International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
SJIF (2025): 8.036

A Counterfactual Diagnosis Method for Power Transformers Based on a Differentiable Structural Causal Model

Lichuan Lei

North China Electric Power University, School of Control and Computer Engineering,
No. 2 Beinong Road, Changping District, Beijing, China
Email: *2086562624[at]qq.com*

Abstract: ***With the increasing intelligence of power systems, multimodal monitoring data of transformers have become increasingly abundant. However, existing diagnosis methods mostly rely on statistical correlations, lack explicit causal reasoning capability, and struggle to handle unseen fault combinations. To address these issues, this paper proposes a counterfactual diagnosis method for power transformers based on a differentiable structural causal model (DSCM-CFD). The method explicitly constructs a causal generative mechanism from vibration to temperature to audio, and introduces path consistency loss and counterfactual consistency loss, jointly optimized within a variational autoencoder framework. The model achieves high classification accuracy, accurate root cause localization, and effective zero-shot diagnosis. Experimental results demonstrate that the proposed method outperforms existing methods in diagnostic accuracy, zero-shot generalization, and causal interpretability, providing a new technical pathway for intelligent operation and maintenance of power equipment.***

Keywords: causal reasoning; differentiable structural causal model; multimodal fusion; power transformer fault diagnosis

1. Introduction

The safe and stable operation of power systems depends on the condition monitoring and accurate diagnosis of key equipment such as transformers [1]. In recent years, multi-source sensors can synchronously acquire multimodal data including vibration/current time series, temperature time series, and audio spectra, providing a data foundation for intelligent diagnosis [2]. Traditional single-modal methods cannot fully capture complex fault characteristics, and although deep learning methods have made progress, most are still based on single-modal data [3].

Multimodal fusion has become an important research direction. Transformer-based multimodal data fusion methods have shown promising potential in fault diagnosis [4], but they primarily learn statistical correlations rather than causal relationships, making them vulnerable to distribution shifts. In fact, power equipment faults follow a clear causal chain: mechanical wear → abnormal vibration → temperature rise → audio spectrum change. Existing methods fail to explicitly model this causal structure, lacking true reasoning ability and interpretability [5].

To overcome these limitations, this paper proposes a counterfactual diagnosis method for power transformers based on a differentiable structural causal model (DSCM-CFD). The main contributions are: 1) constructing a causal generative architecture that transforms diagnosis into causal intervention and counterfactual reasoning; 2) designing path consistency loss and counterfactual consistency loss to enforce causal chain transitivity and representation stability; 3) achieving causal effect evaluation and root cause localization; 4) realizing zero-shot diagnosis through counterfactual reasoning.

2. Related Work

Intelligent Fault Diagnosis: Early methods relied on mechanistic models. For example, Tavner and Penman [1] systematically summarized the theoretical foundations of condition monitoring for electrical machines. Machine learning methods such as Simões et al. [6] combined support vector machines with wavelet transform for transformer fault detection. Among deep learning methods, Jia et al. [3] reported the powerful feature extraction capability of deep neural networks for rotating machinery fault diagnosis, but most of the above methods are based on single-modal data.
Multimodal Fusion: Multimodal fusion significantly improves diagnostic performance. Baltrusaitis et al. [2] reviewed the trends in multimodal machine learning, pointing out that fusing heterogeneous sensing modalities enhances decision robustness. Ramachandram and Taylor [7] emphasized the importance of deep multimodal learning. Representative methods include the Transformer-based zero-shot fault diagnosis method proposed by Huang et al. [4] and the multimodal Transformer by Tsai et al. [8]. However, these methods are essentially correlation modeling and lack causal constraints.

Application of Causal Reasoning in Fault Diagnosis: Li et al. [5] systematically reviewed the progress of causal inference in industrial fault diagnosis, noting that causal reasoning can reveal fault propagation paths and improve interpretability. Yang et al. [9] proposed CausalVAE, which combines structural causal models with variational autoencoders, but mostly as an independent module. This paper deeply integrates causal generation with multimodal fusion to achieve counterfactual reasoning and zero-shot diagnosis.

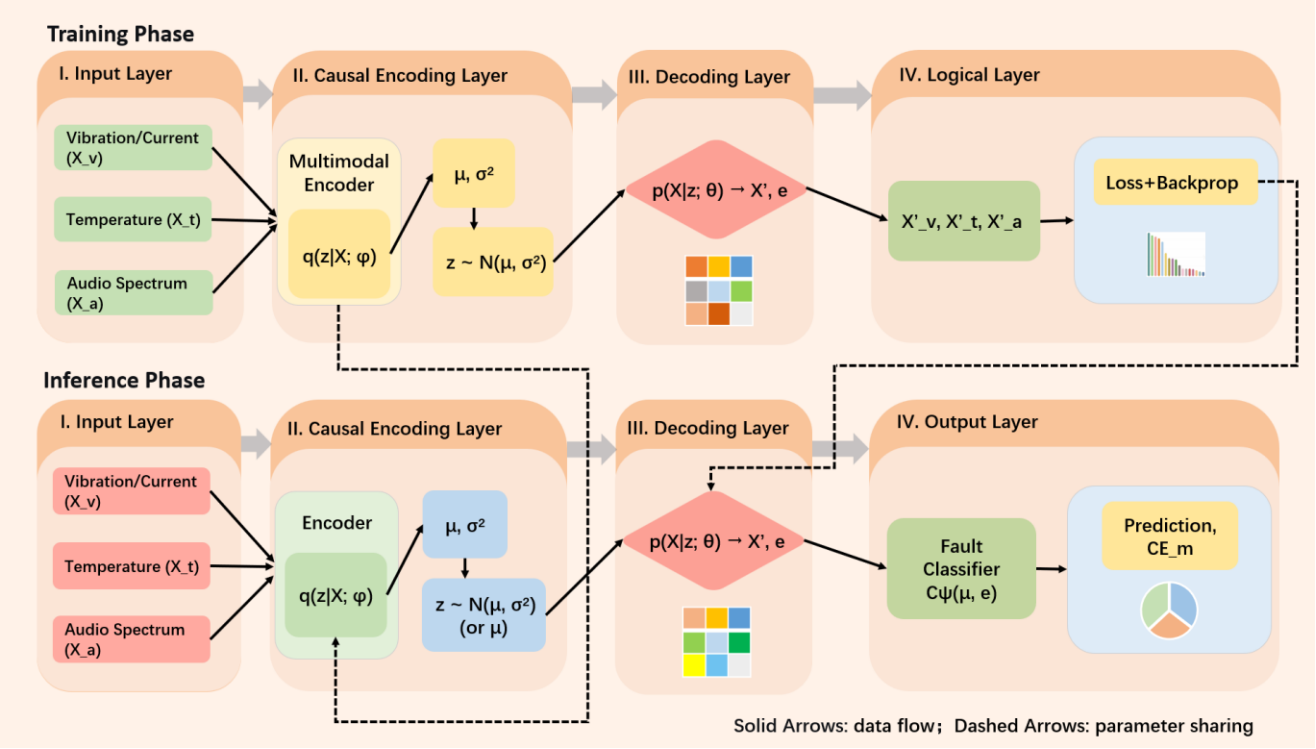

Figure 1: Overall architecture of DSCM-CFD

3. Proposed method

Problem Formulation. Consider three modalities: vibration/current $\mathbf{X}^{(1)} \in \mathbb{R}^{T\times d_1}$ (where T is the time window length and d_1 the feature dimension), temperature time series $\mathbf{X}^{(2)} \in \mathbb{R}^{T\times d_2}$, and audio spectrum features $\mathbf{X}^{(3)} \in \mathbb{R}^{d_3}$. The label is $y \in \{0,1,\dots,C\}$, where C is the number of fault classes. There exists a latent causal variable $\mathbf{z} \in \mathbb{R}^K$ (with K the latent dimension), generated in the causal order $\mathbf{z} \to \mathbf{X}^{(1)} \to \mathbf{X}^{(2)} \to \mathbf{X}^{(3)}$.

Overall Architecture. The model consists of an encoder $q_\phi(\mathbf{z} \mid \mathbf{X})$ (parameters ϕ), a generator $p_\theta(\mathbf{X} \mid \mathbf{z})$ (parameters θ), a classifier $\mathcal{C}_\psi$ (parameters ψ), and an intervention/counterfactual module. Using a variational autoencoder (VAE) as the framework, we add path consistency loss and counterfactual consistency loss. Figure 1 shows the data flow during training and inference.

Structural Causal Model Parameterization. The prior of the latent variable is a standard normal distribution:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}_K) \tag{1}$$

The structural equations of the generator are as follows, where $\epsilon_1, \epsilon_2, \epsilon_3$ are independent exogenous noise variables (following standard normal distributions with dimensions matching the corresponding modalities), and f_1, f_2, f_3 are neural network parameterized functions:

$$\mathbf{X}^{(1)} = f_1(\mathbf{z}, \epsilon_1), \mathbf{X}^{(2)} = f_2(\mathbf{z}, \mathbf{X}^{(1)}, \epsilon_2), \mathbf{X}^{(3)} = f_3(\mathbf{X}^{(2)}, \epsilon_3) \tag{2}$$

The observation noise is assumed to be Gaussian with variance σ_m^2 as a hyperparameter. The encoder outputs the posterior Gaussian distribution:

$$q_\phi(\mathbf{z} \mid \mathbf{X}) = \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{X}), \mathrm{diag}\,(\boldsymbol{\sigma}_\phi^2(\mathbf{X}))) \tag{3}$$

where $\boldsymbol{\mu}_\phi(\cdot)$ and $\boldsymbol{\sigma}_\phi(\cdot)$ are outputs of neural networks, representing the posterior mean and standard deviation, respectively.

Training Objectives. The variational lower bound (ELBO) loss consists of reconstruction loss and KL divergence:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{rec}} + \beta \mathcal{L}_{\text{KL}} \tag{4}$$

where $\mathcal{L}_{\text{rec}}$ is the expected reconstruction error, $\mathcal{L}_{\text{KL}}$ is the KL divergence between the posterior and the prior, and β is a balancing coefficient. The intervention consistency loss $\mathcal{L}_{\text{int}}$ enforces the single-step causal edges $1 \to 2$ and $2 \to 3$. The classification loss $\mathcal{L}_{\text{cls}}$ is cross-entropy. The path consistency loss is defined as:

$$\mathcal{L}_{\text{path}} = \| \Delta_{v\to a} - \Delta_{v\to t} \odot \Delta_{t\to a} \|_2^2 \tag{5}$$

Here, $\Delta_{v\to a}$ denotes the change in audio caused by direct intervention on vibration, $\Delta_{v\to t}$ the change in temperature caused by vibration intervention, $\Delta_{t\to a}$ the change in audio caused by temperature intervention, and $\odot$ the Hadamard product. This loss enforces the transitivity of the causal chain $X_v \to X_t \to X_a$. The counterfactual consistency loss is:

$$\mathcal{L}_{\text{cf}} = \mathbb{E}_{\mathbf{z},k}[\textstyle\sum_{j\neq k} \| z_j - z_{\text{enc},j}^{cf} \|_2^2] \tag{6}$$

where k is the intervened latent dimension, z_j is the j-th dimension of the factual encoding, and $z_{\text{enc},j}^{cf}$ is the j-th dimension after re-encoding the counterfactual sample. This loss ensures that when a single causal factor is intervened, other factors remain unchanged. The total loss function is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{VAE}} + \lambda_{\text{int}}\mathcal{L}_{\text{int}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{path}}\mathcal{L}_{\text{path}} + \lambda_{\text{cf}}\mathcal{L}_{\text{cf}} \tag{7}$$

where $\lambda_{\text{int}}, \lambda_{\text{cls}}, \lambda_{\text{path}}, \lambda_{\text{cf}}$ are hyperparameters.

Causal Intervention Diagnosis. For a test sample, define the

causal effect CE_m of modality m as the absolute change in fault probability after intervening on that modality. The modality with the largest CE_m is identified as the root cause.

Counterfactual Reasoning and Zero-shot Diagnosis. Counterfactual samples are generated in three steps: abduction (encoding to obtain $\mathbf{z}$), action (modifying a modality or a latent dimension), and prediction (re-generating descendant modalities). For zero-shot diagnosis, missing causal chains are completed via counterfactual reasoning and then fed into the classifier.

Theoretical Comparison. Table 1 summarizes the characteristics of different methods.

Table 1: Theoretical comparison of different diagnosis *methods*

Method	Causal Utilization	Zero-shot Capability	Output Explanation
Traditional attention fusion	None	No	Attention weights
Causal attention network	Causal graph mask	No	Causal graph
DSCM-CFD (proposed)	Structural causal model + path consistency + counterfactual consistency	Yes	Causal effect + counterfactual sample

4. Experiments

Dataset. The experimental dataset is a private, non-public collection acquired from an online monitoring system of power transformers at a substation. The monitoring system employs accelerometers, temperature sensors, and audio sensors to synchronously capture vibration/current signals, temperature signals, and audio signals. Each recording lasts 10 seconds with a sliding step of 2 seconds, yielding a total of 10,116 valid samples. Each sample consists of a vibration/current time window, a temperature average, and 13-dimensional MFCC audio features. The data cover normal condition and four fault types (bearing wear, electrical fault, cooling failure, and insulation degradation), and are split into training, validation, and test sets in an 8:1:1 ratio while preserving class proportions.

Experimental Setup. Hardware: NVIDIA RTX 3080 GPU; Software: PyTorch 2.3.1. Latent dimension $K = 8$, learning rate 10^{-3}, batch size 64. Hyperparameters: $\beta = 1$, $\lambda_{\text{int}} = 0.1$, $\lambda_{\text{cls}} = 1$, $\lambda_{\text{path}} = 0.5$, $\lambda_{\text{cf}} = 0.5$. Comparison methods include: Single-Best (vibration), Concat-MLP, Adv-Domain, CausalVAE [9], and CF-CL.

Classification Performance. Table 2 shows the classification results on the test set. DSCM-CFD achieves 95.02% accuracy and 95.76% F1-score, outperforming all baselines. Compared to the best baseline CF-CL (89.23% accuracy), the proposed method improves by about 5.8 percentage points. The confusion matrix shows that the misclassification rate for each fault type is below 5%, with clear separation between easily confused faults such as cooling failure and insulation degradation.

Table 2: Classification performance of different methods on the test set (%)

Method	Accuracy	Recall	F1-score
Single-Best (vibration)	71.22	71.15	72.15
Concat-MLP	80.71	80.30	80.32
Adv-Domain	81.34	82.50	82.10
CausalVAE	85.07	86.15	86.92
CF-CL	89.23	88.40	89.11
DSCM-CFD (proposed)	95.02	94.15	95.76

Zero-shot Diagnosis. Three unseen scenarios were constructed: only vibration abnormal, only temperature abnormal, and only audio abnormal. Figure 2 shows the zero-shot diagnosis accuracy. DSCM-CFD achieves 86.7%, 88.2%, and 85.4% in the three scenarios, respectively, while all baselines are below 75% (Concat-MLP ~50%, CausalVAE ~69%). This verifies the fundamental advantage of generative causal models in zero-shot scenarios.

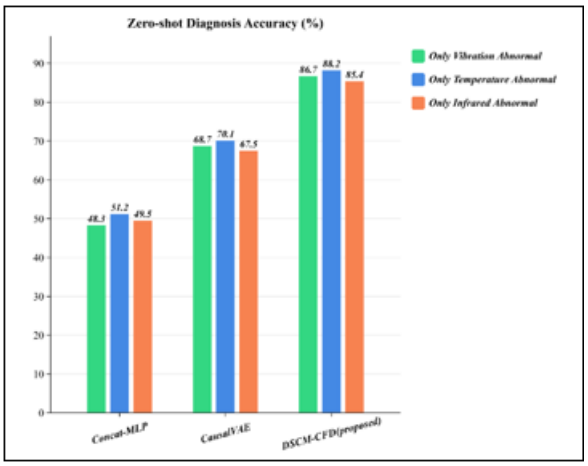

Figure 2: Zero-shot diagnosis accuracy comparison (bar chart)

Root Cause Localization. The Top-1 root cause localization accuracy of DSCM-CFD is 94.3%, significantly higher than CausalVAE (68.2%) and CF-CL (74.5%). In ablation experiments, removing the path consistency loss reduces the accuracy to 82.1%, and removing the counterfactual consistency loss reduces it to 80.5%, demonstrating the critical role of both losses in learning causal direction.

Counterfactual Generation Quality. The full model achieves an FID (Frechet Inception Distance) of 27.8 and a CCE (counterfactual consistency error) of 0.19, both better than the ablated versions (w/o path: 32.4/0.28; w/o cf: 36.7/0.41). Compared to CausalVAE (45.2/0.52), the FID is reduced by about 38%, proving the effectiveness of the counterfactual consistency loss.

Ablation Study. Removing L_{path}, L_{cf}, or L_{int} one by one causes significant drops in all metrics. Removing all causal losses reduces the model to a plain VAE classifier with 90.45% accuracy, 54.2% root cause localization, and 60.3% zero-shot performance. This indicates that causal constraints are the core of the model's superior performance.

Computational Efficiency. The inference latency is 12.3 ms per sample (encoder 2.1 ms, generator 6.5 ms, classifier 0.8 ms, CE computation 2.9 ms), far below the typical monitoring period of 100 ms, satisfying real-time diagnosis requirements.

5. Conclusion

This paper proposes a counterfactual diagnosis method for power transformers based on a differentiable structural causal model (DSCM-CFD). By explicitly constructing a causal generative mechanism from vibration to temperature to audio, the method elevates fault diagnosis from traditional pattern classification to causal intervention and counterfactual reasoning. The path consistency loss enforces causal chain transitivity, while the counterfactual consistency loss ensures representation stability under single interventions; their synergy enhances the model's causal reasoning capability and generalization performance. Experiments on real-world transformer multimodal samples demonstrate that DSCM-CFD significantly outperforms existing baselines in classification accuracy, zero-shot diagnosis, and root cause localization. Ablation studies confirm the critical contributions of each causal loss term. This work provides a new framework for multimodal intelligent diagnosis of power equipment that combines high accuracy, strong generalization, and interpretability. Future work will explore automatic causal discovery and extension to more modalities.

References

[1] Tavner, P.J.; Penman, J. Condition Monitoring of Electrical Machines. 1987.
[2] Baltrusaitis, T., et al. Multimodal machine learning: A survey. IEEE TPAMI 2018, 41, 423–443.
[3] Jia F, Lei Y, Lin J, et al.Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data [J]. Mechanical Systems and Signal Processing, 2016, 72-73:303-315.
[4] Q. Huang, X. Wei, Z. Wang and H. Wang, "Transformer zero-sample fault diagnosis based on multimodal data fusion," 2022 China Automation Congress (CAC), Xiamen, China, 2022, pp. 763-768.
[5] Li, B., et al. Research, application, and challenges of causal inference in industrial fault diagnosis: A survey. Eng. Appl. Artif. Intell. 2025, 142, 111200.
[6] Simões, L.D., et al. A power transformer differential protection based on support vector machine and wavelet transform. Electr. Power Syst. Res. 2021, 197, 107297.
[7] D. Ramachandram and G. W. Taylor, "Deep Multimodal Learning: A Survey on Recent Advances and Trends," in IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 96-108, Nov. 2017.
[8] Tsai, Y.H., et al. Multimodal Transformer for Unaligned Multimodal Language Sequences. In ACL 2019. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6558–6569.
[9] Yang, M., et al. CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models. In CVPR 2021.
[10] Kingma D P, Welling M. Auto-Encoding Variational Bayes [J]. arXiv.org, 2014.
[11] Zhao, C., et al. Multimodal unified generalization and translation network for intelligent fault diagnosis under dynamic environments. Eng. Appl. Artif. Intell. 2025, 142, 112559.
[12] Runge, J., et al. Detecting and quantifying causal associations in large nonlinear time series datasets. Sci. Adv.5, eaau4996(2019).
[13] Hochreiter S, Schmidhuber J. Long Short-Term Memory [J]. Neural Computation, 1997, 9(8):1735-1780.