Article

Top-Oil Temperature Prediction of Power Transformer Based on Long Short-Term Memory Neural Network with Self-Attention Mechanism Optimized by Improved Whale Optimization Algorithm

Dexu Zou [1,2], **He Xu** [3], **Hao Quan** [3,*], **Jianhua Yin** [3], **Qingjun Peng** [2], **Shan Wang** [2], **Weiju Dai** [2] **and Zhihu Hong** [2]

1   School of Electrical Engineering, Chongqing University, Chongqing 400044, China; zoudexu@csg.cn
2   Electric Power Research Institute, China Southern Power Grid Yunnan Power Grid Co., Ltd.,
     Kunming 650217, China; penggingjun@csg.cn (Q.P.); wangshan@csg.cn (S.W.); daiweiju@csg.cn (W.D.);
     hongzhihu@csg.cn (Z.H.)
3   School of Automation, Nanjing University of Science and Technology, Nanjing 210094, China;
     xuhe1@njust.edu.cn (H.X.); yinjhhust@163.com (J.Y.)
*   Correspondence: quanhao@njust.edu.cn

Abstract: The operational stability of the power transformer is essential for maintaining the symmetry, balance, and security of power systems. Once the power transformer fails, it will lead to heightened instability within grid operations. Accurate prediction of oil temperature is crucial for efficient transformer operation. To address challenges such as the difficulty in selecting model hyperparameters and incomplete consideration of temporal information in transformer oil temperature prediction, a novel model is constructed based on the improved whale optimization algorithm (IWOA) and long short-term memory (LSTM) neural network with self-attention (SA) mechanism. To incorporate holistic and local information, the SA is integrated with the LSTM model. Furthermore, the IWOA is employed in the optimization of the hyper-parameters for the LSTM-SA model. The standard IWOA is improved by incorporating adaptive parameters, thresholds, and a Latin hypercube sampling initialization strategy. The proposed method was applied and tested using real operational data from two transformers within a practical power grid. The results of the single-step prediction experiments demonstrate that the proposed method significantly improves the accuracy of oil temperature prediction for power transformers, with enhancements ranging from 1.06% to 18.85% compared to benchmark models. Additionally, the proposed model performs effectively across various prediction steps, consistently outperforming benchmark models.

Keywords: power transformer; top-oil temperature prediction; self-attention mechanism; whale optimization algorithm; long short-term memory networks

1. Introduction

Power transformers undertake a vital role in the symmetrical operation of power systems [1]. They serve as critical infrastructure for power transmission and distribution, with extensive applications in various other fields, such as transportation [2]. Once the power transformer fails, it can severely disrupt the normality of the power system operation, potentially causing widespread power outages and significant economic losses [3]. As a vital component of the power system, the stable operation of the transformer is fundamental to maintaining the symmetry and balance of the power system [4,5].

Top oil temperature is significant for determining whether the transformer can maintain normal operation. In practice, the transformer internal faults rely on the trend of the oil temperature to make judgments [6,7]. Therefore, the good performance of oil temperature prediction helps professionals find problems promptly in the transformer's daily operation and maintenance. By reliably forecasting oil temperature, we can not only prevent

unexpected failures but also optimize maintenance schedules, reduce operational risks, and extend the transformer's lifespan. Effective oil temperature prediction enhances the overall reliability and efficiency of the power system, making it an essential component in maintaining the symmetrical operation of the electrical grid.

Researchers generally study the prediction of transformer oil temperatures through mathematical and data-driven models [8–10]. Zhao et al. used the least squares method to establish a parameter identification algorithm [11], and this mathematical model can effectively predict the top oil temperature but lacks strong generalization ability. Wang et al. establish a thermal circuit model to simulate the changes in the transformer temperature over time, but it has a lengthy computation time [12].

With the development of intelligent algorithms, artificial intelligence technologies have been applied to the field of power system forecasting. Interesting studies can be found in the fields of load forecasting [13], vehicle-to-grid (V2G) scheduling prediction [14], and solar irradiance forecasting [15]. There have been some research efforts focused on predicting transformer oil temperature using these algorithms. Qing et al. developed a model based on artificial neural networks for forecasting the top oil temperature of transformers [16], and this model significantly reduces the computational time but ignores the selection of optimal hyperparameters. Tan et al. proposed a forecast model that considers path analysis and similar moments [17], but the validation dataset is small and the adaptability is difficult to confirmed. Li et al. introduced a regression model with enhanced particle swarm optimization (PSO) for transformer top oil temperature forecast [18]. However, the large sampling interval of data caused the substandard performance. Based on a similar day, Tan et al. introduced a method to predict top oil temperature. The above approach relies solely on single-day similarity for prediction and deteriorates the model prediction performance [19]. To sum up, these studies do not fully consider temporal information of different input features, thus failing to combine global and local information within transformer operational data. In addition, the optimal hyper-parameters of the model are difficult to determine.

To tackle the issues mentioned, this paper introduces a novel method: an improved whale optimization algorithm (IWOA) optimized long short-term memory (LSTM) neural network with self-attention (SA) mechanism model. The proposed method comprehensively addresses challenges related to the difficulty in selecting hyperparameters for the oil temperature prediction model and the insufficient consideration of temporal information. It integrates SA with LSTM and utilizes the IWOA to obtain the optimal hyper-parameters for the LSTM-SA model, resulting in high prediction accuracy. Finally, the proposed method is tested with actual operating data in a practical power grid. The results demonstrate that the proposed method has better forecasting performance.

The remaining sections of this paper are as below: Section 2 discusses the power transformer and top-oil temperature. Section 3 introduces the LSTM-SA model and the IWOA. Section 4 presents a case study that shows the superiority of the IWOA for optimization and the effectiveness of the proposed method for predicting top-oil temperature. Finally, conclusions and discussions are presented in Section 5.

2. Power Transformer and Top-Oil Temperature

The top oil temperature of a transformer is a crucial indicator for measuring the reliability of transformer operation, monitoring the internal insulation status. Accurately predicting the top oil temperature of the power transformer is of great significance for analyzing potential faults, carrying out transformer operation and maintenance, maintaining the symmetry and balance of the power system, and achieving early warning of transformer failures. It is a key factor in limiting the transformer's load capacity and assessing its operational lifespan.

There are two merits to considering top oil temperature as the subject of study. First, researchers can easily access real-time monitoring data for the transformer's top oil temperature, thanks to advanced sensor technologies and the widespread implementation

of smart grids. This accessibility facilitates continuous monitoring and data collection, which are essential for accurate prediction and timely intervention. Second, the hot spot temperature that is difficult to obtain can be calculated from the transformer top oil temperature. Hot spot temperature is crucial, as it represents the highest temperature within the transformer and is a direct indicator of the condition of the transformer's insulation. Accurate estimation of this temperature is vital for predicting the remaining life of the insulation and planning maintenance activities.

The above advantages have made the top oil temperature highly favored by researchers, and it has now become a hot research topic [20]. The basic construction of an oil-immersed transformer is graphically represented in Figure 1. This paper focuses on improving the accuracy of oil temperature prediction, particularly in addressing the challenges posed by the nonlinearity and time-series characteristics of the data.
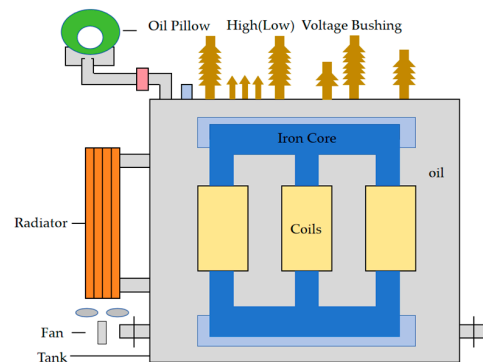

Figure 1. The basic construction of an oil-immersed transformer.

3. The Proposed IWOA-LSTM-SA Method for Top-Oil Temperature Prediction

3.1. Framework

In this study, IWOA-LSTM-SA has been developed for transformer oil temperature forecasting, in which IWOA has been employed to precisely search optimal input hyperparameters and LSTM-SA as the forecasting model to combine global and local information. The flowchart is presented in Figure 2.
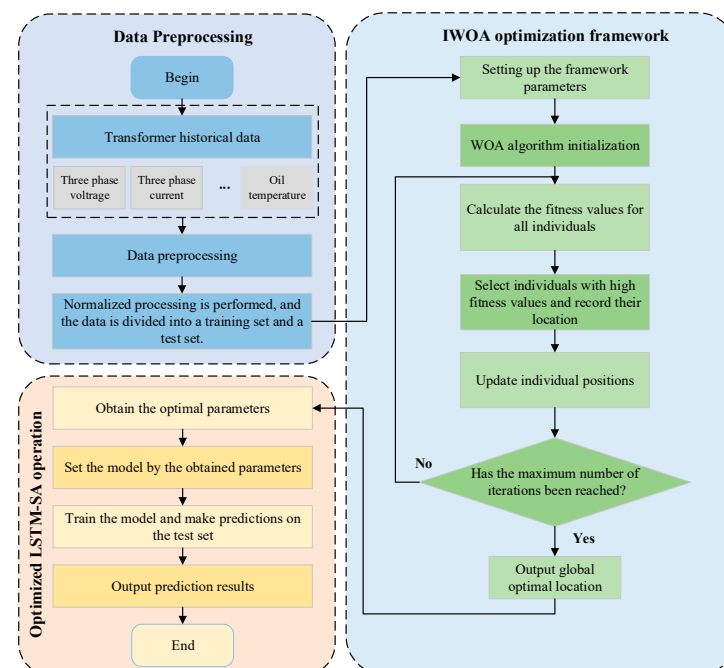

Figure 2. Flow chart of IWOA-LSTM-SA.

The main phases of the IWOA-LSTM-SA will be detailed in the following sections.

3.2. LSTM Integrated by SA

LSTM is a specialized type of recurrent neural network (RNN), specifically designed to process temporal data sequences. On the basis of traditional RNN, LSTM introduces the concept of "gating", which not only overcomes the gradient vanishing but also selects samples. Therefore, LSTM is more suitable for solving nonlinear temporal structure problems. Each memory block of an LSTM comprises one or more self-connected memory cells and three gating units: the input gate, the output gate, and the forget gate. The specific structure of the gate is shown in Figure 3. The forgetting gate is responsible for deciding which information should be discarded from the cell state, effectively determining the extent to which the previous cell state is preserved within the current cell state. The calculation equation is as below:

$$m_t = \sigma(W_m \times [r_{t-1}, x_t] + p_m) \tag{1}$$

The input gate controls which the current input is stored in the unit state. The formulas for input gates and candidate cell states is as below:

$$s_t = \sigma(W_s \times [r_{t-1}, x_t] + p_s) \tag{2}$$

The output gate regulates the current output and decides the output information. The formula for calculation is given below:

$$g_t = \sigma(W_g \cdot [r_{t-1}, x_t] + p_g) \tag{3}$$

$$r_t = o_t \cdot \tanh(C_t) \tag{4}$$

The formula for calculating the cell state is as below:

$$\widetilde{C}_t = \tanh(W_C \times [r_{t-1}, x_t] + p_C) \tag{5}$$

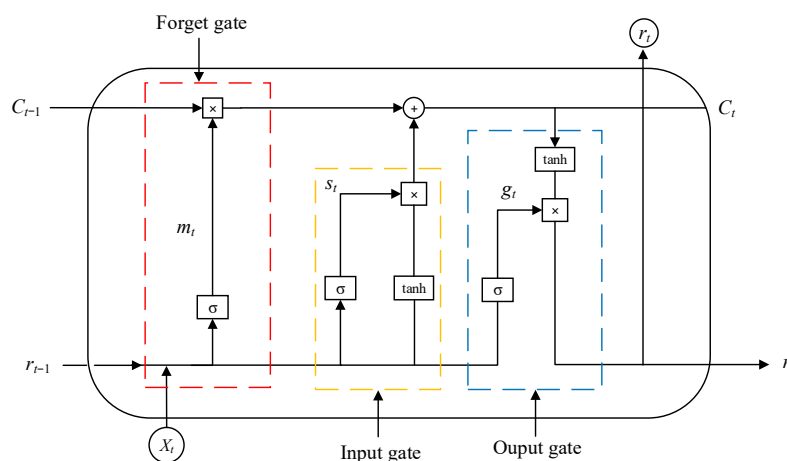$$C_t = m_t \cdot C_{t-1} + s_t \cdot \widetilde{C}_t \tag{6}$$

Figure 3. LSTM structure diagram.

In summary, LSTM is suitable for processing time series data, so this paper uses LSTM to establish a temperature prediction model. Furthermore, it is difficult to process long sequence data for the LSTM model that we introduce SA to solve this problem. This method considers both local and global information.

It consists of three components. Firstly, the data that come from the LSTM model is the input of the SA layer. Secondly, the matrices q, k, and v are calculated using the weight matrices W_q, W_k, and W_v. Thirdly, $a^{1,2}$ is the dot product between q_1 and k_2, and $a^{2,2}$ is the

dot product between q_2 and k_2. The attention matrix M means the correlation between different time steps. The structure is shown in Figure 4.
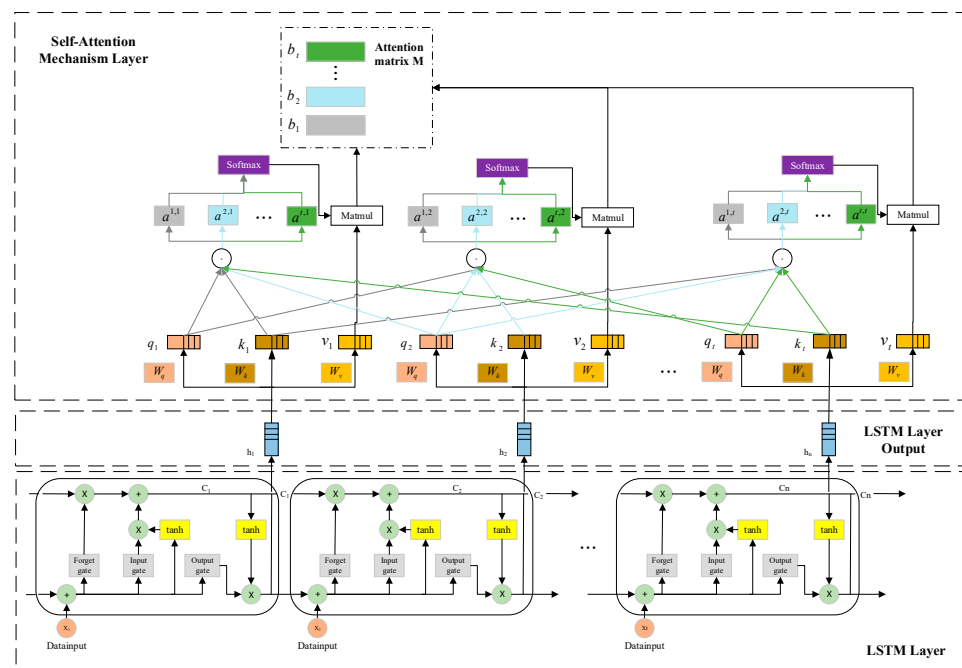

Figure 4. LSTM-SA structure.

3.3. Hyper-Parameters Optimization by IWOA

The Whale Optimization Algorithm (WOA) was introduced to deal intricate optimization problems by Mirjalili et al. [21,22]. The WOA can be formulated as the following steps: encircling prey, bubble-net attacking method and search for prey.

3.3.1. Encircling Prey

Humpback whales can identify and encircle their prey. In the population, the remaining whales will try to adjust their positions towards the direction of the best search agent as defined by the equation:

$$\vec{G}(t+1) = \vec{G}^*(t) - \vec{A}\left|\vec{C}\vec{G}^*(t) - \vec{G}(t)\right| \tag{7}$$

where t denotes the current iteration; $\vec{G}$ is a vector indicating the position; $\vec{G}^*$ is the place vector of the best solution acquired yet, $\vec{A}$ and $\vec{C}$ are calculated from the following:

$$\vec{A} = 2\vec{a}\,\vec{r}_1 - \vec{a} \tag{8}$$

$$\vec{C} = 2\vec{r}_2 \tag{9}$$

where $\vec{a}$ is an adjustment vector and $\vec{a}$ is linearly decreasing from 2 to 0; the vectors $\vec{r}_1$ and $\vec{r}_2$ are random vectors that fall within the range of [0, 1].

3.3.2. Bubble-Net Attacking Method

Humpback whale predation consists of two main mechanisms: shrinkage bracketing mechanism and the spiral updating location.

(1) Shrinkage bracketing mechanism: As $\vec{a}$ decreases, $\vec{A}$ represents an any value within the range of [−1, 1]. The new position is determined by the distance between its

original position and the position of the currently best-so-far whale. The equation for calculation is as below:

$$\vec{a} = 2 \times \left(1 - \frac{t}{t_{\max}}\right) \tag{10}$$

(2) Spiral updating location: the WOA uses spiral updating location to launch attacks on prey, and the spiral hunting equation is as below:

$$\vec{G}(t+1) = e^{bl}\cos(2\pi l) \cdot \left|\vec{G}^{*}(t) - \vec{G}(t)\right| + \vec{G}^{*}(t) \tag{11}$$

where l is a random count within the interval $[-1, 1]$ and b represents a constant. They approach the prey using two mechanisms: a shrinking circle and a spiral-shaped path. The updated equations are as follows.

$$\vec{G}(t+1) = \begin{cases} \vec{G}^{*}(t) - \vec{A}\left|\vec{C} \cdot \vec{G}^{*}(t) - \vec{G}(t)\right|, p < 0.5 \\ e^{bl}\cos(2\pi l) \cdot \left|\vec{G}^{*}(t) - \vec{G}(t)\right| + \vec{G}^{*}(t), p \geq 0.5 \end{cases} \tag{12}$$

where p falls within the range of [0,1].

3.3.3. Search for Prey

Humpback whales search for their prey randomly, with their locations varying relative to each other. In this stage, the position of a searching whale is modified according to the position of a randomly selected whale, as opposed to being updated based on the current best whale. The calculation formula is as listed below:

$$\vec{G}(t+1) = \vec{G}_{\text{rand}}(t) - \vec{A} \cdot \left|\vec{C}\vec{G}_{\text{rand}}(t) - \vec{G}(t)\right| \tag{13}$$

where $\vec{G}_{\text{rand}}$ denotes the random location of a whale.

3.3.4. Improved Whale Optimization Algorithm

The original WOA faces certain limitations, particularly in terms of inadequate local search capabilities and insufficient population diversity. Therefore, it is necessary to further improve the strategy and adjust the algorithm [23]. For example, Naderi et al. proposed a Whale Optimization Algorithm enhanced by wavelet mutation, aimed at improving the algorithm's convergence characteristics to address the complex trade-off between generation costs and water consumption [24]. In this study, an approach takes a different direction by introducing three key improvements: Latin Hypercube Sampling for more diverse and uniform population initialization, an adaptive selection threshold to dynamically adjust the whale's movement strategy, and a nonlinear parameter adjustment to enhance local search capabilities. These modifications are designed to address different aspects of the original WOA's limitations. The specific improvements are as follows:

(1) Latin Hypercube Sampling (LHS) initialization of population: as stated in [25], population initialization plays a crucial role in swarm intelligence optimization algorithms. In WOA, population initialization follows a random approach. However, it can lead to uneven population distribution and individual overlap [26]. Therefore, it is necessary to optimize the population initialization. IWOA incorporates LHS to increase the diversity of initial population, and this method can initialize population more uniformly and efficiently.

(2) Adaptive selection threshold: in WOA, the whales choose either encircling activity or spiral movement with 50% probability. However, this method prevents the whale population from choosing the appropriate movement for the current population [27,28]. In

this paper, an adaptive selection threshold is used to replace the fixed threshold. The method automatically adjusts the threshold according to the problem's characteristics throughout the search process. The calculation is given by the following formula:

$$p_a = 1 - \left[\frac{t}{(L+f)t_{\max}} \times \left(L \times \frac{e^t}{e^{t_{\max}}} + f \times \frac{t^f}{t_{\max}{}^f} \right) \right] \tag{14}$$

where t denotes the current iteration, while $t_{\max}$ denotes the maximum iteration count; L, f are control parameters, and their values are 2 and 4, respectively.

In our method, when the threshold is larger in the initial stage, the whale will preferentially choose the encircling movement strategy. With the increasing of iterations, the threshold decreases, thus the whale is more likely to choose the spiral motion strategy. Equation (12) is updated to Equation (15).

$$\vec{G}(t+1) = \begin{cases} \vec{G}^*(t) - \vec{A} \left| \vec{C} \cdot \vec{G}^*(t) - \vec{G}(t) \right|, p < p_a \\ e^{bl} \cos(2\pi l) \cdot \left| \vec{G}^*(t) - \vec{G}(t) \right| + \vec{G}^*(t), p \geq p_a \end{cases} \tag{15}$$

(3) Adaptive parameter: in traditional method, $\vec{a}$ decreases linearly from 2 to 0. In order to enhances local searching ability, this study uses a nonlinear strategy to adjust b in Equation (16), which influences the shape of the logarithmic spiral. It can significantly improve the effectiveness of local search and the speed of global search, thereby enhancing overall accuracy [29]. At the same time, we establish a relationship between b and t to achieve adaptive adjustment. Equation (10) is updated to Equation (16).

$$\begin{cases} \vec{a}(t) = 2 \times \left(1 - \tanh\left(\sqrt[k]{\frac{t}{t_{max}}} \right) \right) \\ b(t) = v - \left(\frac{v}{t_{max}} \right) \times t \end{cases} \tag{16}$$

where k, v are control parameters, and their values are 4 and 10, respectively.
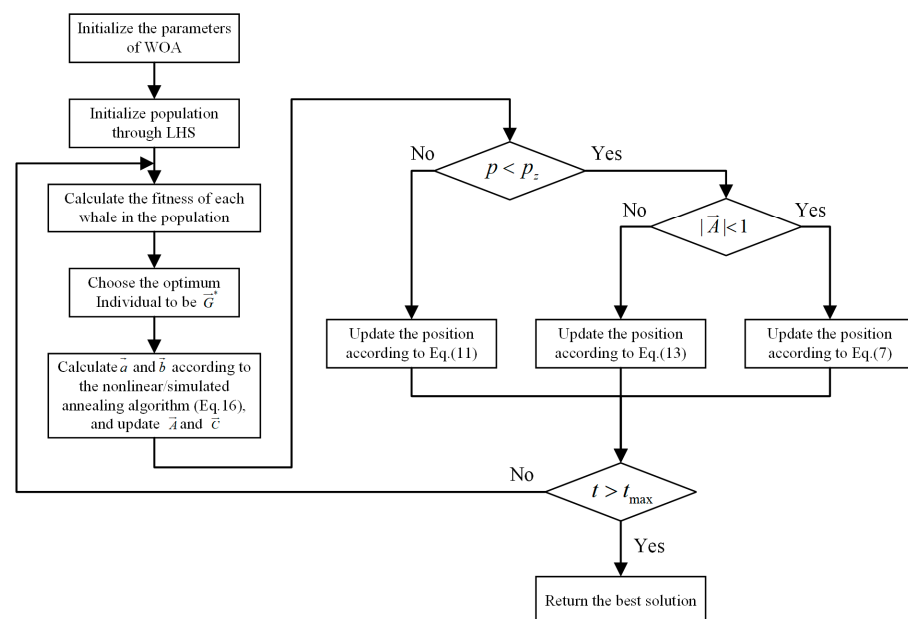
The IWOA flowchart is illustrated in Figure 5.

Figure 5. Flow chart of the IWOA.

4. Case Studies and Results Analysis

4.1. Data Source

This study includes two datasets. Dataset 1 consists of transformer operation data collected from a 500 kV substation from 1 April to 30 June in 2022, with a sampling period of half an hour. In total, there are 4368 samples. The characteristic parameters include high-voltage-side three-phase current (A_I, B_I, C_I), active and reactive power (P, Q), high-voltage-side three-phase voltage (A_U, B_U, C_U), and top-oil temperature (T). This paper used the Pearson correlation coefficient method to select features, and the results are shown in Table 1. Dataset 2 consists of transformer operation data collected from a 220 kV substation from 10 February 2021 to 10 February 2022, with a sampling period of half an hour. In total, there are 17,518 samples.

Table 1. Correlation matrix.

	A_I	B_I	C_I	P	Q	A_U	B_U	C_U	T
A_I	1.000	0.999	0.999	0.999	0.925	−0.862	−0.866	−0.835	0.371
B_I	0.999	1.000	0.999	0.999	0.924	−0.863	−0.866	−0.835	0.371
C_I	0.999	0.999	1.000	0.999	0.925	−0.862	−0.866	−0.835	0.371
P	0.999	0.999	0.999	1.000	0.925	−0.857	−0.859	−0.828	0.369
Q	0.925	0.924	0.925	0.925	1.000	−0.842	−0.844	−0.823	0.372
A_U	−0.862	−0.863	−0.862	−0.857	−0.842	1.000	0.979	0.964	−0.346
B_U	−0.866	−0.866	−0.866	−0.859	−0.844	0.979	1.000	0.981	−0.342
C_U	−0.835	−0.835	−0.835	−0.828	−0.823	0.964	0.981	1.000	−0.339
T	0.371	0.371	0.371	0.369	0.372	−0.346	−0.342	−0.339	1.000

As shown in Table 1, the correlation coefficient between the top-oil temperature and the high-voltage side three-phase current is 0.371, and the correlation coefficients with active power and reactive power are 0.369 and 0.372, respectively, indicating a positive correlation. The correlation coefficients between the top-oil temperature and the high-voltage side three-phase voltage are −0.346, −0.342, and −0.339, respectively, indicating a negative correlation with the top-oil temperature. This also suggests that the high-voltage side three-phase voltage, current, and active and reactive power have some influence on the transformer oil temperature. Similarly, a correlation analysis of the input features of Dataset 2 based on the Pearson correlation coefficient method is conducted. Ultimately, this paper selects high-voltage-side current, active and reactive power, voltage, and top-oil temperature as input features. The dataset is split into training and test sets, in which 80% is used for training and 20% for testing.

4.2. Comparison of Algorithm Optimization Results

This paper compared the performance of IWOA with traditional methods, which consist of GA, PSO, and the original WOA. Appendix A, Table A1 presents the ten test functions employed for evaluation, which are derived from the studies conducted in [30,31].

In Appendix A, Table A1: Each function has a dimension of 30, and the minimum value is 0. To ensure the fairness of the comparison, the iteration is set to 500. The crossover probability of GA is set to 1, and the variance probability is 0.1. Meanwhile, the learning factor c1 = c2 = 2 for PSO, and b is 10 for WOA. Each algorithm runs independently 30 times. The average and the best results are utilized for comparison, as shown in Table 2. The average convergence curve of each algorithm is shown in Figure 6.

In Table 2, the optimal value reaches 0 in the F_5, F_6 and F_8 functions, and the average values also show significant improvement. As shown in Figure 6, IWOA exhibits better convergence performance compared to traditional algorithms. These findings confirm the effectiveness of the enhancement strategies for WOA.

Table 2. Comparison of test results for each algorithm.

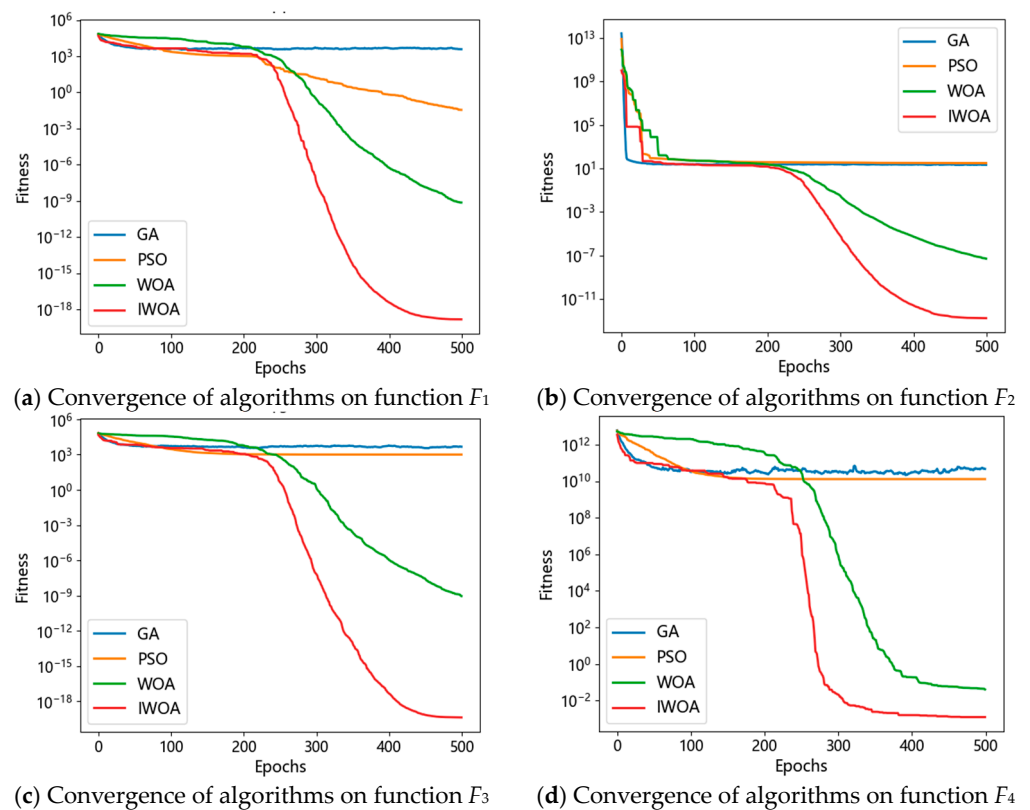| Function	Evaluation Index	GA	PSO	WOA	IWOA
F_1	Mean	3602.311	0.035	7.21×10^{-10}	1.46×10^{-19}
	Best	1454.955	0.001	3.32×10^{-13}	1.17×10^{-24}
F_2	Mean	21.197	32.013	5.16×10^{-9}	1.73×10^{-13}
	Best	13.936	0.081	5.12×10^{-9}	2.24×10^{-15}
F_3	Mean	3477.958	0.047	8.98×10^{-10}	4.16×10^{-20}
	Best	1771.241	0.001	1.68×10^{-12}	1.42×10^{-22}
F_4	Mean	1.432	5.176	0.015	0.00075
	Best	0.413	0.065	0.003	0.00014
F_5	Mean	28.474	51.152	0	0
	Best	5.522	0	0	0
F_6	Mean	91.831	127.257	0.462	1.78×10^{-16}
	Best	64.795	69.170	6.78×10^{-11}	0
F_7	Mean	11.337	2.028	3.936	1.49×10^{-11}
	Best	9.197	0.023	8.06×10^{-7}	1.35×10^{-12}
F_8	Mean	77.000	551.976	0.988	0
	Best	35.494	185.625	0	0
F_9	Mean	75.910	727.867	−0.898	−0.829
	Best	28.593	479.302	−0.967	−0.986
F_{10}	Mean	73.449	596.665	−0.890	−0.796
	Best	26.910	332.989	−0.980	−0.899

(**a**) Convergence of algorithms on function F_1

(**b**) Convergence of algorithms on function F_2

(**c**) Convergence of algorithms on function F_3

(**d**) Convergence of algorithms on function F_4

Figure 6. *Cont.*

(**e**) Convergence of algorithms on function F_5

(**f**) Convergence of algorithms on function F_6

(**g**) Convergence of algorithms on function F_7

(**h**) Convergence of algorithms on function F_8

(**i**) Convergence of algorithms on function F_9
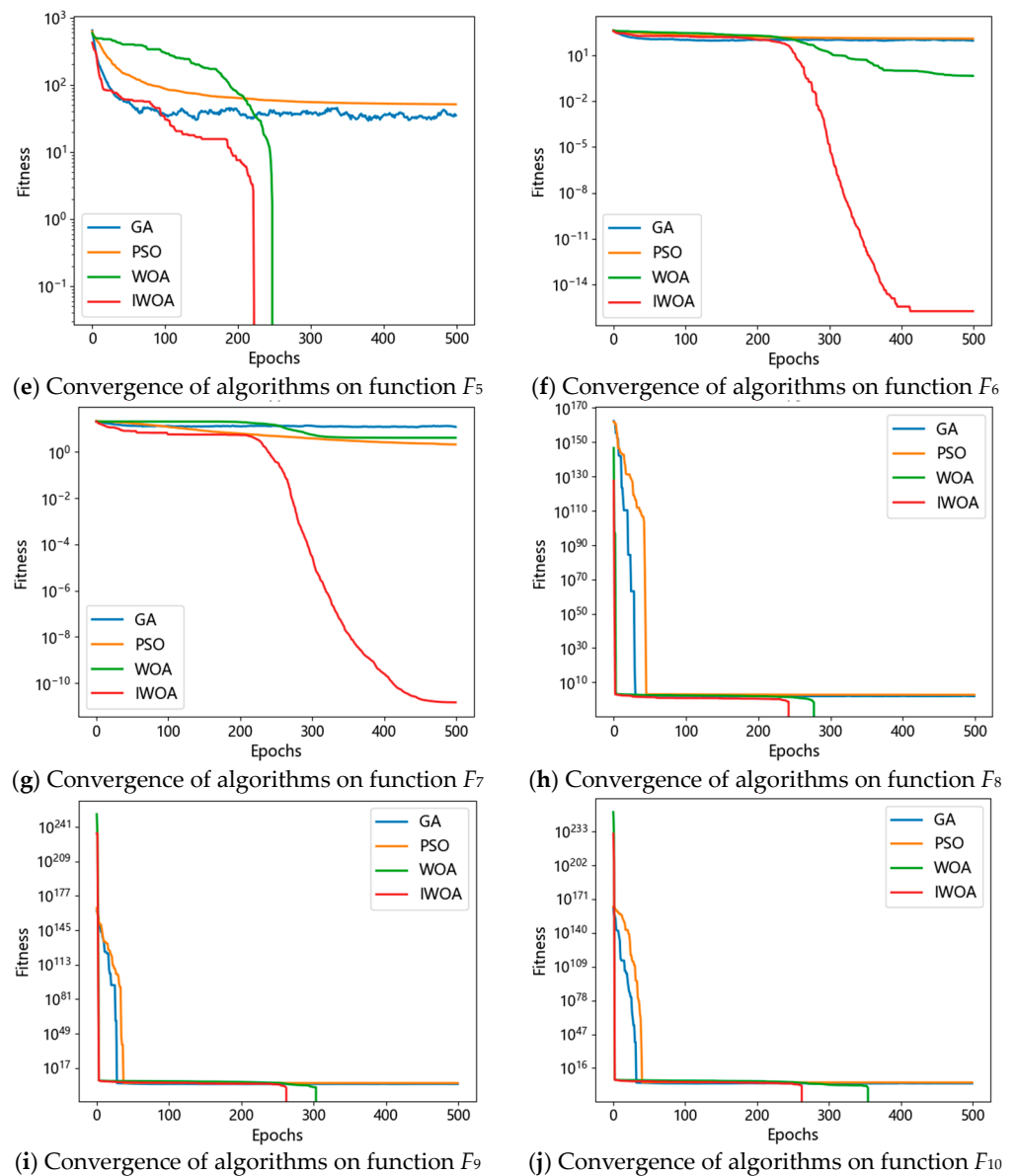

(**j**) Convergence of algorithms on function F_{10}

Figure 6. Average convergence curves for each algorithm.

4.3. One-Step Prediction

Single-step oil temperature prediction involves forecasting the transformer's top oil temperature for the next time step using historical data. In this experiment, the prediction is for 30 min into the future. To balance the training and testing errors, we introduced L2 regularization and dropout during the model training. Specifically, a dropout rate of 0.1 was applied, along with L2 regularization using a factor of 0.01. The prediction results for Dataset 1, demonstrating the effectiveness of the method, are presented in Figure 7. To further illustrate the trade-off between training and testing errors, Figure 8 provides a comparison of the training and testing errors.
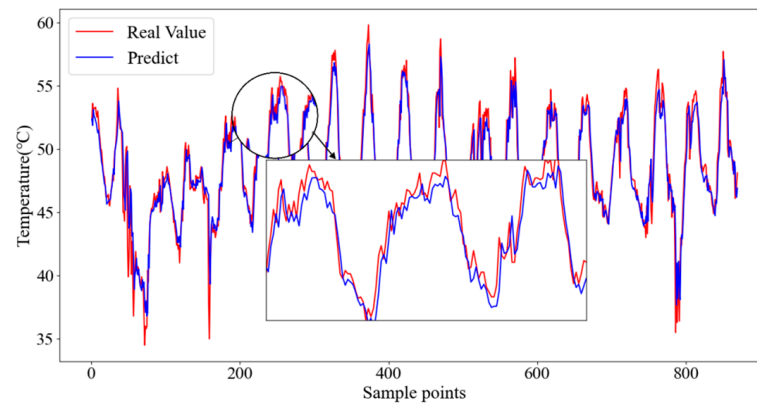
Figure 7. The prediction results of IWOA-LSTM-SA.
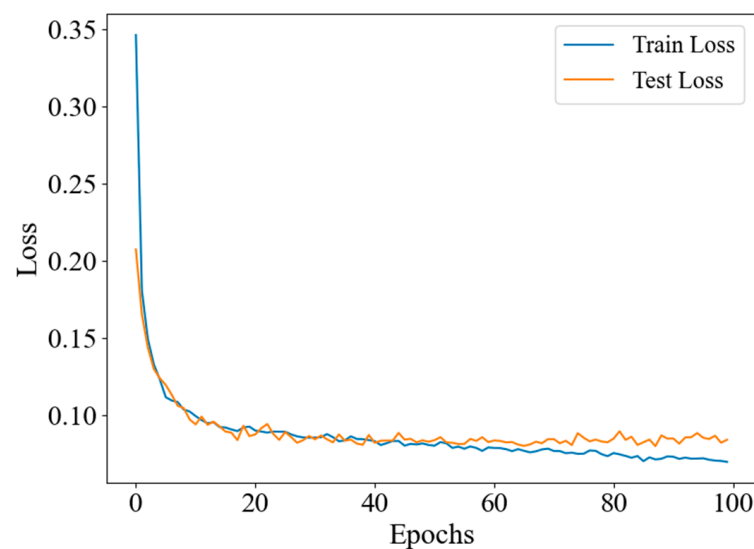

Figure 8. Training and testing errors over iterations.

Theoretically, when there is a significant gap between training and test errors, it usually indicates over-fitting, where the model performs well on the training data but struggles to generalize to unseen data. As illustrated in Figure 8, both the training and test losses decrease rapidly during the initial epochs and then converge to similar values as training progresses. This suggests that we have achieved a well-balanced trade-off between training and testing errors. This balance was successfully attained by applying regularization techniques, such as L2 regularization and dropout, which helped control model complexity, mitigate over-fitting, and enhance the model's generalization capabilities.

To assess the performance of this method, this paper compared it with benchmark methods, including BP, gate recurrent unit (GRU), convolutional neural networks (CNN), LSTM, LSTM-SA, and WOA-LSTM-SA models. In order to reduce the accidental error, this paper conducted 10 repeated experiments and averaged the results to show the forecasting performance. Figure 9 displays the prediction results for each model on Dataset 1. It is evident that the proposed model shows the best prediction result compared to all benchmark models. The reason is that the proposed approach not only combines both local and global information but also utilizes IWOA to determine the optimal hyper-parameters. Table 3 presents the comparative results.
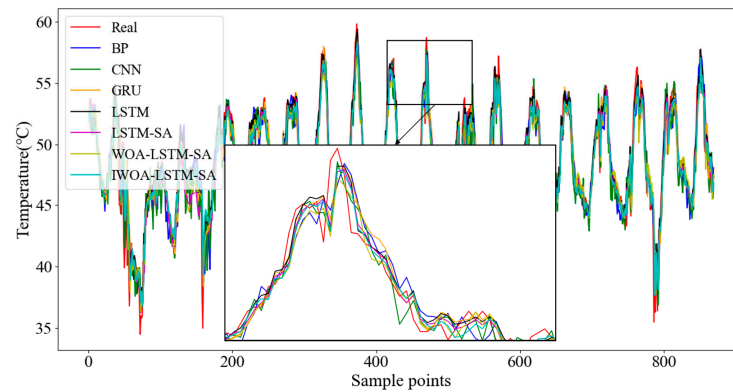
Figure 9. Performance comparison across models.

Table 3. Model prediction evaluation indexes.

	Model	RMSE	MAE	MAPE (%)	R_2	Time (s)
	BP	1.698	1.228	2.581	0.825	13.287
	CNN	1.646	1.170	2.462	0.836	32.317
	GRU	1.553	1.011	2.144	0.854	96.109
Dataset 1	LSTM	1.633	1.022	2.175	0.838	129.666
	LSTM-SA	1.537	1.031	2.253	0.861	174.497
	WOA-LSTM-SA	1.462	0.998	2.103	0.870	11,058.906
	IWOA-LSTM-SA	1.438	0.989	2.089	0.873	10,083.375
	BP	0.923	0.715	2.428	0.974	38.216
	CNN	0.824	0.596	1.929	0.979	80.746
	GRU	0.758	0.544	1.772	0.982	165.984
Dataset 2	LSTM	0.874	0.643	2.129	0.977	234.946
	LSTM-SA	0.809	0.576	1.890	0.980	383.995
	WOA-LSTM-SA	0.757	0.535	1.739	0.982	13,016.477
	IWOA-LSTM-SA	0.749	0.524	1.703	0.983	11,075.689

From Table 3, it is evident that our method does not have an advantage in terms of computation time compared to traditional machine learning models. Therefore, in scenarios where prediction accuracy is not a primary concern, traditional machine learning models can still be considered for top oil temperature prediction of transformers. The prediction model proposed in this paper, however, places a greater emphasis on improving prediction accuracy. To analyze and compare each model more comprehensively, this paper includes a residual plot. Using Dataset 1 as an example, in the residual plot (Figure 10), the true values are shown on the horizontal axis, while the vertical axis represents the residual values (percentage).
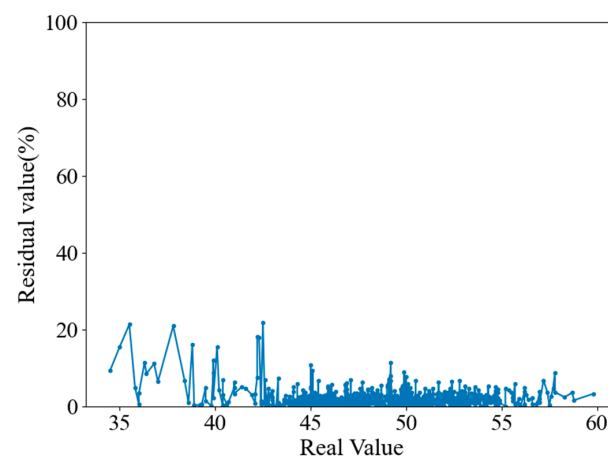

Figure 10. Model residuals.

The residual percentage is relatively higher for the data between 30 and 43 °C and 55 to 60 °C. The reason is as follows: there are about 4000 sample points within the temperature range of 43 to 55 °C, whereas the temperature ranges of 30~43 °C and 55~60 °C each contain approximately 200 sample points. This unbalanced distribution leads to low accuracy on sparse samples.

4.4. Ablation Experiment

To comprehensively validate the effectiveness of each component of the proposed method (IWOA-LSTM-SA), ablation experiments were conducted. Specifically, the experiments compared the following models: LSTM, LSTM-SA, WOA-LSTM, IWOA-LSTM, and WOA-LSTM-SA, with the LSTM model serving as the benchmark for comparison and analysis. Results are shown in Table 4.

Table 4. Ablation experiment evaluation metrics.

		LSTM	**LSTM-SA**	**WOA-LSTM**	**IWOA-LSTM**	**WOA-LSTM-SA**	**IWOA-LSTM-SA**
Dataset 1	RMSE	1.633	1.537	1.596	1.517	1.462	1.438
	MAPE	2.175	2.253	2.141	2.106	2.103	2.089
Dataset 2	RMSE	0.874	0.809	0.837	0.782	0.757	0.749
	MAPE	2.129	1.890	2.042	1.814	1.739	1.703

As shown in Table 4, the proposed model demonstrates higher prediction accuracy compared to the baseline model LSTM and other comparative models. Compared to LSTM, the RMSE of LSTM-SA decreased by 5.88% on Dataset 1 and by 7.44% on Dataset 2; the MAPE increased by 3.59% on Dataset 1 but decreased by 11.23% on Dataset 2. This validates the effectiveness of combining the SA algorithm with LSTM. Compared to LSTM-SA, the RMSE of WOA-LSTM-SA and IWOA-LSTM-SA decreased by 4.88% and 6.44% on Dataset 1, and by 6.43% and 7.42% on Dataset 2, respectively. The MAPE decreased by 6.66% and 7.28% on Dataset 1, and by 7.99% and 9.89% on Dataset 2, respectively. This validates the effectiveness of the optimization algorithms proposed in the models. Additionally, compared to WOA-LSTM and IWOA-LSTM, the RMSE of the proposed model decreased by 9.89% and 5.21% on Dataset 1, and by 10.51% and 4.22% on Dataset 2, respectively. The MAPE decreased by 2.43% and 0.81% on Dataset 1, and by 16.60% and 6.12% on Dataset 2, respectively.

In summary, compared to using optimization algorithms or SA individually, combining them results in a greater improvement in the performance of the prediction model.

4.5. Multi-Step Forecasting

The multi-step prediction model refers to a model that predicts a series of values rather than a single value. Multi-step prediction is more important in real-world power system operations because it provides longer-term temperature trend forecasts, which help to identify potential issues in advance. Therefore, this section conducts a multi-step prediction analysis, where the prediction steps are set to 3 steps (90 min) and 5 steps (150 min). The evaluation metrics are shown in Table 5, and the prediction results (for one week) are presented in Figure 11.

From Table 5, it can be seen that the error increases as the prediction step increases across all models. By comparing the RMSE metric, it can be concluded that the proposed model exhibits better accuracy across different prediction steps compared to the baseline model. Specifically, in Dataset 1 and Dataset 2, for the 3 step prediction, the RMSE of the proposed model is 1.537 and 1.015, respectively. This represents reductions of 12.83% and 38.65% compared to the BP model, 6.98% and 20.89% compared to the CNN model, 3.75% and 13.62% compared to the GRU model, 4.24% and 27.16% compared to the LSTM model, 1.60% and 17.93% compared to the LSTM-SA model, and 1.16% and 4.34% compared to the WOA-LSTM-SA model. For the 5 step prediction, the RMSE of the proposed model is 1.714

and 1.634, representing reductions of 12.60% and 11.11% compared to the BP model, 7.61% and 15.89% compared to the CNN model, 6.49% and 17.30% compared to the GRU model, 5.19% and 14.14% compared to the LSTM model, 4.56% and 12.82% compared to the LSTM-SA model, and 3.06% and 1.80% compared to the WOA-LSTM-SA model. By analyzing the multi-step prediction metrics, we conclude that the proposed model demonstrates good performance across different prediction steps compared to traditional models.

Table 5. Multi-step prediction evaluation metrics.

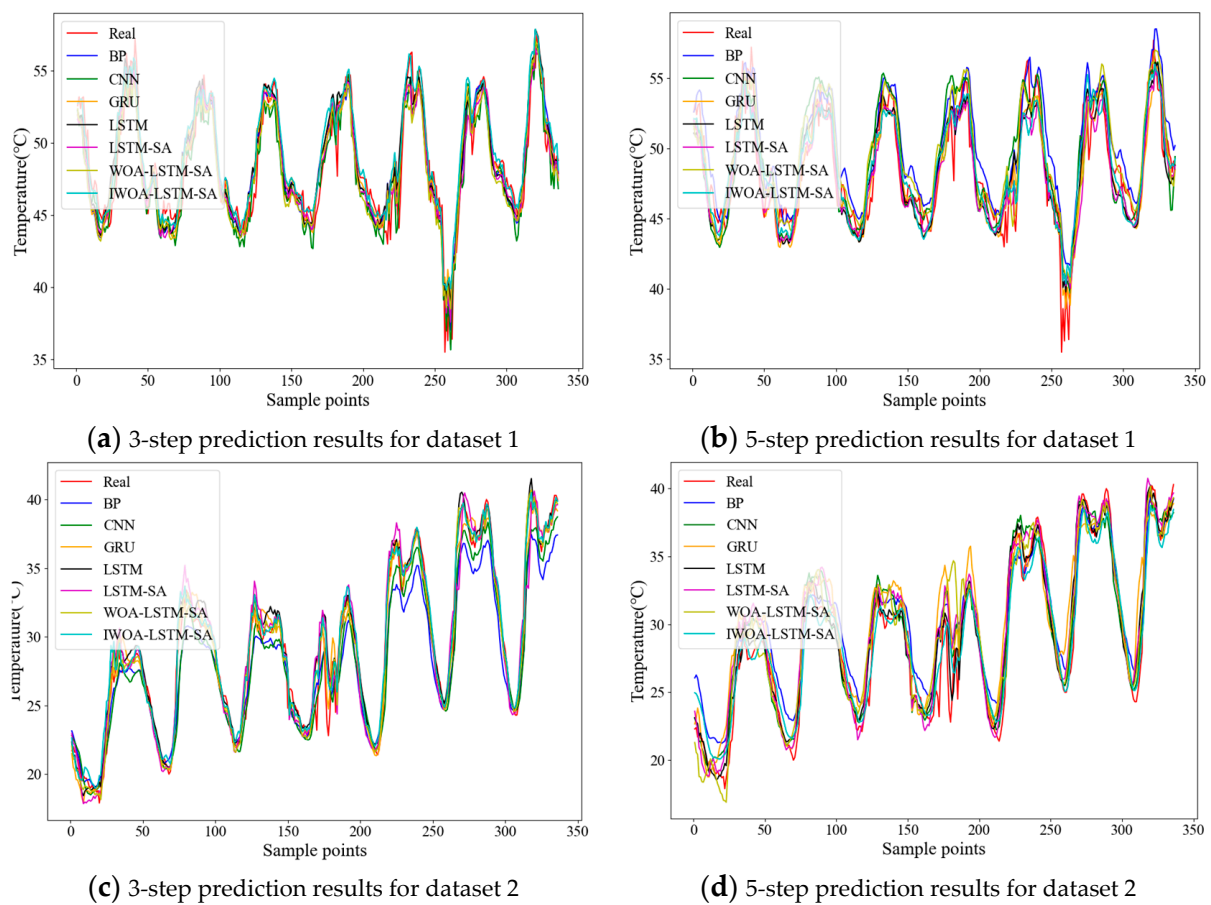|	Step	Model	RMSE	MAE	MAPE (%)	Time (s)
Dataset 1	1 (30 min)	BP	1.698	1.228	2.581	13.287
		CNN	1.646	1.170	2.462	32.317
		GRU	1.553	1.011	2.144	96.109
		LSTM	1.633	1.022	2.175	129.666
		LSTM-SA	1.537	1.031	2.253	174.497
		WOA-LSTM-SA	1.462	0.998	2.103	11,058.906
		IWOA-LSTM-SA	1.438	0.989	2.089	10,083.375
	3 (90 min)	BP	1.763	1.382	2.873	14.082
		CNN	1.652	1.221	2.557	22.572
		GRU	1.597	1.133	2.409	95.775
		LSTM	1.605	1.164	2.453	179.898
		LSTM-SA	1.562	1.162	2.448	229.012
		WOA-LSTM-SA	1.555	1.102	2.311	11,746.135
		IWOA-LSTM-SA	1.537	1.088	2.308	10,149.217
	5 (150 min)	BP	1.961	1.611	3.351	13.617
		CNN	1.855	1.411	2.973	21.579
		GRU	1.833	1.387	2.943	98.763
		LSTM	1.808	1.367	2.878	197.507
		LSTM-SA	1.796	1.345	2.832	240.519
		WOA-LSTM-SA	1.768	1.352	2.859	12,212.086
		IWOA-LSTM-SA	1.714	1.294	2.702	10,778.976
Dataset 2	1 (30 min)	BP	0.923	0.715	2.428	38.216
		CNN	0.824	0.596	1.929	80.746
		GRU	0.758	0.544	1.772	165.984
		LSTM	0.874	0.643	2.129	234.946
		LSTM-SA	0.809	0.576	1.890	383.995
		WOA-LSTM-SA	0.757	0.535	1.739	13,016.477
		IWOA-LSTM-SA	0.749	0.524	1.703	11,075.689
	3 (90 min)	BP	1.654	1.124	4.225	37.313
		CNN	1.283	1.012	3.166	79.190
		GRU	1.175	0.831	2.821	229.788
		LSTM	1.394	1.080	3.674	320.336
		LSTM-SA	1.237	0.923	3.111	433.645
		WOA-LSTM-SA	1.061	0.833	2.746	13,623.563
		IWOA-LSTM-SA	1.015	0.750	2.537	11,284.158
	5 (150 min)	BP	1.838	1.568	4.854	37.081
		CNN	1.943	1.403	4.933	77.883
		GRU	1.976	1.387	4.801	264.860
		LSTM	1.903	1.414	4.765	171.239
		LSTM-SA	1.874	1.365	4.810	414.213
		WOA-LSTM-SA	1.664	1.249	4.298	12,823.645
		IWOA-LSTM-SA	1.634	1.229	4.162	10,984.776

(**a**) 3-step prediction results for dataset 1

(**b**) 5-step prediction results for dataset 1

(**c**) 3-step prediction results for dataset 2

(**d**) 5-step prediction results for dataset 2

Figure 11. Multi-step prediction performance comparison across models (one week).

5. Conclusions

Oil temperature prediction can effectively prevent symmetrical and asymmetrical faults in transformers. This paper adopts a novel approach to improve the performance of top-oil temperature prediction during transformer operations. The proposed model has been tested using actual data, and some conclusions can be obtained as follows:

(1) To verify the efficacy of the IWOA, this paper conducts tests with eight test functions. The findings demonstrate that the IWOA outperforms GA, PSO, and WOA in terms of convergence speed and accuracy.

(2) To verify the effectiveness of the proposed model, extensive experiments were conducted using actual operating data. The experimental results indicate that the proposed approach outperforms current state-of-the-art methods. On Dataset 1, the model achieved reductions in RMSE of 15.31%, 12.64%, 7.41%, 11.94%, 6.44%, and 1.98% compared to the BP, CNN, GRU, LSTM, LSTM-SA, and WOA-LSTM-SA methods, respectively. Similarly, on Dataset 2, the model demonstrated significant improvements, with RMSE reductions of 18.85%, 9.09%, 1.19%, 14.29%, 7.42%, and 1.06% compared to the same benchmark methods.

(3) The proposed model performs effectively across various prediction steps compared to benchmark models. Specifically, for the 3-step prediction, the RMSE of the proposed model is 1.537 and 1.015 for Dataset 1 and Dataset 2, respectively, reflecting reductions of 12.83% and 38.65% compared to the BP model, 6.98% and 20.89% compared to the CNN model, 3.75% and 13.62% compared to the GRU model, 4.24% and 27.16% compared to the LSTM model, 1.60% and 17.93% compared to the LSTM-SA model, and 1.16% and 4.34% compared to the WOA-LSTM-SA model. For the 5-step prediction, the RMSE of the proposed model is 1.714 and 1.634, representing reductions of 12.60% and 11.11% compared to the BP model, 7.61% and 15.89% compared to

the CNN model, 6.49% and 17.30% compared to the GRU model, 5.19% and 14.14% compared to the LSTM model, 4.56% and 12.82% compared to the LSTM-SA model, and 3.06% and 1.80% compared to the WOA-LSTM-SA model.

Author Contributions: D.Z. led the conceptualization, methodology, software development, and original draft preparation. Validation was carried out by D.Z., H.X. and H.Q., while H.X. and D.Z. handled formal analysis. H.Q. managed the investigation, and Z.H. and W.D. provided resources. S.W. was responsible for data curation. Writing—review and editing involved D.Z., H.X., H.Q., Q.P. and J.Y., with visualization by D.Z., H.X. and J.Y. Supervision was provided by D.Z. and H.Q., project administration by Q.P. and S.W., and funding acquisition by D.Z. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: Data are contained in the article.

Conflicts of Interest: Authors Dexu Zou, Qingjun Peng, Shan Wang, Weiju Dai, and Zhihu Hong were employed by the company China Southern Power Grid Yunnan Power Grid Co., Ltd. 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.

Appendix A

Table A1 displays the ten test functions used in this study.

Table A1. Test functions.

Function	Range
$F_1(x) = \sum_{n=1}^{k} x_n^2$	$[-100, 100]$
$F_2(x) = \sum_{n=1}^{k} |x_n| + \prod_{n=1}^{k} |x_n|$	$[-10, 10]$
$F_3(x) = \sum_{n=1}^{k} \left(\sum_{i-1}^{n} x_i \right)^2$	$[-100, 100]$
$F_4(x) = \sum_{n=1}^{k} n x_n^4 + random[0, 1)$	$[-1.28, 1.28]$
$F_5(x) = 1 + \frac{1}{4000} * \sum (x_n^2) - \prod \left(\cos \left(\frac{x_n}{\sqrt{n}} \right) \right)$	$[-600, 600]$
$F_6(x) = [x_n^2 - 10\cos(2\pi x_n) + 10]$	$[-5.12, 5.12]$
$F_7(x) = 20 - 20\exp\left(-0.2\sqrt{\frac{1}{k}\sum_{n=1}^{k} x_n^2} \right) - \exp\left[\frac{1}{k}\sum_{n=1}^{k} \cos(2\pi x_n) \right] + e$	$[-32, 32]$
$F_8(x) = \frac{\pi}{k}\left\{ 10\sin(\pi y_1) + \sum_{n=1}^{k-1}(y_n - 1)^2[1 + 10\sin^2(\pi y_{n+1})] + (y_n - 1)^2 \right\} + \sum_{n=1}^{k} \mu(x_n, 10, 100, 4)$	$[-50, 50]$
$F_9(x) = \sum_{i=d}^{d}\left(-x_i \times \sin\left(\sqrt{|x_i|} \right) \right) + 418.98288727243369 \times d$	$[-500, 500]$
$F_{10}(x) = \sum_{i=1}^{d}\left((\ln(x_i - 2))^2 + (\ln(10 - x_i))^2 \right) - \left(\prod_{i=1}^{10} x_i \right)^{0.2}$	$[2, 10]$

References

1. Xu, X.; He, Y.; Li, X.; Peng, F.; Xu, Y. Overload Capacity for Distribution Transformers with Natural-Ester Immersed High-Temperature Resistant Insulating Paper. *Power Sys. Technol.* **2018**, *42*, 1001–1006.
2. Wang, S.; Gao, M.; Zhuo, R. Research on high efficient order reduction algorithm for temperature coupling simulation model of transformer. *High Volt. Appar.* **2023**, *59*, 115–126.
3. Liu, X.; Xie, J.; Luo, Y. A novel power transformer fault diagnosis method based on data augmentation for KPCA and deep residual network. *Energy Rep.* **2023**, *9*, 620–627. [CrossRef]
4. Chen, T.; Chen, Y.; Li, X. Prediction for dissolved gas concentration in power transformer oil based on CEEMDAN-SG-BiLSTM. *High Volt. Appar.* **2023**, *59*, 168–175.
5. Zang, C.; Zeng, J.; Li, P. Intelligent diagnosis model of mechanical fault for power transformer based on SVM algorithm. *High Volt. Appar.* **2023**, *59*, 216–222.
6. Ji, H.; Wu, X.; Wang, H. A New Prediction Method of Transformer Oil Temperature Based on C-Prophet. *Adv. Power Syst. Hyd. Eng.* **2023**, *39*, 48–55.

7. Tan, F.; Xu, G.; Zhang, P. Research on Top Oil Temperature Prediction Method of Similar Day Transformer Based on Topsis and Entropy Method. *Elect. Power Sci. Eng.* **2021**, *37*, 62–69.
8. Amoda, O.A.; Tylavsky, D.J.; McCulla, G.A.; Knuth, W.A. Acceptability of three transformer hottest-spot temperature models. *IEEE Trans. Power Deliv.* **2011**, *27*, 13–22. [CrossRef]
9. Zhou, L.; Wang, J.; Wang, L.; Yuan, S.; Huang, L.; Wand, D.; Guo, L. A Method for Hot-Spot Temperature Prediction and Thermal Capacity Estimation for Traction Transformers in High-Speed Railway Based on Genetic Programming. *IEEE Trans. Transp. Electrif.* **2019**, *5*, 1319–1328. [CrossRef]
10. Deng, Y.; Ruan, J.; Quan, Y.; Gong, R.; Huang, D.; Duan, C.; Xie, Y. A Method for Hot Spot Temperature Prediction of a 10 kV Oil-Immersed Transformer. *IEEE Access* **2019**, *7*, 107380. [CrossRef]
11. Zhao, B.; Zhang, X. Parameter Identification of Transformer Top Oil Temperature Model and Prediction of Top Oil Tempeature. *High. Volt. Eng.* **2004**, *30*, 9–10.
12. Wang, H.; Su, P.; Wang, X. Prediction of Surface Temperatures of Large Oil-Immersed Power Transformers. *J. Tsinghua Univ. Sci. Technol.* **2005**, *45*, 569–572.
13. Tan, M.; Hu, C.; Chen, J.; Wang, L.; Li, Z. Multi-node load forecasting based on multi-task learning with modal feature extraction. *Eng. Appl. Artif. Intell.* **2022**, *112*, 104856. [CrossRef]
14. Shang, Y.; Li, S. FedPT-V2G: Security enhanced federated transformer learning for real-time V2G dispatch with non-IID data. *Appl. Energy* **2024**, *358*, 122626. [CrossRef]
15. Bai, M.; Yao, P.; Dong, H.; Fang, Z.; Jin, W.; Yang, X.; Liu, J.; Yu, D. Spatial-temporal characteristics analysis of solar irradiance forecast errors in Europe and North America. *Energy* **2024**, *297*, 131187. [CrossRef]
16. Qing, H.; Jennie, S.; Daniel, J. Prediction of top-oil temperature for transformers using neural network. *IEEE Trans. Power Deliv.* **2000**, *15*, 1205–1211.
17. Tan, F.; Chen, H.; He, J. Top oil temperature forecasting of UHV transformer based on path analysis and similar time. *Elect. Power Autom. Equip.* **2021**, *41*, 217–224.
18. Li, S.; Xue, J.; Wu, M.; Xie, R.; Jin, B.; Zhang, H.; Li, Q. Prediction of Transformer Top-oil Temperature with the Improved Weighted Support Vector Regression Based on Particle Swarm Optimization. *High Volt. Appar.* **2021**, *57*, 103–109.
19. Tan, F.L.; Xu, G.; Li, Y.F.; Chen, H.; He, J.H. A method of transformer top oil temperature forecasting based on similar day and similar hour. *Elect. Power Eng. Tech.* **2022**, *41*, 193–200.
20. Yi, Y. Research on Prediction Method of Transformer Top-Oil Temperature Based on Assisting Dispatchers in Decision-Making. Master's Thesis, Southwest Jiaotong University, Chengdu, China, 2017.
21. Gharehchopogh, F.S.; Gholizadeh, H. A comprehensive survey: Whale Optimization Algorithm and its applications. *Swarm Evol. Comput.* **2019**, *48*, 1–24. [CrossRef]
22. Brodzicki, A.; Piekarski, M.; Jaworek-Korjakowska, J. The whale optimization algorithm approach for deep neural networks. *Sensors* **2021**, *21*, 8003. [CrossRef] [PubMed]
23. Mostafa Bozorgi, S.; Yazdani, S. IWOA: An improved whale optimization algorithm for optimization problems. *J. Comput. Des. Eng.* **2019**, *6*, 243–259. [CrossRef]
24. Naderi, E.; Azizivahed, A.; Asrari, A. A step toward cleaner energy production: A water saving-based optimization approach for economic dispatch in modern power systems. *Electr. Power Syst. Res.* **2022**, *204*, 107689. [CrossRef]
25. Gao, W.; Liu, S.; Huang, L. Inspired artificial bee colony algorithm for global optimization problems. *Acta Electron. Sin.* **2012**, *40*, 2396.
26. Shi, X.; Li, M.; Wei, Q. Application of Quadratic Interpolation Whale Optimization Algorithm in Cylindricity Error evaluation. *Metrol. Meas. Tech.* **2019**, *46*, 58–60.
27. He, Q.; Wei, K.; Xu, Q. Mixed strategy based improved whale optimization algorithm. *Appl. Res. Comput.* **2019**, *36*, 3647–3651.
28. Qiu, X.; Wang, R.; Zhang, W.; Zhang, Z.; Zhang, Q. Improved Whale Optimizer Algorithm Based on Hybrid Strategy. *Comput. Eng. Appl.* **2022**, *58*, 70–78.
29. Chen, Y.; Han, B.; Xu, G.; Kan, Y.; Zhao, Z. Spatial Straightness Error Evaluation with Improved Whale Optimization Algorithm. *Mech. Sci. Technol. Aero. Eng.* **2022**, *41*, 1102–1111.
30. Xu, J.; Yan, F. The Application of Improved Whale Optimization Algorithm in Power Load Dispatching. *Oper. Res. Manag. Sci.* **2020**, *29*, 149–159.
31. Naderi, E.; Mirzaei, L.; Pourakbari-Kasmaei, M.; Cerna, F.V.; Lehtonen, M. Optimization of active power dispatch considering unified power flow controller: Application of evolutionary algorithms in a fuzzy framework. *Evol. Intell.* **2024**, *17*, 1357–1387. [CrossRef]