An Anomaly Diagnosis and Prediction Method for Dissolved Gases in Transformer Oil Based on Multi-Expert Learning

Yang Xu[1,2], Zhu Ye[1,2], Zhang Fengda[1,2], Zhou Zhengqin[1,2,†], Li Mengqi[1,2], Luo Ziqiu[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
† Corresponding author: zhou_zhengqin@163.com

***Abstract*—As a core component of power grids, the operational status of oil-immersed transformers is directly linked to the safety and stability of the entire power system. Dissolved Gas Analysis (DGA) has been widely used to assess internal faults in transformers. However, existing methods often rely on single-model learning, which struggles to capture the complex couplings and nonlinear dynamics among multiple gas components, leading to limited prediction accuracy and anomaly detection capability. To address these limitations, this paper proposes a novel anomaly diagnosis and concentration prediction method for dissolved gases in transformer oil based on a multi-expert learning mechanism. The proposed system integrates three complementary expert models: the anomaly-aware expert identifies and repairs abnormal points in historical gas concentration sequences; the temporal modeling expert captures the dynamic evolution of gas concentrations over time; and the context-aware expert models the latent interactions among different gas components. A gated fusion mechanism is designed to adaptively assign weights to each expert according to the input context, enabling robust multi-dimensional feature integration and reliable prediction. Experimental results demonstrate that the proposed method outperforms traditional single-model approaches under various transformer operating conditions, offering strong support for equipment condition monitoring and intelligent maintenance.**

***Index Terms*—Dissolved Gas Analysis, Mixture of expert.**

I. Introduction

Oil-immersed transformers are critical to the stable operation of power grids [1], [2]. Under long-term high voltage, heavy load, and harsh environmental conditions, internal insulating materials such as transformer oil and paper gradually degrade, leading to thermal decomposition, electrical breakdown, and partial discharge. These processes generate gases like 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), which dissolve in transformer oil. Monitoring the concentration of these gases provides valuable insight into internal faults, making Dissolved Gas Analysis (DGA) one of the most effective diagnostic tools widely used for condition monitoring and maintenance in the power industry [3].

The accuracy of DGA-based diagnosis, however, depends heavily on data processing methods and modeling algorithms.

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

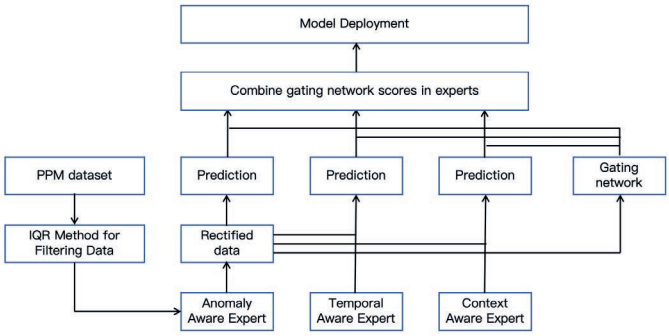

Fig. 1. Flowchart of the dissolved gas anomaly diagnosis and prediction method based on multi-expert learning.

Traditional approaches, such as rule-based methods, the Duval Triangle [4], [5], and Rogers Ratio [6], are simple but lack adaptability to complex scenarios, fluctuating gas concentrations, and multi-factor coupling effects. Rooted in static thresholds, these methods struggle to distinguish nuanced fault patterns—for example, gradual hydrogen accumulation versus abrupt spikes. Their rigidity often leads to misdiagnoses in non-stationary operating conditions.

Driven by the rise of data-driven methods, machine learning and deep learning have been increasingly applied to DGA. Models ranging from traditional algorithms like SVM, Decision Trees, and KNN to advanced deep learning methods such as CNN-GRUT [7] and EMD-gcForest [8] have shown improved fault detection capabilities. Nevertheless, key challenges persist. Dissolved gas data is a high-dimensional time series with frequent fluctuations and complex inter-gas relationships, making it difficult for single models to capture global patterns, local anomalies, and gas coupling effects simultaneously. Moreover, the presence of outliers often degrades model performance, and many existing studies overlook data anomalies.

To address these challenges, this paper proposes a dissolved gas anomaly diagnosis and concentration prediction method for oil-immersed transformers based on a mixture of expert [9], [10] learning mechanism. The core idea is

to integrate multiple expert modules with distinct modeling capabilities, enabling collaborative learning and information fusion to achieve deep understanding and accurate modeling of multi-component gases. Specifically, the anomaly-aware expert targets anomaly detection, evaluating potential abrupt changes in the time series and repairing/compensating abnormal data to ensure the reliability of subsequent modeling—a critical step for improving data quality and reducing error propagation. The temporal modeling expert focuses on mining the inherent evolutionary trends of gas concentrations over time, leveraging a structurally optimized Gated Recurrent Unit (GRU) network to enhance the model's perception of dynamic changes and long-term dependencies. Meanwhile, context-aware experts model interactive coupling relationships between different gases, constructing contextual semantics among gas components via a Bi-LSTM network to improve the recognition of local gas combination patterns.

To fully exploit the complementary advantages of these modules, a gating fusion mechanism is designed to dynamically assign weights to each expert's output. This ensures the final prediction not only captures global temporal modeling capabilities but also accounts for local anomaly detection and gas coupling relationships, with the gating module learning the importance of different experts across scenarios based on the input sequence to enhance adaptability and robustness.

Extensive experiments validate the method using operational data from multiple substations, covering diverse operating conditions, fault types, and time periods. Evaluations focus on regression prediction accuracy, anomaly detection capability, and stability under noisy data. Results demonstrate that the proposed multi-expert fusion model significantly outperforms traditional single-model methods and existing ensemble learning strategies across multiple metrics, enabling more accurate characterization of dissolved gas evolution trends and fault features in oil.

II. Proposed Method

A. Anormaly Aware Expert

In dissolved gas analysis (DGA) of transformer oil, factors such as sensor noise, environmental disturbances, or communication errors often lead to irregular anomalies in the historical gas concentration sequences, including sudden spikes, drifts, or missing values. These abnormal data points not only compromise the accuracy of fault diagnosis but may also mislead the training of subsequent predictive models. To effectively detect and repair such anomalies, this paper introduces an Anomaly-Aware Expert module, designed as a self-correcting sequence modeling network. Its purpose is to capture abnormal patterns within the time series and output reconstructed, "clean" gas sequences, thus providing more reliable input for the downstream expert modules.

The Transformer model [11], originally developed for natural language processing tasks, offers strong parallel computation capabilities and excels at modeling long-range dependencies. Compared with traditional recurrent neural networks, the Transformer processes the entire sequence simultaneously

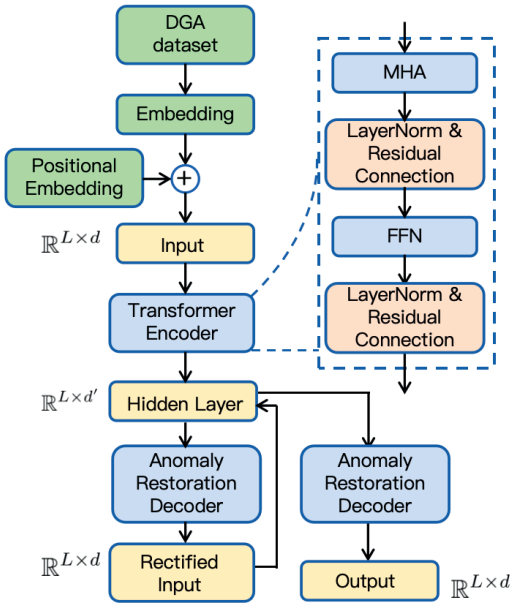

Fig. 2. Architecture diagram of the anormaly aware expert.

without relying on sequential calculations. Its self-attention mechanism enables flexible modeling of dependencies between different time steps, making it particularly suitable for capturing complex patterns in gas concentration sequences, such as abrupt changes, periodic fluctuations, and localized anomalies. Moreover, the Transformer architecture supports deep stacking and multi-head attention mechanisms, allowing the model to capture multi-scale features from various levels and perspectives.

1) Model Architecture Design: In our design, the Anomaly-Aware Expert is responsible for two critical tasks: Detecting and repairing abnormal fluctuations in the historical gas concentration sequence; Providing clean and reliable data for accurate future gas evolution prediction. To achieve this, we propose a dual encoder framework consisting of an encoder, anomaly repair detector, and future prediction detector, as illustrated in the following.

We first perform a linear transformation on the original gas concentration sequence, mapping each timestamp's gas concentration vector into a fixed-dimensional feature space, forming the initial embedding representation. The input oil-gas sequence is defined as:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L] \in \mathbb{R}^{L\times m} \tag{1}$$

Here, L represents the sequence length, and $m = 7$, which corresponds to the seven typical dissolved gases in transformer oil: H_2, CH_4, C_2H_6, C_2H_4, C_2H_2, CO, and CO_2. ze first project the input sequence into a high-dimensional embedding space using a linear transformation:

$$\mathbf{H}_0 = \mathbf{X}\mathbf{W}_0 + \mathbf{b} \in \mathbb{R}^{L\times d} \tag{2}$$

Where d is the embedding dimension used consistently across Transformer layers. Since the Transformer lacks inherent temporal ordering capability, we incorporate a fixed positional encoding $\mathbf{P}$ into the embedding, ensuring the model can distinguish both sequence order and relative positions:

$$\mathbf{H}_0' = \mathbf{X}\mathbf{W}_0 + \mathbf{P} \tag{3}$$

After the linear transformation and positional encoding, the features are passed to the encoder for further extraction and anomaly-aware representation learning.

2) Encoder Module: The role of the encoder is to project the gas sequence features into a high-dimensional space, capturing temporal dependencies and potential abnormal interactions hidden within the sequence. To fully model the nonlinear temporal variations and complex relationships between normal and abnormal patterns, we adopt a Transformer-based encoder structure. Specifically, the Transformer encoder consists of L stacked layers, each containing a Multi-Head Attention (MHA) mechanism and a Feedforward Neural Network (FFN). The first layer takes the embedded sequence $\mathbf{H}_{l-1}$ as input, and computes global dependencies via:

$$\text{head}_i = \text{Attention}(Q, K, V) = \frac{QK^T}{\sqrt{d_k}}V \tag{4}$$

Where Q, K, and V represent the query, key, and value matrices, respectively, obtained by linear transformations of the input:

$$Q_i = \mathbf{H}_{l-1}\mathbf{W}_i^Q, \quad K_i = \mathbf{H}_{l-1}\mathbf{W}_i^K, \quad V_i = \mathbf{H}_{l-1}\mathbf{W}_i^V \tag{5}$$

Here, $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, and $\mathbf{W}_i^V$ are learnable parameters for the i-th attention head, with $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{d\times d_k}$, where d is the embedding dimension and d_k is the attention subspace dimension. The outputs from all attention heads are concatenated and passed through a linear layer:

$$z = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}_o + \mathbf{b} \tag{6}$$

Subsequently, z is fed into a position-wise Feedforward Neural Network (FFN), which applies non-linear transformations independently to each time step:

$$\text{FFN}(z) = \text{ReLU}(z\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \tag{7}$$

This design enhances the model's expressive power, allowing it to learn complex nonlinear relationships and distinguish hidden anomalies or subtle variations in the gas sequence.

3) Anomaly Restoration Decoder Module: In the monitoring data of dissolved gases in oil, abnormal fluctuations (e.g., abrupt changes, drifts, or missing values) often occur due to sensor limitations, external interference, or system anomalies. These anomalies not only reduce the accuracy of fault detection but also disrupt subsequent trend modeling and prediction. Thus, detecting anomalies alone is insufficient; we must also repair or replace them to ensure models use complete, reliable data. The goal of this module is to replace detected anomalies such that the restored sequence preserves the original dynamic trend while eliminating abnormal disturbances.

In the anomaly perception expert model, after encoding gas concentrations at each time step, we calculate an anomaly score using the latent state vector to identify anomalies. We use a prediction-residual-based scoring mechanism:

$$\widehat{x}_t = h_t W_p + b_p \tag{8}$$

The anomaly score is the mean squared error between the prediction and the observed value:

$$s_t = \|\mathbf{x}_t - \widehat{x}_t\|_2^2 \tag{9}$$

We set a correction threshold τ. if $s_t > \tau$, the time step is anomalous and requires correction. The threshold is adaptively adjusted as follows:

$$\tau = \mu_s + \lambda \cdot \sigma_s \tag{10}$$

For anomalous time points $T = [t_1, t_2, \ldots, t_k]$, we reconstruct observations into smooth values $\widehat{x}_t$ to replace anomalies.

4) Future Prediction Decoder Module: We propose a context-aware sequence imputation method using transformer-based interpolation. For anomalous positions, we replace values with [MASK], which is a learnable vector:

$$\widehat{\mathbf{X}} = [x_1, x_2, \ldots, [\text{MASK}], \ldots, x_L] \tag{11}$$

$$\widehat{x}_t = \text{MLP}(H_o) \in \mathbb{R}^{L\times d} \tag{12}$$

5) Training Objective Functions: The model replaces anomalies with [MASK] and optimizes predictions against original values via regression:

$$\mathcal{L}_1 = \frac{1}{|T|}\sum_{t\in T}\|\widehat{x}_t - x_t\|^2 \tag{13}$$

Beyond imputation, the model predicts future concentrations (K steps ahead) to enhance long-term trend modeling:

$$\mathcal{L}_2 = \frac{1}{K}\sum_{k=1}^{K}\|\widehat{x}_{T+k} - x_{T+k}\|^2 \tag{14}$$

This loss encourages accurate future trend capture, guiding decisions like early warning and maintenance. Jointly optimizing imputation and prediction strengthens the model's anomaly handling and time-series modeling capabilities.

B. Temporal Modeling Expert

1) Model Architecture Design: In the evolution process of dissolved gas concentrations in oil, gas components typically exhibit trends and phased changes over time, such as slow accumulation, abrupt growth, or periodic fluctuations. To effectively model this dynamic evolution, the temporal modeling expert aims to extract deep time-dependent features from historical gas sequences to predict future concentration trends. Considering the limited length, sparsity, and non-stationarity of dissolved gas data in oil, this paper adopts the Gated Recurrent Unit (GRU) [12] as the core modeling framework, which balances modeling capability while reducing parameter complexity and overfitting risk. The model structure is shown in the following figure:

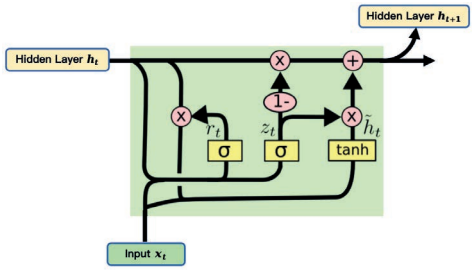

Fig. 3. Architecture diagram of the temporal modeling expert.

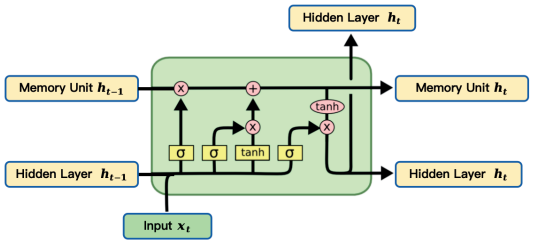

Fig. 4. Architecture diagram of the context aware expert.

2) Gated Recurrent Unit: GRU controls information forgetting and retention via a reset gate and an update gate. Specifically, the update gate z_t determines how much historical information from the previous hidden state h_{t-1} is retained in the current hidden state h_t:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{15}$$

If z_t is large, the model tends to retain new information from the current input; otherwise, it relies more on historical states. The reset gate r_t controls the information interaction between the current input and the previous hidden state, deciding whether to forget part of the previous information to avoid redundant or irrelevant content interfering with the model's judgment:

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{16}$$

If r_t is close to 0, the model "resets" the influence of historical states, using only the current input. Notably, when a state abrupt change is detected, the reset gate helps the model quickly adjust its dependence on historical information.

$$\widetilde{h_t} = \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \tag{17}$$

Here, the reset gate "trims" the previous state, retaining only parts helpful for the current input, which then fuses with the new input to generate a candidate state. The $\tanh$ activation constrains values to $[-1, 1]$.

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h_t} \tag{18}$$

Here, z_t controls the fusion ratio of old and new information. A small z_t means the current input is less important, so more historical information is retained; a large z_t indicates the current input carries key trends, so old states are replaced promptly.

3) Objective Function Optimization: In the prediction phase, we use the last hidden state h_T as the initial condition and predict future α steps via the decoder:

$$\begin{aligned} h_{T+1} &= \text{GRU}(h_T) \\ \widehat{x}_{T+1} &= W_o h_{T+1} + b_o \end{aligned} \tag{19}$$

The final predicted sequence $[\widehat{x}_{T+1}, \ldots, \widehat{x}_{T+\alpha}]$ is generated step-by-step through this decoding process and serves as the target for comparison with true values in the loss function. The loss function uses mean squared error (MSE) to measure the distance between predictions and true values:

$$\mathcal{L}_3 = \frac{1}{|\alpha d|} \sum_{t=T+1}^{T+\alpha} \sum_{i=1}^{\text{sd}} \|\widehat{x}_t^{(i)} - x_t^{(i)}\|^2 \tag{20}$$

C. Context Aware Expert

1) Model Architecture Design: To capture the complex dependencies among gases, we propose a context-aware expert model based on the LSTM (Long Short-Term Memory) [13] architecture. The core goal of this module is to model the collaborative evolution patterns among different gas components at the same time step. Unlike traditional univariate modeling, we introduce the idea of "cross-variable utilization" for modeling coupling relationships, enabling the model to not only understand the temporal trends of gases but also capture the interactive relationships among gases at the current moment, thereby enhancing the model's ability to represent potential fault patterns.

In the context-aware expert model, we treat the vector composed of all gas components at each time step t as the basic input unit, defined as:

$$\mathbf{x}_t = \left[x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(7)}\right]^\top \in \mathbb{R}^7 \tag{21}$$

where the superscripts denote gas component types (e.g., H_2, CH_4, C_2H_6, etc.). As shown in Equation (2), we map $\mathbf{x}_t$ to a high-dimensional embedding $\mathbf{e}_t^{(i)}$ to enhance its context representation capability. Subsequently, we feed the entire gas sequence $\left[\mathbf{e}_t^{(1)}, \mathbf{e}_t^{(2)}, \ldots, \mathbf{e}_t^{(7)}\right]$ into the LSTM model, leveraging its powerful sequence modeling ability to model the joint distribution of gases over time. The model structure is illustrated in Figure 3.

2) Context Modeling Design: The internal structure of LSTM consists of three gates (forget gate, input gate, output gate) and a memory cell. For some gas components, their changes may have little relation to the target gas. The forget gate automatically reduces the influence of such irrelevant information to avoid noise interference:

$$f_i = \sigma\left(W_f[\mathbf{h}_{i-1}, \mathbf{e}_i] + b_f\right) \tag{22}$$

where W_f and b_f are learnable parameters, and $\mathbf{h}_{i-1}$ is the hidden state from the previous step. The input gate determines

how much new information is "written" into the current memory cell:

$$I_i = \sigma\left(W_i[\mathbf{h}_{i-1}, \mathbf{e}_i] + b_i\right) \tag{23}$$

Next, we calculate the candidate memory state and update the memory cell, allowing the model to retain long-term patterns while incorporating cross-gas features at the current moment:

$$\widetilde{c}_i = \tanh\left(W_c[\mathbf{h}_{i-1}, \mathbf{e}_i] + b_c\right) \tag{24}$$

$$c_i = f_i \odot c_{i-1} + I_i \odot \widetilde{c}_i \tag{25}$$

Finally, the output gate decides how much activated memory content to pass to the next layer or final output. In downstream tasks, the current memory and output gate are used to select the true output vector:

$$O_i = \sigma\left(W_o[\mathbf{h}_{i-1}, \mathbf{e}_i] + b_o\right) \tag{26}$$

$$\mathbf{h}_i = O_i \odot \tanh(c_i) \tag{27}$$

3) Objective Optimization Design: Similar to Section 2.3, we use MSE to measure the difference between predicted gas values and true values. The formula is as follows:

$$\mathcal{L}_4 = \frac{1}{|I|} \sum_{i=1}^{|I|} \sum_{t=1}^{T} \left\| \widehat{x}_t^{(i)} - x_t^{(i)} \right\|^2 \tag{28}$$

We calculate the loss for each gas type, directly optimizing the independent prediction errors of all gases, which helps improve the overall prediction accuracy.

III. Mixture of Experts Model

In the preceding sections, we constructed three expert models targeting different modeling objectives: The anomaly perception expert detects and repairs anomalies in gas concentration sequences, the temporal modeling expert captures the temporal evolution trends of a single gas and the context-aware expert models the complex cross-gas dependencies at the same time step. However, in actual power equipment monitoring scenarios, the evolution of gas concentrations is often jointly influenced by anomalies, historical trends, and component interactions. Relying on a single expert is insufficient to fully characterize this dynamic process.

To address this, we propose a multi-expert fusion framework (Mixture of Experts, MoE), which integrates the advantages of the three experts and uses a dynamic gating mechanism to achieve multi-perspective information fusion and collaborative prediction. Specifically, we use a multi-layer perceptron as the gating network, which generates expert weights in real time based on the characteristics of the input data. The formula is:

$$g(\mathbf{x}) = \text{softmax}\left(W_g \mathbf{x} + b_g\right) \tag{29}$$

where $g(\mathbf{x})$ represents the output weights of each expert, and $\mathbf{x}$ is the actual output of each expert. The final prediction result is the weighted sum of the outputs of all experts.

IV. Experiments

A. Dataset Introduction

To comprehensively evaluate the proposed model's capability in modeling and generalizing complex multivariate time series data, this study constructs an experimental dataset based on transformer online monitoring data collected from 37 typical substations and 471 high-voltage transmission lines across the country. The dataset spans several years, ensuring broad representativeness and practical application value. After data cleaning and pre-processing, more than 2.56 million valid high-quality time series samples were obtained, providing a solid data foundation for model training and validation.

B. Dataset Preprocessing

To address measurement errors and extreme outliers in field-collected data, systematic data cleaning and normalization were performed before modeling. First, anomalies were detected using the interquartile range (IQR) method: for each gas, the first quartile (Q_1) and third quartile (Q_3) were calculated, with the IQR defined as$\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_1$. Observations outside the range$[\mathrm{Q}_1 - 1.5\mathrm{IQR}, \mathrm{Q}_3 + 1.5\mathrm{IQR}]$were identified as outliers and removed. Missing values were imputed using the Exponential Weighted Moving Average (EWMA), which assigns higher weights to recent data.To construct input-output pairs for sequence modeling, a sliding-window mechanism was adopted: a window size of 32 time steps was used to input historical data and predict the next time step, with a step size of 16 to generate large-scale data pairs. The cleaned data were split into training, validation, and test sets at an 8:1:1 ratio. Optimal model parameters were selected using the validation set, and the test set was used to evaluate generalization ability.

C. Evaluation Metrics

To comprehensively assess the model's performance in multi-variable time-series prediction, two metrics were used: Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). Lower values indicate smaller deviations. The formulas are as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\hat{y}_i - y_i\right)^2} \tag{30}$$

$$\text{MAPE} = \frac{100\%}{N} \sum_{i=1}^{N} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \tag{31}$$

D. Experimental Setup

The Adam [14] optimizer was used to update model weights, with an initial learning rate of 10^{-5}, a dropout rate of 0.1, a hidden dimension of 128, a batch size of 128, and 500 training epochs. All encoders had 4 layers; the anomaly-aware expert model used 8 attention heads, and the decoder had 1 layer.

TABLE I
COMPARISON OF RMSE RESULTS BETWEEN SINGLE-EXPERT AND MULTI-EXPERT MODELS

Model	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	CO	CO_2
MLP	6.74	3.77	0.95	0.64	0.38	98.65	56.36
Anomaly-Aware Expert	4.40	3.76	1.87	0.24	0.41	47.96	98.58
Temporal Modeling Expert	3.00	3.25	1.89	0.43	0.62	49.52	62.87
Context-Aware Expert	2.57	3.22	1.88	0.52	0.91	51.59	28.03
Multi-Expert Learning Model (Ours)	**1.87**	**1.89**	**0.43**	**0.16**	**0.34**	**7.42**	**25.29**

TABLE II
COMPARISON OF MAPE RESULTS BETWEEN SINGLE-EXPERT AND MULTI-EXPERT MODELS

Model	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	CO	CO_2
MLP	76.76	47.80	89.14	60.83	32.93	55.34	57.13
Anomaly-Aware Expert	84.09	47.72	65.05	62.22	38.52	74.41	95.85
Temporal Modeling Expert	57.23	41.14	65.60	130.14	14.97	74.55	95.95
Context-Aware Expert	48.97	44.85	65.47	157.78	46.71	74.75	95.99
Multi-Expert Learning Model (Ours)	**35.55**	**40.97**	**60.88**	**24.60**	**4.14**	**73.98**	**96.12**

E. Experimental Results

To evaluate the multi-expert learning model for dissolved gas time-series regression, comparative experiments were designed to compare single-expert and multi-expert models. Training configurations and input-output formats were standardized to ensure fairness, with results rounded to two decimal places. The best-performing results are highlighted in bold in the tables.

F. Experimental Analysis

From the results in Table, the multi-expert learning model demonstrates significant advantages in dissolved gas regression. In detail, RMSE metric: The multi-expert model achieves the lowest errors across all gases (e.g., H_2, CH_4, CO), showcasing stronger fitting ability, reduced bias, and better generalization. MAPE metric: The multi-expert model outperforms others for most gases, especially fault-sensitive gases (e.g., C_2H_4, C_2H_2), indicating superior adaptability and robustness.

Single-expert models capture only partial features, while the multi-expert mechanism integrates anomaly detection, temporal modeling, and context-aware capabilities, enabling collaborative feature learning and improving prediction accuracy.

V. CONCLUSION

This paper proposes a prediction model for time series forecasting of dissolved gases in transformer oil based on a multi-expert learning mechanism. The model combines anomaly-aware, temporal modeling, and context-aware experts to achieve comprehensive modeling of complex industrial signals. Experiments on multiple gas indicators show that the proposed model significantly outperforms conventional multilayer perceptron and single-expert models in both RMSE and MAPE metrics. The results demonstrate that the multi-expert model achieves lower prediction errors and effectively adapts to diverse signal patterns, including trends, sudden changes, and contextual dependencies, highlighting strong generalization and robustness. Compared with traditional methods, the model is better suited for equipment condition monitoring and anomaly detection in complex environments, with high engineering application value.

Future work will explore enhanced collaboration among expert models through dynamic weight adjustment or reinforcement learning, and extend the method to broader intelligent maintenance tasks such as equipment life assessment.

REFERENCES

[1] C. Tie, C. Yifu, L. Xianshan, L. Haowei, and C. Weidong, "Prediction of dissolved gas concentration in transformer oil based on sds-ssa-lstm," *Electronic Measurement Technology*, 2022.

[2] X. Guomin and L. Xiaoyu, "Transformer fault diagnosis method based on improved ssa optimized mds-svm," *Control and Decision*, 2023.

[3] L. Xuan, M. Fei, S. Haoyuan, D. Qili, and Z. Jianyong, "State evaluation of uhvdc converter valve considering sample imbalance and its influencing factors analysis," *Proceedings of the CSEE*, 2022.

[4] O. E. Gouda, S. H. El-Hoshy, and H. H. EL-Tamaly, "Condition assessment of power transformers based on dissolved gas analysis," *IET Generation, Transmission & Distribution*, 2019.

[5] H. Ma, Z. Li, P. Ju, J. Han, and L. Zhang, "Diagnosis of power transformer faults on fuzzy three-ratio method," in *2005 International Power Engineering Conference*. IEEE, 2005.

[6] S. Souahlia, K. Bacha, and A. Chaari, "Svm-based decision for power transformers fault diagnosis using rogers and doernenburg ratios dga," in *10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13)*, 2013.

[7] W. Nana, L. Wenyi, and L. Jianqiu, "Prediction of dissolved gas content in transformer oil based on variational mode decomposition-cuckoo search-support vector regression model," *EMT*, 2024.

[8] Z. Pengkun, Y. Jin, L. Bo, S. Changji, and Z. Jing, "Prediction method of dissolved gas concentration in transformer oil based on emd-gcforest model," *Electric Power Science and Engineering*, 2023.

[9] D. J. Miller and H. Uyar, "A mixture of experts classifier with learning based on both labelled and unlabelled data," *Advances in neural information processing systems*, vol. 9, 1996.

[10] S. E. Yuksel, J. N. Wilson, and P. D. Gader, "Twenty years of mixture of experts," *IEEE transactions on neural networks and learning systems*, 2012.

[11] 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.

[12] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (gru) neural networks," in *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, 2017.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, 1997.

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