NONDESTRUCTIVE TESTING AND EVALUATION
https://doi.org/10.1080/10589759.2025.2544893

Semi-supervised transfer learning preserving spatial homogeneity for gearbox diagnostics in extraneous transient noise

Chaojun Xu[a,b], Peng Chen[a,c], Jia Gao[a], Yaqiang Jin[b,d] and Meng Rao[d]

[a]College of Engineering, Shantou University, Shantou, Guangdong, P.R. China; [b]School of Qilu Transportation, Shandong University, Jinan, Shandong, P.R. China; [c]Key Laboratory of Intelligent Manufacturing Technology, Ministry of Education, Shantou, Guangdong, P.R. China; [d]Qingdao Mingserve Tech, Qingdao, Shandong, P.R. China

ABSTRACT

Planetary gearboxes are critical components in a wide range of applications, including electric motors, automotive systems, and wind turbines. However, current fault diagnosis methods face significant challenges in accurately detecting faults due to the prevalence of large amounts of unlabelled data and interference from transient events. To address these issues, this paper proposes a novel Semi-Supervised Transfer Learning (SSTL) approach. SSTL integrates semi-supervised techniques with transfer learning and introduces innovative normalisation strategies to overcome the limitations of traditional methods, particularly in environments characterised by transient interference and limited labelled data. The proposed approach offers several key contributions: (1) a semi-supervised transfer learning model that effectively leverages unlabelled data while mapping the source to the target domain, thereby enhancing fault detection accuracy, (2) a label migration and matching strategy that assigns pseudo labels to transient signals, addressing persistent challenges in signal processing, and (3) a limiting normalisation strategy designed to mitigate the effects of transient interference and stabilise vibration signals, thus improving the robustness of the model. Two case studies are presented to validate the effectiveness of SSTL, demonstrating its superiority over existing methods in terms of fault diagnosis accuracy and reliability, especially in scenarios with limited labelled data and frequent transient interference.

ARTICLE HISTORY
Received 1 June 2025
Accepted 25 July 2025

KEYWORDS
Gearbox; fault diagnosis; semi-supervised learning; transfer learning; vibration signal; transient noise

1. Introduction

Planetary gearboxes are crucial components used in a wide range of applications, including electric motors, the automotive and aviation industries, wind turbines, and various other industrial environments. Their widespread use is primarily attributed to key advantages, such as high transmission ratios, compact design, and significant load-carrying capacity [1–3]. However, the gearboxes in these applications are often subject to varying speeds, dynamic loads, and harsh operating conditions, which can consequently

CONTACT Peng Chen pengchen@alu.uestc.edu.cn

lead to localised failures such as wear, cracks, broken or damaged teeth, and pitting. Therefore, detecting and addressing these faults promptly is critical [4–7]. As a result, the development of advanced and effective fault diagnosis techniques is imperative. Not only are such techniques crucial for the early detection of potential issues, but they also ensure safe and reliable system operation, thereby enhancing overall system reliability. Furthermore, these advanced diagnostic methods can potentially reduce maintenance costs and extend the lifespan of planetary gearboxes across various industrial applications.

In recent years, data-driven-based fault diagnosis methods [8–12], particularly those incorporating deep learning techniques, have seen rapid and significant advancements. These advancements can be primarily attributed to the superior capacity of deep learning to automatically extract high-level feature representations from raw signals, thereby facilitating high-precision diagnostic predictions in an end-to-end manner. Over the past decade, numerous impressive algorithms based on deep learning have emerged, including but not limited to convolutional neural networks [13], generative networks [14,15], semi-supervised learning models [16,17], and self-supervised models [18]. To illustrate the effectiveness of these methods, several studies have significantly contributed to the field. For instance, Jamil et al. [19] propose a novel deep boosted transfer learning method for wind turbine gearbox fault detection, which effectively mitigates negative transfer by selectively focusing on relevant information from the source machine. This approach leads to enhanced accuracy compared to traditional deep learning and deep transfer learning methods. Similarly, Zhang et al. [20] introduce a nearly end-to-end deep learning approach for diagnosing wind turbine gearbox faults using vibration signals. Their method integrates Empirical Mode Decomposition (EMD) to improve model efficiency and generalisation, particularly under nonstationary working conditions. Furthermore, Chen et al. [21] present a physics-informed hyperparameter selection strategy for Long Short-Term Memory (LSTM) neural networks, which aims to enhance fault detection in gearboxes by focusing on maximising the discrepancy between healthy and faulty states, rather than solely minimising validation mean squared error. Additionally, Zhang et al. [22] propose an intelligent fault diagnosis method based on an adaptive intraclass and interclass convolutional neural network (AIICNN). This method improves sample distribution and diagnostic accuracy by addressing variable working conditions through adaptive constraints.

Despite the impressive performance of deep learning-based methods, these approaches necessitate substantial amounts of labelled data for training, which presents challenges for fault diagnosis in real-world industrial applications. Consequently, extensive research has been conducted on semi-supervised learning, aiming to leverage large quantities of unlabelled data to support and enhance the performance of deep learning models with limited labelled data. Zhang et al. [23] propose the Semi-Supervised Momentum Prototype Network (SSMPN), an advanced few-shot semi-supervised learning approach designed to improve gearbox fault diagnosis in scenarios with limited labelled samples. This method effectively utilises prototype networks to capture feature mappings, employs Monte Carlo uncertainty for refined pseudo-labelling, and incorporates momentum-based prototype fine-tuning to enhance model performance. Similarly, Zhao et al. [24] introduce a novel two-stage hybrid semi-supervised learning framework that integrates grouped pseudo-labelling with consistency regularisation. This method addresses challenges related to teacher model accuracy and the limitations of data augmentation techniques for 1-D vibration signals, thereby significantly improving fault diagnosis accuracy in rotating machinery despite the constraints of limited labelled data. Furthermore, Xiao et al. [25] present a semi-supervised hybrid framework aimed at diagnosing converter transformers with limited labelled data. Their framework combines multi-feature graph generation, which encodes vibration signals into time, frequency, and energy graphs, with a blend of unsupervised and supervised learning strategies and soft voting decision-making. This comprehensive approach enhances intelligent fault diagnosis capabilities, even under challenging dataset conditions. Despite the notable effectiveness of semi-supervised learning-based methods in addressing the challenges associated with limited labelled data in fault diagnosis, three significant challenges persist, which warrant further investigation and methodological improvements.

(1) Pseudo-labelling Inaccuracy: The generation of pseudo labels for unlabelled data through data enhancement techniques often results in substantial deviations from true labels. Consequently, these inaccurate pseudo labels fail to positively influence the model's feature extraction process and may even introduce noise or bias into the learning algorithm. This challenge underscores the need for more robust and accurate pseudo-labelling techniques that can better approximate the underlying data distribution.
(2) Limited Dataset Expansion: While methods for constructing data matching pairs partially address data scarcity, they offer only modest dataset expansion and fail to introduce entirely new samples. This limitation constrains the model's ability to generalise to unseen data, particularly in scenarios where fault patterns exhibit high variability or complexity. As a result, the overall utility of these methods in enriching dataset diversity remains limited.
(3) Insufficient Leverage of Unlabelled Data: Despite advancements in semi-supervised learning, current methods do not fully exploit the potential of unlabelled data in fault diagnosis tasks. This inefficiency stems from the combined effects of pseudo-labelling inaccuracies and limited dataset expansion. Consequently, there is a pressing need for innovative approaches that can more effectively leverage unlabelled data while simultaneously enhancing the quality and diversity of the training dataset.

Given the inherent limitations of semi-supervised learning, researchers have increasingly turned their attention to transfer learning, particularly those involving the fine-tuning of pre-trained models to achieve improved diagnostic performance. This shift in focus is largely due to the potential of transfer learning to overcome the constraints associated with limited labelled data and to leverage knowledge from related domains. For instance, Xiang et al. [26] introduce the Classifier Constrained Domain Adaptation Network (CCDAN), an innovative transfer unsupervised learning method designed to enhance rotor fault diagnosis by extracting transferable features from simulated samples and improving classification accuracy through the use of classifier constraints and multiple-kernel maximum mean discrepancy (MK-MMD). Furthermore, Sun et al. [27] propose a novel Cross-Domain Transfer Learning with Fine-Tuning Mechanism (CTL-FTM) for gearbox fault diagnosis, which effectively addresses challenges associated with imbalanced datasets and the complexities of hyperparameter tuning by leveraging pre-trained models and shallow networks, leading to enhanced diagnostic accuracy and generalisation capability. However, under certain operating conditions, instantaneous fluctuations may arise, manifesting as transient disturbances within the vibration signal. These disturbances can adversely affect the stability and reliable performance of the equipment. It is crucial to acknowledge that when vibration signals encounter interference, especially transient interference, the effectiveness of transfer learning methods frequently fails to meet anticipated results. This shortcoming highlights the imperative for continued research and development in this domain, as the current methods may not fully address the complexities introduced by such interferences.

However, if this unlabelled data can be effectively leveraged alongside a limited set of labelled data, there is significant potential to enhance the model's performance, accuracy, and reliability. This approach aligns with the growing trend in machine learning towards leveraging large amounts of unlabelled data to improve model robustness and generalisation. In the light of these challenges and the growing demand for more adaptive and efficient learning paradigms in industrial mechanical systems, a novel approach is proposed: SSTL. This method is specifically designed to address the issue of limited labelled data in the presence of transient interferences by integrating semi-supervised learning and transfer learning with novel normalisation strategies. The SSTL approach represents a synthesis of multiple machine learning paradigms, aiming to harness the strengths of each while mitigating their individual weaknesses. By combining the ability of semi-supervised learning to leverage unlabelled data with the knowledge transfer capabilities of transfer learning, SSTL offers a promising solution to the persistent challenges in fault diagnosis for planetary gearboxes.

The primary contributions of this paper can be summarised as follows:

(1) A novel SSTL framework is proposed, which adeptly incorporates the characteristics of unlabelled data to enhance model training while simultaneously leveraging transfer learning techniques to map the source domain to the target domain. This innovative approach effectively addresses the critical challenge of low pseudo-label reliability in unlabelled data, thereby significantly improving the model's detection accuracy. Furthermore, this framework bridges the gap between supervised and unsupervised learning paradigms, offering a robust solution for scenarios where labelled data are scarce.
(2) A label migration and matching strategy is introduced to facilitate label transfer and alignment between homologous signals. This strategy effectively addresses the critical challenge of accurately assigning pseudo labels to transient interference signals, which has long been a bottleneck in signal processing and machine learning applications. By employing this method, the model achieves a higher degree of precision in identifying and categorising transient phenomena, thus enhancing its overall performance and reliability.
(3) A novel limiting normalisation strategy is proposed to mitigate the impact of transient interference on the model and stabilise the characteristic scale of vibration signals. This innovative approach not only enhances the efficiency of the model training process but also enables the effective development of a fault detection model capable of withstanding transient interference. Consequently,

this strategy significantly improves the model's robustness and generalisability, making it particularly suitable for real-world applications where signal noise and interference are prevalent.

The paper is organised as follows: In Section 2, a comprehensive review of the relevant literature on semi-supervised learning and transfer learning is provided, establishing the foundational concepts that underpin this research. Section 3 then elaborates on the details of the proposed SSTL approach, with a focus on preserving spatial homogeneity. Following this, Section 4 presents the experimental results obtained from two fault diagnosis datasets, offering a thorough analysis and comparison of the findings. Finally, Section 5 synthesises the key insights derived from the research and provides the concluding remarks of the study.

2. Preliminaries

2.1. Semisupervised learning

For the semi-supervised learning-based fault diagnosis, it is crucial to consider the nature and composition of the available data. Typically, the collected signals comprise two distinct datasets: a limited labelled dataset $\{x_i^{label}\}_{i=1}^{N}$ and a substantial unlabelled dataset $\{x_i^{unlabel}\}_{i=1}^{M}$. Within this framework, x_i represents individual time series samples, each with a sequence length l, where $i \in \{1, l\}$. The label $\{l_j^{label}\}_{j=1}^{N} \in \{1, 2, \ldots, C\}$ associated with the labelled dataset $\{x_i^{label}\}_{i=1}^{N}$ represents C distinct gear states. It is important to note that N and M denote the sizes of the labelled and unlabelled datasets, respectively, with the relationship $N \ll M$ holding true, emphasising the scarcity of labelled data relative to unlabelled data.

To quantify the proportion of labelled data in the semi-supervised learning context, we introduce the labelling rate α, defined as the ratio of N to $N + M$. One of the primary objectives in this domain is to effectively leverage the abundant unlabelled data in $\{x_i^{unlabel}\}_{i=1}^{M}$ to enhance the model's ability to fit the limited labelled data in $\{x_i^{label}\}_{i=1}^{N}$, thereby achieving performance levels that surpass those attainable through conventional supervised learning approaches. Concurrently, there is a strong emphasis on minimising α, as this directly translates to reduced time and cost associated with manual data annotation, an often resource-intensive process in real-world applications.

The landscape of semi-supervised learning methodologies can be broadly categorised into three distinct approaches, each with its own merits and challenges. The first approach, unsupervised pre-training, as exemplified by the works of Wang et al. [28] and Zhu et al. [29], involves an initial phase where the model learns representations from unlabelled data using unsupervised techniques, followed by a fine-tuning phase utilising the available labelled data. This approach leverages the abundance of unlabelled data to establish a robust foundation for feature extraction before refining the model with task-specific labelled data. The second category encompasses co-training methods, as demonstrated in the research of Zhang et al. [30], Li et al. [31], and Lee et al. [32]. These methods simultaneously train models using both labelled and unlabelled data, resulting in a composite loss function that combines a supervised loss L_s and an unsupervised loss L_u. The unsupervised loss L_u typically quantifies the discrepancy between data distributions, computed from a large sample of data. The final loss L is formulated as a weighted sum of these components: $L = L_s + \omega L_u$, where ω serves as a hyperparameter to balance the contributions of supervised and unsupervised learning objectives. The third approach, known as self-training, is exemplified by the work of Jiao et al. [33] and Pu et al. [34]. This iterative method begins by training the network on a small subset of labelled data. Subsequently, the trained model is employed to classify unlabelled samples, with those classified with high confidence being incorporated into the training set. This process is then repeated, gradually expanding the effective labelled dataset and refining the model's performance.

2.2. Transfer learning

In the domain of transfer learning, existing methodologies can be broadly categorised into two primary approaches: statistically based methods and adversarially based methods. While both aim to improve the transferability of knowledge across domains, they employ distinct strategies to achieve this goal. The fundamental principle underlying statistically based transfer learning methods, as elucidated by Chen et al. [35] and Zhang et al. [36], is the pursuit of domain-invariant representations. This is typically accomplished by minimising the distribution divergence between the source and target domains. By doing so, these methods strive to create a shared feature space that is less sensitive to domain-specific variations, thus facilitating more effective knowledge transfer. On the other hand, adversarially based methods, such as those proposed by He et al. [37] and Wang et al. [38], draw inspiration from the innovative framework of Generative Adversarial Networks (GANs). GANs, characterised by their zero-sum game dynamics, have emerged as a promising machine learning paradigm. In the context of transfer learning, adversarial approaches leverage this competitive mechanism to align features across domains.

To further illustrate the application of these concepts, Wang et al. [39] proposed a novel approach that utilises labelled data from both the source domain and a limited subset of the target domain. This method employs supervised training techniques for the feature extractor and classifier components. Additionally, to promote the learning of domain-invariant features, a discriminator is incorporated to align latent representations across domains. Addressing scenarios with extremely limited fault data, particularly in single-sample instances, Han et al. [40] introduced an innovative multi-domain discriminator. This enhancement aims to improve domain-invariant feature extraction, consequently boosting fault diagnostic performance in resource-constrained environments. In contrast to the approach presented by Han et al. [40], Li et al. [41] developed a method that leverages multiple classifiers, utilising label information for more accurate fault prediction. Furthermore, their approach incorporates a discriminator to align features between the source and target domains, thereby enhancing the overall transferability of the learned representations.

3. Semi-supervised transfer learning preserving spatial homogeneity

The proposed architecture is composed of four distinct yet intricately interconnected blocks, each fulfilling a critical function within the overarching framework: (1) a teacher-student model, (2) the construction of pseudo labels, (3) data matching with pseudo-labelling, and (4) normalisation with amplitude-limited. This innovative approach is designed to effectively leverage both labelled and unlabelled data in scenarios characterised by a paucity of labelled instances, thereby enhancing model performance and generalisation capabilities. The integration of the teacher-student model paradigm, in conjunction with the pseudo-labelling technique, enables the architecture to efficiently utilise both labelled and unlabelled data. This methodology is particularly advantageous in contexts where labelled data is scarce or prohibitively expensive to obtain, as it facilitates the exploitation of abundant unlabelled data to augment performance and generalisation capabilities. Moreover, the iterative nature of this process, wherein the student model has the potential to assume the role of the teacher in subsequent iterations, engenders continuous refinement and adaptation of the model to evolving data distributions.

The teacher-student model serves as the cornerstone of knowledge transfer within the architecture, facilitating the propagation of learned representations from a more experienced model (the teacher) to a less experienced one (the student). This transfer of knowledge accelerates the learning process and enhances the student model's ability to generalise from limited labelled data. Concurrently, the construction of pseudo labels represents a crucial step in leveraging unlabelled data. By assigning probabilistic labels to unlabelled instances, this component effectively expands the training set, allowing the model to learn from a broader range of examples. The data matching process with pseudo-labelling further enhances the model's ability to learn from unlabelled samples. By aligning the distributions of labelled and unlabelled data, this component ensures that the knowledge gained from pseudo-labelled instances is consistent with the underlying distribution of the labelled data. This alignment is critical for maintaining the integrity of the learning process and preventing potential biases that may arise from discrepancies between labelled and unlabelled data distributions. Finally, the normalisation with amplitude-limited components plays a crucial role in maintaining the stability and consistency of the learning process. By constraining the range of values within the network, this module mitigates the risk of exploding or vanishing gradients, which can impede effective learning. This normalisation process ensures that the model remains robust and stable throughout the training process, even when dealing with diverse and potentially noisy data sources.

The synergistic interaction among these components engenders a robust framework capable of effectively learning from both labelled and unlabelled data. This, in turn, potentially leads to improved performance across various machine learning. The proposed approach may prove particularly beneficial in domains where the acquisition of labelled data is challenging or resource-intensive, such as medical imaging, natural language processing, or autonomous systems. The holistic framework of this architecture, elucidating the interconnections and flows between the four primary components, is visually represented in Figure 1, while the comprehensive algorithm is elaborated in Algorithm 1 and Algorithm 2. They provide a comprehensive overview

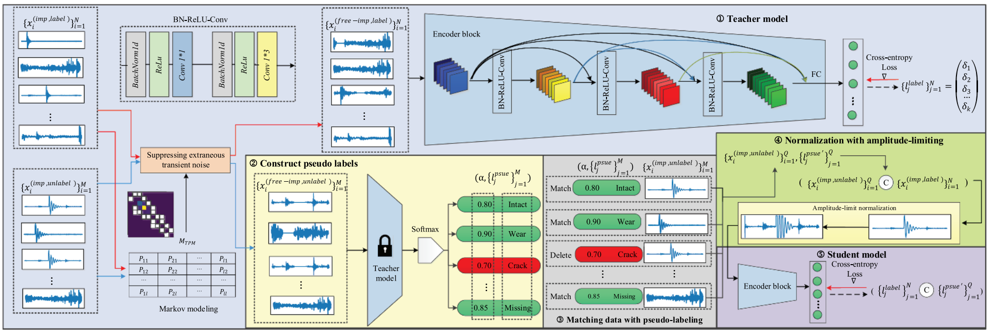

Figure 1. The proposed SSTL preserving spatial homogeneity.

of the proposed methodology and its constituent parts, facilitating a deeper understanding of the architecture's operational mechanics and the interplay between its various elements.

Semi-supervised learning scenarios involve labelled and unlabelled datasets contaminated by random transient noise interference. These datasets are denoted as $\{x_i^{(imp,label)}\}_{i=1}^{N}$ and $\{x_i^{(imp,unlabel)}\}_{i=1}^{M}$ respectively, where x_i represents an individual sample in the dataset, characterised by a length of L, the value of each data point is denoted as R, such that $x_i|R^{1\times L}$. The variables N and M represent the cardinality of the labelled and unlabelled datasets, respectively. Furthermore, the ratio $N/(N+M)$ represents the proportion of labelled data to the total data, referred to as the labelling rate, denoted by α. This parameter α serves as a critical hyper-parameter, where a higher value of α indicates an enhanced capacity of the model to extract meaningful features from the data.

The methodology begins with a crucial pre-processing step, wherein transient noise interference is systematically removed from both the labelled datasets $\{x_i^{(imp,label)}\}_{i=1}^{N}$ and the unlabelled datasets$\{x_i^{(imp,unlabel)}\}_{i=1}^{M}$. This noise removal process is essential for improving the signal-to-noise ratio and enhancing the quality of the input data, thereby facilitating more accurate subsequent analyses. Following this pre-processing, these refined datasets are denoted as $\{x_i^{(free-imp,label)}\}_{i=1}^{N}$ and $\{x_i^{(free-imp,unlabel)}\}_{i=1}^{M}$, respectively.

Subsequently, the dataset $\{x_i^{(free-imp,label)}\}_{i=1}^{N}$ is employed to train a teacher model using supervised learning techniques. The trained teacher model, leveraging its knowledge acquired from the labelled data, is then utilised to classify the dataset $\{x_i^{(free-imp,unlabel)}\}_{i=1}^{M}$. This classification process generates confidence scores, denoted by β_c, and pseudo labels, denoted by $\{l_j^{psue}\}_{j=1}^{M}$, for each sample in the unlabelled dataset.

The next phase involves a critical matching process, wherein the confidence scores β_c, pseudo labels $\{l_j^{psue}\}_{j=1}^{M}$, and the original noise-contaminated unlabelled dataset $\{x_i^{(imp,unlabel)}\}_{i=1}^{M}$ are combined to create a new dataset $\{x_i^{(imp,unlabel)}\}_{i=1}^{Q}$. This matching process is crucial for leveraging information from both labelled and unlabelled data, thereby enhancing the overall learning process. The matching algorithm employs a threshold-based approach to select high-confidence pseudo-labelled samples, ensuring that only the most reliable predictions from the teacher model are incorporated into the student model's training data.

In the final stage, a student model is trained using both the original labelled dataset $\{x_i^{(imp,label)}\}_{i=1}^{N}$ and the newly obtained dataset $\{x_i^{(imp,unlabel)}\}_{i=1}^{Q}$. This approach allows the student model to benefit from the knowledge distilled by the teacher model, as well as the additional information provided by the pseudo-labelled data. The student model's training process incorporates a carefully designed loss function that balances the contributions of labelled and pseudo-labelled samples, ensuring optimal learning from both sources.

3.1. Teacher model

To identify and characterise transient interference in the datasets $\{x_i^{(imp,label)}\}_{i=1}^{N}$ and $\{x_i^{(imp,unlabel)}\}_{i=1}^{M}$, a first-order Markov model [4] is employed for signal analysis. This approach begins by discretising the continuous signal $x_i|R^{1\times L}$ into a finite number of states. For each value R in the signal, the corresponding interval number j is determined, which is then treated as the state S_i of the signal at that particular moment. This process effectively transforms each continuous state into an independent state sequence, as defined by the following equation:

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

where $k \in (1, L)$. n_{bin} represents a predefined number of states. It is important to note that the size of n_{bin} directly affects the discrete state density, thus influencing the granularity of the analysis. After partitioning x_i into n_{bin} states, each state S_j is defined by the following formula.

$$\begin{gathered} S_j = [min(x_i) + (j-1)\cdot\Delta, \mathrm{min(x_i)} + \mathrm{j}\cdot\Delta] \\ \Delta = \frac{\mathrm{max(x_i)} - \mathrm{min(x_i)}}{\mathrm{n_{bin}}} \end{gathered} \tag{2}$$

where Δ represents the equidistant length of the partition, ensuring uniform state intervals. Subsequently, based on the state sequence S_i, a statistical analysis is performed to determine the frequency of state transitions. This analysis is used to construct a Markov transition matrix, MTM_{ij}, which is mathematically expressed as:

$$MTM_{ij} = \sum_{k=1}^{L-1} \delta(S_n = i, S_{n+1} = j) \tag{3}$$

where $\delta(x, y)$ denotes the Kronecker Delta function, which equals 1 when $x = y$ and 0 otherwise. To derive meaningful transition probabilities, MTM_{ij} is normalised by rows, resulting in the Markov transition probability matrix, M_{TPN}. This matrix represents the probability of each state i transitioning to state j:

$$\begin{gathered} P_{ij} = \frac{MTM_{ij}}{\Sigma_{\mathrm{j=1}}^{\mathrm{N}}\mathrm{MTM_{ij}}} \\ M_{TPM}(i,j) = P_{ij} \end{gathered} \tag{4}$$

The element $M_{TPM}(i,j)$ represents the likelihood of the system transitioning from the current state to another state. Under normal conditions, the signal's state transitions should exhibit smooth and coherent characteristics, with the discretised signal highly concentrated around the predefined discrete intervals. However, in the presence of transient interference, the signal's state transitions may experience abrupt changes. Therefore, $M_{TPM}(i,j)$ serves as a powerful tool for localising transient interference within the signal.

To identify areas of extreme transition, a threshold t_p is established for P_{ij}. This threshold is a critical hyperparameter in the analysis. When P_{ij} exceeds this threshold, the corresponding region is classified as an area of extreme transition, indicating the presence of transient interference. The datasets $\{x_i^{(imp,label)}\}_{i=1}^{N}$, $\{x_i^{(imp,unlabel)}\}_{i=1}^{M}$, and the transition probability matrix $M_{TPM}(i,j)$ are then input into a supporting extraneous transient noise module. This module is designed to eliminate transient interference using a set-to-zero processing method, defined as:

$$x_i[i] = \begin{cases} 0, & P_{ij} \geq t_p \\ x_i[i], & P_{ij} < t_p \end{cases} \tag{5}$$

This process yields a labelled dataset $\{x_i^{(free-imp,label)}\}_{i=1}^{N}$ and an unlabelled dataset $\{x_i^{(free-imp,unlabel)}\}_{i=1}^{M}$, both free from transient interference.

Following the noise removal process, a backbone Encoder block $E(\cdot)$ is employed to extract high-level semantic features $z_i = E(x_i)$ from each sample. These features are then input into a classification header $C(\cdot)$, which outputs a probability distribution vector $p_i = C(z_i)$ representing the predicted health status. The model's performance is evaluated using cross-entropy loss L_s, calculated based on the prediction results p_i and the true labels $\{l_j^{label}\}_{j=1}^{N}$:

$$L_s = -\frac{1}{N}\sum_{i=1}^{N} l_i^{label} \cdot log(p_i) \tag{6}$$

This loss function is used to train the teacher model, after which the model parameters are locked to preserve the learned knowledge.

3.2. Construct pseudo labels

In the second stage of the process, the set of unlabelled, interference-free samples, denoted as $\{x_i^{(free-imp,unlabel)}\}_{i=1}^{M}$, undergoes classification using the pre-trained teacher model. Subsequently, the resulting output is processed through a SoftMax function, which can be mathematically expressed as follows:

$$\beta_c = Soft\,Max(p_i) \tag{7}$$

The SoftMax function serves a crucial role in this context, as it normalises the elements of an input vector to values between 0 and 1, while ensuring that the sum of all elements equals 1. This normalisation property is particularly useful for probabilistic interpretations. The output of the SoftMax function is defined as the confidence score β_c, which represents the probability that a given sample belongs to a specific category. It is important to note that the threshold β for β_c is a hyper-parameter that requires careful tuning. In the proposed methodology, pseudo labels

$\{l_j^{psue}\}_{j=1}^{M}$ with confidence scores $\beta_c \geq 0.9$ are considered to be reliable and are treated as true labels. This threshold selection is critical for maintaining the quality of the pseudo-labelling process. When the model's predicted probability for a given sample reaches or exceeds 0.9, the softmax output exhibits a pronounced peak, indicating a high degree of certainty in the classification. According to the Maximum A Posteriori (MAP) estimation principle in probability theory, such high-confidence predictions are more likely to correspond to the true class label, as the posterior probability of the predicted category substantially exceeds that of alternative categories. Under these conditions, the assignment of pseudo-labels can be performed with a relatively low error rate, thereby enhancing the reliability of semi-supervised learning. An essential observation is that $\{x_i^{(free-imp,unlabel)}\}_{i=1}^{M}$ is derived from $\{x_i^{(imp,unlabel)}\}_{i=1}^{M}$ after the removal of interference noise. Consequently, the labels for these two sets of data should be identical, both represented by $\{l_j^{psue}\}_{j=1}^{M}$. This consistency in labelling is crucial for maintaining the integrity of the dataset throughout the noise removal process.

The third stage of the procedure involves pairing the corresponding confidence scores β_c, pseudo labels $\{l_j^{psue}\}_{j=1}^{M}$, and the original unlabelled samples with interference $\{x_i^{(imp,unlabel)}\}_{i=1}^{M}$. Subsequently, pairs where the confidence score falls below the threshold (i.e. $\beta_c < \beta$) are eliminated from the dataset. This filtering process results in the creation of new, refined datasets: $\{x_i^{(imp,unlabel)}\}_{i=1}^{Q}$ and its corresponding set of pseudo-labels $\{l_j^{psue'}\}_{j=1}^{Q}$. These refined datasets are expected to contain more reliable samples and labels, which can potentially improve the performance of subsequent machine learning tasks.

3.3. Normalisation with amplitude-limiting

In the fourth stage of the process, both the labelled samples with interference $\{x_i^{(imp,label)}\}_{i=1}^{N}$ and the unlabelled samples with interference $\{x_i^{(imp,unlabel)}\}_{i=1}^{Q}$ undergo amplitude-limited normalisation, constraining their values to the range of 0 to 1. This normalisation step is crucial for ensuring consistency in the data representation and facilitating subsequent analysis. Figure 2 provides a comprehensive visual representation of the various stages of signal processing. Specifically, Figure 2(a) illustrates the original signal samples containing transient interference, while Figure 2(b) depicts the corresponding samples after the removal of transient interference. A comparative analysis of these figures reveals the significant impact of transient interference on the signal amplitude. Notably, samples with transient interference obscure the underlying vibration data fluctuations, whereas the removal of such interference unveils the intrinsic amplitude variations of the vibration signal. Figure 2(c,d) demonstrates the outcomes of conventional signal normalisation techniques. Although this approach confines the data within the (0,1) range, it fails to adequately mitigate the effects of transient interference. In cases where the amplitude of transient interference is substantial, traditional normalisation may compress the vibration characteristics to

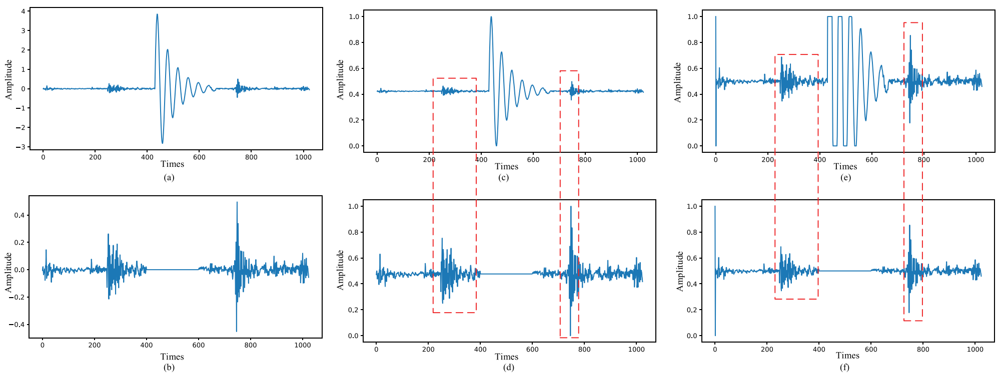

Figure 2. (a) Original sample containing transient interference; (b) processed sample after removal of transient interference from (a); (c) samples produced via the application of traditional normalisation to the raw signal (a), with amplitude constrained to the [0, 1] range; (d) samples generated through traditional normalisation applied to signal (b), confined to the [0, 1] amplitude range; (e) samples generated through amplitude-limited normalisation of signal (a); and (f) samples produced by amplitude-limited normalisation of signal (b).

approximately 0.42, severely impeding the model's ability to extract meaningful gear vibration features.

To address these limitations, a new Amplitude-limited Normalisation (ALN) method is proposed. This approach constrains the vibration signal amplitude within a predefined range of maximum ALT and minimum $-ALT$, where Amplitude-limited Threshold (ALT) is a carefully selected hyper-parameter. The selection of the ALT must be carefully tailored to the characteristics of each dataset in order to optimise model accuracy and enhance training efficiency. The mathematical formulation of this technique is expressed as follows:

$$x_i[j] = \begin{cases} x_i[0] = ALT, & j = 0 \\ x_i[1] = -ALT, & j = 1 \\ -ALT, & x_i[j] \leq -ALT \\ x_i[j], & -ALT < x_i[j] \leq ALT \\ ALT, & x_i[j] > ALT \end{cases} \tag{8}$$

where $j \in \{1, l\}$, l is defined as the sample length.

The amplitude normalisation ensures that vibration signals of the same category exhibit features of comparable scale post-normalisation. As illustrated in Figure 2 (e) and Figure 2(f), this method effectively constrains transient interference within the $(-ALT, ALT)$ range while preserving the characteristics of the vibration signal in regions unaffected by interference. It is important to note that the amplitude of xi may not always reach ALT. In the traditional normalisation procedure for processing vibration signals subjected to random transient noise, there is a corresponding variation in amplitude scale. Consequently, signals of the same type may exhibit diverse amplitude scales, potentially leading to inaccuracies in feature extraction and obscuring the identification of useful or latent features. To address this issue, the amplitude of the vibration signal is constrained prior to normalisation, specifically setting $x_i[0] = ALT$ and $x_i[1] = -ALT$ in Equation 8.

This modification ensures that signals with exceptionally small vibration amplitudes maintain consistent feature scales after normalisation. The selection of an appropriate ALT value is critical and requires empirical testing, as different datasets may necessitate different ALT values. Setting ALT too high may fail to adequately limit the impact of amplitude on normalisation, while setting it too low risks losing essential characteristics of the vibration signal. Figure 2 provides a comprehensive visual comparison of the various signal processing stages. It is crucial to emphasise that amplitude-limited normalisation is not applied prior to training the teacher model. This decision is based on the fact that the teacher model is trained using data without transient interference, and amplitude-limited normalisation, while not eliminating transient interference, may result in the loss of significant vibration features.

3.4. Student model

In the final stage of this process, the encoder function $E(\cdot)$ is employed to extract high-dimensional features from the normalised input x_i^{nor}. X_i^{nor} represents the output of the amplitude-limited normalisation process applied to both$\{x_i^{(imp,unlabel)}\}_{i=1}^{Q}$ and $\{x_i^{(imp,label)}\}_{i=1}^{N}$. This step is crucial for capturing the intricate characteristics of the vibration signals, which are inherently complex in nature. It is important to note that transient interference in vibration signals is typically localised, affecting only a small portion of the time series rather than permeating the entire signal. Consequently, the data characteristics of $\{x_i^{(imp,unlabel)}\}_{i=1}^{Q}$ and $\{x_i^{(imp,label)}\}_{i=1}^{N}$ remain consistent in regions unaffected by transient interference. This property allows for the inclusion of$\{x_i^{(imp,label)}\}_{i=1}^{N}$ in the training set of the student model, even after amplitude-limited normalisation has been applied.

The encoder's output, denoted as $z_i^{student} = E(x_i^{nor})$, is subsequently processed through a fully connected layer to produce the distribution vector $p_i^{student} = C(z_i^{student})$. The cross-entropy loss is then calculated using the following equation:

$$L_s^{student} = -\frac{1}{N+Q}\sum_{i=1}^{N+Q} l_i^{label'} \cdot log(p_i^{student}) \tag{9}$$

This loss function is used to perform gradient updates on the model weights, thereby training the student model. It is worth noting that $l_i^{label'}$ is composed of $\{l_j^{label}\}_{j=1}^{N}$ and $\{l_j^{psue'}\}_{j=1}^{Q}$, encompassing both the original labels and the refined pseudo labels.

4. Experimental validation and comparative analysis

4.1. Case study I

4.1.1. Specifications for data description and test-rig

The gearbox dataset, meticulously collected from a sophisticated gear transmission system, provides a comprehensive representation of various operational conditions and

Algorithm 1 Training teacher model

Initialize parameters, epoch $\leftarrow E$.
Input: $\{x_i^{(imp,label)}\}_{i=1}^{N}$ and $\{x_i^{(imp,unlabel)}\}_{i=1}^{M}$.
Locate transient pulses position (TPP) TPP_{label} and $TPP_{unlabel}$.
$\{x_i^{(imp,label)}\}_{i=1}^{N}(TPP_{label}) = 0$ to get $\{x_i^{(free-imp,label)}\}_{i=1}^{N}$.
$\{x_i^{(imp,unlabel)}\}_{i=1}^{M}(TPP_{unlabel}) = 0$ to get $\{x_i^{(free-imp,unlabel)}\}_{i=1}^{M}$.
for each epoch in $1 : E$ **do**
    $\{x_i^{(free-imp,label)}\}_{i=1}^{N}$ was used to train Encoder block.
end forreturn $\{x_i^{(free-imp,label)}\}_{i=1}^{N}$, $\{x_i^{(free-imp,unlabel)}\}_{i=1}^{M}$ and teacher model $\Theta_t(\cdot)$.

Algorithm 2 Training student model

Initialize parameters, $\beta \leftarrow \beta$, $ALT \leftarrow ALT$, epoch $\leftarrow E$, teacher model $\Theta_t(\cdot)$.
$\{l_j^{psue}\}_{j=1}^{M} = \Theta(\{x_i^{(free-imp,unlabel)}\}_{i=1}^{M})$.
Matches $\{x_i^{(imp,unlabel)}\}_{i=1}^{M}$ and $\{l_j^{psue}\}_{j=1}^{M}$.
if $\{l_j^{psue}\}_{j=1}^{M} < \beta$ **then**
    Delete this sample, get $\{x_i^{(imp,unlabel)}\}_{i=1}^{Q}$ and $\{l_j^{psue'}\}_{j=1}^{Q}$.
end if
Dataset $D = \{x_i^{(imp,unlabel)}\}_{i=1}^{Q} \mathbb{C}\{x_i^{(imp,label)}\}_{i=1}^{N}$.
if $x_i\,[j] > ALT$ **then**
    $x_i\,[j] = ALT$.
end if
if $x_i\,[j] < -ALT$ **then**
    $x_i\,[j] = -ALT$.
end if
Get dataset D'
Normalize D' to [0,1] and get dataset D_nor.
for each epoch in $1 : E$ **do**
    D_{nor} was used to train Encoder block.
end forreturn A gear fault diagnosis model $\Theta_s(\cdot)$ with resistance to transient disturbance.

fault types. This system, as illustrated in Figure 3, comprises several principal components, including a tachometer, driven motor, torque transducer, two-stage parallel gearbox system, load gearboxes, and load motor. The placement of the accelerometer is especially noteworthy, as it is affixed to a separate disk. For a closer examination, a detailed view of this configuration is provided in the zoomed section of Figure 3. To ensure a high-fidelity representation of the system's dynamics, the dataset is sampled at a frequency of 12.8 kHz. Furthermore, it encompasses a range of operational conditions, with rotational speeds systematically varied from 1600 to 2400 r/min. In addition to normal operating conditions, the dataset incorporates five common gear fault types, illustrated in Figure 4, namely: miss (missing tooth), chip (cracked teeth), root (crack at tooth root), surface (wear on gear surface), and eccentric (misaligned geometric and rotational centres). The gear meshing configuration is depicted in Figure 5(a), while Figure 5(b) illustrates the internal configuration of the parallel gearbox system. In the latter, the faulty gear is clearly demarcated with a dotted box for ease of identification.

To facilitate in-depth gear diagnosis analysis, vibration data is collected along the x-axis of the accelerometer while the gear rotates at a constant speed of 1600 rpm. Each category, including the healthy condition, comprises 768,000 data points gathered over

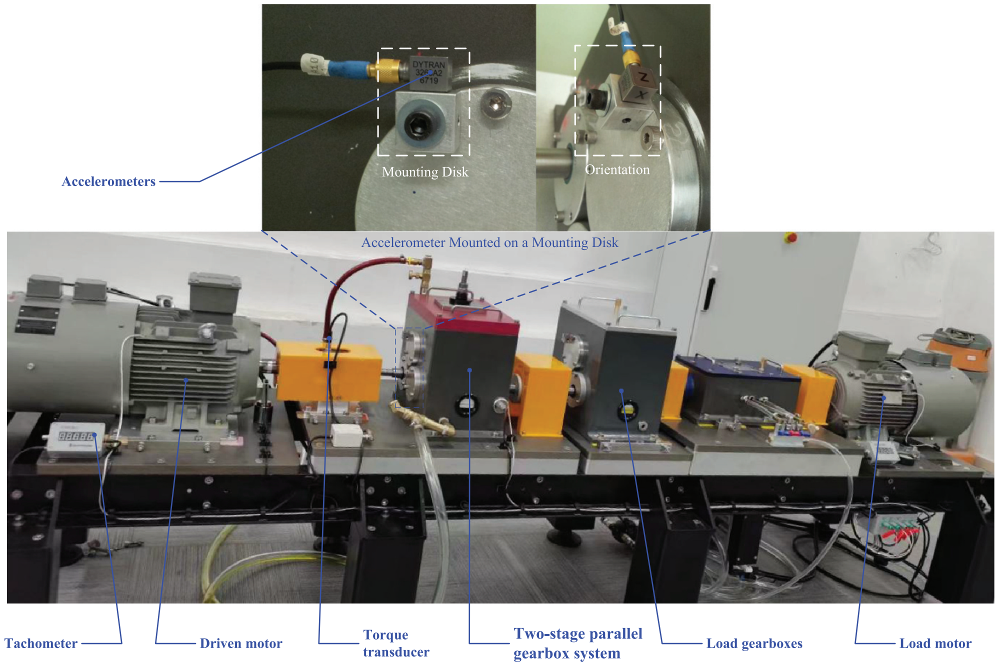

Figure 3. Experimental test-rig of gear transmission system.

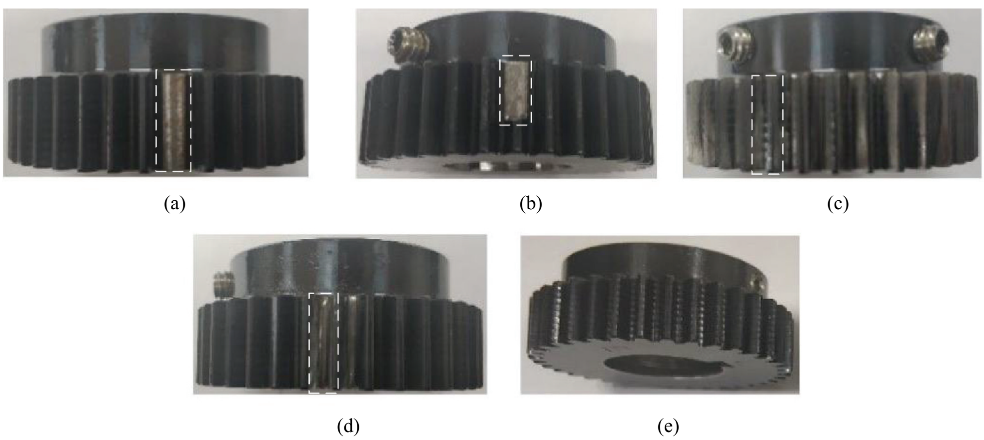

Figure 4. (a) Miss (missing tooth), (b) chipped (cracked teeth), (c) surface (wear on gear surface), (d) root (crack at tooth root), (e) eccentric (misaligned geometric and rotational centres).

a 60-second duration. This extensive dataset provides a robust foundation for the development and validation of fault diagnostic algorithms, enabling researchers to explore a wide range of operational conditions and fault types within a controlled experimental setting.

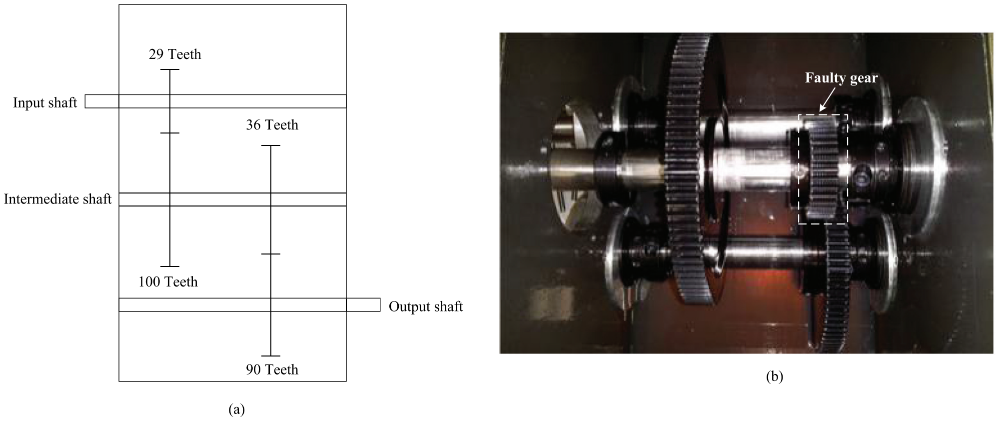

Figure 5. (a) Gear meshing, (b) Internal configuration of parallel gearbox system.

4.1.2. Comparative networks and results analysis

The primary objective of this research is to examine and analyse the performance variations exhibited by gearboxes when subjected to transient noise disturbances across a diverse range of environmental conditions and fault scenarios. In order to conduct this research with rigorous methodology, external impacts such as transient noise are carefully measured and systematically combined with the gearbox vibrations during the data collection process, thereby obtaining a realistic representation of transient noise disturbances. To ensure the integrity of the testing procedure and mitigate potential bias, the dataset is initially partitioned into two distinct subsets: a training set comprising 80% of the data, and a test set encompassing the remaining 20%. Subsequently, with the aim of simulating real-world scenarios of label scarcity, the number of labelled samples is carefully determined based on the parameter α, which is previously introduced in section 3.2 For any given set of N samples representing an identical state, only a fraction $(\alpha \times N)$ of these samples retain their original labels, while the labels for the remaining samples are systematically removed. This approach allows for a controlled simulation of varying degrees of label availability. To comprehensively assess the model's performance under different levels of label scarcity, the dataset is evaluated using three distinct label rates: 5%, 10%, and 15%.

To assess the superiority of the proposed method, it is rigorously compared with state-of-the-art (SOTA) techniques, which include three supervised learning methods and three semi-supervised learning methods. These methods include: DenseNet (Impulse noise), a supervised approach that exclusively utilises $\{x_i^{(imp,label)}\}_{i=1}^{N}$ to train the Encoder block and classification head; DenseNet (Denoise), another supervised approach that trains the Encoder block and classification heads using only $\{x_i^{(free-imp,label)}\}_{i=1}^{N}$; and DenseNet (Amplitude limited), which trains the Encoder block and classification heads using Amplitude-limited $\{x_i^{(imp,label)}\}_{i=1}^{N}$. Additionally, the comparison encompasses SimCLR [42], a contrastive learning method that employs additional negative samples and a projection layer for training; Fast-MoCo [43], an advanced contrastive learning method that increases the number of negative pairs using momentum encoders and a memory bank and is jointly trained with supervised learning; and ITSSL [44], a semi-supervised method that utilises time-amplitude data augmentation techniques for training. This comprehensive comparison allows for a thorough evaluation of the proposed method's effectiveness across various learning paradigms and techniques.

The successful training of a deep learning model is largely contingent upon the selection of appropriate hyper-parameters. To validate the generality of the proposed semi-supervised transfer (SSTL) framework, both teacher and student models employed the widely adopted DenseNet121 architecture. DenseNet121 features a densely connected structure that facilitates efficient extraction of sample features and has demonstrated excellent performance across a variety of models. Initially, the learning rate is set to 0.0001, with the customary practice of commencing training gradually and making adjustments as the process unfolds. Subsequently, the learning rate for each epoch is halved, a strategy aimed at efficiently converging the model. The optimiser employed in this architecture is Adam, a popular optimisation algorithm renowned for its adaptive learning rate adjustments across different parameters. Adam ingeniously combines the principles of RMSprop and momentum optimisation, rendering it suitable for a diverse array of deep learning tasks. The loss function designated for the training procedure is CrossEntropyLoss, which is commonly utilised for classification problems to minimise the discrepancy between predicted and actual class labels. This function measures the differences between probability distributions, making it an apt choice for training classification models. Furthermore, the framework incorporates the Gaussian Error Linear Unit (GELU) activation function. GELU is a non-linear activation function that has garnered attention in recent years due to its capacity to enhance the performance of deep neural networks. To mitigate overfitting and improve generalisation, a dropout rate of 0.1 is applied. Dropout is a regularisation technique that randomly sets a small fraction of input units to zero during training, thereby preventing the model from over-relying on specific features and enhancing its ability to generalise to unseen data. Lastly, the training process is conducted over 50 epochs. An epoch refers to a complete pass through the entire training dataset, and training over multiple epochs allows the model to iteratively update its parameters and learn more thoroughly from the data. The highest test accuracies achieved across all datasets are presented in Table 1, providing a comprehensive overview of the model's performance under various conditions and configurations.

The comparative results, indicated by accuracies across all datasets as presented in Table 1, reveal several significant insights into the performance of various learning methods. A notable inverse relationship between label rate and model performance is observed, with all methods experiencing a precipitous decline in accuracy as the label rate decreases, thus highlighting the detrimental impact of label scarcity. Among the evaluated methods, SSTL consistently demonstrates superior performance, achieving the highest accuracy across various label rates and showcasing its robustness in diverse data scenarios. When comparing semi-supervised learning approaches with amplitude-limited normalisation and supervised learning, the latter two exhibit lower training accuracy, suggesting that amplitude-limited normalisation may excessively attenuate time-related features in supervised learning contexts. Conversely, semi-supervised training, while capable of extracting sufficient time features, is more susceptible to noise interference. Interestingly, within the semi-supervised learning paradigm, amplitude-

Table 1. Comparative analysis of experimental results for case study I.

Accuracy(%) / Label rate / Model	5%	10%	15%
DenseNet (Impulse noise) [45]	28.8±0.18	48.7±0.03	68.3±0.08
DenseNet (Denoise)	73.5±0.5	85.5±0.11	90.5±0.15
DenseNet (Amplitude-limited)	32.3±0.96	54.6±0.06	75.4±0.07
SimCLR [42]	83.3±0.35	88.6±0.5	97.1±0.35
Fast-MoCo [43]	84.6±0.31	91.5±0.16	**97.4±0.49**
ITSSL [44]	**87.1±0.37**	**93.4±0.39**	95.5±0.57
SSTL (Ours)	**95.3±0.24**	**97.5±0.34**	**98.5±0.15**

limited normalisation strikes a balance by sacrificing some time characteristics without compromising model accuracy, while effectively mitigating the impact of transient noise. Furthermore, semi-supervised learning methods consistently outperform their supervised counterparts, with ITSSL achieving commendable accuracy but still falling short of SSTL's performance. This discrepancy can be attributed to ITSSL's inability to completely eliminate transient noise and its vulnerability to challenging samples, whereas the proposed method incorporates a confidence-based sample selection mechanism, resulting in a more stable model fitting process characterised by faster convergence and higher accuracy.

For further comparative analysis, a 15% labelling rate is employed in Case Study I for experimental validation, with results depicted in Figure 6. The proposed SSTL demonstrated a remarkable improvement in diagnostic accuracy by 3%. While the supervised learning method such as Densenet (Denoise) achieved an accuracy of 90.5%, and the semi-supervised learning method reached 97.4%, SSTL attained an impressive 98.5%. These results emphatically underscore SSTL's superiority over both supervised and other semi-supervised methods, demonstrating the efficacy of the proposed SSTL in capturing underlying features of unlabelled signals and consequently enhancing diagnostic performance in scenarios with limited labelled data. Furthermore, Figure 6 illustrates accuracy fluctuations of SSTL and other comparative methods in the SSL with a 5% labelled rate. In this most challenging semi-supervised learning scenario, SSTL achieves remarkable performance, attaining an average accuracy of 95.3%. Compared to traditional supervised learning methods, SSTL's performance improves by 66.5%, 21.8%, and 63%, respectively. In relation to other semi-supervised learning methods, SSTL's performance increases by 12%, 10.7%, and 8.2%, respectively. Moreover, SSTL exhibits stable performance across experiments, with accuracy consistently ranging from 93% to 97%, significantly surpassing results obtained from most supervised and semi-supervised learning methods. These experimental results provide compelling evidence that the proposed framework not only outperforms traditional supervised learning methods and other semi-supervised approaches but also effectively leverages underlying features from unlabelled data. Consequently, it significantly enhances diagnostic performance in scenarios characterised by data scarcity, thus addressing a critical challenge in machine learning applications.

4.1.3. Comparison of ablation experimental results

(a) Influence of Aptitude Limiting Threshold (ALT)

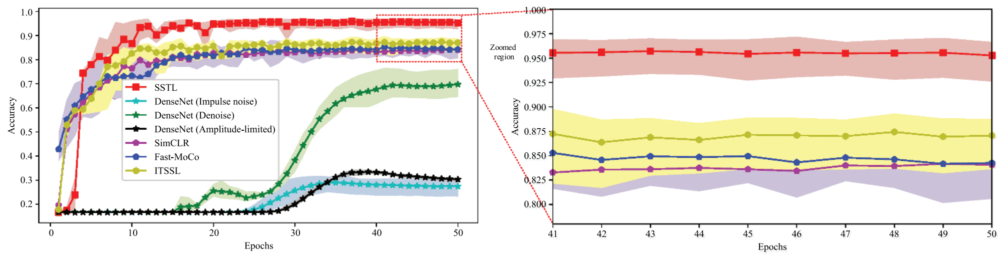

Figure 6. Experimental results for various methods applied to case study I with 5% of the data labelled.

In order to comprehensively investigate the impact of the Amplitude Limiting Threshold (ALT), a crucial hyper-parameter that influences the amplitude scale of vibration signals, a meticulous ablation study is conducted. Figure 7 presents a detailed illustration of the test accuracy results obtained from various ALT configurations of the SSTL model at a 5% labelled rate. Moreover, to facilitate a more in-depth analysis of the model's performance stability, a comparative analysis of the accuracy convergence is provided in the highlighted and zoomed region in Figure 7. The results of this study consistently demonstrate that SSTL implementations incorporating the amplitude-limited normalisation strategy exhibit significantly superior performance compared to their counterparts that do not employ this strategy. This observation strongly suggests that the amplitude-limited normalisation strategy effectively mitigates the detrimental impact of transient interference on vibration signals, thereby enhancing the overall performance of the model.

Furthermore, a notable trend emerges as the ALT value decreases: the test accuracy gradually improves, and concurrently, the range of fluctuation in test accuracy becomes markedly smaller. This inverse relationship between ALT and performance metrics indicates that the reduction of ALT serves a dual purpose. Firstly, it amplifies the amplitude features of vibration signals, making them more pronounced and discernible. Secondly, it constrains these features to a uniform amplitude scale, which, in turn, substantially enhances the SSTL model's feature extraction capabilities.

(b) Influence of β

To evaluate the impact of the hyper-parameter β on the proposed SSTL model, a series of ablation experiments are conducted. These experiments are designed to systematically investigate the degree of influence that β exerts on the model's performance. Specifically,

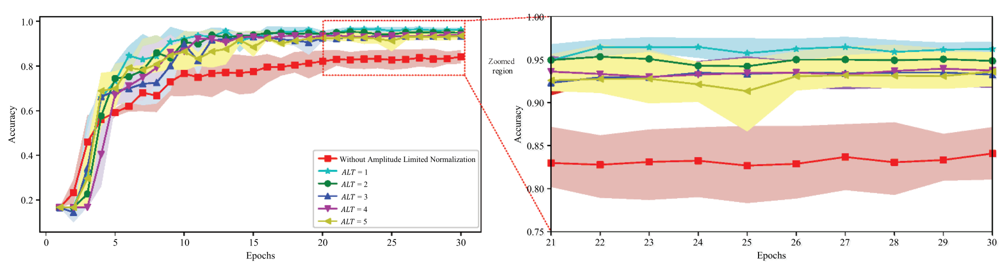

Figure 7. Impact of amplitude-limited.

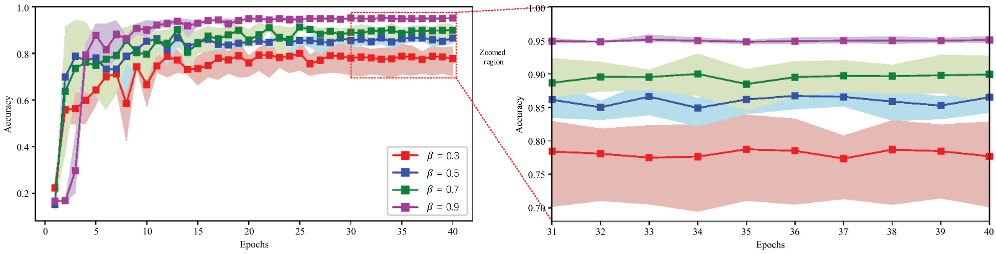

Figure 8. Impact of β.

a range of β values are tested to quantitatively assess their impact on SSTL's efficacy and stability. Figure 8 presents the test accuracy of SSTL for various values at a 5% label rate. Additionally, a zoomed region in Figure 8 is provided to offer a more detailed view of the model's performance during the crucial accuracy convergence stages of training.

The experimental results reveal a notable correlation between the magnitude of β and the model's performance. When β is assigned a relatively small value, the test accuracy of SSTL exhibits significant fluctuations, indicating a lack of stability in the learning process. This instability can be attributed to the low credibility of pseudo labels generated when β is small, which consequently results in a higher proportion of samples that deviate from the true labels. As a result, this phenomenon increases the complexity and difficulty of SSTL training. Conversely, as β increases, the test accuracy of SSTL demonstrates markedly reduced fluctuations, ultimately approaching a consistent level of nearly 95%. This observation suggests that a larger β value is effective in extracting high-quality samples from the dataset. Furthermore, it mitigates the challenges associated with insufficient labelled data, thereby supporting more robust and effective SSTL training. In conclusion, these findings underscore the critical role of β in optimising SSTL performance.

4.1.4. Comparative analysis of visualisations

To conduct a more rigorous quantitative analysis of the diagnostic results, a Confusion Matrix (CM) is employed in this case study. The results for a 5% labelled rate are illustrated in Figure 9, thereby providing a visual representation of the model's performance across various fault types. Upon careful examination of the data, it becomes evident that the proposed SSTL method, as depicted in Figure 9(g), successfully identifies all fault types with remarkably high accuracy. Notably, it achieves 100% accuracy for Chipped faults and impressive 98.8% for Eccentric faults, thus demonstrating its superior diagnostic capabilities. In contrast, the supervised models, specifically DenseNet (Impulse noise) illustrated in Figure 9(a) and DenseNet (Denoise) shown in Figure 9(c), exhibit significant limitations in their ability to accurately identify gear faults. While DenseNet (Denoise) demonstrates a rudimentary capacity to differentiate between gear faults, it only achieves precise diagnosis for chipped and surface faults, thereby highlighting its restricted applicability in comprehensive fault detection scenarios.

Furthermore, the semi-supervised models, including SimCLR (Figure 9(d)), Fast-MoCo (Figure 9(e)), and ITSSL (Figure 9(f)), display improved fault detection

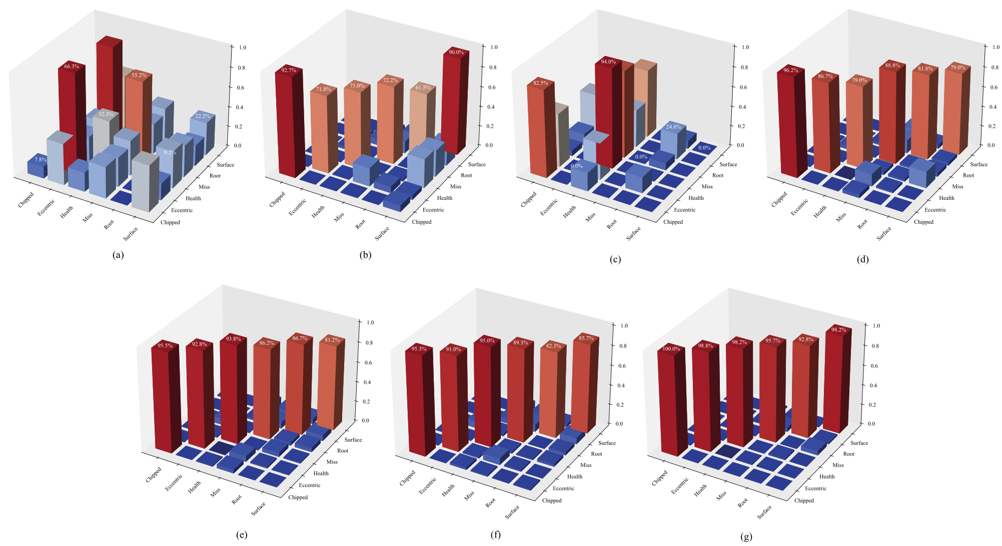

Figure 9. Classification performance via confusion matrix: (a) DenseNet (Impulse noise), (b) DenseNet (denoised), (c) DenseNet (amplitude-limited), (d) SimCLR, (e) Fast-MoCo, (f) ITSSL, (g) SSTL (Ours).

capabilities when compared to their supervised counterparts. However, it is important to note that their average accuracy remains below 87.1%, indicating room for improvement in their diagnostic precision. In stark contrast to these aforementioned models, the SSTL method demonstrates exceptional performance across the board. It accurately diagnoses all six fault types with an impressive average accuracy of 95.3%, and notably achieves over 95% accuracy for five of the fault types. This remarkable performance underscores the robustness of the SSTL framework against transient interference and its effectiveness in fault diagnosis tasks, particularly in scenarios where the availability of training data is limited.

In order to conduct a comprehensive comparative analysis of the captured features across various models, this study employs the t-distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction technique. This advanced visualisation method allows for a more intuitive understanding of the high-dimensional feature spaces. The comparative results of this analysis are presented in Figure 10. The visualisation results for the typical models, including DenseNet (Impulse noise), DenseNet (Denoised), and DenseNet (Amplitude-limited), are illustrated in Figure 10(a-c). These visualisations reveal that the scatter points representing different damage categories are largely clustered together, with minimal separation between classes. This clustering suggests that these models struggle to effectively differentiate between various failure types, indicating a limited capacity for fault identification. In contrast, the models employing semi-supervised learning approaches, namely SimCLR, Fast-MoCo, and ITSSL, demonstrate an enhanced ability to distinguish between different gear failures, as evidenced in Figure 10(d-f). This improvement can be attributed to the additional information leveraged through the semi-supervised learning paradigm. Specifically, the Fast-MoCo model, as depicted in Figure 10(e), exhibits a notable capability to identify chipped and eccentric gear failures. This is evident from the well-clustered feature

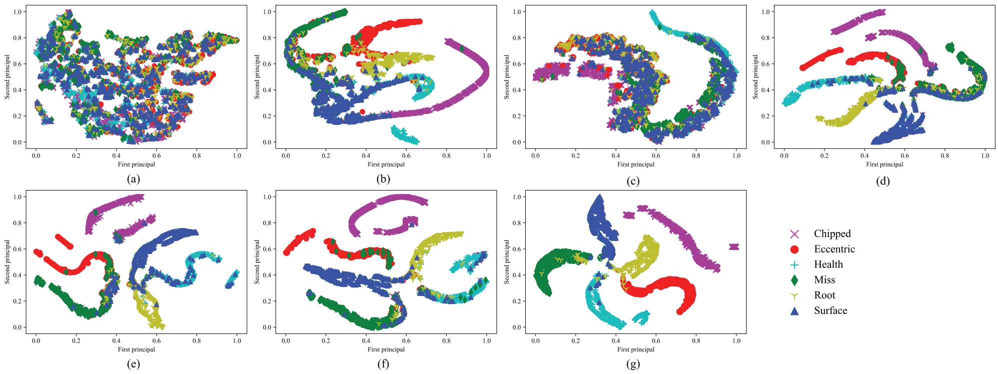

Figure 10. Feature visualisation using t-SNE: (a) DenseNet (Impulse noise), (b) DenseNet (denoised), (c) DenseNet (amplitude-limited), (d) SimCLR, (e) Fast-MoCo, (f) ITSSL, (g) SSTL (Ours).

manifolds corresponding to these failure types. However, it is important to note that the model still faces challenges in differentiating between root, surface, and missing tooth failures, as indicated by the overlapping feature manifolds for these categories. The proposed SSTL method, illustrated in Figure 10(g), demonstrates superior performance in feature discrimination. The feature manifold distributions for almost all failure types are clearly distinguishable, with scatter points that are either well-clustered or distinctly separable. This visual evidence strongly suggests that the SSTL model possesses a remarkable ability to accurately capture and discriminate the latent characteristics associated with different failure scenarios.

4.2. Case study II

4.2.1. Specifications for data description and test-rig

In order to further validate and rigorously test the proposed method, an experimental apparatus known as the Drivetrain Prognostics Simulator (DPS), as illustrated in Figure 11, is employed for case study II. This test-rig, manufactured by SpectraQuest Inc., is specifically chosen for its ability to provide complex drivetrain dynamics under controlled conditions. The DPS comprises several intricately interconnected components, each of which plays a crucial role in the overall system functionality. These components include: a variable speed drive motor, which provides the primary motive force; a planetary gearbox system, which offers a compact and efficient means of power transmission; a two-stage parallel gearbox system, which allows for further speed and torque modifications; resistance-load gear boxes coupled with a resistance-load inducing electric load motor, which simulate various operational loads; and an electric control unit that orchestrates and manages the entire configuration. In this experimental protocol, the signal sampled from the planetary gearbox transmission system is selected for analysis. This choice is motivated by the complex dynamics exhibited by planetary gearboxes and their widespread use in various industrial settings. The data acquisition process is carefully designed to ensure high-quality, high-resolution data collection. Specifically, the horizontal position signal is captured at a sampling frequency of 30,720 Hz, which

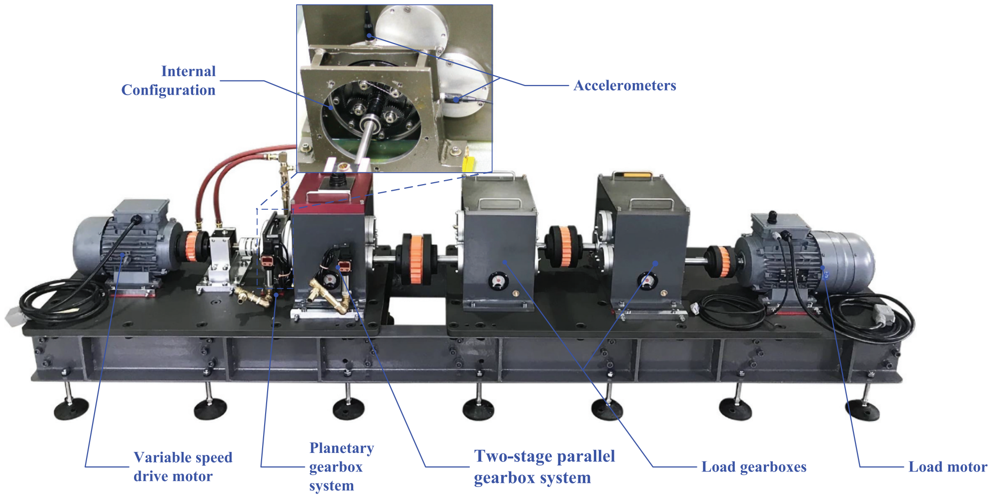

Figure 11. Illustration of the drivetrain prognostics simulation (DPS).

provides a detailed temporal resolution for subsequent analysis and allows for the capture of high-frequency components that may be critical for fault detection and diagnosis.

Each data category in the experimental dataset encompasses 196,608 data points, collected over a period of 6.4 seconds. This substantial dataset size ensures statistical robustness and allows for the application of advanced signal processing and machine learning techniques. Moreover, the 6.4-second duration for each data category strikes a balance between capturing sufficient system dynamics and maintaining computational feasibility in subsequent analyses. This comprehensive experimental setup and data collection protocol are designed to rigorously test the proposed method under conditions that closely simulate real-world industrial drivetrain operations. By doing so, the study aims to enhance the practical applicability and validity of the research findings, ultimately contributing to the advancement of prognostics and health management in industrial systems.

4.2.2. Comparative networks and results analysis

To further explore the effectiveness of the proposed SSTL method, as previously discussed in case study I, this section compares it with various established techniques, including DenseNet (Impulse noise), DenseNet (Denoise), DenseNet (Amplitude-limited), SimCLR, Fast-MoCo, and ITSSL. The diagnostic results obtained from the DPS datasets are detailed in Table 2 and visually represented in Figure 12. Figure 12 illustrates the experimental results of various methods applied to the DPS dataset, using 10% labelled data. In this study, the proposed SSTL approach is compared with established supervised learning models, including DenseNet (Impulse noise), DenseNet (Denoise), and DenseNet (Amplitude-limited). The results indicate that SSTL significantly enhances performance relative to these models. Specifically, with a labelling rate of 10%, SSTL achieves an accuracy of 91.2%, substantially outperforming the supervised learning models DenseNet (Impulse noise), DenseNet (Denoise), and DenseNet (Amplitude-limited), which

Table 2. Comparative analysis of experimental results for case study II.

Accuracy(%) Label rate / Model	10%	15%	20%
DenseNet (Impulse noise)	84.5±0.21	85±0.17	88±0.36
DenseNet (Denoise)	83.3±0.19	85.5±0.19	86.8±0.18
DenseNet (Amplitude-limited)	21.4±0.2	48.7±0.16	52±0.22
SimCLR	**87.6±0.61**	88.2±0.46	88.7±0.89
Fast-MoCo	82.1±3.94	87.4±3.49	88±3.39
ITSSL	84.9±0.2	**89.7±0.4**	**90.7±0.4**
SSTL (Ours)	**91.2±0.53**	**93±0.61**	**93.1±0.65**

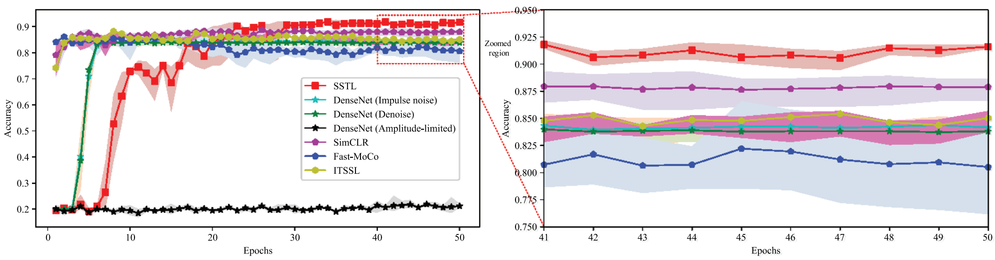

Figure 12. Experimental results of different methods on UESTC dataset with 10% labelled data.

recorded accuracies of 84.5%, 83.3%, and 21.4%, respectively. Furthermore, when the labelling rate is increased to 20%, SSTL demonstrates even more impressive results, attaining a diagnostic accuracy of 93.1%. In contrast, the accuracy of the supervised methods remains below 88%, despite the increase in labelled data. This observation highlights the robustness and scalability of the SSTL method across different labelling rates.

In addition, this study conducts a rigorous comparison between the proposed Self-Supervised Transfer Learning (SSTL) methods and several well-established semi-supervised learning approaches, including SimCLR, Fast-MoCo, and ITSSL. The results consistently demonstrate the superiority of the SSTL models. Notably, when operating under a constrained labelled data scenario with only 10% of the dataset labelled, the SSTL approach achieves a remarkable accuracy of 91.2%. This performance significantly surpasses that of its counterparts, with SimCLR attaining 87.6%, Fast-MoCo reaching 82.1%, and ITSSL achieving 84.9%. These findings not only highlight the efficacy of SSTL but also provide robust and compelling evidence that the incorporation of SSTL networks can markedly enhance the accuracy of fault diagnosis systems.

4.2.3. Comparative analysis of visualisations

In this case study, a confusion matrix is employed for the quantitative analysis of diagnosis results, as illustrated in Figure 13. The results demonstrate that the proposed Self-Supervised Transfer Learning (SSTL) method exhibits remarkable efficacy in identifying all fault types with high precision, even when trained on a dataset with a mere 10% labelled rate. This performance is especially remarkable when compared to traditional supervised models like DenseNet variants (which are fine-tuned for impulse noise, denoising, and amplitude-limited scenarios), as

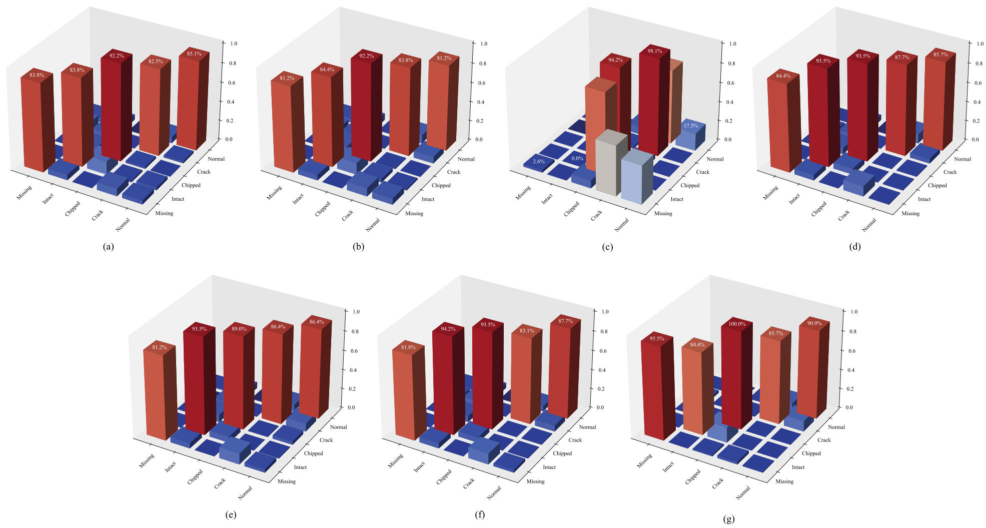

Figure 13. The classifying performance through confusion matrix. (a) DenseNet(ImpulseNet(Impulse noise). (b) DenseNet(Denoise).Net(Denoise). (c) DenseNetNet (amplitude-limited). (d) SimCLR.CLR. (e) Fast-MoCo. (f) ITSSL. (g) SSTL(Ours).

they are limited to detecting only one type of fault at a time. Moreover, although semi-supervised models like SimCLR, Fast-MoCo, and ITSSL are capable of generally detecting all types of gear failures, their accuracy in distinguishing specific fault categories is still not ideal. In contrast, the SSTL method showcases exceptional performance, not only accurately identifying all five types of faults but also achieving an impressive average accuracy of 91.2%. These compelling results highlight the strong effectiveness of the SSTL framework in fault diagnosis, especially in difficult situations where training data is scarce and transient interference is present.

The T-SNE results presented in Figure 14 provide further evidence of SSTL's notable advantages in differentiating between various fault signal types. Specifically, when constrained to a 10% labelled rate for training, traditional supervised learning approaches exhibit marked difficulties in distinguishing between different fault signals. Although semi-supervised models show marginal improvements in this regard, their overall performance remains limited. In contrast, SSTL demonstrates a robust and superior ability to differentiate between nearly all fault types, with only minimal confusion observed between the Crack and Normal fault categories. These visualisation results provide strong corroborative evidence supporting the conclusion that the SSTL framework not only effectively extracts salient features from vibration signals under conditions of transient interference but also successfully separates signal features of different fault types within the high-dimensional feature space.

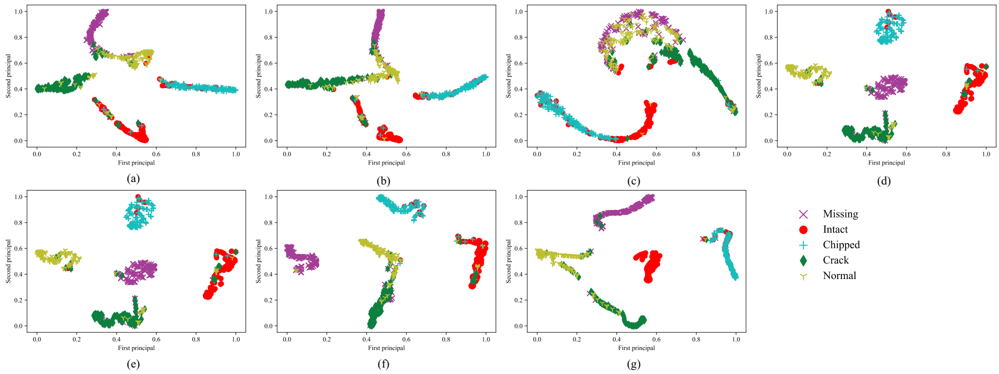

Figure 14. The feature visualisation through t-SNE. (a) DenseNet(ImpulseNet(Impulse noise). (b) DenseNet(Denoise).Net(Denoise). (c) DenseNetNet (amplitude-limited). (d) SimCLR.CLR. (e) Fast-MoCo. (f) ITSSL. (g) SSTL(Ours).

5. Conclusion

In conclusion, this research presents a novel SSTL approach that significantly enhances the accuracy and reliability of fault diagnosis in planetary gearboxes, particularly in challenging environments with limited labelled data and transient interference. By integrating semi-supervised learning with transfer learning and introducing innovative strategies such as label migration and matching, along with a limiting normalisation technique, SSTL addresses key limitations of traditional fault diagnosis methods. The results from two case studies confirm the effectiveness of the proposed approach, demonstrating its clear advantages over existing methods in terms of both fault detection accuracy and robustness.

Future work could build upon the findings of this study by exploring the application of the SSTL approach to other types of rotating machinery beyond planetary gearboxes. Additionally, further investigation into the scalability of SSTL in real-time applications with varying levels of transient interference and unlabelled data is warranted. Developing more advanced techniques for pseudo-labelling and domain adaptation could also improve SSTL's performance in even more complex and dynamic environments, potentially leading to broader industrial adoption of this approach.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by the National Natural Science Foundation of China (Grant 52105111), the Guangdong Basic and Applied Basic Research Foundation (Grant 2025A1515012256), the Shantou University (STU) Scientific Research Initiation Grant (NTF21029), the Industry-Academia Cooperation Project from the Guangdong Institute of Special Equipment Inspection and Research Shunde Branch (XTJ-KY01-202503-030), and the Enterprise Collaboration Project from the National Excellent Engineer Innovation Research

Institute for Advanced Manufacturing Industry in Foshan of Guangdong-Hong Kong-Macao Greater Bay Area (NSJH2025008).

References

[1] Shi J, Peng D, Peng Z, et al. Planetary gearbox fault diagnosis using bidirectional-convolutional LSTM networks. Mech Syst Signal Process. 2022;162:107996. doi: 10.1016/j.ymssp.2021.107996
[2] 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. doi: 10.1016/j.knosys.2024.112863
[3] Shao H, Ming Y, Liu Y, et al. Small sample gearbox fault diagnosis based on improved deep forest in noisy environments. Nondestr Test Evaluation. 2024;40(8):1–22. doi: 10.1080/10589759.2024.2404489
[4] 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;73:1–8. doi: 10.1109/TIM.2024.3383497
[5] Peng D, Yazdanianasr M, Mauricio A, et al. Physics-driven cross domain digital twin framework for bearing fault diagnosis in non-stationary conditions. Mech Syst Signal Process. 2025;228:112266. doi: 10.1016/j.ymssp.2024.112266
[6] Chen P, Wu Y, Xu C, et al. Interference suppression of nonstationary signals for bearing diagnosis under transient noise measurements. IEEE Transactions on Reliability. 2025. p. 1–15. doi:10.1109/TR.2025.3527739
[7] Chen P, Wang K, Zuo MJ, et al. An ameliorated synchroextracting transform based on upgraded local instantaneous frequency approximation. Measurement. 2019;148:106953. doi: 10.1016/j.measurement.2019.106953
[8] Chen P, Ma J, 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. doi: 10.1016/j.eswa.2025.128305
[9] Yin Y, Liu Z, Qin Y. A high-fidelity symbolization method for reciprocating pump vibration monitoring data. IEEE Sensors Journal. 2025;25:11613–11621. doi:10.1109/JSEN.2025.3541740
[10] 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. doi: 10.1016/j.oceaneng.2025.120798
[11] Xin G, Chen Y, Li L, et al. Complex symplectic geometry mode decomposition and a novel time-frequency fault feature extraction method. IEEE Transactions on Instrumentation and Measurement. 2025;74:1–10. doi:10.1109/TIM.2024.3522417
[12] 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. doi: 10.1088/1361-6501/adc031
[13] Chen P, Li Y, Wang K, et al. An automatic speed adaption neural network model for planetary gearbox fault diagnosis. Measurement. 2021;171:108784. doi: 10.1016/j.measurement.2020.108784
[14] 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;72:1–11. doi: 10.1109/TIM.2023.3267522
[15] Chen P, Li Y, Wang K, et al. A threshold self-setting condition monitoring scheme for wind turbine generator bearings based on deep convolutional generative adversarial networks. Measurement. 2021;167:108234. doi: 10.1016/j.measurement.2020.108234
[16] Zhou K, Diehl E, Tang J. Deep convolutional generative adversarial network with semi-supervised learning enabled physics elucidation for extended gear fault diagnosis

under data limitations. Mech Syst Signal Process. 2023;185:109772. doi: 10.1016/j.ymssp.2022.109772
[17] Chen P, Li Y, Wang K, et al. A novel knowledge transfer network with fluctuating operational condition adaptation for bearing fault pattern recognition. Measurement. 2020;158:107739. doi: 10.1016/j.measurement.2020.107739
[18] 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(13):23045–23054. doi: 10.1109/JIOT.2024.3382878
[19] Jamil F, Verstraeten T, Nowé A, et al. A deep boosted transfer learning method for wind turbine gearbox fault detection. Renewable Energy. 2022;197:331–341. doi: 10.1016/j.renene.2022.07.117
[20] Zhang L, Fan Q, Lin J, et al. A nearly end-to-end deep learning approach to fault diagnosis of wind turbine gearboxes under nonstationary conditions. Eng Appl Artif Intel. 2023;119:105735. doi: 10.1016/j.engappai.2022.105735
[21] Chen Y, Rao M, Feng K, et al. Physics-informed LSTM hyperparameters selection for gearbox fault detection. Mech Syst Signal Process. 2022;171:108907. doi: 10.1016/j.ymssp.2022.108907
[22] Zhao X, Yao J, Deng W, et al. Intelligent fault diagnosis of gearbox under variable working conditions with adaptive intraclass and interclass convolutional neural network. IEEE Trans Neural Netw Learn Syst. 2022;34(9):6339–6353. doi: 10.1109/TNNLS.2021.3135877
[23] Zhang X, Su Z, Hu X, et al. Semisupervised momentum prototype network for gearbox fault diagnosis under limited labeled samples. IEEE Trans Ind Inf. 2022;18(9):6203–6213. doi: 10.1109/TII.2022.3154486
[24] Zhao B, Cheng C, Zhao S, et al. Hybrid semi-supervised learning for rotating machinery fault diagnosis based on grouped pseudo labeling and consistency regularization. IEEE Trans Instrum Meas. 2023;72:1–12. doi: 10.1109/TIM.2023.3269112
[25] Xiao R, Zhang Z, Dan Y, et al. Multifeature extraction and semi-supervised deep learning scheme for state diagnosis of converter transformer. IEEE Trans Instrum Meas. 2022;71:1–12. doi: 10.1109/TIM.2022.3168929
[26] Xiang L, Zhang X, Zhang Y, et al. A novel method for rotor fault diagnosis based on deep transfer learning with simulated samples. Measurement. 2023;207:112350. doi: 10.1016/j.measurement.2022.112350
[27] Sun Q, Zhang Y, Chu L, et al. Fault diagnosis of gearbox based on cross-domain transfer learning with fine-tuning mechanism using unbalanced samples. IEEE Trans Instrum Meas. 2024;73:1–10. doi: 10.1109/TIM.2024.3415783
[28] Wang B, Wei Y, Liu S, et al. Unsupervised joint subdomain adaptation network for fault diagnosis. IEEE Sensors J. 2022;22(9):8891–8903. doi: 10.1109/JSEN.2022.3163425
[29] Zhu Y, Liang X, Wang T, et al. Multi-information fusion fault diagnosis of bogie bearing under small samples via unsupervised representation alignment deep Q-learning. IEEE Trans Instrum Meas. 2022;72:1–15. doi: 10.1109/TIM.2022.3225008
[30] Zhang T, Chen J, He S, et al. Prior knowledge-augmented self-supervised feature learning for few-shot intelligent fault diagnosis of machines. IEEE Trans Ind Electron. 2022;69(10):10573–10584. doi: 10.1109/TIE.2022.3140403
[31] Li J, Huang R, Chen J, et al. Deep self-supervised domain adaptation network for fault diagnosis of rotating machine with unlabeled data. IEEE Trans Instrum Meas. 2022;71:1–9. doi: 10.1109/TIM.2022.3218574
[32] Lee D, Kim S, Kim I, et al. Contrastive regularization for semi-supervised learning. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA; 2022 (IEEE). p. 3910–3919 doi:10.1109/CVPRW56347.2022.00436.
[33] Jiao J, Li H, Lin J. Self-training reinforced adversarial adaptation for machine fault diagnosis. IEEE Transactions on Industrial Electronics. 2023;70:11649–11658. doi:10.1109/TIE.2022.3229344

[34] Pu X, Li C. Meta-self-training based on teacher–student network for industrial label-noise fault diagnosis. IEEE Trans Instrum Meas. 2022;72:1–11. doi: 10.1109/TIM.2022.3205681
[35] Chen X, Yang R, Xue Y *et al*. Deep transfer learning for bearing fault diagnosis: a systematic review since 2016. IEEE Transactions on Instrumentation and Measurement. 2023;72:1–21. doi:10.1109/TIM.2023.3244237
[36] Zhang W, Zhang T, Cui G, et al. Intelligent machine fault diagnosis using convolutional neural networks and transfer learning. IEEE Access. 2022;10:50959–50973. doi: 10.1109/ACCESS.2022.3173444
[37] He J, Ouyang M, Chen Z, et al. A deep transfer learning fault diagnosis method based on WGAN and minimum singular value for non-homologous bearing. IEEE Trans Instrum Meas. 2022;71:1–9. doi: 10.1109/TIM.2022.3160533
[38] Wang Z, Cui J, Cai W, et al. Partial transfer learning of multidiscriminator deep weighted adversarial network in cross-machine fault diagnosis. IEEE Trans Instrum Meas. 2022;71:1–10. doi: 10.1109/TIM.2022.3216413
[39] Wang X, Liu F, Zhao D. Cross-machine fault diagnosis with semi-supervised discriminative adversarial domain adaptation. Sensors. 2020;20(13):3753. doi: 10.3390/s20133753
[40] Han T, Liu C, Wu R, et al. Deep transfer learning with limited data for machinery fault diagnosis. Appl Soft Comput. 2021;103:107150. doi: 10.1016/j.asoc.2021.107150
[41] Li X, Zhang W, Ding Q, et al. Diagnosing rotating machines with weakly supervised data using deep transfer learning. IEEE Trans Ind Inf. 2019;16(3):1688–1697. doi: 10.1109/TII.2019.2927590
[42] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. PMLR; 2020 119. p. 1597–1607.
[43] Ci Y, Lin C, Bai L, et al. Fast-MoCo: boost momentum-based contrastive learning with combinatorial patches. In: European Conference on Computer Vision, Tel Aviv, Israel. Springer; 2022. p. 290–306.
[44] Hu C, Wu J, Sun C, et al. Interinstance and intratemporal self-supervised learning with few labeled data for fault diagnosis. IEEE Trans Ind Inf. 2022;19(5):6502–6512. doi: 10.1109/TII.2022.3183601
[45] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA; 2017 (IEEE). p. 2261–2269 doi:10.1109/CVPR.2017.243.