Original Research Article

SHM Structural Health Monitoring

Structural Health Monitoring
1–25

DOI: 10.1177/14759217251414344
journals.sagepub.com/home/shm
Sage

Semi-Supervised Transfer Graph Representation Learning with Few-Shot Adaptation for gearbox diagnostics under extraneous transient noise

Peng Chen[1,2], Jia Gao[1], Yuhao Wu[1], Changbo He[3], Ge Xin[4], Shuai Fan[5] and Junyu Qi[6]

Abstract

Gearboxes are critical mechanical components widely deployed in industrial applications, where their reliable operation directly impacts system safety and efficiency. However, conventional fault diagnostic approaches face significant challenges when operating under extraneous transient noise conditions, particularly with limited labeled fault samples. These challenges manifest as performance degradation with extremely sparse labeled datasets, vulnerability in pseudo-label generation mechanisms under intense transient noise, and inconsistent feature scale representations due to noise-induced interference. Furthermore, existing methods struggle to maintain diagnostic accuracy when confronted with both data scarcity and transient disturbances, often resulting in compromised model generalization and unreliable fault classification. To address these limitations, this research proposes the Semi-Supervised Transfer Graph Representation Learning with Few-Shot Adaptation (SSTGRL-FSA) framework, featuring three innovative components: a novel pseudo-label reliability enhancement mechanism leveraging systematic knowledge transfer from established source domains, an advanced label transmission and matching strategy exploiting homologous signal patterns across operational domains, and an integrated first-order Markov state probability transition matrix with amplitude-constrained scaling. SSTGRL-FSA significantly advances the field by effectively handling both labeled data scarcity and transient noise interference while enhancing model robustness through sophisticated temporal dependency modeling and stable feature scale maintenance, ultimately providing a more reliable and practical solution for industrial gearbox fault diagnosis under challenging operational conditions.

Keywords

Gearbox, fault diagnosis, semi-supervised learning, transfer learning, graph representation learning, transient noise interference, pseudo-label generation

Introduction

In industrial applications, gearbox systems constitute fundamental components that are integral to the operational efficiency of vehicles, electric motors, and wind turbines.[1–6] These critical mechanisms routinely operate under exceptionally demanding conditions characterized by extreme load variations and high-velocity operations. Such intensive operational parameters subject the constituent components to substantial cyclical stress patterns and thermal variations over extended periods. Furthermore, these sophisticated mechanical systems exhibit inherent vulnerabilities to multiple failure modalities, encompassing progressive

[1]College of Engineering, Shantou University, Shantou, P.R. China
[2]Key Laboratory of Intelligent Manufacturing Technology, Shantou University, Shantou, P.R. China
[3]College of Electrical Engineering and Automation, Anhui University, Hefei, P.R. China
[4]School of Traffic and Transportation, Beijing Jiaotong University, Beijing, P.R. China
[5]School of Mechanical and Electrical Engineering, Chengdu University of Technology, Chengdu, P.R. China
[6]Electronics & Drives, Reutlingen University, Reutlingen, Germany

Corresponding authors:
Changbo He, College of Electrical Engineering and Automation, Anhui University, Hefei 230601, Anhui, P.R. China.
Email: changbh@ahu.edu.cn

Peng Chen, College of Engineering, Shantou University, Shantou 515063, Guangdong, P.R. China.
Emails: pengchen@alu.uestc.edu.cn; dr.pengchen@foxmail.com

wear patterns, crack propagation phenomena, and catastrophic tooth fractures—primarily arising from sustained high-speed operations and dynamic loading conditions. The timely detection of incipient defects thus becomes paramount in preventing cascading secondary damage and averting costly operational interruptions. Additionally, the presence of transient noise phenomena introduces substantial complexity to fault diagnosis procedures, potentially compromising both the operational longevity and reliability of gearbox systems.[2,7–11] Therefore, the development of robust fault diagnosis methodologies emerges as a critical imperative for ensuring sustained equipment reliability.

The convergence of advanced sensor technologies and internet connectivity has catalyzed remarkable progress in data-driven methodological approaches.[10,12–14] These techniques leverage comprehensive operational datasets to identify subtle patterns and anomalies indicative of mechanical degradation. While traditional fault diagnosis methods such as envelope analysis[15–19] and spectral kurtosis[20] have demonstrated effectiveness in specific scenarios, they suffer from significant limitations. These methods require extensive manual feature engineering by domain experts and are highly sensitive to operational variations such as speed fluctuations and load changes. Additionally, traditional methods like wavelet transform,[21] empirical mode decomposition,[22] and Hilbert–Huang transform[23] struggle with transient noise interference, as their feature extraction processes are highly susceptible to noise contamination.

Deep learning, as an advanced computational paradigm, has demonstrated exceptional utility in mechanical fault diagnosis applications.[24–26] Notable developments include Lu et al.'s[27] innovative optimization architecture, which integrates attention mechanisms with chaotic quantum particle swarm optimization and multi-layer denoising autoencoders for precise fault diagnosis in wind turbine planetary gearboxes. Similarly, Cui et al.[28] implemented a dual-stream convolutional neural network architecture to address data imbalance challenges in centrifugal pump rotor fault diagnosis, while Joung et al.[29] developed an integrated Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) framework for bearing fault detection utilizing operational degradation data from injection molding machinery. However, these advanced computational approaches predominantly rely on supervised learning algorithms, necessitating extensive labeled datasets for effective model training.[30] Within industrial contexts, the acquisition of such labeled data presents significant challenges, proving both resource-intensive and time-consuming due to the predominance of unlabeled operational data. Consequently, this fundamental limitation significantly impedes the development and implementation of reliable fault diagnosis systems based exclusively on supervised methodologies.

In response to these constraints, the research community has increasingly gravitated toward semi-supervised learning (SSL) approaches, effectively utilizing both limited labeled data and abundant unlabeled samples during model training.[31] Within SSL-based diagnostic frameworks, specialized neural network architectures have demonstrated remarkable success in gear and bearing fault diagnosis applications.[32–34] Exemplary contributions include Chen et al.'s[35] pseudo-label assisted semi-supervised adversarial learning methodology, which incorporates metric learning enhancement for gearbox fault diagnosis, substantially reducing annotation requirements. Additionally, Huang et al.[36] developed a semi-supervised prototype network incorporating similarity-based pseudo-labeling for wind turbine gearbox fault diagnosis, while Cai et al.[37] introduced a spatio-temporal graph attention contrastive learning framework achieving exceptional diagnostic accuracy with minimal labeled data. Notably, graph neural networks (GNNs) have emerged as a particularly powerful paradigm in data-driven fault diagnosis, effectively capturing system complexity through sophisticated node-edge representations.[38] While GNNs demonstrate significant advantages in modeling structured data and fault propagation patterns, challenges persist in effectively integrating temporal evolution patterns and managing high-dimensional unlabeled datasets within SSL-based diagnostic frameworks.

Moreover, GNN architectures inherently rely on explicit graph structures, which become problematic when node relationships cannot be definitively established in complex mechanical systems, leading to potential reliability issues.[39] Addressing these persistent challenges in SSL, particularly regarding domain shift vulnerability and computational efficiency, researchers have increasingly adopted transfer learning methodologies. These approaches fundamentally reconstruct feature spaces to align source and target distributions through domain-invariant representations. Notable advancements include Yue et al.'s[40] physics-informed dual guidance methodology, which integrates physical envelope harmonic distribution with transfer learning for efficient gear fault classification. Similarly, Luo et al.[41] developed an integrated Convolutional Neural Network and Long Short-Term Memory (CNN-LSTM) transfer learning framework for comprehensive fault diagnosis across power levels in Small Modular Reactors, while Dai et al.[42] proposed an innovative hybrid framework combining optimized decomposition techniques with transformer-based transfer learning. However, despite these significant advances, the diagnostic robustness of existing methodologies remains susceptible to noise interference,[43,44] particularly when impulsive noise transients introduce spurious feature correlations that compromise learned domain-invariant representations during operational inference.

Analysis of current fault diagnosis methods reveals persistent challenges that limit system robustness and reliability, particularly in industrial settings with data constraints and harsh operating environments. The key limitations are:

1. Performance degradation with extremely sparse labeled datasets, where model bias becomes pronounced despite transfer learning and unlabeled data integration, compromising pseudo-label generation and diagnostic accuracy.
2. Vulnerability in pseudo-label generation mechanisms under intense transient noise operating conditions, leading to error cascade effects throughout the learning pipeline and degraded system performance.
3. Inconsistent feature scale representations due to transient noise-induced interference, resulting in model overfitting, reduced generalization capabilities, and compromised optimization dynamics in transfer learning implementations.

In response to these critical challenges, this research introduces an innovative diagnostic framework: Semi-Supervised Transfer Graph Representation Learning with Few-Shot Adaptation (SSTGRL-FSA). This comprehensive methodology represents an advancement in addressing the fundamental challenges of limited labeled data availability under conditions of transient interference, through the strategic integration of SSL principles and transfer learning mechanisms. Specifically designed for planetary gearbox fault diagnosis in harsh operational environments, SSTGRL-FSA synergistically leverages complementary learning paradigms to enhance diagnostic robustness. The framework's principal technical innovations and contributions can be enumerated as follows:

1. A novel SSTGRL-FSA framework that enhances pseudo-label reliability in few-shot learning (FSL) scenarios through systematic knowledge transfer from established source domains, improving diagnostic performance under data-constrained conditions.
2. An advanced label transmission and matching strategy that ensures precise pseudo-label alignment across operational domains by exploiting homologous signal patterns, effectively mitigating noise-contaminated signal effects.
3. Integration of a first-order Markov state probability transition matrix with amplitude-constrained scaling (ACS), providing robust temporal dependency modeling while maintaining stable feature scales and enhanced training efficiency.

The structural organization of the remainder of this paper proceeds in a systematic manner through several interconnected sections. In section "Related work," we thoroughly examine and critically analyze the theoretical foundations underlying three fundamental components: the Dynamic Graph Attention Network (DGAT), which enables adaptive feature learning, and Pseudo-Labeling Algorithm, which utilizes a teacher model to construct pseudo labels. Subsequently, section "The proposed SSTGRL-FSA" presents a detailed exposition of the novel methodology, namely the SSTGRL-FSA, along with its architectural components and theoretical underpinnings. Furthermore, section "Experimental results and comparative analysis" offers a comprehensive presentation of our experimental findings, accompanied by an in-depth comparative analysis with state-of-the-art approaches and ablation studies. Building upon these empirical results, section "Ablation study" systematically investigates the individual contributions of core components through ablation studies, specifically analyzed the impact of transient noise processing and label transfer modules on the framework. In addition, section "In-depth analysis of SSTGRL-FSA performance and optimization strategies" analyzed the sensitivity of hyperparameters and the performance of the model in industrial noise environments. Finally, section "Conclusion" synthesizes our key contributions, draws meaningful conclusions, and delineates promising directions for future research endeavors in this field.

Related work

This section systematically reviews two key theoretical foundations supporting the SSTGRL framework. The review first examines recent developments in DGATs, which effectively capture complex relationships in structured data through adaptive feature learning mechanisms. The theoretical basis of pseudo-labeling algorithms is then analyzed, with particular attention to how teacher-student model architectures leverage knowledge transfer to address scarce labeled data scenarios.

Dynamic Graph Attention Network

Despite the Graph Attention Network's (GAT) notable ability to discern the relative significance of information between adjacent nodes through iterative weight adjustments,[14,16] this architecture exhibits inherent limitations stemming from its static attention mechanism. Most notably, the framework's inability to account for the temporal dynamics of inter-node connections significantly constrains its expressive capacity and representational capabilities, which can be formally demonstrated through the following analysis.

$$e_{ij} = LeakyReLU\left(F_1 W f_i + F_2 W f_j\right) \tag{1}$$

where e_{ij} is the importance of node V_j to node V_i, f_i and f_j are feathers of nodes V_i and V_j, W represents weight matrix, F_1 and F_2 are the shared attention mechanism.

One of the primary constraints inherent in the traditional GAT architecture stems from its global parameter-sharing mechanism, whereby the transformation matrices W, F_1, and F_2 in Equation (1) are universally shared across all nodes and implemented sequentially. As a consequence of this design, these operations can be mathematically reduced to a singular linear transformation layer, thereby resulting in a static attention mechanism with limited adaptive potential. In response to these architectural limitations, the DGAT introduces an advanced aggregation mechanism for central node processing. Specifically, the DGAT employs a refined computational workflow wherein the transformation matrix W is strategically applied following concatenation, while the unified attention mechanism $F = [F_1 \parallel F_2]$ is positioned subsequent to the *Leakyrelu* activation function. This strategic reorganization of the architecture enables the network to achieve genuinely dynamic attention capabilities, thereby facilitating more flexible and context-sensitive feature processing. The comprehensive computational workflow within the DGAT layer is meticulously detailed in Figure 1. Furthermore, the dynamic attention score in DGAT is determined through the following mathematical framework:

$$e_{ij} = F \cdot LeakyReLU\left(W \cdot [f_i \parallel f_j]\right) \tag{2}$$

$$\alpha_{ij} = \mathrm{Softmax}\left(e_{ij}\right) = \frac{\exp\left(e_{ij}\right)}{\sum_{k \in N(i)} \exp\left(e_{ik}\right)} \tag{3}$$

where F represents a shared attention mechanism, and $N(i)$ represents the set of all nodes adjacent to node V_i.

Building upon these architectural innovations, the attention computation process is fundamentally anchored in the distinctive characteristics of individual nodes while methodically incorporating the interconnected relationships with neighboring nodes. This holistic approach substantially augments the model's capacity to capture and represent intricate graph structures. To elucidate this process more precisely, we can examine the feature learning procedure at the l-th layer of DGAT, which can be mathematically formulated as:

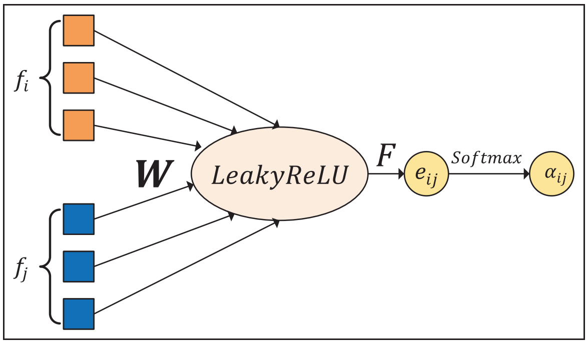


Figure 1. Schematic illustration and computational workflow of the DGAT layer.

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} \alpha_{ij} W^{(l)} h_j^{(l)}\right) \tag{4}$$

where $\sigma(\cdot)$ denotes activation function, $h_i^{(l+1)}$ represents the updated vector after node V_i aggregates the information of layer l, and $W^{(l)}$ is its learnable weight matrix.

Pseudo-labeling algorithm

In the field of SSL methodologies, pseudo-labeling has emerged as a particularly robust and efficacious algorithm, as comprehensively illustrated in Figure 2. This algorithmic approach, which has garnered significant attention in recent years due to its practical applicability and theoretical foundations, operates fundamentally on bifurcated datasets that are systematically partitioned into two distinct categories: the labeled sample set, denoted as $D_l = \{(x_i, y_i)\}_{i=1}^{N_l}$, which encompasses data points with known classifications, and the unlabeled sample set, represented as $D_u = \{x_i\}_{i=1}^{N_u}$, comprising data points whose classifications remain to be determined. Within this framework, the initial phase of the algorithm centers on the development and training of teacher models, whose learning process is exclusively dependent on and constrained to the labeled samples, thereby establishing a foundational basis for subsequent pseudo-labeling operations.

$$f_{\theta_T}^{*} = \min_{\theta_T} \frac{1}{N_l} \sum_{i=1}^{N_l} \mathcal{L}(f_{\theta_T}(x_i), y_i) \tag{5}$$

where θ_T parameterizes the teacher model f_{θ_T}, and $\mathcal{L}$ represents the supervised loss function.

For the neural network architecture, when implementing the Softmax activation function in the output layer, it is typically paired with cross-entropy loss as the supervised loss function. This fundamental

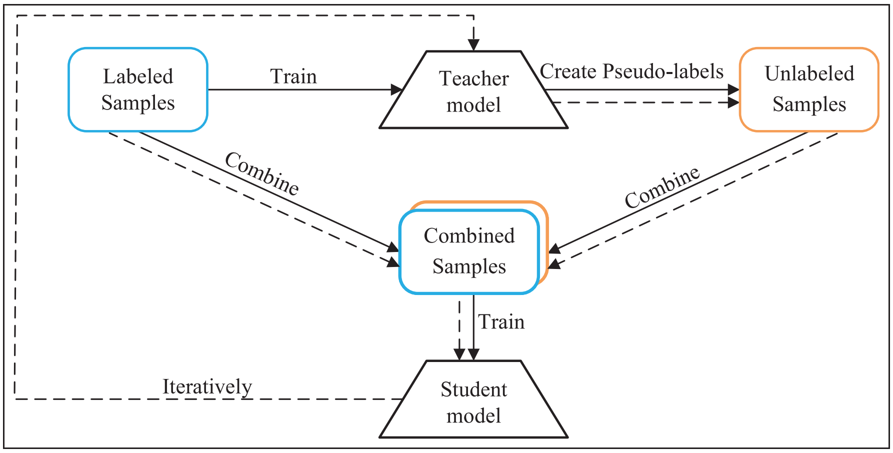


Figure 2. Procedure of the classical pseudo-labeling algorithm.

combination can be mathematically defined through a specific formula, establishing the foundation for the model's learning process.

$$\mathcal{L}(f_{\theta_T}(x_i), y_i) = -\sum_{j=1}^{K} y_{i,j} \log \left(f_{\theta_T}(x_i)_j\right) \tag{6}$$

where K is the number of dimension of network output, $y_{i,j} \in \{0, 1\}$ indicates whether x_i belongs to class j, $f_{\theta_T}(x_i)_j$ is the predicted probability of class j by the teacher model.

Subsequently, once the training phase is completed, the optimized teacher model, denoted as $f^*_{\theta_T}$, assumes a crucial role in generating pseudo-labels for the unlabeled data points. This pedagogical process is initiated through the computation of class probability distributions, which are obtained via forward propagation mechanisms, thereby facilitating the transfer of knowledge from teacher to student model.

$$\begin{gathered} p_i = f^*_{\theta_T}(\mathbf{x}_i) = [p_{i1}, p_{i2}, \ldots, p_{iK}], \\ \text{where } \sum_{j=1}^{K} p_{ij} = 1. \end{gathered} \tag{7}$$

pseudo-labels $\hat{y}_k$ are assigned by selecting the class index corresponding to the maximum Softmax probability:

$$\hat{y}_k = \arg\max_j p_{ij} \tag{8}$$

To ensure the robustness and quality of the learning process, a significant refinement step is introduced through the implementation of a confidence threshold, represented as τ. This threshold serves as a discriminative filter, effectively eliminating predictions that fall below the desired confidence level, thus maintaining the integrity of the pseudo-labeling process.

$$\mathcal{D}_{\text{pseudo}} = \left\{ (\mathbf{x}_i, \hat{y}_k) \mid \max_j p_{kj} \geqslant \tau \right\}_{i=1}^{M} \tag{9}$$

The student model f_{θ_S} is then trained on the augmented dataset $\mathcal{D}_{\text{aug}} = \mathcal{D}_l \cup \mathcal{D}_{\text{pseudo}}$ combining original labeled data and filtered pseudo-labeled examples. The optimization objective minimizes a composite loss function balancing supervised and unsupervised components:

$$\begin{aligned} f_{\theta_S} = \min_{\theta_S} \frac{1}{|D_{\text{aug}}|} \Bigg[& \underbrace{\sum_{(\mathbf{x}_i, y_i) \in D_l} \mathcal{L}(f_{\theta_S}(\mathbf{x}_i), y_i)}_{\text{supervised}} \\ & + \lambda \underbrace{\sum_{(\mathbf{x}_k, y_k) \in D_{\text{pseudo}}} \mathcal{L}(f_{\theta_S}(\mathbf{x}_k), \hat{y}_k)}_{\text{unsupervised}} \Bigg] \end{aligned} \tag{10}$$

where the weighting coefficient λ prioritizes labeled data during initial training phases. Following the successful training of the student model f_{θ_S} through Equation (10), the process enters an iterative phase where the newly trained student model assumes the role of the teacher. This cyclical progression continues as the model undergoes repeated training iterations through Equations (7)–(10), establishing a self-improving feedback loop that enhances the overall model performance.

The proposed SSTGRL-FSA

To effectively address the multifaceted challenges inherent in mechanical signal processing, particularly the persistent issues of non-Gaussian interference patterns—such as transient noise contamination in vibration signals—and the inherent unreliability of one-dimensional data pseudo-labeling mechanisms, this research proposes a novel methodological framework termed SSTGRL-FSA. This innovative approach systematically integrates confidence-based pseudo-labeling techniques with ACS methodologies, thereby establishing a robust foundation for enhanced signal processing capabilities. As systematically illustrated in Figure 3, the proposed framework encompasses three fundamental yet interconnected components: firstly, the meticulous training of a teacher model utilizing strategically denoized data sets; secondly, the implementation of a sophisticated prediction mechanism for generating high-confidence pseudo labels; and finally, the systematic training of a student model through the application of ACS procedures.

Data preprocessing and teacher model architecture implementation

Real-world vibration signals from multi-component transmission systems are frequently contaminated by transient noise from unknown interference sources, exhibiting amplitude anomalies and irregular temporal patterns that compromise statistical consistency and hinder discriminative feature extraction for fault diagnosis. To address this challenge, the proposed method employs a first-order Markov transition matrix to systematically identify and eliminate transient noise components, thereby restoring signal integrity and enhancing feature extraction reliability in vibration-based diagnostic applications.

The proposed methodological framework specifically targets transient noise in both labeled and unlabeled datasets, denoted as $\left\{x_i^{(\text{imp, label})}\right\}_{i=1}^{N}$ and $\left\{x_i^{(\text{imp, unlabel})}\right\}_{i=1}^{M}$, respectively, through the utilization of first-order Markov transition matrix analysis.[18,19] This approach initiates with a fundamental transformation process, wherein continuous signals $x_i|R^{1\times L}$ undergo discretization into a finite set of states. Throughout this process, each value R within the signal is systematically assigned an interval number j, which subsequently becomes the signal's state S_i at that particular temporal point. This transformation effectively facilitates the conversion of continuous states into discrete state sequences, governed by a specific mathematical formulation:

$$S_i = \begin{cases} j, & R[k] \in S_j \\ n_{\text{bin}} - 1, & R[k] = \max(x_i) \end{cases} \tag{11}$$

where $k \in (1, L)$, n_{bin} is a predefined number of states. Notably, the size of n_{bin} will affect the granularity of analysis as it is closely related to the density of discrete states. Following the division of x_i into n_{bin} discrete states, each state S_j is characterized according to the formula given below:

$$\begin{aligned} S_j = \{R[k] \mid R_{\min} + (j-1)\Delta \leqslant R[k] \\ < R_{\min} + j\Delta\},\ j = 1, 2, \ldots, n_{\text{bin}} \end{aligned} \tag{12}$$

$$\Delta = \frac{\max(x_i) - \min(x_i)}{n_{\text{bin}}} \tag{13}$$

where Δ represents the width of each interval. Subsequently, the state sequence $S = \{S_1, S_2, \ldots, S_L\}$ undergoes empirical frequency statistics to quantify inter-state transitions. This process generates a Markov count matrix $C \in \mathbb{R}^{N\times N}$, formally defined as:

$$C(i,j) = \sum_{k=1}^{L-1} \mathbb{I}(S_k = i \wedge S_{k+1} = j) \tag{14}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. To derive valid stochastic transitions, C is normalized row-wise to obtain the Markov Transition Probability Matrix P:

$$P(i,j) = \frac{C(i,j)}{\sum_{k=1}^{N} C(i,k)},\quad \forall i,j \in \{1, 2, \ldots, N\} \tag{15}$$

$$P = \operatorname{diag}(C1)^{-1}C \tag{16}$$

where the transition matrix $P(i,j)$ emerges as a crucial analytical tool, where each element encodes the conditional probability $P(S_{t+1} = j \mid S_t = i)$, effectively characterizing the system's dynamic evolution patterns. Under normal operating conditions, these transition patterns typically manifest as concentrated distributions along the matrix diagonal, reflecting the natural tendency for state persistence within predetermined discrete boundaries. However, the presence of transient interference introduces distinctive anomalous off-diagonal transitions, disrupting the coherent evolution of states. Consequently, the transition matrix P serves as a powerful discriminative tool for identifying and precisely localizing transient disturbances through systematic deviation analysis from expected transition patterns.

$$x_i[k] = \begin{cases} 0, & \text{where } P_{ij} \geqslant t_p \\ x_i[k], & \text{otherwise} \end{cases} \tag{17}$$

To localize transient interference segments, the preset threshold t_p is applied to the transition probabilities P_{ij}. Regions where P_{ij} surpasses t_p are flagged as exhibiting anomalous state transitions, indicative of transient noise contamination. These identified regions

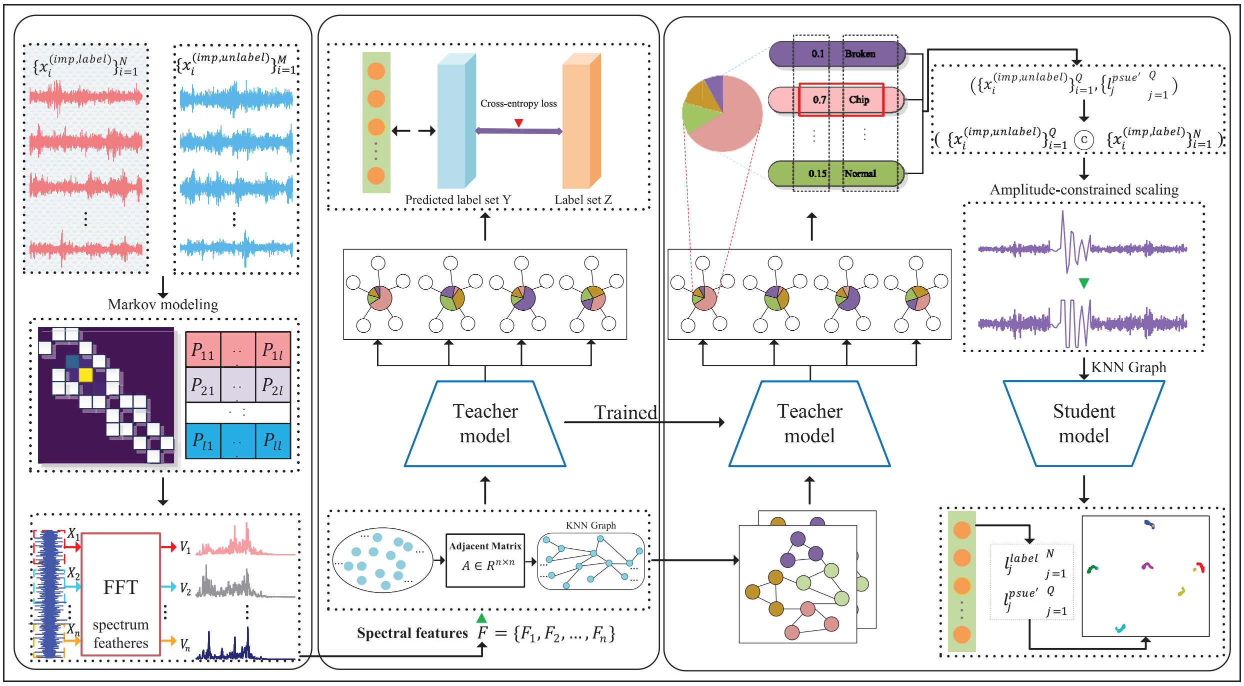


Figure 3. The proposed SSTGRL-FSA.
SSTGRL-FSA: Semi-Supervised Transfer Graph Representation Learning with Few-Shot Adaptation.

within the originally impaired labeled and unlabeled datasets ($\{x_i^{(\text{imp, label})}\}_{i=1}^N$, $\{x_i^{(\text{imp, unlabel})}\}_{i=1}^M$) are then processed by the supporting transient noise removal module. The module suppresses the interference by setting the corresponding signal samples to zero:

Following the successful denoising process, the methodology progresses to a crucial transformation phase where the time-series data undergoes conversion into the frequency domain through Fast Fourier Transform (FFT). This transformation yields spectral representations $f_i \in \mathbb{R}^d$ for each sample, establishing the foundation for subsequent graph topology construction. The topological framework is developed through the computation of pairwise Euclidean distances between spectral features, culminating in the formation of a k-nearest neighbor (KNN) graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

$$\mathcal{E} = \left\{ (i,j) \;\middle|\; j \in \underset{j' \in \mathcal{V}\setminus\{i\}}{\operatorname{argtopk}_k} \left(-\|f_i - f_{j'}\|_2 \right) \right\} \cup \{(i,i) | i \in \mathcal{V}\} \tag{18}$$

where the architectural framework culminates in the formal definition of adjacency relationships, wherein argtopk_k operates as a selective mechanism for identifying indices of the k smallest Euclidean distances, with explicit inclusion of self-loops (i, i). Within this structure, nodes $\mathcal{V}$ establish correspondence with individual samples, while edges $\mathcal{E}$ facilitate connections between nodes and their k most spectrally similar neighbors, including self-referential connections. This sophisticated graph structure, in conjunction with the spectral node features f_i, serves as the foundational input for the DGAT, which employs an advanced dynamic attention mechanism as defined in Equation (2). After obtaining the denoized labeled dataset, the training process for the teacher model, denoted as $f^*_{\theta_T}$, is carried out by employing the specific methodologies outlined in Equations (5) and (6). These equations provide a structured framework for guiding the optimization process, ensuring the model learns effectively from the refined dataset. The application of such methodical approaches not only facilitates the reduction of noise-induced discrepancies in the labeled data but also enhances the overall robustness and accuracy of the teacher model.

Pseudo-label generation and ACS methodology

Initially, the unlabeled interference-free samples, denoted as $\{x_i^{(\text{free}-\text{imp, unlabel})}\}_{i=1}^M$, undergo spectral transformation through FFT. These frequency-domain representations are subsequently organized according to the formulation presented in Equation (18), followed by classification using the pre-trained teacher model. The resulting output is then processed through a SoftMax function, which can be mathematically expressed as follows.

$$\beta_c = SoftMax(f_t(\mathrm{FFT}(x_j^{(\mathrm{free-imp,\,unlabel})}))) \tag{19}$$

The SoftMax transformation serves as a fundamental component in this framework, facilitating the conversion of input vectors into normalized probability distributions. This operation effectively maps each element to the interval [0,1] while ensuring that the output components sum to unity, thereby enabling robust probabilistic interpretation. The resulting value β_c quantifies classification confidence, representing the estimated probability of class membership. Of particular importance is the selection criterion β, a tunable hyper-parameter that demands meticulous optimization. Within our framework, pseudo-labels $\{l_j^{\mathrm{psue}}\}_{j=1}^{M}$ demonstrating confidence scores $\beta_c \geqslant 0.95$ are considered sufficiently reliable to serve as surrogate ground-truth annotations. It is noteworthy that the denoized samples $\{x_i^{(\mathrm{free-imp,\,unlabel})}\}_{i=1}^{M}$ originate from their noisy counterparts $\{x_i^{(\mathrm{imp,\,unlabel})}\}_{i=1}^{M}$ post-interference suppression, necessitating identical labeling assignments $\{l_j^{\mathrm{psue}}\}_{j=1}^{M}$ to maintain data integrity throughout the noise elimination pipeline.

To ensure the robustness of the training process, samples with insufficient pseudo-label confidence are excluded from subsequent training stages through a threshold evaluation mechanism. This quality control measure can be mathematically represented by establishing a threshold β, where the evaluation operation is formally defined.

$$\mathcal{X}^{\mathrm{ref}} = \{x_i \mid s_i \geqslant \beta\},\ \mathcal{L}^{\mathrm{ref}} = \{l_i \mid s_i \geqslant \beta\} \tag{20}$$

where s_i represents the probability score of the predicted category for the i-th sample, $\mathcal{X}^{\mathrm{ref}} = \{x_i^{(\mathrm{imp,\,unlabel})}\}_{i=1}^{Q}$ and $\mathcal{L}^{\mathrm{ref}} = \{l_j^{\mathrm{psue}'}\}_{j=1}^{Q}$ represent refined sample set and refined pseudo label set, respectively.

In parallel with pseudo-label generation, both interference-contaminated labeled data $\{x_i^{(\mathrm{imp,\,label})}\}_{i=1}^{N}$ and their unlabeled counterparts $\{x_i^{(\mathrm{imp,\,unlabel})}\}_{i=1}^{Q}$ undergo ACS to the [0,1] interval. This crucial preprocessing step establishes uniform data characteristics essential for analytical consistency. As illustrated in Figure 4, the signal transformation pipeline reveals distinct stages: Figure 4(a) showcases raw signals with transient disturbances, while Figure 4(c) demonstrates their purified versions following interference excision.

Further analysis of conventional scaling outcomes, as depicted in Figure 4(b) and (d), reveals confinement to (0,1) ranges without effective noise suppression. This limitation becomes particularly problematic when transient interference dominates, resulting in the compression of vibration signatures to approximately 0.42, thereby significantly compromising the extraction of gear vibration features.

To address these limitations, we introduce ACS, an innovative approach that constrains vibration amplitudes within $[-A_{\max}, A_{\max}]$, where $A_{\max}$ represents a carefully tunable hyperparameter. The mathematical framework employs sophisticated clipping functions to achieve this constraint.

$$\begin{gathered} x_i'[j] = \mathrm{clip}(x_i[j],\ -A_{\max}, A_{\max}), \\ \text{where}\,\mathrm{clip}(z, a, b) = \max(a, \min(b, z)). \end{gathered} \tag{21}$$

with specialized initialization:

$$x_i'[0] = A_{\max} \tag{22}$$

$$x_i'[1] = -A_{\max} \tag{23}$$

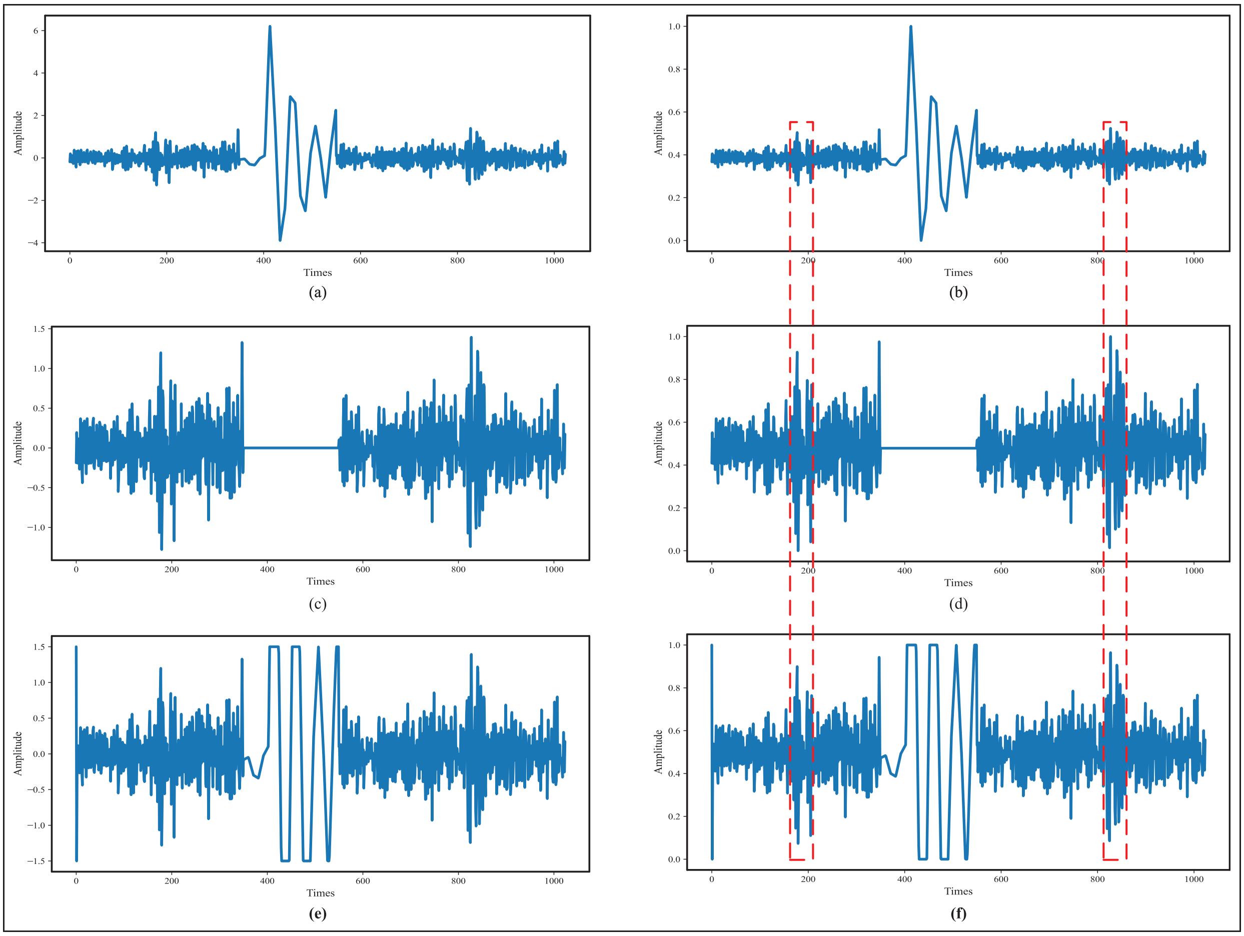
The implementation spans $j \in \{0, 1, \ldots, l-1\}$, with l representing the sample length. This formulation ensures that category-specific vibration features maintain proportional scaling post-normalization. As evidenced by Figure 4(e) and (f), ACS successfully contains transient disturbances within the specified $[-A_{\max}, A_{\max}]$ range while preserving uncontaminated vibration characteristics.

The limitations of traditional scaling approaches become particularly apparent when processing noise-polluted vibration signals, as they often result in inconsistent amplitude scaling within categories, thereby hindering accurate feature identification. Our innovative pre-normalization amplitude bounding approach, implemented through Equations (20)–(23), ensures uniform feature scaling even for minimally expressed vibrations.

Comprehensive visual validation of this processing chain is provided in Figure 4, demonstrating the superior capability of our method in retaining distinct characteristic signals. It is worth noting that ACS is intentionally excluded from the teacher model training phase, as the transient-free training data inherently requires no amplitude constraints, and the application of ACS might potentially compromise critical vibration signatures.

Student model optimization and domain adaptation

During the terminal processing phase, the amplitude-limited normalized signals, formally represented as $\mathcal{X}^{\mathrm{nor}} = \{X_i^{\mathrm{nor}}\}_{i=1}^{N+Q}$, are systematically derived through the integration of both labeled impulse signals $\{x_i^{(\mathrm{imp,\,label})}\}_{i=1}^{N}$ and their unlabeled counterparts



(a) (b) (c) (d) (e) (f)

Figure 4. (a) Original sample containing transient interference; (b) Samples produced via the application of traditional normalization to the original signal, with amplitude constrained to the [0, 1] range; (c) Processed sample after removal of transient interference from original signal; (d) Samples generated through traditional normalization applied to signal (c), confined to the [0, 1] amplitude range; (e) Samples generated by clipping the original signal; (f) Samples produced by amplitude-limited normalization of signal (e).

$\{x_i^{(\mathrm{imp,\,unlabel})}\}_{i=1}^{Q}$. Subsequently, these signals undergo FFT processing, whereby they are transformed into a comprehensive node feature matrix essential for graph construction. Within this framework, each individual signal $X_i^{\mathrm{nor}} \in \mathbb{R}^d$ manifests as a distinctive feature vector corresponding to node v_i within the established graph topology.

Following the feature extraction process, the graph representation $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is methodically established through KNNs association, as rigorously formulated in Equation (18). This meticulously constructed graph structure subsequently serves as the primary input to the Dynamic Graph Attention Network (DGAT). In this context, the student model $f_s : \mathbb{R}^{|\mathcal{V}| \times d} \to \mathbb{R}^C$ operates within a supervised learning paradigm, wherein the supervised loss function is specifically defined over the labeled node subset $\mathcal{V}_L \subseteq \mathcal{V}$. $y_k^{(c)} \in \{0, 1\}$ represents the one-hot encoded ground truth label for node v_k and class c, with $y_k^{(c)} = 1$ indicating that node v_k belongs to class c and 0 otherwise.

$$\mathcal{L}_{\mathrm{sup}} = -\frac{1}{|\mathcal{V}_L|} \sum_{v_k \in \mathcal{V}_L} \sum_{c=1}^{C} y_k^{(c)} \log\left(\sigma_c\left(f_s\left(X_k^{\mathrm{nor}}\right)\right)\right) \tag{24}$$

where $\sigma_c(\cdot)$ denotes the softmax activation function for class c, $\sigma_c(\cdot)$ also represents the softmax activation function for class c defined as $\sigma_c(z) = \frac{\exp(z_c)}{\sum_{j=1}^{C} \exp(z_j)}$, where z is the vector of logits and C is the total number of classes, $f_s^{(c)}\left(X_k^{\mathrm{nor}}\right) = \sigma_c\left(f_s\left(X_k^{\mathrm{nor}}\right)\right)$ represents the class-specific probability distribution, $\mathcal{V}_L$ contains both originally labeled samples and validated pseudo-labeled samples and $|\mathcal{V}_L|$ denotes the cardinality of the labeled node set.

In accordance with established pseudo-labeling methodologies, the iterative retraining protocol for the

Algorithm 1. The proposed algorithm of SSTGRL-FSA.

```
Require: Labeled data $\mathcal{D}_l = \{x_i^{(\text{imp, label})}, y_i\}_{i=1}^{N}$; Unlabeled data $\mathcal{D}_u = \{x_j^{(\text{imp, unlabel})}\}_{j=1}^{M}$
Ensure: Robust student model $\Theta_s$
  Stage 1: Data Preprocessing and Teacher Model Architecture Implementation
for each $x \in \mathcal{D}_l \cup \mathcal{D}_u$ do
    Discretize signal: $S \leftarrow \text{StateSequence}(x, n_{\text{bin}})$    ▷ Equation (11)–(14)
    Compute Markov matrix: $C(i,j) \leftarrow \sum_{k=1}^{L-1} \mathbb{I}(S_k = i \wedge S_{k+1} = j)$
    Normalize: $P \leftarrow \text{diag}(C1)^{-1} C$    ▷ Equation (16)
    Detect noise: $\text{TPP} \leftarrow \{k \ \exists\ P_{ij}(k) \geqslant t_p\}$
    Denoise: $x^{(\text{free}-\text{imp})} \leftarrow x \cdot \mathbb{I}(\text{TPP}^c)$    ▷ Zero-out TPP regions
end for
Extract spectral features: $f_i \leftarrow \text{FFT}(x_i^{(\text{free}-\text{imp})})$
Build graph: $\mathcal{G} \leftarrow \text{KNN} - \text{Graph}(\{f_i\}, k)$    ▷ Equation (18)
Train DGAT teacher: $\Theta_t \leftarrow \arg\min_\theta \sum_{i=1}^{N} \mathcal{L}_{\text{CE}}(f_\theta(\mathcal{G}), y_i)$    ▷ Equation (5)
  Stage 2: Pseudo-Labeling & ACS
Predict: $\beta_j, \hat{y}_j \leftarrow \text{Softmax}(\Theta_t(\text{FFT}(x_j^{(\text{free-imp, unlabel})})))$    ▷ Equation (19)
Filter: $\mathcal{X}^{\text{ref}} \leftarrow \{x_j^{(\text{imp, unlabel})} \mid \beta_j \geqslant 0.95\}$, $\mathcal{L}^{ref} \leftarrow \{\hat{y}_j \mid \beta_j \geqslant 0.95\}$    ▷ Equation (20)
Apply ACS:
for each $x \in \mathcal{D}_l \cup \mathcal{X}^{\text{ref}}$ do
    $x'[0] \leftarrow A_{\max}, x'[1] \leftarrow -A_{\max}$    ▷ Equations (22)–(23)
    $x'[j] \leftarrow \text{clip}(x[j], -A_{\max}, A_{\max}) \ \forall j \geqslant 2$    ▷ Equation (21)
    $x^{\text{nor}} \leftarrow \text{Normalize}(x', [0, 1])$    ▷ Figure 4(f)
end for
  Stage 3: Student Model Optimization and Domain Adaptation
Extract features: $f_k^{\text{nor}} \leftarrow \text{FFT}(x_k^{\text{nor}})$
Build graph: $\mathcal{G}' \leftarrow \text{KNN-Graph}(\{f_k^{\text{nor}}\}, k)$
Define node sets: $\mathcal{V}_L \leftarrow \{\text{labeled}\} \cup \{\text{pseudo-labeled}\}$
Train DGAT student: $\Theta_s \leftarrow \arg\min_\phi \frac{1}{|\mathcal{V}_L|} \sum_{v_k \in \mathcal{V}_L} \mathcal{L}_{\text{CE}}(f_\phi(\mathcal{G}'), y_k)$    ▷ Equations (10) and (24)
```

SSTGRL-FSA: Semi-Supervised Transfer Graph Representation Learning with Few-Shot Adaptation.

student model f_s adheres to the formulations presented in Equations (19)–(24). Throughout this recursive process, the optimized student model f_s^* from each iteration systematically transitions into the role of teacher model f_t^* for the subsequent cycle. Although the classification framework fundamentally relies on this iteratively refined student model f_s^*, our implementation deliberately employs the non-iterative variant of the student model for test sample prediction. This strategic decision is predominantly influenced by two crucial considerations: firstly, the substantial reduction in computational complexity, and secondly, the demonstrated efficacy of our SSTGRL-FSA as evidenced through comprehensive empirical evaluations.

For a more detailed exposition of the aforementioned methodology, we present a structured algorithmic formulation in Algorithm 1, which encapsulates the essential computational steps and operational workflow of our proposed framework.

Experimental results and comparative analysis

The effectiveness of the proposed methodology was validated through detailed analyses of two complementary case studies, shown in Section "Quantitative evaluation and comparison of diagnostic results", each providing distinct insights into both the methodology's performance capabilities and its practical applications across different scenarios. These selected studies not only demonstrate the framework's operational reliability but also illuminate its versatility in addressing diverse analytical challenges.

Case study I

Experimental setup and data acquisition protocol

The comprehensive gearbox dataset, obtained through measurements from an advanced gear system, provides a representation of various operational states and fault classifications. The experimental apparatus, illustrated in Figure 5, comprises several critical components, including a tachometer, drive motor, torque sensor, dual-stage parallel gear assembly, loading gear units, and load motor. The accelerometer's strategic placement on an isolated disk represents a key design element, with its specific mounting configuration elaborated in the figure's inset. To ensure the capture of high-fidelity dynamic behavior, data acquisition was conducted at a sampling frequency of 12.8 kHz.

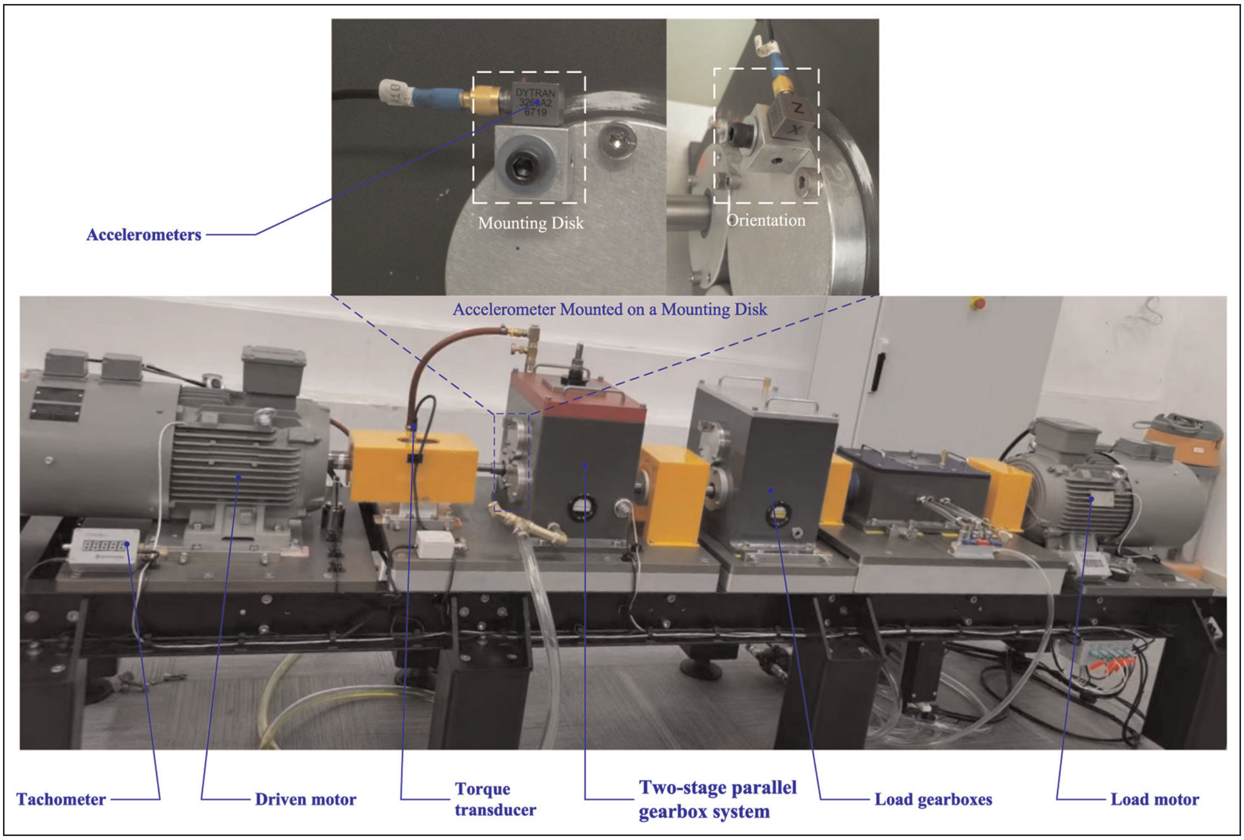


Figure 5. Experimental test rig of gear transmission system.

Furthermore, the operational parameters encompass rotational velocities ranging from 1600 to 2400 r/min, thereby facilitating a systematic investigation across multiple working regimes. In addition to normal operational conditions, the dataset incorporates five distinct gear defect categories, as demonstrated in Figure 6: missing tooth (complete absence), chipped tooth (partial fracture), root damage (fundamental structural failure), surface deterioration (degradation of contact surfaces), and eccentricity (misalignment between geometric and rotational centers).

The intricate gear engagement relationships are depicted in Figure 7(a), whereas Figure 7(b) elucidates the parallel gearbox's internal configuration, wherein dashed borders distinctly indicate the presence of defective components. To facilitate comprehensive fault analysis, vibration measurements were systematically acquired along the accelerometer's x-axis during steady-state operation at 1600 rpm. Each operational condition, including the baseline healthy state, comprises 768,000 data points collected over continuous 60-s intervals. Consequently, this extensive data repository serves as a robust foundation for the development and validation of diagnostic algorithms, thereby enabling systematic investigation of diverse operational scenarios and fault modes under controlled laboratory conditions.

The experimental methodology encompasses the construction and thorough validation of a comprehensive dataset. Specifically, the training set incorporates 300 samples per health state, culminating in 1800 samples across six distinct health conditions, with each sample containing inherent transient noise. During the training phase, the dataset undergoes systematic partitioning into labeled and unlabeled samples according to various predetermined ratios, with ten nodes allocated for validation purposes. In the subsequent performance evaluation phase, 450 independent samples, entirely distinct from the training dataset, are utilized, comprising 75 samples per health state, with each sample containing 2048 nodes. The detailed architectural specifications of the implemented DGAT model are comprehensively documented in Table 1.

In order to ensure the rigor and reproducibility of statistics, a comprehensive experimental plan was implemented, including eight repeated trials of all methodological methods. The architecture configuration and hyperparameter specifications of all GNNs are standardized for fair comparison. The training program consists of two different stages: the teacher model training stage involves 200 iterations, followed by the fine-tuning stage, which continues for another 200 iterations, both utilizing the Adam optimization

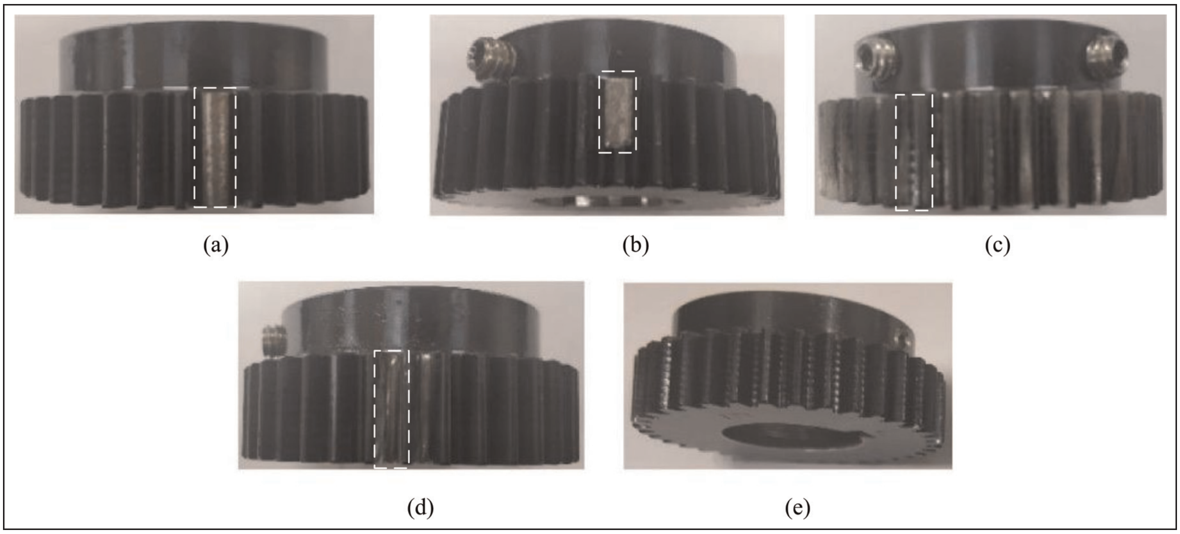


Figure 6. (a) Miss (missing tooth), (b) Chipped (cracked teeth), (c) Surface (wear on gear surface), (d) Root (crack at tooth root), (e) Eccentric (misaligned geometric and rotational centers).

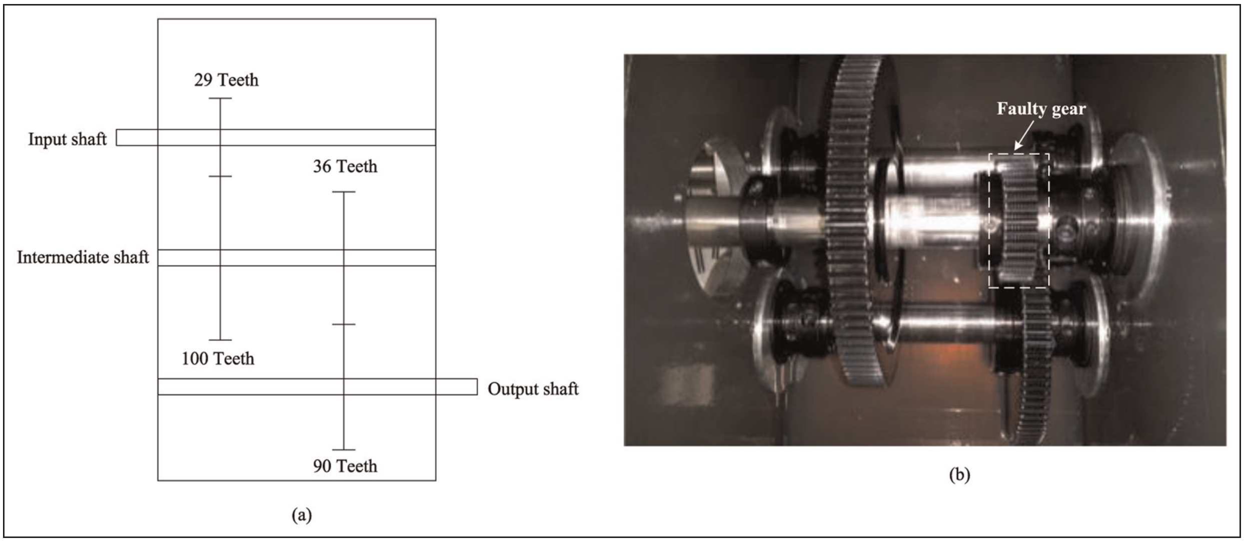


Figure 7. (a) and (b) are internal configuration of parallel gearbox system.

Table 1. Structure of the DGAT model.

Layer	Input channels	Output channels	Parameters
GATv2Conv1	feature	1024	heads = 4
Linear1	1024	1024	—
BatchNorm1	1024	1024	—
GATv2Conv2	1024	1024	heads = 4
Linear2	1024	1024	—
BatchNorm2	1024	1024	—
Fully Connected Layer (FCL) 1	1024	512	Act: ReLU, inplace = True
Dropout1	512	512	$p = 0.2$
FCL2	512	out_channel	—

DGAT: Dynamic Graph Attention Network.

Table 2. Comparative analysis of classification accuracy for case study I.

	Rate of labeled samples		
Model	1%	5%	10%
HyperFast	59.56 ± 5.46	90.52 ± 2.47	93.55 ± 3.01
TabPFN	56.67 ± 2.52	89.78 ± 2.40	91.63 ± 1.81
GRL-CSG	65.56 ± 4.44	93.33 ± 1.98	95.47 ± 1.12
Meta-SGC	58.44 ± 3.81	90.77 ± 2.11	92.13 ± 3.24
DGAT	52.67 ± 10.24	86.67 ± 3.36	92.67 ± 2.91
GraphSAGE	62.44 ± 7.89	86.22 ± 7.07	54.89 ± 6.21
CARDensenet	53.78 ± 8.32	88.67 ± 3.11	89.56 ± 3.46
AS-DWResnet	54.89 ± 6.21	87.11 ± 2.86	92.89 ± 2.91
Ours	83.71 ± 4.21	95.78 ± 1.56	97.41 ± 0.84

GRL-CSG: Graph representation learning and component space graph; DGAT: Dynamic Graph Attention Network; Meta-SGC: Meta Learning and Simplified Graph Convolutional.

algorithm. The learning rate was initialized to 0.01 and then modulated through a gradual decay plan, resulting in a decrease of 0.1 times in both the 100th and 150th iterations. In addition, the neighborhood depth parameter k is determined to be 3, the threshold of the Markov probability transition matrix t_p is set to 10%, the label threshold *beta* is set to 0.95, and the amplitude limit value $A_{\max}$ is set to 1 according to the data characteristics. For validation purposes, the test dataset was constructed by randomly selecting instances from each category, ensuring complete independence from the training corpus.

Quantitative evaluation and comparison of diagnostic results

To evaluate the effectiveness of the proposed SSTGRL-FSA, an extensive and comprehensive comparative experiment was designed and conducted, encompassing both conventional approaches and state-of-the-art diagnostic methodologies. Within this robust experimental framework, the comparative methods encompass well-established traditional approaches, including HyperFast,[45] which has been specifically optimized for rapid tabular data classification, and TabPFN,[46] which innovatively leverages causal inference mechanisms for prior data fitting. Additionally, advanced GNNs were represented through the inclusion of the DGAT[47] and the graph sample and aggregate (GraphSAGE)[48] methodologies. Meanwhile, a graph comparison learning framework based on graph representation learning and component space graph (GRL-CSG) was also introduced,[49] as well as one bearing fault diagnosis method using semi-supervised meta learning and simplified graph convolutional neural networks (Meta-SGC).[50] Moreover, the temporal dimension was addressed through the incorporation of two contemporary time series classification models, namely the causal dilated convolution-based residual DenseNet with channel attention (CARDenseNet)[51] and the adaptive symmetric loss in dynamic wide-kernel residual network (AS-DWResNet).[52]

The experimental protocol was carefully designed to evaluate model performance across varying data availability scenarios, wherein all methodologies were subjected to training regimens utilizing few-shot labeled samples at three distinct labeling ratios: 1%, 5%, and 10%, respectively. Subsequently, the trained models were rigorously validated against a comprehensive test set comprising 450 samples, ensuring robust performance evaluation. It is noteworthy that, within this experimental framework, the semi-supervised GNNs were permitted to leverage additional unlabeled samples from the training set to enhance their learning capabilities. The resultant performance analysis reveals compelling and statistically significant trends across multiple evaluation metrics, as illustrated in both Table 2 and Figure 8, which present the classification accuracy results obtained under varying labeled sample rates per health state category. For enhanced clarity and interpretation, values demonstrating superior performance metrics across different methodological approaches under specific experimental conditions are prominently highlighted in bold. While Table 2 provides precise numerical data accompanied by corresponding standard deviations, these statistical variations are visually represented through error bars in Figure 8, thereby offering complementary perspectives on model performance stability and reliability. Furthermore, to facilitate a more nuanced and comprehensive assessment of the model's diagnostic capabilities, Table 3 presents an extensive evaluation framework incorporating multiple performance indicators, including Precision, Recall, F1 score metrics, and computational efficiency metrics, specifically focusing

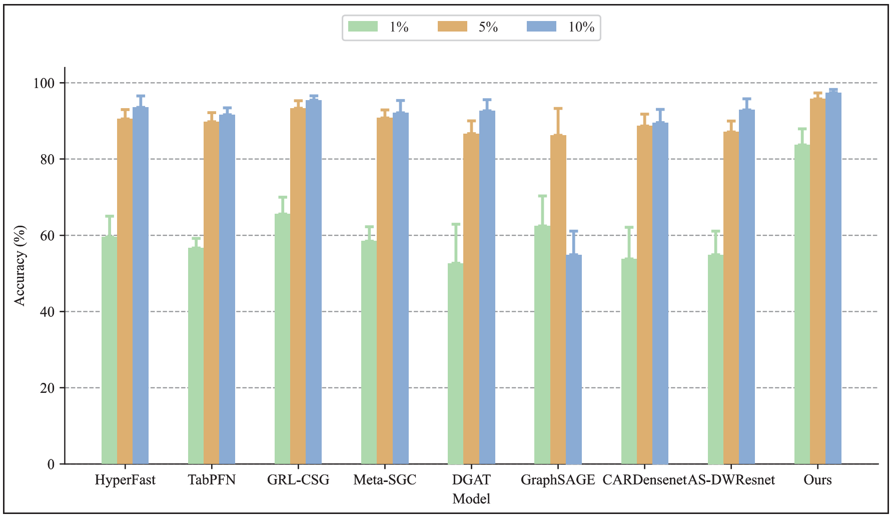


Figure 8. Evaluation of classification accuracy (%) across various model architectures in case study I, where experiments were conducted using 1%, 5%, and 10% labeled training samples per fault category.
Error bars represent the standard deviation based on $n = 8$ independent trials.

Table 3. Comparative analysis results of key metrics (%) and total running times for case study I.

	10% labeled samples			
Method	Precision	Recall	F1 score	Time (s)
HyperFast	94.60 ± 2.04	93.12 ± 3.01	93.60 ± 2.98	8.46
TabPFN	93.23 ± 1.46	91.87 ± 1.81	91.25 ± 2.02	7.48
GRL-CSG	93.51 ± 1.70	91.99 ± 2.61	92.51 ± 2.61	10.32
Meta-SGC	90.76 ± 3.05	89.09 ± 2.05	89.87 ± 3.22	9.98
DGAT	93.36 ± 1.51	92.49 ± 1.34	92.96 ± 1.38	11.14
GraphSAGE	93.51 ± 1.70	90.51 ± 3.12	90.99 ± 2.86	9.68
CARDensenet	91.56 ± 1.01	90.22 ± 1.41	90.84 ± 1.36	8.52
AS-DWResnet	92.75 ± 1.32	91.79 ± 1.77	91.48 ± 1.74	9.29
Ours	97.61 ± 0.68	97.09 ± 0.69	97.42 ± 0.82	13.02

GRL-CSG: Graph representation learning and component space graph; DGAT: Dynamic graph attention network; Meta-SGC: Meta learning and simplified graph convolutional.

on scenarios where 10% labeled samples per health state were utilized during the training phase. The analytical framework is further enhanced through the inclusion of Figure 9, which demonstrates T-SNE dimensionality reduction visualization, and Figure 10, which presents detailed confusion matrix analysis.

A detailed examination of the efficiency metrics presented in Table 3 reveals that SSTGRL-FSA, while incurring moderate runtime increases attributable to its sophisticated two-stage knowledge distillation process, demonstrates a highly favorable performance-to-efficiency ratio. The comprehensive temporal analysis indicates that the proposed methodology SSTGRL-FSA requires 13.02 s for execution—representing an increase of 5.54 s compared to the fastest benchmark TabPFN, and an elevation of 4.56 s relative to HyperFast. However, this additional computational investment yields disproportionate performance dividends, as evidenced by substantial improvements across key performance metrics. These improvements include a remarkable 3.01% absolute precision enhancement over HyperFast, a significant 3.97% recall improvement, and a notable 3.82% F1-score elevation. Furthermore, when considering the temporal-cost-to-performance-gain ratio—achieving a 3.82% F1 score improvement at the expense of 4.56 s of additional processing time compared to the nearest high-performing alternative—the methodology demonstrates compelling practical viability in industrial diagnostic applications where diagnostic accuracy is

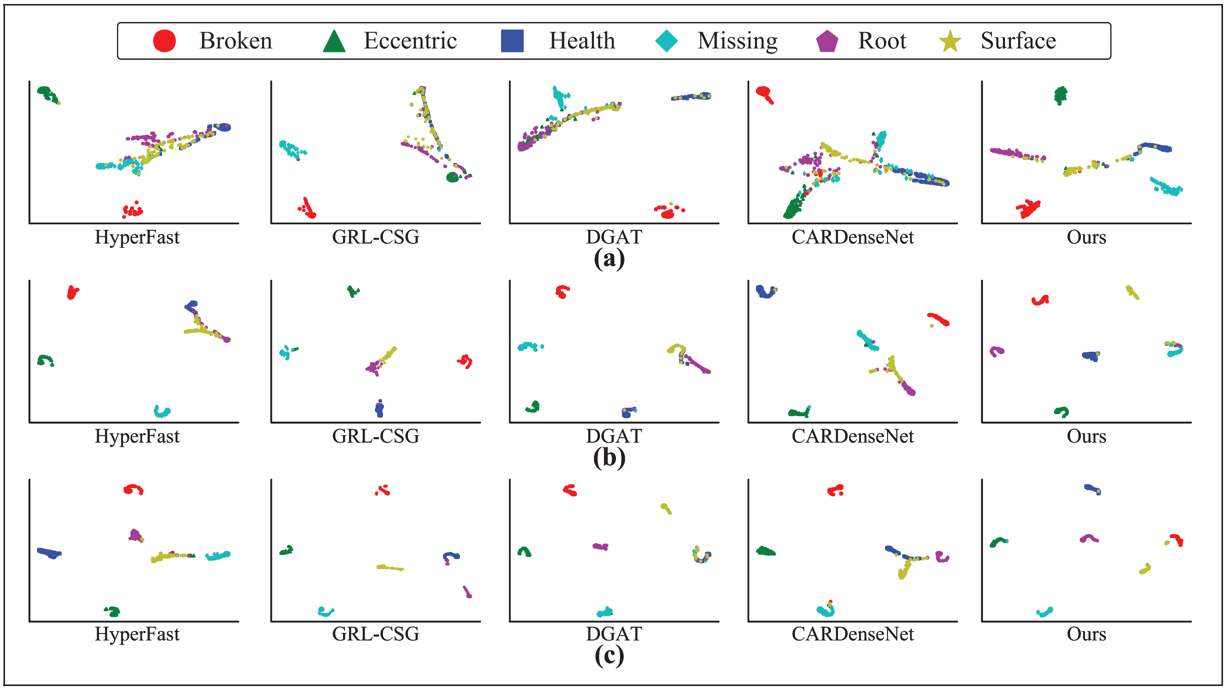


Figure 9. The T-Distributed Stochastic Neighbor Embedding (T-SNE) visualization of feature distributions for various methods in case study I: (a), (b), (c) represent models trained with 1%, 5% and 10% labeled samples per class, respectively.

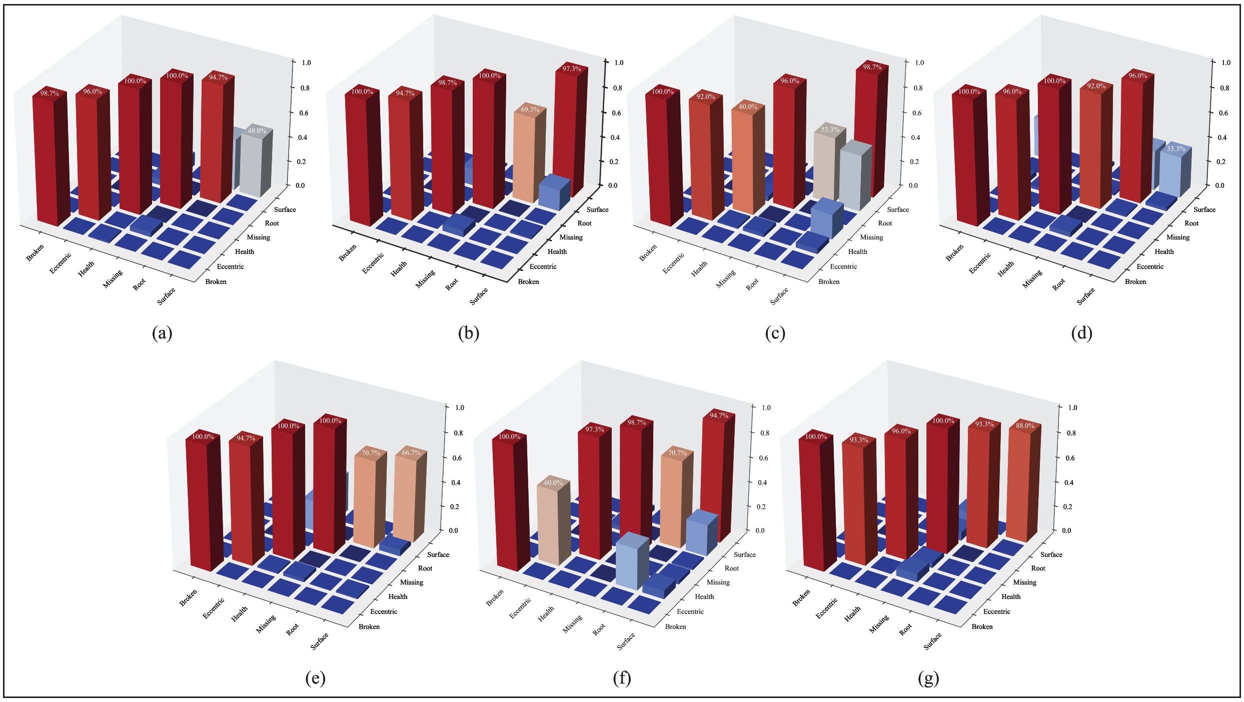


Figure 10. Comparative analysis of confusion matrices for selected methods in case study I (5% labeled samples per class): (a) HyperFast, (b) GRL-CSG, (c) DGAT, (d) GraphSAGE, (e) CARDensenet, (f) AS-DWResnet, (g) Ours.
GRL-CSG: Graph representation learning and component space graph; DGAT: Dynamic Graph Attention Network.

paramount. This trade-off becomes particularly justified in safety-critical gearbox monitoring systems, where marginal increases in computational latency are vastly outweighed by the substantial reduction in false negatives and false positives that could otherwise lead to catastrophic equipment failures or unnecessary maintenance interventions.

Throughout the experimental evaluation, the proposed model consistently demonstrates exceptional performance across all labeled sample ratios, achieving remarkable accuracy rates that significantly surpass existing benchmarks. Specifically, the model attains accuracy rates of 83.71%, 95.78%, and 97.41% under conditions of 1%, 5%, and 10% labeled samples, respectively. Although contemporary approaches such as HyperFast and TabPFN exhibit gradual performance improvements with increased sample availability, their performance remains substantially inferior under few-shot conditions. At the critical 1% sample ratio, SSTGRL-FSA achieves a commanding 18.15% absolute advantage over the nearest competitor, GRL-CSG, which achieves 65.56% accuracy. Moreover, the systematic pattern of performance gaps across increasing sample sizes—from 18.15% at 1% to 2.45% at 5%, and further narrowing to 1.94% at 10%—further underscores SSTGRL-FSA's unique efficiency in extracting maximum value from minimal supervisory signals. The consistent outperformance over specialized graph-based methods like GRL-CSG and Meta-SGC particularly validates our innovative integration of Markov denoising, ACS, and dual-graph adaptive transfer mechanisms in addressing the fundamental challenges of few-shot fault diagnosis. The sophisticated T-SNE dimensionality reduction analysis, as visualized in Figure 9, provides additional evidence of SSTGRL-FSA's enhanced feature discrimination capabilities, demonstrating superior separability across all operational states while maintaining robust cluster cohesion within each state category—a visual confirmation of the quantitative performance advantages reflected in our metric evaluations.

The detailed confusion matrix visualization presented in Figure 10 provides a comprehensive analysis of classification performance across 6 distinct health states, with each state represented by 75 test samples. Within this analytical framework, the horizontal coordinate represents the predicted label, while the vertical coordinate corresponds to the true label, thereby enabling precise evaluation of classification accuracy and error patterns. The results definitively demonstrate the superior diagnostic capabilities of our proposed approach, while specifically validating the effectiveness of two distinct noise cancellation techniques in enhancing fault diagnosis under extremely low labeling rates.

Case study II

Experimental apparatus and data acquisition protocol

The experimental dataset focusing on fixed-shaft gearbox dynamics[53] was collected at the University of Alberta's mechanical testing facility in Edmonton, Alberta. The test apparatus, illustrated in Figure 11, comprises an intricate power transmission system consisting of multiple interconnected components: a primary drive motor, a precision-engineered bevel gearbox, dual-stage planetary gearbox assemblies (first and second stage), accompanied by two speed-up fixed-shaft gearboxes in sequence, and culminating in a driven motor. The entire system's rotational velocity was precisely regulated through a variable frequency drive, with particular emphasis on the second speed-up gearbox (distinctly highlighted within a rectangular demarcation). This focal gearbox employs a three-shaft configuration, incorporating an input shaft, an intermediate (middle) shaft, and an output shaft, as detailed in Figure 12. The experimental setup was augmented with a strategic sensor array, including four high-precision accelerometers for vibration data acquisition, with three sensors positioned on the middle shaft's bearing housing to capture horizontal, vertical, and gear meshing dynamics, while the fourth was mounted atop the gearbox cover. Additionally, an encoder interfaced with the load motor shaft enabled accurate speed monitoring, with the output shaft's velocity synchronized to the motor speed through a coupling mechanism, thereby facilitating the calculation of the middle shaft's rotational velocity based on the established transmission ratio.

The experimental protocol encompassed a comprehensive investigation of five distinct operational scenarios, systematically examining both optimal functionality and progressive deterioration states characterized by varying severities of tooth root cracks, as documented in Figure 13. Data acquisition in Figure 14 was conducted under strictly controlled laboratory conditions. Horizontal vibration measurements were recorded at a high-fidelity sampling rate of 24 kHz during steady-state operation at a controlled input frequency of 20 Hz. This methodologically rigorous framework facilitated the development of case study II's dataset, which encompasses 200 unique samples for each identified health state and transient noise condition. Given the 5 discrete health states under investigation, the aggregate dataset comprises 1000 comprehensive samples, with each sample containing 2048 discrete data points to ensure statistical robustness and reliability. Furthermore, during the analytical phase, both labeled and unlabeled samples were

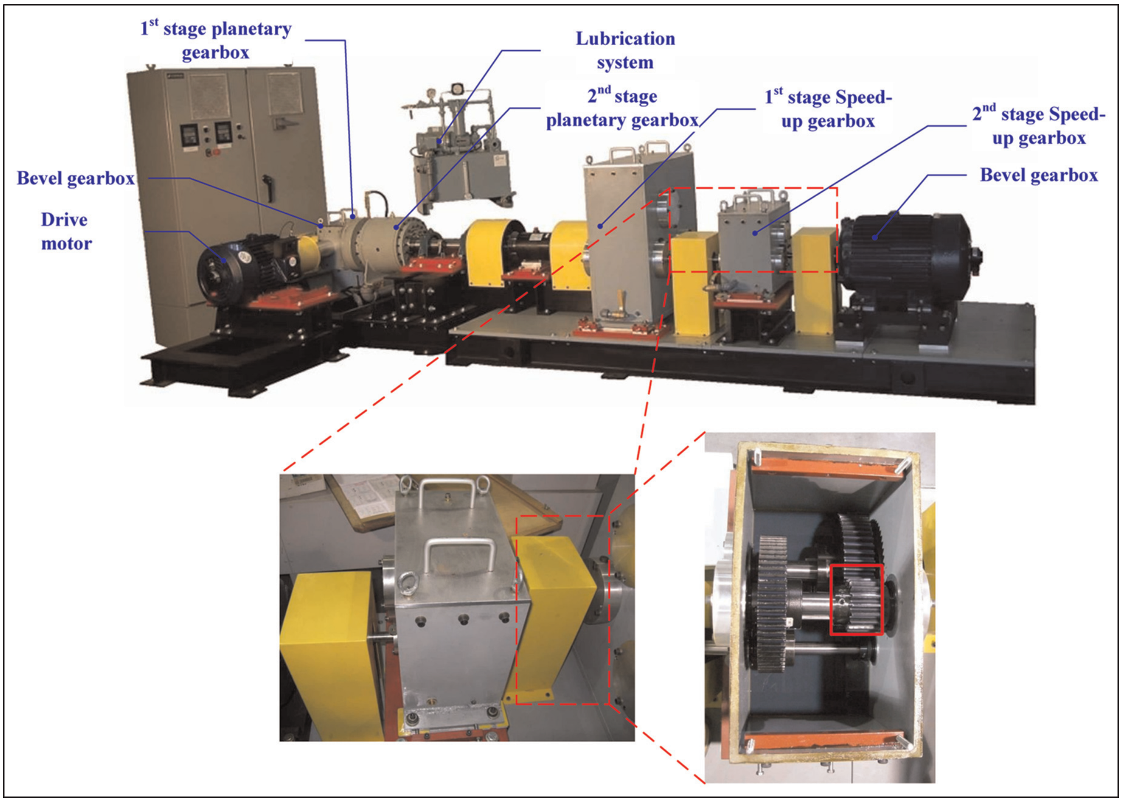


Figure 11. Experimental gearbox test rig configuration at the University of Alberta Reliability Laboratory.[53]

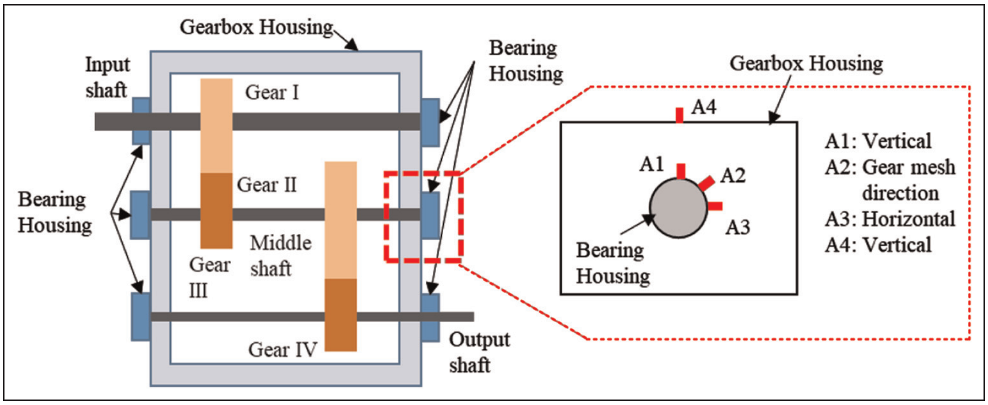


Figure 12. Cross-sectional schematic of the fixed-shaft gearbox with sensor placements.

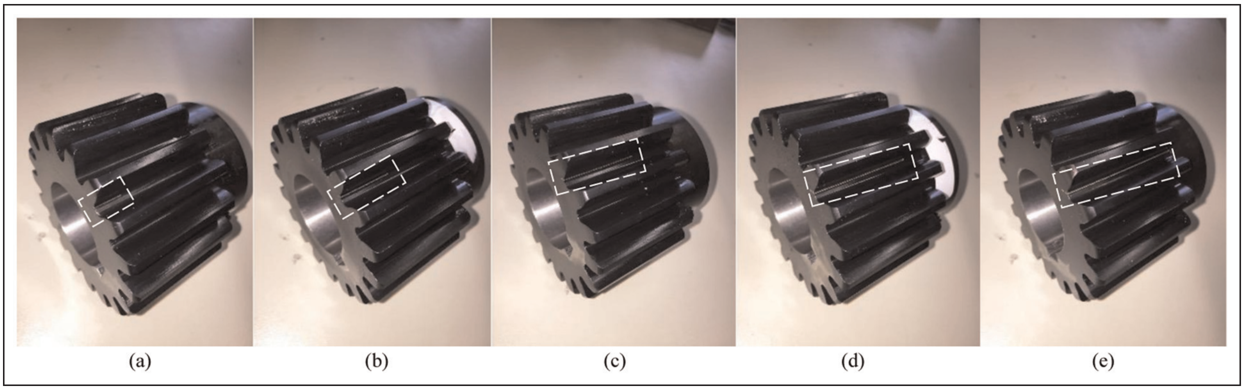


Figure 13. Sequential progression of gear tooth root crack severity, panels (a) through (e).

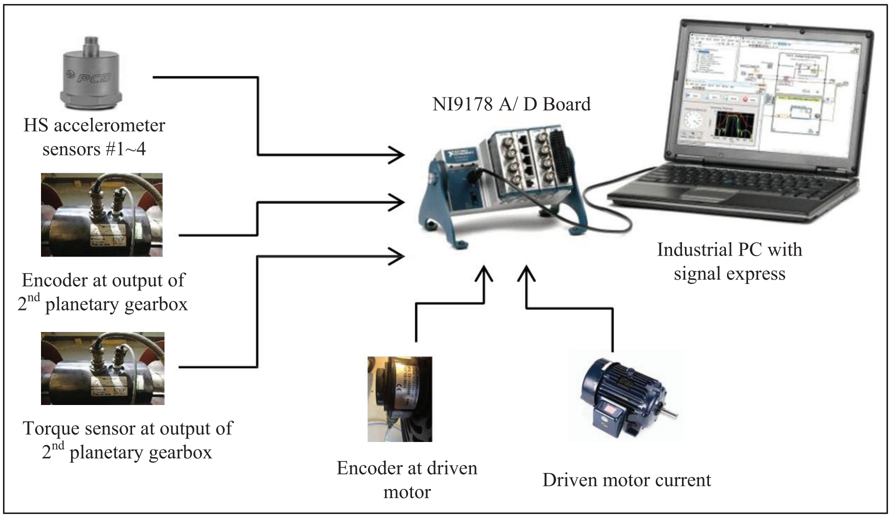


Figure 14. Architecture of the data acquisition system.

Table 4. Comparative analysis of classification accuracy (%) for case study II.

	Rate of labeled samples	
Model	10%	15%
HyperFast	92.98 ± 0.56	94.41 ± 0.46
TabPFN	80.00 ± 4.49	83.73 ± 2.08
GRL-CSG	86.13 ± 1.07	91.73 ± 1.27
Meta-SGC	92.27 ± 1.64	92.53 ± 2.88
DGAT	92.00 ± 0.27	92.62 ± 1.47
GraphSAGE	91.64 ± 1.89	93.07 ± 0.30
CARDensenet	92.27 ± 1.32	93.87 ± 1.55
AS-DWResnet	89.60 ± 0.98	90.93 ± 2.07
Ours	96.85 ± 0.96	97.42 ± 1.20

GRL-CSG: Graph representation learning and component space graph; DGAT: Dynamic Graph Attention Network; Meta-SGC: Meta learning and simplified graph convolutional.

systematically distributed according to varying proportions, with 10 nodes specifically allocated for validation purposes. The parametric configurations for additional critical model variables maintained consistency with those established in case study I, ensuring methodological continuity and comparative validity.

Comparative analysis of diagnostic performance metrics

Building upon the analytical framework established in case study I, this section presents an exhaustive evaluation of the proposed methodology through a rigorous comparative analysis with well-established diagnostic techniques, namely HyperFast, TabPFN, DGAT, GRL-CSG, Meta-SGC, GraphSAGE, CARDensenet, and AS-DWResnet. In order to facilitate a comprehensive understanding, the comparative diagnostic accuracy results for the dataset in case study II have been illustrated and presented in Table 4, whereas supplementary performance metrics are extensively detailed in Table 5. Moreover, to ensure statistical transparency, the standard deviation values presented in Table 4 are visually represented through error bars in Figure 15, wherein 10% and 15% specifically denote the proportions of labeled samples utilized for each fault type during the teacher model training phase. Additionally, to enhance the interpretability of our findings, we conducted sophisticated visualization analyses through T-SNE dimensionality reduction techniques, as meticulously illustrated in Figure 16, which are further complemented by comprehensive receiver operating characteristic (ROC) curve analyses presented in Figure 17.

The comparative analysis conducted in case study II encompasses an extensive examination of methodological performance across multiple critical metrics, including precision, recall, and F1 score, as comprehensively documented in Table 5. Notably, the proposed methodology demonstrates substantial and consistent performance advantages over existing state-of-the-art methods. Specifically, when utilizing 10% labeled samples, SSTGRL-FSA achieved an exceptional precision of 96.98%, which significantly surpasses not only HyperFast's 93.64% and DGAT's 93.06%, but also the graph-based GRL-CSG's 90.09% and meta-learning approach Meta-SGC's 91.56%. In terms of recall performance, our approach demonstrated remarkable effectiveness by attaining 96.82%, considerably exceeding both HyperFast's 92.25% and the strongest graph-based competitor GRL-CSG's 89.05%. Furthermore,

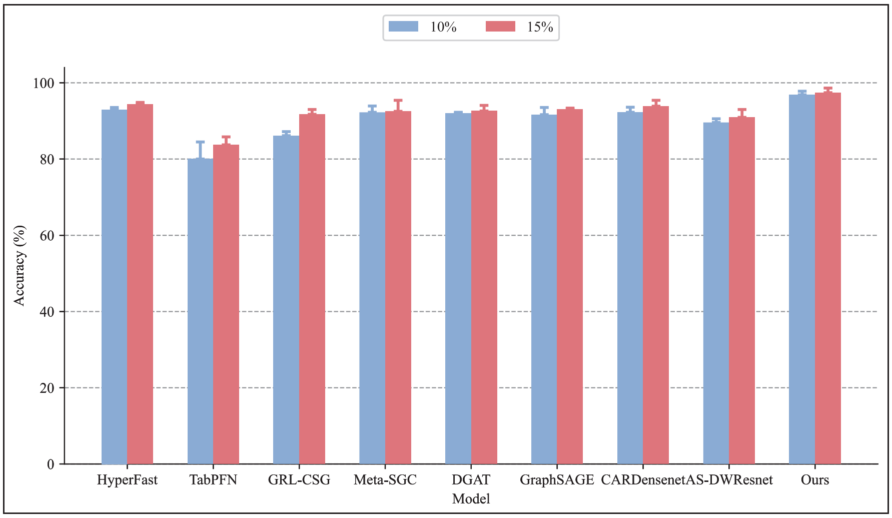


Figure 15. Evaluation of classification accuracy (%) across various model architectures in case study II, where experiments were conducted using 10% and 15% labeled training samples per fault category.
Error bars represent the standard deviation based on *n* = 8 independent trials.

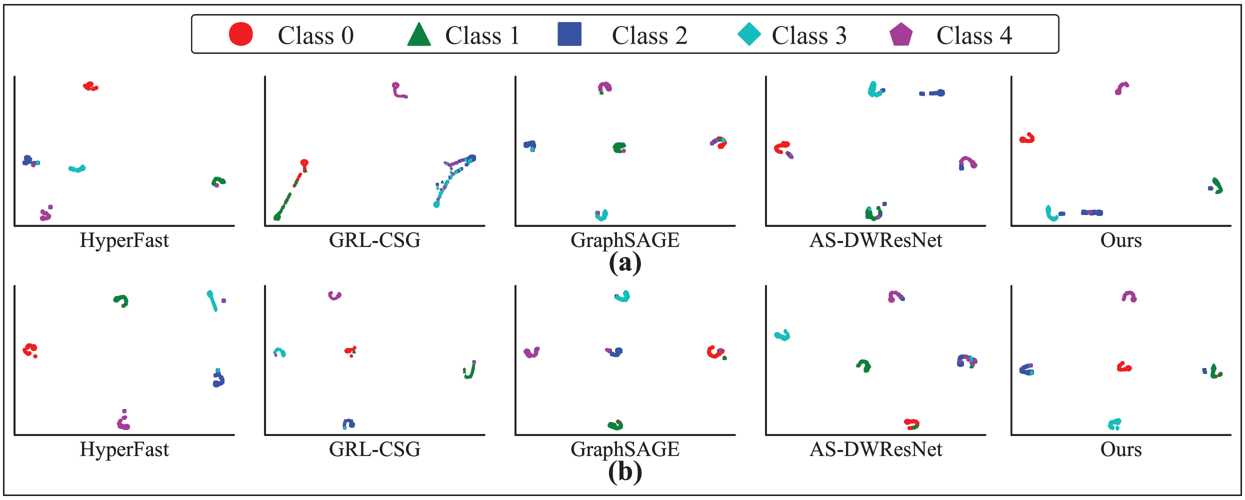


Figure 16. The T-Distributed Stochastic Neighbor Embedding (T-SNE) visualization of feature distributions for various methods in case study II: (a), (b) represent models trained with 10% and 15% labeled samples per class, respectively.
Fault categories 0∼4 correspond to root cracks of varying severity, respectively.

Table 5. Comparative analysis results of key metrics (%) and total running times for case study II.

	10% labeled samples			
Method	Precision	Recall	F1 score	Time (s)
HyperFast	93.64 ± 0.51	92.25 ± 0.73	93.11 ± 0.55	7.68
TabPFN	82.33 ± 3.59	81.02 ± 4.49	79.76 ± 4.67	7.11
GRL-CSG	90.09 ± 2.05	89.05 ± 2.31	91.04 ± 1.05	9.23
Meta-SGC	91.56 ± 3.11	90.03 ± 2.47	90.09 ± 1.88	7.95
DGAT	93.06 ± 0.26	92.06 ± 0.43	92.19 ± 0.27	8.13
GraphSAGE	92.11 ± 1.74	91.76 ± 1.93	91.59 ± 2.02	8.27
CARDensenet	93.51 ± 1.37	92.21 ± 1.46	92.03 ± 0.89	9.21
AS-DWResnet	90.31 ± 1.55	89.66 ± 1.01	89.48 ± 0.61	8.92
Ours	96.98 ± 0.85	96.82 ± 0.87	96.81 ± 0.96	11.02

GRL-CSG: Graph representation learning and component space graph; DGAT: Dynamic Graph Attention Network; Meta-SGC: Meta learning and simplified graph convolutional.

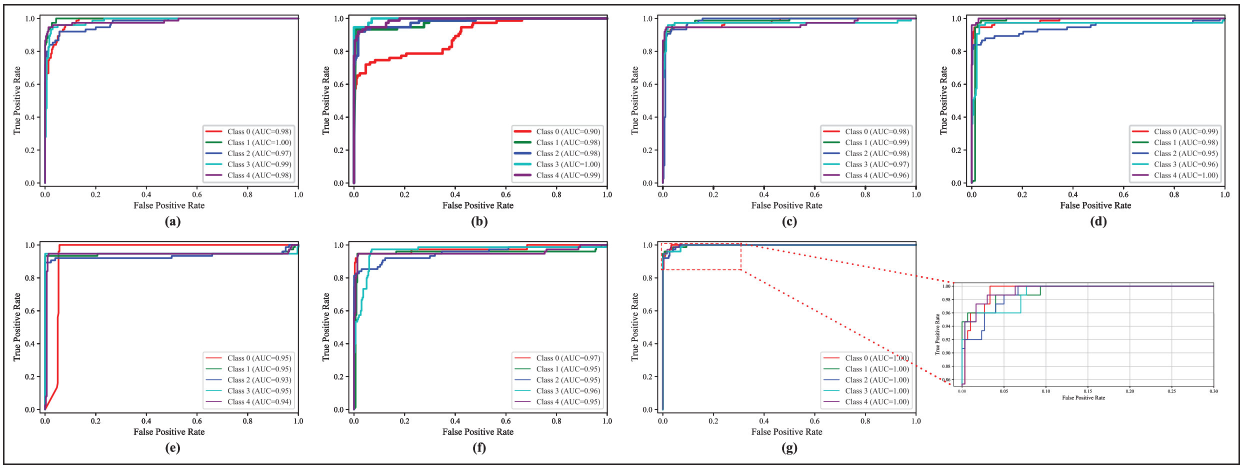


Figure 17. The ROC curves for comparative methods using 10% labeled samples per class in case study II: (a) HyperFast, (b) GRL-CSG, (c) DGAT, (d) GraphSAGE, (e) CARDensenet, (f) AS-DWResnet, (g) Ours.
Fault categories 0∼4 correspond to root cracks of varying severity.
ROC: Receiver operating characteristic; GRL-CSG: Graph representation learning and component space graph; DGAT: Dynamic Graph Attention Network.

the achieved F1 score of 96.81% substantiates the superior performance of our methodology when compared to all benchmarks, with absolute improvements of 3.70% over HyperFast, 4.62% over GRL-CSG, and 6.72% over Meta-SGC.

Although SSTGRL-FSA necessitates increased computational resources, requiring 11.02 s for execution, this computational overhead is justified by the framework's sophisticated hierarchical architecture, which systematically integrates multi-scale feature extraction with advanced attention mechanisms to optimize discriminative representations. The temporal-cost-to-performance-gain ratio proves particularly favorable when compared to graph-based methods: achieving a 4.62% F1 score improvement over GRL-CSG at the cost of only 1.79 s (19.4%) additional processing time. This efficiency-performance profile presents a compelling trade-off in industrial diagnostic applications where predictive accuracy directly impacts maintenance decisions and operational safety. The substantial performance margins over specialized graph-based methods like GRL-CSG and Meta-SGC particularly validate our innovative dual-graph adaptive transfer mechanism's effectiveness in capturing complex fault propagation patterns across heterogeneous wind turbine components.

When examining performance at the 10% labeled sample ratio, our methodology achieves a commanding accuracy of 96.85%, representing a substantial 3.87% absolute advantage over the nearest competitor, HyperFast (92.98%), and a remarkable 10.72% superiority over the graph-based approach GRL-CSG (86.13%). Furthermore, as the proportion of labeled samples increases to 15%, SSTGRL-FSA maintains its dominant position with an impressive 97.42% accuracy. While our method's absolute accuracy increase from 10% to 15% labeled data is modest at 0.57%, this characteristic actually demonstrates SSTGRL-FSA's exceptional data efficiency—achieving near-optimal performance even with extremely limited supervision. Notably, our framework maintains significant performance advantages across all conditions, with a consistent 4.35% lead over GRL-CSG even at the higher 15% labeling rate.

The comprehensive visual analysis of diagnostic performance through T-SNE dimensionality reduction, as meticulously illustrated in Figure 16, definitively demonstrates that features extracted through the proposed methodology exhibit markedly superior class separation and discriminative characteristics when compared to alternative approaches. Furthermore, the detailed ROC curve analysis presented in Figure 17 provides an exhaustive evaluation of diagnostic performance across all methodologies under consideration. While area under the curve (AUC) values theoretically range from 0.5 (indicating random classification) to 1.0 (perfect discrimination), with higher values signifying superior diagnostic capability, our proposed methodology demonstrates optimal performance across all evaluation criteria. This is evidenced by all four ROC curves exhibiting two crucial characteristics: firstly, maximal proximity to the ideal upper-left corner relative to all comparative methods, and secondly, consistently superior AUC scores across all test conditions. This dual evidence provides robust and comprehensive validation of both the methodological advancement

Table 6. Ablation study on component contributions across different label ratios (%).

Model variant	1% labels	5% labels	10% labels
Full SSTGRL-FSA (all components)	82.74	94.48	97.52
Markov and ACS module variants			
w/o Markov transition matrix	76.48	90.84	94.11
w/o ACS	77.44	91.26	93.07
w/o Full Markov module	65.32	75.04	84.04
Semi-supervised learning variants			
w/o Pseudo-label enhancement	67.18	80.36	88.75
w/o Label matching strategy	70.52	85.24	91.33
Baseline (DGAT + basic pseudo-labeling)	63.51	73.76	82.59

DGAT: Dynamic graph attention network; ACS: Amplitude-constrained scaling; SSTGRL-FSA: Semi-Supervised Transfer Graph Representation Learning with Few-Shot Adaptation.

and operational effectiveness of our proposed approach in real-world diagnostic scenarios.

Ablation study

To rigorously quantify the contribution of each component in our SSTGRL-FSA framework, we conducted a comprehensive ablation study across varying label availability scenarios (1%, 5%, and 10% labeled samples). Table 6 presents the systematic evaluation of our framework's key innovations: (1) the Markov and ACS denoising module, and (2) the SSL components, including pseudo-label enhancement and label matching strategy.

The results reveal critical insights into the functional necessity of each component. Starting with the Markov module variants, the complete removal of this module (w/o Full Markov module) achieves only 65.32% accuracy with 1% labeled data, representing a mere 1.81% improvement over the baseline. This minimal gain confirms that without proper noise handling, the framework cannot effectively leverage limited labeled data. However, when either the Markov transition matrix or ACS is preserved, performance increases substantially to 76.48% and 77.44% respectively, demonstrating their complementary roles in noise suppression. Their combined effect is particularly pronounced in low-label scenarios, where they collectively contribute 13.97% accuracy improvement over the baseline at 1% label ratio.

The SSL components demonstrate equally significant contributions. Removing the pseudo-label enhancement mechanism reduces accuracy to 67.18% (1% labels), only 3.67% above baseline and 15.56% below the complete framework. This dramatic degradation highlights its critical role in preventing error propagation when learning from few labeled samples. Similarly, eliminating the label matching strategy yields 70.52% accuracy, confirming its 7.01% contribution to cross-domain knowledge transfer and its 12.22% importance to the full framework's performance.

The synergistic integration of all components produces multiplicative benefits that exceed individual contributions. With only 1% labeled data, the complete SSTGRL-FSA framework achieves 82.74% accuracy—19.23% higher than baseline and 5.30%–15.56% better than any single-component removal variant. This performance advantage becomes even more pronounced with increased label availability, reaching 97.52% at 10% label ratio. Notably, the performance gap between the full framework and component-removal variants widens as label ratio increases from 1% to 10%, confirming that our integrated approach not only enables FSL but also maximizes the utility of available labeled data.

These findings provide empirical validation for our theoretical framework: effective gearbox fault diagnosis under transient noise requires coordinated signal denoising (Markov module), feature stabilization (ACS), and efficient label utilization (pseudo-label enhancement and label matching). The systematic performance degradation observed when removing any component—ranging from 5.30% to 15.56% at 1% label ratio—demonstrates that SSTGRL-FSA's effectiveness stems from the carefully integrated combination of theoretically grounded mechanisms rather than any isolated technique. This comprehensive analysis establishes both the necessity of each component and their synergistic interaction, addressing a critical requirement for practical deployment where labeled fault data remains inherently scarce.

In-depth analysis of SSTGRL-FSA performance and optimization strategies

Empirical investigation of neighborhood parameter dynamics and their impact on classification performance

In graph-based analysis, selecting the optimal neighborhood parameters is a crucial consideration factor

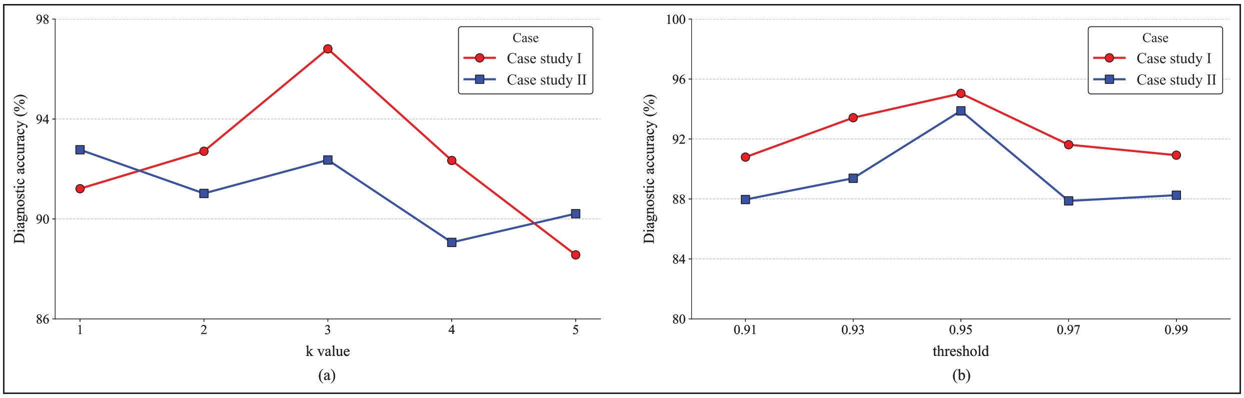


Figure 18. Comprehensive performance analysis: (a) Classification accuracy across varying neighborhood parameters (k-value). (b) Performance metrics under different thresholds of pseudo-label.

that requires systematic empirical research. Figure 18(a) shows the diagnostic performance indicators of the proposed method under different k-value configurations. The experimental results indicate that the selection of neighborhood parameters has a significant impact on the diagnostic performance.

As shown in Figure 18, in case study I, the model achieved optimal performance at $k = 3$, with a classification accuracy of 96.81%; In case study II, the suboptimal result occurred at $k = 3$, with an accuracy of 92.36%. A noteworthy phenomenon is the relationship between changes in k values and diagnostic accuracy: as k values increase, classification accuracy exhibits a unique oscillating downward trend. This phenomenon is mainly attributed to the inherent limitations in FSL scenarios, where larger k values may unintentionally increase the probability of establishing edge connections between samples belonging to different class labels, thereby potentially weakening the model's discriminative ability.

Optimization and analysis of confidence threshold

The selection of an optimal confidence threshold when generating pseudo-labels using a teacher model is crucial for ensuring the reliability and accuracy of the labels assigned to unlabeled data. In our experiment, we meticulously explored how varying this threshold impacts diagnostic accuracy across two distinct case studies as shown in Figure 18(b). Our findings reveal that the choice of threshold significantly influences the performance outcomes, with a clear peak in accuracy observed at a threshold value of 0.95.

In case study I, the diagnostic accuracy ascended steadily as the threshold increased, reaching its apex at 95.04% when the threshold was set to 0.95. This trend indicates that, within this specific context, setting a relatively high threshold helps in filtering out less confident predictions, thereby enhancing overall accuracy. However, continuing to raise the threshold beyond this point resulted in a slight decline in accuracy, suggesting that overly stringent criteria may lead to the exclusion of potentially correct labels. Similarly, in case study II, the highest diagnostic accuracy of 93.88% was achieved at the same threshold level of 0.95. This consistency across both cases underscores the importance of fine-tuning the threshold to optimize model performance. It also highlights a delicate balance between precision and recall; while a higher threshold improves the quality of pseudo-labels by increasing the confidence required for their acceptance, it simultaneously risks reducing the quantity of usable data.

Robustness analysis under various noise conditions

To thoroughly assess signal processing resilience, the research introduced multiple intensities of Gaussian noise with signal-to-noise ratios (SNRs) of 0 dB, 5 dB, and 10 dB into the case study II dataset. These noise intensities were determined using the conventional SNR equation: $SNR(dB) = 10\log_{10}\left(\frac{P_{\text{signal}}}{P_{\text{noise}}}\right)$, where P_{noise} indicates noise power and P_{signal} represents signal power.

As shown in Table 7, the experimental results reveal that the proposed approach preserves stable performance when SNR values are above 5 dB, yet experiences significant performance reduction at 0 dB SNR. This decline in performance stems primarily from the graph structure's essential dependence on similarity relationships between samples; particularly, substantial noise injection can significantly reduce these similarity relationships, even between samples with the same class labels, thus undermining the reliability of the sample connectivity network and ultimately deteriorating the overall graph quality.

Table 7. Experimental results under various noise conditions.

Evaluation metric	SNR (dB)			
	10	5	0	No noise
Accuracy (%)	95.41 ± 0.93	93.02 ± 1.11	90.68 ± 1.21	96.85 ± 0.96
F1 Score (%)	96.25 ± 1.01	92.67 ± 1.53	91.42 ± 1.22	96.81 ± 0.96
Precision (%)	95.17 ± 1.44	92.81 ± 0.65	90.31 ± 2.01	96.98 ± 0.85

SNR: Signal-to-noise ratio.

Conclusion

The SSTGRL-FSA framework represents an advancement in gearbox fault diagnosis by effectively addressing two critical challenges: data scarcity and transient noise interference. Through its innovative three-component architecture, the framework demonstrates superior diagnostic capabilities compared to conventional methods. The pseudo-label reliability enhancement mechanism ensures robust fault classification even with limited labeled data, while the advanced label transmission strategy effectively bridges the gap between different operational domains. The integration of the Markov state probability transition matrix further strengthens the framework's ability to maintain stable feature representations under noise conditions.

The framework's robust performance under challenging operational conditions makes SSTGRL-FSA a practical solution for industrial applications, potentially improving maintenance efficiency, reducing downtime, and enhancing overall system reliability. The successful implementation of this framework could lead to significant cost savings in industrial maintenance programs and improved safety standards across various sectors utilizing planetary gearboxes. Future research directions could include extending this framework to other mechanical systems, exploring real-time implementation strategies for dynamic industrial environments, and investigating the potential integration with emerging technologies such as edge computing and Industrial Internet of Things (IIoT) platforms. Additionally, the framework's adaptability suggests promising applications in related fields requiring robust fault diagnosis under similar challenging conditions.

ORCID iDs

Peng Chen https://orcid.org/0000-0002-3265-3079
Junyu Qi https://orcid.org/0000-0003-1669-8987

Author contributions

Peng Chen and Jia Gao: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing—Original Draft, Writing—Review & Editing. Yuhao Wu: Data curation, Methodology, Software. Changbo He and Ge Xin: Funding acquisition, Resources, Supervision, Visualization, Writing—review & editing. Shuai Fan and Junyu Qi: Data curation, Funding acquisition, Project administration.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the National Natural Science Foundation of China (Grants 52105111, 52305085, 52375078, U23B20104 and 52405009), the Guangdong Basic and Applied Basic Research Foundation (Grant 2025 A1515012256), the China Postdoctoral Science Foundation (Grant 2023M740021), and the Natural Science Foundation of Anhui Province (Grant 2108085QE229).

Data availability

Data will be made available on request.

References

1. Bachar L, Matania O, Cohen R, et al. Anomaly detection of gear wear: introducing a novel health indicator through experimentation and physical investigation. *Struc Health Monitor*. Epub ahead of print 13 January 2025. DOI: 10.1177/14759217241307837.
2. Chen P, Ma J, He C, et al. Semi-supervised consistency models for automated defect detection in carbon fiber composite structures with limited data. *Meas Sci Technol* 2025; 36(4): 046109.
3. Xu C, Chen P, Gao J, et al. Semi-supervised transfer learning preserving spatial homogeneity for gearbox diagnostics in extraneous transient noise. *Nondestruct Test Eval*. Epub ahead of print 28 August 2025. DOI: 10.1080/10589759.2025.2544893.
4. Chen P, Xu C, Ma Z, et al. A mixed samples-driven methodology based on denoising diffusion probabilistic model for identifying damage in carbon fiber composite structures. *IEEE Trans Instrum Meas* 2023; 72: 3513411.

5. Wang Y, Chen P, Wei Q, et al. Multi-channel fusion scale transformed signals with magnetic leakage for damage detection in steel wire ropes. *Nondestr Test Eval*. Epub ahead of print 9 August 2025. DOI: 10.1080/10589759.2025.2543513.
6. Ahmad H, Cheng W, Xing J, et al. Deep learning-based fault diagnosis of planetary gearbox: a systematic review. *J Manuf Syst* 2024; 77: 730–745.
7. Chen P, Zhang R, Fan S, et al. Step-wise contrastive representation learning for diagnosing unknown defective categories in planetary gearboxes. *Knowl Based Syst* 2024; 309: 112863.
8. Ghanbari Kohyani N, Riyazi Y, Shirazi FA, et al. Noise-robust gearbox fault detection: a deep learning approach. *J Theor Appl Vib Acoust* 2024; 10(1): 54–66.
9. Li D, Li C, Yang J, et al. Bayesian optimization-attention-feedforward neural network based train traction motor-gearbox coupled noise prediction. *Measurement* 2024; 238: 115323.
10. Chen P, Zhang R, He C, et al. Progressive contrastive representation learning for defect diagnosis in aluminum disk substrates with a bio-inspired vision sensor. *Expert Syst Appl* 2025; 289: 128305.
11. Zhang Q, Song C and Yuan Y. Fault diagnosis of vehicle gearboxes based on adaptive wavelet threshold and LT-PCA-NGO-SVM. *Appl Sci* 2024; 14(3): 1212.
12. Wang Z, Gao Y, Yu J, et al. Data-driven fault diagnosis of PEMFC water management with segmented cell and deep learning technologies. *Int J Hydrogen Energy* 2024; 67: 715–727.
13. Chen P, Ma Z, Xu C, et al. Self-supervised transfer learning for remote wear evaluation in machine tool elements with imaging transmission attenuation. *IEEE Internet Things J* 2024; 11: 23045–23054.
14. Chen P, Gao J, Jin Y, et al. Physics-aware digital twin and few-shot learning for self-supervised gearbox fault diagnosis. *Meas Sci Technol* 2025; 36(9): 096106.
15. El Idrissi A, Derouich A, Mahfoud S, et al. Bearing faults diagnosis by current envelope analysis under direct torque control based on neural networks and fuzzy logic—a comparative study. *Electronics* 2024; 13(16): 3195.
16. Chen P, Gao J, Zhang R, et al. Metric-guided graph contrastive learning: an unsupervised approach for few-shot gearbox fault diagnosis. *Meas Sci Technol* 2025; 36(7): 076110.
17. Chen P, Wu Y, Xu C, et al. Interference suppression of nonstationary signals for bearing diagnosis under transient noise measurements. *IEEE Trans Reliab* 2025; 74: 4047–4061.
18. Chen P, Wu Y, Fan S, et al. Adaptive signal regime for identifying transient shifts: a novel approach toward fault diagnosis in wind turbine systems. *Ocean Eng* 2025; 325: 120798.
19. Chen P, Wu Y, Xu C, et al. Markov modeling of signal condition transitions for bearing diagnostics under external interference conditions. *IEEE Trans Instrum Meas* 2024; 73: 3518308.
20. Vashishtha G, Chauhan S, Zimroz R, et al. Optimization of spectral kurtosis-based filtering through flow direction algorithm for early fault detection. *Measurement* 2025; 241: 115737.
21. Chin C, Abdullah S, Ariffin A, et al. A review of the wavelet transform for durability and structural health monitoring in automotive applications. *Alexandria Eng J* 2024; 99: 204–216.
22. Yuzgec U, Dokur E and Balci M. A novel hybrid model based on empirical mode decomposition and echo state network for wind power forecasting. *Energy* 2024; 300: 131546.
23. Indragandhi V, Kumar RS and Saranya R. Hilbert-Huang Transform and machine learning based electromechanical analysis of induction machine under power quality disturbances. *Results Eng* 2024; 24: 103075.
24. Keshun Y, Puzhou W, Peng H, et al. A sound-vibration physical-information fusion constraint-guided deep learning method for rolling bearing fault diagnosis. *Reliab Eng Syst Saf* 2025; 253: 110556.
25. Bilal H, Obaidat MS, Aslam MS, et al. Online fault diagnosis of industrial robot using IoRT and hybrid deep learning techniques: an experimental approach. *IEEE Internet Things J* 2024; 11(19): 31422–31437.
26. Amin AA, Mubarak A and Waseem S. Application of physics-informed neural networks in fault diagnosis and fault-tolerant control design for electric vehicles: a review. *Measurement* 2025; 246: 116728.
27. Lu L, Wang S, Yu X, et al. An advanced fault diagnosis approach for wind turbine planetary gearbox based on optimized multi-layer attention denoising autoencoders. *Meas Sci Technol* 2025; 36(2): 026132.
28. Cui D, Zhou R, Li H, et al. Rotor fault diagnosis of centrifugal pumps in nuclear power plants based on CWGAN-GP-CNN for imbalanced dataset. *Prog Nucl Energy* 2025; 178: 105500.
29. Joung BG, Nath C, Li Z, et al. Bearing anomaly detection in an air compressor using an LSTM and RNN-based machine learning model. *Int J Adv Manuf Technol* 2024; 134(7): 3519–3530.
30. Obaido G, Mienye ID, Egbelowo OF, et al. Supervised machine learning in drug discovery and development: algorithms, applications, challenges, and prospects. *Mach Learn Appl* 2024; 17: 100576.
31. Gui J, Chen T, Zhang J, et al. Tao. A survey on self-supervised learning: algorithms, applications, and future trends. *IEEE Trans Pattern Anal Mach Intell* 2024; 46(12): 9052–9071.
32. Liang G, Li F, Pang X, et al. A gear fault diagnosis method based on reactive power and semi-supervised learning. *Meas Sci Technol* 2024; 35(12): 126107.
33. Ye T, Yuan X, Yang X, et al. A semi-supervised intelligent fault diagnosis method for bearings under low labeled rates. *IEEE Trans Instrum Meas* 2024; 73: 3531111.
34. Su H, Yao Q, Xiang L, et al. Semi-supervised temporal meta-learning framework for wind turbine bearing fault diagnosis under limited annotation data. *IEEE Trans Instrum Meas* 2024; 73: 3512309.

35. Chen X, Chen Z, Guo L, et al. Pseudo-label assisted semi-supervised adversarial enhancement learning for fault diagnosis of gearbox degradation with limited data. *Mech Syst Signal Process* 2025; 224: 112108.
36. Huang Q, Li C, Han Y, et al. Semi-supervised prototype networks with similarity information selection for fault diagnosis of wind turbine gearboxes. *IEEE Trans Instrum Meas* 2025; 74: 3507012.
37. Cai W, Zhao D and Wang T. Spatial-temporal graph attention contrastive learning for semi-supervised bearing fault diagnosis with limited labeled samples. *Comput Ind Eng* 2025; 204: 111106.
38. Jia M, Yao Y and Liu Y. Review on graph neural networks for process soft sensor development, fault diagnosis, and process monitoring. *Ind Eng Chem Res* 2025; 64(17): 8543–8564.
39. Dang X, Li Y, Ma W, et al. Towards exploring the limitations of test selection techniques on graph neural networks: an empirical study. *Empir Softw Eng* 2024; 29(5): 112.
40. Yue K, Wang L, Ding X, et al. Physics-informed dual guidance method using physical envelope harmonic distribution and transfer learning for few-shot gear fault classification. *Eng Appl Artif Intell* 2025; 153: 110956.
41. Luo R, Li Y, Guo H, et al. Cross-operating-condition fault diagnosis of a small module reactor based on CNN-LSTM transfer learning with limited data. *Energy* 2024; 313: 133901.
42. Dai X, Yi K, Wang F, et al. Bearing fault diagnosis based on POA-VMD with GADF-Swin Transformer transfer learning network. *Measurement* 2024; 238: 115328.
43. Dong X, Cheng M, Wang H, et al. A potential solution to insufficient target-domain noise data: transfer learning and noise modeling. *IEEE Trans Geosci Remote Sens* 2023; 61: 1–15.
44. Zhan X, Liu C, Zhang A, et al. Prediction of compressor aerodynamic noise based on transfer learning Random Forest. *J Acoust Soc Am* 158(1): 336–351.
45. Bonet D, Montserrat DM, Giró-i Nieto, et al. Hyperfast: instant classification for tabular data. *Proc AAAI Conf Artif Intell* 2024; 38: 11114–11123.
46. Hoo SB, Müller S, Salinas D, et al. The tabular foundation model TabPFN outperforms specialized time series forecasting models based on simple features. In: *NeurIPS workshop on time series in the age of large models*, https://arxiv.org/pdf/2501.02945 (2024).
47. Yan S, Shao H, Xiao Y, et al. Semi-supervised fault diagnosis of machinery using LPS-DGAT under speed fluctuation and extremely low labeled rates. *Adv Eng Inform* 2022; 53: 101648.
48. Huang K and Chen C. Subgraph generation applied in GraphSAGE deal with imbalanced node classification. *Soft Comput* 2024; 28(17): 10727–10740.
49. Guo J, Liu C, Liu S, et al. Graph contrastive learning for semi-supervised wind turbine fault diagnosis with few labeled SCADA data. *Measurement* 2025; 245: 116531.
50. Liu Z and Peng Z. Few-shot bearing fault diagnosis by semi-supervised meta-learning with graph convolutional neural network under variable working conditions. *Measurement* 2025; 240: 115402.
51. Li J, Ding W, Mao W, et al. Causal dilated Convolution-Based residual DenseNet with channel attention for RUL prediction of rolling bearings. *Measurement* 2024; 235: 115012.
52. Qin G, Zhang K, Lai X, et al. An adaptive symmetric loss in dynamic wide-kernel ResNet for rotating machinery fault diagnosis under noisy labels. *IEEE Trans Instrum Meas* 2024; 73: 3517512.
53. Rao M. *Development of deep learning-based methods for rotating machinery fault diagnosis under varying speed conditions*. PhD thesis, University of Alberta, 2023.