10.37965/jdmd.2026.1597

A Generalizable and Interpretable Deep Transfer Learning Framework for Industrial Condition Monitoring under Noise-Induced Domain Shift

Adrián Rodríguez Ramos[1,*], José Manuel Bernal de Lázaro[1,2], Antônio J. Silva Neto[1] and Orestes Llanes-Santiago[1,3]

[1]Laboratório LEMA-LEMEC & Programa de Pós-Graduação em Modelagem Computacional, Instituto Politécnico-Universidade do Estado do Rio de Janeiro (IPRJ/UERJ), Rua Bonfim, 25 - Campus UERJ, Vila Amélia, Nova Friburgo, 28625-570, RJ, Brazil

[2]Escuela Superior de Ingeniería, Universidad Internacional de La Rioja (UNIR), Avenida de la Paz, 137, Logroño, 26006, La Rioja, España

[3] Departamento de Automática y Computación, Universidad Tecnológica de La Habana José Antonio Echeverría, CUJAE, Calle 114 No. 11901, entre Ciclovía y Rotonda, Marianao, 19390, La Habana, Cuba

*Corresponding author: adrian.ramos@iprj.uerj.br

Received 30 June 2026; Revised 31 July 2026; Accepted 04 August 2026; Published online 11 August 2026

Abstract: Industrial condition monitoring systems increasingly operate under heterogeneous deployment conditions where the statistical distribution of monitoring data differs from that observed during model development. Such heterogeneous domain shifts, arising from both cross-process variability and realistic industrial noise, significantly degrade the performance and reliability of conventional deep learning models. Addressing this challenge constitutes the primary objective of this work.

This paper proposes a generalizable and interpretable monitoring framework based on an adaptive Transfer Learning (TL) strategy combined with a hybrid CNN-Kolmogorov-Arnold Network (KAN) architecture. First, the impact of realistic industrial noise (Gaussian, Colored, Impulsive, and Mixed) is statistically characterized using both univariate and multivariate tests, demonstrating significant domain shifts between noise-free and noisy conditions. Subsequently, a pre-trained model is adapted using original (noise-free) data from the target domain, while validation and online evaluation are performed under noisy conditions to emulate realistic deployment scenarios.

The proposed CNN-KAN model combines convolutional layers for robust temporal feature extraction with KAN-based functional representations that enhance interpretability and structured learning. Additionally, an adaptive transfer mechanism dynamically selects the optimal fine-tuning configuration, effectively balancing stability and plasticity during domain

10.37965/jdmd.2026.1597

adaptation. Experimental evaluations were conducted on two heterogeneous industrial processes (DAMADICS and GPN) involving both fault diagnosis and cyberattack detection scenarios. Results demonstrate that the proposed framework achieves high classification performance under nominal conditions while maintaining stable degradation behavior under severe noisy environments. Furthermore, quantitative interpretability analyses based on latent feature importance, class-wise contribution maps, and gradient-based feature attribution confirm that the CNN-KAN architecture provides consistent and structured explanations aligned with the monitoring process. Comparative analyses against existing transfer learning and domain adaptation approaches further demonstrate the effectiveness of the proposed framework for industrial condition monitoring under complex domain-shifting conditions.

Overall, the obtained results demonstrate that the proposed adaptive transfer learning framework provides an effective and interpretable solution for industrial condition monitoring under heterogeneous domain-shifting conditions.

Keywords: Industrial condition monitoring, Transfer learning, Domain shift, Generalizability, Interpretability

I. Introduction

Industrial condition monitoring has become an essential component of modern industrial systems, enabling the early detection of abnormal operating conditions, improving operational reliability, and supporting predictive maintenance and industrial cybersecurity strategies [9, 2, 23]. Recent advances in deep learning have significantly improved the capability of data-driven monitoring systems to diagnose physical faults and detect cyberattacks from multivariate process data [55, 56, 8, 14]. However, the practical deployment of these models in real industrial environments remains challenging because the statistical characteristics of operational data frequently differ from those observed during model development. Such heterogeneous domain shifts, arising from both cross-process variability and realistic industrial noise, substantially degrade the generalization capability of conventional deep learning models. Consequently, developing adaptive transfer learning strategies capable of maintaining reliable monitoring performance under these conditions has become a major research challenge. In this work, the term adaptive refers to the automatic optimization of the transfer learning configuration during the model adaptation stage prior to deployment.

Recently, Digital Twin (DT) technologies have emerged as a complementary paradigm for industrial condition monitoring by enabling continuous interaction between physical assets and their virtual counterparts. DT-driven approaches have demonstrated significant potential for real-time health assessment, predictive maintenance, and degradation monitoring in cyber-physical industrial systems. Nevertheless, these methodologies primarily focus on virtual asset synchronization and system-level monitoring, whereas the proposed framework focuses on improving the robustness and generalization capability of deep transfer learning models operating under heterogeneous noise-induced domain shifts [69, 70].

10.37965/jdmd.2026.1597

Within industrial condition monitoring, two of the most critical applications are physical fault diagnosis and cyberattack detection. These phenomena may coexist and manifest themselves through complex temporal patterns in multivariate process data, making unified monitoring a demanding task [4, 12, 7]. In the technical literature, approaches for fault diagnosis and cyberattack detection are commonly categorized into model-based and data-driven methods [53, 58, 54]. Model-based techniques rely on accurate mathematical representations of the underlying processes, which require in-depth domain knowledge and are difficult to obtain in modern large-scale and highly complex systems [3, 32]. In contrast, data-driven approaches leverage historical measurements to learn operational patterns without requiring explicit physical models [11, 24, 19].

In reality, industrial measurements are affected by various types of noise, such as sensor noise, communication disturbances, colored noise from dynamic processes, and impulsive disturbances caused by faults or malicious interventions. These disturbances induce domain shifts between noise-free training data and noisy operational data [43, 17]. When the implementation domain differs from the training domain, the generalizability of deep models cannot be adequately assessed under classical assumptions. Therefore, before proposing a generalizable monitoring system, it is necessary to formally demonstrate that industrial noise can create statistically distinct domains. In this work, we validate this hypothesis using comprehensive univariate and multivariate statistical analyses.

To address domain mismatch, Transfer Learning (TL) has become a promising strategy [15, 10]. However, conventional TL approaches often rely on fixed adaptation schemes, which may not be optimal at different levels and types of distribution change. Motivated by this limitation, we propose an adaptive transfer learning scheme capable of determining the appropriate adaptation configuration based on domain characteristics. Therefore, adaptive transfer learning strategies capable of dynamically selecting the most appropriate adaptation configuration are required to ensure reliable monitoring under heterogeneous deployment conditions.

Furthermore, interpretability is becoming a fundamental requirement for industrial AI systems [1]. Deep black-box networks limit confidence, validation, and regulatory acceptance in safety-critical environments. Recently, Kolmogorov-Arnold Networks (KANs) have been introduced as a functional representation paradigm that replaces fixed activation functions with learnable univariate functions, enabling improved interpretability and robustness [36].

In this paper, we integrate convolutional neural networks with KAN layers to form a hybrid CNN-KAN architecture. The convolutional component extracts hierarchical time features, while the KAN module provides structured functional representations that enhance interpretability and potentially robustness under noisy conditions.

Based on the above discussion, this work addresses the problem of adaptive transfer learning for industrial condition monitoring under heterogeneous domain-shifting conditions. Specifically, the proposed framework simultaneously considers two complementary sources of distribution mismatch: (i) cross-process discrepancies between source and target industrial systems and (ii) noise-induced distribution

shifts caused by realistic sensing and communication disturbances. Rather than addressing these challenges independently, the proposed framework formulates them within a unified adaptive transfer learning strategy. To this end, the hybrid CNN-KAN architecture provides robust and interpretable feature learning, while the proposed Hybrid Adaptive Fine-Tuning and Structural Expansion (HAFSE) strategy automatically determines the optimal transfer configuration according to the characteristics of the target domain. The proposed framework aims to improve the generalization capability of transfer learning across heterogeneous industrial processes when limited labeled target data is available.

The main contributions of this work is the adaptive transfer learning framework described in (1), which constitutes the central objective of this study. Contributions (2)-(5) correspond to complementary methodological components that enable and support this central objective, as detailed below:

1) An adaptive transfer learning framework specifically developed for industrial condition monitoring under heterogeneous domain-shifting conditions, simultaneously addressing cross-process discrepancies and realistic noise-induced distribution changes.
2) A Hybrid Adaptive Fine-Tuning and Structural Expansion (HAFSE) strategy that automatically determines the optimal transfer configuration by jointly optimizing the number of trainable layers and structural expansion according to the characteristics of the target domain.
3) A hybrid CNN-KAN architecture that combines deep temporal feature extraction with intrinsically interpretable functional decision modeling, providing transparent and robust industrial monitoring.
4) A unified transfer learning methodology applicable to both fault diagnosis and cyberattack detection, allowing different industrial monitoring tasks to be addressed within the same adaptation framework.
5) Extensive experimental validation on heterogeneous industrial benchmarks under multiple realistic noise conditions, demonstrating its effectiveness, robustness, and generalization capability.

The remainder of this article is organized as follows. Section 2 presents a review of works related to existing approaches to transfer learning and domain adaptation. Section 3 introduces the CNN, KAN, and hybrid CNN-KAN models, along with transfer learning concepts and strategies. Section 4 presents the proposed adaptive transfer learning framework and experimental setup. Section 5 discusses the results and interpretability analysis. Finally, Section 6 concludes the article and outlines future research directions.

II. Related Works

A. Transfer Learning and Domain Adaptation in Industrial Monitoring

Transfer Learning (TL) has been extensively studied as an effective mechanism to mitigate data scarcity and distribution mismatch across domains [38, 46]. In industrial condition monitoring, TL enables the reuse of knowledge acquired from a source process to improve diagnostic performance in target environments with limited labeled data and varying operating conditions.

10.37965/jdmd.2026.1597

Early domain adaptation approaches focused on learning domain-invariant representations through statistical alignment or adversarial learning mechanisms. Representative methods include discrepancy-based techniques relying on Maximum Mean Discrepancy (MMD) and correlation alignment methods such as CORAL and DeepCORAL [45, 64, 65]. Adversarial approaches, including Domain-Adversarial Neural Networks (DANN), further improved transferability by promoting indistinguishable latent feature distributions between source and target domains [39]. Although these methods demonstrated promising results under moderate domain discrepancies, they generally assume relatively stable operating conditions and single-source distribution shifts.

Recent research has progressively shifted toward adaptive and deployment-oriented domain adaptation strategies capable of handling realistic industrial scenarios characterized by noisy measurements, evolving operating conditions, and heterogeneous fault patterns. In this context, Smart Filter-Aided Domain Adversarial Neural Networks (SFDANN) were introduced to improve robustness under noisy industrial environments through adaptive filtering and adversarial learning mechanisms [40]. Continual Unsupervised Domain Adaptation (CoUDA) frameworks further extended industrial transfer learning toward dynamic operational conditions by incorporating continual adaptation and catastrophic forgetting mitigation strategies [41].

Similarly, recent source-free domain adaptation approaches avoid direct access to source-domain data during deployment, improving practical applicability in industrial systems with restricted data-sharing constraints [42]. Open-set and progressive self-training approaches, such as Dual-Path Adversarial Progressive Self-Training Networks (DAPSN), address cross-condition adaptation under unknown operational states and category mismatch scenarios [43]. In parallel, Interactive Residual Domain Adaptation Networks (IRDAN) and related partial adaptation frameworks have explored adaptive residual transfer mechanisms to improve industrial fault diagnosis under heterogeneous domain discrepancies [44].

Despite these advances, most existing approaches remain specialized for isolated adaptation scenarios, such as noise-aware adaptation, continual adaptation, or source-free transfer. In real industrial environments, however, domain variability is often multi-factorial, simultaneously involving cross-process discrepancies, measurement degradation, sensor disturbances, and evolving operational conditions. Recent surveys highlight that handling such heterogeneous and compound domain shifts remains an open challenge in industrial transfer learning [47, 48].

Therefore, there is still a need for adaptive transfer learning frameworks capable of jointly addressing heteroge-neous domain discrepancies while maintaining robustness, adaptability, and interpretability under realistic industrial operating conditions.

Unlike most recent approaches that specialize in a particular adaptation scenario, the proposed framework aims to provide a unified adaptive transfer strategy capable of handling heterogeneous cross-process and noise-induced domain shifts under realistic industrial monitoring conditions.

B. Joint monitoring for fault diagnosis and cyberattack detection.

Fault diagnosis and cyberattack detection have traditionally been treated as separate problems, mainly due to their origins in Operations Technology (OT) and Information Technology (IT), respectively [49, 50]. However, recent trends emphasize the integration of both perspectives to enhance situational awareness and operational efficiency in industrial systems [51, 52].

Within this unified context, existing approaches can be broadly classified into model-based and data-driven methods. Model-based techniques offer high interpretability by leveraging explicit process knowledge, but their applicability is limited by the need for accurate system models, sensitivity to noise, and poor transferability across different processes [53, 54]. In contrast, data-driven approaches learn directly from historical data and are more flexible, although they often rely on assumptions such as signal stationarity, struggle with unseen events, or depend on expert-defined features, which can limit scalability and robustness [55, 56].

Overall, developing unified condition monitoring frameworks capable of jointly addressing fault diagnosis and cyberattack detection while remaining robust to domain shift, noise, and unseen events remains an open challenge. In particular, the integration of adaptive transfer learning with online monitoring for heterogeneous industrial data is still insufficiently explored.

C. Adaptive Fine-Tuning Strategies

Conventional TL approaches typically rely on fixed adaptation schemes, such as freezing the first few layers and fine-tuning the upper layers, or retraining the entire network based on the size of the target dataset. While these strategies are effective for moderate changes in distribution, they often fail to generalize across different degrees and types of discrepancies between domains.

Recent research has highlighted the need for more flexible adaptation mechanisms that can dynamically adjust to domain characteristics [5, 18]. In particular, the effectiveness of fine-tuning depends heavily on factors such as the similarity between the source and target domains, the presence of noise, and the availability of labeled data. Therefore, static adaptation strategies are insufficient in scenarios involving complex and heterogeneous domain changes.

To overcome these limitations, several studies have explored adaptive and automated approaches, including hyperparameter optimization and meta-learning techniques. Evolutionary algorithms, such as Differential Evolution (DE), have been used to optimize training configurations due to their robustness and global search capabilities. However, existing approaches typically focus on optimizing either parametric adaptation (e.g., learning rates, number of trainable layers) or structural adaptation (e.g., network depth or width) independently [57, 58].

In contrast, the joint optimization of structural and parametric adaptation remains underexplored, particularly in the context of industrial monitoring under domain-changing conditions. This limitation justifies the need for adaptive transfer learning frameworks capable of automatically determining the most appropriate adaptation configuration based on domain characteristics.

D. Interpretability in Industrial System and Kolmogorov-Arnold Networks

On the other hand, interpretability has become a fundamental requirement in

industrial systems, especially in safety-critical applications where transparency, traceability, and regulatory compliance are essential [31]. However, most deep learning models used for condition monitoring systems are inherently black-box, providing limited insight into how predictions are generated.

To mitigate this limitation, post-hoc explanation techniques, such as feature attribution and surrogate models, have been widely adopted [59, 60]. While these methods provide useful approximations of model behavior, they do not fully capture the internal decision-making mechanisms and can introduce additional uncertainty into the interpretation. Consequently, there is growing interest in developing intrinsically interpretable models that integrate transparency directly into their architecture [61, 62].

Unlike conventional neural networks that rely on fixed activation functions, KAN networks replace them with trainable univariate functions, enabling a more structured and interpretable representation of nonlinear relationships [58, 35]. This formulation allows each neuron to explicitly model functional transformations, facilitating interpretability while maintaining high expressive capacity. Furthermore, preliminary studies have shown that KAN-based models can achieve performance competitive with or superior to that of traditional deep neural networks, offering greater robustness and transparency. By integrating KAN into an adaptive transfer learning process, the proposed approach aims to provide transparent and robust decision-making under complex industrial conditions characterized by heterogeneous data distributions and domain changes.

In contrast to earlier methods that mainly focus on static domain alignment, recent approaches increasingly address more realistic industrial scenarios involving noisy measurements, evolving operating conditions, source-free deployment, and open-set adaptation. However, most existing methods remain specialized for specific adaptation settings and do not simultaneously address heterogeneous domain variability, adaptive structural transfer, and interpretable decision modeling within a unified framework.

III. Materials and methods

A. Convolutional Neural Networks (CNN)

CNN are deep learning architectures that utilize convolution operations in place of general matrix multiplication in certain layers [6] (See Fig. 1). Given the temporal nature of the data in this study, a CNN1D is applied.

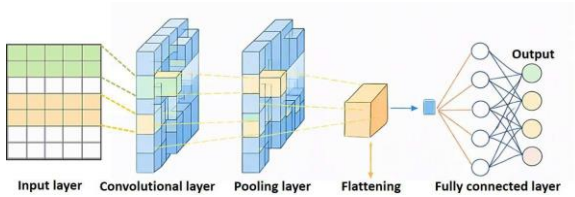


Figure. 1: General CNN architecture.

The typical architecture of a CNN for multi-class classification consists of a sequence of convolutional blocks, followed by a dense classifier. Each convolutional block usually contains a convolution layer, a ReLU activation function, and a pooling layer. The final output of these blocks is then flattened into a single vector. This vector serves as the input to the fully connected layers, which perform the classification, with the output layer employing a softmax activation function to generate class probabilities.

10.37965/jdmd.2026.1597

B. Kolmogorov - Arnold Networks (KAN)

Kolmogorov - Arnold Networks (KAN) are grounded on the Kolmogorov - Arnold representation theorem, which states that any continuous multivariate function can be expressed as a finite superposition of continuous univariate functions and sums [58]. Unlike conventional MLPs, KAN replaces linear weighted connections with learnable univariate functions. Let the input vector be:

$$\mathbf{x} = [x_1, x_2, \dots, x_m]^\top \in \mathbb{R}^m \qquad (1)$$

The KAN models a nonlinear mapping $F\colon \mathbb{R}^m \to \mathbb{R}^q$ through a functional decomposition composed of three main stages: (i) internal functional transformation, (ii) summation layer, (iii) external functional transformation.

Internal Functional Layer

Each input variable x_i is independently transformed by a set of learnable univariate functions $f_{ij}(\cdot)$:

$$f_{ij}(x_i) = \sum_{k=1}^{K} c_{ijk} B_k(x_i), i = 1, \dots, m, \quad j = 1, \dots, q \qquad (2)$$

where: $B_k(\cdot)$ are spline basis functions, c_{ijk} are trainable coefficients, and K is the number of spline basis elements. This layer replaces the traditional linear weighted connections of MLPs by learnable nonlinear functional mappings.

Summation Layer

For each functional unit *j*, the outputs of the internal functions are added according to the Equation (3). This summation implements the additive structure proposed in the Kolmogorov–Arnold representation theorem.

$$u_j = \sum_{i=1}^{m} f_{ij}(x_i) \qquad (3)$$

External Functional Layer

Each aggregated value u_j is further transformed by a nonlinear external function $\phi_j(\cdot)\colon y_j = \phi_j(u_j)$. The external functions can also be parameterized using spline bases:

$$\phi_j(u_j) = \sum_{l=1}^{L} d_{jl} \tilde{B}_l(u_j) \qquad (4)$$

where: $\tilde{B}_l(\cdot)$ are spline basis functions, d_{jl} are trainable coefficients, and L is the number of spline basis elements in the external layer.

Output Layer

The final network output is obtained as:

$$F(\mathbf{x}) = \sum_{j=1}^{q} y_j \qquad (5)$$

or, in vector form:

$$\mathbf{y} = [\phi_1(\sum_{i=1}^{m} f_{i1}(x_i)), \dots, \phi_q(\sum_{i=1}^{m} f_{iq}(x_i))]^\top \qquad (6)$$

The KAN architecture can be summarized according to the Equation (7) and shown in the Fig. 2, which explicitly highlights the functional decomposition structure.

$$F(\mathbf{x}) = \sum_{j=1}^{q} \phi_j(\sum_{i=1}^{m} \sum_{k=1}^{K} c_{ijk} B_k(x_i)) \qquad (7)$$

C. Hybrid CNN-KAN approach

The CNN-KAN architecture combines the hierarchical feature extraction capabilities of Convolutional Neural Networks (CNNs) with the nonlinear and interpretable functional modeling framework of Kolmogorov-Arnold Networks (KANs).

10.37965/jdmd.2026.1597

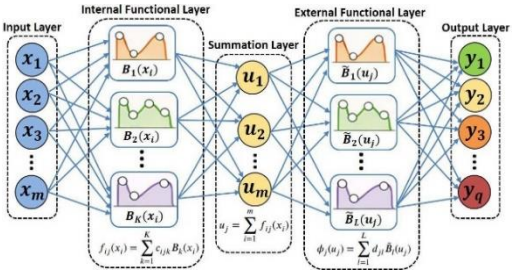


Figure 2: General KAN architecture.

The CNN module extracts latent representations, while the KAN module performs functional modeling on these latent variables, enabling the analysis of industrial time series data under heterogeneous operating conditions.

First, a CNN processes signals from multivariable sensors to extract compact latent representations, capturing local time dependencies, transient patterns, and noise-robust structural features. This stage performs dimensionality reduction, preserving the discriminative dynamics of the process. Subsequently, the KAN network replaces the conventional fully connected classifier with a functional layer in which learnable spline-based transformations model nonlinear interactions between latent features. Unlike standard MLP layers, which rely on fixed activation functions and scalar weights, KAN implements learnable univariate functional mappings and structured summation, enabling a more fluid and expressive approach to complex relationships. This hybrid approach decouples feature extraction from functional decision modeling, improving flexibility, interpretability, and robustness, particularly in scenarios involving domain shifting, measurement noise, or changing operating regimes.

CNN Feature Extraction

Let the input be a multivariate time-series matrix: $\mathbf{X} \in \mathbb{R}^{T\times d}$, where: T is the number of temporal observations and d is the number of measured variables. Then, the CNN block computes: $z = \Phi_{\text{CNN}}\mathbf{X}$, where $z \in \mathbb{R}^p$ is a latent feature vector. Each convolutional layer is defined as:

$$h^{(l)} = \sigma(W^{(l)} * h^{(l-1)} + b^{(l)}) \quad (8)$$

Where * denotes convolution and σ is a nonlinear activation. After pooling and final projection: $\mathbf{z} = \text{Flatten}(\mathbf{h}^{(L)})$.

KAN Functional Classifier

Each connection between latent variable z_i and functional unit j is modeled by a learnable univariate function:

$$f_{ij}(z_i) = \sum_{k=1}^{K} c_{ijk} B_k(z_i) \quad (9)$$

The intermediate aggregation within each functional unit is:

$$u_j = \sum_{i=1}^{m} f_{ij}(z_i) \quad (10)$$

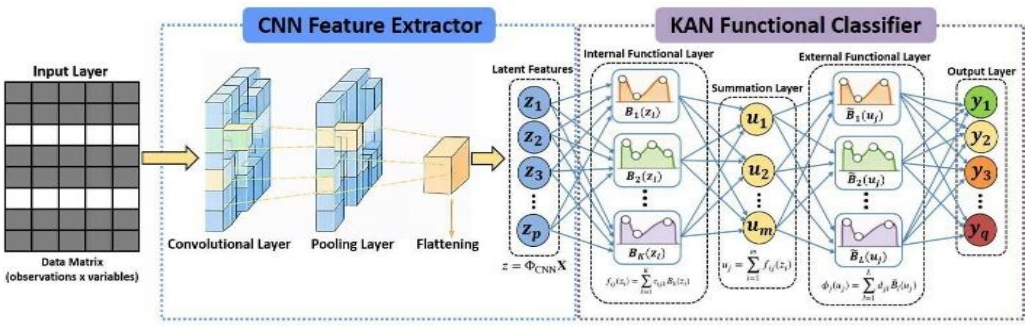


Figure 3: General CNN-KAN architecture.

Each aggregated variable u_j is processed through an external nonlinear functional mapping: $y_j = \phi_j(u_j)$, with $\phi_j(u_j)$ defined in Equation (4). Therefore, the output of the complete CNN-KAN model is defined in Equation (11). This hybrid architecture can be observed in Fig. 3.

$$F(\mathbf{X}) = \sum_{j=1}^{q} \phi_j\left(\sum_{i=1}^{m} f_{ij}(z_i)\right) \quad (11)$$

D. Transfer learning approach

General concepts

Deep learning models can be developed either through training from scratch or via transfer learning (TL) using pre-trained models. While building models from scratch is ideal for highly specialized tasks, this approach demands substantial datasets to achieve satisfactory accuracy. Therefore, a TL-based approach can achieve higher accuracy in less time.

Following the definitions in [21, 34], TL implementations fall into four main categories. Instance-based TL enhances learning in the target domain by re-weighting and reusing source data. Feature-based TL adapts the feature representation from the source for use in the target domain. Parameter-based TL, also known as model-based TL, transfers knowledge by sharing model parameters (weights) across domains under the assumption of shared underlying patterns. Finally, relation-based TL focuses on extracting and applying relational knowledge, such as logical rules, from the source to the target data. In this work, we employ parameter-based TL, specifically using pre-training and fine-tuning techniques, for detecting and localizing anomalous operational behaviors. This approach transfers knowledge by sharing critical model structure information across domains. According to the definitions proposed by [27], TL can be formally described in terms of *Domain* and *Task.*

Domain: A *Domain* $\mathcal{D}$ is defined as a pair composed of a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X = \{x_1, x_2, \ldots, x_n\} \in \mathcal{X}$. Thus, a domain can be expressed as: $\mathcal{D} = \{\mathcal{X}, P(X)\}$.

Task: A *Task* $\mathcal{T}$ is defined as a pair consisting of a label space $\mathcal{Y}$ and an objective predictive function $f(\cdot)$ that is learned from the training data, which maps input samples *X* to their corresponding labels *Y*. Formally: $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$.

Transfer Learning: *Transfer Learning* aims to improve the learning of the target predictive function $f_T(\cdot)$ in a target domain $\mathcal{D}_T$ and task $\mathcal{T}_T$, by leveraging the knowledge contained in a source domain $\mathcal{D}_s$ and task T_S. This occurs under the condition that either the domains or the tasks are different, i.e., $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$.

Strategy

For the transfer learning, the strategy was considered, described below. To find the optimal number of *N* layers to thaw and the optimal number of *M* dense layers to add, an optimization algorithm is applied. Across numerous scientific fields, optimization algorithms inspired by biological processes have shown remarkable success in solving challenging optimization problems by efficiently approximating global optima within reasonable computational limits.

This study employs the Differential Evolution (DE) algorithm for optimal estimation of the parameters *N* and *M* due to its simple implementation, fast execution speed, and strong robustness [28]. Preliminary experiments were carried out to establish an appropriate range: $1 \leq (N, M) \leq 10$. In this case, the objective

function applied is the Mean Squared Error (MSE), associated with the classification error obtained., where S is the number of observations, Z_i is the ideal accuracy (Acc = 1), and $\hat{Z}_i$ is the accuracy obtained during model training. Subsequently, the optimization problem is formulated as:

$$min\{MSE\} = \frac{1}{S}\sum_{i=1}^{S} (Z_i - \hat{Z}_i)^2$$

subject to:

$1 \leq (N, M) \leq 10$

The primary objective of DE is to generate a population of new candidate solutions by perturbing individuals from the current population, which has evolved to a given iteration. This evolutionary process is governed by three fundamental operators: mutation, crossover, and selection. A detailed description of the algorithm can be found in [29]. The stopping criteria used are: (1) Maximum number of iterations (2) Value of the objective function.

It is important to distinguish between the training loss used for model learning and the objective function employed during the Differential Evolution (DE) optimization process. During CNN-KAN training, the network parameters are optimized using the categorical cross-entropy loss, which is appropriate for multi-class classification tasks. In contrast, the Mean Squared Error (MSE) is employed only within the DE-based optimization stage to evaluate the deviation between the ideal classification performance and the accuracy obtained by each candidate adaptation configuration. Therefore, MSE is not used for network training, but rather as a meta-level objective function guiding the hyperparameter and adaptation search process.

Strategy 1: Freeze N layers and retrain only the latest ones. In Strategy 1, the goal is to find the optimal number of *N* layers to unfreeze (retrain) and obtain maximum performance. The hypothesis of this experiment is that retraining more layers will improve performance up to a certain point, after which it could lead to overfitting. This leverages the knowledge acquired from the source dataset by applying it to the target dataset. The model adapts its deeper layers to the new task, preserving the older layers for feature extraction. The operators of the optimization algorithm described earlier are applied for this purpose. After freezing, we retrain the remaining layers with a very low learning rate. This is crucial to avoid "destroying" the already effective weights generated in the pretraining stage. This process continues iteratively until the algorithm converges.

Strategy 2: Delete the final layer and add M new layers. In Strategy 2, the goal is to find the optimal number of *M* dense layers to add to achieve maximum performance. The hypothesis of this experiment is that adding new layers will improve performance up to a certain point, after which it may lead to overfitting. These additional layers increase the model's capacity and flexibility, allowing it to better adapt to new data, thereby improving its ability to capture patterns relevant to the target task. The operators of the optimization algorithm described earlier are applied for this purpose. The modified architecture is subsequently retrained to ensure seamless integration between the existing and new layers, thus optimizing performance for the target task. This process iteratively until the algorithm converges.

The effectiveness of transfer learning strongly depends on the degree of heterogeneity between the source and target tasks. When both tasks share similar data distributions and feature characteristics, a large portion of the pretrained feature

extractor can be preserved, requiring only minor adaptation of the higher network layers. In contrast, heterogeneous industrial scenarios involving different physical processes or severe noise-induced distribution shifts require greater adaptation because high-level feature representations become increasingly task-specific. Consequently, the optimal number of frozen layers and newly added layers cannot be assumed to be fixed, but should instead be determined according to the characteristics of the target domain to balance knowledge preservation and adaptation while mitigating negative transfer.

Proposed Strategy: Hybrid Adaptive Fine-Tuning and Structural Expansion (HAFSE). Motivated by the above observations, the proposed Hybrid Adaptive Fine-Tuning and Structural Expansion (HAFSE) strategy was developed to automatically determine the transfer configuration that best matches the heterogeneity of the target task. In this Strategy, the goal is to obtain a fine-tuned model that combines progressive adaptation (fine-tuning) with structural expansion (new layers), optimizing performance without losing generalization. As can be seen, the proposed methodology shown in Fig. 4, integrates strategies 1 (Fine-Tuning → Stability) and 2 (Structural Expansion → Flexibility) into a single optimized strategy: (i) Optimize number of unfrozen layers (N) (ii) Optimize number of new layers (M) (iii) Retrain final model with (N^*, M^*).

First, Strategy 1 is implemented, starting with the last KAN layer and continuing until convergence is reached. This step optimizes the number of unfrozen layers (N). Next, Strategy 2 is applied, optimizing the number of added dense layers (M). Finally, the model is retrained with the optimal configurations N and M layers, with the aim of consolidating the adjustments obtained from both strategies into a single unified model. With this proposed strategy, an expanded and refined final CNN-KAN model is obtained, achieving maximum positive transfer between domains by combining stability (through freezing) and flexibility (through structural expansion).

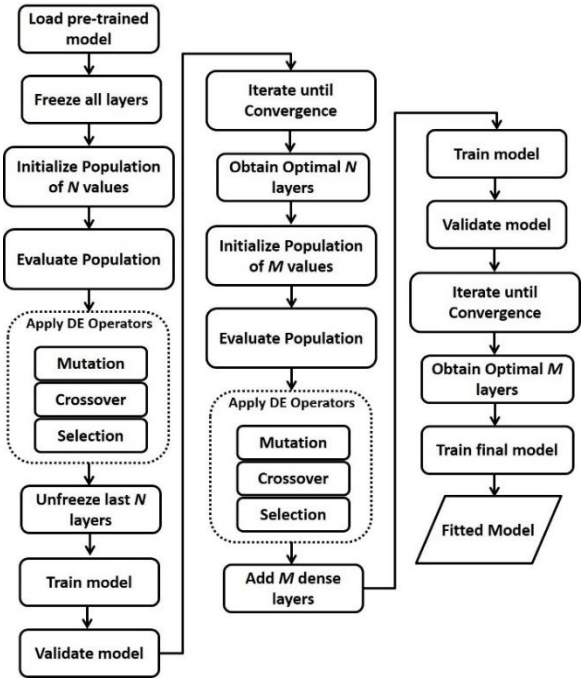


Figure 4: Methodology for optimal adjustment.

The chosen optimization order is primarily motivated by practical considerations rather than theoretical optimality. Optimizing the shared parameters N first enables stable alignment of transferable representations with the target domain while mitigating abrupt representation drift and catastrophic forgetting. Once this alignment is achieved, the expanded parameters M are optimized to capture domain-specific characteristics, reducing overfitting risks under limited target-domain data. Alternative strategies, such as alternating optimization or prioritizing M, were considered but may introduce gradient interference or unstable

10.37965/jdmd.2026.1597

adaptation, particularly in data-scarce industrial scenarios.

Mathematical Formulation of the HAFSE Strategy

Let a pretrained model on the source domain be: $f_S(X;\theta_S)$, where θ_S represents learned parameters. The target model is defined as: $f_T(X;\theta) = f_T(X;\theta_f,\theta_t,\theta_n)$, where θ_f: frozen parameters (knowledge preservation), θ_t: fine-tuned parameters (adaptation), and θ_n: newly introduced parameters (structural expansion). The target-domain loss is defined as:

$$\mathcal{L}_T(\theta_f,\theta_t,\theta_n) = \frac{1}{|D_t|} \sum_{(x_i,y_i)\in D_t} \ell(f_T(x_i;\theta_f,\theta_t,\theta_n),y_i) \quad (12)$$

where $D_t = (x_i, y_i)$: target dataset and $\ell(\cdot)$: loss function. In classification tasks, corresponds to categorical cross-entropy. The adaptation problem is formulated as:

$$\min_{\theta_t,\theta_n} \quad \mathcal{L}_T(\theta_f,\theta_t,\theta_n) + \lambda\|\theta_t - \theta_S\|^2 \quad \text{s.t.} \quad \theta_f = \theta_S^{(f)},\ \theta_t \subseteq \theta_S,\ \theta_n \sim \mathcal{I} \quad (13)$$

where $\mathcal{L}_T$: target-domain loss (adaptation) and λ: fine-tuned parameters. $\lambda > 0$ is a fixed regularization hyperparameter controlling the trade-off between adaptation and knowledge preservation. The joint optimization problem is defined as:

$$(N^*,M^*) = \arg\min_{N,M} \mathcal{L}_T\left(f_T^{(N,M)}\right) \quad \text{s.t.} 1 \le (N,M) \le K \quad (14)$$

where K is the upper limit of the search space. The regularization coefficient λ is treated as a fixed hyperparameter controlling the strength of knowledge retention from the source model. Unlike structural parameters (N,M), which are optimized via Differential Evolution, λ is selected independently using validation procedures. This separation avoids coupling structural search with regularization strength, improving optimization stability and interpretability. The proposed HAFSE Strategy is detailed in Algorithm 1.

Algorithm 1 HAFSE Strategy

Input: Source dataset $\mathcal{D}_S$, Target dataset $\mathcal{D}_T$, Pretrained model f_S with parameters θ_S.
Output: Adapted model f_T.
Train base model f_S on source domain $\mathcal{D}_S$,.
Initialize target model with pretrained weights
Step 1: Adaptive Fine-Tuning
Define search space for N
For each candidate N:
for $N = 1$ to $N = K$ **do**
    Unfreeze last N layers → θ_t
    Freeze remaining layers → θ_f
    Fine-tune model on target data $\mathcal{D}_T$
    Evaluate validation loss.
end for
Select optimal N^*.
Step 2: Structural Expansion
Define search space for M
For each candidate M:
for $M = 1$ to $M = K$ **do**
    Add M new layers → θ_n.
    Train model on target data $\mathcal{D}_T$.
    Evaluate validation loss.
end for
Select optimal M^*.
Step 3: Joint Training
Construct final model with (N^*, M^*)
Train final model minimizing Equation (13)

Although alternative hyperparameter optimization methods, such as Bayesian Optimization, Grid Search, or Random Search, could also be employed to optimize the transfer configuration, Differential Evolution (DE) was adopted in this work because of its effectiveness in solving low-dimensional, mixed-integer, and non-differentiable optimization problems. In addition, DE provides an efficient global

search mechanism without requiring gradient information, exhibits strong robustness against local optima, and has been extensively applied in machine learning and industrial optimization problems, offering a favorable balance between optimization performance and computational cost [67, 68].

Unlike conventional domain adaptation methods that rely on fixed alignment mechanisms, the proposed framework formulates transfer learning as an adaptive optimization problem via the HAFSE module. Instead of enforcing a predefined strategy, HAFSE dynamically selects the optimal adaptation configuration for each source-target pair by jointly optimizing structural and parametric components, including the number of trainable layers and key hyperparameters.

This enables the model to modulate the degree of knowledge transfer according to the nature and severity of the domain shift, which is particularly critical in industrial setting characterized by multi-factor variability, including operating conditions and noise perturbations. As a result, the proposed approach departs from static alignment paradigms by adapting the transfer mechanism itself, leading to improved robustness and generalization under heterogeneous and dynamically evolving conditions.

IV. Proposed Adaptive Transfer Learning Framework and Experimental Setup

A. Proposed TL-based approach for condition monitoring

This section presents the proposed adaptive transfer learning framework (Fig. 5), which constitutes the central contribution of this work. The framework is specifically designed to maintain reliable industrial condition monitoring performance under heterogeneous domain-shifting conditions arising from cross-process variability and realistic industrial noise. The scheme is designed to explicitly address the double distribution mismatch previously demonstrated: (i) cross-process domain shift and (ii) noise-induced domain shift. An architecture based on a CNN-KAN hybrid model is adopted to construct robust and interpretable reference models. The overall strategy consists of two stages: (1) Off-line stage (pre-training and domain adaptation), and (2) Online stage (validation).

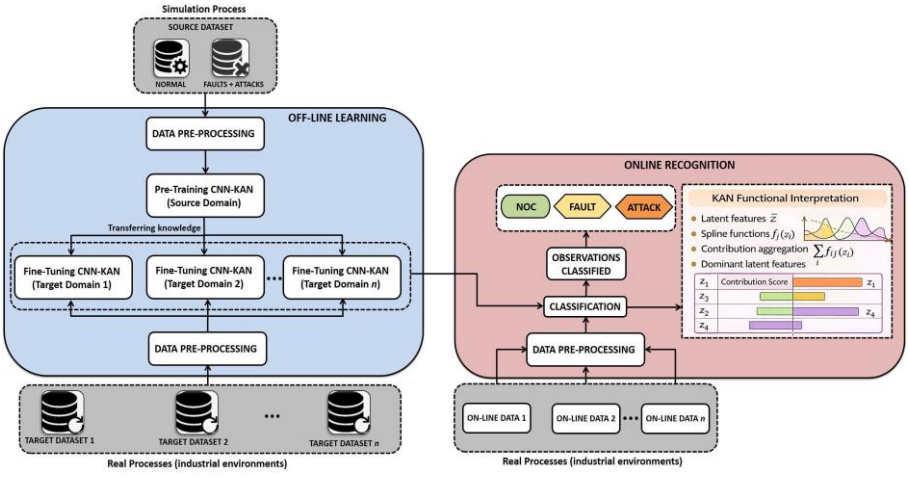


Figure 5: Proposed TL-based CM scheme**.**

10.37965/jdmd.2026.1597

Off-line stage: Pre-training and Domain Adaptation

The fundamental stage of the condition monitoring scheme is training. The development pipeline for an artificial neural network (ANN) capable of detecting and identifying industrial faults and cyber-attacks consists of four critical phases: a) data preprocessing, b) architecture design, c) model training, and d) model validation.

The preprocessing phase begins with data labeling. The dataset is then normalized to ensure that variables with larger scales do not dominate those with smaller ones, thereby retaining all relevant information for effective training. The data is split into an 80% training set and a 20% test set. Within the training set, 70% of the total data is used for direct training, while the remaining 10% is reserved for validation during the training process. The validation subset is exclusively employed during the DE optimization process to evaluate each candidate transfer configuration using the validation Mean Squared Error (MSE) as the optimization objective. After determining the optimal values of the transfer parameters, the final model is trained using the selected configuration and subsequently evaluated on the independent test set.

The training and knowledge transfer process is the heart of the proposed condition monitoring scheme. The goal is to train a robust model that can identify different operating states of multiple real-world industrial processes. The key is that a model is first trained exhaustively in a simulated environment (with all possible data), and then its knowledge is transferred and tuned to operate in specific real-world environments, where data is scarce.

The process begins with a large, labeled source dataset obtained from a simulation process. This dataset include normal operating data, faults, and cyberattack patterns. This is advantageous because it does not require a production shutdown, and faults or cyberattacks that would be too dangerous or costly to induce in real-world industrial environments can be simulated.

Source-Domain Pre-training. In this step, a deep neural network based on the CNN-KAN architecture is trained using data from the source domain. Let $\mathcal{D}_S = \{\mathcal{X}_S, P_S(X)\}$ denote the source domain. The model learns the general characteristics and patterns that distinguish between normal operation and anomalous states. Subsequently, the learned weights (knowledge) are transferred from the pre-trained CNN-KAN in the simulation to initialize new networks for real-world processes. Instead of starting from scratch (random initialization), the network starts from a baseline with a degree of intelligence. This pre-training stage allows the network to learn generic representations of industrial dynamics, fault signatures, and attack patterns.

Target-Domain Adaptation (Noise-Free Data). In the fine-tuning phase, the transfer learning strategy presented in the previous section are applied and therefore the pre-trained model is transferred to the target domain. Let $\mathcal{D}_T^{clean}$ denote the noise-free target-domain dataset. Due to cross-process domain shift $(P_S(X) \neq P_T^{clean}(X))$, direct deployment of the source model is suboptimal. Therefore, fine-tuning is performed using noise-free target-domain data.

This adaptation phase aligns the learned representations with the statistical characteristics of the target process while preserving transferable knowledge from the source domain. The result is n specialized models (one for each real-world process), each tailored to its specific environment but

10.37965/jdmd.2026.1597

leveraging the general knowledge learned in the simulation. This approach allows for more practical use in industrial environments, where collecting sufficient data, especially labeled data about faults or cyberattacks, requires a considerable investment of time, labor, and expertise from plant technicians.

Online stage: Validation Under Noisy Conditions

In real industrial environments, measurements are rarely noise-free. As will be demonstrated in the next section, industrial noise can induce an additional distribution shift: $P_T^{clean}(X) \neq P_T^{noisy}(X)$.

Therefore, in the online validation phase, the adapted model is evaluated using noisy data from the target domain $\mathcal{D}_T^{noisy}$. This implementation scenario reflects realistic operating conditions, where sensor disturbances, communication noise, and environmental disturbances alter the measurement distributions. The goal is to assess whether the adapted CNN-KAN model maintains robust performance despite the noise-induced domain shift.

In the online phase, the proposed CNN-KAN framework incorporates an intrinsic interpretability mechanism through the KAN-based decision module. Once the CNN extracts high-level representations of the incoming process data, the KAN block models the final decision using adaptive nonlinear basis functions associated with each input feature, enabling transparent mapping between the latent descriptors and the predicted output. Unlike conventional deep architectures that operate as black-box models, this structure allows for individual analysis of the contribution and sensitivity of each feature, facilitating the identification of the most influential variables in detecting faults and cyberattacks. This structural interpretability is particularly relevant in monitoring industrial conditions, where operational transparency and decision traceability are crucial. Furthermore, since the interpretability arises from the functional form of the KAN mappings, no additional post-hoc explanation tools are required, preserving computational efficiency and real-time capability during implementation.

B. Domain used

For this study, in the pre-training (Source Domain), the Tennessee Eastman (TE) process was used, considered a complete plant divided into several critical subprocesses [16]. On the other hand, as the Target Domain, the DAMADICS (Development and Application of Methods for Actuator Diagnosis in Industrial Control Systems) process benchmark [30] and the critical infrastructure of the Gas Pipeline Network (GPN) [25] were used to evaluate the proposed scheme. These domains represent real-world industrial environments and are thus characterized by measurement noise and operational disturbances.

In the experimental setup, the source and target domains differ in both system characteristics and task definitions. The source domain, based on the Tennessee Eastman (TE) process, includes a relatively large number of fault and attack categories, while the target domains comprising a intelligent electro-pneumatic system and a gas pipeline network exhibit fewer categories with distinct physical and operational properties. These differences are intentional and reflect realistic industrial deployment scenarios, where pre-trained models are often transferred between systems with heterogeneous fault taxonomies and attack manifestations. In this context, the goal of transfer learning is not to preserve class identity across domains. Rather, it aims to reuse learned representations that capture generic temporal

patterns, system dynamics, and anomaly-related features, applying them to multiple domains.

From a methodological perspective, this setting corresponds to a partial and heterogeneous task transfer scenario. While the output spaces differ between source and target domains, the underlying monitoring objective joint detection of abnormal physical and cyber behaviors remains consistent. The proposed framework therefore transfers feature extraction and temporal modeling knowledge rather than explicit class boundaries.

Source Dataset: Tennessee Eastman (TE)
The source domain dataset consists of the Tennessee Eastman (TE) process. This industrial plant is made up of five interconnected subprocesses: reactor, condenser, vapor-liquid separator, product separator, and compressor.

The Tennessee Eastman (TE) process involves 52 measured variables, with a dataset that captures both Normal Operating Conditions and 20 distinct fault scenarios. It is essential to monitor the input variables, as deviations in these can signal a cyber-attack particularly one designed to mimic normal operation. Each data sample represents a 48-hour simulation where faults are introduced at the 8-hour mark, and measurement noise is present throughout. The 20 faults employed for the analysis and evaluation of the proposed methodology are detailed in Table 1. The dataset was partitioned such that the training set contains 500 samples for each fault and the Normal Operating Condition (NOC), and the testing set contains 960 observations per class. Table 2 details the cyber-attacks, which assume a knowledgeable hacker capable of manipulating sensor measurements arbitrarily [16]. For each fault, attack, and the NOC, 500 simulations were conducted. These were split into training (350 samples, 70%), validation (50 samples, 10%), and testing (100 samples, 20%). Consequently, the overall dataset sizes are 175000 (training), 25000 (validation), and 96000 (testing) samples.

Target Dataset 1: DAMADICS Process.
The proposed method was evaluated on the DAMADICS benchmark. This reference actuator system belongs to the class of intelligent electro-pneumatic devices widely used in industrial plants. The experimental data from the DAMADICS used in this work were obtained from http://diag.mchtr.pw.edu.pl/damadics/. The general structure of this actuator comprises a set of components: a control valve, a pneumatic servomotor with spring and diaphragm, and a positioner.

10.37965/jdmd.2026.1597

Table 1. Simulated faults in the TE process.

Fault	Description	Type
F-1	A/C feed ratio, B composition constant (Stream 4)	Step
F-2	B composition, A/C ratio constant (Stream 4)	Step
F-3	D feed temperature (stream 2)	Step
F-4	Reactor cooling water inlet temperature	Step
F-5	Condenser cooling water inlet temperature (Stream 2)	Step
F-6	A feed loss (Stream 1)	Step
F-7	C header pressure loss reduced availability (Stream 4)	Step
F-8	Reactor cooling water inlet	Random variation
F-9	D feed temperature	Random variation
F-10	C feed temperature (stream 4)	Random variation
F-11	Reactor cooling water inlet temperature	Random variation
F-12	Condenser cooling water inlet temperature	Random variation
F-13	Reaction kinetics	Slow drift
F-14	Reactor cooling water	Valve sticking
F-15	Condenser cooling water valve	Valve sticking
F-16	Deviations of heat transfer within stripper	Random variation
F-17	Deviations of heat transfer within reactor	Random variation
F-18	Deviations of heat transfer within condenser	Random variation
F-19	Recycle valve of compressor, underflow stripper and steam valve stripper	Valve sticking
F-20	Unknown	Random variation

Table 2. Simulated attacks in the TE process.

Attack	Attacked sensor	Symptom variables	Description	Impact
A-1	XME-1 [+2.35]	XME-1,3,7,8	Within 3 hrs the actual value rose by 2.35	High reactor pressure or low stripper level shutdown
A-2	XME-14 [+7]	XME-7,8,12,14,15	In 2.88 hrs the actual value grew by 7	High stripper Level shutdown
A-3	XME-14 [-7]	XME-7,8,12,14,15	In 2.02 hrs the actual value is reduced by 7	Low stripper Level shutdown
A-4	XME-14 [22.9]	XME-7,12,15	Set the value to 22.9 for 1.9 hours	Low separator Level shutdown
	XME-1: A feed	XME-7: Rector pressure	XME-12: Separator level	XME-14: Separator underflow
	XME-15: Stripper level	XMV-3: A feed flow	XMV-7: Separator flow	XMV-8: Stripper flow

The faults considered are described in Table 3 and are representative of the different actuator devices. Measurements of four process variables were recorded with a sampling time of 1 second: Process control external signal (CV), Stem displacement (X), Liquid flow rate (F), and Process value (PV). Considering the six states of the process (NOC and the five faults) 600

Table 3. Faults simulated in the DAMADICS.

Class	Description
Normal (NOC)	Normal operation condition
Fault 1 (F-1)	Valve obstruction
Fault 7 (F-7)	Critical flow
Fault 12 (F-12)	Electro-pneumatic transducer fault
Fault 15 (F-15)	Positioner spring fault
Fault 19 (F-9)	Flow rate sensor fault

10.37965/jdmd.2026.1597

observations were stored (100 for each class), to evaluate the DTL-based strategy. Although publicly available, the DAMADICS benchmark was generated from a real industrial actuator benchmark.

Target Dataset 2: Gas Pipeline Network (GPN).
The proposed method was evaluated on a real gas pipeline network. The system comprises three pipelines operating at high, medium, and low pressures, each regulated by pressure converters and switches. Sensors continuously monitor key parameters including temperature, flow rate, and pressure to ensure system safety and stability.

Three types of attacks, including Denial-of-Service (DoS) attack (A-1), Tamper attack (A-2), and Wiretap attack (A-3), are simulated in this process. Few samples (800: 200 for each class) were used to evaluate DTL-based strategy. As explained above, this dataset is used to evaluate the accuracy and adaptability of the proposed scheme in practical situations. Similarly, the GPN dataset represents measurements acquired from a real industrial gas pipeline process.

The selected source-target configuration was intentionally designed to reproduce a representative industrial transfer learning scenario. The Tennessee Eastman Process constitutes a data-rich source domain containing a relatively large number of fault and cyberattack categories, whereas DAMADICS and GPN represent heterogeneous target domains with different physical processes, operating conditions, and considerably fewer labeled categories. This heterogeneous configuration reflects practical industrial deployment scenarios, where pretrained models are commonly transferred to new systems for which only limited annotated operational data are available. Consequently, the objective of the proposed framework is not to preserve identical class definitions across domains, but to transfer generic temporal and anomaly-related representations that can be effectively adapted to heterogeneous industrial environments.

C. Domain Shift

This section rigorously demonstrates the presence of domain shift in two complementary scenarios: i) structural cross-process discrepancies between industrial benchmarks, where the Tennessee Eastman (TE) process is adopted as the Source Domain and the DAMADICS and GPN processes are considered as Target Domains, and (ii) noise-induced domain shift, which evaluates the impact of measurement disturbances within the same industrial process, specifically affecting the DAMADICS and GPN datasets as Target Domains.

Domain Definition. A domain is formally defined as: $\mathcal{D} = \{\boldsymbol{\mathcal{X}}, \boldsymbol{P}(\boldsymbol{X})\}$, where $\boldsymbol{\mathcal{X}}$ denotes the feature space and $\boldsymbol{P}(\boldsymbol{X})$ the marginal probability distribution. Let $\mathcal{D}_{\boldsymbol{S}}$ denote the source domain (Tennessee Eastman process) and $\mathcal{D}_{\boldsymbol{T}}$ denote a target domain (GPN or DAMADICS process). Domain shift occurs when: $\boldsymbol{P_S}(\boldsymbol{X}) \neq \boldsymbol{P_T}(\boldsymbol{X})$. In this work, two levels of shift are analyzed: (i) Cross-process domain shift: differences between TEP and each target process, (ii) Noise-induced domain shift: differences between noise-free and noisy measurements within the same target process.

Statistical Tests. Industrial processes generate high-dimensional, non-Gaussian, and dynamically correlated data. Therefore, comparing distributions requires both univariate and multivariate statistical measures to ensure the comprehensive detection of both local and global distribution discrepancies. The following

10.37965/jdmd.2026.1597

criteria guided the selection of tests:

Kolmogorov - Smirnov (KS) test: nonparametric test sensitive to differences in cumulative distributions, suitable for detecting changes in distribution shape. Let $\{x_i\}_{i=1}^{n} \sim P$ and $\{y_j\}_{j=1}^{m} \sim Q$ be two independent samples drawn from cumulative distribution functions (CDFs) $F_P(x)$ and $F_Q(x)$, respectively. The two-sample Kolmogorov--Smirnov test evaluates the null hypothesis: $H_0: F_P(x) = F_Q(x), \forall x$. Let the empirical distribution functions (EDFs) be:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{x_i \le x\}} \quad (15)$$

$$\hat{G}_m(x) = \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}_{\{y_j \le x\}} \quad (16)$$

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function. The KS statistic is defined as the maximum absolute difference between the two empirical CDFs:

$$D_{n,m} = \sup_x |\hat{F}_n(x) - \hat{G}_m(x)| \quad (17)$$

Mann - Whitney (MW) test: robust nonparametric test for differences in central tendency when normality cannot be assumed. The Mann-Whitney U test (also known as Wilcoxon rank-sum test) evaluates whether two independent samples originate from distributions with equal central tendency. The null hypothesis is $H_0: P(X > Y) = 0.5$. Let the combined sample $\{z_k\}_{k=1}^{n+m}$ be obtained by merging $\{x_i\}$ and $\{y_j\}$ and assigning ranks $R(x_i)$. The Mann--Whitney statistic is defined as:

$$U = \sum_{i=1}^{n} R(x_i) - \frac{n(n+1)}{2} \quad (18)$$

Energy Distance (ED): metric-based multivariate measure capable of detecting arbitrary distribution differences. Let $X \sim P$ and $Y \sim Q$ be random vectors in $\mathbb{R}^d$ with finite first moments. The Energy Distance between distributions P and Q is defined as:

$$ED(P,Q) = 2\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\| \quad (19)$$

where: X, X' are independent and identically distributed (i.i.d.) random variables drawn from P, Y, Y' are i.i.d. random variables drawn from Q and $\|\cdot\|$ denotes the Euclidean norm. This estimator is consistent and suitable for multivariate distribution comparison in high-dimensional industrial data.

Maximum Mean Discrepancy (MMD): kernel-based statistic effective for high-dimensional distribution comparison. Let $\mathcal{H}$ be a Reproducing Kernel Hilbert Space (RKHS) associated with a positive definite kernel function $k(\cdot,\cdot)$. The Maximum Mean Discrepancy between P and Q is defined as:

$$MMD(P,Q) = \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \le 1} (\mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{Y\sim Q}[f(Y)]) \quad (20)$$

where: $X, X' \sim P$ independently, $Y, Y' \sim Q$ independently, $k(\cdot,\cdot)$ is typically chosen as a Gaussian (RBF) kernel. This kernel-based measure is particularly effective for detecting high-dimensional and nonlinear distributional discrepancies.

Cross-Process Domain Shift.

The Tennessee Eastman process represents a nonlinear chemical plant with reaction dynamics, recycle streams, and complex fault/attack scenarios. In contrast, the GPN process models gas transport dynamics governed by pressure-flow relationships. On the other hand, the DAMADICS process

focuses on actuator behavior and control valve faults. These processes differ in physical governing equations, dynamic response characteristics, measurement dimensionality, fault and cyberattack manifestations.

For each feature x_i, the hypothesis $H_0: F_S(x_i) = F_T(x_i)$ is tested using KS and MW tests at significance level $\alpha = 0.05$. Multivariate discrepancy is quantified ED y MMD. Significant values confirm structural domain mismatch.

Noise-Induced Domain Shift.

Although the DAMADICS and GPN datasets correspond to real industrial processes, publicly available versions containing naturally corrupted field measurements with annotated noise disturbances are not available. Therefore, synthetic perturbations were incorporated to systematically evaluate robustness under controlled conditions. The adopted perturbation models (Gaussian, Colored, Impulsive, and Mixed noise) were selected because they represent characteristic disturbance mechanisms frequently reported in industrial sensing and communication systems. Gaussian noise approximates thermal sensor noise and electronic measurement uncertainty, colored noise models temporally correlated disturbances introduced by industrial dynamics, filtering, and communication channels, impulsive noise represents sporadic perturbations caused by electromagnetic interference, switching events, sensor malfunctions, or packet transmission errors, whereas mixed noise combines these effects to emulate more realistic industrial operating conditions where multiple disturbance sources coexist. Although these models do not fully reproduce the complete spectral characteristics of every industrial installation, they provide a representative, reproducible, and widely adopted approximation for evaluating the robustness of industrial monitoring algorithms under heterogeneous disturbance scenarios.

Gaussian White Noise: It represents the sensor's thermal noise and electronic disturbances: $n_G(t) \sim \mathcal{N}(0, \sigma^2)$.

Colored Noise: This simulates correlated disturbances present in industrial dynamics. Modeled as a first-order autoregressive process: $n_C(t) = \rho n(t-1) + \epsilon(t)$, where $\epsilon(t) \sim \mathcal{N}(0, \sigma^2)$ and $|\rho| < 1$.

Impulsive Noise: This models sporadic spikes caused by communication errors, where A_k is a high-amplitude random perturbation.

$$n_I(t) = \begin{cases} A_k & \text{with probability } p \\ 0 & \text{otherwise} \end{cases} \quad (21)$$

Mixed Noise: Combining Colored and impulsive components to emulate realistic hybrid disturbances: $n_M(t) = n_C(t) + n_I(t)$

On the other hand, noise amplitude is defined as a percentage of the signal standard deviation: $\sigma_n = \lambda \cdot \sigma_x$, where $\lambda \in \{0.02, 0.05, 0.10\}$, corresponding to 2%, 5%, and 10% noise levels.

The synthetic perturbations were generated by adding the selected noise model independently to each sensor channel of the original measurements before the data normalization stage. Subsequently, the same normalization procedure was applied to both the noise-free and noisy datasets. This strategy preserves identical preprocessing conditions while allowing the influence of the injected disturbances to be systematically isolated.

Throughout this work, the term noise-free data refers to the original datasets prior to

synthetic noise injection, whereas noisy data denotes the corresponding datasets after the addition of synthetic perturbations.

Statistical Validation of Noise-Induced Shift.

For each target domain, noise-free and noisy datasets are compared using KS, MW, ED, and MMD. Noise-induced domain shift is formally established if: $P_{clean}(X) \neq P_{noisy}(X)$, and discrepancy metrics increase monotonically with λ. The combined analysis demonstrates: $P_S(X) \neq P_T^{clean}(X) \neq P_T^{noisy}(X)$.

Therefore, industrial deployment involves a double domain shift: cross-process and noise-induced. This statistically validated mismatch justifies the need for an adaptive transfer learning strategy capable of maintaining performance under heterogeneous and noisy operational conditions.

D. Design of the experiments performed

Domain Evaluation

Cross-Process Domain Shift. The statistical analysis confirms the presence of significant cross-process domain shift between the source domain (Tennessee Eastman Process, TEP) and both target domains (Gas Pipeline Network, GPN, and DAMADICS). The assessment was conducted through complementary univariate and multivariate statistical tests in order to characterize both marginal and joint distribution discrepancies.

From a univariate perspective, the Kolmogorov-Smirnov (KS) and Mann--Whitney (MW) tests were applied independently to each monitored variable under a significance level of $\alpha = 0.05$. For both TEP-GPN and TEP-DAMADICS comparisons, the null hypothesis $H_0: F_S(x_i) = F_T(x_i)$ was rejected for all evaluated variables. This systematic rejection indicates that the marginal distributions of all features differ significantly between the source and each target process. The agreement between KS, which is sensitive to global distributional differences, and MW, which captures shifts in central tendency, suggests that the discrepancy involves both changes in distributional shape and median location. Therefore, the domain shift is global and structurally embedded rather than localized to specific measurements. To complement the marginal analysis, multivariate discrepancy was quantified using Energy Distance (ED) and Maximum Mean Discrepancy (MMD). The obtained results are summarized in Table 4.

Table 4. Multivariate domain discrepancy measured using ED and MMD.

Domain Pair	ED	MMD
TEP-GPN	11284.41	1.9222
TEP-DAMADICS	11475.33	1.6390
GPN-DAMADICS	174.77	0.93398

The large Energy Distance values for TEP-GPN and TEP-DAMADICS indicate substantial differences in their joint distributions. Since ED measures the expected pairwise Euclidean discrepancy between samples, these magnitudes confirm structural separation in the feature space. Similarly, the strictly positive MMD values demonstrate that the kernel mean embeddings of the domains differ in reproducing kernel Hilbert space, implying discrepancies beyond first-order statistics and affecting higher-order dependencies among variables.

In contrast, the GPN-DAMADICS comparison exhibits a markedly smaller Energy Distance (174.77) and a lower MMD value (0.93398). Although univariate tests still reject equality in all variables (4/4 KS

10.37965/jdmd.2026.1597

and 4/4 MW rejections), the magnitude of the multivariate discrepancy is dramatically reduced compared to the source target comparisons. This indicates that, while GPN and DAMADICS are statistically different processes, their joint distributions occupy relatively closer regions in the feature space than either does with respect to TEP.

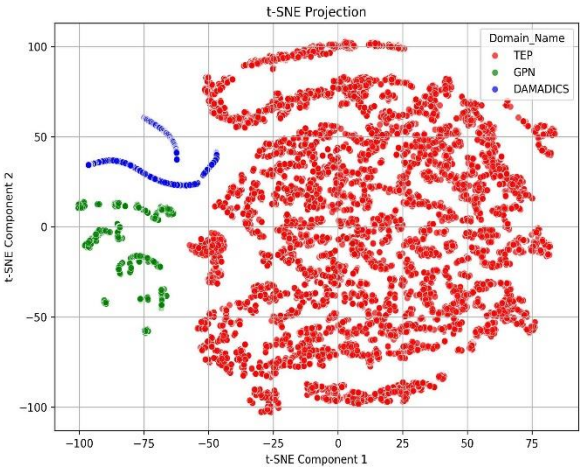


Figure 6. t-SNE projection between the source and target domains.

To better characterize the structural relationship between domains, a nonlinear manifold projection was obtained using stochastic neighbor embedding with a t-distribution (t-SNE). This technique is particularly well-suited for geometrically visualizing complex multivariate distributions and identifying cluster separation under domain shifting. Fig. 6 shows the two-dimensional embedding of TEP, GPN, and DAMADICS samples after feature standardization. The projection reveals a clear geometric separation between the source domain (TEP) and both target domains. The TEP samples occupy a large and distinct region of the embedded space, with no appreciable overlap with GPN or DAMADICS. This behavior is consistent with the large multivariate discrepancies previously reported (Energy Distance greater than 11,000 and MMD greater than 1.6 for source-target comparisons), confirming that TEP is located in a statistically distant region of the feature space. On the other hand, GPN and DAMADICS appear significantly closer to each other in the embedding, forming compact and relatively localized clusters.

Overall, the convergence of univariate and multivariate evidence demonstrates that the domain shift between processes is statistically significant and affects both marginal and joint distributions. The source domain is structurally distant from both target domains, and although the latter may show relative proximity, they remain statistically distinct processes. Additionally, the t-SNE visualization provides intuitive geometric evidence of domain switching between processes. The source domain is structurally isolated, while the two target domains show relative proximity despite remaining statistically distinct. These findings quantitatively justify the adoption of an adaptive transfer learning strategy, since direct implementation of the model without adaptation would likely suffer performance degradation under such distributional discrepancies.

Noise-Induced Domain Shift. To evaluate the impact of measurement disturbances within the same industrial process, controlled noise injections were applied to the GPN and DAMADICS datasets using four noise models (Gaussian, impulsive, colored, and mixed) and three intensity levels (2%, 5%, and 10%). The resulting noise-free-noisy pairs were compared using univariate (KS and MW) and multivariate (ED and MMD) statistical tests.

Considering the DAMADICS process, for all noise types and severity levels, the Kolmogorov-Smirnov (KS) and Mann-Whitney (MW) tests consistently rejected the null hypothesis for all monitored variables. This behavior was observed even

at the lowest noise level (2%), indicating that minimal measurement degradation is sufficient to alter the marginal distributions of the process variables. The systematic rejection across all variables confirms that the noise does not merely perturb second-order statistics (e.g., variance), but modifies the overall distributional structure. The MW results further indicate that the central tendency and rank ordering of observations are significantly affected.

Fig. 7 illustrates the evolution of the Energy Distance (ED) and Maximum Mean Discrepancy (MMD) as a function of noise type and noise level for the DAMADICS process. The results reveal a clear and systematic increase in distributional discrepancy between noise-free and noisy measurements as the noise severity increases. From a multivariate perspective, the structural deviation exhibits a nearly monotonic growth with respect to noise level for Gaussian, colored, and mixed noise. In particular, colored noise produces the greatest structural deviation. Gaussian noise shows intermediate behavior, while mixed noise follows a similar trend. Conversely, impulsive noise produces negligible ED values at all noise levels, indicating a limited impact on the overall geometry of the data distribution

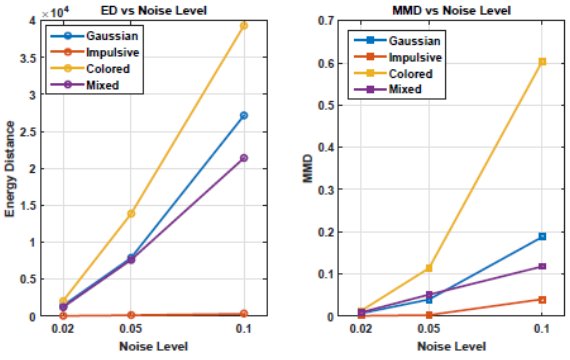


Figure 7. Multivariate Domain Shift Metrics (DAMADICS process).

The MMD results confirm this trend from the perspective of a reproductive kernel Hilbert space (RKHS). The high sensitivity to colored noise suggests that the DAMADICS dynamics amplify correlated perturbations, generating coherent distortions in the multivariate feature space. Since ED measures pairwise Euclidean discrepancies and MMD evaluates distributional divergence in a kernel space, the agreement between the two metrics reinforces the conclusion that the change is not merely geometric, but fundamentally probabilistic.

For the GPN process, from a univariate perspective, gaussian, colored, and mixed noise produce systematic changes in the distribution. For these noise types, the Kolmogorov-Smirnov (KS) test rejects the null hypothesis for most variables, especially at 5% and 10% noise levels, where extremely low p-values are observed. The Mann-Whitney (MW) test show similar behavior, confirming that both the shape of the distribution and the central tendency are affected. In contrast, impulsive noise produces minimal univariate rejection across all tested levels. KS rejection rates remain low and p-values are frequently close to unity, indicating that sparse impulsive perturbations do not significantly modify the global empirical distributions, despite affecting isolated samples.

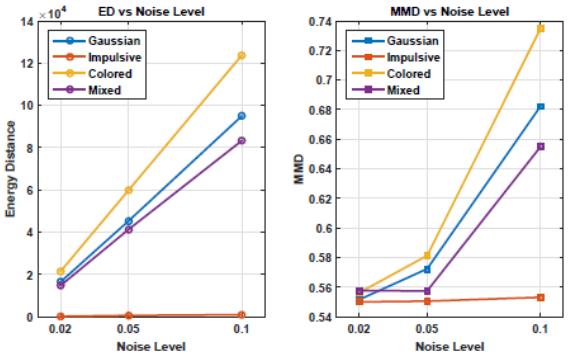


Figure 8. Multivariate Domain Shift Metrics (GPN process).

Fig. 8 shows the evolution of ED and MMD as a function of the noise severity level for the GPN process under four industrial noise models. Multivariate analysis provides a clearer quantification of structural divergence. A clear, monotonic increase in the ED is observed for Gaussian, colored, and mixed noise configurations as the noise level increases from 2% to 10%. Among the evaluated disturbances, colored noise produces the largest discrepancies. This behavior indicates that structured and temporally correlated disturbances induce substantial geometric distortion in the feature space. Gaussian noise exhibits a similar, but slightly less pronounced, trend, and mixed noise follows the same monotonic pattern. In contrast, impulsive noise produces negligible variation in the DE at all severity levels. This suggests that, under the current configuration, scattered outliers do not significantly alter the overall pairwise distance structure of the dataset. The MMD metric confirms the severity-dependent domain shift. Colored noise again produces the largest increase. Gaussian and mixed noise show a similar progression. Impulsive noise remains almost constant, indicating minimal divergence in the Hilbert space representation of the replay kernel.

These findings confirm that even intra-process variations in measurement quality can lead to statistically distinct operational domains, thereby justifying the need for adaptive Transfer Learning mechanisms capable of handling noise-induced distributional mismatch during online deployment. Overall, the results demonstrate that noise-induced domain shift in the GPN process depends strongly on the type and intensity of the noise gaussian, colored, and mixed noise generate progressive and structurally significant divergence, while impulsive noise produces a comparatively weak global shift. These findings reinforce the need for adaptive Transfer Learning mechanisms capable of handling heterogeneous noise conditions during online deployment.

Overall, the results confirm that $P_{\text{clean}}(X) \neq P_{\text{noisy}}(X)$, particularly for correlated noise and moderate-to-high severity levels. Therefore, even without modifying the underlying physical process, measurement degradation alone is sufficient to induce a structured domain shift, which may significantly affect model generalization performance. The DAMADICS results corroborate those obtained for GPN, confirming that noise-induced domain shift is systematic and process-independent. Therefore, incorporating adaptive Transfer Learning mechanisms becomes necessary to ensure robustness and generalizability under realistic industrial noise conditions.

Off-line stage: Pre-training and Domain Adaptation

The development environment used was Google Colab, using Keras, a Python library that runs on TensorFlow. Each model was trained using a computer with the following characteristics: Intel Core i7-6600U 2.6 - 2.81GHz, memory: 16GB DDR3L. Before training, the data is normalized to standardize value ranges and labeled for the supervised learning models.

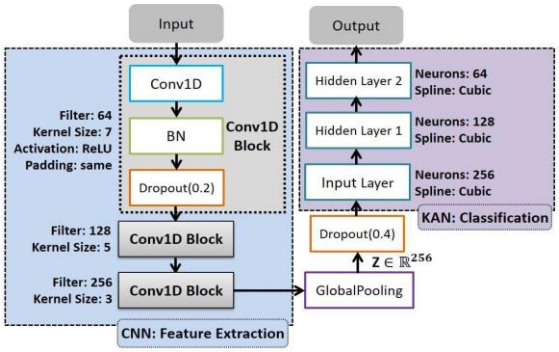


Figure 9. Detailed architecture of the proposed hybrid CNN-KAN model

10.37965/jdmd.2026.1597

Subsequently, the neural network architecture is constructed by defining the parameters for each layer. Fig. 9 shows the parameters used for the CNN-KAN model

The adopted hybrid CNN-KAN architecture, consisting of three convolutional layers followed by three KAN layers, was selected after extensive empirical evaluation and architectural exploration. Different configurations varying the number of convolutional and recurrent layers were systematically tested, and the selected structure consistently achieved the most favorable balance between representational capacity, generalization, and computational efficiency.

The input layer contains the sequential data, followed by three convolutional blocks, each with a Conv1D layer, Batch Normalization (BN), and Dropout. Increasing numbers of filters (64, 128, and 256) with decreasing kernel sizes (7, 5, and 3) are used to capture patterns at different timescales, from broad patterns to very fine details. The gradual increase in the number of filters (from 64 to 256) allows the network to learn increasingly diverse feature maps while preserving computational efficiency. Each convolutional layer is followed by Batch Normalization and a ReLU activation function. Batch Normalization stabilizes training by reducing internal covariate shifting, enabling higher learning rates and improving robustness under domain shifting conditions. ReLU introduces nonlinearity while preserving computational efficiency. The Dropout function is applied after nonlinear activations to mitigate overfitting and reduce co-fitting between filters, which is especially important in transfer learning scenarios where the source and target data distributions differ. Instead of using a fully connected flattening operation, a Global Average Pooling layer (AdaptiveAvgPool1d(1)) is employed. This operation produces a compact latent vector $\mathbf{z} \in \mathbb{R}^{256}$, reduces the number of trainable parameters, introduces translational invariance along the temporal axis, and improves generalization performance.

Prior to the KAN classifier, a dropout layer is applied to the latent vector $\mathbf{z}$. This design choice plays a critical regularization role at the interface between feature extraction and nonlinear functional modeling. Although the latent representation is already compact, it may still contain correlated or dominant feature activations. Given the high modeling flexibility of KAN layers, where each connection is represented by learnable spline functions direct propagation of the full latent vector could encourage overfitting.

The latent dropout introduces stochastic feature suppression during training, preventing the KAN classifier from over-relying on specific latent dimensions. This encourages distributed representation learning and improves robustness under domain variability, noise perturbations, and unseen operating regimes. The relatively higher dropout rate (0.4) is intentionally selected because this stage directly feeds the high-capacity nonlinear decision block.

The latent representation is fed into a Kolmogorov-Arnold network to capture sequential dependencies for classification. Progressive dimensionality reduction $256 \rightarrow 128 \rightarrow 64 \rightarrow C$ (C denotes the number of output classes) applies structured compression while preserving nonlinear separability. Two intermediate hidden layers are used to allow higher-order nonlinear interactions between latent features. Three layers in KAN offer greater functional capacity than a single layer without incurring excessive complexity.

In KAN, each connection is modeled as a spline function. In this work, $k=3$ (cubic

splines) is selected to ensure sufficient flexibility for nonlinear modeling and numerical stability without excessive oscillatory behavior.ntinuity, sufficient flexibility for nonlinear modeling, and numerical stability without excessive oscillatory behavior. The random seed is fixed to ensure experimental reproducibility and stable initialization of spline coefficients. The final KAN layer outputs unnormalized logits. Class posterior probabilities are obtained via the softmax function during inference. Cross-entropy loss is employed during training, thereby avoiding explicit softmax application within the forward pass and ensuring numerical stability.

The CNN block performs robust multiscale feature extraction and noise reduction from raw time-series data, producing a compact latent embedding. The KAN block then models structured nonlinear relationships within this latent space through functional transformations rather than simple affine mappings. This combination enhances representational capacity while maintaining interpretability at the classifier level, making the architecture particularly suitable for industrial condition monitoring under heterogeneous operating conditions and potential domain shift scenarios.

The adopted architectural configuration directly influences the monitoring performance under domain-shift and noisy conditions. In particular, the multiscale CNN structure improves robustness by capturing degradation patterns at different temporal resolutions, while the KAN classifier enhances nonlinear representation capability through functional modeling in the latent space. The progressive feature compression and regularization mechanisms further improve generalization and reduce sensitivity to noise perturbations. These design choices are experimentally validated in the ablation study, where the proposed CNN-KAN architecture consistently outperforms CNN, LSTM, CNN-LSTM, and standalone KAN models across multiple operating conditions and noise scenarios.

To implement the DE algorithm, three strategies were considered: i) one favoring diversification (exploration), ii) one favoring intensification (exploitation), and iii) one exhibiting a balanced trade-off between diversification and intensification. As the Friedman non-parametric test revealed no significant differences among the three strategies, the third strategy was chosen. The parameters in DE are the population size Z, the crossover constant C_R, and the scaling factor F_S. The parameter values for the DE algorithm were then: $C_R = 0.5, F_S = 0.1,\ Z = 10, Itr\_max = 100,$ and $MSE = 0.0001$. To reduce the influence of the stochastic nature of Differential Evolution, the optimization procedure was independently executed 10 times for each transfer learning scenario. The transfer configuration yielding the best validation performance was adopted for the final experimental evaluation.

In the experiments, an evaluation was performed using the source and target data sets to assess the performance of each pre-trained model. To do so, well-known metrics such as accuracy, recall, precision, and the F1 score were used to evaluate the performance of the proposed method.

Online stage: Validation Under Noisy Conditions.

The online validation stage is designed to evaluate the robustness of the adapted CNN-KAN model under realistic measurement disturbances. After transfer learning adaptation using noise-free data from each target process, the resulting model is deployed in an online inference scenario where input measurements are corrupted by

synthetic but physically motivated noise. Two target processes are considered: the Gas Pipeline Network (GPN) process and the DAMADICS actuator benchmark. For each process, three representative noise structures are evaluated: Gaussian, colored, and mixed noise. Noise intensity levels are defined as 2%, 5%, and 10% of the standard deviation of each measured variable, there by preserving scale consistency across heterogeneous features. For each target process and each noise configuration (3 noise types × 3 levels), the following experimental protocol is applied:

1. The CNN-KAN model previously adapted using noise-free target-domain data is kept fixed (no additional retraining is performed).
2. Noisy measurements are injected at the input stage to simulate real-time acquisition conditions.
    1. The model performs inference in a streaming fashion, generating class predictions for faults and cyberattack scenarios.
2. Performance metrics are computed and compared against the noise-free-data baseline.

The primary objective is to quantify performance degradation as a function of both noise structure and intensity. Standard classification metrics are computed, including accuracy, precision, recall, and F1-score. By conducting the same protocol on both DAMADICS and GPN processes, it becomes possible to assess whether robustness behavior is process-dependent or structurally consistent across heterogeneous industrial dynamics. Since these processes differ significantly in physical governing equations and measurement characteristics, consistent robustness patterns would indicate structural generalization capability of the proposed CNN-KAN architecture.

Importantly, interpretability is preserved during the online phase. The KAN-based decision module allows inspection of the adaptive nonlinear basis functions associated with each latent feature. Consequently, variations in feature contribution under increasing noise levels can be analyzed without the need for external post-hoc explanation tools. This capability is particularly relevant in industrial monitoring applications, where traceability of decision mechanisms is essential for operational acceptance. Overall, this online validation framework provides a realistic and rigorous assessment of whether the transfer-learned CNN-KAN model maintains reliable diagnostic performance under noise-induced domain shift conditions representative of real industrial environments.

V.Results and discussion

A. Off-line stage

Source-Domain Pre-training

The results were analyzed using metrics to evaluate the performance of the CNN-KAN model in the transfer learning strategy used (HAFSE). Table 5 shows the excellent performance obtained on our source dataset, with an accuracy of 98.02%, a completeness of 97.94%, a precision of 98.32%, and an F1 score of 98.13%. Furthermore, it exhibited acceptable training times of 2726.88 seconds. Fig. 10 shows the CNN-KAN model's performance in terms of training and validation accuracy, as well as test and validation loss.

Table 5. Performance of CNN-KAN model on Source Dataset (TEP process).

Architecture	Accuracy(%)	Recall(%)	Precision(%)	F1-score(%)	Train time(sec)
CNN-KAN	97.86	97.75	98.12	97.03	2726.88

10.37965/jdmd.2026.1597

Table 6. Performance of CNN-KAN model on Target Datasets (Without DTL strategy).

Architecture	Data Target	Without DTL				
		Accuracy(%)	Recall(%)	Precision(%)	F1-score(%)	Train time(sec)
CNN-KAN	1	95.63	95.61	95.74	95.67	2024.64
	2	93.62	93.60	93.81	93.86	1896.42

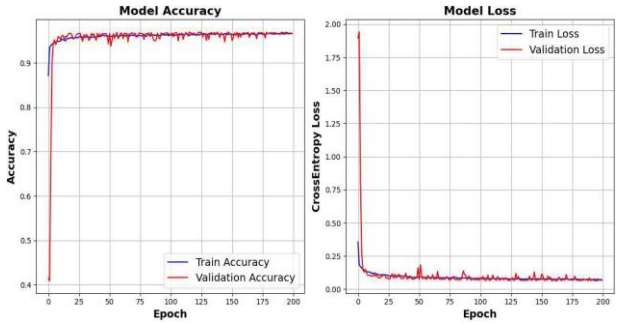


Figure 10. Training and Validation Performance (%) of the CNN-KAN model on Source Dataset.

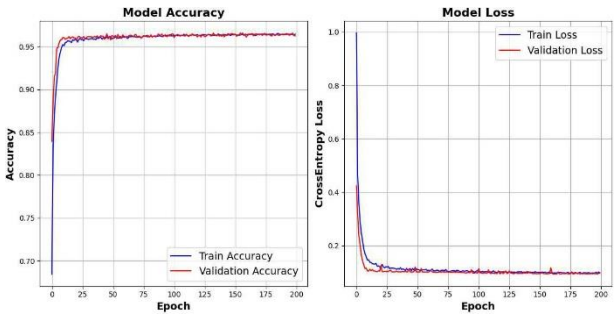


Figure 11. Training and Validation Performance (%) of the CNN-KAN model on Target Dataset 1 (Without DTL strategy).

Target-Domain Adaptation (Noise-Free Data)

Table 6 shows good performance of the CNN-KAN model on both target datasets, without applying any DTL strategy. For Target Dataset 1 (DAMADICS), CNN-KAN achieved an precision and F1 score of 96.74% and 98.67%, respectively. Furthermore, CNN-KAN performed excellently for Target Dataset 1, with a high recall rate of 96.61%. This algorithm also achieved a high accuracy of 96.63%.

For Target Dataset 2 (GPN), it performed well in terms of accuracy and recall, with 93.44% and 93.39%, respectively. CNN-KAN also achieved an precision of 93.67% and an F1 score of 93.53%. Training times were acceptable on both target datasets. Figs. 11 and 12 shows the performance of the CNN-KAN model (without DTL strategy) in terms of training and validation accuracy, and test and validation loss.

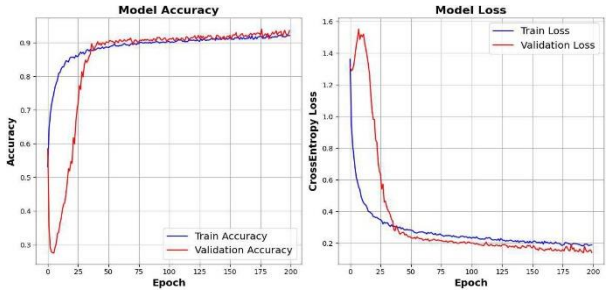
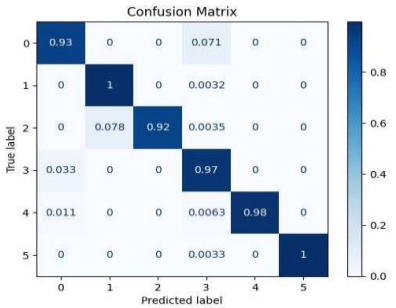


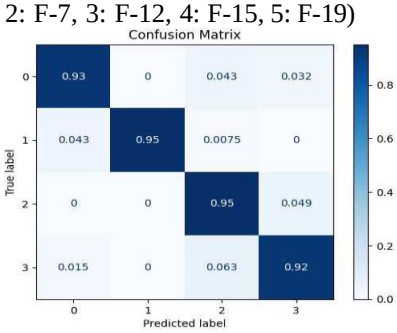
Figure 12. Training and Validation Performance (%) of the CNN-KAN model on Target Dataset 2 (Without DTL strategy).

The confusion matrix is a standard tool for evaluating classifier performance, where the diagonal represents correct predictions and off-diagonal cells represent errors. Figure 13 shows the confusion matrices obtained with the CNN-KAN model using Target Dataset 1 and Target Dataset 2 (without DTL strategy).

10.37965/jdmd.2026.1597



(a)Target Dataset 1 (0: NOC, 1: F-1, 2: F-7, 3: F-12, 4: F-15, 5: F-19)



(b) Target Dataset 2 (0: NOC, 1: A-1, 2: A-2, 3: A-3)

Figure 13. Confusion matrix of CNN-KAN model without using DTL strategy.

On the other hand, when applying the methodology described in Fig. 4 to both target datasets, the best HAFSE Strategy is obtained for N = 2 and M = 1. Fig. 14 reflects a rapid convergence of the objective function (MSE), which constitutes a significant advantage of DE as an optimization algorithm. Table 7 shows the performance of the CNN-KAN model on the two target datasets, applying the proposed strategy. The obtained optimization results indicate that different target tasks require distinct transfer configurations, confirming that task heterogeneity significantly influences the optimal balance between preserving pretrained representations and learning target-specific features.

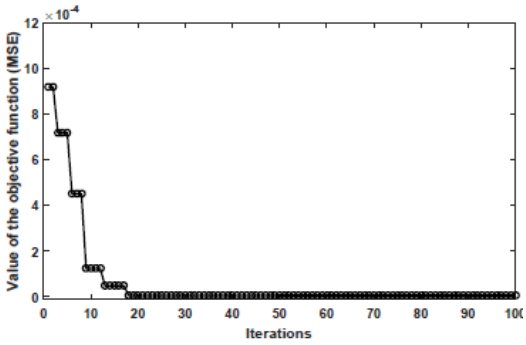


Figure 14. Value of the objective function (MSE).

In all metrics used, the proposed HAFSE Strategy showed the best performance compared to the scheme without a DTL strategy (see Table 7). On the one hand, these results validate the hypothesis that the proposed HAFSE Strategy achieves a balance between stability (through freezing) and adaptability (through structural expansion), as it prevents the degradation of pre-trained knowledge while adapting to new domain characteristics and improving its ability to capture relevant patterns.

It is very important to mention that these results were obtained using minimal training data, resulting in a more practical condition monitoring system for industrial environments. Furthermore, the presented TL-based approach achieves higher performance in less time.

Table 7. Performance of CNN-KAN model on Target Datasets (With DTL strategy).

Architecture	Data Target	With DTL				
		Accuracy(%)	Recall(%)	Precision(%)	F1-score(%)	Train time(sec)
CNN-KAN	1	97.46	97.43	97.58	97.50	909.73
	2	96.94	96.90	97.04	96.97	858.40

10.37965/jdmd.2026.1597

The HAFSE strategy based on the CNN-KAN architecture demonstrates consistent improvements across all evaluation metrics on the target datasets. As shown in Table 8, the proposed approach achieves performance gains in terms of Accuracy, Recall, Precision, and F1-score for both target processes.

Table 8. Improvement in metrics for DTL strategy on the target datasets.

Metric	Data Target	Strategy HAFSE
Accuracy	1	1.83%
	2	3.32%
Recall	1	1.82%
	2	3.30%
Precision	1	1.84%
	2	3.23%
F1-score	1	1.83%
	2	3.11%
Training time savings	1	1114.91 sec
	2	1038.02 sec

For Target Dataset 1, the improvements range between 1.82% and 1.84%. Although moderate, these improvements are consistent across all evaluation indicators, demonstrating that transfer learning provides quantifiable benefits. Furthermore, a small reduction in missed detections or false alarms can translate into earlier intervention and reduced downtime. In contrast, Target Dataset 2 exhibits a larger improvement, with gains between 3.11% and 3.32%. These results confirm that the proposed transfer learning procedures better leverage cross-domain knowledge when the distribution of the target data differs substantially from that of the source data, highlighting their robustness and adaptability.

In addition, the proposed strategy also provides significant computational benefits. The results show training time savings of approximately 1114.91 seconds and 1038.02 seconds for datasets 1 and 2, respectively. These reductions highlight the efficiency of the transfer learning framework, which avoids training the model entirely from scratch by leveraging previously learned representations. Such computational savings are particularly valuable in industrial environments where models may require frequent updates or deployment under limited computational resources.

The improvements obtained in classification performance and training cost validate the relevance of transfer learning for real industrial monitoring scenarios, where changes in data and its distribution are frequent. Therefore, the proposed DTL framework not only improves performance, but also strengthens the generalizability of the CNN-KAN model in heterogeneous environments. Furthermore, Figs. 15 and 16 shows the behavior of the accuracy and loss during training and validation (with DTL strategy). On the other hand, Fig. 17 shows the confusion matrices obtained for each model (using DTL strategy) using the target datasets 1 and 2, respectively.

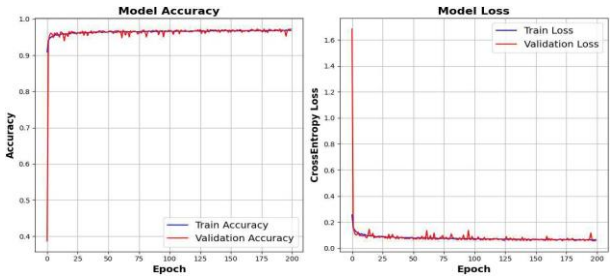


Figure 15. Training and Validation Performance (%) of the CNN-KAN model on Target Dataset 1 (With DTL strategy).

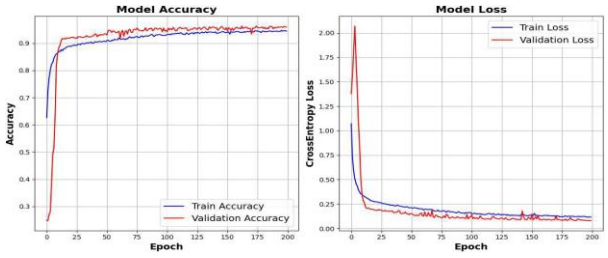
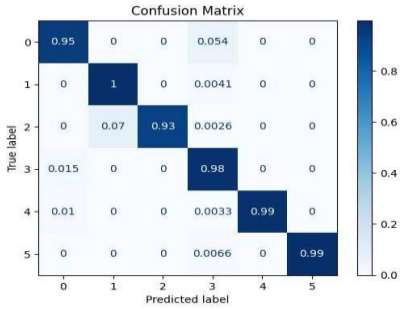
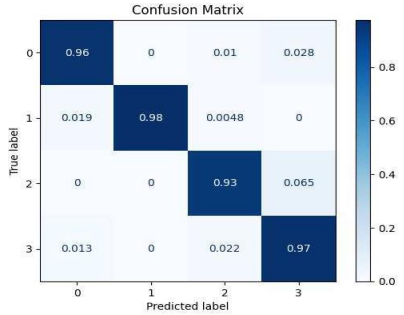


Figure 16. Training and Validation Performance (%) of the CNN-KAN model on Target Dataset 2 (With DTL strategy).

10.37965/jdmd.2026.1597



(a)Target Dataset 1 (0: NOC, 1: F-1, 2: F-7, 3: F-12, 4: F-15, 5: F-19)



(b) Target Dataset 2 (0: NOC, 1: A-1, 2: A-2, 3: A-3)

Figure 17. Confusion matrix of CNN-KAN model using DTL strategy.

Ablation Study of Backbone Architectures

To evaluate the impact of the backbone architecture and the proposed adaptive transfer learning strategy, an ablation study was conducted considering five model configurations: CNN, LSTM, CNN-LSTM, KAN, and CNN-KAN. Each model was evaluated under two conditions: (i) without transfer learning and (ii) with the proposed HAFSE strategy. Experiments were carried out on two datasets (TD1: DAMADICS and TD2: GPN), and performance was assessed using Accuracy, Recall, Precision, and F1-score.

The objective of this ablation study is not to compare neural network architectures under identical model capacity, but rather to investigate the influence of representative backbone architectures on the effectiveness of the proposed HAFSE transfer learning framework widely adopted in industrial condition monitoring. Therefore, each backbone was implemented using its standard architecture without artificially modifying its structure to match the number of trainable parameters, thereby preserving its representative characteristics and enabling a realistic evaluation of the proposed transfer learning strategy.

To improve the transparency of the ablation study, Table 9 summarizes the computational complexity of the evaluated backbone architectures, including the number of trainable parameters, training time, inference latency, and qualitative complexity level. All measurements were obtained under the same hardware configuration, identical data partitions, and the same training protocol. These results provide additional context for interpreting the predictive performance of each backbone while preserving the original objective of the ablation analysis.

The complexity analysis indicates that the evaluated backbone architectures exhibit different computational requirements and model capacities. Although the hybrid CNN--KAN architecture presents the highest computational complexity, it also achieves the best predictive performance while maintaining inference latencies compatible with real-time industrial condition monitoring applications

Table 9. Computational complexity analysis of the evaluated backbone architectures in the HAFSE ablation study

Method	Trainable Parameters	Training Time TD1 (sec)	Training Time TD2 (sec)	Inference TD1 (ms/sample)	Inference TD2 (ms/sample)	Complexity Level
CNN	1.42M	438.26	401.73	0.81	0.77	Low
LSTM	1.87M	526.84	487.61	1.06	1.01	Moderate
CNN-LSTM	2.94M	704.35	657.28	1.39	1.32	Moderate-High
KAN	2.36M	648.72	601.49	1.28	1.22	Moderate
CNN–KAN	3.76M	909.73	858.40	1.73	1.66	High

As shown in Figs. 18 a) and (c), when no transfer learning is applied, all models exhibit consistent behavior across metrics.

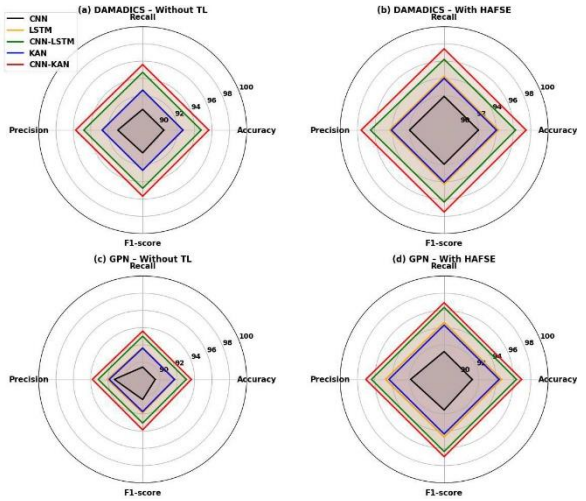


Figure 18. Radar-based comparison of backbone architectures across both target datasets, with and without transfer learning. (a) Target Dataset 1 (DAMADICS) without transfer learning, (b) Target Dataset 1 (DAMADICS) with HAFSE, (c)Target Dataset 2 (GPN) without transfer learning, and (d) Target Dataset 2 (GPN) with HAFSE.

Among the baseline architectures, CNN achieves the lowest performance (≈ 90.6% on TD1 and ≈ 90.3% on TD2 in F1-score), highlighting its limited capacity to capture temporal dependencies and complex nonlinear relationships. LSTM improves performance (≈ 92.7% on TD1 and ≈ 91.9% on TD2), confirming the importance of temporal modeling. Similarly, KAN achieves comparable results ≈ 92.7% on TD1 and ≈ 91.7% on TD2), demonstrating its ability to model nonlinear functional relationships with improved interpretability.

The hybrid CNN-LSTM model further improves performance (≈ 94.8% on TD1 and ≈ 93.1% on TD2), indicating that combining spatial and temporal feature extraction is beneficial. However, the best results are consistently obtained by the CNN-KAN architecture, achieving ≈ 95.7% on TD1 and ≈ 93.9% on TD2. This demonstrates that integrating convolutional feature extraction with the functional representation capability of KAN provides a more expressive and robust model.

On the other hand, the incorporation of HAFSE leads to systematic improvements across all architectures, as shown in Figs. 18 (b) and (d). For instance, CNN improves from ≈ 90.6% to ≈ 92.0% on TD1 and from ≈ 90.3% to ≈ 91.6% on TD2, confirming the effectiveness of transfer learning even for simple architectures. More advanced models benefit further from HAFSE. CNN-LSTM improves from ≈ 94.8% to ≈ 96.3% on TD1 and from ≈ 93.1% to ≈ 96.4% on TD2, while KAN improves from ≈ 92.7% to ≈ 94.0% on TD1 and from ≈ 91.7% to ≈ 94.3% on TD2.

Notably, the proposed CNN-KAN model achieves the highest performance under HAFSE, reaching ≈ 97.5% on TD1 and ≈ 97.0% on TD2. This corresponds to an improvement of approximately +1.8% (TD1) and +3.1% (TD2) over its non-transfer counterpart. Although CNN-KAN exhibits the highest predictive performance among the evaluated backbone architectures, it also presents the largest model capacity and computational complexity. Therefore, these results should be interpreted as evidence of the effectiveness of the proposed HAFSE framework when combined with representative backbone architectures rather than as a strict capacity-controlled demonstration of the intrinsic superiority of CNN-KAN.

B. Online stage

Validation Under Noisy Conditions: Target Dataset 1 (DAMADICS)

The results presented in Table 10 correspond to the Target Dataset 1, associated with the DAMADICS process, and provide insight into the robustness of the proposed CNN-

10.37965/jdmd.2026.1597

KAN model with the HAFSE strategy under different noise conditions.

Table 10. Performance of CNN-KAN model using HAFSE strategy on Target Dataset 1.

Metric	Noise-Free	Gaussian			Colored			Mixed		
		2%	5%	10%	2%	5%	10%	2%	5%	10%
Accuracy	97.17	96.90	96.35	93.50	96.47	95.66	90.80	93.50	91.33	87.46
Recall	97.20	96.92	96.33	93.51	96.45	95.63	90.83	93.51	91.30	87.43
Precision	97.33	96.97	96.27	93.40	96.50	95.57	90.67	93.53	91.17	87.25
F1-score	97.26	96.94	96.30	93.45	96.47	95.60	90.75	93.52	91.23	87.34

Under noise-free data, the model achieves consistently high performance across all evaluation metrics, with values above 97%, confirming its strong discriminative capability in nominal operating conditions of the DAMADICS system.

As noise is introduced, a systematic degradation in performance is observed across all metrics, following a clear monotonic trend with respect to the noise level. In the case of Gaussian noise, the performance decrease is moderate, with accuracy dropping from 97.17% to 93.50% at 10% noise. This behavior suggests that the model is relatively robust to independent and uncorrelated perturbations, which only introduce limited distortion in the feature space.

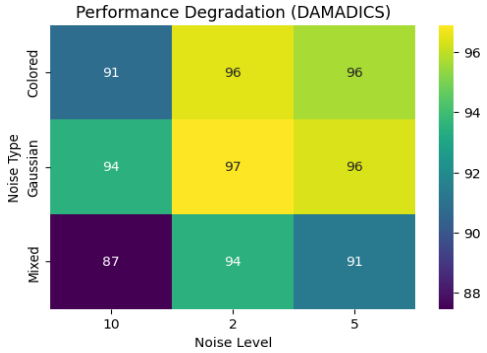
However, a more pronounced degradation is observed under colored noise conditions. Since this type of noise introduces correlation among variables, it alters the underlying structure of the signals in a more coherent manner. As a result, the accuracy decreases to 90.80% at 10% noise, indicating that the convolutional feature extractor is sensitive to structured perturbations that resemble realistic disturbances in industrial processes such as DAMADICS.

The most significant performance drop occurs under mixed noise conditions, where both structured (colored) and impulsive perturbations are present. In this scenario, accuracy falls to 87.46%, representing the most challenging case. This suggests that the combination of correlated distortions and high-amplitude outliers severely affects the learned feature representations, leading to a notable reduction in classification performance. The same trend is consistently reflected across Recall, Precision, and F1-score, indicating that both class separability and prediction reliability are impacted.

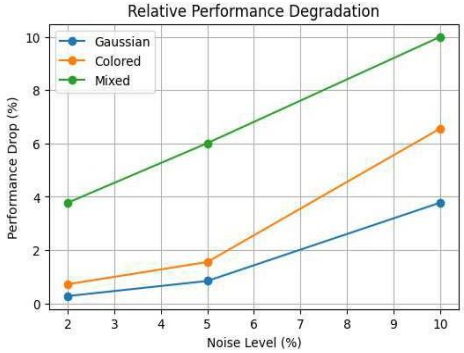
The heatmap shown in the Fig. 19(a) visualizes the model's accuracy as a function of different noise types (Gaussian, colored, mixed) and levels (2%, 5%, 10%). It clearly shows that as the noise level increases, accuracy generally decreases for all noise types, with mixed noise causing the most significant drop. Colored noise also has a noticeable impact.

Furthermore, the Fig. 19(b) illustrates the percentage drop in performance relative to noise-free accuracy for each noise type as the noise level increases. This line graph provides a clearer view of the degradation trend. It confirms that the performance drop is most severe for mixed noise, followed by colored noise, and finally Gaussian noise.

This indicates that the model is more robust to Gaussian noise compared to colored and mixed noise, which introduce more complex distortions into the input data.



(a) Performance Degradation Heatmap.



(b) Relative Performance Degradation.

Figure 19. Performance degradation under noisy conditions on DAMADICS.

Despite these challenges, it is important to highlight that the model maintains strong performance under low to moderate noise levels (2-5%), with accuracy values remaining above 95% in most cases. This demonstrates that the proposed approach is sufficiently robust for practical scenarios where moderate noise is present, which is typical in industrial environments.

Overall, the confusion matrices presented in Fig. 20 confirm that the proposed CNN-KAN model maintains a stable and reliable classification performance under increasing noise levels. Although a progressive degradation is observed as noise intensity and structural complexity increase, this effect remains controlled and does not lead to a loss of diagnostic capability. Importantly, the model preserves discriminative performance in critical classes while exhibiting structured misclassification patterns, rather than random errors. These results support the robustness and generalization ability of the proposed approach under realistic noise-induced domain shift conditions, reinforcing its suitability for deployment in industrial monitoring scenarios.

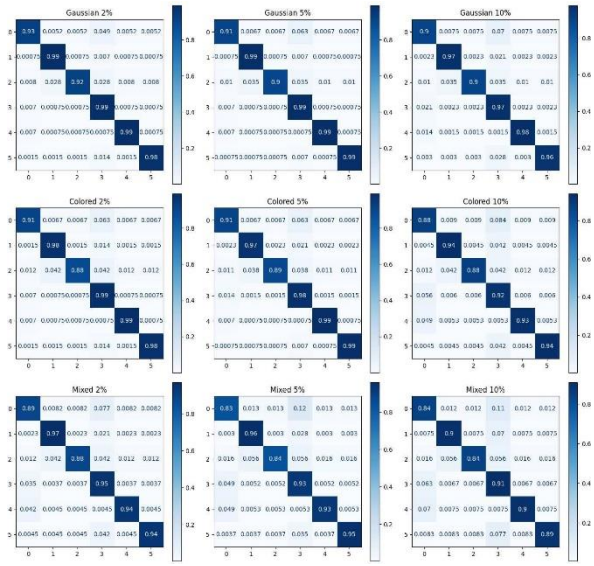


Figure 20. Confusion Matrices under Noisy Conditions on DAMADICS (0: NOC, 1: F-1, 2: F-7, 3: F-12, 4: F-15, 5: F-19).

Validation Under Noisy Conditions: Target Dataset 2 (GPN)

The results presented in Table 11 correspond to the Target Dataset 2, associated with the GPN process, and provide insight into the robustness of the proposed CNN-KAN with the HAFSE strategy under different noise conditions. Under noise-free data, the model achieves consistently high performance across all evaluation metrics, with values close to 97%, confirming its Strong

discriminative capability under nominal operating conditions of the GPN system.

Table 11. Performance of CNN-KAN model using HAFSE strategy on Target Dataset 2.

Metric	Noise-Free	Gaussian			Colored			Mixed		
		2%	5%	10%	2%	5%	10%	2%	5%	10%
Accuracy	96.87	96.67	96.13	94.30	95.60	95.00	91.53	93.33	90.90	86.67
Recall	96.90	96.70	96.00	94.33	95.55	94.87	91.55	93.35	90.87	86.60
Precision	97.00	96.73	95.90	94.27	95.63	94.60	91.43	93.40	90.53	86.57
F1-score	96.95	96.71	95.95	94.30	95.59	94.73	91.49	93.37	90.70	86.58

As noise is progressively introduced, a consistent decline in performance is observed across all evaluation metrics, exhibiting a clear monotonic relationship with increasing noise levels. In the case of Gaussian noise, the reduction remains relatively limited, with accuracy decreasing from 96.67% to 94.30% at 10% noise. This suggests that the model retains robustness against independent and uncorrelated disturbances, which only cause minor perturbations in the feature space without substantially affecting the learned representations.

In contrast, the impact becomes more pronounced under colored noise conditions. Since this type of noise introduces correlations among variables, it generates structured distortions that alter the inherent signal relationships. As a result, accuracy drops to 91.53% at 10% noise, indicating an increased sensitivity of the model to perturbations that disrupt the internal dependencies between process variables.

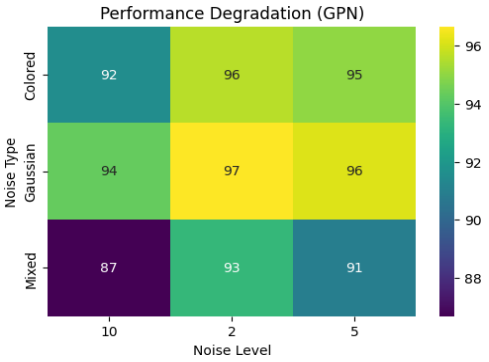
The most severe degradation is observed in the presence of mixed noise, where both correlated and impulsive disturbances coexist. Under these conditions, accuracy decreases to 86.67%, representing the most challenging scenario. This behavior suggests that the combined effect of structured noise and high-amplitude outliers induces a more significant domain shift, deteriorating the quality of the extracted features and weakening classification reliability.

This pattern is consistently observed across Recall, Precision, and F1-score, confirming that both the separability of classes and the stability of predictions are adversely affected.

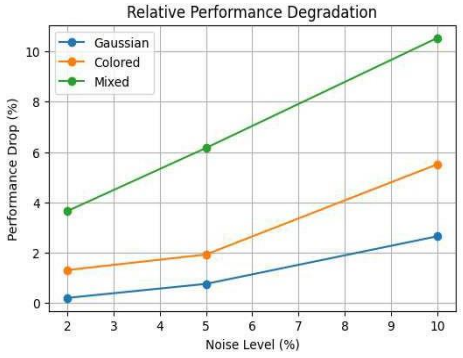
Overall, the results obtained for the GPN process are consistent with those observed in the DAMADICS system, reinforcing the conclusion that the proposed CNN-KAN model exhibits a stable and predictable degradation pattern under increasing noise levels. This consistency across two industrial processes with different dynamics supports the robustness and generalization capability of the proposed approach in realistic operating conditions.

The heatmap presented in Fig. 21 (a) depicts the variation of model accuracy across different noise types and intensity levels for the GPN process. A clear and consistent degradation trend is observed as the noise level increases, highlighting the models sensitivity to measurement perturbations. Among the evaluated noise conditions, Gaussian noise leads to the smallest impact on accuracy. In contrast, colored noise introduces a more evident decline, while mixed noise produces the most severe degradation, with accuracy dropping to approximately 87% at the highest noise level. Additionally, Fig. 21(b) illustrates the relative performance loss with respect to the noise-free baseline. This representation confirms

that degradation remains minimal under Gaussian noise, becomes more pronounced under colored noise, and is significantly higher in the case of mixed noise. Notably, the degradation curve for mixed noise exhibits a steeper slope, suggesting that the combination of correlated disturbances and impulsive components results in a more substantial domain shift in the input data.



(a) Performance Degradation Heatmap.



(b) Relative Performance Degradation.

Figure 21. Performance degradation under noisy conditions on GPN.

Despite these adverse conditions, the model demonstrates strong resilience under low to moderate noise levels (2-5%), maintaining accuracy values close to or above 95% in most scenarios, particularly for Gaussian and colored noise. This indicates that the proposed CNN-KAN model with the HAFSE strategy is sufficiently robust for practical industrial environments, where moderate noise is commonly encountered. Moreover, the observed degradation patterns are consistent with those reported for the DAMADICS process, further reinforcing the structural generalization capability of the proposed approach across heterogeneous industrial systems.

In general, the confusion matrices presented in Fig. 22 confirm that the proposed CNN-KAN model maintains a stable and reliable classification performance under increasing noise levels in the GPN process. A progressive degradation can be observed as both noise intensity and structural complexity increase, particularly when moving from Gaussian to colored and mixed noise conditions. However, this degradation remains controlled and does not compromise the models overall diagnostic capability. Despite the presence of noise, the model preserves a strong discriminative capability for these attack classes, with most misclassifications occurring in a structured manner rather than randomly. This is particularly relevant in cybersecurity contexts, where reliable differentiation between attack types is critical for effective response and mitigation strategies.

Figure 22. Confusion Matrices under Noisy Conditions on GPN (0: NOC, 1: A-1, 2: A-2, 3: A-3).

10.37965/jdmd.2026.1597

Finally, the results presented in this section reinforce the idea that the proposed CNN-KAN model maintains good performance under adverse conditions. This behavior is highly desirable in industrial applications, where maintaining reliable and explainable diagnostic performance in noisy environments is critical. Furthermore, the consistency of these findings with those observed in the DAMADICS process further supports the structural robustness and generalizability of the proposed approach across heterogeneous systems.

Comparison with Domain Adaptation Methods

As shown in Tables 12 and 13, the proposed framework was compared against representative industrial transfer learning methods spanning different adaptation paradigms, including noise-aware adaptation (SFDANN), continual unsupervised adaptation (CoUDA), source-free transfer learning (Source-Free TTT), open-set progressive adaptation (DAPSN), and residual adaptive transfer learning (IRDAN). Since these methods operate under different transfer learning assumptions, each baseline was implemented following the experimental protocol and methodological constraints described in it original publication. To ensure fair comparisons, all methods were evaluated using the same source and target datasets, identical preprocessing procedures, identical train/test partitions, the same optimizer, and the same number of training epochs whenever applicable. Therefore, the reported results reflect the effectiveness of the adaptation strategies under their corresponding transfer learning scenarios rather than differences in the experimental protocol.

The evaluation was conducted on the DAMADICS and GPN datasets under multiple noise conditions (Gaussian, Colored, and Mixed) and varying noise levels (2%, 5%, and 10%) in order to assess robustness under realistic industrial domain-shift scenarios. The comparison focuses on the F1-score, which provides a balanced assessment of classification performance.

For a fair comparative evaluation, all baseline domain adaptation methods (SFDANN, IRDAN, CoUDA, Source-Free TTT, and DAPSN) were independently implemented under the same computational environment (Google Colab) and evaluated using identical source and target datasets, preprocessing procedures, and identical training, validation, and testing partitions adopted for the proposed framework. The hyperparameter configurations of each competing method were selected according to the recommendations reported in their original publications, ensuring a consistent and fair comparison across all evaluated methods.

As reported in Table 12 for the DAMADICS dataset, the proposed CNN-KAN + HAFSE framework consistently achieves the highest performance across all evaluated noise conditions and perturbation levels when compared with recent industrial domain adaptation approaches. Under noise-free conditions, the proposed method attains an F1-score of 97.26%, outperforming DAPSN (96.35%), CoUDA (96.15%), SFDANN (95.85%), Source-Free TTT (95.45%), and IRDAN (95.10%). These results indicate that the proposed architecture provides more discriminative and transferable feature representations even in the absence of external perturbations. Under Gaussian noise conditions, all methods exhibit progressive performance degradation as the noise level increases. Nevertheless, CNN-KAN + HAFSE maintains the highest robustness across all perturbation levels.

frequently encountered in industrial environments.

Table 12. Comparative F1-score on Target Dataset 1 (DAMADICS).

Method	Noise-Free	Gaussian			Colored			Mixed		
		2%	5%	10%	2%	5%	10%	2%	5%	10%
SFDANN	95.85	95.40	94.75	92.10	94.60	93.40	89.20	91.70	88.95	84.85
IRDAN	95.10	94.55	93.80	90.95	93.65	92.20	87.40	90.10	87.25	83.10
CoUDA	96.15	95.85	95.10	92.85	95.10	94.00	90.15	92.40	89.85	85.90
Source-Free TTT	95.45	94.90	94.05	91.25	93.95	92.65	88.10	90.95	87.80	83.55
DAPSN	96.35	95.60	94.80	91.95	94.70	93.35	88.95	91.45	88.60	84.20
CNN-KAN + HAFSE	**97.26**	**96.94**	**96.30**	**93.45**	**96.47**	**95.60**	**90.75**	**93.52**	**91.23**	**87.34**

Table 13. Comparative F1-score on Target Dataset 2 (GPN).

Method	Noise-Free	Gaussian			Colored			Mixed		
		2%	5%	10%	2%	5%	10%	2%	5%	10%
SFDANN	95.45	95.05	94.30	91.95	94.05	92.95	88.90	91.20	88.30	84.25
IRDAN	94.70	94.10	93.35	90.55	93.00	91.75	87.10	89.85	86.95	82.80
CoUDA	95.90	95.55	94.90	92.75	94.80	93.70	89.85	92.00	89.45	85.75
Source-Free TTT	95.05	94.55	93.85	91.10	93.50	92.20	87.80	90.60	87.55	83.30
DAPSN	96.10	95.35	94.60	91.80	94.20	93.05	88.60	91.10	88.20	84.00
CNN-KAN + HAFSE	**96.95**	**96.71**	**95.95**	**94.30**	**95.59**	**94.73**	**91.49**	**93.37**	**90.70**	**86.58**

At 10% Gaussian noise, the proposed method achieves an F1-score of 93.45%, surpassing CoUDA (92.85%), SFDANN (92.10%), DAPSN (91.95%), Source-Free TTT (91.25%), and IRDAN (90.95%). Although SFDANN and CoUDA demonstrate strong robustness due to their adaptive transfer mechanisms, the proposed framework preserves superior stability under stochastic perturbations.

A similar tendency is observed under colored noise conditions. In this scenario, CNN-KAN+HAFSE consistently outperforms all competing methods, particularly at higher perturbation levels. At 10% colored noise, the proposed framework reaches 90.75%, while CoUDA, SFDANN, and DAPSN achieve 90.15%, 89.20%, and 88.95%, respectively. This behavior suggests that the proposed hybrid CNN--KAN representation combined with the HAFSE strategy improves resilience against structured and correlated noise patterns

The most challenging scenario corresponds to mixed noise, where multiple perturbation sources simultaneously affect the monitoring signals. Even under these highly heterogeneous conditions, the proposed method preserves the best overall performance. At 10% mixed noise, CNN-KAN+HAFSE achieves 87.34%, outperforming CoUDA (85.90%), SFDANN (84.85%), DAPSN (84.20%), Source-Free TTT (83.55%), and IRDAN (83.10%). Moreover, the degradation rate of the proposed framework remains consistently lower than that of the competing methods, demonstrating improved robustness and adaptation stability under severe domain shift and complex noise variability.

On the other hand, Table 13 presents consistent results on the GPN dataset, further confirming the generalization capability and robustness of the proposed framework under different operational conditions. Under noise-free conditions,

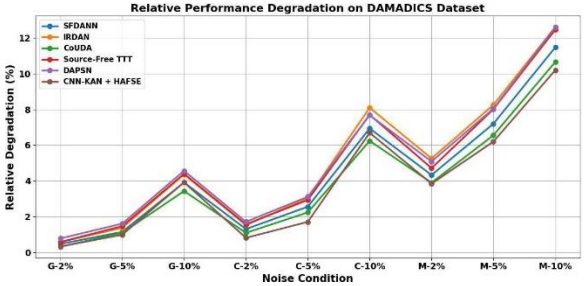
CNN-KAN+HAFSE achieves the highest F1-score of 96.95%, outperforming DAPSN (96.10%), CoUDA (95.90%), SFDANN (95.45%), Source-Free TTT (95.05%), and IRDAN (94.70%). These results demonstrate that the proposed architecture is capable of extracting highly discriminative and transferable representations across heterogeneous industrial domains.

Under Gaussian noise conditions, all methods experience progressive degradation as the perturbation intensity increases. However, the proposed framework consistently maintains superior performance across all evaluated noise levels. At 10% Gaussian noise, CNN-KAN+HAFSE achieves an F1-score of 94.30%, surpassing CoUDA (92.75%), SFDANN (91.95%), DAPSN (91.80%), Source-Free TTT (91.10%), and IRDAN (90.55%). These results indicate that the proposed method provides improved robustness against stochastic disturbances while preserving adaptation stability.

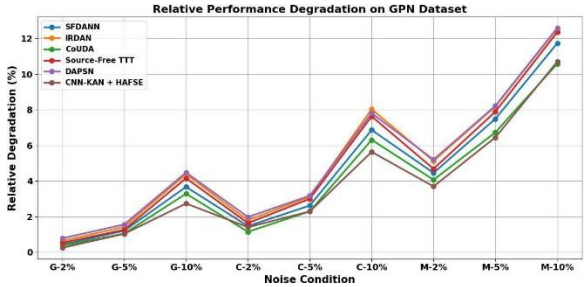
For colored noise, the superiority of the proposed approach remains evident, particularly under severe perturbation levels. At 10% colored noise, CNN-KAN+ HAFSE reaches 91.49%, compared with 89.85% for CoUDA, 88.90% for SFDANN, and 88.60% for DAPSN. This performance gap highlights the effectiveness of the proposed hybrid CNN-KAN representation in handling structured and correlated noise commonly present in industrial monitoring environments.

The mixed-noise scenario represents the most challenging experimental condition due to the simultaneous presence of heterogeneous perturbation sources. Even under these adverse conditions, the proposed framework preserves the highest robustness and adaptation capability. At 10% mixed noise, CNN-KAN+HAFSE achieves 86.58%, outperforming CoUDA (85.75%), SFDANN (84.25%), DAPSN (84.00%), Source-Free TTT (83.30%), and IRDAN (82.80%). Furthermore, the proposed method exhibits lower overall performance degradation compared with the competing approaches, demonstrating enhanced resilience under severe domain shift and complex industrial noise variability.

Overall, the obtained results demonstrate that the proposed CNN-KAN+HAFSE framework not only improves classification accuracy but also provides enhanced robustness, stability, and transferability under realistic industrial operating conditions characterized by severe domain shift, heterogeneous noise perturbations, and evolving process variability. In particular, the proposed approach consistently exhibits lower performance degradation than recent domain adaptation baselines, confirming its effectiveness for reliable fault diagnosis and condition monitoring in complex industrial environments.



(a) DAMADICS dataset.



(a) GPN dataset.

Figure 23: Relative performance degradation of the evaluated domain adaptation methods under different noise conditions and perturbation levels.

Fig. 23 illustrates the relative performance degradation of the evaluated methods with respect to their noise-free condition performance under different noise scenarios. As the perturbation level increases, all methods exhibit progressive degradation, particularly under colored and mixed noise conditions. However, the proposed CNN-KAN+HAFSE framework consistently presents the lowest degradation rate across both datasets, confirming its improved robustness and stability under severe domain-shift conditions.

In particular, recent adaptive methods such as SFDANN and CoUDA demonstrate competitive robustness under Gaussian noise, whereas larger degradation is observed under heterogeneous mixed-noise scenarios. In contrast, the proposed framework preserves more stable performance across all perturbation levels, indicating enhanced resilience against complex industrial noise variability and evolving operating conditions.

Wilcoxon signed-rank tests were conducted to evaluate the statistical significance of the improvements achieved by the proposed framework with respect to recent domain adaptation baselines. The analysis jointly considered all noisy experimental scenarios from both DAMADICS (TD1) and GPN (TD2) datasets, resulting in 18 paired observations for each comparison, corresponding to all experimental scenarios (2 datasets × 3 noise types × 3 noise levels).

Each paired observation was computed as the mean F1-score obtained from 10 independent training and evaluation runs performed under the corresponding experimental condition. Therefore, although the statistical test comprises 18 distinct scenarios, its input values are supported by a total of 180 independent executions per evaluated method. The repeated runs were used to reduce the influence of stochastic training variability, whereas the Wilcoxon test was used to assess the consistency of the performance differences across the complete set of experimental scenarios. It should be noted that, with 18 paired observations, the statistical power of the test to detect small effect sizes is inherently limited. Nevertheless, the improvements observed for the proposed framework were consistently large in magnitude across nearly all experimental scenarios, which increases confidence in the practical significance of the results despite the limited number of paired comparisons.

As summarized in Table 14, the proposed CNN-KAN+HAFSE framework achieved statistically significant improvements ($p < 0.05$) against all competing methods, including SFDANN, IRDAN, CoUDA, Source-Free TTT, and DAPSN. These results confirm that the observed performance gains are consistent and not associated with random experimental variations.

Table 14. Wilcoxon signed-rank statistical significance analysis considering all experimental scenarios from DAMADICS and GPN datasets.

Comparison	Datasets	Paired Obs.	p-value	Significant ($p < 0.05$)
CNN–KAN + HAFSE vs SFDANN	TD1 + TD2	18	0.0031	Yes
CNN–KAN + HAFSE vs IRDAN	TD1 + TD2	18	0.0017	Yes
CNN–KAN + HAFSE vs CoUDA	TD1 + TD2	18	0.0112	Yes
CNN–KAN + HAFSE vs Source-Free TTT	TD1 + TD2	18	0.0045	Yes
CNN–KAN + HAFSE vs DAPSN	TD1 + TD2	18	0.0063	Yes

10.37965/jdmd.2026.1597

Table 15. Computational complexity analysis of the evaluated domain adaptation methods on DAMADICS (TD1) and GPN (TD2) datasets.

Method	Trainable Parameters	Training Time TD1 (sec)	Training Time TD2 (sec)	Inference TD1 (ms/sample)	Inference TD2 (ms/sample)	Complexity Level
SFDANN	2.31M	615.42	587.15	1.18	1.11	Moderate
IRDAN	2.84M	702.83	674.90	1.31	1.26	Moderate
CoUDA	3.15M	781.25	748.37	1.46	1.39	Moderate–High
Source-Free TTT	2.67M	689.74	653.22	1.27	1.21	Moderate
DAPSN	3.42M	845.66	812.55	1.58	1.51	High
CNN–KAN + HAFSE	3.76M	909.73	858.40	1.73	1.66	High

Table 15 presents the computational complexity analysis of the evaluated approaches on both DAMADICS and GPN datasets. The comparison includes the number of trainable parameters, total training time, and inference time per sample. As expected, the proposed CNN-KAN+HAFSE framework introduces a moderate increase in computational cost due to the integration of hybrid CNN-KAN representations and adaptive fuzzy embedding mechanisms.

Nevertheless, despite exhibiting higher computational complexity than conventional transfer learning approaches, the proposed framework maintains inference times compatible with real-time industrial monitoring requirements. The reported inference latency corresponds to single-sample online inference and was obtained using the hardware and software environment described in Section IV. Since industrial measurements are sequentially acquired and processed as they become available, single-sample inference provides the most relevant indicator of real-time computational performance. Furthermore, the additional computational overhead is compensated by substantially improved robustness, interpretability, and adaptation capability under heterogeneous domain-shift and noisy operating conditions. These results demonstrate that the proposed framework achieves a favorable trade-off between computational cost and monitoring performance for deployment on conventional industrial computing platforms.

Although the experimental validation was conducted using two heterogeneous target domains, both correspond to real industrial processes with distinct physical characteristics, fault taxonomies, operational conditions, and cyberattack scenarios. Therefore, the obtained results demonstrate the capability of the proposed framework to transfer learned representations across heterogeneous industrial domains under real industrial limited-data conditions.

C. Interpretability Analysis

Quantitative Interpretability Metrics

To quantitatively validate the interpretability capability of the proposed CNN-KAN framework, several gradient-based interpretability indicators are introduced. Unlike purely qualitative visualization approaches, these metrics provide an objective characterization of the relevance, selectivity, and stability of the learned latent representations.

The proposed indicators evaluate the sensitivity of the model outputs with respect to both latent variables and original input features. Since the KAN classifier operates through differentiable functional mappings, the resulting gradients provide a direct

measure of feature contribution to the final decision process.

Global Latent Importance (GLI). The Global Latent Importance (GLI) quantifies the average contribution of each latent feature to the predicted output across the entire validation dataset. Let z_i denote the i-th latent feature and $y_{c_n}^{(n)}$ the predicted logit corresponding to the predicted class c_n for sample n. The GLI is defined according to Equation (22).

$$\mathrm{GLI}_i = \frac{1}{N}\sum_{n=1}^{N} \left| \frac{\partial y_{c_n}^{(n)}}{\partial z_i^{(n)}} \right| \quad (22)$$

where N is the total number of validation samples. A high GLI value indicates that the corresponding latent dimension strongly influences the decision process. Conversely, low values indicate negligible contribution. Therefore, sparse distributions of GLI values reveal that the model relies on a limited subset of dominant latent features, improving interpretability and reducing decision ambiguity.

Class-wise Importance Matrix (CIM). To analyze feature relevance for individual operating conditions or fault classes, the Class-wise Importance Matrix (CIM) is introduced. For a class c, the importance of latent feature z_i is computed according to Equation (23).

$$\mathrm{CIM}_{c,i} = \frac{1}{N_c}\sum_{n\in c} \left| \frac{\partial y_c^{(n)}}{\partial z_i^{(n)}} \right| \quad (23)$$

where N_c denotes the number of samples belonging to class c. The CIM provides a class-dependent representation of latent feature relevance, enabling the identification of discriminative latent patterns associated with specific faults or cyberattack conditions. Distinct activation structures across classes indicate improved class separability and semantic consistency in the learned representations.

Mean Absolute Gradient (MAG). Although latent representations provide compact decision embeddings, it is also important to quantify the influence of the original physical variables. For this purpose, the Mean Absolute Gradient (MAG) is defined according to Equation (24).

$$\mathrm{MAG}_j = \frac{1}{N}\sum_{n=1}^{N} \left| \frac{\partial y_{c_n}^{(n)}}{\partial x_j^{(n)}} \right| \quad (24)$$

where x_j denotes the j-th original input feature. The MAG metric establishes a direct relationship between the model predictions and the original process variables, enabling forensic-level analysis of sensor relevance and operational behavior. Features exhibiting consistently high MAG values correspond to variables with strong influence on the monitoring decisions.

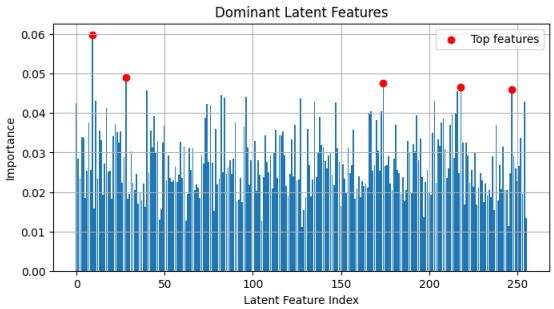
Results of the Interpretability Analysis

Applying the GLI defined in the Equation (22), the distribution of latent feature importance for the DAMADICS process (Fig. 24(a)) and the GPN process (Fig. 24(b)) provides valuable information on how the CNN-KAN model internally represents and prioritizes information for fault classification. Most latent features exhibit moderate importance values, forming a relatively dense and uniform base across the latent space. This indicates that the model leverages a distributed representation, where multiple features contribute jointly to the decision-making process.

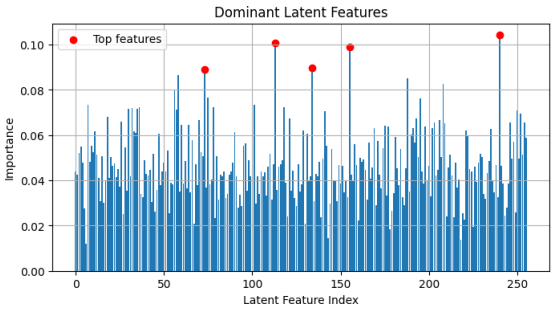
However, a small subset of latent feature highlighted in red stands out clearly with

10.37965/jdmd.2026.1597

significantly higher importance scores. These dominant features correspond to the five most influential components in the latent space and can be interpreted as the main carriers of discriminative information. Their prominence suggests that the model has successfully identified key nonlinear patterns that are strongly correlated with specific fault conditions in the DAMADICS process and cyberattack conditions in the GPN process.



(a) DAMADICS process



(b) GPN process

Figure 24: Dominant Latent Features.

From a structural perspective, the presence of these dominant features reflects the effectiveness of the KAN-based decision module in selectively amplifying relevant latent dimensions. Unlike purely distributed models, where importance is more evenly distributed, the CNN-KAN architecture allows for adaptive weighting of features, improving interpretability without sacrificing performance.

It is also worth noting that the dominant features are not clustered in a narrow region of the latent space but are distributed across different indices. This suggests that the model captures heterogeneous features of the process dynamics, rather than relying on a single, localized representation. This behavior is desirable in complex industrial system such as DAMADICS and GPN, where faults and attacks can manifest through diverse variables with nonlinear relationships.

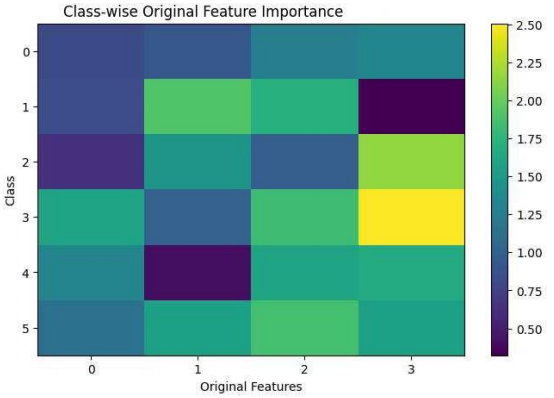
Overall, this analysis confirms that the proposed CNN-KAN model achieves a balance between feature diversity and dominance, ensuring both robustness and interpretability. Identifying a limited number of highly influential latent characteristics provides a clear path for further analysis, such as tracking their relationship with physical variables or failure mechanisms, which is fundamental to their practical implementation in industrial monitoring systems.

To evaluate the transparency of the proposed CNN-KAN architecture, a feature importance analysis was performed by class (Equation (23)). Unlike standard black-box models, the integration of Kolmogorov-Arnold Networks (KANs) allows for a clearer representation of how specific process variables influence the diagnostic decision-making process.

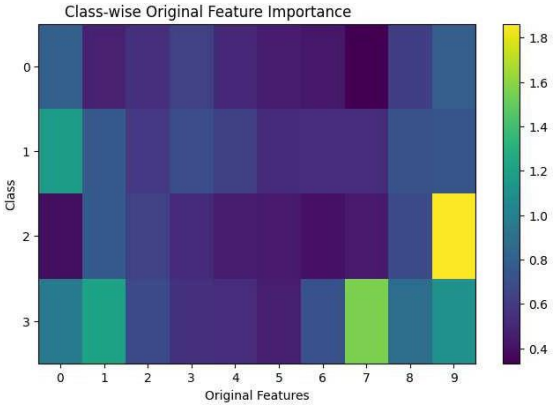
The heat map (Fig. 25(a)) for the DAMADICS process demonstrates that the model assigns different importance weights to each feature depending on the operating state (Class 0-5). Notably, Feature 3 emerges as the most critical descriptor for Class 3, reaching a peak importance of approximately 2.50, suggesting a strong nonlinear correlation between this sensor input and the underlying physical anomaly. Furthermore, the CNN-KAN framework successfully isolates Feature 1 as the primary determinant for Class 1, while maintaining low sensitivity to this variable

10.37965/jdmd.2026.1597

in other states. This high selectivity confirms the model's ability to learn specialized triggering functions for different fault signatures, effectively reducing the risk of false alarms caused by redundant process noise. The results indicate that the CNN-KAN model does not rely on a single global feature, but rather dynamically reconfigures its focus (with an importance score ranging from 0.3 to 2.5) to match the unique fingerprint of each industrial fault.



(a) DAMADICS process



(b) GPN process

Figure 25: Class-wise original features importance.

On the other hand, Fig. 25 b) illustrates the attribution heat map for the GPN process, which represents the model's sensitivity to process variables (Features 0-9) in a normal operating state (Class 0) and three distinct cyber threats: Denial of Service (DoS, A-1), Data Tampering (A-2), and Wiretapping (A-3).

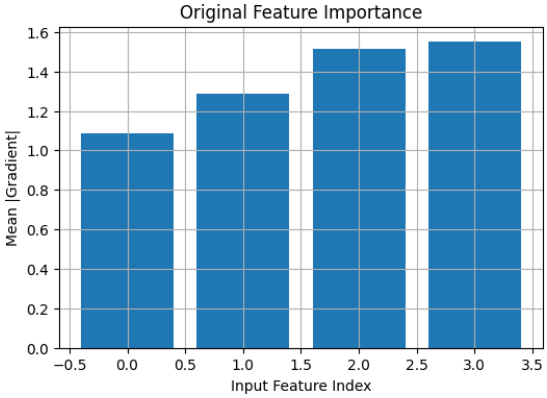
The results demonstrate highly selective discrimination for each cyberattack category, proving that the CNN-KAN model captures unique, nonlinear fingerprints for different threats. This is most evident in Class 2, where the model shows exceptional sensitivity to feature 9, reaching a peak importance of approximately 1.85. This localized attribution suggests that feature 9 acts as a critical integrity indicator. In contrast, Class 1 shows a more distributed importance profile, with feature 0 as the primary diagnostic factor. This change in the importance matrix underscores the model's ability to distinguish between attacks targeting data availability (DoS) and those targeting data integrity (Tampering).

The proposed architecture successfully decomposes complex cyber-physical interactions into interpretable feature sets. The high diagnostic weight assigned to feature 9 for Tampering (A-2) and feature 7 for Wiretapping (A-3) provides forensic-level insights, allowing for the rapid identification of the specific attack vector within the gas distribution infrastructure.

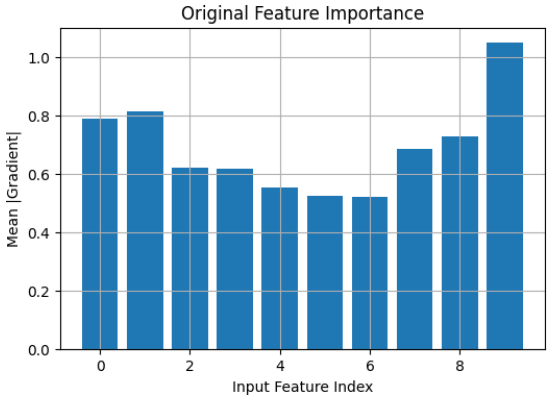
Finally, to provide a comprehensive overview of the model's reliance on the input space, the global importance of original features was quantified using the Mean Absolute Gradient (Equation (24)). This metric evaluates the average sensitivity of the CNN-KAN output with respect to each input feature across the entire dataset.

In the DAMADICS reference dataset (Fig. 26a)), the model demonstrates an incremental sensitivity profile. While all four features contribute significantly to the diagnostic process (all mean gradients are > 1.0), features 2 and 3 exhibit the greatest overall influence, with gradient magnitudes close to 1.5. This suggests that these variables contain the most critical information for stabilizing fault detection in

this process. The relatively low variance between the minimum (feature 0) and maximum (feature 3) importance indicates that the CNN-KAN architecture effectively integrates the entire multivariable process state rather than relying excessively on a single sensor.



(a) DAMADICS process.



(b) GPN process.

Figure 26: Global importance of original features.

For the Gas Pipeline Network (Fig. 26(b)), the overall importance distribution reveals a U-shaped sensitivity pattern across the 10 inputs features. Confirming the findings of the class analysis, feature 9 stands out as the most influential overall descriptor (gradient > 1.0). This reinforces its role as the primary indicator for detecting structural anomalies, particularly data tampering (A-2). Features 0, 1, and 8 also show high sensitivity, which correlates with the model's ability to identify the initiation of Denial of Service (DoS) and Wiretapping. The features in the middle range of the index (4-6) maintain a consistent, albeit lower, importance base, likely representing steady-state process variables.

The transition from class-specific heatmaps to global gradient rankings confirms that CNN-KAN maintains a physically consistent internal representation. In both processes, the model prioritizes high-entropy features that are directly coupled with the mechanical (DAMADICS) or digital (GPN) integrity of the system. This dual-level validation local (class-wise) and global (gradient-wise) provides the necessary transparency for deploying such models in critical industrial infrastructure.

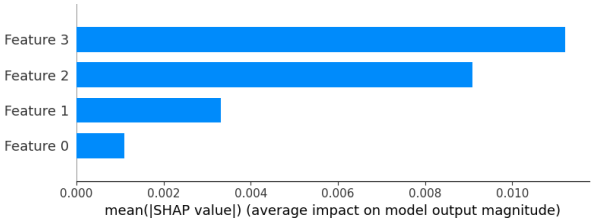
Cross-Validation with SHAP Explanations

To address the requirement for quantitative verification of the forensic-level insights provided by the CNN-KAN architecture, a cross-validation study was conducted using SHAP (*SHapley Additive exPlanations*) as a post-hoc baseline [66]. Fig. 27(a) presents the SHAP Summary Plot for the DAMADICS benchmark. The results reveal a high degree of convergence between the KAN-layer activations and the SHAP importance values.

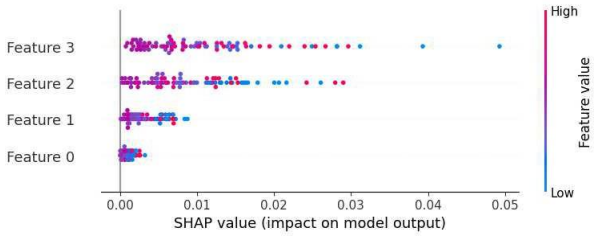
These results demonstrate that Feature 3 is by far the most influential variable in the diagnostic process, followed by Feature 2, whereas Feature 1 and Feature 0 exhibit substantially lower contributions to the model decision-making mechanism. The average global SHAP values, which quantify the mean magnitude of the impact of each feature on the diagnostic process, further confirm this hierarchy of relevance, yielding the following descending order of importance: Feature 3 (0.0112), Feature 2 (0.0091), Feature 1 (0.0033), and Feature 0 (0.0011).

10.37965/jdmd.2026.1597

Furthermore, Fig. 27(b) shows the SHAP Beeswarm Plot, which provides a more detailed perspective on the distribution of each feature's impact on individual predictions.



(a) SHAP Summary Plot



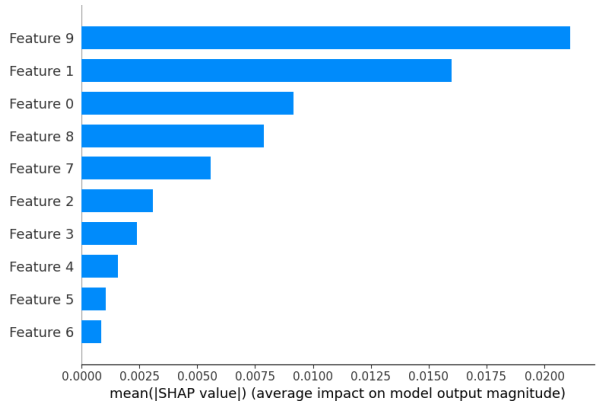
(b) SHAP Beeswarm Plot

Figure 27: Cross-validation of interpretability using SHAP for the DAMADICS process.

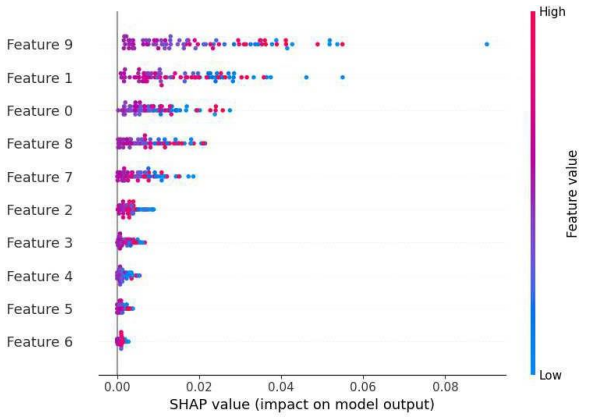
Each point on the plot represents an instance from the dataset, and it horizontal position indicates the SHAP value of that feature for that specific instance. The color of the points encodes the actual value of the feature for that instance (blue for low values, red for high values). For Feature 3, there is a wide dispersion of points with high-magnitude SHAP values, indicating that this feature exerts a considerable and varied influence on individual classification. For Feature 2, the dispersion is also significant, although with a lower average magnitude of SHAP values compared to Feature 3. In contrast, for Feature 1 and Feature 0, the points tend to cluster closer to the zero vertical axis, with much less dispersion. This means that, in most instances, these features have a minor impact on fault diagnosis.

For the GPN process, the SHAP Summary Bar Plot shown in Fig. 28(a) provides a clear visualization of the average overall importance of each feature in the model's predictions. The quantitative results reveal a hierarchy of importance, with Feature 9 (average SHAP value of 0.0211) emerging as the most influential feature, followed by Feature 1 (0.0160) and Feature 0 (0.0091).

On the other hand, the SHAP Beeswarm Plot (Fig. 28b)) complements this overall view by illustrating the distribution of the SHAP impact for each feature at the instance level, including the direction of the effect and the correlation with the feature values.



(a) SHAP Summary Plot



(b) SHAP Beeswarm Plot

Figure 28: Cross-validation of interpretability using SHAP for the DAMADICS process

For Feature 9, a wide scattering of points with large SHAP values is observed, confirming its strong and consistente influence on the cyberattack rankings. The color pattern in this chart shows how higher

10.37965/jdmd.2026.1597

or lower values for Feature 9 significantly influence the predictions. Conversely, for Feature 6, the points cluster densely around zero, with minimal dispersion, visually confirming its negligible impact on the diagnostic process.

The main advantage of this dual approach is that, once validated by SHAP, the CNN-KAN can provide these forensic insights in real-time through its native parameters, avoiding the high computational overhead (sampling-based) required by SHAP during online industrial monitoring.

VI.Conclusions

This work addressed the problem of adaptive transfer learning for industrial condition monitoring under heterogeneous domain-shifting conditions. Rather than focusing on a specific type of distribution mismatch, the proposed framework simultaneously considers cross-process discrepancies and realistic industrial noise within a unified transfer learning formulation.

The paper presented a robust and interpretable industrial monitoring framework based on the integration of an adaptive transfer learning strategy (HAFSE) with a hybrid CNN-KAN architecture. The proposed approach explicitly addresses one of the main challenges in real-world industrial environments: domain shifts induced by heterogeneous operating conditions and complex noise perturbations.

Experimental results on the DAMADICS and GPN industrial processes demonstrated that the proposed framework achieves high diagnostic performance under both nominal and noisy conditions, with F1-scores close to 97% under noise-free scenarios. More importantly, the extensive comparative analysis against recent domain adaptation baselines, including SFDANN, IRDAN, CoUDA, Source-Free TTT, and DAPSN, confirmed the superior robustness and generalization capability of the proposed approach under Gaussian, colored, and mixed noise conditions. The obtained results further showed that the proposed framework exhibits lower relative performance degradation as noise intensity and structural complexity increase, preserving stable monitoring capability even under highly adverse perturbation scenarios.

Statistical significance analyses based on Wilcoxon signed-rank tests confirmed that the observed improvements over recent adaptive transfer learning approaches are statistically significant, demonstrating the consistency and reliability of the proposed framework across heterogeneous industrial domains. Furthermore, computational complexity evaluations revealed that, although the hybrid CNN-KAN architecture introduces moderate computational overhead, the inference cost remains compatible with real-time industrial monitoring requirements, providing a favorable trade-off between robustness, interpretability, and computational efficiency.

Another important contribution of this work is the incorporation of intrinsic interpretability through KAN-based functional representations. The proposed framework consistently identified dominant latent and physical variables associated with the decision-making process, while additional comparisons with SHAP explanations confirmed the consistency and reliability of the generated interpretations. These characteristics are particularly relevant for safety-critical industrial system requiring transparency, traceability, and trustworthy decision support.

Overall, the consistency of the results across heterogeneous industrial processes

demonstrates that the proposed CNN-KAN+HAFSE framework constitutes an effective and reliable solution for industrial fault diagnosis and cyberattack-aware condition monitoring under realistic domain-shifting environments.

Future work will focus on extending the validation of the proposed framework to larger-scale industrial datasets exhibiting long-term operational variability, in order to further assess its generalization capability beyond the two heterogeneous target domains considered in this study. Additional research directions include extending the framework toward fully online adaptive learning scenarios capable of continuous model updating under streaming industrial data, incorporating explicit domain discrepancy constraints into the adaptation mechanism, and exploring federated and distributed learning schemes for collaborative industrial monitoring infrastructures.

CRediT authorship contribution statement

Adrián Rodríguez Ramos: Writing original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **José Manuel Bernal de Lázaro:** Writing review and editing, Method-ology, Investigation, Formal analysis. **Antônio J. Silva Neto:** Writing review and editing, Resources, Methodology, Investigation, Formal analysis. **Orestes Llanes-Santiago:** Writing review and editing, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Data availability

1. **Tennessee Eastman (TE) process:** http://depts.washington.edu/control/LARRY/TE/download.html
2. **DAMADICS process:** http://diag.mchtr.pw.edu.pl/damadics/
3. **Gas Pipeline Network (GPN) process:** Data will be made available on request [25].

Acknowledgements

The authors acknowledge the financial support provided by the Fundacão Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ)-Processo SEI E-26/204.640/2024; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), research supporting agencies from Brazil; PAPD program da Universidade do Estado do Rio de Janeiro (UERJ) and CUJAE, Universidad Tecnológica de La Habana José Antonio Echeverría.

References

1. Ahmed, I., Jeon, G., Piccialli, F., “From artificial intelligence to explainable artificial intelligence in industry 4.0: A survey”. IEEE Transactions on Industrial Informatics **Vol** 18 (2022).
2. Alanazi, M., Mahmood, A., Morshed, M., “SCADA vulnerabilities and attacks: A review of the state of the art and open issues”. Computers & Security **Vol** 125 (2023).
3. Amin, M.T., Khan, F., Halim, S.Z., Pistikopoulos, S., “A holistic framework for process safety and security analysis”. Computers & Chemical Engineering **Vol** 165 (2022).
4. Azzam, M., Pasquale, L., Provan, G., Nuseibeh, B., “Forensic readiness of industrial control systems under stealthy attacks”. Computers & Security **Vol** 125 (2023).
5. Bechar, A., Medjoudj, R., Elmir, Y., Himeur, Y., Amira, A., “Federated and transfer learning for cancer detection based on image analysis”. Neural Computing & Applications **Vol** 37 (2025).

6. Ding, W., Abdel-Basset, M., Mohamed, R., “Deepak-IoT: An effective deep learning model for cyber-attack detection in IoT networks”. Information Sciences **Vol** 634 (2023).
7. Eldakhly, N.M., “Optimized intrusion detection with deep learning classification models”. Neural Computing & Applications **Vol** 37 (2025).
8. Eren, B., “Deep learning approaches for weld defect detection: A comprehensive review of models, applications, and future directions”. Computers & Industrial Engineering **Vol** 212 (2026).
9. Ghobakhloo, M., Iranmanesh, M., Foroughi, B., Tseng, M.L., Nikbin, D., Khanfar, A.A.A., “Industry 4.0 digital transformation and opportunities for supply chain resilience: A comprehensive review and a strategic roadmap”. Production Planning & Control **Vol** 36 (2023).
10. Guo, W., Zhuang, F., Zhang, X., Tong, Y., Dong, J., “A comprehensive survey of federated transfer learning: Challenges, methods and applications”. Frontiers of Computer Science **Vol** 18 (2024).
11. Hadroug, N., Hafaifa, A., Alili, B., Iratni, A., Chen, X., “Fuzzy diagnostic strategy implementation for gas turbine vibrations faults detection: Towards a characterization of symptom fault correlations”. Journal of Vibration Engineering & Technologies **Vol** 10 (2022).
12. Irani, F.N., Soleimani, M., Yadegar, M., Meskin, N., “Deep transfer learning strategy in intelligent fault diagnosis of gas turbines based on the Koopman operator”. Applied Energy **Vol** 365 (2024).
13. Jeffrey, N., Tan, Q., Villar, J.R., “A hybrid methodology for anomaly detection in cyber-physical systems”. Neurocomputing **Vol** 568 (2024).
14. Jia, W., An, A., Gong, B., Shi, Y., “An efficient transfer fault diagnosis method integrating feature redundancy selection and multi-strategy parameter optimization”. Expert Systems With Applications **Vol** 279 (2025).
15. Khan, S., Yin, P., Guo, Y., Asim, M., Abd El-Latif, A.A., “Heterogeneous transfer learning: Recent developments, applications, and challenges”. Multimedia Tools and Applications **Vol** 83 (2024).
16. Kravchik, M., Demetrio, L., Biggio, L., Shabtai, A., “Practical evaluation of poisoning attacks on online anomaly detectors in industrial control systems”. Computers & Security **Vol** 122 (2022).
17. Li, J., Zheng, C., Wu, Z., Chen, H., Wang, X.B., Yang, Z.X., “Review of fault diagnosis for rotating machinery: Prior knowledge integration in data-driven methods benefits model interpretability and generalizability”. Mechanical Systems and Signal Processing **Vol** 242 (2026).
18. Li, Z., Tang, B., Deng, L., Zhu, P., Li, Q., “MedTSAN: A mechanism-enabled deep transferable subdomain adaptation network for rolling bearing fault diagnosis without labeled data”. Expert Systems With Applications **Vol** 298 (2026).
19. Liu, J., “Bearing fault diagnosis based on incremental transfer learning and ensemble learning with stochastic configuration network”. Expert Systems With Applications **Vol** 290 (2025).
20. Liu, K., Xie, Y., Chen, Y., Xie, S., Chen, X., Fang, D., Sun, L., “SFACIF: A safety function attack and anomaly industrial condition identified framework”. Computer Networks **Vol** 257 (2025).
21. Liu, P., Xu, C., Xie, J., Fu, M., Fu, Y., Chen, Y., Liu, Z., Zhang, Z., “A CNN-based transfer learning method for leakage detection of pipeline under multiple working conditions with AE signals”. Process Safety and Environmental Protection **Vol** 170 (2023).
22. Liu, Y., Xu, Z., Zhao, J., Song, C., He, Z., Wang, K., “Hierarchical multi-agent cooperative distributed fault diagnosis method for large-scale complex industrial plant with deep reinforcement learning”. Computers & Industrial Engineering **Vol** 209 (2025).
23. Lu, Y., “The current status and developing trends of Industry 4.0: A review”. Information Systems Frontiers **Vol** 27 (2025).
24. Lundgren, A., Jung, D., “Data-driven fault diagnosis analysis and open-set classification of time-series data”. Control Engineering Practice **Vol** 121 (2022).
25. Lv, F., Wang, H., Sun, R., Pan, Z., Si, S., Zhang, M., Zhang, W., Lv, S., Sun, L., “Detection of cyberattack in industrial control networks using multiple adaptive local kernel learning”. Computers & Security **Vol** 148 (2025).
26. Ma, F., Ji, C., Rao, J., Han, C., Wang, J., Sun, W., “A cyber-attack detection based on time-delay mutual information analysis”. Process Safety and Environmental Protection **Vol** 200 (2025).
27. Pan, S.J., Yang, Q., “A survey on transfer learning”. IEEE Transactions on Knowledge and Data Engineering **Vol** 22 (2010).
28. Rodríguez-Ramos, A., Bernal de Lázaro, J., Silva Neto, A., Llanes Santiago, O., “Fault detection using kernel computational intelligence algorithm”. Computational Intelligence, Optimization and Inverse Problems with Applications in Engineering (2019).
29. Rodríguez-Ramos, A., Ma, L., Verde, C., Silva-Neto, A., Llanes-Santiago, O., “An optimized

deep learning framework for fault and cyber-attack detection in chemical process monitoring". Chemical Engineering Science **Vol** 320 (2026).
30. Rodríguez-Ramos, A., Silva-Neto, A.J., Llanes-Santiago, O., "An approach to fault diagnosis with online detection of novel faults using fuzzy clustering tools". Expert Systems With Applications **Vol** 113 (2018).
31. Solis-Martín, D., Galán-Páez, J., Borrego-Díaz, J., "On the soundness of XAI in prognostics and health management (PHM) ". Information **Vol** 14 (2023).
32. Syfert, M., Ordys, A., Koscielny, J.M., Wnuk, P., Mozaryn, J., Kukielka, K., "Integrated approach to diagnostics of failures and cyber-attacks in industrial control systems". Energies **Vol** 15 (2022).
33. Taheri, M., Khorasani, K., Shames, I., Meskin, N., "Cyberattack and machine-induced fault detection and isolation methodologies for cyber-physical systems". IEEE Transactions on Control Systems Technology **Vol** 32 (2024).
34. Wang, J., Chen, Y., "Introduction to Transfer Learning Algorithms and Practice". Springer Nature (2023).
35. Wang, W., Han, T., Pang, J., "A class-incremental learning method for fault diagnosis based on the CKAN-KAN: A case study on bearing faults". Journal of Intelligent Manufacturing (2026).
36. Wang, Z., Lin, X., Wang, D., Cui, C., Hao, X., "Research directions on Kolmogorov-Arnold Networks: A comprehensive review". Symmetry **Vol** 18 (2026).
37. Yang, J., Yin, S., Chang, Y., Li, Y., "Causal meta-learning framework: A co-evolutionary approach to disentanglement and generalization for fault diagnosis of rotating machinery". IEEE Transactions on Instrumentation and Measurement **Vol** 75 (2026).
38. Pan, S.J., Yang, Q., "A survey on transfer learning". IEEE Transactions on Knowledge and Data Engineering **Vol** 22 (2010).
39. Ganin, Y., Lempitsky, V., "Domain-adversarial training of neural network". Journal of Machine Learning Research **Vol** 17 (2016).
40. Dai, B., Frusque, G., Li, T., Li, Q., Fink, O., "Smart Filter Aided Domain Adversarial Neural Network for Fault Diagnosis in Noisy Industrial Scenarios". Engineering Applications of Artificial Intelligence **Vol** 126 (2023).
41. Chen, B., Zhang, X., Shen, C., Li, Q., Song, Z., "CoUDA: Continual Unsupervised Domain Adaptation for Industrial Fault Diagnosis Under Dynamic Working Conditions". IEEE Transactions on Industrial Informatics **Vol** 21 (2025).
42. Zhang, W., Liu, P., Chen, H., Wang, R., "A Source-Free Domain Adaptation Framework for Rotary Machine Fault Diagnosis". Sensors **Vol** 25 (2025).
43. Yang, X., Ge, M., Wang, B., Li, Z., Sun, K., "Dual-Path Adversarial and Progressive Self-Training Network for Cross-Condition Open-Set Domain Adaptation Fault Diagnosis". Simulation Modelling Practice and Theory **Vol** 138 (2026).
44. Chen, G., Zhao, L., Xu, Y., Li, F., "Interactive Residual Domain Adaptation Networks for Industrial Fault Diagnosis Under Partial Domain Shift". Mechanical Systems and Signal Processing **Vol** 213 (2024).
45. Wang, M., Deng, W., "Deep visual domain adaptation: A survey". Neurocomputing **Vol** 312 (2018).
46. Zhuang, F., Qi, Z., Duan, K., et al., "A comprehensive survey on transfer learning". Proceedings of the IEEE **Vol** 109 (2021).
47. Li, X., Zhang, W., Ding, Q., "Domain adaptation for fault diagnosis: A review". Measurement **Vol** 188 (2022).
48. Wang, Y., Zhao, R., Wang, J., et al., "Domain generalization for fault diagnosis: A survey". Mechanical Systems and Signal Processing **Vol** 182 (2023).
49. Li, G., Ren, L., Pradhan, O., ONeill, Z., Wen, J., Yang, Z., Fu, Y., Chu, M., Huang, J., Wu, T., Candan, K.S., Adetola, V., Zhu, Q., "Emulation and detection of physical faults and cyber-attacks on building energy systems through real-time hardware-in-the-loop experiments". Energy and Buildings **Vol** 320 (2024).
50. Polat, H., Turkoglu, M., Polat, O., Sengur, A., "A novel approach for accurate detection of the DDoS attacks in SDN-based SCADA systems based on deep recurrent neural network". Expert Systems With Applications **Vol** 197 (2022).
51. Shilenge, M.C., Telukdarie, A., "Optimization of operational and information technology integration towards Industry 4.0". IEEE International Symposium on Industrial Electronics (2022).
52. Kok, A., Martinetti, A., Braaksma, J., "The impact of integrating information technology with operational technology in physical assets: A literature review". IEEE Access **Vol** 12 (2024).
53. Taheri, M., Khorasani, K., Shames, I., Meskin, N., "Cyberattack and machine-induced fault detection and isolation methodologies for cyber-physical systems". IEEE Transactions on Control Systems Technology **Vol** 32 (2024).
54. Liu, K., Xie, Y., Chen, Y., Xie, S., Chen, X., Fang, D., Sun, L., "SFACIF: A safety function

attack and anomaly industrial condition identified framework". Computer Networks **Vol** 257 (2025).
55. Jeffrey, N., Tan, Q., Villar, J.R., "A hybrid methodology for anomaly detection in cyber-physical systems". Neurocomputing **Vol** 568 (2024).
56. Ma, F., Ji, C., Rao, J., Han, C., Wang, J., Sun, W., "A cyber-attack detection based on time-delay mutual information analysis". Process Safety and Environmental Protection **Vol** 200 (2025).
57. Jia, W., An, A., Gong, B., Shi, Y., "An efficient transfer fault diagnosis method integrating feature redundancy selection and multi-strategy parameter optimization". Expert Systems With Applications **Vol** 279 (2025).
58. Liu, J., "Bearing fault diagnosis based on incremental transfer learning and ensemble learning with stochastic configuration network". Expert Systems With Applications **Vol** 290 (2025).
59. Tursunalieva, A., Alexander, D., et al., "Making sense of machine learning: A review of interpretation techniques and their applications". Applied Sciences **Vol** 14 (2024).
60. Hassija, V., Chamola, V., et al., "Interpreting black-box models: A review on explainable artificial intelligence". Cognitive Computation **Vol** 16 (2024).
61. Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., Rinzivillo, S., "Benchmarking and survey of explanation methods for black box models". Data Mining and Knowledge Discovery **Vol** 37 (2023).
62. Yang, R., Hu, J., Li, Z., Mu, J., Yu, T., Xia, J., Li, X., Dasgupta, A., Xiong, H., "Interpretable machine learning for weather and climate prediction: A survey". arXiv preprint arXiv:2403.18864 (2024).
63. Muhammad, G., Kleijn, W.B., Zhang, M., "Domain adaptive neural network for object recognition". Lecture Notes in Computer Science **Vol** 8862 (2014).
64. Sun, B., Feng, J., Saenko, K., "Return of frustratingly easy domain adaptation". AAAI Conference on Artificial Intelligence (2016).
65. Sun, B., Saenko, K., "Deep CORAL: Correlation alignment for deep domain adaptation". Lecture Notes in Computer Science **Vol** 9915 (2016).
66. Bernal-de Lázaro, J.M., Rodríguez-Ramos, A., Silva-Neto, A.J., Llanes-Santiago, O., "A three-layer explainable stacking ensemble for monitoring subtle anomalies in chemical processes". Chemometrics and Intelligent Laboratory Systems **Vol** 268 (2026).
67. Awad, N., Mallik, N., & Hutter, F. "DEHB: Evolutionary Hyperband for Scalable, Robust and Efficient Hyperparameter Optimization". Proceedings of IJCAI. (2021).
68. Xu, H., Zhang, Z., Deng, Q., et al. "A segmented differential evolution with enhanced diversity and semi-adaptive parameter control". Complex & Intelligent Systems, 11, 251 (2025).
69. Qiying L., Yunze H., Xuefeng G., Yiming N., Jianbin S., Xiuluo L., Baoyuan D., Qiang Z., Kai Z., "Digital twin-driven condition monitoring methods for power electronic devices: Application and challenges". Renewable and Sustainable Energy Reviews, **Vol** 235 (2026), 116983.
70. Sergio de L.D., Roberto M.L., Francisco Javier R.S., Edel D.L, Emilio José B.P., "A real-time digital twin approach on three-phase power converters applied to condition monitoring". Applied Energy, **Vol** 334 (2023), 120606.