Electrical Power and Energy Systems 172 (2025) 111227

Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

journal homepage: www.elsevier.com/locate/ijepes

From limited data to reliable diagnosis: an interpretable deep learning framework for transformer fault analysis

Jiajian Lin [a,*], Lit Yen Yeo [a], Hadi Nabipour Afrouzi [b], Mehran Motamed Ektesabi [c], Jalal Tavalaei [a,*]

[a] Faculty of Engineering, Computing and Science, Swinburne University of Technology, Sarawak Campus, 93350 Kuching, Malaysia
[b] Faculty of Computing, Engineering and The Built Environment, Birmingham City University, Birmingham B5 5JU, UK
[c] School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Hawthorn, Victoria, Australia

ARTICLE INFO

Keywords:
Power Transformer
Fault Diagnosis
Generative Adversarial Networks
Goose Optimization
Attention Mechanism

ABSTRACT

The reliable diagnosis of power transformer faults is important for ensuring the safety and stability of modern power systems. However, existing fault identification techniques suffer from limited diagnostic accuracy due to insufficient feature representation, inadequate handling of data imbalance in Dissolved Gas Analysis datasets, and suboptimal model generalization. Furthermore, the absence of comprehensive theoretical investigations into the underlying fault mechanisms and model interpretability has significantly constrained the development of robust and explainable diagnostic frameworks. To address these problems, the GAN-CNN-BiLSTM-Attention-GOOSE framework was proposed to overcome the limitations of traditional transformer fault diagnosis, address data scarcity challenges and provide new avenues for future transformer protection research. In this study, an attention-based deep learning model was developed to improve the accuracy and reliability of transformer fault diagnosis. To figure out the limitations posed by insufficient data, a generative adversarial network was introduced to enrich the training samples. A GOOSE optimization algorithm was employed to fine-tune the learning process and enhance overall performance. This integrated approach yielded higher classification accuracy compared to conventional methods. To further interpret the model's predictions, an explainability technique was applied to analyze the input gas data. The analysis revealed clear patterns linking specific gas compounds in transformer oil to operational faults. In particular, two key indicators, C_2H_2 and C_2H_4, were found to be strongly associated with high-energy arcing and thermal faults, respectively. These findings highlight the importance of adequate training data and careful model calibration in achieving accurate and interpretable fault identification.

1. Introduction

Power transformers are vital in modern power systems, ensuring the effective transmission and distribution of electrical energy while maintaining grid stability and reliability [1]. Any unexpected failure in these transformers may lead to severe power outages and significant economic losses [2]. Therefore, early detection and diagnosis of faults have drawn substantial attention in industrial and academic communities. As one of the most used diagnostic techniques, dissolved gas analysis (DGA) measures various gases generated inside transformers, offering valuable insights into their operating conditions and facilitating fault identification. The principle is to identify internal faults through the absorption of important gases such as CO, CO_2, H_2, C_2H_6, C_2H_4, C_2H_2, and CH_4, using various classical methods, and to evaluate the status based on test sample results [3]. By interpreting the DGA data, engineers can predict incipient faults more accurately, minimizing downtime and preventing catastrophic failures. Beyond dissolved gas analysis, health indices that integrate multiple diagnostic criteria can enhance transformer health assessment. Multi-criteria decision-making frameworks, combining parameters such as dielectric strength, oil quality, furan content, and electrical tests, offer more comprehensive evaluations than single-criterion methods. However, their reliance on manual weighting and statistical formulations limits adaptability under noisy or incomplete data. This highlights the need for intelligent deep learning models that automatically learn feature importance and generalize across diverse fault conditions [4].

* Corresponding authors.
E-mail addresses: jlin@swinburne.edu.my (J. Lin), jtavalaei@swinburne.edu.my (J. Tavalaei).

https://doi.org/10.1016/j.ijepes.2025.111227
Received 2 July 2025; Received in revised form 15 September 2025; Accepted 30 September 2025
Available online 8 October 2025

Deep learning models have been widely adopted to process DGA data for fault diagnosis, owing to their capability for automatic feature extraction and powerful nonlinear approximation. The proposed convolutional neural network demonstrated high effectiveness in transformer fault diagnosis by accurately identifying health states based on dissolved gas, oil quality, and depolarization features. Through oversampling to address data imbalance, the model achieved an overall accuracy of 98.53 % with complete recognition of the critical poor state, significantly surpassing conventional machine learning approaches [5]. A semi-supervised transfer learning framework using deep neural networks has achieved an accuracy of over 95 % in transformer fault type classification, demonstrating that deep learning techniques can improve transformer diagnosis performance [6]. Besides, a novel method based on Graph Convolutional Network (GCN) effectively represented the similarity measure between unknown samples and labeled samples using an adjacency matrix. The results indicated that the GCN model outperforms traditional methods such as CNN, multilayer perception, and SVM, achieving higher diagnostic accuracy in various input features and data volumes [7]. However, several research gaps exist within this domain. First, most current deep learning models lack an autonomous mechanism to emphasize critical features, which may reduce their effectiveness in accurately identifying transformer fault patterns [8]. Second, real-world DGA datasets are often incomplete or exhibit noise due to sampling errors, sensor malfunctions, or diverse operating conditions, making it difficult for models to learn robust fault representations [9]. Furthermore, the hyperparameters in deep learning approaches are often tuned manually based on subjective expertise, leading to suboptimal model performance and potentially higher training costs.

Incorporating attention mechanisms into deep learning models offers a promising solution to the challenges of fault diagnosis, particularly in complex and noisy industrial environments. By adaptively emphasizing salient features, attention mechanisms significantly enhance the accuracy of fault identification. Specifically, the cooperative attention module strengthens diagnostic performance by directing focus toward critical features and increasing the diversity among base classifiers, thereby improving the model's resilience to noise and its sensitivity to minority fault classes [10]. Meanwhile, the attention-guided graph isomorphism learning framework captures task-relevant feature dependencies through self-attention, facilitating effective knowledge sharing between fault diagnosis and remaining useful life (RUL) prediction [11]. This integration enhances predictive accuracy while reducing the computational cost and complexity of deploying separate models. The frequency attention mechanism also contributes to diagnostic robustness by highlighting essential frequency-domain features in vibration signals, enabling accurate fault detection even in noisy or data-limited scenarios [12]. Collectively, these attention-based strategies advance the reliability, adaptability, and efficiency of intelligent fault diagnosis systems.

In parallel, data augmentation methods, particularly those driven by generative models—have effectively handled limited or noisy datasets by producing plausible synthetic samples. The multiple path alignment generative adversarial network (GAN) enhances fault diagnosis under limited data conditions by generating high-quality synthetic signals aligned with real data in both time and frequency domains while filtering out redundant components and improving data diversity and discriminability through auxiliary classification [13]. Moreover, sophisticated optimization algorithms that leverage bio-inspired or *meta*-heuristic strategies have emerged to automatically search for high-performing hyperparameter configurations, thereby alleviating the labor-intensive and error-prone manual tuning process [14]. The graph optimization algorithm enhances fault diagnosis by refining the graph structure through dual-scale spectral feature extraction and similarity optimization, enabling robust representation of feature relationships and improving diagnostic accuracy under noisy and low-label conditions [15]. The Barabási-Albert model-enhanced genetic algorithm improves fault diagnosis by optimizing feature subset selection through complex network topology and evolutionary strategies, thereby enhancing the accuracy, noise resistance, and generalization ability of machine learning models for diagnosing multivariate time-series faults in ship power grids [16].

Recent advances in intelligent fault diagnosis have increasingly focused on addressing challenges such as insufficient samples, noisy environments, and high computational costs. For instance, uncertainty-aware metric learning frameworks have been developed to explicitly model data uncertainty and improve diagnostic reliability under limited and noisy domain conditions [17]. The framework integrates a multi-scale cross-feature extraction module to mine key discriminative cues under noise, an uncertainty modeling of the query–prototype similarity in metric space with a tailored loss for joint optimization of similarity and uncertainty, trained in an episodic N-way K-shot manner, and a colony-based class activation mapping module to provide reliable, focused visual explanations of the model's decisions. Similarly, adaptive evolutionary reconstruction networks (AERMN) have effectively identified unknown fault types by leveraging domain adaptation strategies and robust feature reconstruction [18]. The proposed AERMN integrates a feature embedding and clustering stage for known fault identification, a reconstruction-based stage for unknown fault rejection, an embedding self-evolving regularization strategy to dynamically adjust feature importance, and a reinforcement learning-based adaptive threshold mechanism to improve robustness against variations in unseen domains. In parallel, energy-efficient diagnostic methods based on neural-dynamics-inspired spiking architectures have been proposed to tackle the dual challenges of accuracy and resource constraints in Industrial Internet of Things (IIoT) scenarios, achieving high diagnostic accuracy with significantly reduced computational overhead [19]. The proposed framework incorporates a multiscale mask spiking self-attention mechanism to efficiently extract spatiotemporal features, a rate encoding metric classifier to bridge spiking representations with prototype-based decision boundaries, and a neural-dynamics-inspired backpropagation strategy to achieve stable and effective training under few-sample and noisy conditions. Collectively, these recent studies highlight the transition of fault diagnosis research toward interpretable, noise-robust, and resource-efficient models, thereby pointing to promising directions for future applications in complex and data-limited industrial environments.

In response to these challenges, this work proposes an enhanced CNN-BiLSTM-Attention framework to improve power transformer fault diagnosis accuracy and reliability based on DGA data. The key contributions of this study are as follows:

1. Data augmentation: a GAN was employed to address the limited and low-quality datasets. GAN generated synthetic yet realistic data samples, enriching the dataset and mitigating the negative effects of data sparsity and imbalance.
2. Hyperparameter optimization: the GOOSE optimization algorithm was used to fine-tune the model's hyperparameters. This *meta*-heuristic approach systematically identified the optimal parameter set, ensuring robust and efficient network training for fault diagnosis tasks.
3. Feature contribution analysis: SHAP analysis was performed on the DGA dataset to investigate the relationship between key features and transformer fault types. This interpretable approach revealed the nuanced dependencies of each dissolved gas component, aligning the model's decisions with domain-specific diagnostic knowledge and enhancing trust in its predictions.

2. Feature extraction

This section presents the feature extraction framework used to enhance transformer fault diagnosis. It combines a CNN-BiLSTM-Attention model for capturing key patterns, a generative approach to

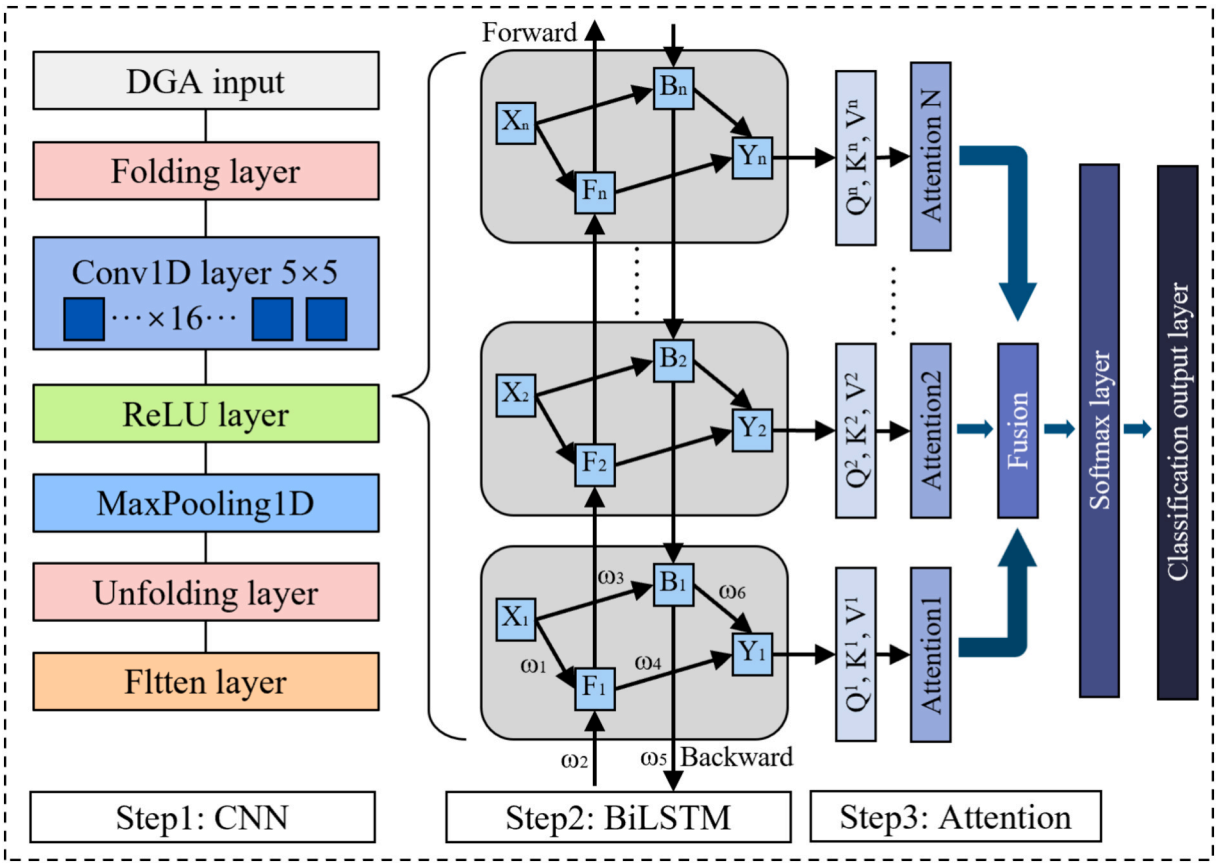

Fig. 1. Classification framework for CNN-BiLSTM-Attention.

enrich data diversity, and an optimization algorithm to fine-tune model performance.

2.1. CNN-BiLSTM-Attention

An intelligent classification framework was proposed to analyze DGA data, integrating convolutional layers for spatial feature extraction, bidirectional long short-term memory networks for temporal sequence modelling, and an attention mechanism to enhance the interpretability and focus on key features. The overall architecture, including its core components and data processing flow, is illustrated in Fig. 1.

CNN are highly valued in power system research because they can automatically extract local spatial features from complex input data, thereby improving the accuracy and efficiency of fault detection and classification tasks. CNN is mainly composed of convolutional layers and pooling layers, where the convolutional layer uses convolutional kernels for effective nonlinear local feature extraction of power load data, and the pooling layer is used to compress the extracted features and generate more important feature information to improve generalization ability. The CNN model employed in this study comprised several key layers for effective feature extraction. It incorporates a Folding Layer to reshape the input data for subsequent processing. Then, a Convolution1D layer with 16 convolution kernels is utilized to capture local patterns and extract essential features from the input sequence.

Each convolution operation was followed by a ReLU Layer for introducing non-linearity and enhancing the model's expressive power, resulting in $X_l^{(relu)}$. The next step was to introduce the MaxPooling1D layer to reduce data dimensionality, preserve salient features, reduce computational complexity, and enhance the model's generalization ability.

$$X_l^{pool}[i,f] = \max_{j=1}^{k}\left(X_l^{(relu)}[i\cdot s+j,f]\right) \tag{1}$$

Where i is index of the output feature map; f is the f^{th} feature channel; k is pooling window size; s is stride. After the Maxpooling operations, an Unfolding Layer was used to restore the data structure. Subsequently, a Flatten Layer flattened the output to prepare it for subsequent processing in the neural network. The training objective for the CNN model was to minimize the cross-entropy loss function, defined as:

$$L = -1\Big/N\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\left(\hat{y}_{i,c}\right) \tag{2}$$

Where $y_{i,c}$ is true label for class c in sample i; $\hat{y}_{i,c}$ is predicted probability for class c in sample i.

The Bidirectional Long Short-Term Memory (BiLSTM) network consists of two LSTM layers—a forward LSTM and a backward LSTM—working in parallel to capture past and future dependencies in sequential data. Initially, feature extraction is performed using a CNN to capture local spatial patterns from the data. The extracted features are then passed into the BiLSTM network, where both the forward and backward layers analyze the sequence to preserve temporal dependencies.

In Fig. 1, the corresponding input data is $x_1, x_2, \cdots, x_n$. Forward and backward iterations respectively generate hidden state $F_1, F_2, \cdots, F_n$ and $B_1, B_2, \cdots, B_n$. Then, output the corresponding data $Y_1, Y_2, \cdots, Y_n$, where the weights of each layer are $\omega_1, \omega_2, \omega_3, \cdots, \omega_6$. In BiLSTM, the update status and output of the hidden layer are as follows:

$$\begin{cases} F_i = f_1(\omega_1 x_i + \omega_2 F_{i-1}) \\ B_j = f_2(\omega_3 x_i + \omega_5 B_{i+1}) \\ Y_i = f_3(\omega_4 A_i + \omega_6 B_i) \end{cases} \tag{3}$$

Where f_1, f_2, and f_3 are the activation functions between different layers, respectively.

To enhance model stability and accelerate convergence, layer normalization was applied after each BiLSTM layer, mitigating sensitivity to parameter initialization. In addition, global norm clipping was employed during backpropagation to prevent gradient explosion, ensuring numerical stability. In addition, L2 regularization is added to the loss function to penalize excessive weight values, and a standard gradient descent optimizer was used.

The CNN effectively extracts local spatial features from raw sensor data, while the BiLSTM captures long-term temporal dependencies across the feature sequences. However, not all extracted features contribute equally to the final diagnosis, and sometimes, steps or spatial regions hold more critical information for identifying fault patterns. To address this issue, a self-attention mechanism was introduced after CNN-BiLSTM. This mechanism dynamically recalibrated the feature representations by assigning attention scores, ensuring that more critical features exert a stronger influence on the final output.

Multiply the feature matrix Y output by BiLSTM with the weight matrix W to obtain the Query (Q), Key (K), and Value (V) matrices:

$$\begin{cases} Q = Y \cdot W_Q \\ K = Y \cdot W_K \\ V = Y \cdot W_V \end{cases} \tag{4}$$

Next, the attention scores are computed using the scaled dot-product attention:

$$Attention(Q, K, V) = softmax\left(QK^T \Big/ \sqrt{d_k}\right)V \tag{5}$$

Where d_k is dimensionality of the Key vector; the softmax function normalizes the attention scores across all features.

2.2. Generative adversarial networks

GAN comprises two key components: the Generator (G) and the Discriminator (D) [20]. The Generator aims to produce high-quality synthetic data by learning patterns from historical load data. At the same time, the Discriminator evaluates the authenticity of the generated data and distinguishes it from real data samples. The overall objective function is:

$$\min_G \max_D V(D, G) = E_{x \sim P_{data}(x)}[logD(x)] + E_{z \sim P_z(z)}[log(1 - D(G(z)))] \tag{6}$$

Where x is real data samples; z is random noise samples; $D(x)$ is the discriminative probability of the discriminator on real data; $D(G(z))$ is the discriminative probability of the discriminator on the generated data; $P_{data}(x)$ is real data distribution; $P_z(z)$ is distribution of noise.

The Generator aims to produce high-quality synthetic fault data by learning patterns from historical transformer operational data, effectively expanding and diversifying the training dataset. This is particularly important when certain fault types are underrepresented in the original dataset. Meanwhile, the Discriminator evaluates the authenticity of the generated data, distinguishing between real fault data and synthetic samples. Therefore, the loss function of the generator and the Binary Cross Entropy Loss function (L_{CE}) can be defined as:

$$\begin{cases} L_G = -E_{z \sim P_z(z)}[logD(G(z))] \\ L_G = L_{CE}(1, D(G(z))) \end{cases} \tag{7}$$

At the early stages of training, the Discriminator often struggles to accurately differentiate real from generated data, resulting in significant prediction errors. These errors are then utilized to iteratively optimize the Discriminator, enhancing its ability to detect subtle anomalies and patterns within fault data. As the Discriminator improves, the Generator faces increasing challenges in producing fault samples capable of deceiving the Discriminator. Therefore, the goal of the discriminator is to accurately distinguish between real data and generated data as much as possible. Its loss consists of two parts:

$$\begin{cases} L_{D_{real}} = L_{CE}(1, D(x)) \\ L_{D_{fake}} = L_{CE}(0, D(G(z))) \\ L_D = \frac{1}{2}\left(L_{D_{real}} + L_{D_{fake}}\right) \end{cases} \tag{8}$$

where 0 and 1 are represent the desire for the discriminator to determine all generated data as false or the desire for the discriminator to determine real data as true; L_D is the mean of the two. During the training process, gradients are calculated through backpropagation, and the Adam optimizer is used to update the parameters of the generator and discriminator. Subsequently, the Generator was optimized based on feedback from the Discriminator, enabling it to produce increasingly realistic fault patterns. This adversarial training process ensures that the synthetic data generated by the GAN aligns closely with the statistical and chemical characteristics of real transformer fault data.

2.3. Goose algorithm

The GOOSE algorithm is recognized for its robust balance between exploration and exploitation, making it particularly effective in solving high-dimensional and complex optimization problems [21]. The GOOSE algorithm effectively avoids local optima and accelerates convergence towards global optima by employing stochastic search strategies and adaptive parameter adjustments. Furthermore, its flexibility and robustness in handling multidimensional optimization tasks allow it to efficiently explore complex feature spaces and identify optimal parameter combinations.

It initializes the geese population's positions as the X matrix. To update their positions within the solution space, it employs three key variables: RND (Random Variable), PRO (Probability Variable), and WEIGHT (Weight of the Stone). The RND, uniformly distributed in [0,1], determines the search focus, triggering either exploration ($RND < 0.5$) or exploitation ($RND \geq 0.5$). PRO, also in [0,1], refines exploitation, with $PRO > 0.2$ enabling more intense local refinement, while $PRO \leq 0.2$ encourages broader local adjustments. WEIGHT, representing a randomly assigned "stone" weight (5,25), influences positional updates, with larger weights (weight $\geq$ 12) driving more precise adjustments. After each update, the fitness of all agents is evaluated, and the best fitness and corresponding position are recorded, iteratively guiding the algorithm toward the global optimum.

3. Data preparation and model optimization

This chapter describes the preparation of the DGA dataset, the use of GANs to generate additional synthetic samples for improved data balance, and the application of the GOOSE algorithm to optimize model hyperparameters, ensuring better accuracy and generalization in transformer fault diagnosis.

3.1. DGA dataset

The DGA used in this research is sourced from a publicly available GitCode platform dataset containing 570 samples with five key features (H_2, C_2H_6, CH_4, C_2H_2, and C_2H_4) and encompasses normal operating conditions and seven distinct fault types. Gas analysis was performed in accordance with IEC 60599 oil testing standards, with repeated sampling to reduce random measurement error. All gas concentration values were normalized into the [0,1] range to ensure comparability across different operating conditions. The power transformer model is ODFSZ-250000/500 with a voltage ratio of $525/\sqrt{3}$ / $230/\sqrt{3} \pm 8 \times 1.25$ % / 36 kV and a rated capacity configuration of 250/250/80 MVA (high/medium/low). The measured short-circuit impedances are 44 % (±10

Table 1
The number of samples for each operating condition.

Status of transformer	Quantity	Type
Low and medium temperature overheating (LMT)	113	1
High temperature (HT)	92	2
Normal	140	3
Partial discharge (PD)	83	4
Low-energy discharge (LD)	45	5
High-energy discharge (HD)	77	6
Spark discharge (SD)	10	7
Arc discharge (AD)	10	8

%) between high and low voltage windings, 14 % (±7.5 %) between high and medium, and 29 % (±10 %) between medium and low windings. For 500 kV transformers, temperature probes are typically positioned 120 mm ± 10 mm into the oil tank to ensure accurate thermal condition assessment. All gas concentration data were normalized into the [0,1] range prior to model training to address sensor noise and environmental fluctuations. These fault types are categorized into thermal faults and discharge faults. Thermal faults include low and medium temperature overheating (LMT) and high temperature (HT) overheating. In contrast, discharge faults consist of partial discharge (PD), low energy discharge (LD), high energy discharge (HD), spark discharge (SD), and arc discharge (AD). The number of samples for each operating condition is shown in Table 1.

3.2. GAN model for synthetic data generation

The process began with data preprocessing, where the input features were normalized to a range of 0 and 1 to facilitate model training. The GAN architecture comprises two key components: a Generator and a Discriminator. The Generator was designed as a fully connected neural network with two hidden layers (128 and 256 neurons, respectively) and a Sigmoid output layer to produce synthetic feature data. The Discriminator, a fully connected network, consisted of two hidden layers (256 and 128 neurons) and a single Sigmoid output neuron to classify data as real or synthetic. Both models were optimized using the Adam optimizer with a learning rate of 0.0002, and the binary cross-entropy loss function was employed to guide the training process. During training, the generator and discriminator were updated iteratively over 5000 epochs, with each epoch involving the generation of synthetic data from random noise, evaluation by the discriminator, and gradient-based optimization of both models. The training process was monitored to ensure stability by tracking the generator and discriminator losses, which were logged every 100 epochs. Upon completion of training, the generator was used to produce a large volume of synthetic data, which was then rescaled to the original feature range and combined with randomly sampled fault labels. The fidelity of the simulated data is illustrated in Fig. 2.

Fig. 2 illustrates the two-dimensional projection of high-dimensional DGA data using t-distributed Stochastic Neighbor Embedding (t-SNE), a widely adopted nonlinear dimensionality reduction technique. The primary objective of t-SNE is to map high-dimensional data into a lower-dimensional space while preserving the pairwise similarity relationships among the data points. In the high-dimensional space, the similarity between samples is modelled using conditional probabilities derived from a Gaussian distribution centred at each data point. This probabilistic framework allows t-SNE to capture local structural information, effectively revealing clusters or groupings in the original data.

The Fig. 2 shows that the blue points represent 570 original samples, while the orange points denote 2,280 synthetic samples generated by a GAN. The t-SNE projection reveals two dominant density regions corresponding to the distributional characteristics of the real and generated data. Importantly, a significant overlapping region is observed between the real and generated samples, suggesting that the GAN was able to learn and replicate the underlying distribution of the original DGA dataset. This distributional alignment between synthetic and real samples is a key indicator of generation fidelity. This implies that the GAN has successfully captured essential statistical patterns and local structures in the real data.

KL divergence was computed between the probability distributions of the real DGA samples and the synthetic dataset across the 5 key gas features to further validate the fidelity of the GAN-generated data [22]. As shown in Fig. 3, the divergence values for all gases are less than 0.015, indicating a near-perfect overlap between the real and synthetic distributions. This outcome demonstrates that the GAN could faithfully capture the statistical characteristics of the original dataset, including the mean concentration levels and the variance structure of the

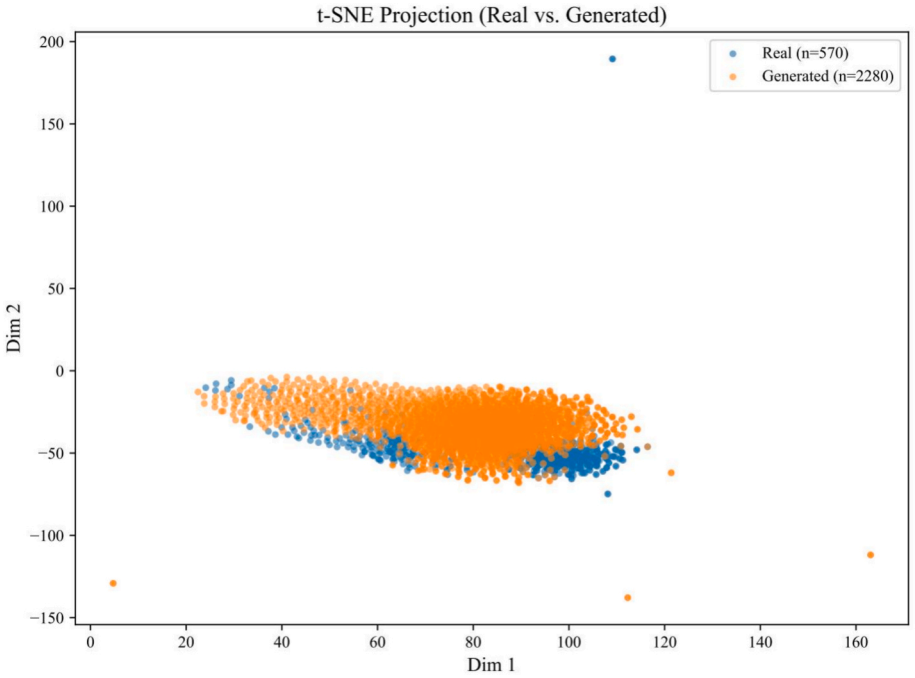

Fig. 2. t-SNE result (2D projection).

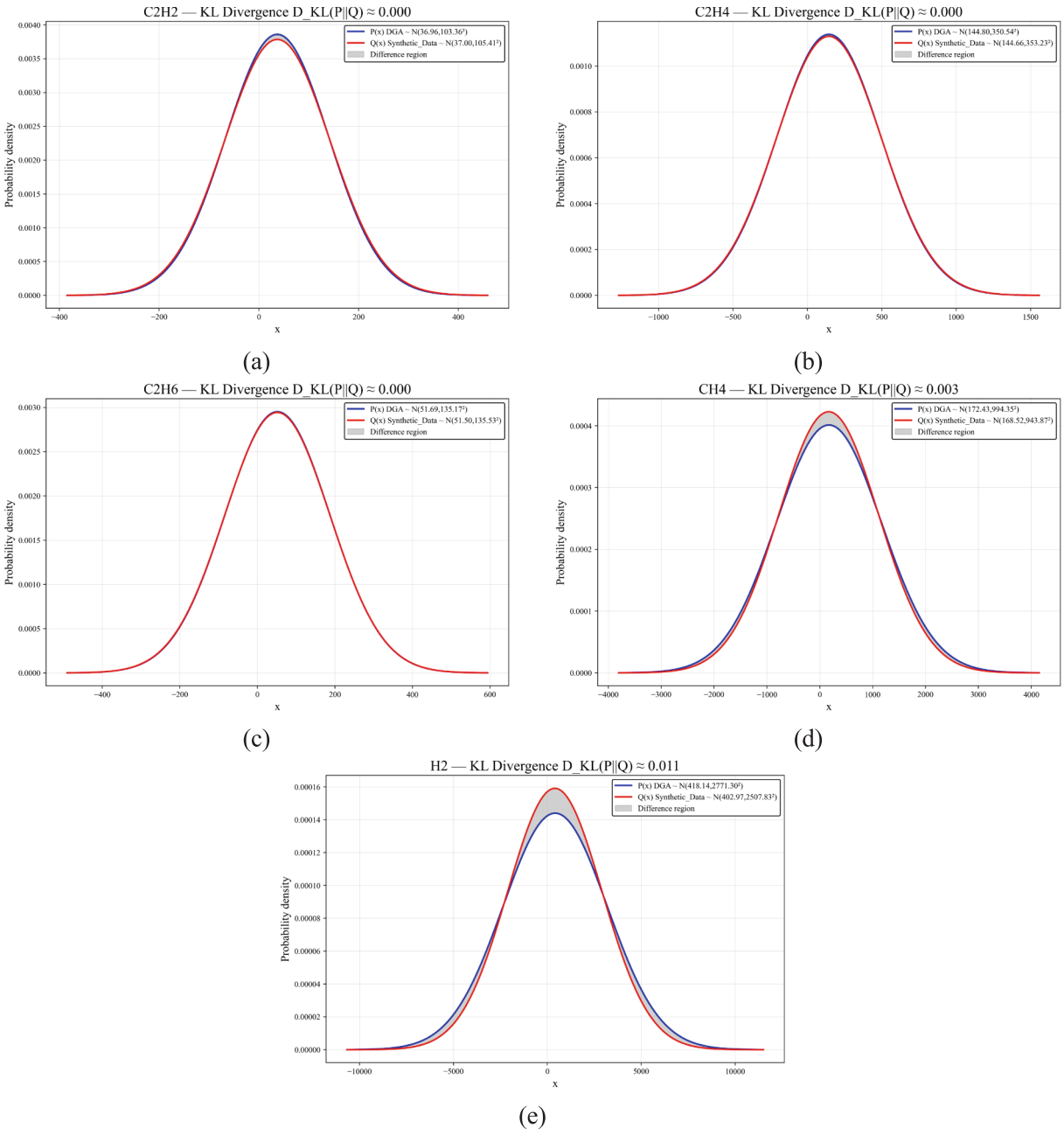

Fig. 3. Key features KL divergence.

dissolved gases.

3.3. Hyperparameters based on GOOSE

The GOOSE algorithm is a metaheuristic optimization approach inspired by the behaviour of geese during foraging and rest periods. This algorithm initializes a population of search agents within specified bounds and evaluates their fitness based on the validation loss of the CNN-BiLSTM-Attention model. It employed a combination of exploration and exploitation strategies to update the positions of these agents, thereby optimizing the model's hyperparameters. The primary hyperparameters optimized in this study include the learning rate, convolutional kernel size, and the number of BiLSTM hidden units, with respective ranges of (0.005, 0.020), (0, 25), and (55, 75).

As illustrated in Fig. 4, the resulting fitness landscape over this hyperparameter space exhibits a smooth, convex bowl-shaped topology centred near the optimal configuration. This optimal setting was determined to be (0.01, 5, 66.4447), indicating a balance between learning stability, feature extraction capacity, and sequence modelling complexity. The key parameters of the proposed model are summarized in Table 2. Then, divide the training and testing sets into a 7:3 ratio and input them into the classification framework.

Fig. 5 shows the nine convergence curves comparing the GOOSE algorithm under different initialization conditions. Overall, the most stable and efficient convergence is achieved when the number of search agents and epochs is set to 10, as the trajectories closely approach the optimum with minimal oscillation. In contrast, when the parameters are enlarged, particularly at 20 search agents with 20 epochs, the convergence rate is noticeably reduced and the curves exhibit clear instability. This trend becomes more evident as the search agents and epochs increase, suggesting that excessive parameter scaling does not enhance convergence but introduces fluctuations and inefficiency. These results demonstrate that GOOSE can achieve reliable convergence with relatively small parameter settings, highlighting its efficiency in balancing

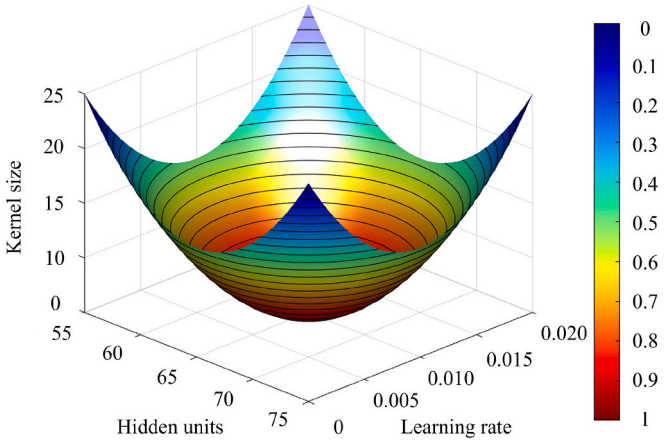

Fig. 4. Fitness function for optimization.

Table 2
Parameter settings of the proposed model.

	Parameters	Values
CNN	**convolutional kernel size**	**5 × 5**
	number of convolution kernels	16
	activation function	ReLu
BiLSTM	**hidden units**	**66.4447**
	output mode	last
Attention	heads	2
GOOSE	max search agent	30
	learning rate	**0.01**
	maximum iterations	50
	Batch size	50
	epoch	50
	activation function	Softmax
	optimizer	Adam

solution quality and computational cost.

4. Result and analysis

4.1. Comparative experiment

This study conducted a series of comparative experiments to further investigate the impact of GAN to increase the number of samples by two, three, and four times the original samples and optimization algorithms on the CNN-BiLSTM-Attention framework, as detailed in Table 3.

This study compared the impact of eight optimization algorithms on the proposed classification framework, including the GOOSE algorithm, RIME (Rime-Ice Optimization Algorithm), PO (Parrot Optimizer Algorithm), BKA (Black-winged Kite Algorithm), CPO (Crested Porcupine Optimizer), FVIM (Four Vector Intelligent Metaheuristic), NRBO (Newton-Raphson-based Optimizer), and BFO (Bitterling Fish Optimization). During the experimental runs, the number of Search Agents was uniformly set to 30, and the number of iterations was fixed at 10 for all optimization algorithms to ensure the fairness of the experiments. The experimental results demonstrated that the classification framework optimized with GOOSE exhibited the best performance, with 96.29 %, 96.24 %, 96.38 %, and 96.30 % for Accuracy, Recall, Precision, and F1 Score, respectively. Compared to the lowest-performing BFO, these figures represented improvements of 8.48 %, 8.45 %, 8.30 %, and 8.59 %, respectively. This indicated the importance of appropriate parameter settings for neural networks, which could significantly enhance the model's diagnostic capabilities.

When the proposed model was fed with raw data as input, the metrics of accuracy, recall, precision, and F1 score were not satisfactory, resulting in a significant margin of error in fault diagnosis. This indicated that the model's initial capacity to detect and classify faults was limited, potentially due to insufficient training data that constrained its learning process. However, upon increasing the sample size to 2280, the proposed model's performance improved, with enhancements of 9.38 %, 9.35 %, 9.27 %, and 9.29 % in accuracy, recall, precision, and F1 score, respectively. This substantial increase in performance metrics suggested that the expanded dataset significantly improved the model's diagnostic capabilities. Apart from the GOOSE algorithm, the performance of all other models also improved with the increase in data volume.

The comparative analysis of computational time demonstrates that the proposed framework achieves the most favorable balance between diagnostic accuracy and efficiency. As shown in Table 3, the execution time of all methods increases monotonically with larger sample sizes, reflecting the expected computational cost of processing additional data. While algorithms such as RIME and PO occasionally achieved marginally shorter runtimes at specific data scales, their diagnostic accuracies remained consistently lower than that of the proposed model. In contrast, the proposed framework attained the highest accuracy across all evaluation metrics. It maintained execution times at 8.92 s, nearly identical to the fastest RIME at 8.87 s. It should also be emphasized that all optimization algorithms were executed under the same experimental settings with a population size of 30 and a maximum of 50 iterations. This indicates that the proposed method provides dual advantages: superior predictive performance and competitive computational efficiency.

4.2. Comparison of performance with other methods

To evaluate the superiority of our proposed method comprehensively, we compared it with several recent approaches, as illustrated in Table 4.

Our proposed model achieved an accuracy of 96.29 %, which was the highest among the methods listed. Notably, the most used models in this domain were neural networks, and the CNN-based models significantly outperformed GCN. This indicated that deep convolutional models were more effective in feature extraction for fault diagnosis tasks. Including attention mechanisms further enhanced the model's performance by focusing on the most relevant features. While optimizing feature extraction models was crucial, improving the quality of training data and fine-tuning parameters remained key research trends in advancing fault diagnosis frameworks.

As summarized in Table 4, the proposed GAN-CNN-BiLSTM-Attention-GOOSE framework demonstrates several architectural advantages compared with existing models. The FCNN–Attention–BiLSTM model integrates convolutional, recurrent, and attention layers, achieving a competitive accuracy of 95 %. However, it lacks a systematic mechanism to handle data imbalance and relies primarily on semi-supervised transfer learning. The BPNN–SVM–Residual Network improves conventional shallow models through residual connections and hybrid classification. Yet, its diagnostic accuracy of 92.7 % is constrained by limited feature representation capability and manual parameter tuning. Similarly, the KPCA–TISOA-SVM model enhances feature extraction and SVM optimization via *meta*-heuristics. Still, its reliance on handcrafted kernel features and the absence of deep hierarchical learning restrict its scalability, with a performance of 91.6 %.

Deep learning-based frameworks such as the Optimized ANN and GCN further improve diagnostic accuracy, reaching 90.0 % and 89.7 %, respectively, but their architectures present limitations. The ANN depends heavily on hyperparameter search while lacking temporal feature modeling, and the GCN requires adjacency matrix construction, which is highly sensitive to graph topology and dataset size. In contrast, the proposed framework integrates five complementary modules: a GAN for data augmentation that resolves class imbalance by generating realistic fault samples, a CNN for extracting localized spatial patterns, a BiLSTM

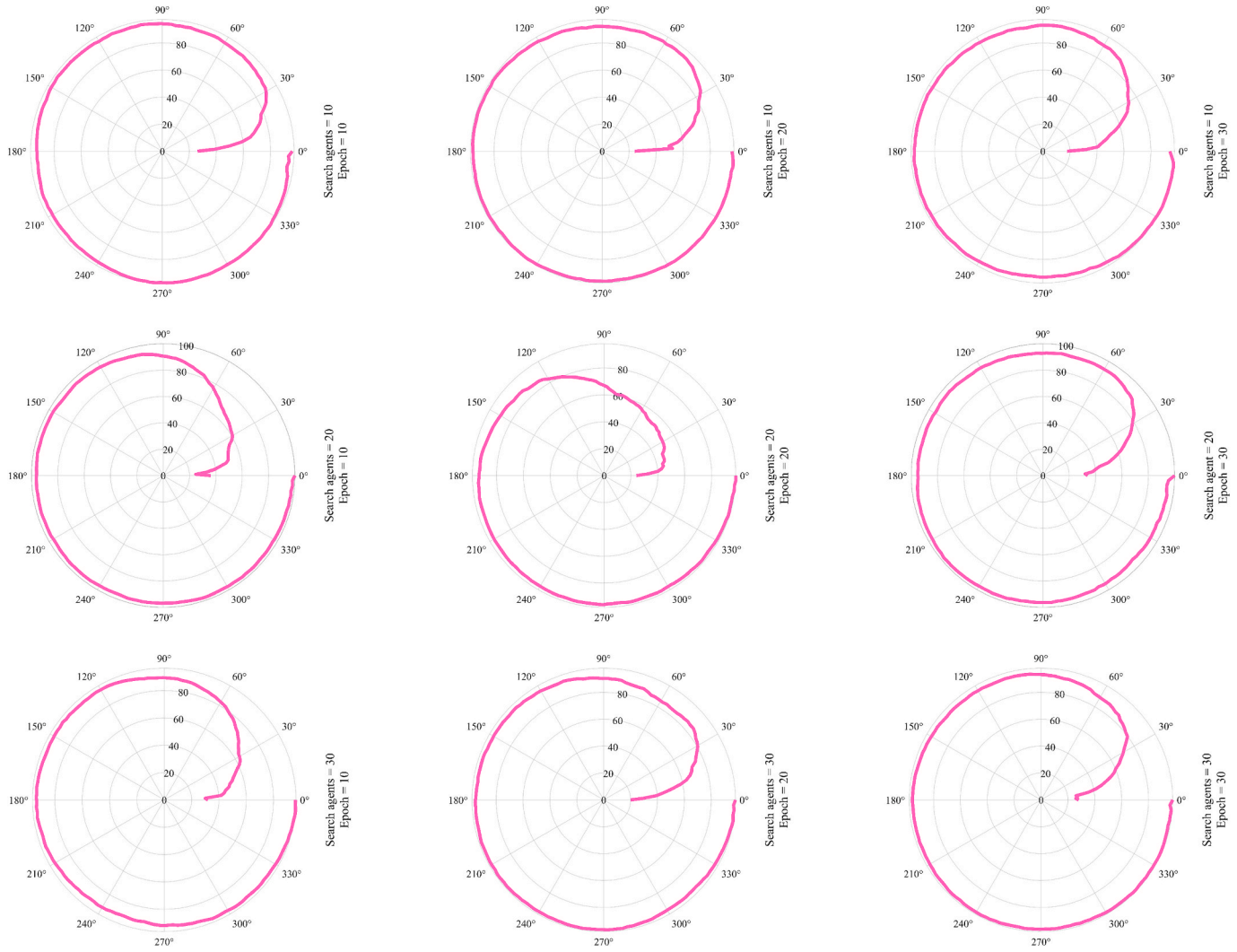

Fig. 5. The convergence performance of GOOSE algorithm under different numbers of search agents and iterations.

for capturing bidirectional temporal dependencies, an attention mechanism to highlight diagnostically salient gases, and the GOOSE algorithm to automatically optimize hyperparameters. This unified design results in superior accuracy of 96.29 %, while simultaneously ensuring interpretability through SHAP analysis that validates model decisions against established diagnostic knowledge of DGA gas signatures.

Table 5 presents the results of pairwise t-tests comparing the proposed GAN–CNN–BiLSTM–Attention–GOOSE framework against five representative baseline methods. To ensure the robustness of the reported results, all experiments were independently repeated ten times under different random initialization conditions. All reported p-values are below the 0.05 threshold, indicating that the improvement observed improvements are statistically significant. The average performance differences range from 1.01 % against the FCNN–Attention–BiLSTM model to 6.12 % against the GCN model, with corresponding 95 % confidence intervals consistently excluding zero. These results provide strong statistical evidence that the proposed framework achieves superior diagnostic performance rather than improvements arising from random variation.

Fig. 6 further visualizes the average differences and their 95 % confidence intervals. It can be observed that all confidence intervals are located entirely on the positive side of the no-improvement reference line, confirming consistent gains over baseline approaches. The largest margins are observed when compared with GCN and Optimized ANN, highlighting the advantage of incorporating data augmentation and attention-enhanced temporal modeling. More moderate yet still significant improvements are achieved over KPCA + TISOA–SVM, BPNN–SVM–Residual Network, and FCNN–Attention–BiLSTM, demonstrating that the proposed integration of GAN and GOOSE optimization provides added value even against advanced hybrid models.

4.3. Ablation experiment

Table 6 shows the results of an ablation experiment to evaluate the impact of various model components on the overall performance of the proposed GAN-CNN-BiLSTM-Attention-GOOSE framework. The experiment gradually removes each component to assess its contribution to the model's accuracy.

As illustrated in Table 6, the proposed model achieved the highest accuracy of 96.29 %. This performance served as a benchmark for evaluating the contributions of individual components. When the GAN component was removed, resulting in the CNN-BiLSTM-Attention-GOOSE model, the accuracy decreased to 86.91 %. This reduction suggested that the GAN component significantly contributed to the model's ability to learn robust features, enhancing fault diagnosis accuracy. Upon removing the optimization algorithm, the accuracy experienced a decline of merely 10.53 %. This observation suggested that while optimal parameter tuning contributed to fault diagnosis performance, the inherent architecture of the model played a more pivotal role in determining its efficacy.

Table 3
Performance of different optimization algorithms and GAN.

Method	Number of samples	Accuracy (%)	Recall (%)	Precision (%)	F1 (%)	Time (s)
Proposed model	570	86.91	86.89	87.11	87.01	4.04
	1140	91.47	91.42	91.77	91.61	4.80
	1710	94.53	94.51	94.27	94.39	7.53
	2280	**96.29**	**96.24**	**96.38**	**96.30**	**8.92**
RIME [23]	570	85.11	85.08	85.42	85.25	3.88
	1140	90.91	90.89	91.23	91.06	4.55
	1710	92.43	92.41	92.63	92.52	7.28
	2280	94.44	94.40	94.74	94.48	8.87
PO [24]	570	83.39	83.35	83.62	83.50	4.05
	1140	88.69	88.63	88.87	88.77	4.73
	1710	90.39	90.36	90.70	90.53	7.42
	2280	93.49	93.44	94.77	93.43	9.09
BKA [25]	570	83.24	83.21	83.44	83.33	4.11
	1140	88.94	88.93	89.28	89.01	4.79
	1710	91.24	91.23	91.53	91.37	7.47
	2280	93.34	93.32	94.40	93.32	9.17
CPO [26]	570	81.74	81.71	81.95	81.83	4.33
	1140	86.54	86.52	86.78	86.66	4.89
	1710	89.24	89.23	89.48	89.36	5.06
	2280	92.74	92.71	93.95	92.53	10.42
FVIM [27]	570	78.54	78.52	78.81	78.66	4.52
	1140	82.04	82.03	82.23	82.12	5.14
	1710	86.54	86.53	86.81	86.65	8.52
	2280	89.74	89.71	90.10	89.75	11.08
NRBO [28]	570	77.18	77.17	77.45	77.30	4.23
	1140	81.78	81.76	81.94	81.85	4.90
	1710	86.27	86.24	86.53	86.39	8.04
	2280	88.49	88.45	88.60	88.50	10.15
BFO [29]	570	76.81	76.79	77.04	76.91	4.17
	1140	81.91	81.88	82.21	82.05	4.82
	1710	85.51	85.50	85.23	85.36	7.53
	2280	87.81	87.79	88.08	87.71	9.28

4.4. Analysis of models based on SHAP method

Previous comparative experiments have revealed that data quality is more significant than model parameter tuning in fault diagnosis. To further explore the impact of the data itself on diagnostic outcomes, SHAP (SHapley Additive exPlanations) technology is employed for in-depth data analysis. Fig. 7 clearly illustrates the importance of five key features: C_2H_4, C_2H_2, H_2, CH_4, and C_2H_6. C_2H_2 (acetylene) and C_2H_4 (ethylene) have the highest mean SHAP values. This indicates that the proposed deep learning model primarily leverages these two features to differentiate between transformer fault types. In the context of DGA, the high contribution of acetylene and ethylene is consistent with their known diagnostic significance: acetylene is a key indicator of arcing and high-energy discharges, while ethylene is typically associated with high-temperature overheating and hot spot formation. This validates the model's ability to learn and apply domain-specific knowledge embedded in these gas signatures.

To investigate each feature's impact on fault diagnosis further, SHAP beeswarm was used to demonstrate the individual contribution of the model prediction, as shown in Fig. 8 [33]. Fig. 7 presents the global average contribution of each dissolved gas, whereas Fig. 8 illustrates the sample-level distribution of gas features across different transformer operating conditions. C_2H_2 (acetylene) and C_2H_4 (ethylene) consistently exhibit the highest SHAP contributions across multiple severe fault classes, such as arc discharge (AD), spark discharge (SD), and high-energy discharge (HD). This is consistent with well-established diagnostic criteria in dissolved gas analysis, which highlight C_2H_2 as a key indicator of high-energy arcing faults and C_2H_4 as an indicator of thermal faults [34].

To further investigate each feature's impact on fault diagnosis, SHAP beeswarm was used to demonstrate the individual contribution of the model prediction, as shown in Fig. 9. C_2H_4 and C_2H_2 exhibited the most significant influence, with their SHAP values widely spread and higher

Table 4
Comparison of performance with other methods.

Method	Parameter setting	Accuracy (%)
Proposed model	**CNN**: {Convolutional kernel size: 5 × 5; Number of convolution kernels: 16; Activation function: ReLU} **BiLSTM**: {Hidden units: 66.4447; Output mode: last} **Attention**: {Heads: 2} **GOOSE:** {Max search agent: 30; Learning rate: 0.01; Maximum iterations: 50} **Training setup:** {Batch size: 50; Epoch: 50; Activation function: Softmax; Optimizer: Adam}	**96.29**
FCNN-Attention-BiLSTM [6]	**FCNN**: {Three 1D convolutional layers; Filters: 128, 64, 32 (length = 1); Stride = 1; Activation function: ReLU; Batch normalization after each layer; Dropout = 0.3} **Attention**: {Dense layer with Softmax activation; Dropout = 0.3} **BiLSTM**: {512 hidden units; Bidirectional processing; Dropout = 0.3} **Training setup**: {Input: 3-phase grounding current (time_steps = 28 after preprocessing); Output: 4 fault classes; Optimizer: Adam; Loss function: categorical cross-entropy; Batch size: 8; Epochs: 40; Learning rate schedule: 1e − 3 (0–10 epochs)}	95.00
BPNN-SVM-Residual Network [30]	**Residual BPNN**: {7 residual modules; Each module = 2 BP layers; Activation: ReLU; Initial weights: Gaussian distribution (mean = 0, std = 0.1); Bias = 0.01; Learning rate = 0.0001; Training epochs = 250} **SVM integration**: {Embedded in residual modules; Selects high-accuracy vectors and increases their weights; Final eigenvector with highest cumulative weight is chosen for diagnosis} **Training setup**: {Training set/test set ratios: 6:4, 7:3, 8:2; Small-sample training verified; Optimization: stochastic gradient descent with backpropagation}	92.70
KPCA and TISOA-SVM [31]	**KPCA**: {Feature extraction from DGA data (H_2, CH_4, C_2H_6, C_2H_4, C_2H_2); Kernel function: RBF; Kernel width = 8; Principal component contribution rate: 95 %; First four principal components extracted as input features} **TISOA**: {Based on Seagull Optimization Algorithm (SOA) with three improvements: (1) Modified Tent map (MTent) for population initialization to enhance diversity; (2) Nonlinear inertia weight to improve convergence speed; (3) Random double helix foraging formula to improve optimization accuracy} **SVM**: {Kernel: RBF; Optimized parameters: C and σ; Optimization method: TISOA; Training/test ratio = 2:1; Normalization: (0, 1)} **Training setup**: {Population size = 30; Max iterations = 100; Parameter bounds: [10^{-3}, 10^{3}]; Evaluation: diagnosis accuracy, time, and convergence performance}	91.67
Optimized ANN [32]	**ANN**: {3 layers; Layer sizes: 100–50–5; Activation function: Tanh; Input features: 4 gas ratios; Output: 5 fault classes} **Training setup**: {Dataset: 400 samples (350 training, 50 testing)} **Optimization methods**: {Hyperparameter search strategies: grid search, Bayesian optimization, random search, manual search; Optimizers compared: Adam, SGDM, RMSprop; Best optimizer: Adam (Learning rate = 0.001; Gradient Decay Factor = 0.999; Batch size = 64; Epochs = 30)}	90.00

(*continued on next page*)

Table 4 (*continued*)

Method	Parameter setting	Accuracy (%)
GCN [7]	Regularization: {L2 regularization applied; λ optimized to reduce overfitting; Gradient threshold method: l2norm; Early stopping based on validation loss} **GraphConv-1**: {Filters: 16; Activation: ReLU; Output shape: 1 × 16} **Dropout-1**: {Rate: 0.25; Output shape: 1 × 16} **GraphConv-2**: {Filters: 8; Activation: ReLU; Output shape: 1 × 8} **Dropout-2**: {Rate: 0.25; Output shape: 1 × 8} Dense (Output layer): {Units: 7; Activation: Softmax; Output shape: 1 × 7}	89.70

Table 5
T-test with other methods.

Methods	T value	P value	Average difference	Standard error	95 % confidence interval of difference
GCN [7]	3.036	0.0024	6.12 %	2.016 %	(2.169 %, 10.071 %)
Optimized ANN [32]	2.765	0.0057	5.05 %	1.827 %	(1.470 %, 8.630 %)
KPCA and TISOA-SVM [31]	2.597	0.0094	3.57 %	1.375 %	(0.876 %, 6.264 %)
BPNN-SVM-Residual Network [30]	2.313	0.0207	2.76 %	1.193 %	(0.422 %, 5.098 %)
FCNN-Attention-BiLSTM [6]	2.256	0.0241	1.01 %	0.448 %	(0.132 %, 1.888 %)

concentrations (red points) corresponding to positive SHAP values, indicating increased prediction scores. H_2 showed a moderate impact with a less variable SHAP value distribution. CH_4 and C_2H_6 had minimal influence, as their SHAP values were tightly clustered around zero, with balanced colour distributions, suggesting negligible effects on the model's predictions regardless of their concentrations. This analysis underscored the critical role of C_2H_4 and C_2H_2 in the model's decision-making process, while CH_4 and C_2H_6 contributed little to the predictive outcomes.

To further investigate the impact of the most critical features on each type of fault, a detailed visualization of the contributions of C_2H_4 and C_2H_2 across different classes was presented, as shown in Fig. 10. Each point in the plots represented a SHAP value for a specific feature, with blue indicating low feature values and red indicating high feature values. For C_2H_4, the SHAP values were widely distributed across various classes, and except for high-temperature overheating, they exhibited a noticeable positive contribution to other faults, suggesting significant variability in their contribution to the model's predictions. In contrast, for C_2H_2, the SHAP values also displayed a broad distribution. However, the majority were negatively contributing, indicating that higher concentrations of C_2H_2 were generally associated with a decrease in the prediction score for most fault types.

5. Model generalization Verification

To evaluate the diagnostic performance of the proposed model in practical transformer applications and assess its robustness under highly imbalanced data conditions, real DGA data from a 15,500 kV three-phase transformer were utilized [35]. This real-world dataset comprises 15 normal operation samples and several verified fault cases, as summarized in Table 7.

The collected samples were input into the proposed diagnostic model and benchmarked against conventional methods: the IEC Three Ratio Method and the Duval Triangle Method. As shown in the results, the traditional approaches struggled with diagnostic clarity in most cases. Specifically, the Duval method labelled 10 out of 15 cases as "Unrecognizable," achieving an overall accuracy of only 13.33 %, while the IEC method exhibited limited precision with an accuracy of 73.33 %. In contrast, the proposed model achieved 100 % diagnostic accuracy, correctly identifying both normal and fault types—including low/medium temperature faults (LMT), high-temperature faults (HT), and partial discharge (PD)—even in complex or borderline conditions. These findings highlight the proposed model's superior interpretability, adaptability, and fault classification capacity when applied to real

Table 6
Ablation experiment.

Method	Accuracy (%)	Recall (%)	Precision (%)	F1 (%)
GAN-CNN-BiLSTM-Attention-GOOSE	**96.29**	**96.24**	**96.38**	**96.30**
CNN-BiLSTM-Attention-GOOSE	86.91	86.89	87.11	87.01
CNN-BiLSTM-Attention	76.38	76.34	76.52	76.48
CNN-BiLSTM	74.15	74.13	74.45	74.37
CNN	71.58	71.55	71.63	71.59

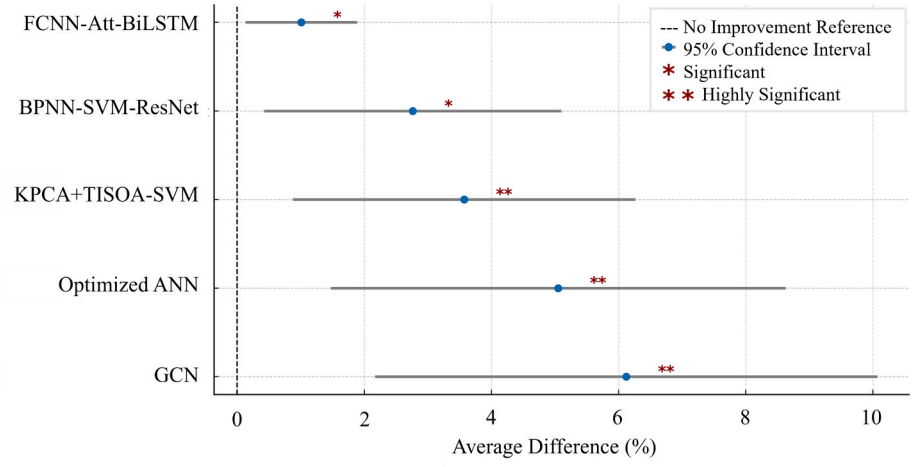

Fig. 6. Model performance improvement with 95% confidence intervals.

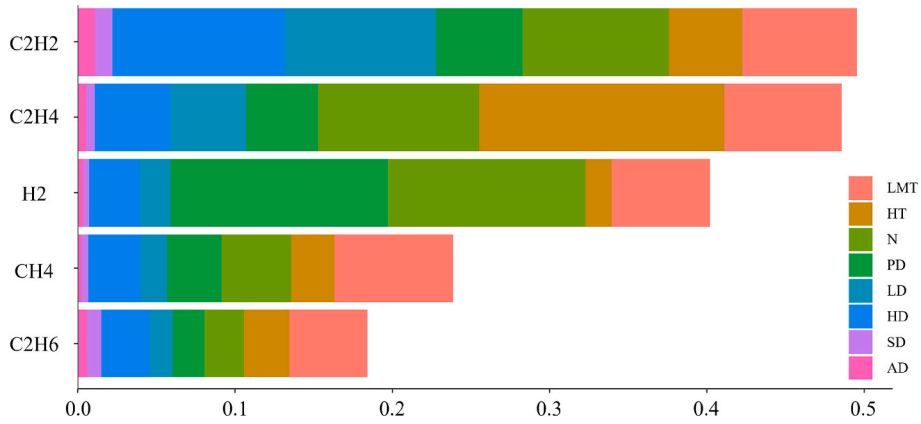

Fig. 7. Global feature importance of dissolved gas analysis dataset.

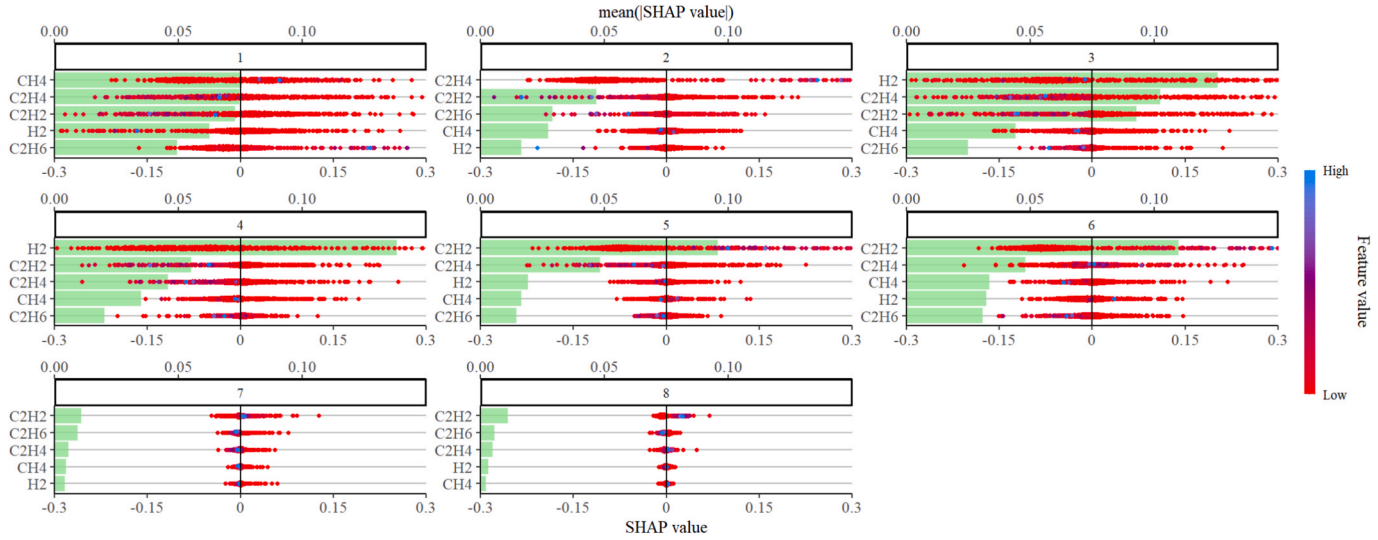

Fig. 8. SHAP beeswarm plot of dissolved gas analysis dataset.

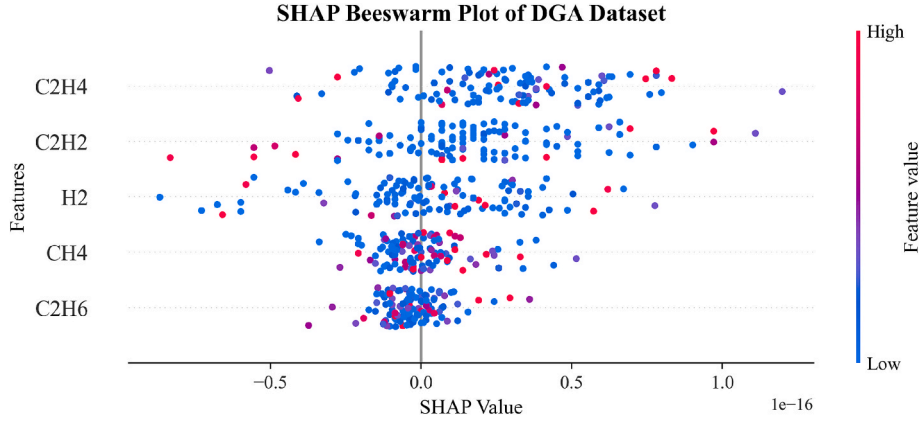

Fig. 9. SHAP beeswarm plot of five key features.

transformer data scenarios.

To further validate the generalization capability of the proposed model, we employed a more authoritative dataset derived from the IEEE DGA datasets [36]. 201 transformer fault samples were collected, from which 60 samples were randomly selected according to fault categories for testing. The diagnostic results are summarized in Table 8.

As shown in Table 8, several patterns can be identified. The Rogers ratio method produced an overall accuracy of only 60 %, with many cases classified as "Unrecognizable." This outcome is unsurprising, since the method relies on fixed ratio thresholds between dissolved gases, which are often too rigid to cope with overlapping or ambiguous gas distributions. When applied to real operating data with more complex patterns, such rules fail to provide reliable results. The implication is that, while simple, ratio-based methods lack robustness and cannot be

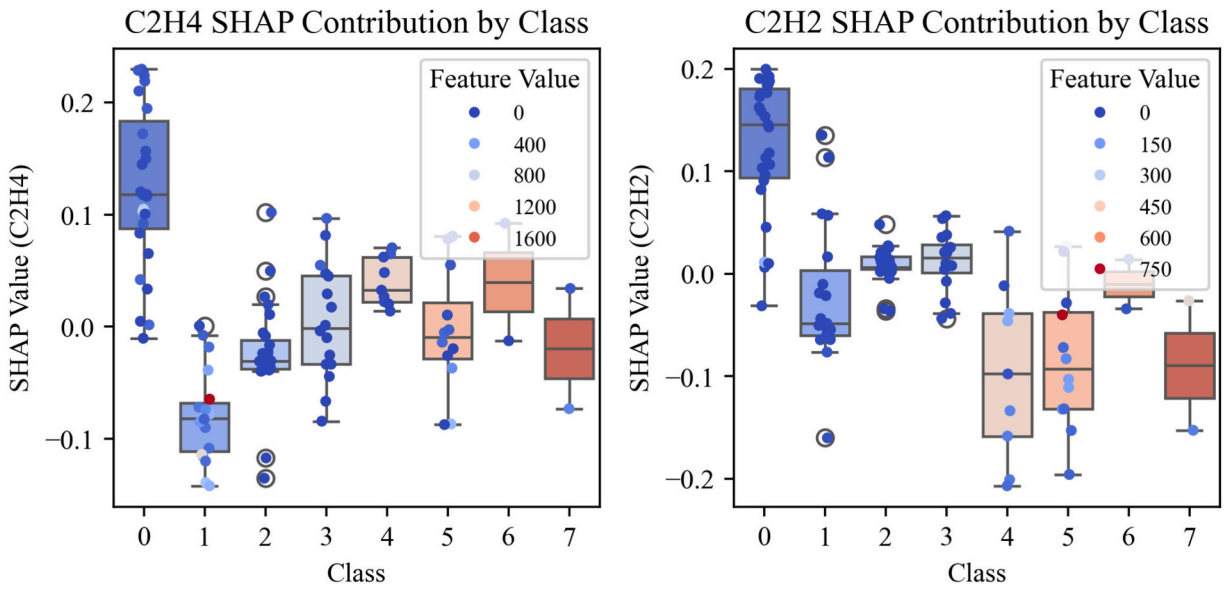

Fig. 10. The SHAP values contributed by C_2H_4 and C_2H_2 by class.

Table 7
Transformer real case fault diagnosis results.

Sample number	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	Fault type	IEC Three ratio method	Duval triangle method	Predict outcomes
1	1.92	4.25	1.98	0.5	0	N	N (√)	Unrecognizable (×)	N (√)
2	2.14	4.4	1.91	0.53	0	N	N (√)	Unrecognizable (×)	N (√)
3	2.07	4.35	1.82	0.52	0	N	N (√)	Unrecognizable (×)	N (√)
4	2.04	4.05	2.19	0.47	0	N	N (√)	Unrecognizable (×)	N (√)
5	1.88	3.85	2.14	0.52	0	N	N (√)	Unrecognizable (×)	N (√)
6	1.9	4.36	2.1	0.52	0	N	N (√)	Unrecognizable (×)	N (√)
7	2.16	4.28	2.03	0.5	0	N	N (√)	Unrecognizable (×)	N (√)
8	2.12	4.18	2.09	0.54	0	N	N (√)	Unrecognizable (×)	N (√)
9	1.9	4.36	2.1	0.52	0	N	N (√)	Unrecognizable (×)	N (√)
10	2.15	3.7	1.82	0.51	0	N	N (√)	Unrecognizable (×)	N (√)
11	28.8	70.1	23.3	163.7	1.4	LMT	HT (×)	HT (×)	LMT(√)
12	43.8	76.9	105	44.8	0.08	LMT	LMT (√)	LMT (√)	LMT(√)
13	8.2	26	1003	2194	4.2	HT	LMT (×)	HT (√)	HT (√)
14	24,101	5813	311	15	0	PD	LMT (×)	HT (×)	PD (√)
15	10,425	803	0.8	22.6	0	PD	HT (×)	HT (×)	PD (√)
Accuracy							73.33 %	13.33 %	100 %

expected to generalize well.

In contrast, the proposed model reached an accuracy of 93.33 %. The improvement stems from its ability to capture nonlinear dependencies among multiple gas features, instead of relying on a few predefined ratios. This allows the model to handle cases where the gas compositions deviate from traditional rule boundaries. The result suggests that learning-based methods incorporating feature interactions are far more effective for fault diagnosis in practical scenarios.

6. Conclusion

To enhance the accuracy of power transformer fault diagnosis, this study proposed a novel diagnostic framework, GAN-CNN-BiLSTM-Attention-GOOSE, which integrates data augmentation, deep learning, and optimization techniques. The base CNN-BiLSTM-Attention model achieved an initial accuracy of 76.38 %. Incorporating a Generative Adversarial Network significantly improved data diversity and quality, resulting in an average performance gain of 9.32 % across key evaluation metrics. Furthermore, applying the GOOSE optimization algorithm for hyperparameter tuning yielded an additional 10.53 % increase in accuracy, outperforming seven benchmark optimization methods and demonstrating its superiority in refining model parameters. Model interpretability was enhanced through SHAP analysis, which identified C_2H_2 and C_2H_4 as the most influential dissolved gas features, aligning with known indicators of arcing and thermal faults and confirming the model's alignment with engineering domain knowledge. The transformer fault diagnosis model proposed in this paper achieved an accuracy of 96.29 %, exhibiting commendable diagnostic performance. It could quickly and accurately diagnose transformer faults and outperform models presented in recent literature in terms of accuracy. This advancement offered a high reference value for the field of transformer fault diagnosis. In addition, a series of results have shown that C_2H_4 and C_2H_2 contribute more to distinguishing transformer fault types among the five key features. However, the research on DGA data in this paper was not exhaustive. Future work will involve more in-depth analysis and study of DGA data feature engineering to further improve the accuracy and stability of fault diagnosis models.

Future work will focus on several promising directions to further advance transformer fault diagnosis and condition assessment. Advanced feature engineering strategies, including ratio-based gas indicators, temporal trend analysis, and composite feature design, will be explored to enrich the representation of DGA data. In addition, cross-domain generalization methods will be developed and validated on datasets from different transformer fleets and operating environments to

Table 8
Transformer real case fault diagnosis results.

(1) Number	(2) H_2	(3) CH_4	(4) C_2H_6	(5) C_2H_4	(6) C_2H_2	(7) Type	(8) Rogers	(9) Result	(10) Ours	(11) Result
1	13	138	83	16	0	1	1	√	1	√
2	762	93	38	54	126	6	6	√	6	√
3	43	116	65	139	0	1	1	√	1	√
4	179	306	73	579	0	1	1	√	1	√
5	57	141	38	51	0	1	1	√	1	√
6	40	8	34	15	0	1	1	√	1	√
7	35	283	121	222	0	1	Unrecognizable	×	3	×
8	15	159	29	87	0	1	Unrecognizable	×	1	√
9	55	159	114	493	0	1	1	√	1	√
10	37	123	67	52	0	1	1	√	1	√
11	723	191	110	293	288	6	6	√	6	√
12	7	15	78	58	0	1	1	√	1	√
13	30	51	12	54	0	1	1	√	1	√
14	31	56	33	77	0	1	1	√	1	√
15	109	226	68	192	0	1	1	√	1	√
16	137	279	66	505	0	1	1	√	1	√
17	59	119	36	70	0	1	1	√	1	√
18	151	242	68	232	0	1	1	√	1	√
19	870	77	73	54	14	5	5	√	5	√
20	376	575	146	1092	0	1	1	√	1	√
21	269	1081	347	1725	25	2	Unrecognizable	×	2	√
22	10	10	8	1	0.01	1	1	√	1	√
23	30	22	14	4.10	0.1	3	3	√	3	√
24	2.90	2	2	0.3	0.1	3	3	√	3	√
25	4	99	82	4	0.1	1	1	√	1	√
26	21	34	5	47	62	6	Unrecognizable	×	6	√
27	50	100	51	305	9	1	1	√	1	√
28	120	17	32	4	23	3	Unrecognizable	×	3	√
29	980	73	58	12	0.01	5	5	√	5	√
30	1607	615	80	916	1294	6	Unrecognizable	×	6	√
(1) Number	(2) H_2	(3) CH_4	(4) C_2H_6	(5) C_2H_4	(6) C_2H_2	(7) Type	(8) Rogers	(9) Result	(10) Ours	(11) Result
31	14.7	3.7	10.5	2.7	0.2	1	Unrecognizable	×	1	√
32	181	262	41	28	0.01	1	Unrecognizable	×	1	√
33	173	334	172	812.5	33.7	1	1	√	1	√
34	127	107	11	154	224	6	1	×	1	×
35	60	40	6.9	110	70	6	1	×	6	√
36	980	73	58	12	0.01	5	6	×	5	√
37	86	187	136	363	0.01	1	6	×	1	√
38	10	24	372	24	0.01	1	Unrecognizable	×	1	√
39	260	3	18	2	0.01	5	5	√	5	√
40	586	19	77	6	0.01	5	1	×	5	√
41	20	175	92	14	0.02	1	1	√	1	√
42	801	87	45	62	150	6	6	√	6	√
43	51	99	75	150	0.03	1	1	√	1	√
44	200	298	69	602	0.05	2	1	×	2	√
45	60	154	41	49	0	1	1	√	1	√
46	40	8	34	15	0.2	3	1	×	3	√
47	45	283	158	199	0	1	Unrecognizable	×	1	√
48	21	159	22	91	0.02	1	Unrecognizable	×	3	×
49	55	159	128	502	0	2	1	×	2	√
50	41	223	71	52	0	6	1	×	6	√
51	689	203	129	301	362	5	6	×	5	√
52	10	24	95	45	0.02	1	1	√	1	√
53	45	69	7	45	0.003	2	1	×	2	√
54	45	59	45	89	0.01	1	1	√	1	√
55	98	198	70	201	0.04	1	1	√	1	√
56	204	302	57	495	0	2	1	×	1	×
57	45	125	48	82	0	1	1	√	1	√
58	201	256	54	224	0	1	1	√	1	√
59	905	83	81	63	12	5	5	√	5	√
60	402	604	99	998	0.02	2	1	×	2	√
Accuracy								60%		93.33%

improve robustness under diverse conditions. Finally, developing online diagnostic capabilities that combine real-time monitoring with adaptive learning will be pursued to support timely, practical, and scalable deployment in smart grid applications.

CRediT authorship contribution statement

Jiajian Lin: Writing – review & editing, Writing – original draft, Visualization, Methodology, Conceptualization. **Lit Yen Yeo:** Writing – original draft, Methodology. **Hadi Nabipour Afrouzi:** Supervision. **Mehran Motamed Ektesabi:** Supervision. **Jalal Tavalaei:** Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Agarwala A, Tahsin T, Ali MF, Sarker SK, Abhi SH, Das SK, et al. Towards next generation power grid transformer for renewables: Technology review. Eng Rep 2024;6. https://doi.org/10.1002/eng2.12848.

[2] Soni R, Mehta B. A review on transformer condition monitoring with critical investigation of mineral oil and alternate dielectric fluids. Electr Pow Syst Res 2023;214:108954. https://doi.org/10.1016/j.epsr.2022.108954.

[3] Soni R, Mehta B. Review on asset management of power transformer by diagnosing incipient faults and faults identification using various testing methodologies. Eng Fail Anal 2021;128:105634. https://doi.org/10.1016/j.engfailanal.2021.105634.

[4] Soni R, Mehta B. Evaluation of power transformer health analysis by internal fault criticalities to prevent premature failure using statistical data analytics approach. Eng Fail Anal 2022;136:106213. https://doi.org/10.1016/j.engfailanal.2022.106213.

[5] Taha IBM. Power Transformers Health Index Enhancement based on Convolutional Neural Network after applying Imbalanced-Data Oversampling. Electronics (Basel) 2023;12:2405. https://doi.org/10.3390/electronics12112405.

[6] Mao W, Wei B, Xu X, Chen L, Wu T, Peng Z, et al. Fault Diagnosis for Power Transformers through Semi-Supervised transfer Learning. Sensors 2022;22:4470. https://doi.org/10.3390/s22124470.

[7] Liao W, Yang D, Wang Y, Ren X. Fault diagnosis of power transformers using graph convolutional network. CSEE Journal of Power and Energy Systems 2021;7:241–9. 10.17775/CSEEJPES.2020.04120.

[8] Machlev R. EV battery fault diagnostics and prognostics using deep learning: Review, challenges & opportunities. J Energy Storage 2024;83:110614. https://doi.org/10.1016/j.est.2024.110614.

[9] Sun H-C, Huang Y-C, Huang C-M. Fault Diagnosis of Power Transformers using Computational Intelligence: a Review. Energy Procedia 2012;14:1226–31. https://doi.org/10.1016/j.egypro.2011.12.1080.

[10] He J, Huang W, Liu Y, Qian C, Ma C, Gao W, et al. Data imbalance fault diagnosis method based on an ensemble multi-scale convolutional attention network. Mech Syst Sig Process 2025;236:112934. https://doi.org/10.1016/j.ymssp.2025.112934.

[11] Qi J, Chen Z, Kong Y, Qin W, Qin Y. Attention-guided graph isomorphism learning: a multi-task framework for fault diagnosis and remaining useful life prediction. Reliab Eng Syst Saf 2025;263:111209. https://doi.org/10.1016/j.ress.2025.111209.

[12] Wang J, Abdullah S, Gao C, Arifin A, Sing SSK. FreqMGCN-Net: an IoT-Integrated Multi-Parallel Graph Convolutional Network with frequency attention for motor bearing fault diagnosis. Alex Eng J 2025;128:175–85. https://doi.org/10.1016/j.aej.2025.05.019.

[13] Wang Y, Lu Z, Zhiwen Z, Liu C, Yu J. Multiple path alignment generative adversarial network for rotating Machinery fault diagnosis with limited data. Adv Eng Inf 2025;67:103550. https://doi.org/10.1016/j.aei.2025.103550.

[14] Yang X-S. Nature-inspired optimization algorithms: challenges and open problems. J Comput Sci 2020;46:101104. https://doi.org/10.1016/j.jocs.2020.101104.

[15] Li Y, Liu X, Hu J, Liang P, Wang B, Yuan X, et al. Graph optimization algorithm enhanced by dual-scale spectral features with contrastive learning for robust bearing fault diagnosis. Knowl Based Syst 2025;315:113275. https://doi.org/10.1016/j.knosys.2025.113275.

[16] Huang K, Li W, Gao F. Barabási-albert model-enhanced genetic algorithm for optimizing LGBM in ship power grid fault diagnosis. Measurement 2025;249:116954. https://doi.org/10.1016/j.measurement.2025.116954.

[17] Wang C, Yang J, Jie H, Tian B, Zhao Z, Chang Y. An uncertainty perception metric network for machinery fault diagnosis under limited noisy source domain and scarce noisy unknown domain. Adv Eng Inf 2024;62:102682. https://doi.org/10.1016/j.aei.2024.102682.

[18] Wang C, Liu X, Yang J, Jie H, Gao T, Zhao Z. Addressing unknown faults diagnosis of transport ship propellers system based on adaptive evolutionary reconstruction metric network. Adv Eng Inf 2025;65:103287. https://doi.org/10.1016/j.aei.2025.103287.

[19] Wang C, Yang J, Jie H, Zhao Z, Wang W. An Energy-Efficient Mechanical Fault Diagnosis Method based on Neural Dynamics-inspired Metric SpikingFormer for Insufficient Samples in Industrial internet of things. IEEE Internet Things J 2024:1. https://doi.org/10.1109/JIOT.2024.3476034.

[20] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ, editors. Adv Neural Inf Process Syst, vol. 27, Curran Associates, Inc.; 2014.

[21] Hamad RK, Rashid TA. GOOSE algorithm: a powerful optimization tool for real-world engineering challenges and beyond. Evol Syst 2024;15:1249–74. https://doi.org/10.1007/s12530-023-09553-6.

[22] Kullback S, Leibler RA. On Information and Sufficiency. Ann Math Stat 1951;22:79–86. https://doi.org/10.1214/aoms/1177729694.

[23] Su H, Zhao D, Heidari AA, Liu L, Zhang X, Mafarja M, et al. RIME: a physics-based optimization. Neurocomputing 2023;532:183–214. https://doi.org/10.1016/j.neucom.2023.02.010.

[24] Lian J, Hui G, Ma L, Zhu T, Wu X, Heidari AA, et al. Parrot optimizer: Algorithm and applications to medical problems. Comput Biol Med 2024;172. https://doi.org/10.1016/j.compbiomed.2024.108064.

[25] Wang J, Wang WC, Hu XX, Qiu L, Zang HF. Black-winged kite algorithm: a nature-inspired meta-heuristic for solving benchmark functions and engineering problems. Artif Intell Rev 2024;57. https://doi.org/10.1007/s10462-024-10723-4.

[26] Abdel-Basset M, Mohamed R, Abouhawwash M. Crested Porcupine Optimizer: a new nature-inspired metaheuristic. Knowl Based Syst 2024;284. https://doi.org/10.1016/j.knosys.2023.111257.

[27] Fakhouri HN, Awaysheh FM, Alawadi S, Alkhalaileh M, Hamad F. Four vector intelligent metaheuristic for data optimization. Computing 2024;106:2321–59. https://doi.org/10.1007/s00607-024-01287-w.

[28] Sowmya R, Premkumar M, Jangir P. Newton-Raphson-based optimizer: a new population-based metaheuristic algorithm for continuous optimization problems. Eng Appl Artif Intel 2024;128. https://doi.org/10.1016/j.engappai.2023.107532.

[29] Zareian L, Rahebi J, Shayegan MJ. Bitterling fish optimization (BFO) algorithm. Multimed Tools Appl 2024;83:75893–926. https://doi.org/10.1007/s11042-024-18579-0.

[30] Jin Y, Wu H, Zheng J, Zhang J, Liu Z. Power Transformer Fault Diagnosis based on improved BP Neural Network. Electronics (Switzerland) 2023;12. https://doi.org/10.3390/electronics12163526.

[31] Wu Y, Sun X, Zhang Y, Zhong X, Cheng L. A Power Transformer Fault Diagnosis Method-based Hybrid improved Seagull Optimization Algorithm and support Vector Machine. IEEE Access 2022;10:17268–86. https://doi.org/10.1109/ACCESS.2021.3127164.

[32] Rokani V, Kaminaris SD, Karaisas P, Kaminaris D. Power Transformer Fault Diagnosis using Neural Network Optimization Techniques. Mathematics 2023;11. https://doi.org/10.3390/math11224693.

[33] Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg U Von, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Adv Neural Inf Process Syst, vol. 30, Curran Associates, Inc.; 2017.

[34] Duval M, dePabla A. Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases. IEEE Electr Insul Mag 2001;17:31–41. https://doi.org/10.1109/57.917529.

[35] Hu D, Yang Y, Dai H, Tang C, Xie J. An interpretable machine learning method for fault diagnosis of oil-immersed transformers based on edge inference. Int J Electr Power Energy Syst 2025;168:110647. https://doi.org/10.1016/j.ijepes.2025.110647.

[36] Li E. Dissolved gas data in transformer oil—Fault Diagnosis of Power Transformers with Membership Degree 2019. 10.21227/h8g0-8z59.