Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira, Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p28973-cd

Leveraging Causal Inference for System Fault Diagnosis in Incremental Operating Conditions

Shuwen Zheng

School of Reliability and Systems Engineering, Beihang University, China.
Energy Department, Politecnico di Milano, Milan, Italy. E-mail:zhengsw@buaa.edu.cn

Jie Liu*

School of Reliability and Systems Engineering, Beihang University, China. E-mail: liujie805@buaa.edu.cn

Yunxia Chen

School of Reliability and Systems Engineering, Beihang University, China. E-mail:chenyunxia@buaa.edu.cn

Enrico Zio

Energy Department, Politecnico di Milano, Milan, Italy.
Center for Research on Risk and Crises, Mines Paris-PSL University, Paris, France. E-mail: enrico.zio@polimi.it

Abstracts While data-driven fault diagnosis is essential for industrial operations, maintaining performance under non-stationary conditions remains a challenge due to catastrophic forgetting. This study introduces a replay-free, causality-based framework designed to mitigate forgetting while ensuring privacy. We theoretically characterize the forgetting phenomenon within a causal graph structure, demonstrating that feature representations from the previous task function as a collider. This collider blocks the information flow between historical and current data distributions. To rectify this, we employ a de-confounding strategy by conditioning on the colliding factor. The framework further incorporates enhanced contrastive learning to enforce local topological consistency and a centroid alignment mechanism to ensure global class separability. Empirical results demonstrate this approach effectively balances the stability-plasticity dilemma, delivering superior diagnostic performance in incrementally evolving industrial environments.

Keywords: Fault diagnosis, Causal inference, Structural causal model, Incremental learning, Data-driven, Bearing.

1. Introduction

As machinery becomes increasingly sophisticated, the requirement for absolute reliability, safety, and operational efficiency has escalated from a competitive advantage to a mandatory baseline (Zio 2024). Distinct faults or undetected anomalies can precipitate catastrophic operational disruptions, leading to severe safety hazards for personnel and substantial economic losses. Consequently, the development of robust fault diagnosis methodologies has become a cornerstone of sustainable industrial safety.

While classical rule-based diagnosis systems have historically served the industry, they increasingly struggle to cope with the complex, non-linear dynamics of modern equipment (Qiao et al. 2024). The exponential growth in sensor deployment and computing power has thus propelled data-driven approaches to the forefront. Deep learning paradigms, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Graph Neural Networks (GNNs), have demonstrated exceptional proficiency in extracting intricate fault patterns from high-dimensional sensor streams (Zheng et al. 2024). Studies such as those by Shao et al. (2020) have showcased the efficacy of these models in identifying faults in induction motors, while Liu (2021) addressed data imbalances in fault classification. These static data-driven models excel in closed-set environments where training and testing data share identical distributions. However, critical vulnerabilities of these celebrated techniques emerge when they are deployed in real-world environments. Industrial operating conditions, characterized by fluctuating

loads, variable speeds, and shifting temperatures, are rarely stationary. When a model trained on historical data is exposed to a new operating regime, the distribution shift often leads to diagnostic accuracy declines (Wang et al. 2024). This sensitivity poses a severe threat to system reliability.

To maintain safety standards, diagnostic models must evolve continuously. Incremental Learning (IL) has emerged as a promising paradigm to address this need, aiming to update models with new data streams without the computational prohibitiveness of retraining from scratch. The central challenge in IL is the "stability-plasticity dilemma": the system must be plastic enough to learn new fault patterns under new conditions but stable enough to retain knowledge of previous states. Failure to maintain this balance results in "catastrophic forgetting", where the refinement of the model on new data obliterates its ability to recognize previously learned faults (Cao et al. 2024).

The research community has proposed various strategies, generally categorized into architecture-based, regularization-based, and replay-based approaches (Oh et al. 2022). Architecture-based methods (Shi et al. 2024) dynamically expand network sub-structures for new tasks but suffer from escalating parameter complexity. Regularization-based methods (Sun et al. 2024) attempt to constrain weight updates to protect important parameters but often struggle when tasks are highly dissimilar. Currently, replay-based approaches are among the most effective, operating on the principle of storing and re-training representative samples from previous tasks (Hu et al. 2024). While effective, their reliance on memory buffers presents significant drawbacks for modern industrial applications. First, the assumption of sufficient storage can be often invalid for resource-constrained devices. Second, data privacy and security regulations often preclude the long-term storage and retention of raw operational data (Roy et al. 2024). In scenarios where data cannot be replayed due to privacy constraints, traditional replay-based incremental learning becomes infeasible.

To overcome these limitations, this study diverges from correlation-based learning and turns to causal inference. While standard deep learning captures statistical associations, causal inference seeks to understand the underlying mechanisms generating the data, offering superior robustness against spurious correlations (Pearl 2009). Recent advancements in Prognostics and Health Management (PHM) have successfully applied causal theory to feature selection and remaining useful life prediction (Zheng et al. 2024), yet its potential in incremental fault diagnosis remains underexplored. In this paper, we propose a novel, replay-free incremental fault diagnosis framework inspired by causal inference. We theoretically characterize the phenomenon of catastrophic forgetting not merely as a overwrite of weights, but as a structural problem in the causal graph: the feature representation learned from previous tasks acts as a "collider" that blocks the information flow from historical data distributions to current predictions. By formally identifying this blockage, we introduce a de-confounding strategy involving a $do(\cdot)$ operator intervention to restore the causal link. Our method employs an Enhanced Contrastive (EC) learning strategy coupled with a Centroid Positioning (CP) constraint. The EC mechanism enforces local topological consistency by aligning new feature representations with their historical counterparts without requiring raw data replay, while the CP constraint ensures global class separability. To validate the effectiveness of the method, a case study concerning bearing fault diagnosis under incremental working conditions is carried out. The results show promising performance of the model in accurate diagnosis and previous knowledge retention.

2. Method

2.1. *Problem formulation*

The primary objective of this study is to facilitate incremental fault diagnosis in non-stationary environments, where data arrives sequentially, and privacy constraints preclude the retention of historical datasets.

Let $\mathfrak{D}_0 = \{(w_0^i, y_0^i)\}_{i=0}^{N_0}$ denote the initial dataset collected under the base operating condition, where w_0^i represents the input sensor signals and y_0^i the corresponding health state labels. A diagnostic model is initially trained on $\mathfrak{D}_0$ to minimize the classification error. As the process evolves, the system encounters a new operating condition, yielding a new dataset $\mathfrak{D}' = \{(w'^j, y'^j)\}_{j=0}^{N'}$. The goal of incremental learning is to update the model parameters such that the refined model achieves

high diagnostic accuracy on the current domain while retaining its discriminative capability on the historical domain. Critically, this update is performed under a privacy-preserving constraint: the historical data is not accessible during the new training phase.

2.2. *Causal interpretation of catastrophic forgetting*

Causal inference enables a rigorous analysis of cause-and-effect relationships through structural interventions (Cui et al. 2020). Central to this methodology is the Directed Acyclic Graph (DAG), where variables are visualized as nodes and their causal dependencies are characterized by directed edges (Dhar et al. 2019). To analyze the mechanism of forgetting in incremental fault diagnosis, we construct a general causal graph incorporating the key variables, as illustrated in Fig. 1.

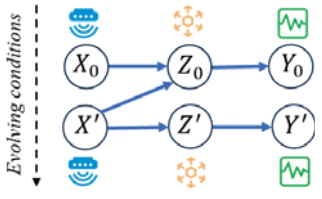

Fig. 1. A general causal graph for incremental fault diagnosis scenario.

In this graphical model, W_0 and W' denote the input data for the historical and current tasks, respectively. The diagnostic process involves extracting feature representations and predicting system states. Specifically, X' represents the features extracted by the updated model from the new input W', which directly influences the current system state prediction Y', denoted by the path $W' \rightarrow X' \rightarrow Y'$. Crucially, we also consider X_0, which represents the features of the *current* input samples extracted within the *old* feature space (i.e., by the frozen old model). The generation of X_0 is causally determined by the current input W' and the historical data W_0, as the latter defined the parameters of the old feature extractor during the initial training phase. This structure implies that X_0 acts as a bridge between historical knowledge and current data.

To quantify the causal influence of these variables, we employ the $do(\cdot)$ operator, a mathematical tool that simulates physical interventions (Pearl 2009). The operation $do(X = x)$ represents an external intervention that fixes variable X to value x, effectively severing all incoming causal edges to X while preserving its outgoing effects. The average causal effect of a manipulated variable X on a target response Y is formulated as the difference between the interventional distribution and a baseline null intervention:

$$Effect_{X\rightarrow Y} = P(Y|do(X = x)) - P(Y|do(X = 0)) \quad (1)$$

in which $do(X = 0)$ denotes a null intervention to X, serving as the un-intervened baseline. The interventional effect is thus the difference between the post-intervention and baseline outcomes.

We thus employ Pearl's do-calculus to quantify the influence of historical data on the current model's prediction. Ideally, to preserve knowledge, the historical data W_0 should exert a non-zero causal effect on the current prediction Y'. This effect is formally defined as the interventional distribution:

$$Effect_{W_0\rightarrow Y'} = P(Y'|do(W_0 = w_0)) - P(Y'|do(W_0 = 0)) \quad (2)$$

The above equation can be further derived as $P(Y'|W_0 = w_0) - P(Y'|W_0 = 0)$ since the manipulated variable W_0 has no parent variables in the graph, and, thus, $P(Y'|do(W_0)) = P(Y'|W_0)$. However, an inspection of the causal graph reveals a fundamental obstruction. The path from W_0 to Y' is intercepted by the collider X_0. According to the rules of *d-separation*, a collider blocks the flow of information between its parents unless it (or one of its descendants) is conditioned upon (Hu et al. 2021). Without intervention, the historical data W_0 and the current prediction Y' are conditionally independent. Consequently, the causal effect vanishes:

$$\begin{aligned} Effect_{W_0\rightarrow Y'} &= P(Y'|W_0 = w_0) - P(Y'|W_0 = 0) \\ &= P(Y') - P(Y') \\ &\rightarrow 0 \end{aligned} \quad (3)$$

The above calculation demonstrates a negligible causal effect of the previous data W_0 on the current prediction Y'. This theoretical result explains catastrophic forgetting: the natural causal structure of incremental learning inherently prohibits the influence of past data, causing the model to optimize solely for W' and disregard W_0.

2.3. *Collider controlling to mitigate forgetting*

Fortunately, causal inference offers a solution to restore the disrupted interactions through controlling the collider. That is, by conditioning on the joint outcome, namely the collider, we induce a dependency between the two otherwise independent variables (Cheng et al. 2024). In this

case, the causal path $W_0 \rightarrow X_0 \leftarrow W'$ becomes active, enabling the interactions between W_0 and Y'. Mathematically, we reformulate the causal effect by conditioning on X_0. Utilizing the law of total probability and the Markov property of the graph, the restored effect can be derived as:

$$\begin{aligned} Effect_{W_0 \rightarrow Y'} &= P(Y'|X_0, do(W_0 = w_0)) \\ &\quad - P(Y'|X_0, do(W_0 = 0)) \\ &= P(Y'|X_0, W_0 = w_0) - P(Y'|X_0, W_0 = 0) \\ &= \sum_{X'} P(Y'|X', X_0, W_0 = w_0)[P(X'|X_0, W_0 = w_0) \\ &\quad - P(X'|X_0, W_0 = 0)] \\ &= \sum_{X'} P(Y'|X')[P(X'|X_0, W_0 = w_0) \\ &\quad - P(X'|X_0, W_0 = 0)] \end{aligned} \quad (4)$$

where the third equality is the result of law of total probability; and due to the Markov property of causal graph, variable Y' is conditionally independent of X_0 and W_0 given X'.

The above equation illuminates the sufficient condition for preventing forgetting. The term $P(Y'|X')$ represents the classification accuracy on the new task, which is standard. The crucial component is the term in the brackets, which implies that the probability of the new representation X', given the collider X_0, must be causally linked to the old data W_0. Since we cannot access W_0 directly due to no replay, we use the old model's representation of the current data as a proxy. The derivation implies that maximizing the alignment between the new representation X' and the old representation X_0 effectively preserves the causal pathway. This theoretical insight justifies the alignment of feature spaces as a necessary condition for replay-free incremental learning. In practice, this motivates a strategy to enforce the new model's representations remain similar to those of the old model, forming the basis for our approach.

2.4. *Replay-free implementation strategy*

2.4.1. *Enhanced contrastive learning*

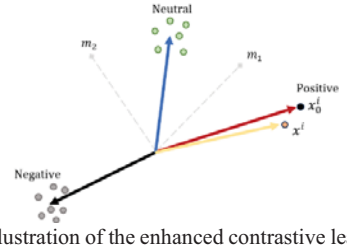

Fig. 2. Illustration of the enhanced contrastive learning.

To align the conditional distributions $P(X'|X_0)$, we introduce an Enhanced Contrastive (EC) learning objective. Unlike standard contrastive methods that rely on data augmentations, our approach utilizes the parallel augmentation provided by the two model versions. For every input sample w'^i, we generate an anchor representation x_0^i from the frozen old model, representing the ground truth location in the historical feature space; and a query representation x^i from the current model, representing the evolving feature space. The EC functions as a geometric regularizer that strictly enforces topological consistency. It categorizes relationships into three distinct tiers:

- **Positive Pair:** The representation of the same instance across the two models (x^i and x_0^i).
- **Neutral Samples:** Other instances belonging to the same class as x^i. Unlike traditional contrastive losses that treat all other samples as negatives, we apply a "soft boundary" to these samples to preserve intra-class compactness without over-penalization.
- **Negative Samples:** Instances belonging to different classes.

The optimization objective is formulated to pull the query x^i towards its historical anchor x_0^i while repelling negative samples and maintaining elastic boundaries for neutral ones:

$$\mathcal{L}_{EC} = -log \frac{\exp(sim(x^i, x_0^i)/\tau)}{\exp(sim(x^i, x_0^i)/\tau) + \sum_{j \in \mathbb{C}_i} \exp(sim(x^i, x_0^j)/\tau) + d_{neu}} \quad (5)$$

$$d_{neu} = \sum_{k \neq i}^{\mathbb{C}_i} \exp((d_{k,i}^{pos} + d_k^{neg})/\tau) \quad (6)$$

$$d_{k,i}^{pos} = \max(0, sim(x^i, x_0^k) - s^{pos} + m_1) \quad (7)$$

$$d_{k,i}^{neg} = \max(0, s^{neg} - sim(x^i, x_0^k) + m_2) \quad (8)$$

where $sim(\cdot,\cdot)$ denotes the cosine similarity; τ is the temperature coefficient; $\mathbb{C}_i$ represents the set of samples belonging to the same class as x^i; s^{pos} represents the similarity between the positive pair; s^{neg} is the average similarity of the negative samples; the term d_{neu} handles the neutral samples with soft thresholds m_1 and m_2 to govern the intra-class variance.

EC coerces the updated representation x^i to converge towards its historical counterpart x_0^i. Concurrently, representations belonging to heterogeneous classes are explicitly pushed to ensure decision boundary separability. Distinct from conventional contrastive learning paradigms (He et al. 2020), our approach introduces a nuanced category of "neutral samples". Instead of enforcing rigid attraction, we apply soft boundary constraints to these samples. This strategy is critical for preserving the granular intra-class structure, as it prevents the collapse of all same-class features into a single point while avoiding

excessive penalization. Consequently, this imposes a hierarchical alignment: the model anchors the specific instance to its past state while maintaining necessary manifold spacing from other intra-class and inter-class embeddings. The schematic of the EC mechanism is depicted in Fig. 2.

2.4.2. *Centroid positioning constraint*

While the EC mechanism effectively preserves the local instance-wise topology, it operates individually on samples and may not fully constrain the global drift of feature distributions. In high-dimensional spaces, it is possible for an entire class cluster to shift significantly even if the relative distances between instances are preserved (Hou et al. 2019). To mitigate this risk, we introduce the Centroid Positioning (CP) constraint, which operates on the macroscopic level of the feature space.

We define the prototype of each fault category as the centroid of its feature distribution. The historical class centroids for class $\mathbb{C}_i$ are estimated in the historical model. Similarly, the centroids in the current evolving feature space $\mu'^{\mathbb{C}_i}$ are calculated using the updated model as:

$$\mu'^{\mathbb{C}_i} = \frac{1}{N_{\mathbb{C}_i}} \sum_{i \in \mathbb{C}_i} x^i \tag{9}$$

where $N_{\mathbb{C}_i}$ is the number of subset samples for class $\mathbb{C}_i$.

The CP loss explicitly enforces the alignment of these global anchors:

$$\mathcal{L}_{CP} = \sum_{\mathbb{C}_i} (1 - sim(\mu'^{\mathbb{C}_i}, \mu_0^{\mathbb{C}_i})) \tag{10}$$

Minimizing $\mathcal{L}_{CP}$ serves two critical functions. First, it anchors the semantic center of each fault type to its historical position, preventing the new model from redefining the meaning of a fault class. Second, it acts as a stabilizing force that complements the EC loss. While EC manages the micro-structure at instance-level, CP locks the macro-structure via cluster location. This hierarchical constraint ensures that the decision boundaries remain globally valid across incremental phases, effectively transferring class-level knowledge without the need for raw data replay.

2.4.3. *Composite optimization objective*

With the above constraints defined to enhance alignment between the new representations X' and the old representations X_0, the incremental update of model is processed with a composite objective that balances both classification accuracy and feature space alignment. Specifically, the final training objective for the incremental update is a composite loss function that balances plasticity and stability as:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \cdot \mathcal{L}_{EC} + \lambda_2 \cdot \mathcal{L}_{CP} \tag{11}$$

$$\mathcal{L}_{CE} = -\frac{1}{N'} \sum_{i=1}^{N'} y_i' \log p_i' \tag{12}$$

where $\mathcal{L}_{CE}$ is the standard cross-entropy loss to ensure diagnostic accuracy on the current operating condition. The hyperparameters λ_1 and λ_2 are two coefficients that balance the strengths of EC and CP, which can be determined by hyperparameter fine-tuning; N' is the number of samples in the new dataset, y_i' is the ground-truth label for the i -th sample, and p_i' is the corresponding predicted probability.

Eventually, such joint optimization guarantees not only classification accuracy in the new diagnosis phase, but similarity of the feature space distributions, thereby maintaining the causal effect of W_0 on Y' to preserve prior knowledge and mitigate catastrophic forgetting. Crucially, this process requires only the current data stream $\mathfrak{D}'$ and the frozen parameters of old model with historical centroids, strictly adhering to the replay-free and privacy-preserving requirements of the industrial deployment.

3. Case study

3.1. *Experimental data and configuration*

To empirically validate the efficacy of the proposed causal incremental framework, we conducted a comprehensive evaluation using the Paderborn University (PU) bearing dataset (Lessmeier et al. 2016), which is widely recognized benchmark for condition monitoring. The experimental rig consists of a modular system featuring a torque-measurement shaft, a rolling bearing test module, a flywheel, and a load motor, designed to simulate the complex dynamics of electromechanical drive systems. High-frequency data acquisition was performed using piezoelectric accelerometers mounted on the bearing housing, capturing vibration signals at a sampling rate of 64 kHz, alongside synchronous motor current measurements.

For this study, we specifically focused on a subset of bearings that underwent accelerated life testing. This selection is critical because artificially induced defects often fail to capture the stochastic progression of material fatigue found in real-world industrial scenarios. Table 1 summarizes the characteristics of the involved bearings in this study.

To simulate the incrementally evolving industrial environment, we designed a three-phase experimental protocol (Phase 0 to Phase 2) characterized by distinct operating conditions, as shown in Table 2. This setup challenges the diagnostic model to adapt to distribution shifts caused by varying rotational speeds and radial forces without accessing raw historical data from previous phases. The data preprocessing phase involved min-max normalization to standardize the signal amplitude, followed by segmentation into non-overlapping windows of 1,024 data points to form the input vectors.

Table 1. Description of the involved bearings.

Label	Bearing element	Combination	Code
0	\	\	K001
1	OR	S	KA04
2	OR	R	KA16
3	IR	M	KI14
4	IR	S	KI18

OR: outer - ring; IR: inner - ring; S: single damage; R: repetitive damage; M: multiple damage.

Table 2. Involved working conditions.

Phase	Rotational speed	Load Torque	Radial force
0	900 rpm	0.7 Nm	1000 N
1	1500 rpm	0.1 Nm	1000 N
2	1500 rpm	0.7 Nm	400 N

3.2. *Implementation details and benchmarking*

The proposed framework was benchmarked against a spectrum of methodologies ranging from naive baselines to advanced continuous learning strategies. The Fine-Tuning (FT) serves as a naive baseline where the model is sequentially updated on new data without any constraints, serving as a lower bound to quantify catastrophic forgetting. Moreover, several popular incremental learning methods, including Elastic Weight Consolidation (EWC), Learning without Forgetting (LwF), and Learning without Memorizing (LwM), were also implemented to facilitate further performance comparisons with the proposed method (Kirkpatrick et al. 2017, Li et al. 2017 and Dhar et al. 2019).

The feature extraction backbone employed is a ResNet-18 architecture (He et al. 2016). Notably, the leaky Rectified Linear Unit (leaky ReLU) was utilized as the activation function in the residue block outputs, which aimed to enhance the output range and alleviate saturation of the representations 0. Furthermore, to stabilize training dynamics under shifting distributions, a cosine normalization layer was integrated into the classifier (Hou et al. 2019). During experiment, the datasets were partitioned into 80% training and validation subsets for hyperparameter tuning, and the remaining 20% was used as the test dataset. During each incremental phase, the training data from the new working condition were incrementally incorporated for model adaptive update. Model validation was conducted using test sets from all previously encountered conditions, to assess both current diagnostic accuracy and model's resilience to catastrophic forgetting. The optimization process utilized the Adam algorithm with a batch size of 256. We implemented a step-wise learning rate decay schedule, initializing the learning rate at 0.002 and reducing it by a factor of 0.2 at 50, 70, and 90 epochs to facilitate convergence.

In order to quantitively assess model in the incremental diagnosis settings, we adopted the average performance A_c to measure the model's diagnosis ability across all learned tasks at the current c-th phase. The performance of the previous conditions is further measured using A_c^{old}, which focuses on the average diagnosis ability in the previous $c-1$ conditions, reflecting its ability to retain prior knowledge, where a_i^c represents the diagnosis accuracy of condition i after training on the c-th condition.

$$\mathcal{A}_c = \frac{1}{c}\sum_{i=0}^{c} a_i^c \tag{13}$$

$$\mathcal{A}_c^{old} = \frac{1}{c-1}\sum_{i=0}^{c-1} a_i^c \tag{14}$$

3.3. *Results and discussion*

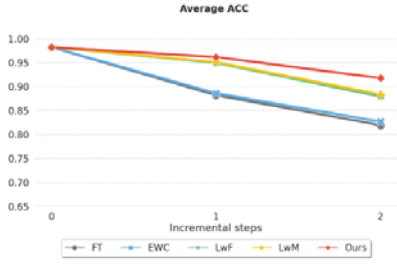

Fig. 3. The average accuracy of different methods in the incremental process.

Table 3. Diagnosis accuracy after learning all conditions.

Methods	$\mathcal{A}_c$	$\mathcal{A}_c^{old}$
FT	0.8195	0.7334
EWC	0.8269	0.7451
LwF	0.8792	0.8269

LwM	0.8837	0.8363
Ours	**0.9178**	**0.8907**

The quantitative trajectory of diagnostic accuracy across the incremental phases is presented in Fig. 3. A comparative analysis reveals distinct patterns among the tested methods. The FT method exhibits a precipitous decline in performance immediately after the first domain shift, corroborating the severity of catastrophic forgetting when no preventive measures are taken. Similarly, while EWC attempts to protect important weights, its reliance on a diagonal approximation of the Fisher information matrix appears insufficient for the complex, high-dimensional shifts inherent in vibration data, resulting in suboptimal retention rates ($\mathcal{A}_c \approx$ 82.69%). Distillation-based methods (LwF and LwM) demonstrate moderate stability during the initial transition (Phase 0 to 1) but struggle to maintain robustness as the task complexity accumulates in Phase 2. This deterioration suggests that merely aligning output probabilities or attention maps is inadequate when the underlying feature distributions diverge significantly due to changing working conditions. In stark contrast, our proposed causality-grounded framework delivers superior performance, achieving a final average accuracy ($\mathcal{A}_c$) of 91.78% and a historical retention rate ($\mathcal{A}_c^{old}$) of 89.07%. As summarized in Table 4, our method outperforms the nearest competitor (LwM) by a significant margin. This performance advantage is further visualized in Fig. 4, which tracks the accuracy on Phase 0 data throughout the learning process. While other methods show a monotonic decay in identifying the initial faults, our approach maintains a high fidelity to the original distribution.

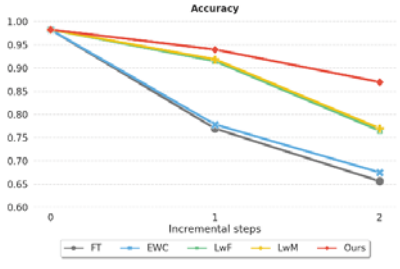

Fig. 4. Performance evolution of phase 0 data.

The robustness of our method can be attributed to the collider adjustment strategy. By theoretically identifying the historical feature representation as a collider node in the causal graph, we effectively unblock the flow of historical information. The enhanced contrastive learning ensures that the topological structure of the feature space remains consistent, locking the relative positions of fault clusters, while the centroid positioning constraint prevents global semantic drift. This dual-regularization mechanism allows the model to plastically adapt to new speed and load conditions without overwriting the decision boundaries established in Phase 0. Consequently, the proposed framework successfully reconciles the conflict between learning new patterns and remembering old ones, adhering to the privacy-preserving, replay-free constraints of industrial deployment.

4. Conclusion

This study addresses the imperative challenge of catastrophic forgetting in the fault diagnosis of safety-critical industrial systems. While data-driven models have revolutionized condition monitoring, their deployment in non-stationary environments is frequently hampered by the "stability-plasticity" dilemma, particularly when privacy regulations preclude the storage and replay of historical data. To resolve this conflict, we proposed a novel causal-inspired framework that shifts the learning paradigm from correlation-based association to causality-based structural preservation. The core contribution of this work lies in identifying the root cause of forgetting not merely as parameter overwriting, but as a structural blockage in the causal graph. Specifically, the historical representation acting as a "collider". By applying the *do*-calculus, we formulated a de-confounding strategy to unblock the information flow between historical data and current predictions without accessing raw past data. This theoretical insight was operationalized through a dual-regularization mechanism. The EC learning utilizes the frozen old model as a surrogate to enforce local topological fidelity, fixing the relative geometric positions of feature clusters. Furthermore, CP constraint anchors the global semantic distribution to prevent class-level drifts. Experimental validation on the PU bearing dataset demonstrated that our framework achieves superior diagnostic accuracy and historical retention ability.

References

Zio, E. (2024). Data-driven prognostics and health management (PHM) for predictive maintenance of industrial components and systems. In Risk-Informed Methods and Applications in Nuclear

and Energy Engineering, 113–137. Academic Press.

Qiao, D., X. Wei, B. Jiang, et al. (2024). Data-driven fault diagnosis of internal short circuit for series-connected battery packs using partial voltage curves. IEEE Transactions on Industrial Informatics 20, no. 4: 6751–6761.

Zheng, S., K. Pan, J. Liu, et al. (2024). Empirical study on fine-tuning pre-trained large language models for fault diagnosis of complex systems. Reliability Engineering & System Safety 252: 110382.

Shao, S., R. Yan, Y. Lu, et al. (2020). DCNN-Based Multi-Signal Induction Motor Fault Diagnosis. IEEE Transactions on Instrumentation and Measurement 69, no. 6: 2658–2669.

Liu, J. (2021). Fuzzy support vector machine for imbalanced data with borderline noise. Fuzzy Sets and Systems 413: 64–73.

Wang, C., and J. Liu. (2024). Anomaly Detection Method for Complex Systems Based on a Domain-Adaptive Causal Decoupling Model. IEEE Transactions on Industrial Informatics.

Cao, X., X. Zheng, G. Wang, et al. (2024). Solving the catastrophic forgetting problem in generalized category discovery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16880–16889.

Oh, Y., D. Baek, and B. Ham. (2022). Alife: Adaptive logit regularizer and feature replay for incremental semantic segmentation. Advances in Neural Information Processing Systems 35: 14516–14528.

Sun, M., X. Xiao, T. Chen, et al. (2024). A Novel Domain Incremental Learning Method for Bearing Fault Diagnosis Based on F&K. IEEE Transactions on Industrial Informatics.

Shi, M., C. Ding, S. Chang, et al. (2024). Cross-Domain Class Incremental Broad Network for Continuous Diagnosis of Rotating Machinery Faults Under Variable Operating Conditions. IEEE Transactions on Industrial Informatics 20, no. 4: 6356–6368.

Hu, K., Q. He, C. Cheng, et al. (2024). Adaptive incremental diagnosis model for intelligent fault diagnosis with dynamic weight correction. Reliability Engineering & System Safety 241: 109705.

Roy, S., V. Verma, and D. Gupta. (2024). Efficient Expansion and Gradient Based Task Inference for Replay Free Incremental Learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 1165–1175.

Pearl, J. (2009). Causality. Cambridge University Press.

Zheng, S., C. Wang, E. Zio, et al. (2024). Fault detection in complex mechatronic systems by a hierarchical graph convolution attention network based on causal paths. Reliability Engineering & System Safety 243: 109872.

Hu, X., K. Tang, C. Miao, et al. (2021). Distilling causal effect of data in class-incremental learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3957–3966.

Cui, P., Z. Shen, S. Li, et al. (2020). Causal inference meets machine learning. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 3527–3528.

Cheng, D., J. Li, L. Liu, et al. (2024). Data-driven causal effect estimation based on graphical causal modelling: A survey. ACM Computing Surveys 56, no. 5: 1–37.

Hou, S., X. Pan, C. C. Loy, et al. (2019). Learning a unified classifier incrementally via rebalancing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 831–839.

Lessmeier, C., J. K. Kimotho, D. Zimmer, et al. (2016). Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. In PHM Society European Conference.

Kirkpatrick, J., R. Pascanu, N. Rabinowitz, et al. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences 114, no. 13: 3521–3526.

Li, Z., and D. Hoiem. (2017). Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, no. 12: 2935–2947.

Dhar, P., R. V. Singh, K. C. Peng, et al. (2019). Learning without memorizing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5138–5146.

He, K., X. Zhang, S. Ren, et al. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778.

Xu, B., N. Wang, T. Chen, et al. (2015). Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853.

He, K., et al. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.