

A Learnable Cross-Modal Adapter for Industrial Fault Detection Using Pretrained Vision Models

Jonne van Dreven , Abbas Cheddad , *Senior Member, IEEE*, Sadi Alawadi ,
Ahmad Nauman Ghazi , Jad Al Koussa , and Dirk Vanhoudt 

Abstract—Automatic fault detection and diagnosis (FDD) are critical for maintaining reliable and efficient industrial systems. However, conventional methods rely heavily on manual inspections or threshold-based techniques, which often fail to capture the dynamic patterns in time series (TS) sensor data. As a result, faults persist for extended periods, leading to suboptimal system operations, increased energy waste, and significant economic losses. This work proposes a cross-modal framework that facilitates the efficient deployment of state-of-the-art pretrained vision models for enhanced FDD, with two novel TS-to-image transformations: first, an adapter deep encoder that learns optimal, task-specific representations from raw sensor data while generating outputs that are input-compliant with pretrained models. Second, an enhanced line plot that creates geometric shapes of two related signals. Comparative experiments against fixed methods, including spectrograms, Gramian angular fields, Markov transition fields, recurrence plots, and five deep learning baseline models, showed substantial performance gains across diverse domains. InceptionTime achieved the highest average baseline performance with an F_1 of 88.6%, while the adapter and shapes achieved 94.4% and 92.4%, respectively. The findings highlight the potential of the cross-modal framework for FDD to facilitate early intervention and efficient system maintenance in industrial settings.

Index Terms—Cross-modal adaptation, deep learning, fault detection and diagnosis (FDD), pretrained vision models, time series (TS), transfer learning (TL).

Received 27 October 2025; revised 13 January 2026; accepted 27 January 2026. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Paper no. TII-25-7566. (Corresponding author: Jonne van Dreven.)

Jonne van Dreven is with the Department of Computer Science, Blekinge Institute of Technology, 37179 Karlskrona, Sweden, also with the Unit Water and Energy Transition, Flemish Institute for Technological Research (VITO), 2400 Mol, Belgium, and also with EnergyVille, 3600 Genk, Belgium (e-mail: jonne.van.dreven@bth.se, jonne.vandreven@vito.be).

Abbas Cheddad is with the Department of Computer Science, Blekinge Institute of Technology, Karlskrona, Sweden, and also with the Institute of Computer Science, University of Tartu, 51009 Tartu, Estonia.

Sadi Alawadi is with the Department of Computer Science, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden.

Ahmad Nauman Ghazi is with the Department of Software Engineering, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden.

Jad Al Koussa and Dirk Vanhoudt are with Unit Water and Energy Transition, VITO, Mol, Belgium, and also with EnergyVille, 3600 Genk, Belgium.

Digital Object Identifier 10.1109/TII.2026.3659264

I. INTRODUCTION

INDUSTRIAL fault detection and diagnosis (FDD) are essential for maintaining the reliability, safety, and operational efficiency of modern infrastructure systems. However, as these systems become increasingly complex and age, traditional FDD approaches often fall short of meeting evolving diagnostic demands. Typically, maintenance relies on manual inspections or threshold-based techniques [1], which often fail to identify faults promptly. As a result, faults persist for extended periods, leading to degraded system performance and economic losses. For example, district heating (DH) networks are highly susceptible to faults [2], [3], with major cities, such as Copenhagen, estimating that more than 50% of substations operate suboptimally [4]. This not only causes substantial energy losses, but a large faulty substation can cost up to 100 000 Euros in extra pumping power per year [5]. Against the backdrop of climate change and the global push toward greener energy solutions, advancing robust FDD approaches for systems like DH is more critical than ever. Recent studies have explored transforming time series (TS) data into 2-D visual representations using different methods, such as Gramian angular fields (GAF) [6], recurrence plots (RP) [7], spectrograms [8], and Markov transition fields (MTF) [6]. However, these methods rely on fixed, hand-crafted transformations that might not capture the most informative features for the downstream task.

While transforming TS-to-images has shown promise, existing approaches rely on fixed, hand-crafted mappings that are agnostic to the downstream task. These transformations implicitly assume that a single predefined mapping is universally optimal, which is unlikely to hold across dynamic, heterogeneous industrial systems and fault mechanisms. The term *dynamic TS* refers to signals where temporal variations encode changes in system states, operating conditions, or fault progression, rather than assuming stationarity. In contrast, this work reframes TS-to-image conversion as a learnable problem. Instead of designing the transformation manually, our model learns a task-optimal mapping (i.e., \mathbb{R}^2 plane mapping) that adapts to the data and objective while remaining strictly input-compliant with pretrained vision backbones. This shift from fixed representations to learned cross-modal adaptation constitutes the core contribution of this article. Moreover, the proposed approach seamlessly leverages the feature extraction capabilities of pretrained vision backbones without requiring extensive retraining. This is particularly advantageous in engineering contexts, where acquiring

large, labeled datasets is challenging. Although pretrained vision models are ubiquitous in image processing, their cross-modal adaptation to TS data in industrial FDD remains relatively unexplored [9]. This work demonstrates, through four industrial datasets, including real-world and lab emulations from various domains (DH, oil, and bearing domains), that such mapping reveals richer latent patterns and significantly improves subsequent FDD performance. This cross-disciplinary perspective suggests that the distinctions between data modalities might be more flexible, opening new avenues for innovative solutions in the TS domain.

The main contributions of this article are as follows.

- 1) A novel end-to-end *trainable adapter* framework that converts raw TS into images optimized through a multitask learning objective for industrial FDD.
- 2) A *model-agnostic* design: the adapter's branch produces exactly the dimensions required by any off-the-shelf vision backbone (e.g., ResNet, DenseNet, ViT) without the need to modify its core.
- 3) A complementary *fixed geometric-shape* TS-to-image transformation, serving as a strong competitor against the adapter's gains.
- 4) Comprehensive experiments on multiple industrial datasets, covering DH, oil, and bearing domains, that demonstrate consistent and substantial FDD performance improvements.

The rest of this article is organized as follows. Section II surveys industrial FDD and TS-to-image methods. Section III details the proposed approach. Section IV provides the theory. Section V describes the experimental setup and datasets. Section VI reports the results. Finally, Section VII concludes this article and outlines future work.

II. RELATED WORK AND RESEARCH GAP

Deep learning approaches for TS analysis have predominantly used architectures, such as convolutional neural networks (CNNs) and long short-term memory networks (LSTMs), where LSTMs [10] are adept in fault detection and forecasting due to their capacity to model long-term dependencies in sequential data. Meanwhile, CNN-based models have also proven effective by capturing local temporal patterns through hierarchical feature extraction, offering both computational efficiency and robust performance in many TS applications [11]. Recent studies have combined CNN and LSTM architectures to harness both local feature extraction and sequence modeling capabilities [12], while others have focused on cross-domain adaptation [13]. These models remain the de facto standard in industrial FDD [14], [15], [16]. Although these approaches have emerged as robust methods, their exclusive reliance on raw TS data may limit their ability to exploit richer, higher level representations, such as images, that can capture more nuanced and discriminative features. Earlier works to transform TS data into image representations have introduced several fixed methods. GAF [6] converts TS into polar coordinates and computes a Gramian matrix, capturing temporal correlations as texture-like images. RP [7] visually depicts the recurrence of

states within TS. MTF [6] represents the transition probabilities between quantized values, thereby encoding the dynamic behavior. Spectrograms [8] compute the short-time Fourier transform of a signal to visualize how its frequency content evolves over time. While these have been shown to enhance deep learning performance in various domains [17], [18], [19], [20], [21], [22], their reliance on fixed transformations may limit their ability to capture the most discriminative features for a given downstream application. More recent studies have demonstrated the potential of leveraging visual representations for TS tasks. For instance, Sood et al. [23] proposed an image-driven framework that converts TS into visual plots and uses an end-to-end convolutional AE to predict future images. Similarly, Semenoglou et al. [24] introduced a method where univariate TS are transformed into grayscale line plots and processed through end-to-end CNNs to generate accurate point forecasts. Recently, Li et al. [25] extended this paradigm to irregularly sampled TS by converting them into line plots and fine-tuning vision transformers, demonstrating robust performance.

Recent years have seen increasing interest in cross-modal learning, including audio-to-image representations [26], text-to-image, and learnable front-ends for speech and audio processing [27], [28]. These approaches demonstrate that replacing fixed signal representations with learned mappings can improve downstream performance. However, such methods are typically designed for specific modalities (e.g., audio spectrogram learning), jointly train both the representation and the backbone, or target tasks fundamentally different from industrial fault detection. In contrast, TS-to-image learning for industrial FDD remains largely under explored, and there is a significant need for improved TS-to-image representations [29], a gap that our work directly addresses.

This work introduces two novel approaches: 1) a learnable adapter deep encoder network that transforms TS into task-optimal image representations and 2) an enhanced line plot approach that transforms two related TS into geometric shapes, thereby exploiting the inherent shape boundary structure detection capabilities of pretrained vision models. Moreover, the adapter's output layer is designed to be input-compliant with any existing pretrained (vision) model, enabling flexible integration with models like DenseNet. The proposed method thereby reduces the reliance on extensive labeled datasets, addressing a major limitation that often slows progress in engineering applications, such as DH [30].

III. PROPOSED METHOD

This section presents the proposed framework for transforming TS data into visual representations for industrial FDD, detailing the implementation of the adapter deep encoder, fusion strategies, and geometric shapes. A Schematic overview of the proposed framework is shown in Fig. 1.

A. Shape Representation

The shape representation is an enhanced line-plot visualization technique that encodes the relationship between two paired TS as a closed geometric shape to explicitly capture

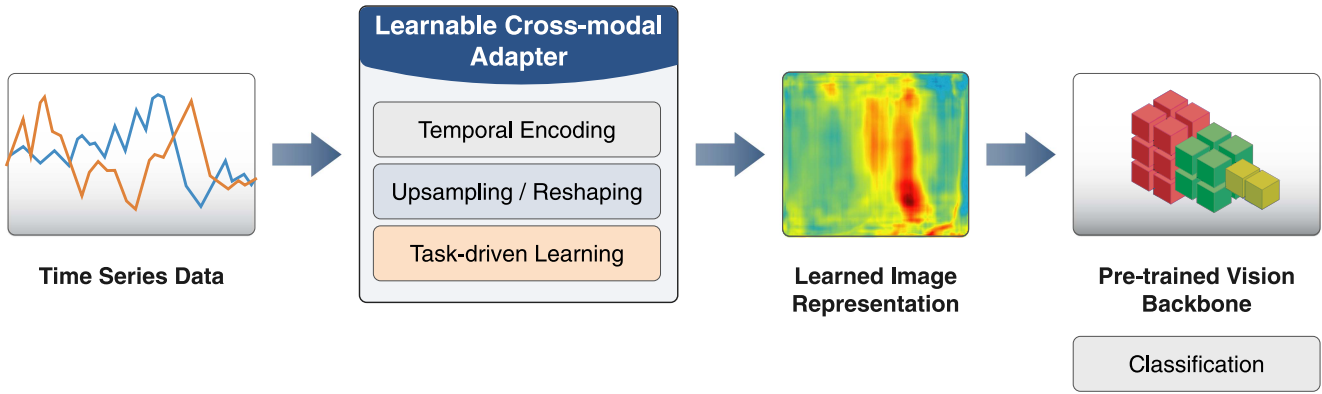


Fig. 1. Schematic of the proposed learnable cross-modal adapter transforming TS data into a learned image representation for pretrained vision models and FDD (classification).

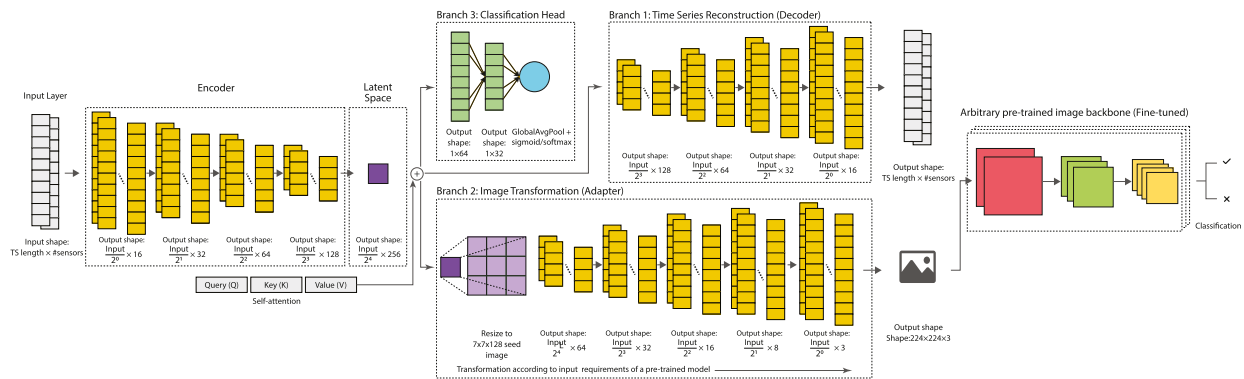


Fig. 2. Architectural overview of the proposed FDD framework. The encoder maps the input TS to a latent code z with transpose-conv blocks, from which three branches operate in parallel: (1) a decoder mirrors the encoder for TS reconstruction, (2) an adapter that maps z to a $7 \times 7 \times 128$ seed, which is then progressively upsampled to a $224 \times 224 \times 3$ image, and (3) a latent classification head enforcing discriminative structure.

dynamic interactions within a system. Dynamics are preserved implicitly through temporal ordering and changes in the relative magnitudes over time. The aim is to leverage the shape boundary structure detection capabilities of pretrained vision models. Suppose the supply temperature (T_s) and return temperature (T_r) are available from a DH substation for a TS window of N time steps. Then, the temperature difference (ΔT), a key indicator of operational irregularities [31], is represented as a geometric shape, with height H and width W , using Algorithm 1.

B. Spectrogram Fusion

Spectrograms encode frequency components' evolution over time, thereby capturing nonstationary and transient dynamics of the underlying system enabling vision models to detect periodic and anomalous fluctuations that may indicate faults. To incorporate the interaction between two features, first generate individual spectrograms, $S_1(t, f)$ and $S_2(t, f)$, and then fuse them into a single representation using Spectrogram Geometric Mean fusion (SGM), which preserves shared temporal–frequency structures while suppressing modality-specific noise, using the following equation:

$$S_{gm}(t, f) = \sqrt{S_1(t, f) \cdot S_2(t, f)} \quad (1)$$

Algorithm 1: Shape Modality.

Require: Paired, temporally aligned sequences

$$T_s = \{s_i\}_{i=1}^N, T_r = \{r_i\}_{i=1}^N; \text{ image size } (H, W)$$

Ensure: Shape image $I \in \{0, 1\}^{H \times W}$

- 1: Define horizontal coordinates $x_i = \frac{i-1}{N-1}(W-1)$
- 2: Compute value bounds:

$$v_{\min} = \min_{i \in \{1, \dots, N\}} \{s_i, r_i\}, \quad v_{\max} = \max_{i \in \{1, \dots, N\}} \{s_i, r_i\}$$

- 3: Define vertical mapping $y(v) = \frac{v-v_{\min}}{v_{\max}-v_{\min}}(H-1)$
- 4: Initialise empty vertex list P
- 5: **for** $i = 1$ to N **do**
- 6: Append $(x_i, y(s_i))$ to P
- 7: **end for**
- 8: **for** $i = N$ down to 1 **do**
- 9: Append $(x_i, y(r_i))$ to P
- 10: **end for**
- 11: Rasterise the filled polygon P into a binary image I

where t denotes the temporal index, f represents the frequency index of the time–frequency representation, S_1 and S_2 are the spectrograms for the first and second features, respectively,

while S_{gm} represents the fused spectrogram. The fusion balances the contributions of both channels. Alternatively, a Spectrogram Principal Component Analysis-based fusion (SPCA) can be used to concatenate both spectrograms (S_1 and S_2) into a feature vector, and project it into its first principal component as shown in the following equation:

$$S_{PCA}(t, f) = \mathbf{w}^\top \begin{bmatrix} S_1(t, f) \\ S_2(t, f) \end{bmatrix} \quad (2)$$

where \mathbf{w} indicates the weight vector corresponding to the first principal component. Both fusion methods enrich the input for deep learning models by capturing the interactions between the two signals.

C. Adapter Deep Encoder

Rather than a fixed TS-to-image transformation rule, this article proposes a trainable adapter deep encoder that 1) learns how to preserve essential TS information in the latent code through a multitask objective and 2) transforms input TS into the exact image dimensions required by a subsequent pre-trained vision model, thereby enabling zero-touch reuse of existing backbones without further modifications. In practice, the encoder learns the sequence dynamics, while the adapter branch handles shape conversion of that latent code into the exact layout demanded by an arbitrary backbone. Fig. 2 illustrates the overall architecture, which comprises three main components: 1) an encoder, 2) dual decoder components [TS reconstruction and image transformation (adapter) branch], and 3) a latent classification head. The encoder compresses the input TS into a compact latent representation that retains the essential temporal information for the downstream task. It consists of four sequential convolutional layers with increasing filter sizes (16,32,64,128) while progressively reducing the temporal resolution of the input signal by half (strides = 2) in each layer. The network incorporates several regularization strategies to mitigate overfitting and ensure the learned representations remain faithful to the input data. For instance, each convolution layer is followed by batch normalization, dropout (0.3) and LeakyReLU activation.

Furthermore, the adapter deep encoder uses a multitask training objective, which is a composite loss described as follows:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{tv}} \mathcal{L}_{\text{tv}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} \quad (3)$$

where, \mathcal{L}_{rec} , \mathcal{L}_{tv} , and \mathcal{L}_{cls} denote the reconstruction, total-variation, and classification losses, respectively, while λ_{rec} , λ_{tv} , and λ_{cls} are scalar weighting coefficients controlling their relative contributions during training. The reconstruction loss optimizes the mean-squared error, where $\mathbf{x} \in \mathbb{R}^{N \times D}$ is the input window (length N , D channels) and $\hat{\mathbf{x}}$ its decoder reconstruction, then the loss is defined as

$$\mathcal{L}_{\text{rec}} = \frac{1}{ND} \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2. \quad (4)$$

The total-variation loss encourages spatial smoothness. Let $\mathbf{I} \in \mathbb{R}^{H \times W \times K}$ be the adapter-generated image before it is fed

to the vision backbone, then

$$\mathcal{L}_{\text{tv}} = \frac{1}{|\Omega|K} \sum_{i=1}^{H-1} \sum_{j=1}^{W-1} \sum_{k=1}^K |I_{i+1,j,k} - I_{i,j,k}| + |I_{i,j+1,k} - I_{i,j,k}| \quad (5)$$

where H represents the height, W represents the width, K represents the number of channels, i and j represent index pixel rows and columns, respectively, and $\Omega := \{1, \dots, H-1\} \times \{1, \dots, W-1\}$ and $|\Omega| = (H-1)(W-1)$. Finally, the classification loss, e.g., fault detection, minimizes the binary cross-entropy between the predicted probability $\hat{y} \in (0, 1)$ and the ground-truth label $y \in \{0, 1\}$

$$\mathcal{L}_{\text{cls}} = -[y \log \hat{y} + (1 - y) \log(1 - \hat{y})]. \quad (6)$$

For multiclass problems, such as fault diagnosis, it is replaced by the categorical cross-entropy over the softmax output. During backpropagation, all three terms are optimized jointly, forcing the latent code to be simultaneously: 1) information-preserving, 2) image-smooth, and 3) fault-discriminative. After computing the latent representation, a self-attention mechanism adjusts features based on contextual interdependencies. Self-attention calculates the pairwise similarity between latent elements, producing an output emphasizing the most informative temporal patterns when added residually to the latent features. The refined latent representation is aggregated via global average pooling and fed into the classification head. It comprises a sequential stack of fully connected layers (64 and 32) to produce a classification probability. In parallel, the reconstruction branch (decoder) mirrors the encoder to upsample the latent representation back to the original TS dimensions, ensuring that the latent features preserve the essential information of the input signal.

The adapter treats the latent code as a 2-D map with time along one axis and channels along the other. Let T denote the reduced temporal length after the encoder and C the latent channel dimension, then the $(T \times C)$ tensor is expanded to $(T \times C \times 1)$, projected via a 2-D convolution to $(T \times C \times C_p)$, where C_p is the number of projection channels. The latent codes are resized to a compact seed (e.g., 7×7), and successive upsampling and convolution stages refine this seed into a $224 \times 224 \times 3$ image that is directly compatible with the vision backbones. This preserves temporal adjacency along one spatial axis prior to projection and allows the learned projection to mix temporal-channel information before upsampling.

Thus, this learned adapter module ‘‘plugs’’ seamlessly into any off-the-shelf pretrained vision model without further modification. Notably, the upsampling stage can be tailored to produce outputs that match the specifications of any pretrained model, thereby representing a universal adapter, enabling broad reuse and consistent FDD gains across industrial settings. Finally, once the adapter deep encoder F_Θ is trained and frozen, applying the FDD pipeline is straightforward: 1) slide an N -sample window over the incoming TS, 2) map each window to a $224 \times 224 \times 3$ image via F_Θ , 3) feed the image to a fine-tuned vision backbone g to obtain $\hat{y} = \sigma(g \circ F_\Theta)$ for fault *detection* or $\hat{y} = \text{softmax}(g \circ F_\Theta)$ for fault *diagnosis*.

IV. THEORETICAL ANALYSIS

The proposed *learnable adapter* is provably *no worse*, and can be strictly better, than any *fixed* TS-to-image transformation. Given a TS window $\mathbf{x} \in \mathbb{R}^{N \times D}$ and label $y \in \mathcal{Y}$, a fixed transformation is $T : \mathbb{R}^{N \times D} \rightarrow \mathbb{R}^{H \times W \times C}$. The adapter is the learnable map F_Θ , $\Theta \in \mathbb{R}^P$, obtained after training. A vision backbone $g \in \mathcal{G}$ is then fine-tuned on the frozen adapter outputs, yielding the final predictor $h = g \circ F_\Theta$. The population risk is $\mathcal{R}(h) = \Pr\{h(\mathbf{x}) \neq y\}$, and $L^p(\mathcal{X})$ represents the space of p -integrable functions $f : \mathcal{X} \rightarrow \mathbb{R}^{H \times W \times C}$ on the compact set $\mathcal{X} \subset \mathbb{R}^{N \times D}$, equipped with the norm

$$\|f\|_{L^p(\mathcal{X})} = \left(\int_{\mathcal{X}} \|f(x)\|_F^p dx \right)^{1/p}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Proposition 1 (Representation): For every exponent $p \in [1, \infty)$, every tolerance $\varepsilon > 0$, and every bounded continuous map $T : \mathcal{X} \rightarrow \mathbb{R}^{H \times W \times C}$, there exist parameters Θ^\dagger such that $\|F_{\Theta^\dagger} - T\|_{L^p(\mathcal{X})} < \varepsilon$.

Proof: A finite ReLU-CNN followed by a linear (1×1) projection is a universal approximator on compact domains in L^p (see [32, Thm. 1]). The adapter has exactly this form. ■

Hence, every fixed transformation is a point in the hypothesis space $\{F_\Theta\}_\Theta$. Then, the *no-worse* risk bound argument follows immediately from the hypothesis-class inclusion

$$\mathcal{H}_T = \{g \circ T : g \in \mathcal{G}\}, \quad \mathcal{H}_F = \{g \circ F_\Theta : g \in \mathcal{G}, \Theta\}.$$

Because T is representable by F_{Θ^\dagger} (see Proposition 1), it follows that $\mathcal{H}_T \subseteq \mathcal{H}_F$ and therefore

$$\inf_{h \in \mathcal{H}_F} \mathcal{R}(h) \leq \inf_{h \in \mathcal{H}_T} \mathcal{R}(h). \quad (7)$$

If a fixed transformation discards label-informative structure that some F_Θ preserves, then the inclusion is strict, as is the inequality in (7), i.e., the minimal achievable risk under the adapter is strictly lower than under any fixed mapping, matching the empirical gains observed in Section VI.

The theoretical analysis establishes that a learnable TS-to-image mapping with sufficient expressive capacity can approximate task-optimal representations and is guaranteed to be no worse than fixed, hand-crafted transformations. Importantly, the theoretical analysis does not prescribe a unique network architecture; rather, it motivates the use of a flexible, nonlinear adapter with sufficient capacity. Guided by this insight, the proposed adapter is implemented as a multilayer network with progressive upsampling stages. This design ensures sufficient representational power to satisfy the universal approximation property in practice, while preserving spatial structure and avoiding information bottlenecks prior to the vision backbone. The number of layers and upsampling stages is chosen to balance expressive capacity and training stability on limited industrial datasets, rather than to maximize model complexity. As a result, the proposed architecture preserves the theoretical guarantees while remaining practically efficient and robust.

V. EXPERIMENTAL SETUP

The study consists of the following three experiments.

- 1) *Experiment 1:* Evaluates fault detection using three baseline image modalities (line plots, spectrograms, and shapes) combined with five pretrained models (see Section V-B) to identify a robust vision model.
- 2) *Experiment 2:* Assesses fault detection performance across seven image modalities (see Section V-A) on multiple industrial TS datasets (see Section V-C).
- 3) *Experiment 3:* Includes an ablation study to disentangle the contributions of the proposed adapter, the backbone, and pretraining.
- 4) *Experiment 4:* Investigates fault diagnosis capabilities by testing the adapter, shape, and baseline approach using Dataset I.

To provide an industrially grounded benchmark, this work includes two end-to-end baselines that remain the de facto standard in engineering and recent surveys [14], [15], [16]: 1) an LSTM with an attention layer and 2) a hybrid CNN-LSTM model. These baselines share the same convolutional inductive bias as the learnable adapter and pretrained vision models, ensuring that any performance differences reflect the TS-to-image mapping framework rather than changes in backbone capacity, and thus do not confound the experimental variable we aim to test. For completeness, the baselines include state-of-the-art TS models, including InceptionTime [33], MiniRocket [34], and time series transformer (TST) [35], implemented from the TSAI library [36]. Moreover, while learnable representations exist in adjacent domains, such as audio-to-image or text-to-image, they are not designed to learn task-aligned TS-to-image mappings compatible with off-the-shelf vision backbones. To the best of the authors' knowledge, no prior work provides a directly comparable baseline for the proposed setting, precluding a direct baseline comparison. All experiments were run in Python 3.9.6 on a 2.6 GHz quad-core Intel Core i7 machine with 16 GB of RAM. The training configuration for the baselines and adapter uses Adam with initial learning rate = 0.001, batch size = 32, epochs = 100, with early stopping (patience = 8), and seed = 42. The adapter uses composite loss weights $\lambda_{\text{rec}} = 0.5$, $\lambda_{\text{iv}} = 0.1$, and $\lambda_{\text{cls}} = 0.4$. Fine-tuning the vision backbone uses a learning rate = 0.001 and 20 epochs. Settings remain consistent across experiments and datasets. All baseline methods were implemented using publicly available reference implementations or official repositories. Unless stated otherwise, the recommended default hyperparameters provided by the original authors were used.

The data are partitioned using stratified fivefold cross-validation (seed = 42): 80% for train/validation (80/20 within fold) and a disjoint 20% hold-out test set. The hold-out set is only used for assessing model performance. This work reports the average and 95% confidence interval (CIs) of precision, recall, F_1 , area under the receiver operating characteristic curve (AUC), and precision-recall curve (PRC).

A. Image Modalities

This study deliberately uses a representative set of fixed mappings (i.e., image modalities) that 1) are widely adopted in

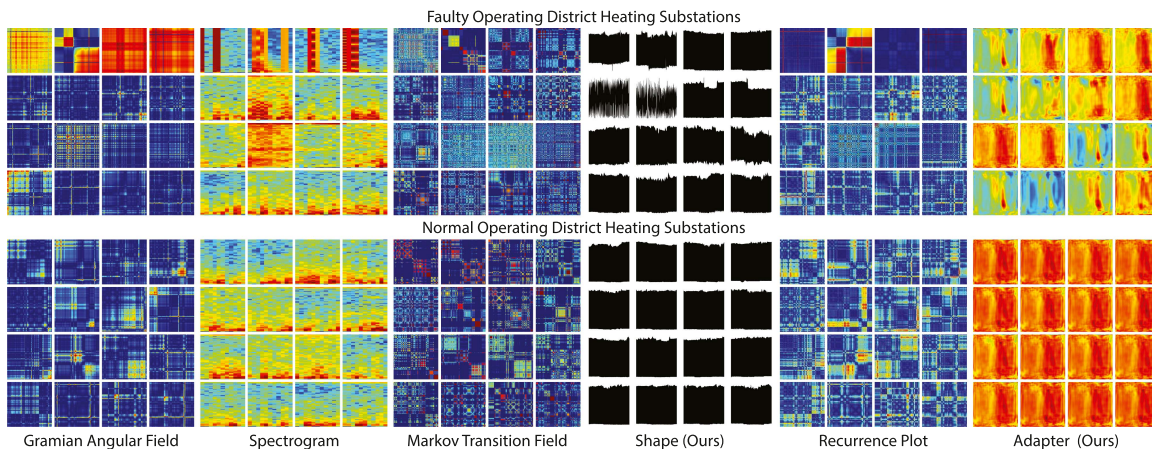


Fig. 3. Examples of TS transformed into different image modalities for faulty and normal operation scenarios.

the literature [37], [38] and 2) each emphasizes a different aspect of the signal; therefore, our evaluation spans a broad range of dataset characteristics. The different mappings are listed below and illustrated (except for line plots) in Fig. 3.

- 1) *Line plots (standard)*: Render the raw sensor trace versus time, preserving amplitude and temporal patterns.
- 2) *GAF*: Encodes long-term temporal correlations, such as slow thermal drifts and gradual baseline shifts.
- 3) *RP*: Marks when the system returns to prior states (e.g., bearing rotations become visible).
- 4) *MTF*: Highlights regime changes or gradual Markovian shifts in process behavior (e.g., load changes).
- 5) *SPCA*: Reveals broadband vibration events and transient bursts; the *fused spectrograms* further combine complementary channels to capture multisensor interactions.

B. Transfer Learning (TL) of Vision Models

This work employs five pretrained deep learning architectures, comprising both recent state-of-the-art models and well-established models, as the classification models: DenseNet121 [39], ResNet18 [40], EfficientNet-B0 [41], MobileNetV3 [42], and ViT-B16 [43]. Each vision model has been pretrained on a large-scale image dataset (e.g., ImageNet) and fine-tuned on the FDD task, i.e., using TL. Specifically, the output layer is replaced with a fully connected layer for classification, with a sigmoid for fault detection and a softmax for fault diagnosis. Then, the pretrained weights are fine-tuned on the TS-derived image modalities.

C. Dataset Description

This work uses four distinct datasets to validate the effectiveness of the proposed framework across different industrial application domains. All TS inputs are processed using standard preprocessing practices. The signals are segmented into fixed-length windows (e.g., 288 samples for daily temperature profiles), ensuring equal-length inputs for the model. The paired streams, such as T_s and T_r , are assumed to be temporally aligned, as provided by the datasets used in this study. Missing values, when present, are handled using linear interpolation to ensure

fully defined inputs. Furthermore, min-max normalization is applied jointly to the paired TS to stabilize training and preserve the relative relationships between paired signals.

- 1) *Datasets I.A & I.B (confidential)*: A multisource real-world dataset from a DH network in the Shandong Province, China, that spans one month, sampled at 5-min intervals (8928 points per sensor). Each substation measures flow (Q), supply temperatures (T_s), return temperatures (T_r), and derived features, such as temperature difference ΔT or energy consumption (E). It captures fault events under actual operational conditions, reflecting the noisy challenges encountered in field deployments. Approximately 32% of substations are in normal operational, and 68% are faulty, including (12.5%) high heat curve (HHC), (25.0%) wrong sensor placement (WSP), (6.3%) large valve (LV), (6.3%) wrong valve setting (WVS), and (18.8%) large secondary leakages (LSL). Dataset I.A contains 5-min measurements, and Dataset I.B is a resampled version at 15 min. Both are segmented into daily windows.
- 2) *Dataset II [44] (confidential)*: A dataset of controlled laboratory emulations of a residential building to provide reliable ground-truth information for method validation. It includes one normal operational scenario and five fault scenarios, each lasting two weeks. Sensor measurements are similar to Dataset I but collected at 10-s intervals (120960 points per sensor). Dataset II is resampled to a 5-min interval and segmented into daily windows.
- 3) *Dataset III [45] (public)*: A dataset of real-world multivariate TS collected from an oil production system, with a 15-min interval for two years (70080 points per sensor), from two regions in China. The continuous oil temperature (OT) signal, segmented into daily windows, is transformed into binary fault labels (roughly 43%–57% split) by computing daily average OT values and comparing them against the overall average. High useful load (HUFL) and low useful load (LUFL) are employed as predictors, with the final predictor defined as the difference between HUFL and LUFL, analogous to ΔT .

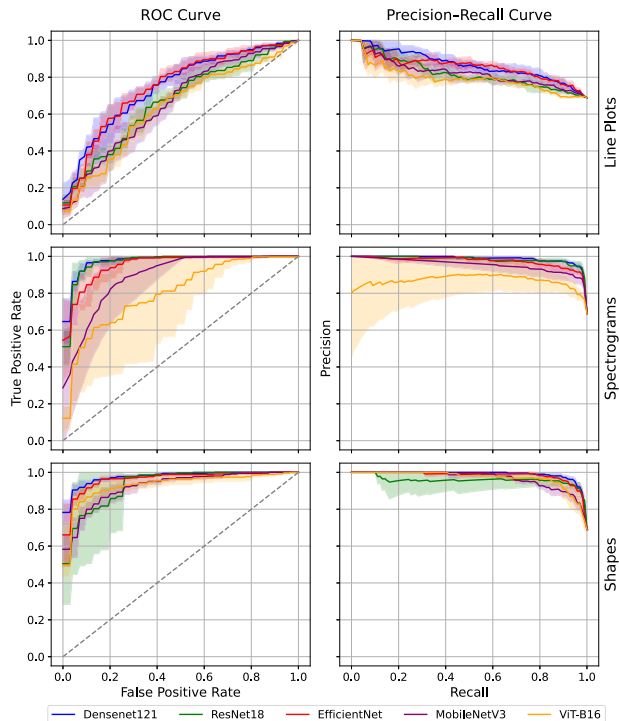


Fig. 4. Comparison of pretrained vision models on Dataset I: for each modality (rows), the ROC curve (left), and PRC (right) are shown with 95% CI. Gray dotted diagonal represents a random classifier.

- 4) *Dataset IV [46] (public)*: A dataset that offers comprehensive test data on ball bearings under normal and fault conditions. This study focuses on the fan-end vibration signals, classifying between bearings with 0.007-inch (four tests) and 0.014-inch faults (four tests), thereby providing a more challenging benchmark. Data are collected at 12000 samples/s (121 168 points per test) and segmented into consecutive 288-sample (≈ 24 ms) windows.

VI. RESULTS AND DISCUSSION

This section presents the results from the four experiments outlined in Section V. Experiment 1 compared the vision models, Experiment 2 benchmarked the best vision backbone on four industrial datasets, Experiment 3 analyzed the contributions of each component, and Experiment 4 investigated the fault diagnosis capabilities.

A. Experiment 1: Impact of Vision Models

This experiment aims to evaluate the capabilities of various pretrained vision models when transferring their knowledge to a new target task that has never been seen before. Rather than selecting a definitive “best” model, this experiment aims to identify a model candidate that demonstrates robust performance across various complex visual representations for subsequent evaluation in Experiment 2. Fig. 4 shows the AUC and PRC for each selected image modality on Dataset I. Meanwhile, Table I summarizes the individual performance metrics. The regular spectrograms showed moderate performance across

TABLE I
PERFORMANCE ACROSS VISION MODELS WITH 0.5 THRESHOLD ON DATASET I

Method	Model	Precision (%)	Recall (%)	F ₁ (%)	AUC
Shape	DenseNet	90.9 \pm 2.7	88.6 \pm 5.5	88.8 \pm 5.2	0.977 \pm 0.006
	ResNet18	87.9 \pm 2.8	85.6 \pm 3.5	85.0 \pm 4.0	0.934 \pm 0.068
	EfficientNet	89.1 \pm 5.6	87.2 \pm 6.8	86.2 \pm 8.8	0.964 \pm 0.007
	MobileNet	76.9 \pm 14.6	75.4 \pm 9.9	72.9 \pm 12.5	0.925 \pm 0.023
	ViT-B16	85.5 \pm 11.3	75.0 \pm 2.7	75.2 \pm 11.6	0.932 \pm 0.031
Spectrogram	DenseNet	68.4 \pm 2.8	64.2 \pm 5.9	62.8 \pm 4.3	0.712 \pm 0.046
	ResNet18	73.0 \pm 6.0	66.0 \pm 6.4	65.4 \pm 4.7	0.703 \pm 0.045
	EfficientNet	68.6 \pm 2.9	66.2 \pm 3.6	66.8 \pm 3.2	0.704 \pm 0.060
	MobileNet	59.9 \pm 11.8	69.2 \pm 1.6	58.9 \pm 2.5	0.630 \pm 0.075
	ViT-B16	63.9 \pm 9.8	64.0 \pm 7.4	57.6 \pm 4.0	0.550 \pm 0.075
Line plot	DenseNet	71.6 \pm 3.2	68.2 \pm 9.1	67.2 \pm 8.2	0.748 \pm 0.036
	ResNet18	71.2 \pm 4.9	51.5 \pm 10.3	47.5 \pm 11.5	0.668 \pm 0.039
	EfficientNet	68.1 \pm 5.6	69.7 \pm 2.7	67.6 \pm 5.6	0.749 \pm 0.027
	MobileNet	52.9 \pm 10.7	70.1 \pm 2.8	59.5 \pm 6.7	0.662 \pm 0.052
	ViT-B16	62.0 \pm 7.3	58.4 \pm 5.7	56.8 \pm 2.6	0.638 \pm 0.019

Best performance is highlighted in bold.

the tested vision models. For instance, ResNet18 achieved the highest precision (73.0%), while MobileNet had the highest recall (69.2%). However, the relatively wide CI and low AUC suggest the models had difficulties generalizing across folds, possibly due to variability in the spectrogram patterns or limited discriminative features captured by this modality. While spectrograms encode helpful information, they may not be sufficiently robust on their own, and the results warrant further investigation into fusion strategies. The subsequent experiments explore the geometric mean and PCA fusion of two spectrograms as a potential approach to combine complementary information from different spectrogram signals. This fusion may better consolidate transient and steady-state features, possibly yielding improved robustness and detection accuracy in the fault detection tasks. The line modality showed intermediate performance comparable to spectrograms, indicating that simple line plots capture some temporal dynamics needed for fault detection. By optimizing the threshold to maximize the F₁, the line plot modality achieved with DenseNet a precision of 76.3%, a recall of 94.1%, and an F₁ of 84.2%. These results indicated that the method is effective at identifying actual faults (high recall) while maintaining a reasonable rate of false alarms (as indicated by the precision). However, compared to shapes (optimal F₁ 95.7%), they lagged in performance with a large effect size ($T = -22.63$; $p < 0.001$; $d = 14.31$). The shape modality demonstrated a substantial performance increase across all tested vision models. DenseNet achieved the highest precision of 87.9% and an AUC of 0.977, significantly outperforming its lightweight counterpart, MobileNet (which achieved a precision of 69.9% and AUC of 0.925). Notably, MobileNet showed a large CI, indicating considerable variability in performance across folds. While ResNet and EfficientNet performed similarly, the differences were still notable, as indicated by moderate effect sizes in the F₁, ResNet ($d = -0.87$) and EfficientNet ($d = -0.38$). In addition, EfficientNet showed high variance across folds, while ResNet exhibited a broader CI in its AUC, particularly at lower thresholds. Even though the performance is largely comparable, DenseNet still maintains a measurable advantage. Interestingly, in contrast to [25], where directly converting TS data into line plots resulted in excellent performance, this work encountered

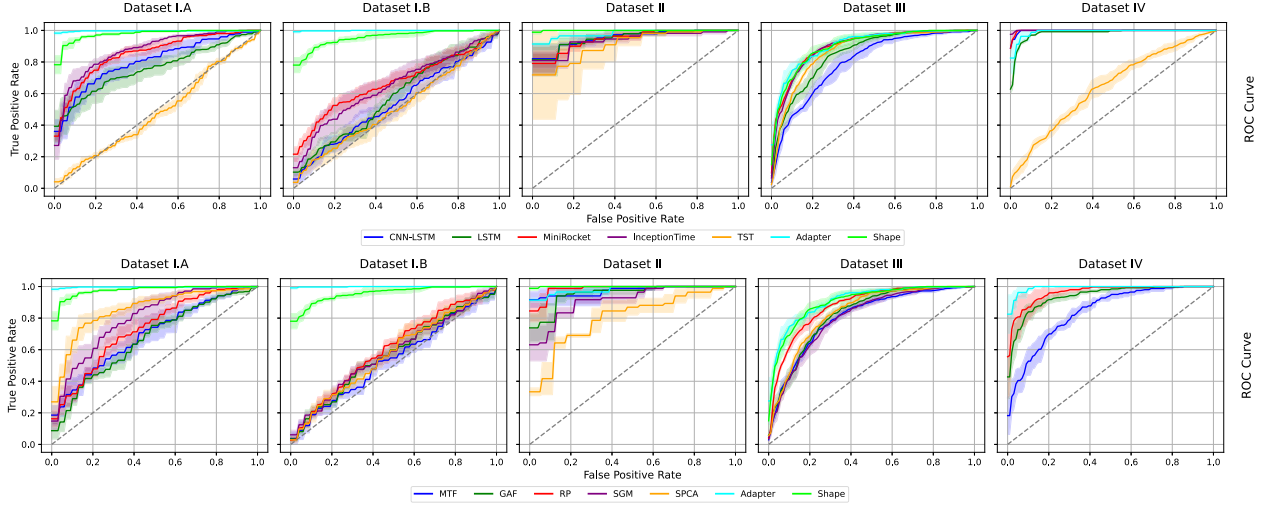


Fig. 5. Comparison of image modalities on each dataset. The shaded area is the 95% CI, and gray dotted diagonal represents a random classifier.

degraded performance. Instead, transforming the TS into shape-based modalities proved more effective. Furthermore, regardless of the underlying transformation, DenseNet, ResNet, and EfficientNet consistently ranked the highest among the architectures evaluated. These results likely stem from the inherent architectural advantages of the deep convolutional networks, which are highly effective at capturing local spatial patterns from the input images. On the other hand, MobileNet, being a lightweight model, may not possess sufficient capacity to extract the detailed discriminative features required for fault detection. Moreover, the underperformance of the vision transformer (ViT-B16) can be attributed to its greater data and training requirements; transformers typically demand larger datasets and longer training durations to fully exploit their capacity for learning global representations [43], conditions that were unmet in the available industrial datasets.

B. Experiment 2: Impact of Image Modalities

This experiment continued with DenseNet on four distinct datasets using the image modalities as outlined in Section V-A. Table II and Fig. 5 present the results categorized per dataset. The baselines in Dataset I.A (see Table II and first column of Fig. 5) achieved moderate performance. Consistent with Experiment 1, the transformer-based architecture, TST, underperforms on all datasets except Dataset II. Given its data demands, even with reduced capacity or stronger regularization, it tends to overfit in small-sample engineering regimes, yielding poor generalization. We report these results for completeness but do not discuss them further. InceptionTime reached the highest baseline performance with an F_1 of 88.4%. In contrast, the proposed adapter method dramatically improves performance, yielding a precision, recall, and F_1 of 99.1%, misclassifying only a few cases. Consequently, the shape modality also demonstrates strong performance with an F_1 of 95.7% and AUC of 0.977. These metrics suggest that the supply and return temperatures in DH systems manifest as distinct geometric patterns that are highly discriminative for fault detection. In contrast, under normal operation, the return

temperature remains relatively stable. This results in unique “shapes” in the visual representation that the model can reliably associate with faults. Conversely, the fixed transformation methods consistently underperformed relative to the adapter and shape methods. Notably, Spectrogram fusion (taking the lower end fusion SPCA) significantly enhanced performance with a large effect size compared to the regular spectrogram ($T = 4.47, p = 0.002, d = 2.83$). This suggests that the fusion of complementary spectral information effectively enriches the feature representation, thereby improving the model’s discriminative capabilities. Moreover, despite Dataset I.B being sampled at 15-min intervals, the adapter and shape transformations maintained robust performance (second column of Fig. 5). In contrast, the performance of all alternative methods degraded significantly. This observation suggests that the learned transformations are more adept at capturing the subtle fault patterns in DH data across varying temporal resolutions, thereby enhancing reliability in fault detection. Finally, Dataset II was used to validate the methods on known ground-truth information from physical lab experiments [44]. All methods demonstrated strong performance on this dataset (third column of Fig. 5), with most achieving F_1 above 90% and high discriminative power (high AUC). These results indicate that the transformation techniques reliably capture the distinct thermal anomalies associated with faulty substations, thereby supporting their use for fault detection. Moreover, the methods were further evaluated on two different application domains: oil (Dataset III) and bearings (Dataset IV). These datasets exhibit different signal characteristics and noise profiles compared to the DH data, thereby providing a robust test of the model’s adaptability. For Dataset III, as observed in the fourth column of Fig. 5, the baseline models exhibit moderate performance. MiniRocket yielded an F_1 of 81.0%; however, with a lower precision of 77.2%, suggesting more false positives and being biased to predict the positive class too often. In contrast, the proposed adapter reached a higher F_1 (82.1%) with a large effect size ($d = 0.93$), and better balanced precision and recall (83.7 and 81.9, respectively). Moreover, the shape modality also performed competitively, with an F_1 of

TABLE II
OPTIMAL PERFORMANCE ACROSS DATASETS AND METHODS

Data	Method	Precision (%)	Recall (%)	F ₁ (%)	AUC	
I.A	CNN-LSTM	81.1 ± 5.6	92.7 ± 5.8	86.1 ± 1.5	0.805 ± 0.059	
	LSTM	75.6 ± 5.8	96.2 ± 5.0	84.2 ± 1.4	0.761 ± 0.065	
	MiniRocket	82.7 ± 3.9	93.6 ± 2.8	87.7 ± 1.3	0.852 ± 0.016	
	InceptionTime	83.2 ± 1.9	94.3 ± 1.6	88.4 ± 0.7	0.873 ± 0.026	
	TST	70.4 ± 0.0	100 ± 0.0	82.7 ± 0.0	0.487 ± 0.022	
	Adapter*	99.1 ± 1.0	99.1 ± 0.6	99.1 ± 0.6	0.996 ± 0.005	
	Shape*	95.0 ± 2.1	96.5 ± 1.0	95.7 ± 0.6	0.977 ± 0.006	
	MTF*	71.9 ± 1.4	98.8 ± 1.5	83.2 ± 0.5	0.693 ± 0.061	
	GAF*	74.2 ± 2.7	94.5 ± 3.5	83.0 ± 0.5	0.666 ± 0.015	
	RP*	75.0 ± 1.3	98.0 ± 1.3	84.9 ± 0.4	0.731 ± 0.026	
	SGM*	79.9 ± 2.6	96.8 ± 2.5	87.5 ± 0.7	0.799 ± 0.051	
	SPCA*	83.8 ± 3.2	91.6 ± 2.5	87.4 ± 1.9	0.855 ± 0.043	
	I.B	CNN-LSTM	70.7 ± 1.6	100 ± 0.0	82.8 ± 1.1	0.828 ± 0.068
		LSTM	71.6 ± 1.3	99.7 ± 0.5	83.4 ± 0.8	0.834 ± 0.052
		MiniRocket	70.9 ± 0.6	99.3 ± 0.9	82.7 ± 0.1	0.666 ± 0.057
InceptionTime		71.9 ± 0.8	98.4 ± 1.3	83.1 ± 0.3	0.649 ± 0.090	
TST		70.9 ± 0.5	100 ± 0.0	82.9 ± 0.3	0.516 ± 0.047	
Adapter*		99.7 ± 0.5	99.7 ± 0.5	99.7 ± 0.3	0.997 ± 0.002	
Shape*		94.2 ± 1.0	91.8 ± 1.4	93.0 ± 0.8	0.930 ± 0.018	
MTF*		69.9 ± 0.7	100 ± 0.0	82.2 ± 0.6	0.822 ± 0.037	
GAF*		69.8 ± 1.3	99.4 ± 1.0	81.9 ± 0.6	0.820 ± 0.023	
RP*		70.1 ± 1.2	98.8 ± 1.5	82.0 ± 0.4	0.820 ± 0.022	
SGM*		69.9 ± 0.9	99.7 ± 0.5	82.2 ± 0.5	0.822 ± 0.029	
SPCA*		69.7 ± 0.7	99.4 ± 1.0	82.0 ± 0.2	0.820 ± 0.030	
II		CNN-LSTM	97.9 ± 3.6	88.2 ± 1.5	92.7 ± 1.6	0.953 ± 0.025
		LSTM	97.2 ± 4.8	87.5 ± 4.1	92.0 ± 3.4	0.952 ± 0.037
		MiniRocket	94.2 ± 6.4	88.2 ± 4.8	90.7 ± 2.4	0.947 ± 0.028
	InceptionTime	92.5 ± 5.4	90.9 ± 5.6	91.3 ± 1.5	0.939 ± 0.030	
	TST	89.0 ± 11.9	98.2 ± 2.0	92.9 ± 7.6	0.912 ± 0.130	
	Adapter*	95.3 ± 6.5	96.5 ± 4.8	95.7 ± 0.1	0.981 ± 0.010	
	Shape*	98.9 ± 1.7	100 ± 0.0	99.4 ± 0.9	0.999 ± 0.001	
	MTF*	98.8 ± 1.8	92.8 ± 8.7	95.5 ± 4.3	0.974 ± 0.041	
	GAF*	90.9 ± 4.7	95.2 ± 5.0	93.0 ± 4.3	0.955 ± 0.036	
	RP*	94.3 ± 1.8	98.8 ± 1.9	96.5 ± 1.6	0.985 ± 0.039	
	SGM*	87.9 ± 6.5	90.4 ± 5.0	88.9 ± 2.2	0.916 ± 0.019	
	SPCA*	76.4 ± 6.7	92.8 ± 8.7	83.4 ± 4.0	0.802 ± 0.037	
	III	CNN-LSTM	62.5 ± 6.2	87.2 ± 3.4	72.4 ± 3.8	0.800 ± 0.033
		LSTM	69.9 ± 2.8	84.4 ± 5.3	76.4 ± 3.7	0.853 ± 0.029
		MiniRocket	77.2 ± 1.8	85.4 ± 2.0	81.0 ± 0.5	0.900 ± 0.006
InceptionTime		75.4 ± 0.9	86.9 ± 1.8	80.7 ± 0.9	0.897 ± 0.009	
TST		70.2 ± 2.3	90.0 ± 2.8	78.8 ± 0.9	0.873 ± 0.009	
Adapter*		83.7 ± 6.4	81.9 ± 7.4	82.1 ± 1.6	0.908 ± 0.016	
Shape*		76.8 ± 3.8	87.8 ± 6.6	81.6 ± 2.6	0.903 ± 0.025	
MTF*		65.7 ± 1.9	82.6 ± 6.3	72.9 ± 1.4	0.805 ± 0.015	
GAF*		67.4 ± 1.8	83.7 ± 2.9	74.6 ± 1.7	0.823 ± 0.024	
RP*		72.9 ± 3.6	84.9 ± 4.2	78.2 ± 0.5	0.875 ± 0.006	
SGM*		67.0 ± 3.2	80.6 ± 4.5	73.0 ± 1.4	0.804 ± 0.031	
SPCA*		67.4 ± 1.8	85.5 ± 3.3	75.3 ± 0.8	0.835 ± 0.010	
IV		CNN-LSTM	96.1 ± 0.3	100 ± 0.0	98.0 ± 0.1	0.998 ± 0.001
		LSTM	92.8 ± 1.5	96.2 ± 1.5	94.4 ± 0.2	0.983 ± 0.001
		MiniRocket	97.3 ± 0.3	100 ± 0.0	98.6 ± 0.2	0.998 ± 0.001
	InceptionTime	98.9 ± 1.0	100 ± 0.0	99.4 ± 0.5	1.000 ± 0.001	
	TST	53.8 ± 3.1	94.7 ± 5.2	68.3 ± 1.1	0.643 ± 0.030	
	Adapter*	94.2 ± 0.2	97.0 ± 1.6	95.6 ± 1.0	0.992 ± 0.001	
	MTF*	69.5 ± 2.1	92.2 ± 3.8	79.1 ± 0.8	0.840 ± 0.012	
	GAF*	87.0 ± 2.5	89.5 ± 3.1	88.1 ± 1.8	0.939 ± 0.016	
	RP*	90.5 ± 4.1	92.5 ± 3.9	91.3 ± 2.7	0.963 ± 0.023	

Bold rows indicate the best-performing method for each dataset.

81.6%; however, it achieved similar precision to MiniRocket. Notably, the RP modality achieved an F₁ of 78.2%, similar to the baselines and higher than the second-best fixed method, SPCA, which yielded an F₁ of 75.3%. The findings throughout the datasets suggest that the RP modality, compared to the other fixed transformation methods, is particularly effective in capturing the discriminative features for fault detection. Finally, Dataset IV contains bearing data with established ground truth from a vibration sensor used in bearing experiments; approaches requiring multivariate inputs are excluded from this assessment. Note that these data are cleaner, meaning there is little noise, drift, or operational variability. Table II and the last column of Fig. 5 show that the baselines perform exceptionally well in this domain. For instance, InceptionTime reached the highest F₁ of 99.4% with both MiniRocket and CNN-LSTM close (98.6% and 98.0%, respectively). In such conditions, specialized

TABLE III
OPTIMAL PERFORMANCE FOR ABLATION WITH DATASET I.A

Method / Class	Precision (%)	Recall (%)	F ₁ (%)	AUC
CNN-LSTM	81.1 ± 5.6	92.7 ± 5.8	86.1 ± 1.5	0.805 ± 0.059
LSTM	75.6 ± 5.8	96.2 ± 5.0	84.2 ± 1.4	0.761 ± 0.065
MiniRocket	82.7 ± 3.9	93.6 ± 2.8	87.7 ± 1.3	0.852 ± 0.016
InceptionTime	83.2 ± 1.9	94.3 ± 1.6	88.4 ± 0.7	0.873 ± 0.026
TST	70.4 ± 0.0	100 ± 0.0	82.7 ± 0.0	0.487 ± 0.022
ID DenseNet	77.7 ± 2.9	91.3 ± 3.4	83.8 ± 0.5	0.738 ± 0.055
Adapter + DenseNet (random init)	97.2 ± 0.8	98.6 ± 0.1	97.9 ± 0.4	0.986 ± 0.006
Adapter + DenseNet (Pre-trained)	99.1 ± 1.0	99.1 ± 0.6	99.1 ± 0.6	0.996 ± 0.005

Best method is highlighted in bold.

end-to-end TS architectures perform exceptionally well because they easily lock onto those fixed patterns without needing to generalize to unseen operating contexts. In contrast, the adapter introduces an additional representation learning step, which, although less direct and slightly lossy on simple, separable data, proves beneficial in complex, noisy, real-world signals. Nevertheless, the adapter performed well with an F₁ of 95.6% and AUC of 0.992. Importantly, the results from Dataset IV highlight the strengths of the adapter in generating representations that are better aligned with the downstream task over the fixed transformation methods. For comparison, the MTF achieved an F₁ of only 79.1% and an AUC of 0.839, while the GAF and RP modalities reached F₁ of 88.1% and 91.3% with corresponding AUCs of 0.939 and 0.963, respectively.

C. Experiment 3: Impact of Architectural Components

The ablation experiment on Dataset I.A (see Table III) disentangles the contributions of each component. The 1-D DenseNet on raw TS (F₁ = 83.8%) matches the other baselines (CNN-LSTM: 86.1%, LSTM: 84.2%, InceptionTime: 88.4, and Minirocket 87.7). Adding the adapter and training DenseNet from random initialization boosts F₁ to 97.9%. The adapter deep encoder produces a task-aligned image representation in a way that subsequent 2-D DensNet can learn extremely well. Notably, the pretrained DenseNet on adapter outputs slightly raises F₁ to 99.1% ($t = 3.47$; $p = 0.008$; $d = 2.19$), improving feature extraction and convergence speed. This makes the backbone's initial weights (whether random or pretrained) less critical, but still meaningful at this level. Pretraining provides a strong initialization, making the pretrained backbone more data-efficient and faster to converge, while maximizing performance.

Taken together, the empirical findings corroborate the analytical formulation (see Section IV) and provide strong evidence that substantial gains over fixed mappings are achievable. In particular, performance improvements can be achieved through the proposed multimodal pipeline, wherein the adapter generates task-aligned, vision-compatible representations from TS data, which synergize with high-capacity image backbones. The adapter consistently outperformed traditional fixed transformation methods, achieving superior metrics across multiple datasets. Consequently, the geometric shapes yielded competitive performance. Furthermore, the discrepancy observed for the fixed representations may stem from a domain gap. Pretrained vision models are optimized for natural images; thus, when applied to inputs with atypical or complex structures, their

TABLE IV
OPTIMAL DIAGNOSIS PERFORMANCE ON DATASET I

Method / Class	Precision (%)	Recall (%)	F ₁ (%)	AUC
CNN-LSTM	69.4 ± 5.8	63.4 ± 3.9	61.2 ± 4.1	0.836 ± 0.037
LSTM	55.4 ± 5.5	48.7 ± 3.2	48.8 ± 3.1	0.752 ± 0.053
MiniRocket	64.1 ± 4.2	44.5 ± 6.4	48.8 ± 6.2	0.786 ± 0.026
InceptionTime	62.5 ± 3.9	48.2 ± 5.6	50.7 ± 5.7	0.811 ± 0.016
TST	30.1 ± 7.5	26.1 ± 5.9	26.4 ± 7.1	0.588 ± 0.465
Adapter*	89.7 ± 2.5	87.8 ± 1.5	87.9 ± 1.8	0.971 ± 0.017
Shape*	88.0 ± 2.6	85.5 ± 5.2	85.4 ± 5.1	0.976 ± 0.007

Best method is highlighted in bold. * Indicates Image Modality+DenseNet.

True Label \ Predicted Label	Shape Modality (a)					Adapter Modality (b)				
	HHC	WSP	LV	LSL	WVS	HHC	WSP	LV	LSL	WVS
HHC	0.91	0.02	0.02	0.06	0.00	0.92	0.00	0.00	0.08	0.00
WSP	0.03	0.83	0.02	0.02	0.09	0.00	0.90	0.00	0.00	0.10
LV	0.03	0.00	0.97	0.00	0.00	0.00	0.00	0.97	0.03	0.00
LSL	0.11	0.03	0.00	0.86	0.00	0.12	0.00	0.07	0.81	0.00
WVS	0.00	0.23	0.00	0.07	0.70	0.00	0.10	0.00	0.00	0.90

Fig. 6. Normalized diagnostic performance of (a) the shape modality and (b) the adapter modality.

performance may be compromised. The geometric shapes and adapter output may be more closely aligned with the natural image distribution, thereby better leveraging the learned features of these models.

D. Experiment 4: Fault Diagnosis

Finally, Experiment 4 focused on fault diagnosis to assess the efficacy of the approach in a more challenging classification setting, namely, classifying specific fault types. Table IV summarizes the diagnostic results. The adapter modality demonstrated robust performance with an F₁ of 87.9% and an AUC of 0.971. The per-class evaluation [contingency table in Fig. 6 with shape results in Fig. 6(a) and adapter results in Fig. 6(b)] shows notable differences. Similarly, the shape modality exhibited an overall optimal performance with an F₁ of 85.4% and an AUC of 0.976. Moreover, Fig. 7 illustrates a compelling observation. Even when the adapter is trained solely for fault detection, the resulting image maps exhibit discernible patterns that correlate with specific fault classes. Although some class overlap in faulty behavior is expected, this finding suggests that the adapter inherently captures class-relevant features in its latent space without explicit class instruction. These visual distinctions become more pronounced when the training objective is expanded to fault diagnosis. Interestingly, substation S06 deviates from this consistent trend observed. Although the precise cause cannot be confirmed, this deviation may indicate an operational change and facilitate early detection.

E. Limitations

While the results are promising, practitioners should be aware of the following limitations.

- 1) *Input-modality requirement*: The shape modality relies on two correlated sensor streams to construct shape patterns.

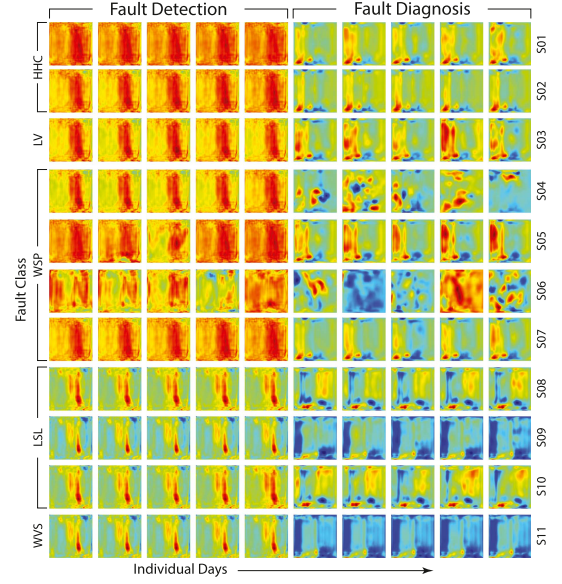


Fig. 7. Daily image maps, with different downstream tasks (detection versus diagnosis), generated by the adapter for various DH substations.

Single-sensor setups (e.g., stand-alone pressure gauges or vibration sensors) cannot leverage this.

- 2) *Interpretability*: While vision backbones excel at discriminative power, it remains challenging to map learned features back to the underlying physical causes. Industries under strict root-cause analysis may require more transparent or explainable methods.
- 3) *Domain specificity*: The evaluation has focused on several industrial domains; however, the generalizability to additional domains with different characteristics is worthwhile.
- 4) *Sampling irregularity*: This work assumes regular TS; however, irregular sampling may arise naturally in some domains (e.g., seismology, healthcare, and sensor networks) with different constraints, requiring specific adaptations.

VII. CONCLUSION

This work introduced a novel cross-modal TS-to-image framework with an adapter deep encoder and shape mapping for FDD in industrial systems. The experiments on various domains (DH, oil, and bearings) show that these learned transformations extract more robust and discriminative features than fixed transformations. Among the fixed transformation methods, RP achieved the best F₁ of 86.6%. In addition, the geometric shape proved to be highly effective, achieving competitive performance that underscores its potential as a low-complexity alternative. On average, InceptionTime yielded the highest baseline F₁ of 88.6%, whereas the adapter and shapes reached 94.4% and 92.4%, respectively. In addition, the high per-class performance of both modalities highlights their value also for fault diagnosis, enabling more precise maintenance decisions in real-world deployments. Future work will explore the method across additional industrial domains.

REFERENCES

- [1] D. Neupane et al., "Data-driven machinery fault diagnosis: A comprehensive review," *Neurocomputing*, vol. 627, Apr. 2025, Art. no. 129588.
- [2] A. Marszal-Pomianowska et al., "Strengths, weaknesses, opportunities and threats of demand response in district heating and cooling systems. from passive customers to valuable assets," *Smart Energy*, vol. 14, May 2024, Art. no. 100135.
- [3] H. Gadd and S. Werner, "Fault detection in district heating substations," *Appl. Energy*, vol. 157, pp. 51–59, Nov. 2015.
- [4] K. Honoré, "The age of digitalization and flexibility – from consumer to flexuser in the district heating system," in *Proc. 9th Int. Conf. Smart Energy Syst.*, Copenhagen, Denmark, Sep. 12–13 2023.
- [5] D. Schmidt et al., Eds., *Guidebook for the Digitalisation of District Heating: Transforming Heat Networks for a Sustainable Future*. Frankfurt am Main, Germany: AGFW Project Company, 2023, p. 67. [Online]. Available: [https://www.iee.fraunhofer.de/content/dam/iee/energiesystemtechnik/en/documents/Presse/2023/IEA_DHC_Annex_TS4_Guidebook_2023\(1\).pdf](https://www.iee.fraunhofer.de/content/dam/iee/energiesystemtechnik/en/documents/Presse/2023/IEA_DHC_Annex_TS4_Guidebook_2023(1).pdf)
- [6] Z. Wang et al., "Encoding time series as images for visual inspection and classification using tiled convolutional neural networks," in *Proc. Workshops at 29th AAAI Conf. Artif. Intell.*, vol. 1, Austin, TX, USA, 2015, pp. 1–7.
- [7] J.-P. Eckmann et al., "Recurrence plots of dynamical systems," in *Turbulence, Strange Attractors Chaos*. Singapore: World Scientific, 1995, pp. 441–445.
- [8] W. Koenig, H. K. Dunn, and L. Lacy, "The sound spectrograph," *J. Acoustical Soc. America*, vol. 18, no. 1, pp. 19–49, 1946.
- [9] A. Saeed et al., "Deep learning based approaches for intelligent industrial machinery health management and fault diagnosis in resource-constrained environments," *Sci. Rep.*, vol. 15, Jan. 2025, Art. no. 1114.
- [10] G. Van Houdt et al., "A review on the long short-term memory model," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5929–5955, 2020.
- [11] H. Ismail Fawaz et al., "Deep learning for time series classification: A review," *Data Mining Knowl. Discov.*, vol. 33, no. 4, pp. 917–963, 2019.
- [12] J. van Dreven et al., "From data scarcity to diagnostic precision: A novel data augmentation and fault diagnosis framework for district heating substations," *Eng. Appl. Artif. Intell.*, vol. 151, Jul. 2025, Art. no. 110662.
- [13] X. Zheng et al., "Exploring informative and highly-transferable features for cross-machine fault diagnosis by convformer-based biconditional domain adaptation method," *IEEE Trans. Ind. Informat.*, vol. 21, no. 4, pp. 3107–3116, Apr. 2025.
- [14] F. Zhang et al., "Deep learning in fault detection and diagnosis of building HVAC systems: A systematic review with meta analysis," *Energy AI*, vol. 12, Apr. 2023, Art. no. 100235.
- [15] W. Li and T. Li, "Comparison of deep learning models for predictive maintenance in industrial manufacturing systems using sensor data," *Sci. Rep.*, vol. 15, no. 1, Jul. 2025, Art. no. 23545.
- [16] A. G. Pereira, G. F. Barbosa, M. G. Filho, S. B. Shiki, and A. L. d. Silva, "Quality control in extrusion-based additive manufacturing: A review of machine learning approaches," *IEEE Trans. Cybern.*, vol. 55, no. 6, pp. 2522–2534, Jun. 2025.
- [17] O. Garibo-i Orts, N. Firbas, L. Sebastiá, and J. A. Conejero, "Gramian angular fields for leveraging pretrained computer vision models with anomalous diffusion trajectories," *Phys. Rev. E*, vol. 107, no. 3, Mar. 2023, Art. no. 034138.
- [18] Z. Wang and T. Oates, "Imaging time-series to improve classification and imputation," in *Proc. 24th Int. Conf. Artif. Intell.*, ser. IJCAI'15, Buenos Aires, Argentina, Jul. 2015, pp. 3939–3945.
- [19] R. K. Tripathy and U. Rajendra Acharya, "Use of features from RR-time series and eeg signals for automated classification of sleep stages in deep neural network framework," *Biocybernetics Biomed. Eng.*, vol. 38, no. 4, pp. 890–902, Jan. 2018.
- [20] N. Hatami et al., "Classification of time-series images using deep convolutional neural networks," in *Proc. 10th Int. Conf. Mach. Vis.*, Apr. 2018, pp. 242–249.
- [21] Z. Wang and T. Oates, "Spatially encoding temporal correlations to classify temporal data using convolutional neural networks," Sep. 2015, *arXiv:1509.07481*.
- [22] A. S. Khan et al., "A spectrogram image-based network anomaly detection system using deep convolutional neural network," *IEEE Access*, vol. 9, pp. 87079–87093, 2021.
- [23] S. Sood et al., "Visual time series forecasting: An image-driven approach," in *Proc. Second ACM Int. Conf. AI Finance*, ser. ICAIF '21, New York, NY, USA, May 2022, pp. 1–9.
- [24] A.-A. Semenoglou et al., "Image-based time series forecasting: A deep convolutional neural network approach," *Neural Netw.*, vol. 157, pp. 39–53, Jan. 2023.
- [25] Z. Li et al., "Time series as images: Vision transformer for irregularly sampled time series," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 49187–49204.
- [26] Z. A. Cheddad and A. Cheddad, "Active restoration of lost audio signals using machine learning and latent information," in *Lecture Notes in Networks and Systems*, vol. 822. Cham, Switzerland: Springer Nature, 2024, pp. 1–16.
- [27] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.
- [28] Y. Yuan, Z. Li, and B. Zhao, "A survey of multimodal learning: Methods, applications, and future," *ACM Comput. Surv.*, vol. 57, no. 7, pp. 167:1–167:34, Feb. 2025.
- [29] J. Ni et al., "Harnessing vision models for time series analysis: A survey," in *Proc. 34th Int. Joint Conf. Artif. Intell.*, ser. IJCAI-2025, Sep. 2025, pp. 10612–10620.
- [30] J. van Dreven et al., "Intelligent approaches to fault detection and diagnosis in district heating: Current trends, challenges, and opportunities," *Electronics*, vol. 12, no. 66, Jan. 2023, Art. no. 1448.
- [31] D. Leiria et al., "Is it returning too hot? Time series segmentation and feature clustering of end-user substation faults in district heating systems," *SSRN*, Jul. 2024, Art. no. 4894104.
- [32] D.-X. Zhou, "Universality of deep convolutional neural networks," *Appl. Comput. Harmon. Anal.*, vol. 48, no. 2, pp. 787–794, Mar. 2020.
- [33] H. Ismail Fawaz et al., "InceptionTime: Finding AlexNet for time series classification," *Data Min. Knowl. Discov.*, vol. 34, no. 6, pp. 1936–1962, Nov. 2020.
- [34] C. W. Tan, A. Dempster, C. Bergmeir, and G. I. Webb, "MultiRocket: Multiple pooling operators and transformations for fast and effective time series classification," *Data Mining Knowl. Discov.*, vol. 36, no. 5, pp. 1623–1646, 2022.
- [35] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based framework for multivariate time series representation learning," in *Proc. 27th ACM SIGKDD Conf. Know. Discov. & Data Mining*, ser. KDD '21, New York, NY, USA, Aug. 2021, pp. 2114–2124. [Online]. Available: <https://dl.acm.org/doi/10.1145/3447548.3467401>
- [36] I. Oguiza, "tsai - A state-of-the-art deep learning library for time series and sequential data," *GitHub*, 2023. [Online]. Available: <https://github.com/timeseriesAI/tsai>
- [37] Y. Zhao et al., "CIR-DFENet: Incorporating cross-modal image representation and dual-stream feature enhanced network for activity recognition," *Expert Syst. With Appl.*, vol. 266, Mar. 2025, Art. no. 125912.
- [38] S. Kumar P and J. Fredo Agastinose Ronickom, "Emotion classification through optimal segments of EDA and texture analysis of time-encoded images with artificial intelligence," *IEEE Trans. Instrum. Meas.*, vol. 74, 2025, Art. no. 2501615.
- [39] G. Huang et al., "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2261–2269.
- [40] K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
- [41] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
- [42] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.
- [43] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.
- [44] J. van Dreven et al., "A systematic approach for data generation for intelligent fault detection and diagnosis in district heating," *Energy*, vol. 307, 2024, Art. no. 132711.
- [45] H. Zhou et al., "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. 35th AAAI Conf. Artif. Intelligence, AAAI 2021, Virtual Conf.*, 2021, pp. 11106–11115.
- [46] C. S. of Engineering, "Data castle bearings dataset." Western Reserve University, Cleveland, OH, USA, Aug. 2021. [Online]. Available: <https://engineering.case.edu/bearingdatacenter>