The Transformer Oil Dissolved Gas Prediction Method Based on Multivariate Variable-Weighting Mechanism

Zhu Ye[1,2], Zhou Zhengqin[1,2,†], Yang Yi[3], Zhan Hao[1,2], Li Mengqi[1,2], Zhou Wen[1,2]
[1] Wuhan NARI Co Ltd., State Grid Electric Power Research Institute,Wuhan Hubei 430206, China,
[2] Nanjing NARI Group Corp., State Grid Electric Power Research Institute, Nanjing Jiangsu 211000, China
[3] State Grid Shandong Electric Power Research Institute, Shandong Jinan 250002, China
[†] Corresponding author: zhou_zhengqin@163.com

***Abstract*—Dissolved gases in transformer oil are important diagnostic indicators of transformer operating conditions. To overcome the limitations of traditional single models in fusion of multi-gas signal information and dynamic response capabilities, this paper proposes an integrated modeling method based on a multivariate variable-weighting mechanism. First, based on the time-series characteristics and nonlinear variation of dissolved gas data, autoregressive moving average (ARMA), gray prediction model (GM), and Transformer-based deep time-series modeling methods are constructed as base prediction models. These models build a multivariate prediction structure from the perspectives of linear trend modeling, gray system approximation, and complex time-series feature extraction. Then, a variable-weighting distribution module based on long short-term memory (LSTM) network is designed, which dynamically outputs the optimal combination weight over time by learning the signal features. Finally, the prediction results of each base model are integrated according to the weights, forming a multivariate variable-weighting combination prediction model. Validation using large-scale real-world data covering multiple substations and operating conditions shows that the proposed multivariate variable-weighting combination prediction model delivers high accuracy and stability.**

***Index Terms*—dissolved gas in transformer oil, Transformer, long short-term memory network, model ensemble**

I. Introduction

Ultra-high voltage transformers are essential for voltage conversion, long-distance transmission, and renewable energy integration in modern power systems. Given their complex structures and harsh operating conditions, failures such as partial discharges, winding short-circuits, and insulation degradation pose serious risks to grid stability. These faults often manifest as abnormal increases in dissolved gases within transformer oil—especially hydrogen (H_2), methane (CH_4), ethane (C_2H_6), ethylene (C_2H_4), acetylene (C_2H_2), carbon monoxide (CO), and carbon dioxide (CO_2). For example, elevated hydrogen and acetylene levels typically indicate arcing, while increases in methane and ethylene often signal overheating. Dissolved Gas Analysis (DGA) is thus widely used for transformer condition monitoring and fault diagnosis.

This work was supported by the Science and Technology Project of State Grid Corporation of China (No. 5108-202218280A-2-350-XG).

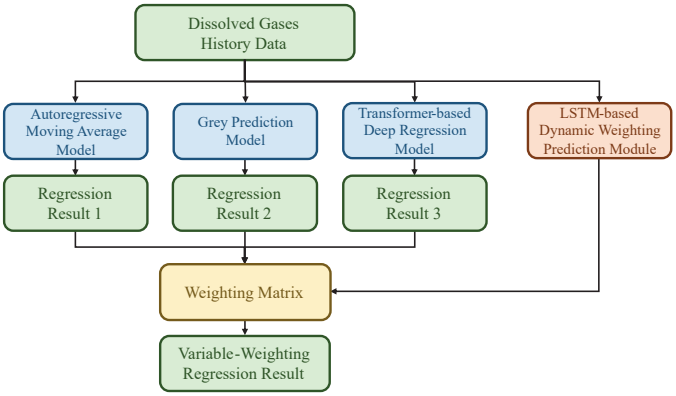

Fig. 1. **Overview of the proposed multivariate variable-weight forecasting framework.** It combines ARMA, GM(1,1), and a Transformer-based model, with an LSTM-based module adaptively assigning dynamic weights for ensemble prediction.

Accurate forecasting of gas trends enables early fault detection and supports preventive maintenance, thereby improving grid safety and operational reliability.

In recent years, substantial research has focused on forecasting dissolved gas concentrations in oil-immersed transformers. Modeling approaches have mainly followed three technological trajectories: statistical modeling, traditional machine learning, and deep learning. These methods address key challenges of DGA time-series data—including nonlinearity, non-stationarity, and high-dimensional coupling—by progressing from linear assumptions to data-driven strategies and from univariate modeling to multi-scale fusion, reflecting a shift toward more intelligent diagnostic frameworks.

Statistical methods remain valuable in practical engineering. For example, [1] combined grey relational analysis with Gaussian Process Regression to model inter-gas dependencies while reducing noise. [2] introduced a seasonally adjusted SARIMA model with external environmental factors to improve stability. While effective for short-term prediction, such methods struggle to capture complex temporal dynamics.

With deep learning advances, models like DBN [3], [4], LSTM [5]–[7], and GRU [8] have demonstrated stronger abili-

ties in nonlinear and temporal modeling. Hybrid methods, such as CEEMDAN-DBN-ELM [9], further enhanced performance via feature decomposition and hierarchical representation.

However, from the perspective of practical application, several notable issues remain in existing approaches. First, single models often struggle to accommodate the high-dimensional heterogeneity inherent in multi-gas data, typically exhibiting only local superiority under specific conditions and lacking robust generalization capabilities. Second, although ensemble strategies have become a prominent means to improve prediction accuracy, most existing methods rely on static weighting schemes, which fail to capture the temporal responsiveness and performance variation across models, thereby limiting the effectiveness of fusion. Moreover, while some deep learning models possess strong feature extraction abilities, they often fall short in capturing linear trends and medium-to-short-term perturbation structures embedded in raw time series data, leaving room for performance improvement.

To address these challenges, this study proposes an ensemble modeling method based on a multivariate dynamic weighting mechanism, aiming to enhance both prediction accuracy and the model's generalization and adaptive capabilities. As illustrated in Fig. 1, the proposed method integrates three heterogeneous base models—namely, the Autoregressive Moving Average (ARMA) model [10], the Grey Model (GM) [11], and a Transformer-based deep temporal model—forming a multi-perspective prediction system that jointly captures linear trends, fuzzy approximations, and complex dynamic dependencies. In addition, a Long Short-Term Memory (LSTM)-based dynamic weighting module is introduced to adaptively adjust the contribution of each base model over time. This results in a fusion forecasting framework characterized by multi-level structure, cross-model coupling, and feature complementarity. Extensive empirical evaluations confirm that the proposed method consistently achieves superior accuracy and robustness across different gas types and operational scenarios, demonstrating its practical potential for intelligent transformer condition assessment and fault prediction.

II. Methodology

A. Autoregressive Moving Average Model

The autoregressive moving average model is a classical and widely applied approach in time series analysis, primarily used to model the linear relationships of stationary time series data. The ARMA model combines the autoregressive (AR) process and the moving average (MA) process, achieving fitting and forecasting of time series by linear regression on historical data and smoothing of random errors.

The AR process assumes a direct linear dependency between the current value and its past observations, which can be expressed as:

$$X_t = c + \sum_{i=1}^{P} \Phi_i X_{t-i} + \epsilon_t \tag{1}$$

where X_t is the observed value at time t, c is a constant term, Φ_i are the autoregressive coefficients, X_{t-i} denotes the lagged observations up to lag i, and ϵ_t is a white noise term with zero mean. The MA process models the current value as a linear combination of past white noise error terms, described by:

$$X_t = \mu + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \tag{2}$$

where μ is the mean of the series, θ_i are the moving average coefficients, ϵ_t and ϵ_{t-i} represent current and past white noise error terms respectively. This process adjusts the current output by weighting historical disturbances, enhancing the model's capability to capture short-term fluctuations. The complete ARMA model integrates both processes as:

$$X_t = c + \sum_{i=1}^{P} \Phi_i X_{t-i} + \sum_{j=1}^{q} \theta_i \epsilon_{t-j} + \epsilon_t \tag{3}$$

B. Grey Prediction Model

The grey prediction model is capable of extracting the latent trend of dissolved gas concentrations through an accumulated generating operation. As a core component of grey system theory, it demonstrates excellent performance in modeling behaviors and forecasting trends in systems characterized by small sample sizes and incomplete information. In this study, the GM(1,1) model is introduced as one of the base submodels. Let the original non-negative time series be denoted as:

$$X^{(0)} = \{x^{(0)}(1), x^{(0)}(2), ..., x^{(0)}(n)\}, n \geq 4 \tag{4}$$

An accumulated generating operation is first applied to the raw sequence to obtain a new series:

$$\begin{gathered} X^{(1)} = \{x^{(1)}(1), x^{(1)}(2), ..., x^{(1)}(n)\} \\ x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i) \end{gathered} \tag{5}$$

To smooth the input for model construction, a neighboring mean series $Z^{(1)}$ is constructed:

$$Z^{(1)}(k) = \frac{1}{2}\left(x^{(1)}(k) + x^{(1)}(k-1)\right), k = 2, 3, ..., n \tag{6}$$

A first-order linear differential equation is then formulated as:

$$\frac{dx^{(1)}(t)}{dt} + ax^{(1)}(t) = b \tag{7}$$

where a is the development coefficient, reflecting the rate of change of the system variable over time, and b is the grey input, indicating the influence of external inputs or baseline trends on the system behavior. This equation can be rewritten in matrix form as:

$$\begin{gathered} Y = B \cdot \theta \\ Y = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(n) \end{bmatrix}, B = \begin{bmatrix} -Z^{(1)}(2) & 1 \\ -Z^{(1)}(3) & 1 \\ \vdots & \vdots \\ -Z^{(1)}(n) & 1 \end{bmatrix}, \theta = \begin{bmatrix} a \\ b \end{bmatrix} \end{gathered} \tag{8}$$

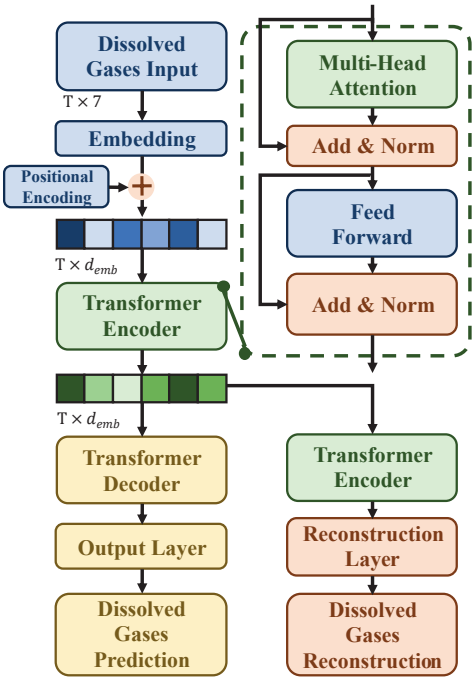

Fig. 2. **Architecture of the Transformer-based regression model.** It consists of a shared encoder and two parallel branches: a regression branch for target forecasting and a reconstruction branch for auxiliary sequence modeling.

The parameter vector θ can be estimated using the least squares method:

$$\hat{\theta} = \left(B^T B\right)^{-1} B^T Y \tag{9}$$

Once the parameters a and b are obtained, the predicted values of the accumulated series can be computed as:

$$\hat{x}^{(1)}(k+1) = \left(x^{(0)}(1) - \frac{b}{a}\right) e^{-ak} + \frac{b}{a} \tag{10}$$

Finally, the predicted values of the original sequence are recovered by:

$$\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k) \tag{11}$$

C. Transformer-based Deep Regression Model

With the increasing complexity of industrial monitoring data and the escalating demands for prediction accuracy, traditional time series modeling methods have gradually exhibited limitations in capturing nonlinearities, long-range dependencies, and multivariate couplings. Since its introduction, the Transformer model [12] has progressively become a mainstream framework for sequence modeling tasks. Leveraging a purely attention-based architecture, it dispenses with the reliance on conventional recurrent structures and demonstrates superior capabilities in capturing long-distance dependencies, enhancing parallel computation efficiency, and modeling complex dynamic relationships. In the domain of time series forecasting, the Transformer's self-attention mechanism effectively uncovers latent dynamic dependency structures within sequences. This global modeling capacity and strong feature representation ability provide significant advantages, especially when dealing with complex operating conditions.

The Transformer model employs a multi-head attention mechanism to model dependencies between arbitrary positions within the input sequence. This avoids the gradient vanishing and inefficient sequential propagation issues commonly faced by traditional recurrent neural networks such as LSTM and GRU when handling long sequences. The mechanism calculates attention weights between each time step in the input sequence, thereby enabling efficient extraction of global semantic information. Its mathematical formulation is expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \tag{12}$$

Here, Q, K, and V represent the linearly projected query, key, and value vectors, respectively, and d_k denotes the dimensionality of the attention subspace. The multi-head attention mechanism performs the above attention operation in parallel across multiple subspaces, concatenating the results to produce the final output:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^O \tag{13}$$

Each attention head is computed as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{14}$$

where W_i^Q, W_i^K, and W_i^V are the projection matrices that map the original Q, K, and V into the i-th subspace.

To thoroughly capture the dynamic structural features and evolutionary trends within the dissolved gas concentration sequences, as well as to enhance the model's generalization ability and prediction robustness, this section proposes a dual-branch Transformer regression architecture incorporating multi-task learning principles. As illustrated in Fig. 2, the model consists of three components: a shared encoder, a reconstruction decoder, and a regression decoder.

In multivariate sequence modeling tasks, effectively extracting latent inter-variable dependencies significantly impacts both predictive performance and generalization capability. To this end, the embedding structure employed in this model maps multiple dissolved gas concentration signals into a unified high-dimensional representation space. Subsequently, the Transformer encoder performs global temporal modeling and joint learning of inter-variable couplings. The input sequence is defined as:

$$X = [x_1, x_2, \ldots, x_T] \in \mathbb{R}^{T \times d} \tag{15}$$

where $x_t = [x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(d)}] \in \mathbb{R}^d$ represents the dissolved gas observation vector at time step t, with $d = 7$ corresponding to seven dissolved gas signals.

To enhance the model's capacity for representing multi-gas time series and fully exploit their nonlinear interaction patterns, the gas signals are first processed by an embedding module composed of a three-layer Multi-Layer Perceptron (MLP) before being input into the Transformer encoder. This

module acts as a high-dimensional semantic transformer for the raw inputs:

$$Z = \mathrm{MLP}(X) \in \mathbb{R}^{T \times d_{\mathrm{model}}} \tag{16}$$

After embedding, the multi-gas embedded vectors Z are fed into the shared encoder module to extract the internal dynamic evolution patterns of the sequence. The encoder consists of L stacked Transformer encoder layers, each comprising multi-head self-attention and feed-forward network structures, supplemented with residual connections and layer normalization to improve training stability. The feature update at each layer is expressed as:

$$H^{(l)} = \mathrm{FFN}\big(\mathrm{MultiHead}(\mathrm{LN}(H^{(l-1)}))\big) + H^{(l-1)} \tag{17}$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization, with initial input $H^{(0)} = Z$. The final output of the shared encoder is $H_{\mathrm{enc}} = H^{(L)} \in \mathbb{R}^{T \times d_{\mathrm{model}}}$.

During the decoding phase, the model adopts two parallel task-specific branches: the reconstruction decoder and the regression decoder, which serve the goals of high-quality input sequence reconstruction and accurate future sequence prediction, respectively. Both branches share the temporal representations H_{enc} produced by the encoder but differ in architectural design and modeling objectives.

The reconstruction decoder employs a nonlinear mapping architecture composed of a Transformer decoder and an MLP, primarily used to compress and restore the high-dimensional semantic representation of the input sequence, facilitating the encoder's learning of key features during training:

$$\hat{X} = \mathrm{Decoder}_{\mathrm{rec}}(H_{\mathrm{enc}}) \in \mathbb{R}^{T \times d} \tag{18}$$

In contrast, the regression decoder adopts a standard Transformer decoder architecture with the primary goal of forecasting the multivariate gas concentration sequence over the next T' time steps. This decoder takes a placeholder sequence for the future time steps as its target-side input and leverages self-attention and cross-attention mechanisms to capture contextual dependencies within the time series and long-range dependencies from the encoder outputs. The final predicted output $\hat{Y}$ is given by:

$$\hat{Y} = \mathrm{Decoder}_{\mathrm{reg}}(Y_{\mathrm{init}}, H_{\mathrm{enc}}) \in \mathbb{R}^{T' \times d} \tag{19}$$

where Y_{init} denotes the initial placeholder sequence at the decoder input.

The two decoding branches are jointly trained by minimizing the Mean Squared Error (MSE) loss functions. The loss functions are defined as follows:

$$L_{\mathrm{rec}} = \frac{1}{Td} \sum_{t=1}^{T} \sum_{j=1}^{d} \left(x_t^{(j)} - \hat{x}_t^{(j)} \right)^2 \tag{20}$$

$$L_{\mathrm{reg}} = \frac{1}{T'd} \sum_{t=1}^{T'} \sum_{j=1}^{d} \left(y_t^{(j)} - \hat{y}_t^{(j)} \right)^2 \tag{21}$$

The overall loss function is the sum of the two losses:

$$L = L_{\mathrm{rec}} + L_{\mathrm{reg}} \tag{22}$$

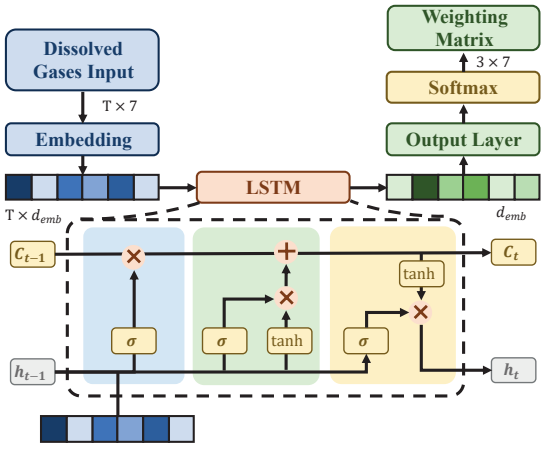

Fig. 3. **Illustration of the LSTM-based adaptive weighting prediction module.** It learns dynamic fusion weights for multiple base models by capturing temporal context, enabling time-varying contribution adjustment and improving overall prediction accuracy.

D. LSTM-based Dynamic Weighting Prediction Module

Although the three base models capture dissolved gas sequences from linear, fuzzy, and deep dynamic perspectives, fixed-weight fusion cannot reflect their varying effectiveness across different time steps, especially under complex and non-stationary industrial conditions. To address this issue, an LSTM-based dynamic weighting module is introduced to learn context-aware weight allocations, allowing adaptive model contributions and improving the flexibility and accuracy of the ensemble framework.

As an improved variant of recurrent neural networks (RNNs), LSTM mitigates the gradient vanishing problem inherent in traditional RNNs when processing long sequences by incorporating gating mechanisms, thus effectively capturing long-range dependencies. The core LSTM structure comprises a forget gate, input gate, and output gate, which are mathematically formulated as follows:

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \tag{23}$$

$$i_t = \sigma\left(W_i[h_{t-1}, x_t] + b_i\right) \tag{24}$$

$$\tilde{c}_t = \tanh\left(W_c[h_{t-1}, x_t] + b_c\right) \tag{25}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{26}$$

$$o_t = \sigma\left(W_o[h_{t-1}, x_t] + b_o\right) \tag{27}$$

$$h_t = o_t \odot \tanh(c_t) \tag{28}$$

where x_t denotes the input vector at time step t, h_t the current hidden state, and c_t the cell memory state; $\sigma(\cdot)$ is the sigmoid activation function; $\odot$ denotes element-wise multiplication; W_f, W_i, W_c, W_o and b_f, b_i, b_c, b_o are model parameters.

Considering that the ensemble performance depends on the individual base models' predictive capabilities under specific inputs, this section applies LSTM to learn a global contextual representation of the time series, which is then used to generate a dynamic fusion weight matrix for the three base models, as

illustrated in Fig. 3. Specifically, the embedded input sequence is denoted as $X \in \mathbb{R}^{T \times d_{\text{emb}}}$, where T is the sequence length and d_{emb} is the embedding dimension per time step. This sequence serves as input to the LSTM network, which updates its hidden states at each time step and finally outputs the hidden state at the last time step $h_T \in \mathbb{R}^{d_h}$, representing the global sequence embedding.

Subsequently, a set of fully connected layers project h_T into a weight matrix space of dimension 3×7, where the rows correspond to the three base models and the columns correspond to the seven gas variables. To ensure that the weights for the three models sum to 1 for each gas dimension, a Softmax function is applied to constrain the weight distribution:

$$W = \text{Softmax}(\text{MLP}(h_T)) \in \mathbb{R}^{3 \times 7}$$
$$\sum_{m=1}^{3} W_{m,j} = 1 \tag{29}$$

Finally, the predictions $\hat{y}_{t,m}^{(j)}$ for the j-th gas by the m-th base model at time step t are weighted and fused according to the dynamic weighting matrix to yield the final prediction output:

$$\hat{y}_t^{(j)} = \sum_{m=1}^{3} W_{m,j} \cdot \hat{y}_{t,m}^{(j)} \tag{30}$$

This mechanism effectively implements a globally-aware weighted function mapping that dynamically assigns higher weights to the most predictive models at each time step, overcoming the limitations of fixed fusion strategies in time series modeling. The LSTM-based dynamic weighting prediction module is also optimized using the MSE loss function.

III. Experiments

A. Experimental Protocol

Dataset. To comprehensively validate the proposed model's capability in modeling and generalizing over complex multivariate time series data, this study utilizes dissolved gas concentration monitoring data collected from transformer online monitoring systems across multiple regions in China. The dataset encompasses a total of 37 representative substations and 471 high-voltage transmission lines nationwide, with data acquisition spanning multiple recent years. This broad spatial and temporal coverage grants the dataset strong representativeness and practical applicability. Each transmission line is equipped with high-precision sensor devices that record dissolved gas concentrations at an hourly frequency. It includes seven key gas components: H_2, CH_4, C_2H_2, C_2H_4, C_2H_6, CO, and CO_2. After cleaning and normalization, over 2.56 million valid samples were obtained.

To align with model input requirements, a sliding window of length 32 and stride 16 was used to generate input-output pairs. The dataset was split into training, validation, and test sets at an 8:1:1 ratio, with the best validation model used for final testing.

Evaluation Metrics. To comprehensively assess the performance of the proposed forecasting model in multivariate time series prediction tasks, this study adopts three representative regression metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These indicators evaluate the deviation between the predicted values and the true observations from the perspectives of squared error, absolute error, and relative error, respectively. The definitions are as follows:

TABLE I
COMPARISON OF RMSE RESULTS OF DIFFERENT PREDICTION MODELS.

Model	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	CO	CO_2
LSTM	1.90	9.07	5.87	1.08	2.81	34.12	151.77
GRU	2.19	4.31	4.02	1.42	1.20	38.09	155.69
Transformer	2.15	2.90	4.38	0.69	1.78	23.19	80.23
Ours	**1.41**	**1.07**	**3.74**	**0.42**	**1.18**	**19.54**	**63.11**

TABLE II
COMPARISON OF MAE RESULTS OF DIFFERENT PREDICTION MODELS.

Model	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	CO	CO_2
LSTM	0.97	2.17	1.64	0.36	0.81	17.72	73.51
GRU	1.05	1.00	1.08	0.43	0.35	19.77	74.87
Transformer	0.84	0.72	0.50	0.21	0.51	9.02	32.27
Ours	**0.62**	**0.45**	**0.43**	**0.16**	**0.34**	**7.42**	**25.29**

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{31}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{32}$$

$$\text{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{33}$$

Here, $\hat{y}_i$ and y_i denote the predicted and true values of a single gas variable at time step i, and N represents the total number of samples.

Implementation Details. For model training, the Adam optimizer [13] is employed for all components. Both the deep regression model and the dynamic weighting prediction module are trained with an initial learning rate of 1×10^{-4}. The total number of training epochs is set to 500, with an early stopping strategy applied to prevent overfitting. The patience for early stopping is set to 50 epochs. The Transformer encoder in the deep regression model consists of 4 stacked layers, each with an embedding dimension of 128 and 8 attention heads. Both decoder branches (reconstruction and regression) are also composed of 4 stacked layers. The dynamic weighting prediction module is implemented with a single-layer LSTM, using an embedding dimension of 128. During training, the batch size is set to 256.

Baselines. To comprehensively evaluate the performance of the proposed multi-source dynamic weighting prediction model on the multivariate time-series regression task of dissolved gas concentrations in transformer oil, a series of

TABLE III
COMPARISON OF MAPE RESULTS OF DIFFERENT PREDICTION MODELS. (UNIT: %)

Model	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	CO	CO_2
LSTM	11.83	37.79	42.50	9.09	2.69	7.45	6.70
GRU	11.54	15.09	29.04	12.50	3.60	8.03	7.01
Transformer	10.41	10.70	9.24	4.77	2.78	4.61	3.40
Ours	**7.31**	**6.97**	**8.58**	**4.12**	**2.57**	**3.15**	**2.56**

comparative experiments were conducted. Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Transformer, which are widely adopted in sequence modeling, were selected as baseline models. To ensure fairness and reliability of the evaluation, all models were trained under identical configurations, with consistent input-output formats.

B. Evaluation Results

First, the RMSE comparison results are shown in Table I. The proposed model consistently achieved superior performance across all gas prediction tasks. In particular, for gases such as H_2, CH_4, and CO_2, the RMSE of the dynamic weighting model was significantly lower than that of the baseline models. For instance, in the case of CO_2, the proposed model attained an RMSE of 63.11, while LSTM and GRU reached 151.77 and 155.69 respectively. This indicates the proposed model's superior ability to capture the underlying data patterns and reduce prediction error. While the Transformer model also demonstrated promising accuracy across most gas types, it still fell short compared to the dynamic weighting approach. These results underscore the strong modeling capability and accuracy of our method in complex time-series regression tasks.

Similarly, as shown in Table II, the Mean Absolute Error (MAE) results further confirm the outstanding prediction performance of the dynamic weighting model. It not only effectively captures the central trend of the data but also maintains stable prediction performance across most time steps. Compared to the baseline models, which exhibited larger deviations for some gases, the proposed model achieved better consistency in prediction error across all variables. This demonstrates its capability to adapt to heterogeneous structures in multivariate data.

MAPE, as a dimensionless metric, reflects model performance from the perspective of relative error. The comparative results in Table III indicate that the proposed model consistently achieved significantly lower MAPE values across all gas types, outperforming LSTM and Transformer. In particular, for critical gases such as CH_4, CO, and CO_2, the MAPE values were consistently below 3%, highlighting the model's ability to maintain high precision. This not only ensures fair cross-variable comparability but also affirms the practical feasibility of our method in high-accuracy industrial applications.

In summary, the proposed multi-source dynamic weighting prediction model achieved remarkable performance improvements across all evaluation metrics. Its superior overall capability in modeling high-dimensional, strongly coupled multivariate time series validates its potential for practical deployment in dissolved gas monitoring and fault trend prediction in power transformer systems.

IV. CONCLUSION

This paper proposes a novel ensemble modeling approach based on a multivariate dynamic weighting mechanism for time-series prediction of dissolved gas concentrations in ultra-high-voltage transformer oil. By integrating linear, fuzzy approximation, and deep dynamic sub-models, a multi-perspective prediction framework is constructed. An LSTM-based dynamic weighting module adaptively adjusts model contributions over time, addressing challenges related to high-dimensional heterogeneity and dynamic complexity in gas data. Experiments on real-world datasets show that the proposed method consistently outperforms mainstream baselines, validating its effectiveness and adaptability in complex time-series scenarios. Future research will explore adaptive feature selection, lightweight deep networks, and graph-based modeling to further enhance multi-source data fusion and fault prediction under diverse operating conditions, contributing to more intelligent and efficient power equipment monitoring.

REFERENCES

[1] S. X. Lu, G. Lin, H. Que, M. J. J. Li, C. H. Wei, and J. K. Wang, "Grey relational analysis using gaussian process regression method for dissolved gas concentration prediction," *International Journal of Machine Learning and Cybernetics*, vol. 10, pp. 1313–1322, 2019.

[2] J. Liu, Z. Zhao, Y. Zhong, C. Zhao, and G. Zhang, "Prediction of the dissolved gas concentration in power transformer oil based on sarima model," *energy reports*, vol. 8, pp. 1360–1367, 2022.

[3] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[4] B. Qi, Y. Wang, P. Zhang, C. Li, and H. Wang, "A novel deep recurrent belief network model for trend prediction of transformer dga data," *IEEE access*, vol. 7, pp. 80 069–80 078, 2019.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] A. K. Lekshmi, D. S. Kumar, and K. S. Beevi, "Trend prediction of power transformers from dga data using artificial intelligence techniques," in *Communication and Intelligent Systems: Proceedings of ICCIS 2021*. Springer, 2022, pp. 1053–1065.

[7] X. Zhang, S. Wang, Y. Jiang, F. Wu, and C. Sun, "Prediction of dissolved gas in power transformer oil based on lstm-ga," in *IOP Conference Series: Earth and Environmental Science*, vol. 675, no. 1. IOP Publishing, 2021, p. 012099.

[8] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[9] W. Zeng, Y. Cao, L. Feng, J. Fan, M. Zhong, W. Mo, and Z. Tan, "Hybrid ceemdan-dbn-elm for online dga serials and transformer status forecasting," *Electric Power Systems Research*, vol. 217, p. 109176, 2023.

[10] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[11] D. Ju-Long, "Control problems of grey systems," *Systems & control letters*, vol. 1, no. 5, pp. 288–294, 1982.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.