Attention-GRU Based Intelligent Prediction of Oil Temperature for the Transformers

Jingting Wang, Gaoji Yang, Zhongjie Cheng, Bao Liu

Abstract—**The oil temperature serves as a crucial indicator for assessing the thermal performance of power transformers. Accurately predicting its temporal variation trend is of great significance for evaluating the operational load capacity and enabling early detection of potential internal overheating faults. This paper proposes an effective oil temperature prediction method for transformers based on the attention mechanism and the gated recurrent unit (Attention-GRU) to tackle the problems of low accuracy and cumbersome feature selection process in oil temperature prediction. Firstly, this article conducts data processing on oil temperature data to reduce computational complexity. Secondly, this article utilizes the GRU to extract data features of oil temperature from transformers at the time scale for long-range predictive modeling. Thirdly, the attention mechanism is introduced to autonomously emphasize the data features that yields substantial effects on the results, replacing the complex and time-consuming data screening process and improving the prediction accuracy. Finally, model outputs prediction results. The experimental results demonstrate that in comparison with existing methods (see, e.g., RNN, LSTM, GRU, BiLSTM, and Transformer), our method has higher prediction accuracy of oil temperature from transformers, which helps to analyze the load capacity of transformers and detect potential faults. At the same time, this paper has also found that the attention mechanism can significantly reduce the dependence of existing methods on the feature selection process, and enhance prediction accuracy of oil temperature in transformers.**

Index Terms—**transformer, oil temperature prediction, attention mechanism, gated recurrent unit, time series prediction**

I. Introduction

As one of the most key power transmission and transformation equipment, transformers are widely used in power systems. The number of system operation accidents caused by transformer failures is also constantly increasing. Therefore, the research on condition monitoring and fault diagnosis of transformers is significant to power systems. According to statistics, among all the fault types of transformers, overheating faults account for the largest proportion, and the rate at which the insulation life of transformers decreases will continue to accelerate with the increase in temperature. Further damage to the transformer will be caused and overheating even poses severe threats to the transformer's safe operation if overheating faults are not dealt with in a timely manner. A substantial quantity of insulating oil must be injected into power transformers to guarantee their insulation and heat dissipation performance. To some extent, the temperature of the insulating oil can indicate the operational status of the transformer [1]-[2]. The implementation of accurate prediction of oil temperature is central to identifying potential transformer faults, performing transformer operation and maintenance, and realizing early warning of transformer failures.

At present, driven by accelerated technological progress, especially in AI and deep learning domains, prediction models utilizing deep learning are widely applied in various industries (e.g., crack length [3], flight control systems [4], tax revenue [5], and prices [6]) and many researchers have also applied it to transformer oil temperature prediction. In [7], the ambient temperature, the active power, and reactive power corresponding to input data, were selected as the input features, and the Long Short Term Memory (LSTM) network was adopted to achieve prediction of the top oil temperature. Reference [8] took into account the factor of three-phase imbalance, decomposed the load factor into three phases, and constructed an LSTM prediction model. Compared with the method of differential calculation using the heat transfer differential equation, the proposed method was more in line with the actual operating conditions of the transformer. Reference [9] put forward a multi-step prediction model of the gated recurrent unit (GRU) network based on multi-output strategy. By changing the structural hyperparameters and training hyperparameters of the model to study the coupling relationship among the hyperparameters, the multi-objective grey wolf optimization algorithm was used to optimize hyperparameters with different prediction result tendencies. Reference [10] used ensemble empirical mode decomposition and wavelet packets to decompose the concentration signal into several sub-signals, and then applied the CNN-GRU model to predict the concentration signal. Reference [11] combined the Radam optimizer and the gated recurrent unit and proposed an improved GRU prediction model for the gas concentration in transformer oil. The experimental results showed that for the LSTM neural network model, the improved GRU model could better predict the changing trend of the dissolved gas

Manuscript received January 22, 2025; revised July 5, 2025.

This work is supported in part by grant for the Key Research and Development Program of Shaanxi (2021GY-131).

Jingting Wang is an associate professor of Information Engineering School, Xi'an FANYI University, Xi’an 710105, China (e-mail: 851969149@qq.com).

Gaoji Yang is the legal representative of Jiugang (Group) Honglian Automation Co., Ltd., Jiayuguan City 735100, China (corresponding author, +86-18909475577, e-mail: YanggaojiJiugang@163.com).

Zhongjie Cheng is a postgraduate student of Electrical and Control Engineering, Xi’an University of Science and Technology, Xi’an 710054, China (e-mail: 2472530209@qq.com).

Bao Liu is an associate professor of Electrical and Control Engineering, Xi’an University of Science and Technology, Xi’an 710054, China (e-mail: xiaobei0077@163.com).

concentration in transformer oil. Reference [12] predicted the oil temperature trend of a certain ±800 kV converter station based on the LSTM and verified the practicability of the method through oil temperature early warning. Reference [13] considers the spatial coupling characteristics existing between neighboring nodes and the impact of the external environment on the load, and adds Transformer to the prediction model to reduce prediction error of transformer load prediction. Reference [14] used the particle swarm optimization algorithm to further optimize relevant hyperparameters of the Bi-directional LSTM (BiLSTM). It advanced diagnostic prediction precision and robustness of the model. However, the randomly generated hyperparameters of the BiLSTM network increased the instability of the model and it tended to fall into local optima. Reference [15] applied LSTM to the prediction of the trend of the dissolved gas concentration. The above methods provided ideas for oil temperature prediction by adding influencing factors related to oil temperature such as load and weather conditions. However, there are numerous available influencing factors for the top oil temperature. The acquisition cost and acquisition cycle need to be considered.

This paper uses the recurrent neural network (RNN) series [16], considering temperature prediction is essentially a prediction based on time series. The traditional RNN has problems such as vanishing gradients, exploding gradients, and at the same time, its ability to process data feature information with a long distance is relatively weak [17]. The LSTM and the GRU networks originate from the RNN. The LSTM uses three gate structures to remember effective information and discard useless interfering information [18]. The GRU reduces structural complexity relative to LSTM networks, but its performance is relatively better. The attention mechanism can well model sequence data with variable lengths, further enhance its ability to capture long-range dependent information, and effectively improve accuracy while reducing the depth of hierarchy [19].

Based on the above analysis, this paper chooses a prediction method combining the GRU and the attention mechanism (Attention-GRU) to achieve higher prediction accuracy of the oil temperature of transformers. The main contributions are as follows:

(1) Aiming at the problems that oil temperature is affected by multiple factors and traditional models show difficulty for adapting to the prediction and require a manual feature selection process, this paper introduces an attention mechanism with self-learning ability to focus on important time features. This mechanism enables the GRU to adaptively get the focus weights of different input data, thereby boosting the model's predictive accuracy.

(2) This paper verifies the effectiveness of this method on two datasets, namely ETTh1 and ETTh2. The results show that contrasting with the existing prediction models (RNN, LSTM, BiLSTM, GRU, and Transformer), Attention-GRU is more suitable for complex data input in practical scenarios and is conducive to achieving accurate prediction of oil temperature.

(3) Aiming at the problem of low prediction accuracy in feature learning process of existing methods, this paper verifies through ablation experiments on ETTh1 dataset and ETTh2 dataset that the attention mechanism can significantly boost forecasting precision of the method. The experimental results show that this mechanism adaptively allocates focus weights and solves the above-mentioned problems.

The other parts of the paper are arranged as follows. The modeling process of oil temperature prediction is introduced in detail in Section II. Section III presents the experimental results and analysis, which confirm the effectiveness of the attention mechanism and Attention-GRU. Section IV provides the conclusions.

II. Attention-GRU Predictive Modeling of Oil Temperature from the Transformer

This section will introduce the Attention-GRU model from three parts: the GRU, the attention mechanism, and Attention-GRU. This paper theoretically analyzes the importance of the attention mechanism.

A. The GRU for temporal feature information extraction

The GRU was first proposed in 2014 [20]. Fig. 1 shows the schematic diagram of the GRU model. As shown in Fig. 1, r_t is the reset gate, x_t is the input at time t, u_t is the update gate, h_t is the output at time t, and h_{t-1} is the hidden state at the previous moment of t.

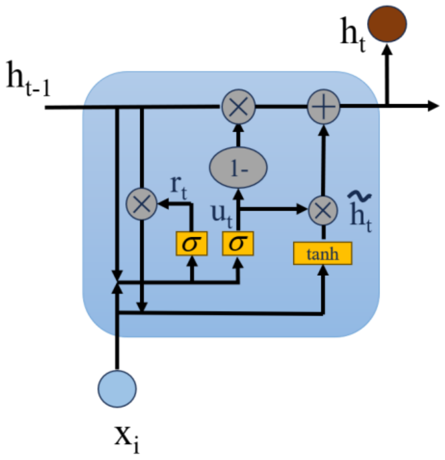


Fig. 1. The schematic diagram of GRU

The working principle of the GRU neural network is as follows:

$$r_t = \sigma\left(W_r h_{t-1} + W_r x_t\right) \tag{1}$$

$$u_t = \sigma\left(W_z h_{t-1} + W_z x_t\right) \tag{2}$$

$$\tilde{h}_t = \tanh\left[r_t \odot \left(W_z' h_{t-1}\right) + W_x' x_t\right] \tag{3}$$

$$h_t = u_t \odot h_{t-1} + \left(1 - u_t\right) \odot \tilde{h}_t \tag{4}$$

$$\tanh = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{5}$$

where W_r , W_z , W_h' , and W_x' are respectively trainable weights, $\odot$ represents Hadamard multiplication, and σ is the sigmoid function.

The reset gate modulates historical information flow into the candidate set: lower values suppress prior-state integration. Conversely, the update gate governs previous-state retention in the current state, where higher values preserve more historical information.

The GRU model computes the weights of input data in a time-series manner and generates the output results when the GRU network processes time series data. The prediction error is obtained after comparing and calculating the output results with the target data. During the next learning process, the prediction error is input into the GRU network to update the weights in the GRU network through the backpropagation of the prediction error toward higher forecasting accuracy.

B. The attention mechanism for focusing on key information

The human brain's signal processing mechanism will focus on some target areas in the current state and reduce or ignore the attention to other areas. Based on the human attention research, experts have put forward the attention mechanism. Essentially, it optimizes information resource allocation [21]. Now the attention mechanism has found extensive applications in many fields of deep learning. By weighting input data, the attention mechanism enables models to focus on information with higher importance for current outputs, thereby enhancing model performance. Initially applied in Natural Language Processing (NLP), attention mechanisms are now widely used in various time-series processing tasks. In the field of data prediction, it addresses several limitations in traditional neural networks, such as the decline in system performance as the input length increases, computational inefficiency arising from suboptimal input ordering, and inadequate feature extraction and enhancement capabilities [22].

The attention mechanism includes multiple types and variants, among which the self-attention mechanism is one type of them. It connects different positions of a single sequence, completes the calculation of sequence weight values, represents relevance by calculating the similarity between vectors, and assigns high weight values to key information to enhance its influence on the results [23]. The structural schematic diagram of self-attention mechanism is shown in Fig. 2.

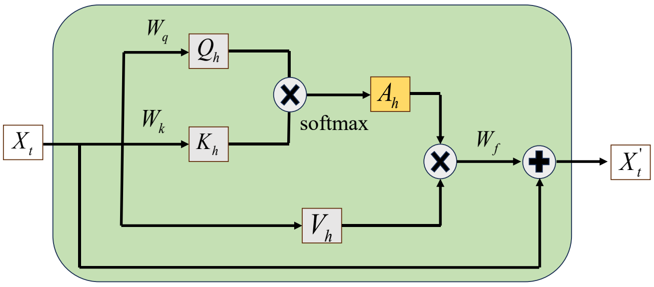


Fig. 2. The Structure diagram of self-attention mechanism

where X_t is the model input, W_q, W_k, and W_f represent weight respectively, Q_h, K_h, A_h, and V_h contain all points as query, key, value, and attention score, respectively, $X_t^{'}$ is the output result, ⊗ represents the matrix multiplication, and ⊕ represents the element wise matrix addition.

The self-attention mechanism focuses on the internal relationships of the input data. The main process of self-attention is as follows:

Step 1 According to the input data, corresponding vectors Q_i, K_i and V_i for each position point of each sample will be generated, where K_i and V_i are a pair of key-value pairs. And then the similarity $e_{i,j}$ by Q_i and the k_j of the j point will be calculated according to equation (6):

$$e_{i,j} = Q_i^T K_j = \left(X_i^T W_q^T\right)\left(W_K X_j\right) \tag{6}$$

where X_i^T and X_j are feature vectors respectively, W is the weight, and different subscripts correspond to different weights, and N is the product of the height and width of the input data.

Step 2 It generates attention scores $\alpha_{i,j}$ by using the softmax function, according to equation (7):

$$\alpha_{i,j} = \frac{\exp e_{i,j}}{\sum_{k=1}^{N} \exp e_{i,k}}, i, j \in \{1,2,\ldots,N\} \tag{7}$$

Step 3 All attention scores can form a matrix or relevant structure, and to obtain the weighted vectors then multiply the attention scores by the value of each sample, according to equation (8). Finally, the characteristics of each point can be obtained.

$$Z_i = \sum_{j=1}^{N} \alpha_{i,j}\left(W_v X_j\right), i, j \in \{1,2,\ldots,N\} \tag{8}$$

Step 4 The characteristics form z, which represents matrices. The weighted vectors encode the effective information of the training samples. Finally, all the weighted vectors are integrated to the input information to derive the output with the global information according to the equation (9):

$$X_t^{'} = W_f Z + X_t, t \in \{1,2,\ldots,N\} \tag{9}$$

The attention mechanism can strengthen the key information and weaken irrelevant information, so as to enhance prediction accuracy of algorithms.

C. The Attention-GRU for predictive modeling

Relative to the LSTM, the GRU performs better in smaller datasets and can better address the problem of gradient explosion or vanishing in long sequence training [24]. Attention mechanism will enable the network to achieve higher efficiency and higher accuracy in information processing. This paper proposes that Attention-GRU can improve the prediction accuracy by discovering deep relationships between input data through attention mechanisms.

Fig. 3 shows the network of Attention-GRU. The attention mechanism weights and calculates the hidden-layer vector expressions output by the GRU. The weight indicates the importance of features at each time point. The structure is composed of an input layer, a GRU layer, an Attention layer, and an output layer. The input is the vector representation of each group of data, x_1, x_2, . . . , x_i. The corresponding outputs k_1, k_2, . . . , k_i will be obtained after vectors are calculated by the GRU model. Then, The attention mechanism is incorporated into the hidden layer in order to compute the attention probability distribution values

$c_1, c_2, \ldots, c_i$ of each input. The equations are as follows:

$$c_i = f_i k_i \tag{10}$$

$$a_i = w_i \times \tan h\left(W_i c_i + b_i\right) \tag{11}$$

$$s_i = \frac{\exp\left(a_i\right)}{\sum_{j=1}^{i} a_j} \tag{12}$$

where f_i is the weight value of each vector, a_i represents the hidden layer state vector at time I, k_i is the output value after the GRU model calculation, w_i and W_i represent respectively the weight coefficient matrices of the i vector, and b_i is the offset corresponding to the i vector.

The feature vector can be calculated through the above equations. Finally, the softmax function normalizes output layer logits into probability distributions correspondingly to evaluate its influence on the output.

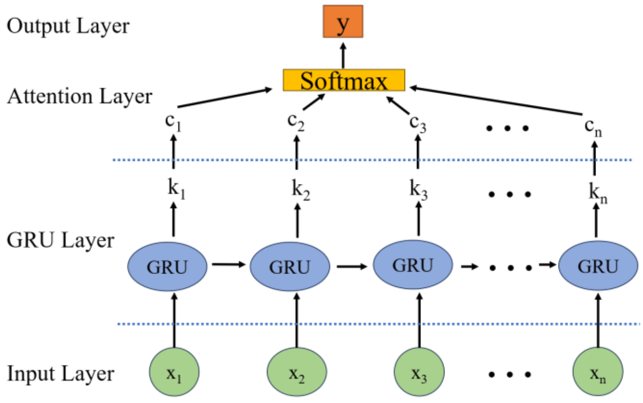


Fig. 3. The structure of Attention-GRU

D. Prediction process

Fig. 4 shows the prediction process figure based on the Attention-GRU model. The steps are as follows:

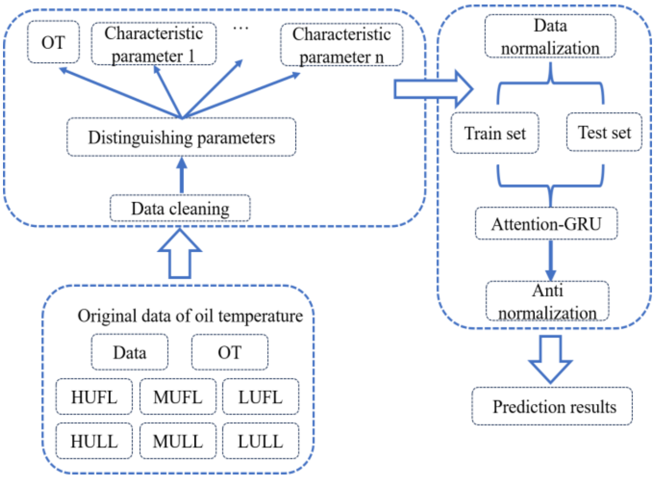


Fig. 4. The Prediction flow chart

Step 1 Model obtains input data of the oil temperature, searches for missing values, and processes time-related data. Then it distinguishes characteristic parameters and selects the target parameter oil temperature of the transformer. The features in the sample cannot be directly passed down when they are input into the model because the absolute values have a large range of variation. Instead, these feature values are processed and then used as the vectors to be passed. The purpose of data processing is to ensure that there are no special samples in the input samples of the network, and to ensure that the range of eigenvalues is between 0 and 1, so that the convergence of the network is not compromised and the speed of processing the data is increased. This paper uses the Min-Max normalization to pre-process the sample data, and the equation is as follows:

$$\bar{X} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{13}$$

where $\bar{X}$ is the normalized sample data, X is the original sample data, $X_{\min}$ is the minimum data of sample, and $X_{\max}$ is the maximum data of sample.

Step 2 We utilize the sliding window processing technique to generate sample data for model training input based on the pre-processed oil temperature data. We divide 0.8 and 0.2 of the dataset into the training set and the test set.

Step 3 To refine oil temperature prediction accuracy and highlight the effectiveness of characteristic indicators, the attention mechanism is incorporated into the GRU model. The Attention-GRU model and its hyperparameters are constructed. The entire model is trained using the training set, and the evaluation indicators between true results and predicted results are calculated. The optimal model parameters are determined through multiple experiments.

Step 4 After determining the ideal model, the test samples will be predicted. After inverse normalization, model obtains the oil temperature values with physical significance at the corresponding moments. Model compares them with the true results and calculate the error evaluation indicators to verify the fitting and generalization capacity.

III. Experimentations and Analysis

In this study, this paper proposes an Attention-GRU prediction method for predicting oil temperature of transformers. To comprehensively verify the effectiveness of this model, experimental comparisons will be carried out between it and several common algorithm models, and the effectiveness of adding the Attention Mechanism will be proved through ablation experiments.

A. Experimental Dataset

A substantial number of operation state monitoring quantities exist for power transformers. These parameters play a crucial role in comprehensively assessing the working condition of power transformers. Some state quantities have a relatively strong correlation with the oil temperature and can accurately reflect the changes in the top-layer oil temperature of the transformers. The Zhou team collected real data over two years and established a power transformer dataset (ETDataset) for researching and predicting the oil temperature of transformers [25]. This dataset consists of three parts in total: ETT-small, ETT-large, and ETT-full. This paper uses the ETT-small dataset for experiments. It collects data every hour from two transformers in two substations to form ETTh1 and ETTh2 datasets. Each dataset contains 17,421 pieces of data, and each data is composed of eight features. The meanings of each feature are shown in Table Ⅰ. The Attention-GRU model utilizes six distinct types of external power supply load characteristics as the input elements for the network, with the oil temperature serving as the output variable.

TABLE I
THE EXPERIMENTAL FEATURE MEANING TABLE

Field	Description
DATE	The Recorded Date
HUFL	High Useful Load
HULL	High Useless Load
MUFL	Middle Useful Load
MULL	Middle Useless Load
LUFL	Low Useful Load
LULL	Low Useless Load
OT	Oil Temperature

"DATE" represents the date and time when the oil temperature is recorded. "HUFL" stands for the highest useful load, generally referring to the highest load of the transformer when it is working under load. "HULL" represents the highest useless load, generally referring to the highest load of the transformer when it is not working under load. "MUFL" stands for the intermediate useful load, generally referring to the intermediate value of the transformer load when it is working under load. "MULL" represents the intermediate useful load, generally referring to the intermediate value of the transformer load when it is not working under load. "LUFL" stands for the lowest useful load, generally referring to the lowest load of the transformer when it is working under load. "LULL" represents the lowest useful load, generally referring to the highest load of the transformer when it is not working under load. "OT" represents the transformer oil temperature detected.

B. Experimental Details

The experiment in this paper is to build a deep learning environment on a computer with AMD Ryzen7 5800H central processing unit and NVIDIA GeForce RTX 3060 laptop graphics processing unit, as shown in Table II. The batch size is 32. The learning rate is 0.01 The Adam algorithm is used as the optimizer. The epoch is 300 and the prediction length is 12.

TABLE II
THE DEEP LEARNING ENVIRONMENT CONFIGURATION

Environment	Configuration
Anaconda	3.0
CUDA	11.5
cuDNN	8.2.4
Python	3.9
Tensorflow	1.11.0
Keras	2.2.4
Numpy	1.19.5
Pandas	1.3.5
Sklearn	1.0.2

C. Evaluation Indicators

The modeling compares true value y_i and predicted value $\hat{y}_i$ on the training set and test set obtained from same dataset in the same division way. To compare prediction accuracy of Attention-GRU model with that of other models, this paper uses the mean squared error, the mean absolute error, and the R-Square to evaluate the prediction results among various methods.

1) The mean square error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{14}$$

where this value range of the MSE is $\left[0,+\infty\right)$. It represents the average squared error between predicted and actual values across all data points. The smaller this value is, the closer the predicted values are to the true values, that is, the better the prediction effect of the model is.

2) The mean absolute error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{15}$$

where this value range of the MAE is $\left[0,+\infty\right)$. It is the average value of the absolute errors and reflects the actual situation of the prediction value errors. The smaller this value is, the closer the predicted values are to the true values, that is, the better the prediction accuracy of the model is.

3) The R-Square (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2} \tag{16}$$

where $\hat{y}_i$ is the average of the actual values. This value range of the R^2 is $\left(-\infty,1\right)$. It embodies the model's fitting quality. The larger this value, the better the model fitting ability. The R^2 is closer to 1 indicates that the fit is more perfect and the prediction is better.

D. Comparative Experiment

To verify the effect of Attention-GRU, it was compared with other models on the ETTh1 dataset and the ETTh2 dataset. Table III and Table IV show the comparison of predicted indicators between Attention-GRU and existing models on the ETTh1 dataset and ETTh2 dataset. The prediction error of RNN model is the largest, and it performs poorly in this experiment. The prediction error of Transformer model is the smallest relative to LSTM, GRU, and BiLSTM. But Attention-GRU has lower MAE and MSE values. On the ETTh1 dataset, the model presented in this paper yielded an MSE of 5.2664, an MAE of 1.6219, and attained an R² of 0.9251. On the ETTh2 dataset, the same model produced identical performance metrics, with an MSE of 13.5504, an MAE of 2.6049, and an R² of 0.9026. These results demonstrate superior performance compared to other models evaluated on both datasets. Specifically on dataset ETTh1, the MAE value of the Attention-GRU model decreased by 2.21%, the MSE decreased by 4.79%, and the R^2 increased by 1.40% against those of the Transformer model. On dataset ETTh2, The MAE and MSE value of the Attention-GRU model decreased by 1.61% and 8.11% respectively, and the R^2 value increased by 1.43% as opposed to the Transformer model. Compared with other models, the attention-GRU model exhibits significant improvement. Fig. 5(a) and Fig. 5(b) visually illustrate the performance differences among various models on the ETTh1 dataset and the ETTh2 dataset across different evaluation metrics. The

MSE and MAE values of BiLSTM, Transformer, and Attention-GRU are relatively close to each other, indicating that these models achieve comparable and relatively high prediction accuracy. However, Attention-GRU demonstrates a higher R^2 score, suggesting superior fitting capability compared to BiLSTM and Transformer. In contrast, the remaining models exhibit larger MSE and MAE values, implying lower prediction accuracy and weaker fitting performance.

The predictive performance of different models is evaluated by comparing their oil temperature forecasting curves against the ground truth measurements on the ETTh1 and ETTh2 test dataset, as showed in Fig. 6(a) and Fig. 6(b) respectively. Despite the significant fluctuations in oil temperature throughout the period, the predicted values of the models generally manage to closely track the trend of the actual values. However, there are notable discrepancies in some individual predictions made by certain models. The prediction results of the Attention-GRU model have a higher degree of fitting with the actual values. Especially in the peak and trough regions of the oil temperature curve, it is closer to the actual oil temperature values. Moreover, the prediction curve caused by the Attention-GRU model basically coincides with the measured curve, and this model can predict the future change trend of the oil temperature more accurately. However, oil temperature prediction ability of other models is relatively weak. Based on the above analysis, Attention-GRU model prediction results are in line with the trend of the oil temperature data.

E. Ablation Experiment

The results of the comparative experiments demonstrate that the RNN model fails to achieve a satisfactory level of alignment between predicted and true values, suggesting its relatively weak predictive performance. One possible reason is that the structure of RNN is relatively simple, which limits its modeling and representation capabilities when dealing with high-dimensional data with strong nonlinear characteristics. As a result, it struggles to fully capture complex patterns in the data. Additionally, RNN suffers from slow convergence, a time-consuming training process, and a tendency to get trapped in local optima. In the ablation experiment, this paper compared the difference of Attention-RNN with RNN on the ETTh1 dataset and ETTh2 dataset to assess attention impact.

As shown in Tables V and Tables VI, contrasting with the RNN model on the ETTh1 dataset, the MAE and MSE of Attention-RNN decreased by 17.46% and 21.95% respectively, and R^2 increased by 4.74%. Contrasting with the RNN model on the ETTh2 dataset, the MAE and MSE of Attention-RNN decreased by 30.86% and 51.21% respectively, and R^2 increased by 20.5%. Attention-RNN model has higher prediction accuracy than RNN model.

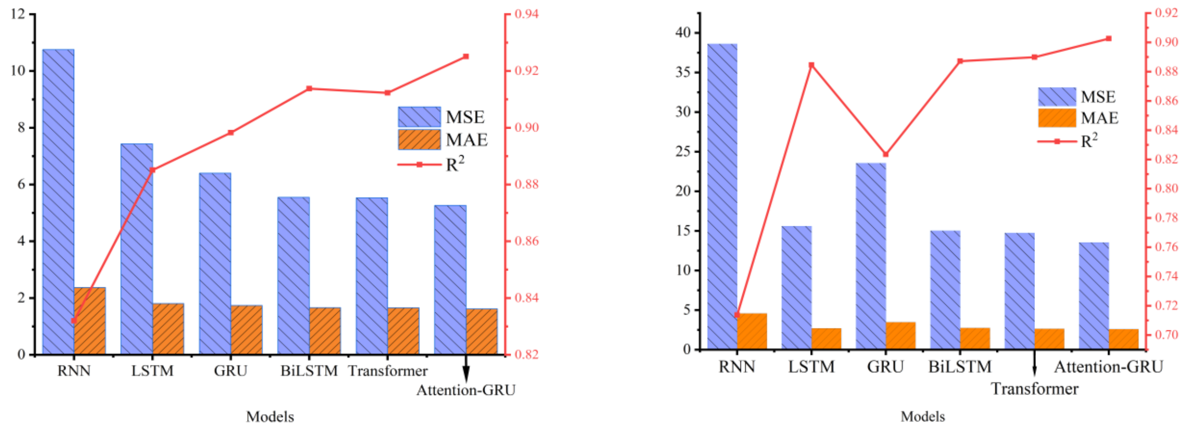
As shown in Fig. 7, the predicted values of the Attention-RNN converge toward ground truth than the RNN after adding the attention mechanism. This result indicates that the attention mechanism focuses on the importance of data features, which will improve the predictive performance.

TABLE III
THE PERFORMANCE COMPARISON OF MODELS ON THE ETTH1 DATASET

Models	MAE	MSE	R^2
RNN	2.3715	10.7561	0.8320
LSTM	1.8054	7.4328	0.8851
GRU	1.7394	6.4016	0.8983
BiLSTM	1.6639	5.5513	0.9138
Transformer	1.6585	5.5316	0.9123
Attention-GRU	**1.6219**	**5.2664**	**0.9251**

TABLE IV
THE PERFORMANCE COMPARISON OF MODELS ON THE ETTH2 DATASET

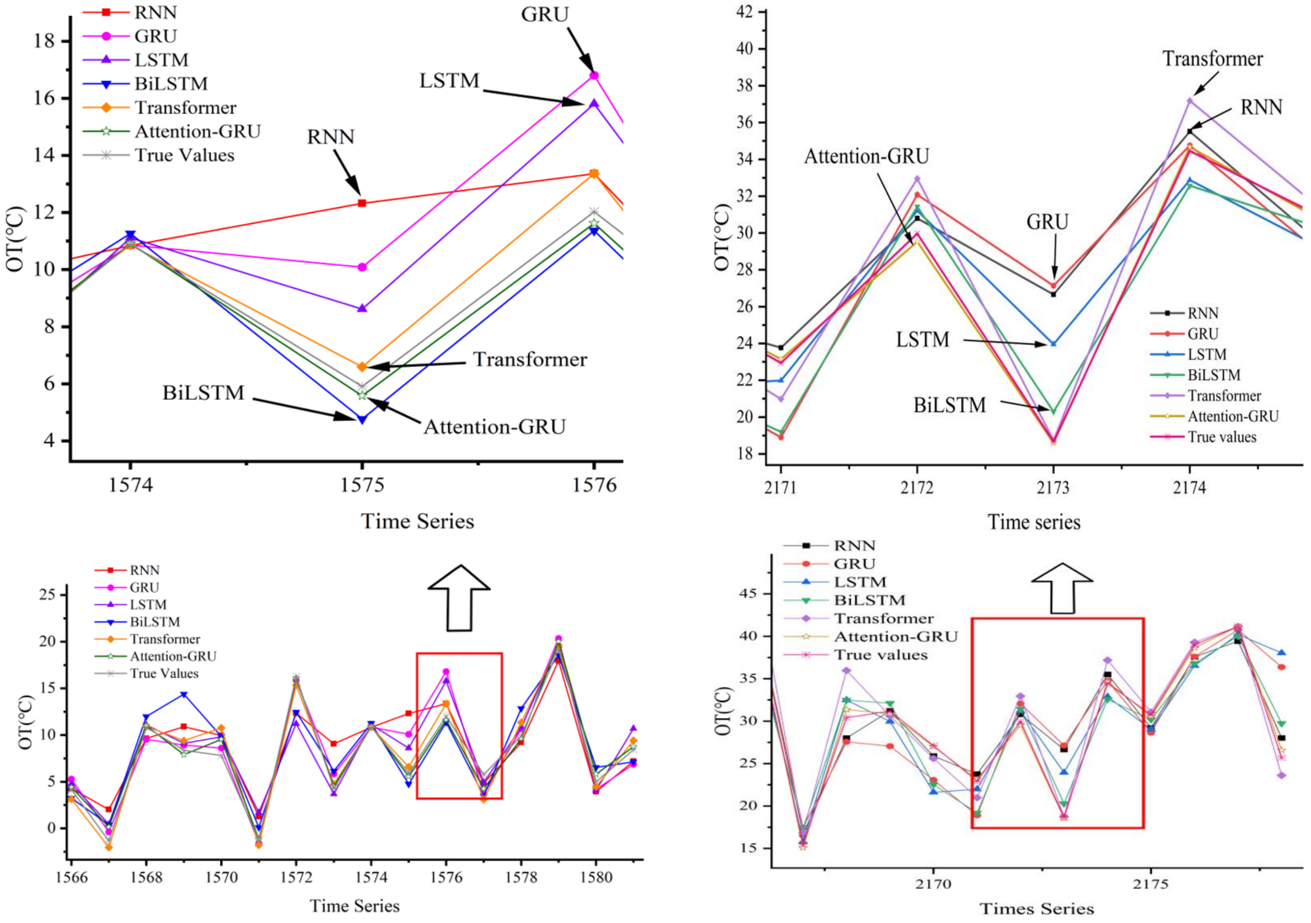
Models	MAE	MSE	R^2
RNN	4.5871	38.6155	0.7137
LSTM	2.7159	15.6125	0.8846
GRU	3.4901	23.5654	0.8235
BiLSTM	2.7545	15.0237	0.8872
Transformer	2.6476	14.7475	0.8899
Attention-GRU	**2.6049**	**13.5504**	**0.9026**



(a) On the ETTh1 dataset

(b) On the ETTh2 dataset

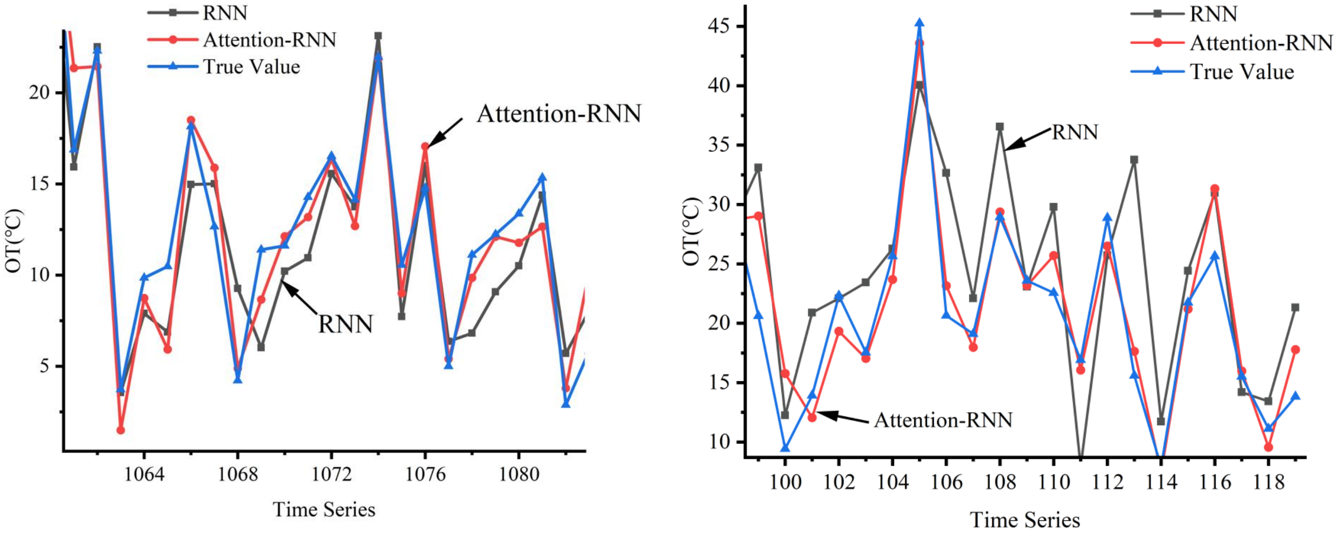
Fig. 5. The comparison of evaluation indicators on different datasets



(a) On the ETTh1 dataset

(b) On the ETTh2 dataset

Fig. 6. The comparison of prediction results between general regression models and Attention-GRU on different dataset



(a) On the ETTh1 dataset

(b) On the ETTh2 dataset

Fig. 7. The prediction results of models with the attention mechanism

TABLE V
THE ABLATION EXPERIMENT ON THE ETTH1 DATASET

Models	MAE	MSE	R^2
RNN	2.3715	10.7561	0.8320
Attention-RNN	**1.9584**	**8.3945**	**0.8715**

TABLE VI
THE ABLATION EXPERIMENT ON THE ETTH2 DATASET

Models	MAE	MSE	R^2
RNN	4.5871	38.6155	0.7137
Attention-RNN	**3.1717**	**18.8415**	**0.8600**

IV. CONCLUSION

This paper contributes to the prediction of transformer oil temperature. The proposed Attention-GRU incorporates the autonomous learning ability of the self-attention mechanism into the time characteristics of GRU. It makes up for the deficiency that ordinary recurrent neural networks cannot perform long-term learning and utilize context information. It enhances the adaptability of the model to complex data and improves the prediction accuracy of the model, reducing the adverse impact of noisy data. Meanwhile, through experiments on the ETTh1 dataset and the ETTh2 dataset, it is proved that the Attention-GRU model proposed in this paper has higher accuracy than RNN, LSTM, GRU, BiLSTM, and Transformer models in different datasets. The MAE and MSE are lower than those of the in comparison with other models, and the R^2 is higher than that of the in comparison with other models, indicating the good prediction performance of this model. It has guiding significance for controlling the operating state of transformers and ensuring the safe operation of transformers. On the other hand, how to effectively integrate resources and make full use of abundant meteorological data to create larger datasets with higher quality, longer time span, and richer features to further improve the prediction accuracy is a problem to be considered in the next step.

REFERENCES

[1] X. G. Xu, Y. W. He, X. Li, F. D. Peng, and Y. Xu, “Overload capacity for distribution transformers with natural ester immersed high tempera ture resistant insulating paper,” *Power System Technology*, vol. 42, no. 3, pp. 1001-1006, 2018.

[2] F. L. Tan, G. Xu, and P. Zhang, “Research on top oil temperature prediction method of similar day transformer based on topsis and entropy method,” *Electric Power Science and Engineering*, vol. 37, no. 6, pp. 62-69, 2021.

[3] Helene Canot, Philippe Durand, Emmanuel Frenod, Dariush Ghorbanzadeh, and Valerie Nassiet, "Neural Networks NARX for Durability Bonded Joint Prediction," Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2021, 20-22 October, 2021, Hong Kong, pp13-17

[4] Qasim Abdul-Aziz, and Hassan H. Hashemi, "Flight Control System using Neural Networks," Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2021, 7-9 July, 2021, London, U.K., pp73-78

[5] N. K. Micheni, E. B. Atitwa, and P. M. Kimani, “Prediction of Domestic Value-Added Tax in Kenya Using SARIMA and Holt-Winters Methods,” *IAENG International Journal of Applied Mathematics*, vol. 55, no. 3, pp. 594-602, 2025.

[6] S. Purwani, R. A. M. Fasa, A. Tsanawafa, and S. Sutisna, “Long-Term Prediction of Oil Palm Fresh Fruit Bunch Prices in Riau Province Post-Pandemic Using a Discrete-Time Markov Chain,” *IAENG International Journal of Applied Mathematics*, vol. 54, no. 6, pp. 1225-1232, 2024.

[7] X. Q. Dong, L. T. Jing, R. Tian, and X. Y. Dong, “Prediction method of transformer top oil temperature based on LSTM model,” *Journal of Electric Power*, vol. 38, no. 01, pp. 38-45, 2023.

[8] Y. An, Z. J. Zhu, H. Y. Wang, M. L. Cai, B. Liu, and Q. Chen, “A thermal evaluation method of distribution transformers,” *High Voltage Apparatus*, vol. 56, no. 4, pp. 48-54, 2019.

[9] J. *Zhao,* S. G. Gao, R. D. He, C. Y. Xiang, Y. F. Rui, Y. L. Wang, and Y. Yin, “Hyper-parameters optimization method for multi-step prediction of acetylene in power transformer oil by Gated Cyclic Unit Network,” *High Voltage Apparatus*, vol. 60, no. 7, pp. 0163-0172, 2024.

[10] Z. C. Tan, J. M. Fan, L. T. Feng, W. J. Mo, M. W. Zhong, “Prediction method of dissolved gas in transformer oil based on modal decomposition and hybrid CNN-GRUT,” *Advanced Technology of Electrical Engineering and Energy*, vol. 43, no. 7, pp. 80-90, 2024.

[11] Y. P. Wei, Y. H. Su, S. L. Wang, X. Zhang, and H. Wang, “Prediction of gas concentration in transformer oil based on improved gated recurrent unit,” *Electrical Engineering*, vol. 23, no. 2, pp. 55-60, 2022

[12] C. Y. Zhou, Y. J. Li, G. W. Deng, Y. N. Liu, and D. Y. Yu, “Multi-dimensional analysis and early warning model of converter transformer,” *Journal of Electrical Engineering*, vol. 15, no. 4, pp. 151-158, 2021.

[13] H. Meng, T. Zhang, J, Wang, J. Y. Zhang, D. Li, and G. R. Shi, “Multi-node Short-term Power Load Forecasting Method Based on Multi-scale Spatiotemporal Graph Convolution Network and Transformer,” *Power System Technology*, vol. 48, no. 10, pp. 4297-4311. 2024.

[14] Q. C. Fan, F. Yu, and M. Xuan, “Power transformer fault diagnosis based on optimised Bi-LSTM model,” *Computer Simulation*, vol. 39, no, 11, pp. 136-140, 2022.

[15] X. Wang, H, N, Rong, J, Wang, and G. X. Ge, “Concentration prediction of dissolved gases in transformer oil based on wavelet decomposition and long short-term memory network,” *Electric Engineering*, no. 9, pp. 24-29,33, 2020.

[16] Z. Y. Rao, C. Y. Feng, and W. J. Liu, “Intelligent road icing early warning system based on machine learning,” *Engineering Letters,* vol. 32, no. 4, pp. 806-811, 2024.

[17] S. Hochreiter, “The vanishing gradient problem during learning recurrent neural nets and problem solutions,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 2, pp. 107-116, 1998.

[18] A. S. Girsang and D. Tanjung, “Fast genetic algorithm for long short-term memory optimization,” *Engineering Letters*, vol. 30, no. 2, pp. 528-536, 2022.

[19] C. D. Zhou, B. Q. Zeng, S. Y. Wang, and Q. Shang, “Chinese summarization research on combination of local attention and convolutional neural network,” *Computer Engineering and Applications*, vol. 55, no. 8, pp. 132-137, 2019.

[20] K. Cho, B. V. Merrieenboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” *Conference on empirical methods in natural language processing*, pp. 1724-1734, 2014.

[21] F. G, Y. L. Su, X. H. Niu, Y. P. Zhao, T. T. Fan, and Q. D. E. J. Ren, “Mongolian-Chinese neural machine translation based on transformer,” *Computer Applications and Software*, vol. 37, no. 2, pp. 141-146,225, 2020.

[22] J. H. Chen, H. Dai, S. L. Wang, and R. Cheng, “Improving accuracy and efficiency in time series forecasting with an optimized transformer model,” *Engineering Letters*, vol. 32, no. 1, pp. 1-11, 2024.

[23] S. Hao, D. H. Lee, and D. Zhao, “Sequence to sequence learning with attention mechanism for short-term passenger flow prediction in large-scale metro system,” *Transportation Research Part C: Emerging Technologies*, vol. 107, pp. 287-300, 2019.

[24] Z. Hu, L. Wang, Y. Luo, Y. Xia, and H. Xiao, “Speech emotion recognition model based on attention CNN Bi-GRU fusing visual information,” *Engineering Letters*, vol. 30, no. 2, pp. 427-434, 2022.

[25] H. Y. Zhou, S. H. Zhang, J. q. Peng, S. Zhang, J. X. Li, H. Xiong, and W. C. Zhang, “Informer: beyond efficient transformer for long sequence time-series forecasting,” *Artificial Intelligence*, vol. 35, no. 12, pp. 11106-11115, 2021.