-real-time limitation of unidirectional information flow. By constructing parallel forward and backward dual hidden layer states, it deeply fuses the historical accumulation and future context of the sequence at the current node. This mechanism fundamentally makes up for the structural defects of unidirectional models in capturing global information, ensuring a complete characterization of the dynamic laws across the entire sequence [30]. Its network structure diagram is shown in Figure 1.

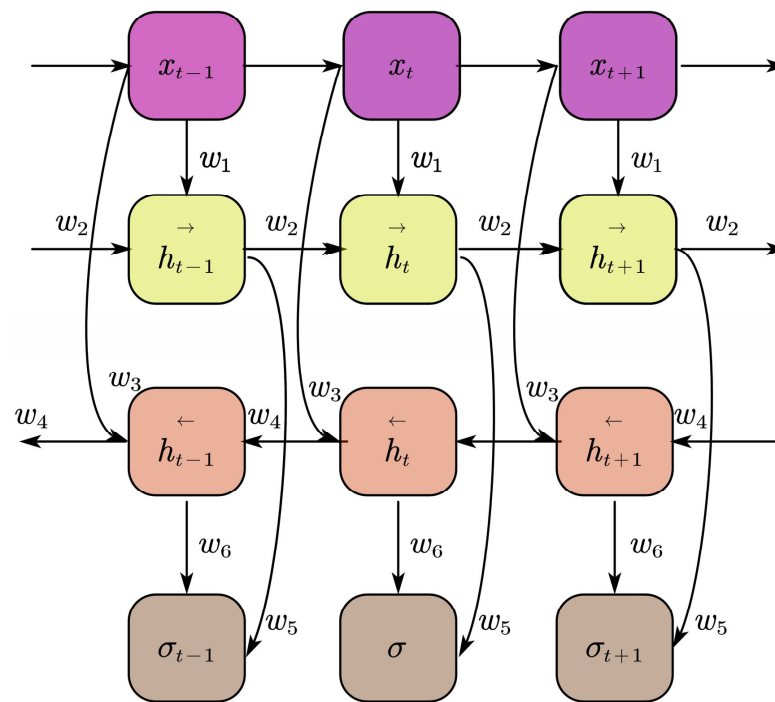


Figure 1. Structure diagram of the BiLSTM network.

The internal calculation formulas of the network structure are as follows:

$$\begin{cases} \vec{h}_t = f(w_1 x_t + w_2 \vec{h}_{t-1}) \\ \leftarrow h_t = f(w_3 x_t + w_4 \leftarrow h_{t+1}) \\ \sigma_t = g(w_5 \vec{h}_t + w_6 \leftarrow h_t) \end{cases} \quad (3)$$

In the above formulas, $\vec{h}_t$, $\leftarrow h_t$, and σ_t are the forward hidden layer state, the backward hidden layer state, and the output at time t , respectively; x_t is the input feature vector at time t ; f and g correspond to the LSTM function and loss function, respectively, and W_i represents the corresponding weights.

2.2.4. Quantum Sparrow Search Algorithm

For high-dimensional, non-convex, and multimodal optimization tasks, traditional heuristic algorithms may suffer from insufficient convergence accuracy or premature convergence. To address this problem, this study uses a quantum-inspired QSSA. QSSA introduces probability-amplitude representation and phase-angle updating to improve population diversity and global exploration. This underlying mechanism highly aligns with the requirement of preventing stagnation in complex solution space exploration. It fundamentally breaks the spatial limitations of conventional searches and effectively

resolves the core pain point of premature loss of population diversity in high-dimensional spaces. Furthermore, compared to other classical swarm intelligence algorithms, QSSA exhibits multiple performance advantages [31].

It should be noted that the QSSA used in this study is a quantum-inspired swarm intelligence optimization algorithm rather than an implementation on a real quantum computer. The term quantum refers to the use of quantum-behaved coding ideas, such as probability-amplitude representation, phase-angle updating, and rotation-gate-inspired position updating, to enhance population diversity and global exploration. Therefore, QSSA does not require quantum hardware, and all computations in this study were performed on a classical computer.

Compared with the Genetic Algorithm (GA), which relies on random crossover and mutation operations, QSSA utilizes quantum rotation gates for positioning, achieving directed optimization and significantly improving iteration efficiency. Unlike PSO, which easily leads to premature convergence due to excessively following historical optimal solutions, the quantum characteristics of QSSA can effectively maintain the independent exploration space of individuals, avoiding blind assimilation of the population. In addition, compared to SA, which relies on a single-path search and incurs high time costs, QSSA not only possesses the search advantages of a population but also achieves a better dynamic balance between global exploration breadth and local exploitation precision through quantum-inspired probability-amplitude coding and adaptive position updating.

The specific steps include five key stages:

Quantum Encoding and Initialization: To solve the problem of unified optimization for mixed parameters (continuous or discrete), a quantum superposition state encoding mechanism is adopted. Each optimization parameter corresponds to a qubit, whose state is represented as:

$$|\psi_i\rangle = \bigotimes_{j=1}^D (\cos \theta_{i,j}|0\rangle + \sin \theta_{i,j}|1\rangle) \quad (4)$$

where $\theta_{i,j}$ is the phase angle and D is the parameter dimension. By designing special mapping rules, continuous parameters adopt the probability projection of $\sin^2(\theta)$, while discrete parameters are determined by the maximum inner product between the quantum state and a preset ground state, thereby achieving a unified quantum representation of the parameter space.

Discoverer Quantum Exploration: In the early stage of iteration, the algorithm endows the top 20% of high-quality individuals (discoverers) with stronger global search capabilities. Their position update strategy is dynamically adjusted based on the environmental safety level.

$$\theta_{i,j}^{t+1} = \begin{cases} \theta_{i,j}^t \exp\left(-\frac{i}{\alpha T}\right) + Q\epsilon & R^2 < ST \\ \theta_{i,j}^t + QL & \end{cases} \quad (5)$$

When the random warning value R^2 is lower than the safety threshold ST , the quantum tunneling perturbation ($\epsilon \sim N(0, 1)$) is executed to simulate the quantum effect of particles penetrating energy barriers. Otherwise, a Lévy flight is performed (where L is the characteristic step size), utilizing the heavy-tailed distribution to enhance cross-regional search capabilities. Q is the quantum perturbation factor, ϵ is the Gaussian noise term, iter_max is the maximum number of iterations, and a is the exponential decay factor.

Follower Collaborative Exploitation: The remaining 80% of individuals act as followers, and their update process reflects the swarm intelligence sharing mechanism. A nonlinear following strategy is constructed by introducing a tangent function:

$$\theta_{i,j}^{t+1} = \theta_{best,j} + \frac{1}{D} \sum_{k=1}^D |\theta_{k,j} - \theta_{k,j}^t| \tan(\pi r) \quad (6)$$

where $\tan(\pi r)$ is a linear scaling term, and $r \sim U(0, 1)$ is a matrix. This design enables followers to automatically adjust their step sizes based on population distribution differences when tracking the optimal solution; a fine search is adopted when local aggregation is high, while a large-step exploration is executed when dispersed.

Cauchy Quantum Mutation Mechanism: To prevent premature convergence, the algorithm introduces an adaptive mutation strategy. A Cauchy perturbation is applied to randomly selected individuals:

$$\theta_{mut,j}^{t+1} = \theta_{best,j} + \gamma (\theta_{best,j} - \theta_{mut,j}^t) e^{-\beta t/T} \quad (7)$$

The long-tail characteristic of the Cauchy distribution ($\gamma \sim \text{Cauchy}(0, 1)$) endows the algorithm with a stronger ability to explore extreme values. γ is the Cauchy perturbation factor, β is the decay coefficient, and t is the current number of iterations. The exponential decay term $e^{-\beta t/T}$ weakens the perturbation intensity in the later stages of iteration to achieve balance.

Dynamic Convergence Condition: The algorithm adopts a dual convergence criterion. Convergence is deemed to have been reached when either the phase angle change rate $\max |\theta_{i,j}^t - \theta_{i,j}^{t-1}| < 10^{-6}$ or the improvement amplitude of the optimal fitness is less than 10^{-4} for 20 consecutive iterations.

2.3. Proposed Model Architecture

2.3.1. MHA-TCN Network Structure

Although TCN can capture short-term dependencies to a certain extent through convolution operations, it still faces challenges in modeling long-term dependencies. Particularly when the sequence length is extremely long, TCN may encounter limitations due to the size of the convolution kernels. Introducing the MHA mechanism into TCN allows for better modeling of global dependencies. In TCN, convolution operations primarily focus on local receptive fields, whereas MHA complements the global dependencies by calculating the attention across all time steps in the time series, making the model more efficient and accurate in capturing long-term dependency relationships. By dynamically adjusting the attention weights, MHA can highlight important time-step information. In this way, when predicting oil production, it can determine which features are more important for the prediction at the current time step and better handle noise in the time-series data. Furthermore, through multi-head parallel processing and residual connections [32], MHA can achieve a certain regularization effect, reduce the possibility of overfitting, and further improve the prediction accuracy of the model. Its structure is shown in Figure 2.

In Figure 2, Kernel 1, Kernel 2, and Kernel 3 represent different convolutional kernels with sizes of 2, 4, and 8, respectively, where the function of Kernel 1 is to increase the data dimensionality. The original TCN generally extracts features at various time scales through convolutional units of different lengths, and then fuses these features using the parallel concatenation strategy designed in this method. The fused features are passed to the improved MHA module, which extracts multi-dimensional data features and finally outputs them.

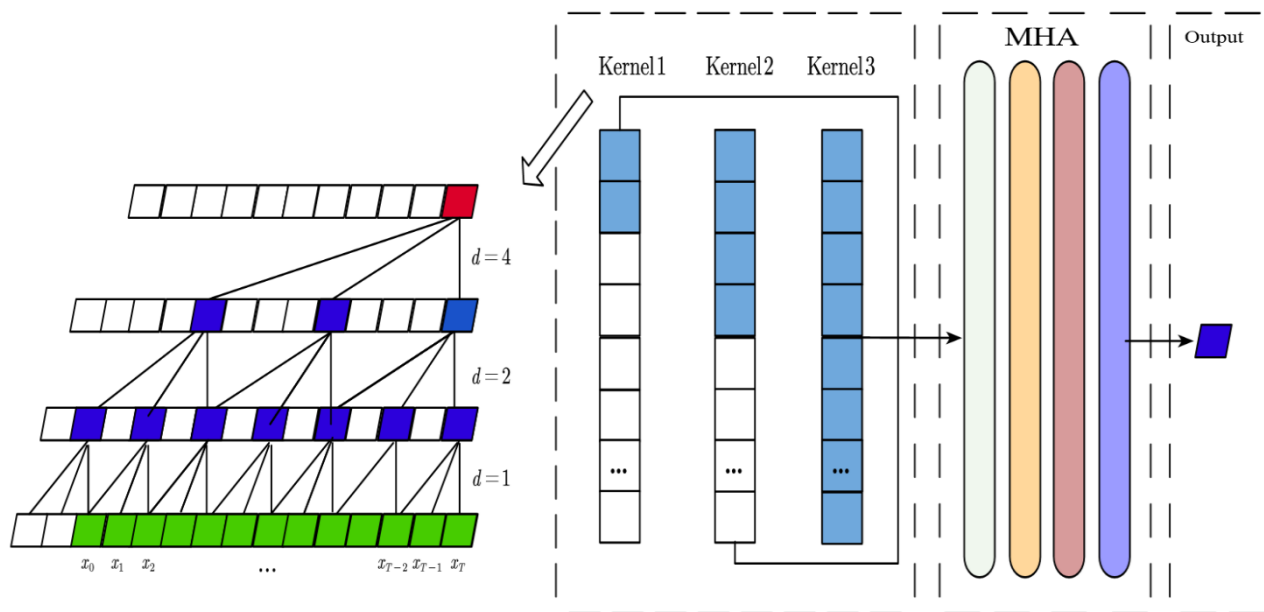


Figure 2. MHA-TCN network structure. The left part illustrates the dilated convolution applied to the green original sequence. The blue blocks in the middle represent the multi-scale feature sequences extracted by three convolutional kernels of different scales. The long bars in the MHA layer correspond to the outputs of different attention heads.

In MHA-TCN network structure, the MHA module is introduced after the multi-scale TCN feature-fusion layer. According to the optimized model setting, the number of filters is 64; therefore, the fused TCN feature representation is projected into a 64-dimensional feature space before entering the MHA module. The query, key, and value dimensions are all set to 64 to maintain dimensional consistency. The number of attention heads is set to 4, and the dimension of each head is 16. The outputs of all attention heads are concatenated and mapped back to the 64-dimensional feature space before being passed to the BiLSTM layer.

2.3.2. QSSA-TCN-BiLSTM Prediction Model

Aiming at the deficiencies of traditional oil production prediction methods in terms of accuracy and timeliness, this paper proposes a QSSA-TCN-BiLSTM tight oil production prediction algorithm incorporating a multi-head attention mechanism. Its overall framework is shown in Figure 3. First, the relationships among different features of the obtained source domain data are analyzed, and the feature factors are selected using Pearson correlation analysis. Data preprocessing is then performed, including the removal of outliers, imputation of missing values, and data normalization, in order to reduce the impact of different data dimensions on the training effect of the neural network. Next, the preprocessed data is sequentially input into the MHA-TCN layer and the BiLSTM layer. The role of MHA-TCN is to capture features at different time scales, while BiLSTM is utilized to better capture the dependencies within the time-series data. Meanwhile, the hyperparameters of the neural network are optimized through the QSSA algorithm. After the training is completed and the iteration process is finished, the model updates its learned parameters and weights, concludes the iterative process, and finally outputs the prediction results.

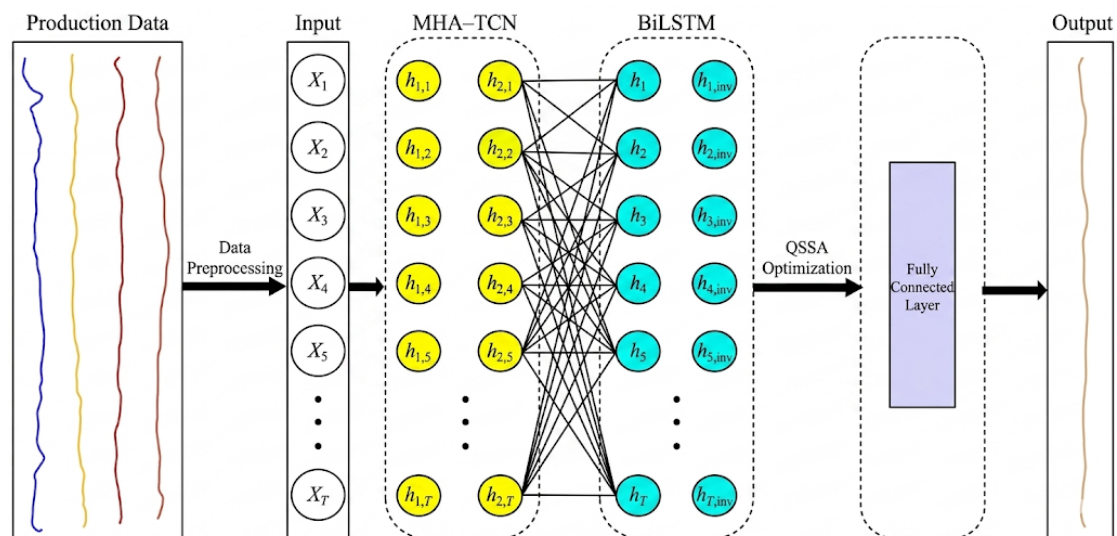


Figure 3. Flowchart of the QSSA-TCN-BiLSTM network structure. Yellow nodes denote MHA-TCN units for extracting multi-scale temporal features and long-range dependencies; cyan nodes represent the forward and reverse hidden states of the BiLSTM layer to model bidirectional temporal correlations. Optimized by the QSSA algorithm, the light-purple fully connected layer generates the final prediction output.

2.4. Data Description and Experimental Setup

The source domain dataset for the experiment in this paper selects the production data of 163 wells from another block (Block X) in this oilfield. Among them, the production data of 100 wells are used as the training set, and the production data of the remaining 63 wells are used as the test set. The target domain dataset for the experiment selects the production data of 50 wells in Block M. The Maximum Mean Discrepancy (MMD) between the two blocks is 0.02137. The data from Block X is used to pre-train the TCN, and then the trained TCN weights are saved and directly transferred to Block M for use.

The source-domain dataset contains production records from 163 wells in Block X, of which 100 wells were used for training and 63 wells were used for testing. The target-domain dataset contains production records from 50 wells in Block M. Each well was treated as a daily oil-production time series. In the prediction experiment, Wells M1 and M2 were selected as representative wells in the target block, and rolling prediction was performed for the next 100 production days. Therefore, the prediction sequence of each representative well contained 100 daily oil-production time points. Because the wells were put into production at different times and had different production durations, the complete historical production-sequence length was not identical for all wells. For the 163 source-domain wells, the production-sequence length ranged approximately from 1200 to 1500 days.

To clarify the comparability between the source and target domains, the geological and engineering backgrounds of Block X and Block M are further described. Block X and Block M are both located in the same Mahu Oilfield system in the Junggar Basin. According to the study area description, the Mahu Oilfield is a large-scale tight conglomerate oilfield. Its reservoirs are mainly distributed in channel and sand-bar structures, and the reservoir rock mass is composed of overlapping sandy conglomerate layers with strong continuity and tightness. The reservoir is characterized by low permeability and strong heterogeneity.

In terms of engineering development, horizontal wells in the study area are stimulated using long-interval multi-cluster fracturing, extreme limited-entry perforating, and intra-stage temporary plugging. The production behavior of horizontal wells is characterized by multi-stage decline, rapid early production decline, and local fluctuations. Block X and

Block M share similar reservoir type, lithological background, fracturing technology, and production-decline behavior. These similarities provide the geological and engineering basis for transferring production-decline knowledge from Block X to Block M.

Due to the different dimensions of the input parameters, the order-of-magnitude differences between the data may affect the calculation speed and prediction accuracy of the model; therefore, data preprocessing is required. We adopted the min-max normalization method for data quantization, and the formula is as follows:

$$X = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (8)$$

where x_i is the original data point, $x_{\max}$ is the maximum value in the data, $x_{\min}$ is the minimum value in the data, and x is the normalized data.

2.4.1. Transfer-Learning and Fine-Tuning Strategy

The transfer-learning process consisted of two stages: source-domain pre-training and target-domain fine-tuning. In the source-domain pre-training stage, the MHA-TCN feature extractor was trained using the production sequences from Block X. The purpose of this stage was to learn transferable multi-scale temporal representations of tight oil production decline. The transferred parameters included the convolution kernel weights, dilation-convolution weights, residual-connection parameters, normalization parameters, and the parameters of the MHA module embedded after the TCN feature-fusion layer. The BiLSTM prediction layer and the final fully connected output layer were not directly reused because their parameters are more sensitive to the production scale and local dynamic characteristics of the target block.

In the target-domain fine-tuning stage, the pre-trained MHA-TCN parameters were loaded into the model for Block M. To avoid overfitting caused by the limited number of target-domain wells, the first two TCN blocks were frozen during the initial fine-tuning stage, whereas the last two TCN blocks, the MHA module, the BiLSTM layer, and the output layer were kept trainable. After the validation loss became stable, all layers were unfrozen and jointly fine-tuned with a smaller learning rate. The learning rate used during transfer fine-tuning was set lower than that in source-domain pre-training, so that the model could preserve the general temporal-production representations learned from Block X while adapting to the local production characteristics of Block M.

The fine-tuning process adopted early stopping based on the validation-set RMSE. Training was stopped if the validation RMSE did not improve for 20 consecutive epochs. The maximum number of fine-tuning epochs was 300, consistent with the optimized iteration setting reported in Table 1. This layer-freezing and gradual-unfreezing strategy was used to balance knowledge retention and target-domain adaptation under small-sample conditions.

Table 1. Experimental settings of the proposed model.

Category	Item	Setting
Experimental Setup	Source domain	Block X
	Source-domain wells	163 wells
	Source-domain split	100 wells for training and 63 wells for testing
	Target domain	Block M
	Target-domain wells	50 wells
	Representative prediction wells	Well M1 and Well M2
	Prediction length	Next 100 production days

Table 1. *Cont.*

Category	Item	Setting
Data preprocessing	/	Outlier removal, missing-value imputation, and min-max normalization
Input Features	Main input factors	Oil layer thickness, formation pressure, permeability, porosity, oil saturation, and production data
TCN Architecture	Layers	4
	Dilation coefficients	[1, 2, 4, 8]
	Multi-scale convolution kernel	2, 4, and 8
	Optimizer	Adam
QSSA Optimization	Activation function	ReLU
	Population size	N = 30
	Maximum search iterations	T = 50

2.4.2. Experimental Setup

The experimental settings are summarized to improve the reproducibility of the proposed model. The source-domain dataset contains production data from 163 wells in Block X, of which 100 wells were used for training and 63 wells were used for testing. The target-domain dataset contains production data from 50 wells in Block M. In the representative prediction cases, Wells M1 and M2 were selected for rolling prediction of daily oil production over the next 100 days.

Before model training, data preprocessing was performed, including outlier removal, missing-value imputation, and min-max normalization. Pearson correlation analysis was used to select the main production-controlling factors. According to the correlation analysis, oil layer thickness, formation pressure, permeability, porosity, and oil saturation were selected and fed into the neural network together with the production data. The parameter settings for the experiment are shown in Table 1.

2.5. Evaluation Metrics

The evaluation metrics selected in this paper are root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R^2). They are defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y'_i)^2} \quad (9)$$

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - y'_i}{y_i} \right| \quad (10)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - y'_i)^2}{\sum_{i=1}^n (y_i - y^*)^2} \quad (11)$$

where n is the number of samples, y_i is the measured production, y'_i is the predicted production, and y^* is the mean value of the measured production.

3. Results and Discussion

3.1. Analysis of Main Controlling Factors

The well production of the tight conglomerate reservoir in the target block is comprehensively affected by multi-dimensional factors. Based on the actual collected and

effectively analyzed production data, the key feature parameters influencing production can be summarized into the following three core dimensions:

- (1) Reservoir geological and physical property parameters: including oil layer thickness, formation pressure, porosity, permeability, oil saturation, as well as shale content and resistivity. These parameters constitute the static geological foundation of the reservoir and directly determine the fluid enrichment degree, natural transport capacity, and well logging response characteristics of the reservoir.
- (2) Engineering and fracturing stimulation parameters: covering total proppant volume, total fluid volume, and average final shut-in pressure. In the development of tight conglomerate reservoirs, these parameters represent the scale of engineering modification and the fracturing intensity, directly relating to the complexity of the artificial fracture network and the drainage volume.
- (3) Wellbore trajectory parameters: including azimuth and deviation angle. The matching degree between the spatial distribution characteristics of the well and the local stress field and natural fracture orientation deeply affects the fracture initiation and extension effects during fracturing.

However, due to the limitations of environmental monitoring instruments and measurement process conditions in the actual production of the Mahu Oilfield, it is difficult to comprehensively acquire all theoretical parameters [36]. To scientifically screen out the main controlling parameters of horizontal well production under existing data conditions, and to provide a reliable feature reference for subsequent intelligent production prediction and production optimization, we conducted a comprehensive Pearson correlation analysis on 13 key indicators targeting the production metric. The collinearity among the parameters and their response sensitivity to production are shown in Figure 4.

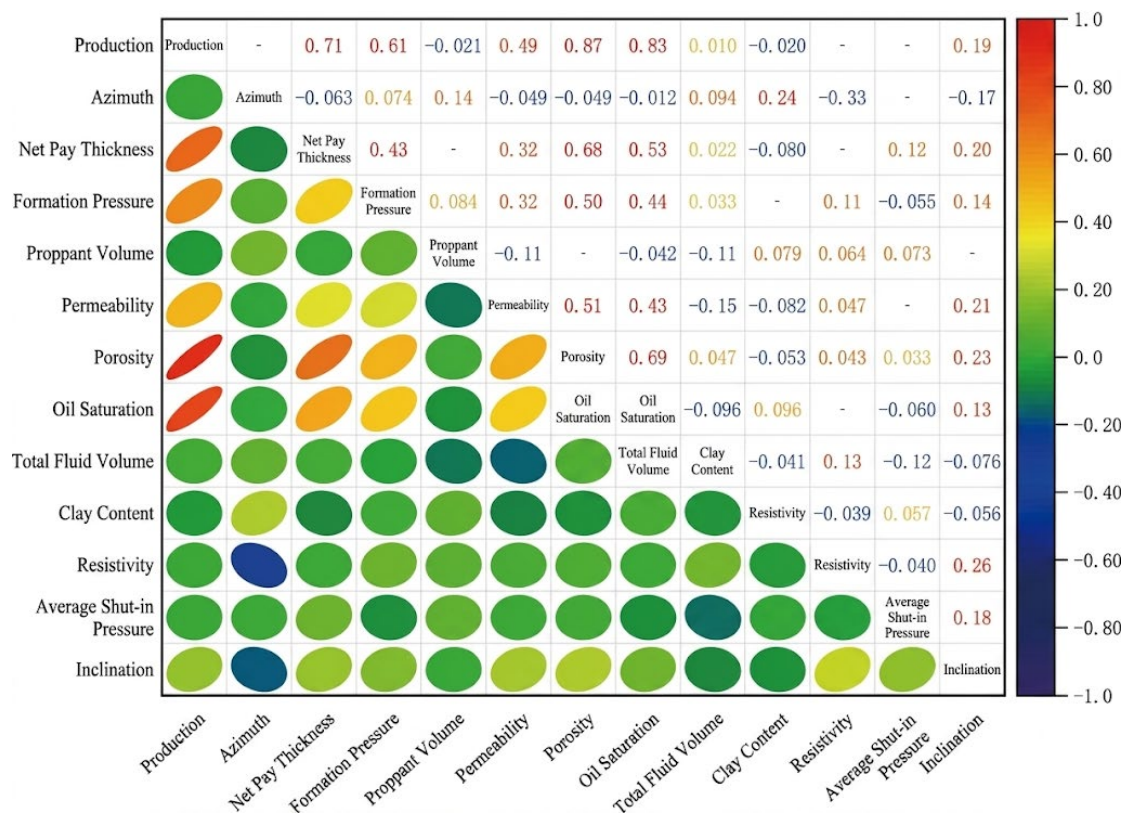


Figure 4. Correlation heatmap. Warmer colors and flatter shapes correspond to a higher degree of correlation.

Figure 4 presents the correlation heatmap; the larger the correlation coefficient, the stronger the correlation between the features. As can be seen from the first column in the figure, the correlation between production and oil layer thickness is 0.71, with formation pressure is 0.61, with permeability is 0.49, with porosity is 0.87, and with oil saturation is 0.83. Therefore, these features are fed into the neural network together with the production data for training.

3.2. Hyperparameter Optimization

To solve the problems of strong subjectivity in manual hyperparameter setting and difficulty in converging to the global optimum in deep neural networks, this paper adopts the improved QSSA to adaptively optimize the key parameters of the BiLSTM layer.

The experimental infrastructure is set up as follows: The TCN layer is configured with a 4-layer structure, and the dilation coefficients are sequentially set to [1, 2, 4, 8] to ensure that the model has the capability of multi-scale temporal feature extraction; the optimizer is Adam, and the activation function is ReLU. For the BiLSTM layer, the QSSA algorithm conducts a global search within a preset search space for parameters, including the number of hidden layer neurons, convolution kernel size, Dropout ratio, learning rate, and the number of iterations. After iterative convergence, the obtained optimal hyperparameter combination is shown in Table 2.

Table 2. Optimization results of the QSSA algorithm.

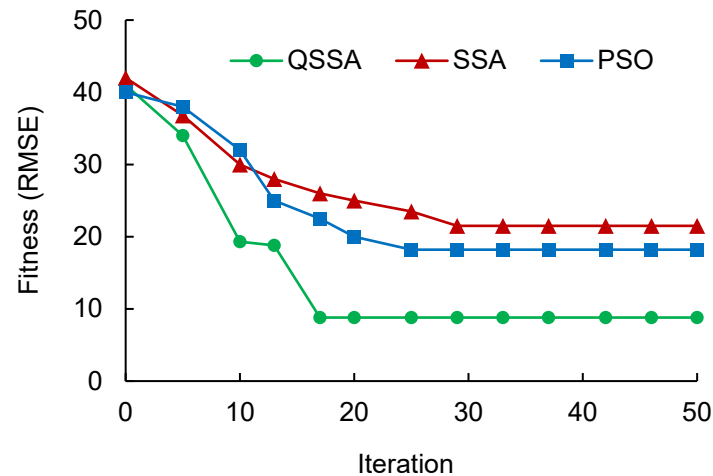
Optimized Parameters	Range	Optimization Results
Kernel size	[2, 7]	4
Number of filters	[12, 128]	64
dropout	[0, 0.5]	0.01
Learning rate	[0.001, 0.01]	0.001
Number of iterations	[100, 500]	300

To adapt the model to the target domain, QSSA was applied after transfer learning rather than only during source-domain pre-training. The transferred MHA-TCN layers provided the initial temporal feature representation, while QSSA searched the hyperparameters that were most sensitive to the target-domain production dynamics. The optimization objective was the validation-set RMSE of Block M. The searched parameters included the convolution kernel size, the number of filters, the dropout ratio, the learning rate, and the maximum number of training epochs. The search was constrained within a moderate range to avoid producing an excessively large model for the small target-domain dataset. Therefore, the target-domain adaptation strategy combined transferred feature extraction with local hyperparameter re-optimization.

To verify the performance advantage of the QSSA algorithm on non-convex optimization problems, while maintaining the consistency of the main network structure (TCN-BiLSTM), it is compared with the standard SSA and the PPSO algorithm. As shown in Table 3, the model optimized by QSSA achieves the minimum values in both RMSE and MAPE, which are 8.72 and 7.24%, respectively. Compared with SSA and PSO, QSSA demonstrates stronger global optimization capability and faster convergence speed, effectively preventing the algorithm from falling into local optimal solutions (Figure 5).

Table 3. Comparison of different optimization algorithms (Results from 10 independent runs).

Optimization Algorithm	MAPE/% (Mean $\pm$ SD)	RMSE (Mean $\pm$ SD)
QSSA	7.24 $\pm$ 0.58	8.72 $\pm$ 0.62
SSA	22.34 $\pm$ 2.15	21.25 $\pm$ 1.82
PSO	17.28 $\pm$ 1.45	18.21 $\pm$ 1.56

**Figure 5.** Convergence curves of different optimization algorithms.

3.3. Comparison of Different Algorithms

To further verify the effectiveness and advancement of the proposed QSSA-TCN-BiLSTM model in the tight oil production prediction task, four mainstream models—Back Propagation (BP) neural network, Gated Recurrent Unit (GRU), LSTM, and standard BiLSTM—were selected as the benchmark control groups in the experiment. Taking Well M1 and Well M2 in the Mahu block as examples, a rolling prediction of the daily oil production for the next 100 days was conducted for each model, respectively. The detailed prediction performance metrics of the different models are shown in Table 4.

Table 4. Comparison results of different baseline models.

	M1			M2			Training Time (s)
	RMSE	R ²	MAPE/%	RMSE	R ²	MAPE/%	
BP	23.82	0.4224	39.22	21.37	0.4524	38.32	45.2
GRU	20.92	0.5211	30.22	16.88	0.4211	33.22	158.4
LSTM	16.43	0.5012	22.82	17.22	0.5512	27.82	182.5
BiLSTM	12.34	0.6633	14.32	14.52	0.6633	14.32	245.6
Proposed model	8.22	0.9212	10.2	8.78	0.8912	7.21	118.5

From the statistical data in Table 4, it can be seen that the prediction accuracy of each model presents an obvious step-like difference. Due to the lack of a processing mechanism for the long-range dependencies of time series, the traditional BP neural network has the largest prediction errors on both wells, with RMSEs as high as 23.82 and 21.37, respectively, which is difficult to meet actual engineering requirements. In contrast, although the GRU and LSTM networks with gating mechanisms alleviate the gradient problem to a certain extent and improve the ability to capture temporal features, their MAPE still remains above 20%, indicating that single recurrent neural networks still have limitations when dealing with non-stationary and strongly nonlinear data like tight oil production. Benefiting from its bidirectional structure, the BiLSTM model can simultaneously utilize both historical and

future (backward) context information, and its prediction accuracy is significantly better than the three aforementioned models. However, the combined model proposed in this paper achieves the best results across all evaluation metrics, demonstrating a significant performance advantage. Taking Well M1 as an example, the RMSE of the proposed model decreases to 8.22, and the coefficient of determination (R^2) is as high as 0.9212. Compared with the best-performing baseline model, BiLSTM, the RMSE and MAPE are reduced by 33.4% and 28.8%, respectively. This result fully demonstrates that by introducing TCN for multi-scale feature extraction and QSSA for hyperparameter optimization, the model can more accurately excavate the deep nonlinear dynamic laws of the production data.

To evaluate the fitting ability of the models more intuitively, Figures 5 and 6 illustrate the comparison between the production prediction curves and the true values of different models on Well M1 and Well M2, respectively. Limitations of shallow and basic models: Because the BP neural network lacks temporal memory capability, its prediction error is the largest (the MAPE for Well M1 reaches 39.22%); although GRU and LSTM introduce gating mechanisms, there is still a certain degree of gradient vanishing when processing non-linear features of long sequences, leading to limited prediction accuracy.

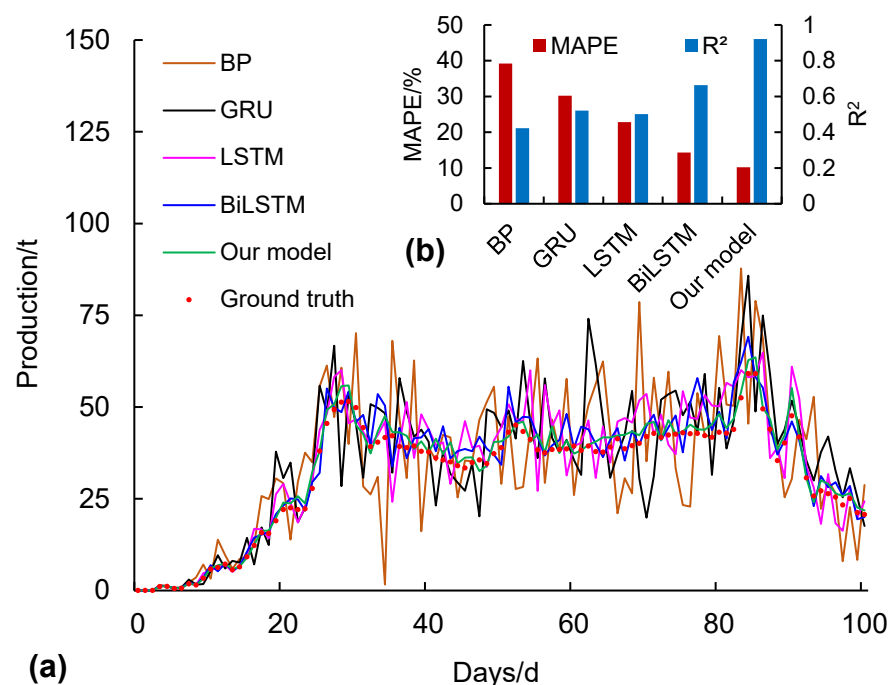


Figure 6. Comparison of production predictions for Well M1. (a) Comparison curves between actual production and predicted values from different algorithms; (b) Visualization results of MAPE and R^2 for different algorithms.

Advantages of the proposed model: Benefiting from the weighting of key time steps by the multi-head attention mechanism and the supplementation of small-sample features by transfer learning, the coefficient of determination (R^2) of the proposed model on Well M1 reaches 0.9212, and the MAPE is controlled within 10.20%. Figures 6 and 7 demonstrate the fitting of the prediction curves of different models against the true production on the two typical wells. It can be seen that the proposed model can not only accurately track the overall trend of production decline but also effectively capture local fluctuation features, exhibiting stronger robustness and generalization ability.

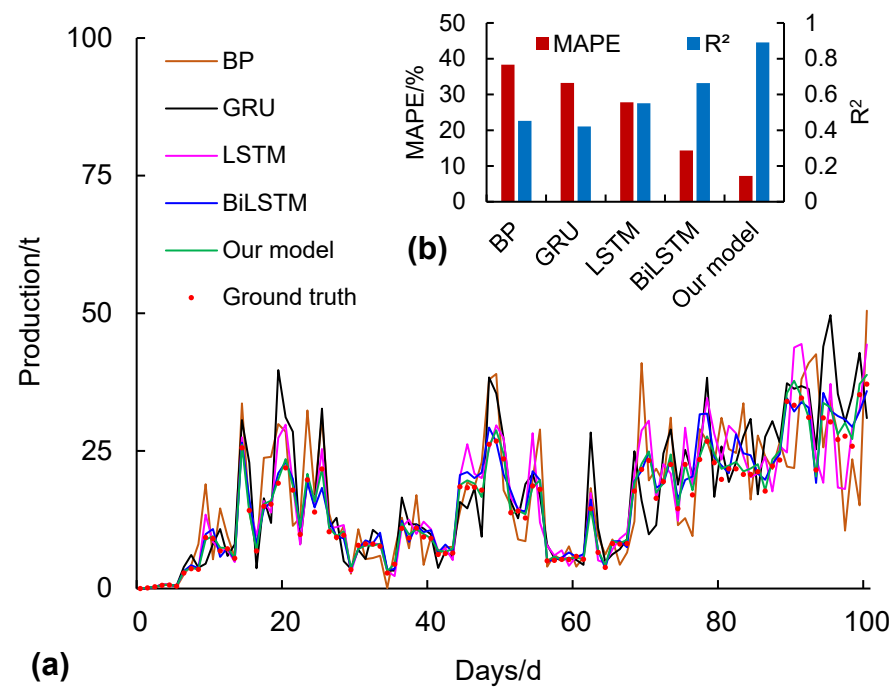


Figure 7. Comparison of production predictions for Well M2. (a) Comparison curves between actual production and predicted values from different algorithms; (b) Visualization results of MAPE and R^2 for different algorithms.

Judging from the morphology of the prediction curves, the prediction results of models such as BP and GRU exhibit obvious hysteresis where the production fluctuates violently, and the fitting deviation gradually increases in the later production decline stage, failing to accurately reflect the production characteristics of tight oil horizontal wells, which typically feature rapid initial decline and large fluctuations in the later stage. Although the standard BiLSTM improves in trend tracking, it is still too smooth in capturing local extremum points, resulting in insufficient detail restoration. On the contrary, the prediction curve of the proposed model has the highest degree of fit with the true production curve. It can not only accurately track the long-term decline trend of production but also maintain a high degree of coincidence when facing local abrupt changes and inflection points in the production curve. This excellent generalization ability is mainly attributed to the adaptive allocation of weights to key time steps by the MHA mechanism, as well as the effective supplementation of target domain sample features by the transfer learning strategy, thereby ensuring the model possesses robust prediction performance under different well conditions.

To verify the specific effectiveness of the best model, the predicted and actual production values of two application wells at typical time points were presented, as shown in Table 5.

Table 5. Comparison between predicted and actual production values at typical time points.

Time (day)	M1			M2		
	Actual (t/d)	Predicted (t/d)	Absolute Error	Actual (t/d)	Predicted (t/d)	Absolute Error
10	3.35	3.46	0.11	9.28	8.77	0.51
30	51.50	53.24	1.74	3.43	3.63	0.20
50	37.25	36.10	1.15	26.78	28.63	1.85
80	41.76	43.20	1.44	22.84	21.55	1.29
100	21.21	22.03	0.82	35.20	37.78	2.58

Table 5 shows that using the model proposed in the article, the prediction deviation of the two wells at typical time points remains within 10%, confirming its actual effectiveness.

3.4. Ablation Study

To systematically deconstruct and quantify the contribution of each core component in the proposed QSSA-TCN-BiLSTM model to the prediction accuracy, as well as their synergistic mechanisms, this paper designs a rigorous progressive ablation experiment. The experiment takes the standard BiLSTM network as the baseline model and observes the evolutionary trajectory of the evaluation metrics by gradually coupling each functional module, thereby verifying the rationality and necessity of the composite model architecture design.

As shown in Table 6, a single BiLSTM model exhibits certain limitations when processing time series with strong non-linearity and long-term dependency characteristics, such as tight oil production. Its RMSE is 22.32, the MAPE is as high as 33.72%, and the goodness of fit (R^2) is only 0.6413, making it difficult to meet high-precision prediction requirements. After introducing the temporal convolutional neural network to construct a hybrid architecture, benefiting from the effective capture of multi-scale local features by dilated convolutions under different receptive fields, the feature extraction capability of the model is significantly enhanced. The RMSE drops to 20.17, and the R^2 increases to 0.6923, which proves the important role of multi-scale spatiotemporal feature fusion in improving the model's robustness. On this basis, the multi-head self-attention mechanism is further embedded. By calculating the correlation matrix within the time series, the model acquires the ability to assign higher weights to key time steps, effectively strengthening the modeling of global dependencies and suppressing noise interference. Consequently, the RMSE further decreases to 19.12, and the MAPE is optimized to 21.82%. Finally, by introducing the improved Quantum Sparrow Search Algorithm for global adaptive optimization of key hyperparameters in the BiLSTM layer, the blindness of manual empirical parameter tuning and the defect of easily falling into local optima are overcome. The model performance achieves a qualitative leap: the RMSE is significantly reduced to 16.32, the MAPE converges to 13.22%, and R^2 reaches 0.8349. The above ablation experiment data clearly confirms that each improved module produces a significant positive gain, and the organic integration among modules constitutes a more robust prediction paradigm, laying a solid model foundation for the subsequent introduction of transfer learning strategies.

Table 6. Results of ablation experiments.

Model	TCN	MHA	QSSA	RMSE	MAPE/%	R^2
BiLSTM				22.32	33.72	0.6413
	✓			20.17	24.25	0.6923
	✓	✓		19.12	21.82	0.7246
	✓	✓	✓	16.32	13.22	0.8349

✓ denotes embedding of the corresponding modules: TCN, MHA, and QSSA.

3.5. Comparison with Alternative Small-Sample Methods

To further justify the rationality and superiority of transfer learning in addressing the small-sample dilemma of Mahu tight oil horizontal wells, three widely adopted small-sample modeling methods are introduced for controlled comparison based on the identical QSSA-TCN-BiLSTM network architecture and Block M target dataset:

- (1) Synthetic Minority Over-sampling Technique (SMOTE) time-series synthesis: Interpolation-based synthetic production samples are generated to expand the target-domain training set;

- (2) Time series augmentation: Three augmentation operations (random Gaussian noise injection, time stretching, sequence flipping) are applied to original production sequences to enlarge sample volume;
- (3) Bayesian BiLSTM: Bayesian probability distribution constraints are embedded into the BiLSTM layer to reduce overfitting via weight regularization, without cross-domain knowledge transfer.

All experiments share consistent data preprocessing, QSSA hyperparameter optimization and rolling prediction settings (100-day daily oil production forecast for M1/M2 wells). The unified prediction indicators of different small-sample strategies are listed in Table 7.

Table 7. Prediction performance of different small-sample processing strategies.

Strategy	RMSE	MAPE/%	R ²
Raw small-sample	16.32	13.22	0.8349
SMOTE data synthesis	13.75	11.06	0.8715
Time series augmentation	12.91	10.37	0.8842
Bayesian BiLSTM regularization	13.28	10.79	0.8786
Proposed transfer learning	8.72	7.24	0.9214

From Table 6, SMOTE and time series augmentation only create virtual samples relying on the limited production fluctuation rules of Block M itself; they fail to learn the general multi-stage decline characteristics of tight conglomerate reservoirs summarized from hundreds of mature wells in Block X. Bayesian BiLSTM merely imposes probabilistic weight constraints to suppress overfitting, lacking external geological-engineering prior knowledge supplementation. Although the three contrast methods slightly improve prediction accuracy compared with the raw small-sample model, their performance gaps with the transfer learning framework remain prominent: the RMSE of transfer learning is at least 32.4% lower than other schemes, and R² increases by more than 4.2%.

Mechanistically, the MMD between source Block X and target Block M is only 0.02137 (high distribution similarity), which provides reliable cross-domain transferable prior knowledge. Transfer learning pre-trains the multi-scale TCN feature extractor using 163 source wells, and migrates the learned universal production temporal feature extraction rules to the target block via layer freezing and gradual unfreezing fine-tuning. This strategy fundamentally breaks the information limitation of single-block scarce data, which cannot be realized by local data synthesis or single-model regularization methods. The above comparative results fully prove that transfer learning is the most effective technical path for small-sample tight oil production prediction in the Mahu study area.

3.6. Analysis of Transfer Effect

The Mahu block suffers from sample scarcity during the early development stage. Directly training a deep learning model on the target-domain data may lead to overfitting and poor generalization. Therefore, this study introduces transfer learning to use prior knowledge from Block X and improve production prediction in Block M. To verify the practical efficacy of this strategy, the experiment first quantified the data distribution distance between the source and target domains based on the Maximum Mean Discrepancy metric. The calculation results show an MMD value of 0.02137, where a value of less than 0.05 determines that the source and target domains are highly similar, indicating a high similarity between the two blocks in terms of geological features and production patterns, which satisfies the basic assumptions and applicable conditions of transfer learning.

On this basis, this study compared the prediction performance of the model before and after the introduction of the transfer strategy. Experimental results indicate that

when relying solely on the small-sample data of the target domain for training, even with the integrated QSSA-TCN-BiLSTM architecture, the model struggles to fully capture the complex nonlinear decline characteristics of tight oil due to the limited richness of training samples. Its RMSE is 16.32, MAPE is 13.22%, and the coefficient of determination (R^2) is only 0.8349.

In contrast, by utilizing the massive data from the source domain to pre-train the TCN network weights and transferring them to the target domain for fine-tuning, the model successfully achieves cross-domain sharing of prior knowledge. Combined with the comparative results of SMOTE, time series augmentation and Bayesian modeling in Section 3.5, it can be seen that only the cross-block transfer learning strategy achieves qualitative performance improvement under the target-domain small-sample condition. After introducing the transfer strategy, the RMSE drops sharply from 16.32 to 8.72. After introducing the transfer strategy, all evaluation metrics demonstrate a significant leap: the RMSE drops sharply from 16.32 to 8.72, a decrease of 46.5%; the MAPE converges from 13.22% to 7.24%; and the R^2 increases from 0.8349 to 0.9214. This significant performance gain strongly confirms the effectiveness of the transfer learning strategy: it not only optimizes the parameter space of the feature extractor through source domain data initialization—effectively avoiding the trap of local optima during the cold-start phase—but also significantly enhances the model's feature representation capability and robustness under the condition of few samples in the target domain. This provides a reliable technical path for accurate production prediction in the Mahu tight oil block.

4. Conclusions

- (1) The selection of input parameters directly affects the prediction accuracy of the model. Utilizing Pearson correlation analysis, five categories of main controlling factors influencing the tight oil production in the Mahu block were screened out in the following order of importance: oil layer thickness > formation pressure > permeability > porosity > oil saturation. This improves the accuracy of the model's prediction results.
- (2) By integrating the multi-head attention mechanism into the QSSA-TCN-BiLSTM model, a novel deep learning model for predicting tight oil production was established, which significantly improved the accuracy of production prediction. Compared with the baseline BiLSTM model, the RMSE decreased from 22.32 to 16.32, the MAPE decreased from 33.72% to 13.22%, and R^2 increased from 0.6413 to 0.8349.
- (3) To solve the potential issue of poor model training effectiveness caused by the limited number of samples in the target domain block, the production data from the experimental domain Block X was utilized to train the model weights, which were then saved and applied directly to Block M. After introducing transfer learning, the RMSE of the established complete model further decreased to 8.72, and the MAPE decreased to 7.24%. R^2 increased to 0.9214, demonstrating better prediction accuracy and providing technical support for the large-scale development of tight oil.
- (4) To overcome the small-sample and multi-scale temporal-dependency challenges in tight oil production forecasting, this study proposes a novel hybrid framework that synergistically combines transfer learning, a multi-head attention-enhanced temporal convolutional network, and a quantum-inspired sparrow search algorithm for BiLSTM hyperparameter optimization. Despite the demonstrated predictive gains, the present investigation is confined to tight conglomerate reservoirs in the Mahu area, and the model's applicability to carbonate or sandstone tight oil blocks remains untested. Furthermore, only static geological attributes and historical production series are employed as inputs, while real-time downhole pressure monitoring data are not

yet incorporated, and the prediction horizon is limited to a 100-day rolling forecast, leaving longer-term (e.g., one-year) productivity predictions unexplored.

- (5) Future work will couple multi-modal time-series data from downhole dynamic monitoring measurements, develop a transfer learning prediction framework applicable across different basins, establish a hybrid prediction approach integrating reservoir numerical simulation and deep learning algorithms, and further develop an online real-time optimization tool for fracturing parameter design.

Author Contributions: Conceptualization, Q.L., T.M., X.L. and Z.G.; Methodology, M.Z., Y.K. and X.S.; Formal analysis, Q.L., M.Z., T.M. and D.G.; Investigation, M.Z., Y.K., X.L. and X.S.; Resources, Q.L. and T.M.; Data curation, Q.L., M.Z. and T.M.; Writing—original draft, M.Z., Y.K., X.L. and X.S.; Writing—review & editing, Q.L., T.M., D.G. and Z.G.; Supervision, Q.L.; Project administration, Z.G.; Funding acquisition, Z.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Open Fund of State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation, grant number PLN2024-29.

Data Availability Statement: The datasets analyzed during the current study are not publicly available due to privacy agreements but are available from the corresponding author on reasonable request.

Conflicts of Interest: Author Qingying Lin, Minghai Zhang and Tiancong Mao were employed by the company Baikouquan Oil Production Plant, PetroChina Xinjiang Oilfield Company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

QSSA	Quantum sparrow search algorithm
TCN	Temporal convolutional network
BiLSTM	Bidirectional long short-term memory
MHA	Multi-head attention
CNN	Convolutional neural network
PSO	Particle swarm optimization
SSA	Sparrow search algorithm
MSSA	Multi-strategy sparrow search algorithm
SA	Simulated annealing
GA	Genetic algorithm
MMD	Maximum mean discrepancy
RKHS	Reproducing kernel Hilbert space
RMSE	Root mean square error
MAPE	Mean absolute percentage error
R ²	Coefficient of determination

References

1. Cortes-Cano, J.M.; Suarez-Dominguez, E.J.; Perez-Sanchez, J.F.; Lozano-Navarro, J.I.; Palacio-Perez, A. Clay-Sand Wettability Evaluation for Heavy Crude Oil Mobility. *Chem. Chem. Technol.* **2022**, *16*, 448–453. [\[CrossRef\]](#)
2. You, L.J.; Xie, Y.S.; Chen, Y.; Kang, Y.L.; Huang, C.; Zhai, H.J. Development directions of formation damage evaluation and fracturing fluids in tight and shale oil reservoirs. *Petroleum* **2025**, *11*, 125–142. [\[CrossRef\]](#)
3. Wang, Q.; Wang, Y.F.; Wang, R.; Zhao, J.Z.; Hu, Y.Q.; Zhao, J. Evolution law of stress induced by pressure depletion in fractured shale reservoirs: Implications for subsequent refracturing and infill well development. *Petroleum* **2025**, *11*, 71–83. [\[CrossRef\]](#)
4. Li, N.Y.; Zhu, S.J.; Li, Y.; Zhao, J.R.; Long, B.; Chen, F.; Wang, E.Z.; Feng, W.T.; Hu, Y.A.; Wang, S.B.; et al. Fracturing-flooding technology for low permeability reservoirs: A review. *Petroleum* **2024**, *10*, 202–215. [\[CrossRef\]](#)

5. Tang, H.Y.; Ruan, Q.; Zhang, L.H.; Hu, D.D.; Qin, J.H.; Zhao, Y.L.; Ye, Y.P. Intelligent optimization of fracturing stage and cluster parameters for tight oil reservoir. *Petroleum* **2025**, *11*, 56–70. [[CrossRef](#)]
6. Wang, H.J.; Qiao, L.; Lu, S.F.; Chen, F.W.; Fang, Z.X.; He, X.P.; Zhang, J.; He, T.H. A novel shale gas production prediction model based on machine learning and its application in optimization of multistage fractured horizontal wells. *Front. Earth Sci.* **2021**, *9*, 726537. [[CrossRef](#)]
7. Fan, Z.; Liu, X.S.; Wang, Z.Q.; Peng, C.; Wang, Y.W. A novel ensemble machine learning model for oil production prediction with two-stage data preprocessing. *Processes* **2024**, *12*, 587. [[CrossRef](#)]
8. Aliyev, S.; Al-Shafloot, T.; Aljawad, M.S.; Abdulraheem, A.; Elkatatny, S. A novel approach to modeling breakdown pressure dynamics using machine learning. *Petroleum* **2025**, *11*, 516–532. [[CrossRef](#)]
9. Zhang, X.R.; Yan, L.; Lai, F.P.; Wang, H.L. Production prediction for shale oil horizontal wells using convolutional neural network–bidirectional long short-term memory neural network model. *ACS Omega* **2025**, *11*, 2206–2220. [[CrossRef](#)] [[PubMed](#)]
10. Pan, Y.; Liu, X.W.; Tian, F.C.; Yang, L.Y.; Gou, X.T.; Jia, Y.P.; Wang, Q.; Zhang, Y.X. Research on shale oil well productivity prediction model based on CNN-BiGRU algorithm. *Energies* **2025**, *18*, 2523. [[CrossRef](#)]
11. Zhu, B.B.; Li, N.; Duan, Y.Q.; Li, G.F.; Liu, G.H.; Qu, F.J.; Long, C.J.; Wang, X.; Liao, Q.Z.; Li, G.S. Well-production forecasting using machine learning with feature selection and automatic hyperparameter optimization. *Energies* **2025**, *18*, 99. [[CrossRef](#)]
12. Li, Q.Y.; Guo, K.L.; Gao, X.C.; Zhang, S.S.; Jin, Y.H.; Liu, J.K. Productivity prediction model of tight oil reservoir based on particle swarm optimization–back propagation neural network. *Processes* **2024**, *12*, 1890. [[CrossRef](#)]
13. Tang, X.Y.; Feng, D.F.; Li, K.Q.; Liu, J.X.; Song, J.Y.; Sheng, V.S. An improved BPNN prediction method based on multi-strategy sparrow search algorithm. *Comput. Mater. Contin.* **2023**, *74*, 2789–2802. [[CrossRef](#)]
14. Qiu, H.; Hu, R.; Chen, J.Q.; Yuan, Z.H. Short-Term electricity load forecasting based on complete ensemble empirical mode decomposition with adaptive noise and improved sparrow search algorithm–convolutional neural network–bidirectional long short-term memory model. *Mathematics* **2025**, *13*, 813. [[CrossRef](#)]
15. He, D.H.; Qu, Y.G.; Sheng, G.L.; Wang, B.; Yan, X.; Tao, Z.; Lei, M. Oil production rate forecasting by SA-LSTM model in tight reservoirs. *Lithosphere* **2024**, *2024*, lithosphere_2023_197. [[CrossRef](#)]
16. Zhang, L.; Dou, H.G.; Wang, T.Z.; Wang, H.L.; Peng, Y.; Zhang, J.F.; Liu, Z.S.; Mi, L.; Jiang, L.W. A production prediction method of single well in water flooding oilfield based on integrated temporal convolutional network model. *Petrol. Explor. Dev.* **2022**, *49*, 1150–1160. [[CrossRef](#)]
17. Pan, S.W.; Yang, B.; Wang, S.K.; Guo, Z.; Wang, L.; Liu, J.H.; Wu, S.Y. Oil well production prediction based on CNN-LSTM model with self-attention mechanism. *Energy* **2023**, *284*, 128701. [[CrossRef](#)]
18. Meng, X.; Liu, X.Y.; Duan, H.C.; Hu, Z.; Wang, M. Optimization of oil well production prediction model based on inter-attention and BiLSTM. *Electronics* **2025**, *14*, 1004. [[CrossRef](#)]
19. Li, F.G.; Ma, G.W.; Ju, C.Q.; Chen, S.J.; Huang, W.B. Data-driven forecasting framework for daily reservoir inflow time series considering the flood peaks based on multi-head attention mechanism. *J. Hydrol.* **2024**, *645*, 132197. [[CrossRef](#)]
20. Sun, S.S.; Bai, W.H.; Shao, Z.Y.; Shi, Z.S.; Zhou, T.Q.; Yu, Y.J.; Sun, J.; Peng, L.S.; Zhang, L. Exploring the predictive performance of deep learning for fracturing fluid flowback and shale gas production. *Sci. Rep.* **2025**, *15*, 42748. [[CrossRef](#)] [[PubMed](#)]
21. Kazemi, A.; Esmaili, M. Reservoir surrogate modeling using U-Net with vision transformer and time embedding. *Processes* **2025**, *13*, 958. [[CrossRef](#)]
22. Desbordes, J.K.; Zhang, K.; Xue, X.M.; Ma, X.P.; Luo, Q.; Huang, Z.Q.; Hai, S.; Jun, Y. Dynamic production optimization based on transfer learning algorithms. *J. Petrol. Sci. Eng.* **2022**, *208*, 109278. [[CrossRef](#)]
23. Qian, Q.J.; Yu, J.H.; Zhan, H.F.; Wang, R. A novel DL-BiGRU multi-feature fusion and deep transfer learning based modeling approach for quality prediction of injection molded products using small-sample datasets. *J. Petrol. Sci. Eng.* **2024**, *120*, 272–285. [[CrossRef](#)]
24. Niu, W.T.; Sun, Y.P.; Zhang, X.W.; Lu, J.L.; Liu, Q.J.; Li, Q.J.; Mu, Y. An ensemble transfer learning strategy for production prediction of shale gas wells. *Energy* **2023**, *275*, 127443. [[CrossRef](#)]
25. Dong, Y.K.; Zhang, Y.; Liu, F.B.; Cheng, X.T. Reservoir production prediction model based on a stacked LSTM network and transfer learning. *Acs Omega* **2021**, *6*, 34700–34711. [[CrossRef](#)] [[PubMed](#)]
26. Al-Alimi, D.; Al-Qaness, M.A.A.; Damaševičius, R. A hybrid transfer learning framework for enhanced oil production time series forecasting. *Comput. Mater. Contin.* **2025**, *82*, 3539–3561. [[CrossRef](#)]
27. Wang, H.C.; Zhang, K.; Chen, N.C.; Zhou, W.S.; Liu, C.; Wang, J.F.; Zhang, L.M.; Yu, Z.G.; Cui, S.T.; Yang, M.C. Better use of experience from other reservoirs for accurate production forecasting by learn-to-learn method. *Petrol. Sci.* **2024**, *21*, 716–728. [[CrossRef](#)]
28. Ning, F.; Lai, F.P. Patch-transformer modeling for production forecasting in tight gas reservoirs: A case study of the Yan'an Field. *Energy Fuels* **2025**, *39*, 432–443. [[CrossRef](#)]

29. Zou, C.N.; Yang, Z.; Cui, J.W.; Zhu, R.K.; Hou, L.H.; Tao, S.Z.; Yuan, X.J.; Wu, S.T.; Lin, S.H.; Wang, L. Formation mechanism, geological characteristics and development strategy of nonmarine shale oil in China. *Petrol. Explor. Dev.* **2013**, *40*, 15–27. [[CrossRef](#)]
30. Song, X.Y.; Liu, Y.T.; Xue, L.; Wang, J.; Zhang, J.Z.; Wang, J.Q.; Jiang, L.; Cheng, Z.Y. Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model. *J. Petrol. Sci. Eng.* **2020**, *186*, 106682. [[CrossRef](#)]
31. Xue, J.K.; Shen, B. A novel swarm intelligence optimization approach: Sparrow search algorithm. *Syst. Sci. Control Eng.* **2020**, *8*, 22–34. [[CrossRef](#)]
32. Ma, X.L.; Zhang, L.; Zhan, J.; Chang, S.L. Application of dual-stage attention temporal convolutional networks in gas well production prediction. *Mathematics* **2024**, *12*, 3896. [[CrossRef](#)]
33. Wang, H.L.; Mu, L.X.; Shi, F.G.; Dou, H.G. Production prediction at ultra-high water cut stage via Recurrent Neural Network. *Petrol. Explor. Dev.* **2020**, *47*, 1084–1090. [[CrossRef](#)]
34. Ahmadi, M.A.; Ebadi, M.; Shokrollahi, A.; Majidi, S.M.J. Evolving artificial neural network and imperialist competitive algorithm for prediction oil flow rate of the reservoir. *Appl. Soft Comput.* **2013**, *13*, 1085–1098. [[CrossRef](#)]
35. Sagheer, A.; Kotb, M. Time series forecasting of petroleum production using deep LSTM recurrent networks. *Neurocomputing* **2019**, *323*, 203–213. [[CrossRef](#)]
36. Kukharova, T.; Martirosyan, A.; Asadulagi, M.A.; Ilyushin, Y. Development of the separation column's temperature field monitoring system. *Energies* **2024**, *17*, 5175. [[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.