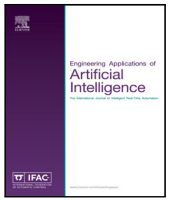
Engineering Applications of Artificial Intelligence 130 (2024) 107980

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Evaluating practical adversarial robustness of fault diagnosis systems via spectrogram-aware ensemble method

Hoki Kim [a], Sangho Lee [b,c], Jaewook Lee [d], Woojin Lee [e,*], Youngdoo Son [b,c,*]

[a] Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul, Republic of Korea
[b] Department of Industrial and Systems Engineering, Dongguk University-Seoul, Pildong-ro 1-gil, Seoul, Republic of Korea
[c] Data Science Laboratory (DSLAB), Dongguk University-Seoul, Pildong-ro 1-gil, Seoul, Republic of Korea
[d] Industrial Engineering, Seoul National University, Gwanakro 1, Seoul, Republic of Korea
[e] School of AI Convergence, Dongguk University-Seoul, Pildong-ro 1-gil, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Bearing fault diagnosis system
Adversarial attack
Spectrogram

ABSTRACT

While machine learning models have shown superior performance in fault diagnosis systems, researchers have revealed their vulnerability to subtle noises generated by adversarial attacks. Given that this vulnerability can lead to misdiagnosis or unnecessary maintenance, the assessment of the practical robustness of fault diagnosis models is crucial for their deployment and use in real-world scenarios. However, research on the practical adversarial robustness of fault diagnosis models remains limited. In this work, we present a comprehensive analysis on rotating machinery diagnostics and discover that existing attacks often over-estimate the robustness of these models in practical settings. In order to precisely estimate the practical robustness of models, we propose a novel method that unveils the hidden risks of fault diagnosis models by manipulating the spectrum of signal frequencies—an area that has been rarely explored in the domain of adversarial attacks. Our proposed attack, Spectrogram-Aware Ensemble Method (SAEM), the hidden vulnerability of fault diagnosis systems through achieving a higher attack performance in practical black-box settings. Through experiments, we reveal the potential dangers of employing non-robust fault diagnosis models in real-world applications and suggest directions for future research in industrial applications.

1. Introduction

The growing complexity of modern industries has yielded numerous advantages, such as enhanced production efficiency and the ability to accommodate diverse customer demands. However, alongside these benefits, the complexity has also introduced challenges in fault diagnosis, requiring careful observation to identify malfunctions and breakdowns. As a result, the monitoring system has become increasingly important in maintaining the stability of industrial operations. Tian et al. (2015). For instance, the ability to diagnose potential faults in basic components widely used across all industries, such as rolling elements, can help us prevent over 50% of system faults (Lanham, 2002). Therefore, fault diagnosis research is significantly important to ensure the efficiency and reliability of industrial systems.

With the advancement of big data and analytic tools, data-driven methods have become common techniques for industrial fault diagnosis. These data-driven fault diagnosis methods are typically trained with collected data and used to prevent future failures during system operations. Building on the success of diverse machine learning techniques in fault diagnosis tasks, including k-nearest neighbor (Tian et al., 2015) and support vector machine (Gryllias and Antoniadis, 2012), recent work has further improved the performance of bearing fault diagnosis systems by adopting deep learning methods (Zhang et al., 2020; Li et al., 2022). Specifically, Zhang et al. (2020) demonstrated that multi-layer perceptron (MLP) can yield high performance on the fault diagnosis of bearings and Zhang et al. (2017) also achieved improved diagnosis performance by adopting wide deep-convolutional neural network (WDCNN). Moreover, to consider the temporal dependency, Li et al. (2022) utilized long short-term memory (LSTM) (Karim et al., 2017) for detecting the faults in rotating machinery. Recently, Chen et al. (2020) and Zheng et al. (2021) proposed the utilization of one-dimensional convolutional neural network (CNN) and temporal convolutional network (TCN) (Bai et al., 2018), achieving the state-of-the-art performance in detecting bearing faults, respectively.

However, previous studies have revealed that these machine learning methods are vulnerable to subtle noises and are further easily

* Corresponding authors.
E-mail addresses: hokikim@cau.ac.kr (H. Kim), sangho218@dgu.ac.kr (S. Lee), jaewook@snu.ac.kr (J. Lee), wj926@dgu.ac.kr (W. Lee), youngdoo@dongguk.edu (Y. Son).

https://doi.org/10.1016/j.engappai.2024.107980
Received 31 August 2023; Received in revised form 3 December 2023; Accepted 23 January 2024
Available online 1 March 2024

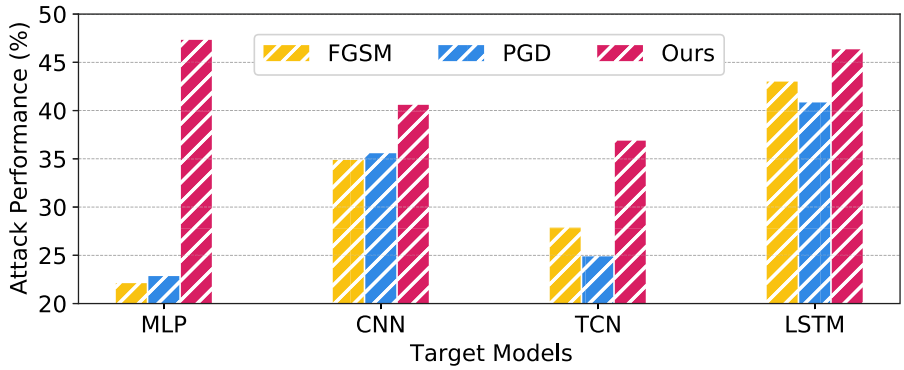


Fig. 1. Comparison of attack performance (%) between different attack methods. Unless specified otherwise, the attack performance (%) implies the degraded accuracy (%) of each target model. All attacks generate adversarial examples using WDCNN (Zhang et al., 2017) as the source model and aim to deceive the target models. Our proposed method achieves the best attack performance against all target models, highlighting that existing methods (FGSM and PGD) **overestimate** the robustness of fault diagnosis systems.

deceived by these maliciously perturbed examples (Szegedy et al., 2013). We call these examples *adversarial examples*. Most importantly, these adversarial examples can be easily generated using a simple method known as an adversarial attack (Goodfellow et al., 2014) that often maximizes a loss value by perturbing the original examples. This adversarial vulnerability has been identified in various industrial applications, including time series classification (Yang et al., 2022), speaker verification (Zhang et al., 2022b; Kim et al., 2023), malware detection (Shaukat et al., 2022), and fault diagnosis systems (Ge et al., 2021; Zhuo et al., 2022). Since ensuring safety and reliability is the key factor of fault diagnosis systems, the potential threat of adversarial attacks not only degrades the reliability of machine learning methods but also leads to substantial economic losses or human casualties by misdiagnosis of system operations. Therefore, identifying the robustness of fault diagnosis models against adversarial attacks is crucial to utilize the models in their applications.

Although some prior work has investigated the robustness of bearing fault diagnosis systems (Ge et al., 2021; Zhuo et al., 2022), these investigations have primarily focused on white-box settings, which assume that the target models are publicly available. However, the white-box settings are not typical in bearing fault diagnosis systems because the target models often are highly secured and not publicly accessible in real-world scenarios. Furthermore, compared to other domains (Zhang et al., 2022b; Kim et al., 2023), we discover that only limited attacks are considered in fault diagnosis systems (Ge et al., 2021; Zhuo et al., 2022), which might lead to unexpected dangers in real-world systems.

Therefore, in this paper, we aim to delve into the practical robustness of bearing fault diagnosis systems in real-world scenarios, which remains largely unexplored. To this end, we first construct the black-box settings (Zhou et al., 2018), where the attackers cannot access the information of the target model, and then investigate the existing attacks with diverse models. In Fig. 1, we summarize the attack performance, which is the degraded accuracy (%) of each target model unless specified otherwise, for each attack. Our experiments reveal that the existing attacks (Ge et al., 2021; Zhuo et al., 2022) are generally ineffective in estimating the practical robustness of bearing fault diagnosis systems. In other words, the existing methods tend to overestimate the adversarial robustness, resulting in the overlooking of unknown risks and hazards.

To address this problem, we propose a novel attack method called Spectrogram-Aware Ensemble Method (SAEM). Based on the domain knowledge and recent finding in bearing fault diagnosis systems (Akhenia et al., 2022; Hendriks et al., 2022), SAEM leverages the spectral information of bearing signals during attack process and achieves a higher attack performance compared to existing attacks and reveal the potential danger associated with deploying bearing fault diagnosis systems in practical settings. Indeed, as illustrated in Fig. 1, our proposed attack demonstrates superior attack performance over other existing attacks.

The remainder of this paper is organized as follows: Section 2 provides background information on bearing fault diagnosis systems and an overview of adversarial attacks. In Section 3, based on the analysis of the spectral information of bearing datasets, we propose a novel adversarial attack method for assessing the practical robustness. In Section 4, we demonstrate the effectiveness of our proposed method through comprehensive experiments conducted on widely used datasets, the Case Western Reserve University (CWRU) bearing dataset (Smith and Randall, 2015), as well as our additional data newly gathered in practical settings that simulate real-world bearing fault diagnosis systems. Finally, Section 5 summarizes our findings and presents concluding remarks.

2. Related work

2.1. Bearing fault diagnosis system

As highlighted in Lanham (2002), bearings serve as fundamental components within industrial engineering systems. Thus, under the system of modern industries, diagnosing the bearing fault is significantly important to prevent faults in system operations and further improve its efficiency. Bearings often consist of different components, such as the inner race, outer race, balls, and cage. Each component can include multiple fault conditions. To gain a deeper understanding of bearing fault diagnosis, various datasets have been developed. Among them, the Case Western Reserve University (CWRU) dataset is widely used in the field of industrial fault diagnosis (Smith and Randall, 2015). The CWRU dataset basically includes one normal state and three fault conditions, including inner race faults, outer race faults, and ball faults. Given a vibration signal $\boldsymbol{x}$ collected from rotating machines, the corresponding machinery state $\boldsymbol{y}$ becomes a target label. In terms of bearing fault diagnosis, the objective is formulated as follows:

$$\underset{f}{\operatorname{argmin}}\, \mathcal{L}(f(\boldsymbol{x}), \boldsymbol{y}), \tag{1}$$

where f is a diagnosis model that determines whether the vibration signal $\boldsymbol{x}$ corresponds to the state $\boldsymbol{y}$ and $\mathcal{L}$ is a loss function for the task. The following cross entropy loss $\mathcal{L}_{CE}$ is often used for this optimization (Zhang et al., 2017; Li et al., 2022),

$$\mathcal{L}_{CE}(f(\boldsymbol{x}), \boldsymbol{y}) = -\sum_{i=1}^{C} y_i \log(f(\boldsymbol{x})_i), \text{ for } C \text{ classes} \tag{2}$$

where y_i is the ith component in $\boldsymbol{y}$.

While the vibration signal $\boldsymbol{x}$ is inherently in the form of a waveform, previous research has discovered the usefulness of frequency transformations in determining its fault state. Since fault-bearing signals exhibit

specific periodic impulse responses in the frequency domain, several approaches have extracted frequency information from the original waveform of $\boldsymbol{x}$ and utilized it as an input for the function f in (1). For instance, the fast Fourier transform (FFT) can be adopted as a pre-processing technique for extracting frequency information. The FFT for a finite length signal $\boldsymbol{x} = [x_1, \dots, x_n]$ is formulated as follows:

$$X_k = \sum_{i=1}^{n} x_n W_n^{ki}, \tag{3}$$

where $k = 1, \dots, n$ and $W_n = \exp(-2\pi j/n)$. Previous works, such as Jia et al. (2016) and Hou et al. (2023), have achieved performance improvement in fault diagnosis systems using the FFT.

More recently, some studies (Pandhare et al., 2019; Akhenia et al., 2022) have discovered that spectral informational can also help the model to classify the bearing faults by capturing their unique fault characteristics. While the FFT provides information about the frequency components of a signal, the spectrogram offers additional advantages by considering a time-dependent representation of the original waveform. The spectrogram initially separates the input signal $\boldsymbol{x}$ into shorter segments, with each frame typically overlapping its neighboring frames. Subsequently, these segments are transformed into the frequency domain using windowing techniques and the FFT. This process is called the short-time Fourier transform (STFT). Finally, the power spectrum calculation is applied to obtain the relative strength and energy contribution of each frequency component. Simply, the spectrogram can be formulated as follows:

$$\texttt{Spec}(\boldsymbol{x}) = \left| \sum_{i=0}^{w-1} \mathcal{W}(i) \boldsymbol{x}_{ni+r} W_n^{ki} \right|, \tag{4}$$

where $\mathcal{W}$ is a window function, such as Hanning window, r is the hop size, and w is the window size. The resulting spectrogram is a two-dimensional matrix for time index n and frequency bin k, leading to a deeper analysis on the original waveform $\boldsymbol{x}$ and further enabling the adoption of convolutional neural networks commonly used in computer vision (Pandhare et al., 2019). Building on these benefits, the spectrogram successfully boosts fault diagnosis systems to achieve high performance in various tasks (Akhenia et al., 2022; Hendriks et al., 2022).

2.2. Adversarial attack

2.2.1. White-box setting

The adversarial attack is a type of method that generates malicious noise (or perturbation) to deceive machine learning models (Szegedy et al., 2013; Goodfellow et al., 2014). Specifically, given a model f, adversarial attacks optimize the following maximization,

$$\underset{\|\boldsymbol{\delta}\| \le \epsilon}{\text{argmax}} \mathcal{L}(f(\boldsymbol{x} + \boldsymbol{\delta}), \boldsymbol{y}), \tag{5}$$

where it is opposite to the general training objective in (1) with the maximum noise size, ϵ, that controls the strength of noises. The optimization output $\boldsymbol{\delta}$ is the adversarial noise that fools the given model f so that f outputs totally different label $\boldsymbol{y}' \neq \boldsymbol{y}$ for the perturbed example $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\delta}$; $\boldsymbol{x}'$ is called an adversarial example. In order to solve the optimization, Goodfellow et al. (2014) proposed a fast gradient sign method (FGSM). By integrating the gradient descent method, FGSM outputs the following noise,

$$\epsilon \cdot \texttt{sign}(\nabla_{\boldsymbol{\delta}} \mathcal{L}(f(\boldsymbol{x} + \boldsymbol{\delta}), \boldsymbol{y})), \tag{6}$$

where ∇ is a vector differential operator and $\texttt{sign}$ is an element-wise sign operator. Furthermore, Madry et al. (2017) proposed project gradient method (PGD), an improved version of FGSM with multi-step optimization, that further degrades the performance of models, which is formalized as follows:

$$\boldsymbol{\delta}^{(t+1)} = \Pi_{\|\boldsymbol{\delta}\| \le \epsilon} [\boldsymbol{\delta}^{(t)} + \alpha \cdot \texttt{sign}(\nabla_{\boldsymbol{\delta}} \mathcal{L}(f(\boldsymbol{x} + \boldsymbol{\delta}^{(t)}), \boldsymbol{y}))], \tag{7}$$

where α is an optimization size for each step, the initial noise $\boldsymbol{\delta}^{(t)}$ is a random vector, and the final noise $\boldsymbol{\delta}$ corresponds to $\boldsymbol{\delta}^{(T)}$ for the number of steps T. Several studies have verified that even these simple algorithms can effectively degrade machine learning models in various domains (Papernot et al., 2017; Choi et al., 2023), including bearing fault diagnosis systems (Ge et al., 2021; Zhuo et al., 2022). The aforementioned attacks, FGSM and PGD, require the information of the model f to minimize its loss as shown in (6) and (7). We call this setting *white-box* since we assume that an attacker can access any information, including the trained models.

2.2.2. Black-box setting

In general, trained models are typically not made public. Especially in industrial fault diagnosis systems, models are only accessible by feeding input, such as querying through Internet of Things (IoT) devices. Thus, *black-box* settings, which assume the target models are unknown, are more practical in industrial fault diagnosis systems. In prior work, several black-box attack methods have been suggested, and among them, the transfer attack is regarded as the most efficient and powerful attack that leads misprediction of models. Under the black-box settings, the transfer attack simply generates adversarial examples with a source model g that can be publicly accessible or self-trained models by attackers. In other words, the adversarial examples $\boldsymbol{x}'$ are generated from the source model g and used to make the target model f incorrect prediction, as follows:

$$f(\boldsymbol{x}') \neq \boldsymbol{y} \text{ where } \boldsymbol{x}' = \boldsymbol{x} + \underset{\|\boldsymbol{\delta}\| \le \epsilon}{\text{argmax}} \mathcal{L}(g(\boldsymbol{x} + \boldsymbol{\delta}), \boldsymbol{y}). \tag{8}$$

By doing this, the attacker can attack the unknown target model f by applying the existing attack method, e.g., FGSM or PGD, on the source model g. Notably, this transfer attack yields a sufficient drop in target model performance in various domains (Papernot et al., 2017; Zhang et al., 2022a).

Since the direct estimation of robustness is not feasible (Madry et al., 2017; Cohen et al., 2019; Li et al., 2019), the robustness of a model is generally approximated as $1 - \mathcal{P}$, where $\mathcal{P}$ represents attack performance. However, in black-box settings, the naive adaptation of FGSM and PGD tends to overestimate the robustness of models in practical settings by developing domain-adaptive adversarial attacks (Xie et al., 2019; Dong et al., 2019). Specifically, Xie et al. (2019) demonstrated that an additional transformation during attack optimization can significantly degrade the performance of the target model f. The key idea was fooling vision models by applying a common visual transformation $\mathcal{T}$ during the optimization, i.e., maximizing $\mathcal{L}(g(\mathcal{T}(\boldsymbol{x}+\boldsymbol{\delta})), \boldsymbol{y})$ rather than $\mathcal{L}(g((\boldsymbol{x} + \boldsymbol{\delta})), \boldsymbol{y})$. The following work (Dong et al., 2019) also found that similar domain-specific knowledge can reveal more precise practical robustness of target models. Recently, Kim et al. (2023) also proposed a new attack that uses the continuous addition of noises during attack process based on sound domain knowledge and achieved high transfer attack performance.

However, there is no work on investigating the practical robustness of industrial fault diagnosis systems in black-box settings. Thus, in this paper, we propose a novel attack method that reveals the practical vulnerability of deployed models in practical settings by adopting spectral transformation during the optimization of adversarial attacks. To the best of our knowledge, it is the first attempt to propose the spectral-aware adversarial attack against industrial fault diagnosis and reveal the true danger of industrial fault diagnosis systems in practical operations.

3. Methodology

In this section, we introduce the overall structure of practical problem setting and the proposed method with its detailed algorithm.

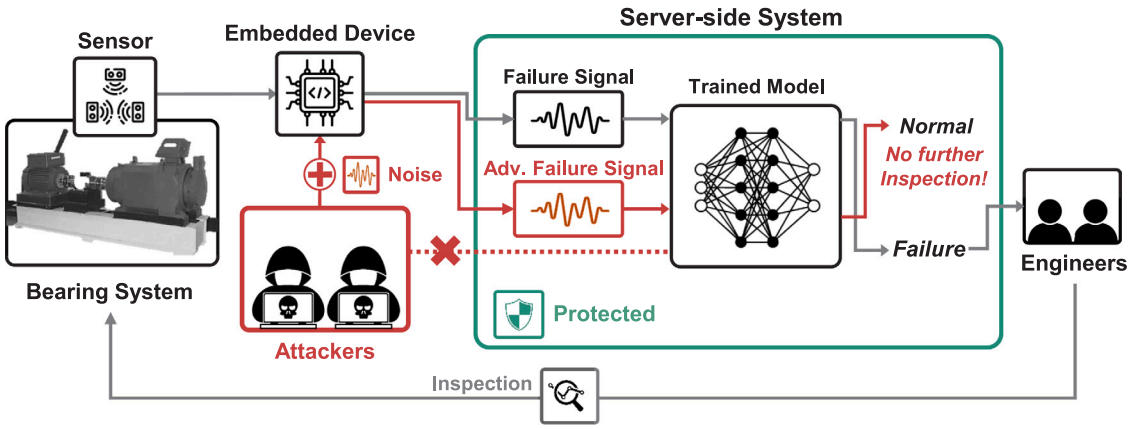


Fig. 2. Problem setting for simulating practical bearing fault diagnosis systems. Given bearing systems such as CWRU (Smith and Randall, 2015), the engineers construct an automatic fault diagnosis system with a sensor, an embedded device, and a trained model. In the typical scenario, server-side systems are well-protected, limiting attackers' direct access to the trained model and making white-box attacks infeasible. Consequently, attackers are forced to use black-box attacks, exploiting less secure devices to inject noise, to deceive the target model.

3.1. Problem setup

In Fig. 2, we present an illustration of the problem setting for simulating practical bearing fault diagnosis systems consisting of various components, including the targeted bearings and sensors. Given this bearing fault diagnosis system, the original signals from each bearing component are monitored by sensors attached to the system. These signals are subsequently transmitted to embedded devices, such as IoT devices. The embedded devices transfer the signals to the server-side system, which is typically well-protected with a high level of security due to its critical role in controlling industrial systems or safeguarding confidential materials. Within the server-side system, the trained model is also secured by the server-side security system as it analyzes the received signals and detects any indications of bearing fault, alerting engineers to perform necessary inspections and maintenance on the industrial system.

For attackers, the trained model in the server-side system becomes the target model. However, since the server-side system is well-protected, the attackers cannot directly access the target model. As a result, the attackers cannot utilize white-box attacks that have been mainly discussed in prior works (Ge et al., 2021; Zhuo et al., 2022). Instead, they are forced to employ black-box attacks that do not require the target model. In this scenario, the less secure embedded devices become the primary targets for injecting adversarial noises. For instance, by injecting such noises into the original fault signals, attackers generate adversarial fault signals (Adv. Fault Signal) and deceive the trained model to misclassify it as normal rather than fault. Consequently, engineers may receive incorrect reports, resulting in insufficient inspection and breakdown of the bearing fault diagnosis system. It is important to note that attackers can also deceive the target model by perturbing normal signals and further generating false fault alarms that lead to unnecessary shutdowns for the maintenance of the bearing fault diagnosis system.

3.2. Noise generation with spectral information

As illustrated in Fig. 2, since white-box attacks are impractical during industrial operations, we should evaluate the robustness of the trained model against black-box attacks rather than white-box attacks in practical settings. Although some work in industrial fault diagnosis focused on the robust evaluation of black-box attacks on diagnosis models (Ge et al., 2021; Zhuo and Ge, 2021), they are limited to adopting the existing methods, such as FGSM and PGD, to generate adversarial signals from the source model g. These naive adaptations of existing adversarial attacks are known to be weak (Dong et al., 2019; Xie et al., 2019) and indeed further overestimate the robustness of models as we observed in Fig. 1. Recent studies (Xie et al., 2019; Kim et al., 2023) have demonstrated that, rather than these naive adaptations, domain-adaptive adversarial attacks show significant attack performance and further reveal the hidden dangers of deployed models. Thus, we here propose a new domain-adaptive noise generation for assessing the adversarial robustness of fault diagnosis models in the practical scenario.

To harness an effective domain-adaptive attack, we focus on the spectral transformation of bearing signals. In prior works, the frequency information of bearing signals has been investigated since they have complex and high-dimensional information. For instance, the spectral transformation extracted by (4) can be effectively used to assist the indication of fault alarms (Pandhare et al., 2019; Akhenia et al., 2022; Hendriks et al., 2022). Indeed, the spectrograms extracted from bearing signals with different states exhibit distinct characteristics even for the human vision system. In Fig. 3, we plot the spectrograms of randomly sampled signals from the CWRU datasets for each label. Here, we use the window size 128 and Hanning window function. Specifically, the outer fault signal exhibits a darker area for high-frequency bins, where high amplitude frequencies are brighter colors. In contrast, the normal signal exhibits more bright colors. Similarly, ball and inner fault show different spectral information between 0 to 20 frequency bins compared to the normal ones. Thus, under the domain of bearing fault diagnosis systems, spectral information can provide effective domain-adaptive knowledge to estimate the precise robustness of diagnosis models.

Inspired by the observation above, we propose a novel regularization technique for incorporating spectral information during the generation of black-box adversarial examples. While the existing methods, such as FGSM and PGD, maximize the loss function $\mathcal{L}(g(\boldsymbol{x}+\boldsymbol{\delta}),\boldsymbol{y})$ with respect to $\boldsymbol{\delta}$ for a given sample $\boldsymbol{x}$ and its corresponding label $\boldsymbol{y}$, they do not guarantee the effectiveness of $\boldsymbol{\delta}$ against an unknown target model f, despite being successful in deceiving the source model g. To precisely estimate the robustness of the target model f by generating domain-adaptive perturbations $\boldsymbol{\delta}$, it is crucial that the resulting adversarial example $\boldsymbol{x}+\boldsymbol{\delta}$ has different spectral information than the benign example $\boldsymbol{x}$, as benign and fault examples have distinct spectral characteristics observed in Fig. 3. For instance, if the spectrogram of a fault signal $\boldsymbol{x}'$ is denoted as $\texttt{Spec}(\boldsymbol{x}')$, a more dangerous adversarial example $\boldsymbol{x}+\boldsymbol{\delta}$ is expected to exhibit a spectrogram similar to $\texttt{Spec}(\boldsymbol{x}')$ rather than $\texttt{Spec}(\boldsymbol{x})$ and, thereby, more likely to cause misclassification as a fault signal. Thus, to maximize the distortion of spectral information, we propose the utilization of the cosine embedding loss function, commonly employed in nonlinear embeddings and semi-supervised learning (Cheng et al., 2023; Patel et al., 2022), ensuring effective

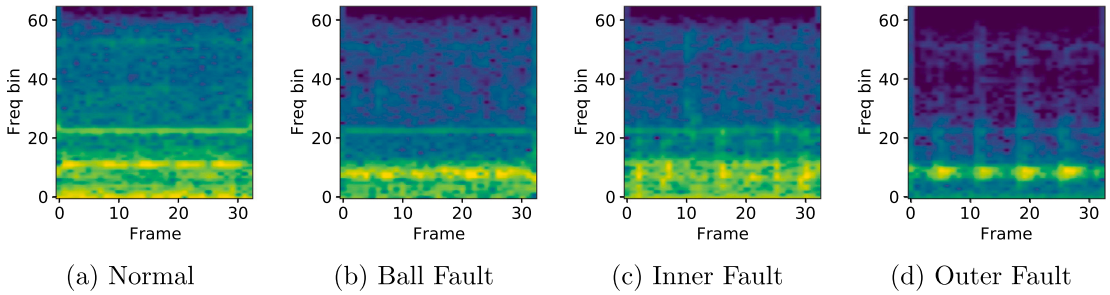


Fig. 3. Spectrograms of (a) normal signals and (b, c, d) fault signals. The time-dependent frequency components obtained from vibration signals corresponding to each fault state exhibit different patterns.

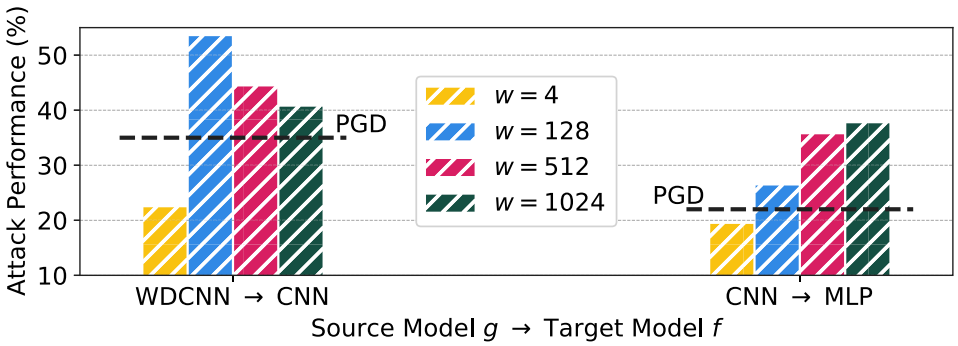


Fig. 4. Impact of the window size (w) on noise generation with spectral information. Noise generated with spectral information demonstrates superior performance compared to PGD (denoted by the black dotted line) when w is sufficiently large. However, when the source model is WDCNN, and the target model is CNN, the window size of 128 exhibits the highest attack performance. Conversely, when the source model is CNN, and the target model is MLP, the window size of 1024 shows the highest attack performance.

manipulation of the spectral manifold in Fig. 3. The objective function is formalized as:

$$\mathcal{L}_{SC}(\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{\delta}) := 1 - \cos(\mathrm{Spec}(\boldsymbol{x}+\boldsymbol{\delta}), \mathrm{Spec}(\boldsymbol{x})) + \max(0, \cos(\mathrm{Spec}(\boldsymbol{x}+\boldsymbol{\delta}), \mathrm{Spec}(\boldsymbol{x}^*)) - \gamma), \quad (9)$$

where γ represents the margin in the cosine embedding loss, and $\boldsymbol{x}^*$ is a target example that has a different label from the one of $\boldsymbol{x}$. For instance, given a normal signal $\boldsymbol{x}$, we perturb its spectrogram by selecting $\boldsymbol{x}^*$ from the fault signals. Maximizing (9) with respect to $\boldsymbol{\delta}$ yields a spectrogram that exhibits similar characteristics to the target signal $\boldsymbol{x}^*$ in the spectral manifold while maintaining distinctiveness from the original spectrogram. Combining this approach with the cross-entropy loss, the generation of adversarial examples becomes:

$$\boldsymbol{\delta}^{(t+1)} = \Pi_{\|\boldsymbol{\delta}\|\leq\epsilon}[\boldsymbol{\delta}^{(t)} + \alpha \cdot \mathrm{sign}(\nabla_{\boldsymbol{\delta}}(\mathcal{L}_{CE}(g(\boldsymbol{x}), \boldsymbol{y}) + \beta\mathcal{L}_{SC}(\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{\delta})))], \quad (10)$$

where α represents the step size, β is the spectral perturbation control parameter, and $\delta^{(T)}$ denotes the final noise after T steps. The target signal $\boldsymbol{x}^*$ can be randomly selected from the training dataset with the corresponding label $\boldsymbol{y}^* \neq \boldsymbol{y}$. By optimizing (10), we can expect a spectrogram-aware noise that effectively deceives the target model f.

3.3. Spectrogram-aware ensemble method

During generating noise with spectral information in the previous subsection, the spectrogram plays a critical role in optimization as it heavily influences the cosine embedding loss. Thus, in this subsection, we further investigate the effect of the window size w during the optimization of (10). Specifically, the window size w determines the frequency resolution and time-frequency trade-off in the spectrogram (Pandhare et al., 2019; Hendriks et al., 2022). A larger window size provides finer frequency resolution but sacrifices time resolution, while a smaller window size yields better time resolution but with coarser frequency information. For example, if the window size is too small, it may fail to capture high-frequency details. On the other hand, if the window size is too large, it may result in a loss of temporal information. Thus, the effect of w should be analyzed to generate more effective adversarial signals under black-box settings.

In Fig. 4, we estimate the attack performance of noise generation with spectral information with varying the window size w. Unless otherwise stated, in this paper, we quantify the attack performance by measuring the percentage of adversarial examples within the test set that successfully deceive the model following (Szegedy et al., 2013; Goodfellow et al., 2014). A higher attack performance indicates a stronger attack. For noise generation, we optimize (10) with a fixed value of $\beta = 100$ (a more detailed analysis of β is provided in Section 4.3). We observe that when $w = 4$, the attack performance is remarkably lower than that of PGD because $w = 4$ is too small to capture high-frequency details identifying transient events or rapid changes in the signal. By contrast, when w is sufficiently large to cover frequency information that distinguishes each bearing state, the performance of our approach is superior to that of PGD. While this result demonstrates the effectiveness of our proposed noise generation with spectral information, the best attack performance according to w depends on the source and target models. Specifically, when the source model is WDCNN and the target model is CNN, $w = 128$ yields the best attack performance, whereas $w = 1024$ yields the best attack performance for the other case. Consequently, we conclude that the window size w can yield different attack performances for different settings, and thus various window sizes should be considered simultaneously.

Therefore, to consider multiple w values during the noise generation simultaneously, we propose the ensemble of multiple window sizes within the spectrogram transformation. By leveraging an ensemble of multiple w values rather than using a single fixed window size, we anticipate that adversarial examples generated using ensembles of window sizes will have a similar effect to using multiple models, which has been proven highly effective in black-box settings (Huang et al., 2019; Kwon and Lee, 2022). For a given set of window sizes $[w_1, \ldots, w_m]$, we generate the noise $\boldsymbol{\delta}$ by maximizing the average of

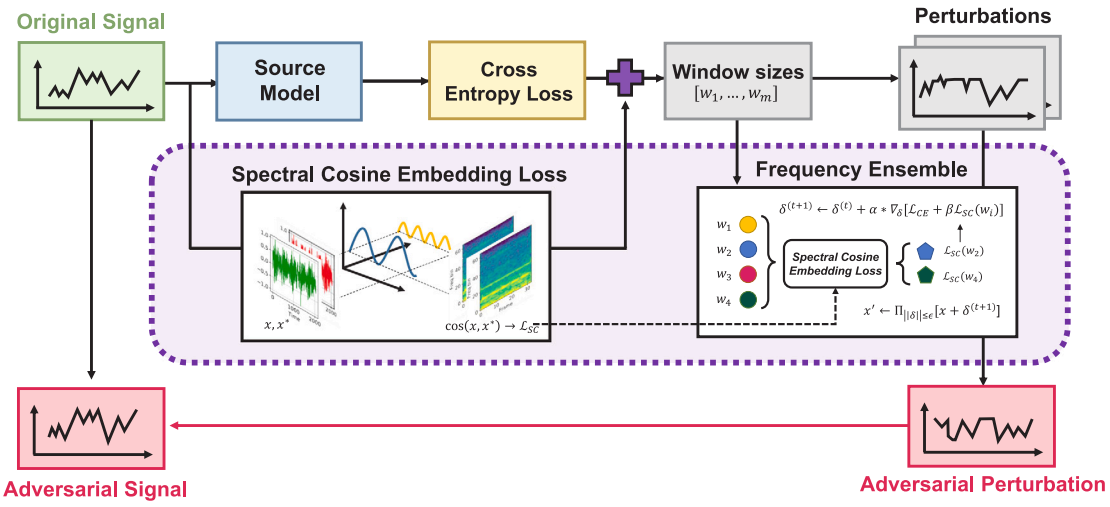


Fig. 5. Illustration of the process of the proposed method. Spectrogram-aware ensemble method (SAEM) uses spectral cosine embedding loss in (11) and further adopt frequency ensemble to generate adversarial signal.

Algorithm 1 Spectrogram-aware ensemble method (SAEM)

Input: a target model f; a signal $\boldsymbol{x}$; a true label $\boldsymbol{y}$; a target signal $\boldsymbol{x}^*$; maximum perturbation ϵ; number of steps T; step-size α; a list of window sizes $[w_1, \cdots, w_m]$; spectral control parameter β;
Output: an adversarial signal $\boldsymbol{x}'$;

$\boldsymbol{\delta}^{(1)} \leftarrow \mathbf{0}$
for t in [1, ..., $T-1$] **do**
  $\ell_{ce} \leftarrow \mathcal{L}_{CE}(g(\boldsymbol{x} + \boldsymbol{\delta}^{(t)}), \boldsymbol{y})$ ▷ Eq. (2)
  for w_i in $[w_1, \cdots, w_m]$ **do**
    $\ell_{w_i} \leftarrow \mathcal{L}_{SC}(\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{\delta}^{(t)}; w_i)$ ▷ Eq. (9)
  end for
  $\boldsymbol{\delta}^{(t+1)} \leftarrow \boldsymbol{\delta}^{(t)} + \alpha * \nabla_\delta \left[\ell_{ce} + \beta \sum_{i=1}^{m} \ell_{w_i}\right]$
  $\boldsymbol{x}' \leftarrow \Pi_{\|\delta\| \le \epsilon}[\boldsymbol{x} + \boldsymbol{\delta}^{(t+1)}]$ ▷ Noise projection
  if $g(\boldsymbol{x}') = \boldsymbol{y}^*$ **then**
    break ▷ Attack Successfully
  end if
end for
Attack the target model f by feeding $\boldsymbol{x}'$.

$\mathcal{L}_{SC}(\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{\delta}; w_i)$ calculated for each w_i in (9) where now $\mathcal{L}_{SC}$ is defined as follows:

$$\mathcal{L}_{SC}(\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{\delta}; w) := 1 - \cos(\mathrm{Spec}(\boldsymbol{x} + \boldsymbol{\delta}; w), \mathrm{Spec}(\boldsymbol{x}; w)) + \max(0, \cos(\mathrm{Spec}(\boldsymbol{x} + \boldsymbol{\delta}; w), \mathrm{Spec}(\boldsymbol{x}^*; w)) - \gamma), \quad (11)$$

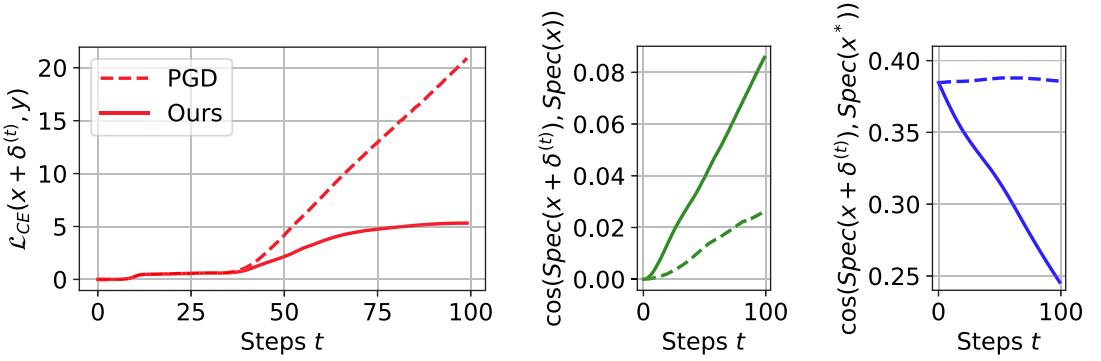
where $\mathrm{Spec}(\boldsymbol{x}; w)$ is the spectrogram of $\boldsymbol{x}$ with the window size w. Following prior works (Zhao et al., 2020; Kumar Dwivedi et al., 2019), the margin γ is set to 0 as a default. Finally, based on Sections 3.2 and 3.3, we denote this novel attack as the Spectrogram-aware ensemble method (SAEM), which is presented in Algorithm 1 and Fig. 5.

4. Experiments

In this section, we present the experimental setup and results to evaluate the effectiveness of our proposed method. In Section 4.1, we begin by outlining our experimental settings, including the models employed in this study and detailed training setups. In subsequent sections (4.2–4.3), we perform a series of analyses to investigate the optimization process and examine the results. Finally, in Section 4.4, we estimate the evaluation of our approach on a benchmark dataset and further provide its practical verification on a newly gathered dataset under a bearing fault diagnosis system. We compare the attack performance of our proposed method with the existing attack methods, demonstrating that our approach enables a more accurate estimation of the robustness of fault diagnosis models.

4.1. Experimental setup

Given the highly protected nature of the server-side system in Section 3.1, where direct access or manipulation of the trained fault diagnosis models by attackers is infeasible, it is essential to consider various possible combinations of source and target models denoted as (g, f) to verify the expected practical robustness of these models. Thus, we incorporate a range of deep learning-based fault diagnosis models. To give a more detailed description of the experiment, we here provide more detail on the models that have been briefly described in the introduction. One of the often employed models is the Multi-Layer Perceptron (MLP), a feedforward neural network consisting of multiple layers of interconnected nodes. Due to its simplicity, MLP has been widely adopted in fault diagnosis systems (Ismail Fawaz et al., 2019). Additionally, a variant of the one-dimensional Convolutional Neural Network (CNN) architecture has been used for fault diagnosis (Chen et al., 2020). The Temporal Convolutional Network (TCN) (Bai et al., 2018) has also been employed by Zheng et al. (2021) for bearing fault diagnosis systems. The TCN architecture incorporates dilated convolutions to capture long-range dependencies in the temporal dimension while keeping the number of model parameters manageable. It typically consists of convolutional layers with dilated convolutions, activation functions such as ReLU, and pooling layers. The Long Short-Term Memory (LSTM) model (Karim et al., 2017) is another architecture for bearing fault diagnosis systems. LSTM is a specific kind of recurrent neural network designed to capture long-term relationships in sequential data, resulting in enhanced fault diagnosis capabilities for rotating machinery, as demonstrated in Li et al. (2022). Finally, the Wide First-Layer Kernels Deep Convolutional Neural Network (WDCNN) (Zhang et al., 2017) is also utilized for this study. WDCNN is specifically designed to handle noisy signals and accurately classify machinery states. The architecture incorporates wide first-layer convolutional kernels, followed by subsequent layers that extract higher-level features and is improved by Gao et al. (2021) for classification in machinery systems. In summary, we constructed and trained five different architectures for bearing fault diagnosis: MLP (Ismail Fawaz et al., 2019), CNN (Chen et al., 2020), TCN (Bai et al., 2018), LSTM (Karim et al., 2017), and WDCNN (Zhang et al., 2017). Following (Ismail Fawaz et al., 2019; Zheng et al., 2021; Li et al., 2022; Gao et al., 2021), we train these models with the original signal to ease comparison with prior works (Ge



(a) Cross Entropy Loss $\mathcal{L}_{CE}(\boldsymbol{x}+\boldsymbol{\delta},\boldsymbol{y})$ (b) Distance to $\texttt{Spec}(\boldsymbol{x})$ (c) Distance to $\texttt{Spec}(\boldsymbol{x}^*)$

Fig. 6. Loss optimization during the attack process. The dotted and solid lines represent PGD and the proposed method, respectively. Both attacks successfully increase the cross-entropy loss ((a)); however, from a spectral perspective, PGD fails to perturb the spectral information, as depicted in (b) and (c). In contrast, our proposed method successfully generates noise $\boldsymbol{\delta}$ that not only maximizes the distance to $\texttt{Spec}(\boldsymbol{x})$ but also minimizes the distance to $\texttt{Spec}(\boldsymbol{x}^*)$.

et al., 2021; Zhuo et al., 2022). Additionally, following (Papernot et al., 2017; Xie et al., 2019; Dong et al., 2019), we assume that the adversary does not have direct access to the target model but can access a public dataset as illustrated in Fig. 2.

In Sections 4.2 to 4.3, we consider the CWRU bearing dataset to ease the comparison between existing methods (Ge et al., 2021; Zhuo et al., 2022). In Section 4.4, we conduct a thorough experiment with the CWRU dataset and an additional dataset, which is a newly gathered with our practical machinery setups, to verify the practical robustness of these models. The CWRU dataset has one normal state and three fault conditions, including inner, outer, and ball faults, as the benchmark dataset. The signals belonging to each bearing state are collected from diverse working conditions with different motor loads. We used the signals collected from the drive-end accelerometer at a sampling rate of 48 kHz and integrated all working conditions and fault diameters. The total number of bearing signals is 8168, and each signal $\boldsymbol{x}$ has its state information $\boldsymbol{y}$. Following (Ge et al., 2021), the fault diagnosis models were trained using signals of length 2048, which were min–max scaled with a maximum value of 1 and a minimum value of -1, and the maximum perturbation size was fixed $\epsilon = 0.1$. Given these trained models, in the subsequent experiments, we investigated all possible combinations of source and target models.

4.2. Optimization analysis

To begin with, we investigate the optimization process involved in perturbation generation for the proposed method. This analysis allows us to gain insights into how the proposed method effectively generates adversarial examples for fault diagnosis models. To ease the illustration of the optimization process, we here fix β to 100 and the window size w to 128 without an ensemble, using the WDCNN model as the source model. Then, we estimate two loss functions in Algorithm 1, $\mathcal{L}_{CE}(\boldsymbol{x}+\boldsymbol{\delta},\boldsymbol{y})$ and $\mathcal{L}_{SC}(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\delta};w)$, during optimization for our method and the representative comparison method, PGD, where $\boldsymbol{x}$ and $\boldsymbol{x}^*$ are selected from the normal and inner race fault states, respectively.

Fig. 6 summarizes the results of the optimization process. In Fig. 6(a), both PGD and our proposed optimization successfully increase the cross-entropy loss, $\mathcal{L}_{CE}(\boldsymbol{x}+\boldsymbol{\delta},\boldsymbol{y})$, leading to the misclassification of the source model. Notably, PGD tends to rapidly increase the cross-entropy loss as the steps increase. However, it is important to note that PGD has a tendency to overfit to the source model (Zhou et al., 2018; Kim et al., 2023). The overfitting phenomenon of PGD in the context of black-box settings refers to a scenario where the adversarial examples generated by PGD become excessively tailored to the target model, resulting in reduced attack performance when these adversarial examples are fed to other models. This is consistent with the result in Fig. 1 and we indeed observe a lower attack performance of PGD, which is further discussed in Section 4.3.

Moreover, as shown in Figs. 6(b) and 6(c), PGD fails to maximize the spectral information. In contrast, our method demonstrates a more rapid increment in the spectral distance between the adversarial examples and the original examples (Fig. 6(b)). In other words, by actively considering the spectral information, our approach produces more distorted examples within the spectral manifold. Furthermore, as shown in Fig. 6(c), we evaluate the spectral distance to the target example $\boldsymbol{x}^*$ and observe that PGD fails to minimize this distance, indicating its insufficient consideration of the manifold of spectrograms. Conversely, our method successfully minimizes the distance to the target example, emphasizing its ability to effectively navigate the spectral manifold.

In Fig. 7, we provide visual verification of the proposed optimization process by plotting the signals and spectrograms of a normal example, a PGD adversarial example, our adversarial example, and an inner race fault example. While the signals (Figs. 7(a)–7(c)) may be difficult to distinguish for the human visual system, the spectrograms exhibit significant differences. We can easily observe that PGD fails to consider the characteristic spectral patterns of bearing signals. Figs. 7(e) and 7(f) demonstrate that although PGD successfully fools the source model, it does not significantly perturb the spectral information. Both the normal and PGD adversarial examples exhibit common patterns near the frequency bin of 10. In contrast, our proposed optimization yields a spectrogram similar to the inner fault example, which is the target example, indicating its ability to effectively consider the spectral information. The spectrogram generated by our method exhibits similar patterns within the low-frequency bins, implying its capability to preserve unique spectral features of bearing signals. Based on these observations, we can conclude that the proposed optimization successfully finds an adversarial example that perturbs the unique spectral information of bearing signals. In the subsequent experiments, we further verify the effectiveness of our optimization in terms of attack performance with varying hyper-parameters.

4.3. Parameter analysis

In this subsection, we investigate the effectiveness of the proposed method in terms of attack performance and analyze the impact of various parameters on its performance. We begin by estimating the effect of the spectral control parameter β in (10). When $\beta = 0$, the proposed method becomes ignorant of spectral information, i.e., reducing to the PGD method. Therefore, investigating the effect of β helps assess the effectiveness of our domain-adaptive loss in evaluating model robustness. Fig. 8 summarizes the attack performance in terms of accuracy, i.e., the drop in accuracy caused by the attack method,

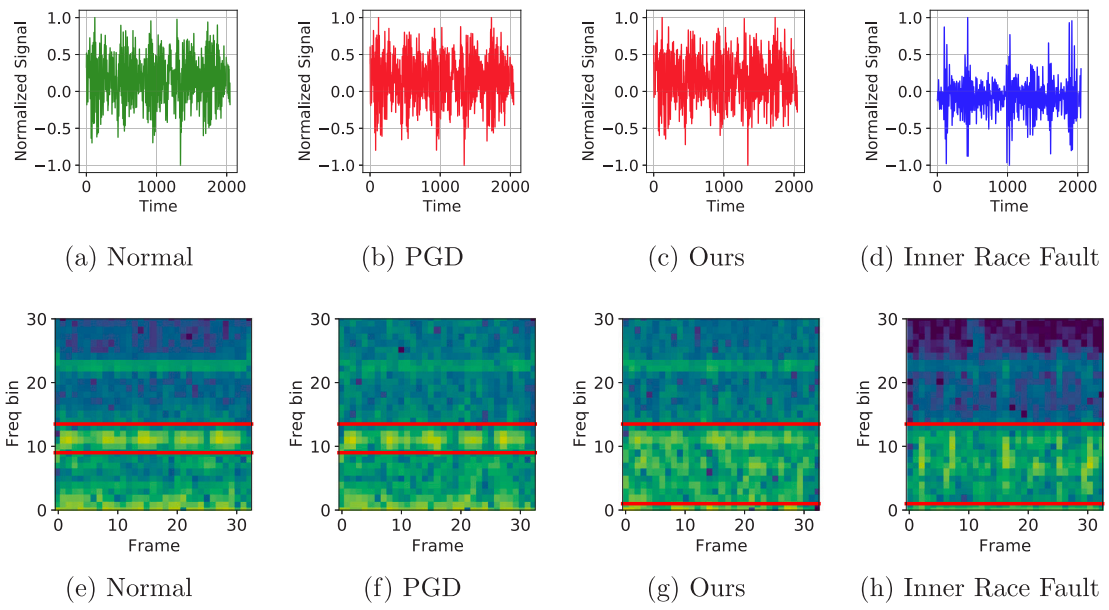


Fig. 7. Signals (a-d) and spectrograms (e-h) for each randomly sampled example: (a, e) normal example; (d, h) inner race fault example; (b, f) adversarial example generated by PGD on the sampled normal example; and (c, g) adversarial example generated by the proposed method on the sampled normal example. As shown in (e, f), the PGD example shares a similar pattern with the normal example near frequency bin 10 (highlighted by the red lines), which implies that PGD fails to perturb the spectrogram. In contrast, our proposed optimization successfully perturbs the spectrogram, where its spectrogram (g) does not exhibit common patterns observed in (e, f). Instead, the spectrogram of (g) exhibits greater similarity to that of (h) as highlighted by the red lines.

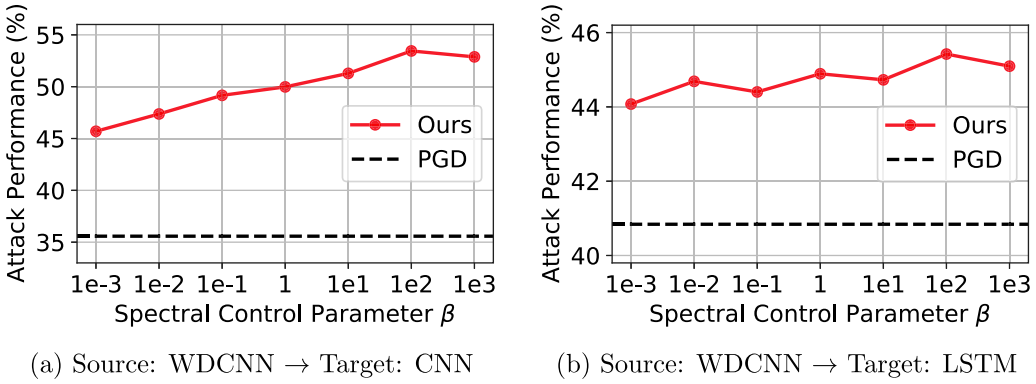
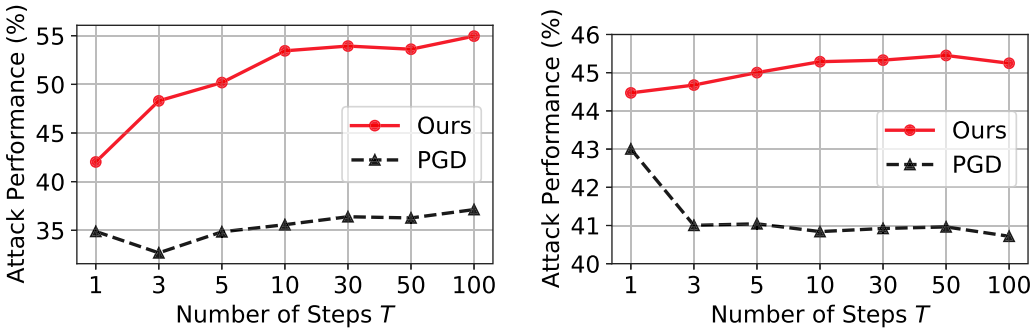


Fig. 8. Attack performance on accuracy (%) for each source and target model with varying the spectral control parameter β for our proposed method. The black line represents the attack performance of PGD. Increasing β effectively enhances the attack performance compared to PGD.

for varying values of β. Although we present the results for the source model WDCNN, similar tendencies were observed for all models, as further validated in Section 4.4. As shown in the figure, our proposed method consistently outperforms PGD, supporting the efficacy of considering domain-adaptive knowledge demonstrated in Section 4.2. The ability of our method to perturb the spectral information, as evidenced in Fig. 7, contributes to its superior attack performance. This finding is consistent with prior works (Dong et al., 2019; Kim et al., 2023) that emphasized the importance of domain knowledge in adversarial attacks.

Furthermore, our domain-adaptive loss helps mitigate the overfitting problem encountered by existing methods. As aforementioned in Fig. 6(a), PGD tends to overfit to the source model as the number of optimization steps increases (Zhou et al., 2018; Kim et al., 2023). Indeed, Fig. 9 illustrates that, in some cases, PGD exhibits lower attack performance as the number of steps T increases. This observation highlights the overfitting characteristic of PGD, which is detrimental to its performance. Notably, this is the first observation of such limitations in existing attack methods, specifically in bearing fault diagnosis systems, due to the overfitting phenomenon. In contrast, our proposed method consistently achieves higher attack performance for various values of T. Furthermore, the attack performance improves with an increase in the number of steps, indicating that our method is robust to the overfitting phenomenon (Zhou et al., 2018).

Next, we investigate the effect of the ensemble of the window sizes during the optimization process. Previously, in Fig. 4, we demonstrate that the choice of the window size, denoted as w, significantly affects the resulting spectrogram for a given input signal $\boldsymbol{x}$. To evaluate its effect on the attack performance of the proposed method, we vary the combination of w values within the spectrogram ensemble. Table 1 summarizes the effects of ensemble sizes for the FFT. To comprehensively assess the effectiveness of the ensemble, we measure the average attack performance across accuracy, recall, precision, and F1 score, which are commonly used measures in fault diagnosis systems with the source model WDCNN. Compared to considering a single w, the use of multiple w values consistently shows higher and more stable attack performance across all measures. Specifically, the combination of [128, 1024] yields the highest performance. This suggests that incorporating both small and large window sizes is beneficial in generating malicious adversarial examples. We note that this finding also aligns with prior works (Pham et al., 2020; Harris et al., 2016; Roy et al., 2020), which



(a) Source: WDCNN → Target: CNN (b) Source: WDCNN → Target: LSTM

Fig. 9. Attack performance on accuracy (%) for each source and target model with varying the number of steps T for PGD and our proposed method. The proposed method consistently outperforms PGD for all number of steps. Furthermore, the attack performance of our proposed method increases as the number of steps increases, whereas the attack performance of PGD decreases, highlighting the overestimation of robustness due to the overfitting phenomenon of PGD.

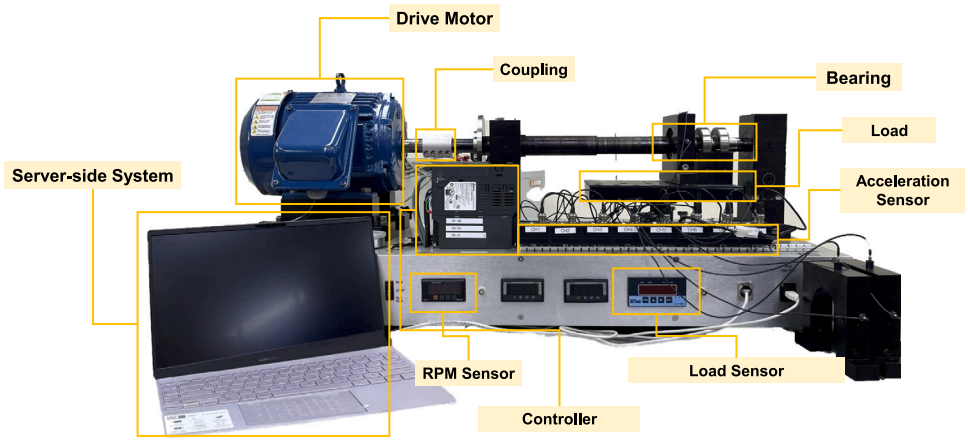


Fig. 10. Experimental devices of the self-constructed bearing fault diagnosis system.

Table 1
Analysis on spectrogram ensemble. Higher values are better for all measures. The highest and second-best results for each measure are highlighted in bold and underlined, respectively.

Spectrogram ensemble		Attack performance			
N	Combinations of the size w	Accuracy	Recall	Precision	F1 score
1	[4]	0.2360	0.1982	0.1648	0.2182
	[128]	0.4103	0.3592	0.3041	0.3737
	[512]	0.4181	0.3679	0.3136	0.3773
	[1024]	0.4137	0.3652	0.3095	0.3748
2	[4, 128]	0.4053	0.3551	0.2998	0.3687
	[4, 512]	0.4181	0.3683	0.3122	0.3776
	[4, 1024]	0.4161	0.3670	0.3145	0.3788
	[128, 512]	0.4181	0.3683	0.3096	0.3778
	[128, 1024]	**0.4273**	**0.3781**	**0.3278**	**0.3887**
	[512, 1024]	<u>0.4203</u>	<u>0.3701</u>	0.3166	<u>0.3817</u>
3	[4, 128, 512]	0.4135	0.3644	0.3116	0.3741
	[4, 128, 1024]	0.4141	0.3657	0.3129	0.3775
	[4, 512, 1024]	0.4178	0.3672	<u>0.3170</u>	0.3805
	[128, 512, 1024]	0.4132	0.3651	0.3137	0.3772
4	[4, 128, 512, 1024]	0.4139	0.3650	0.3136	0.3769

demonstrate the ability of small and large window sizes to extract distinct spectral information from the input signal, enhancing the diversity and effectiveness of the generated adversarial examples.

4.4. Practical evaluation

In this subsection, we provide practical evaluations of the proposed method over various datasets and performance measures to assess its effectiveness. To provide a realistic assessment of the adversarial robustness, we construct a real-world bearing fault diagnosis system under a practical bearing setting, which assumes that the collected signals from an arbitrary bearing are affected by some noises caused by other neighboring bearings, in addition to CWRU. Fig. 10 shows a bearing test rig with three bearings used for comprising our self-gathered bearing dataset, named “practical bearing faults dataset”. The dataset

Table 2
Attack performance on Accuracy and F1 score on the CWRU dataset.

Model		Attack performance on accuracy					Attack performance on F1 score				
Source	Target	FGSM	PGD	TIFGSM	NIA	SAEM	FGSM	PGD	TIFGSM	NIA	SAEM
WDCNN	MLP	0.23	0.24	0.24	0.22	**0.47**	0.19	0.20	0.20	0.19	**0.41**
	CNN	0.23	0.26	0.26	0.35	**0.41**	0.19	0.22	0.22	0.30	**0.35**
	TCN	0.31	0.25	0.25	0.25	**0.38**	0.26	0.21	0.21	0.21	**0.34**
	LSTM	0.37	0.37	0.37	0.40	**0.46**	0.39	0.36	0.36	0.41	**0.46**
MLP	WDCNN	0.09	0.09	0.09	0.19	**0.28**	0.07	0.07	0.07	0.16	**0.24**
	CNN	0.19	0.17	0.17	**0.40**	0.39	0.16	0.14	0.14	**0.37**	0.34
	TCN	0.28	0.19	0.19	0.31	**0.37**	0.23	0.16	0.16	0.31	**0.34**
	LSTM	0.34	0.33	0.33	0.38	**0.47**	0.36	0.33	0.33	0.41	**0.46**
CNN	WDCNN	0.10	0.08	0.08	0.16	**0.28**	0.08	0.07	0.07	0.13	**0.24**
	MLP	0.21	0.19	0.19	0.22	**0.47**	0.17	0.16	0.16	0.18	**0.41**
	TCN	0.39	0.18	0.18	0.33	**0.39**	0.33	0.15	0.15	0.31	**0.35**
	LSTM	0.47	0.41	0.41	**0.53**	0.49	0.45	0.38	0.38	**0.58**	0.50
TCN	WDCNN	0.12	0.12	0.12	0.22	**0.29**	0.10	0.10	0.10	0.18	**0.25**
	MLP	0.21	0.20	0.20	0.21	**0.46**	0.17	0.17	0.17	0.17	**0.40**
	CNN	0.26	0.22	0.22	**0.51**	0.49	0.21	0.19	0.19	**0.48**	0.43
	LSTM	0.41	0.48	0.48	0.43	**0.48**	0.41	0.44	0.44	0.42	**0.46**
LSTM	WDCNN	0.07	0.07	0.07	0.04	**0.28**	0.06	0.05	0.05	0.04	**0.24**
	MLP	0.18	0.18	0.18	0.18	**0.47**	0.15	0.15	0.15	0.17	**0.41**
	CNN	0.19	0.15	0.15	0.32	**0.43**	0.16	0.13	0.13	0.33	**0.40**
	TCN	0.23	0.19	0.19	0.13	**0.37**	0.18	0.16	0.16	0.11	**0.33**
Average		0.24	0.22	0.22	0.29	**0.41**	0.22	0.19	0.19	0.27	**0.37**

Table 3
Attack performance on Recall and Precision on the CWRU dataset.

Model		Attack performance on recall					Attack performance on precision				
Source	Target	FGSM	PGD	TIFGSM	NIA	SAEM	FGSM	PGD	TIFGSM	NIA	SAEM
WDCNN	MLP	0.19	0.20	0.20	0.19	**0.42**	0.19	0.19	0.19	0.19	**0.40**
	CNN	0.19	0.22	0.22	0.29	**0.35**	0.17	0.20	0.20	0.30	**0.34**
	TCN	0.27	0.21	0.21	0.21	**0.33**	0.18	0.15	0.15	0.17	**0.23**
	LSTM	0.31	0.31	0.31	0.34	**0.42**	0.18	0.18	0.18	0.21	**0.35**
MLP	WDCNN	0.07	0.07	0.07	0.16	**0.24**	0.07	0.07	0.07	0.14	**0.22**
	CNN	0.16	0.14	0.14	0.33	**0.33**	0.14	0.13	0.13	**0.37**	0.33
	TCN	0.24	0.16	0.16	0.26	**0.32**	0.17	0.12	0.12	0.21	**0.21**
	LSTM	0.28	0.27	0.27	0.32	**0.43**	0.17	0.16	0.16	0.23	**0.35**
CNN	WDCNN	0.09	0.07	0.07	0.13	**0.24**	0.08	0.07	0.07	0.12	**0.22**
	MLP	0.18	0.16	0.16	0.19	**0.42**	0.17	0.16	0.16	0.18	**0.40**
	TCN	0.33	0.16	0.16	0.27	**0.34**	0.22	0.12	0.12	0.24	**0.26**
	LSTM	0.39	0.35	0.35	**0.47**	0.45	0.31	0.31	0.31	**0.42**	0.39
TCN	WDCNN	0.10	0.10	0.10	0.18	**0.24**	0.09	0.09	0.09	0.16	**0.23**
	MLP	0.17	0.17	0.17	0.17	**0.40**	0.17	0.16	0.16	0.17	**0.39**
	CNN	0.22	0.19	0.19	**0.43**	0.42	0.18	0.16	0.16	**0.49**	0.42
	LSTM	0.34	0.40	0.40	0.35	**0.44**	0.23	0.35	0.35	0.26	**0.36**
LSTM	WDCNN	0.06	0.06	0.06	0.03	**0.24**	0.06	0.05	0.05	0.04	**0.22**
	MLP	0.16	0.15	0.15	0.16	**0.41**	0.15	0.15	0.15	0.17	**0.39**
	CNN	0.16	0.13	0.13	0.28	**0.37**	0.15	0.12	0.12	0.31	**0.40**
	TCN	0.19	0.16	0.16	0.11	**0.33**	0.15	0.12	0.12	0.09	**0.21**
Average		0.20	0.18	0.18	0.24	**0.36**	0.16	0.15	0.15	0.22	**0.32**

includes vibration signals collected under different experimental conditions: various rotating speeds, bearing types, vertical loads, and other factors. Each bearing in the test rig has inner and outer diameters of 35 mm and 80 mm, respectively. In addition, these bearings can take three fault conditions, including ball, inner race, and outer race faults, along with a normal one. The signals were collected at a sampling rate of 10 kHz under all experimental conditions, resulting in 10,000 observations per second. In this experiment, we set the rotating speed and vertical load to 900 rpm and 200 kgf, respectively. In addition, the signals were collected for five minutes, with one-minute pre- and post-heating periods at the beginning and end, respectively, resulting in a total of 1,800,000 observations. We used the signals gathered from one bearing nearest to the drive motor among the three bearings.

Using this practical bearing faults dataset with our practical bearing fault diagnosis system and the CWRU dataset, we evaluate the performance of our proposed method, SAEM, and compare it with other methods. For comparison, we consider four different attacks: FGSM (Goodfellow et al., 2014), PGD (Madry et al., 2017), TIFGSM (Xie et al., 2019) and NIA (Kim et al., 2023). FGSM and PGD are basic adversarial attacks previously studied in recent works on fault diagnosis systems (Ge et al., 2021). TIFGSM and NIA are recently proposed black-box attacks, which have been shown to generate highly malicious adversarial examples. For all methods, we fixed the number of steps at ten and followed the original papers' hyperparameters for each attack. Our attack method utilized $\beta = 100$ and an ensemble of two window sizes, [128, 1024], which exhibited the highest performance in Section 4.3. The target signal $\boldsymbol{x}^*$ is randomly sampled from the signals that have different labels to the label of the original signal $\boldsymbol{x}$. To comprehensively assess the performance of our proposed method, we estimate the attack performance using four different measures: accuracy, precision, recall, and F1 score. These measures are crucial in fault diagnosis, as misclassification can lead to unnecessary inspections or economic losses due to malfunctions.

Table 4
Attack performance on Accuracy and F1 score on the practical bearing faults dataset.

Model		Attack performance on accuracy					Attack performance on F1 score				
Source	Target	FGSM	PGD	TIFGSM	NIA	SAEM	FGSM	PGD	TIFGSM	NIA	SAEM
WDCNN	MLP	0.25	0.24	0.26	0.24	**0.27**	0.24	0.23	0.25	0.24	**0.26**
	TCN	0.19	0.20	0.07	0.21	**0.42**	0.18	0.20	0.07	0.21	**0.39**
	LSTM	0.22	0.26	0.02	0.24	**0.26**	0.22	**0.27**	0.02	0.25	0.26
	CNN	0.19	0.42	0.04	0.44	**0.51**	0.19	0.43	0.05	0.45	**0.48**
MLP	TCN	0.02	0.02	0.01	0.02	**0.17**	0.02	0.02	0.01	0.02	**0.17**
	LSTM	0.05	0.03	0.00	0.03	**0.06**	0.05	0.03	0.00	0.03	**0.06**
	WDCNN	0.00	0.00	0.00	0.00	**0.03**	0.00	0.00	0.00	0.00	**0.03**
	CNN	0.04	0.03	0.01	0.03	**0.10**	0.04	0.03	0.01	0.03	**0.10**
CNN	MLP	0.23	0.22	0.24	0.23	**0.26**	0.22	0.22	0.23	0.22	**0.25**
	TCN	0.42	0.58	0.06	0.61	**0.64**	0.39	0.54	0.05	0.56	**0.59**
	LSTM	0.48	0.72	0.02	**0.79**	0.72	0.48	0.74	0.02	**0.79**	0.70
	WDCNN	0.05	0.06	0.00	0.10	**0.22**	0.05	0.06	0.00	0.10	**0.22**
TCN	MLP	0.25	0.24	0.22	0.26	**0.28**	0.25	0.24	0.22	0.25	**0.27**
	LSTM	0.70	0.79	0.07	0.79	**0.88**	0.70	0.82	0.08	0.83	**0.89**
	WDCNN	0.08	0.04	0.00	0.05	**0.11**	0.08	0.04	0.00	0.05	**0.11**
	CNN	0.65	0.75	0.05	0.77	**0.86**	0.63	0.79	0.05	0.80	**0.87**
LSTM	MLP	0.23	0.22	0.23	0.23	**0.24**	0.22	0.22	0.22	0.22	**0.23**
	TCN	0.65	**0.73**	0.19	0.71	0.64	0.62	**0.76**	0.19	0.75	0.63
	WDCNN	0.21	0.07	0.00	0.03	**0.23**	0.21	0.07	0.00	0.03	**0.23**
	CNN	0.61	0.78	0.06	0.73	**0.79**	0.70	0.78	0.06	0.78	**0.79**
Average		0.28	0.32	0.08	0.33	**0.38**	0.28	0.32	0.08	0.33	**0.38**

Table 5
Attack performance on Recall and Precision on the practical bearing faults dataset.

Model		Attack performance on recall					Attack performance on precision				
Source	Target	FGSM	PGD	TIFGSM	NIA	SAEM	FGSM	PGD	TIFGSM	NIA	SAEM
WDCNN	MLP	0.24	0.23	0.25	0.24	**0.27**	0.24	0.22	0.23	0.23	**0.24**
	TCN	0.18	0.20	0.07	0.21	**0.41**	0.15	0.17	0.07	0.18	**0.30**
	LSTM	0.23	**0.27**	0.02	0.25	0.27	0.13	0.20	0.02	0.20	**0.21**
	CNN	0.20	0.43	0.05	0.46	**0.51**	0.12	0.33	0.04	0.36	**0.38**
MLP	TCN	0.02	0.02	0.01	0.02	**0.17**	0.02	0.02	0.01	0.02	**0.13**
	LSTM	0.05	0.03	0.00	0.03	**0.06**	0.04	0.02	0.00	0.03	**0.05**
	WDCNN	0.00	0.00	0.00	0.00	**0.03**	0.00	0.00	0.00	0.00	**0.03**
	CNN	0.04	0.03	0.01	0.03	**0.10**	0.03	0.02	0.01	0.02	**0.08**
CNN	MLP	0.22	0.22	0.23	0.23	**0.26**	0.21	0.20	0.22	0.21	**0.23**
	TCN	0.42	0.57	0.05	0.60	**0.63**	0.33	0.42	0.05	**0.44**	0.40
	LSTM	0.49	0.73	0.02	**0.80**	0.73	0.24	0.53	0.02	**0.68**	0.53
	WDCNN	0.05	0.06	0.00	0.10	**0.22**	0.05	0.06	0.00	0.10	**0.21**
TCN	MLP	0.25	0.24	0.22	0.25	**0.27**	0.24	0.23	0.21	0.24	**0.25**
	LSTM	0.71	0.79	0.08	0.80	**0.88**	0.45	0.83	0.08	**0.84**	0.67
	WDCNN	0.08	0.04	0.00	0.05	**0.11**	0.08	0.04	0.00	0.05	**0.10**
	CNN	0.66	0.76	0.05	0.78	**0.86**	0.50	0.65	0.05	0.67	**0.73**
LSTM	MLP	0.23	0.22	0.22	0.22	**0.24**	0.21	0.21	0.21	0.21	**0.22**
	TCN	0.64	**0.71**	0.19	0.68	0.62	0.56	**0.78**	0.16	0.77	0.54
	WDCNN	0.21	0.07	0.00	0.03	**0.22**	0.19	0.07	0.00	0.03	**0.19**
	CNN	0.62	**0.80**	0.06	0.75	**0.80**	0.66	**0.85**	0.05	0.83	0.64
Average		0.28	0.32	0.08	0.33	**0.38**	0.22	0.29	0.07	**0.31**	**0.31**

Table 2 and 3 summarize the results of the attack performance on the CWRU dataset for each method. Across all measures, the proposed attack (SAEM) demonstrates the most powerful performance across various combinations. Specifically, SAEM achieves an average accuracy drop of 0.41, while the other methods show drops below 0.3. This trend is consistently observed across all other measures, including F1 score, recall, and precision. Notably, TIFGSM, originally proposed in the vision domain with kernel operations, exhibits the lowest attack performance across all combinations, further supporting the notion that domain-specific knowledge cannot be easily transferred. These results confirm that existing methods, even those well performed in different domains, are insufficient for revealing the practical robustness of fault diagnosis models. Furthermore, the superior performance of our proposed method validates the effective utilization of spectral information for domain-adaptive noise generation.

The proposed method also outperforms the comparison methods on our practical bearing faults dataset, as shown in Table 4 and 5. Notably, except for three cases out of twenty, the proposed method consistently exhibits significantly higher attack performance. Specifically, when the source model is MLP, and the target model is TCN, all other attacks show nearly zero attack performance, while the proposed method achieves a 0.17 drop in both accuracy and F1 score. In summary, the above results demonstrate the efficiency of the proposed method in practical settings and its ability to unveil the practical robustness of fault diagnosis models.

5. Conclusion

This paper introduces a novel domain-adaptive attack method, Spectrogram-aware ensemble method (SAEM), which effectively assesses the robustness of bearing fault diagnosis models. Our experimental results demonstrate that the proposed method achieves superior attack performance under black-box settings, which closely resemble real-world industrial operations but have been inadequately explored.

By emphasizing the significance of domain-adaptive attacks, we highlight the need to address potential vulnerabilities and mitigate the associated dangers in industrial fault diagnosis systems. Based on our findings, further research should focus on exploring practical settings of industrial fault diagnosis systems and identifying potential risks within industrial environments. Future work could involve developing defense mechanisms that leverage spectral information to enhance the robustness of fault diagnosis systems and investigation on more practical settings including accessibility of training datasets, thereby providing trustworthy industrial systems.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Hoki Kim reports a relationship with Chung-Ang University and Seoul National University that includes: employment and travel reimbursement. Jaewook Lee reports a relationship with Seoul National University that includes: employment. Sang ho Lee reports a relationship with Dongguk University that includes: employment. Youngdoo Son reports a relationship with Dongguk University that includes: employment. Woojin Lee reports a relationship with Dongguk University that includes: employment.

Data availability

The data that has been used is confidential.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (Ministry of Science and Information & Communications Technology, MSIT, and Ministry of Education) (Nos. RS-2023-00208412, 2022R1F1A1074393, 2019R1A2C2002358, RS-2023-00271054), by the Chung-Ang University Research Grants in 2024, by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry (IPET) through Smart Agri Products Flow Storage Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) of Korea (No. 322050-3), and by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-2020-0-01789) and the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00254592).

References

Akhenia, P., Bhavsar, K., Panchal, J., Vakharia, V., 2022. Fault severity classification of ball bearing using SinGAN and deep convolutional neural network. Proc. Inst. Mech. Eng. C 236 (7), 3864–3877.

Bai, S., Kolter, J.Z., Koltun, V., 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

Chen, C.-C., Liu, Z., Yang, G., Wu, C.-C., Ye, Q., 2020. An improved fault diagnosis using 1d-convolutional neural network model. Electronics 10 (1), 59.

Cheng, Y., Yang, R., Zhang, Z., Suo, J., Dai, Q., 2023. A mutually boosting dual sensor computational camera for high quality dark videography. Inf. Fusion.

Choi, Y., Park, J., Lee, J., Kim, H., 2023. Exploring diverse feature extractions for adversarial audio detection. IEEE Access 11, 2351–2360.

Cohen, J., Rosenfeld, E., Kolter, Z., 2019. Certified adversarial robustness via randomized smoothing. In: International Conference on Machine Learning. PMLR, pp. 1310–1320.

Dong, Y., Pang, T., Su, H., Zhu, J., 2019. Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4312–4321.

Gao, Y., Kim, C.H., Kim, J.M., 2021. A novel hybrid deep learning method for fault diagnosis of rotating machinery based on extended WDCNN and long short-term memory. Sensors 21 (19), 6614.

Ge, Y., Wang, H., Liu, Z., 2021. Adversarial attack for deep-learning-based fault diagnosis models. In: 2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion. QRS-C, IEEE, pp. 757–761.

Goodfellow, I.J., Shlens, J., Szegedy, C., 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Gryllias, K.C., Antoniadis, I.A., 2012. A support vector machine approach based on physical model training for rolling element bearing fault detection in industrial environments. Eng. Appl. Artif. Intell. 25 (2), 326–344.

Harris, B.W., Milo, M.W., Roan, M.J., 2016. A general anomaly detection approach applied to rolling element bearings via reduced-dimensionality transition matrix analysis. Proc. Inst. Mech. Eng. C 230 (13), 2169–2180.

Hendriks, J., Dumond, P., Knox, D., 2022. Towards better benchmarking using the CWRU bearing fault dataset. Mech. Syst. Signal Process. 169, 108732.

Hou, Y., Wang, J., Chen, Z., Ma, J., Li, T., 2023. Diagnosisformer: An efficient rolling bearing fault diagnosis method based on improved transformer. Eng. Appl. Artif. Intell. 124, 106507.

Huang, Q., Katsman, I., He, H., Gu, Z., Belongie, S., Lim, S.N., 2019. Enhancing adversarial example transferability with an intermediate level attack. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4733–4742.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A., 2019. Deep learning for time series classification: a review. Data Min. Knowl. Discov. 33 (4), 917–963.

Jia, F., Lei, Y., Lin, J., Zhou, X., Lu, N., 2016. Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. Mech. Syst. Signal Process. 72, 303–315.

Karim, F., Majumdar, S., Darabi, H., Chen, S., 2017. LSTM fully convolutional networks for time series classification. IEEE Access 6, 1662–1669.

Kim, H., Park, J., Lee, J., 2023. Generating transferable adversarial examples for speech classification. Pattern Recognit. 137, 109286.

Kumar Dwivedi, S., Gupta, V., Mitra, R., Ahmed, S., Jain, A., 2019. Protogan: Towards few shot learning for action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops.

Kwon, H., Lee, S., 2022. Ensemble transfer attack targeting text classification systems. Comput. Secur. 117, 102695.

Lanham, C., 2002. Understanding the Tests That are Recommended for Electric Motor Predictive Maintenance. Baker Instrument Company.

Li, B., Chen, C., Wang, W., Carin, L., 2019. Certified adversarial robustness with additive noise. Adv. Neural Inf. Process. Syst. 32.

Li, Y., Zou, W., Jiang, L., 2022. Fault diagnosis of rotating machinery based on combination of wasserstein generative adversarial networks and long short term memory fully convolutional network. Measurement 191, 110826.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A., 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

Pandhare, V., Singh, J., Lee, J., 2019. Convolutional neural network based rolling-element bearing fault diagnosis for naturally occurring and progressing defects using time-frequency domain features. In: 2019 Prognostics and System Health Management Conference. PHM-Paris, IEEE, pp. 320–326.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A., 2017. Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. pp. 506–519.

Patel, S., Nargunde, R., Verma, S., Dhage, S., 2022. Summarization and simplification of medical articles using natural language processing. In: 2022 13th International Conference on Computing Communication and Networking Technologies. ICCCNT, IEEE, pp. 1–6.

Pham, M.T., Kim, J.M., Kim, C.H., 2020. Accurate bearing fault diagnosis under variable shaft speed using convolutional neural networks and vibration spectrogram. Appl. Sci. 10 (18), 6385.

Roy, S.S., Dey, S., Chatterjee, S., 2020. Autocorrelation aided random forest classifier-based bearing fault detection framework. IEEE Sens. J. 20 (18), 10792–10800.

Shaukat, K., Luo, S., Varadharajan, V., 2022. A novel method for improving the robustness of deep learning-based malware detectors against adversarial attacks. Eng. Appl. Artif. Intell. 116, 105461.

Smith, W.A., Randall, R.B., 2015. Rolling element bearing diagnostics using the case western reserve university data: A benchmark study. Mech. Syst. Signal Process. 64, 100–131.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R., 2013. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

Tian, J., Morillo, C., Azarian, M.H., Pecht, M., 2015. Motor bearing fault detection using spectral kurtosis-based feature extraction coupled with K-nearest neighbor distance analysis. IEEE Trans. Ind. Electron. 63 (3), 1793–1803.

Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L., 2019. Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2730–2739.

Yang, W., Yuan, J., Wang, X., Zhao, P., 2022. TSadv: Black-box adversarial attack on time series with local perturbations. Eng. Appl. Artif. Intell. 114, 105218.

Zhang, Q., Li, X., Chen, Y., Song, J., Gao, L., He, Y., Xue, H., 2022a. Beyond imagenet attack: Towards crafting adversarial examples for black-box domains. arXiv preprint arXiv:2201.11528.

Zhang, W., Peng, G., Li, C., Chen, Y., Zhang, Z., 2017. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. Sensors 17 (2), 425.

Zhang, X., Zhang, X., Liu, W., Zou, X., Sun, M., Zhao, J., 2022b. Waveform level adversarial example generation for joint attacks against both automatic speaker verification and spoofing countermeasures. Eng. Appl. Artif. Intell. 116, 105469.

Zhang, S., Zhang, S., Wang, B., Habetler, T.G., 2020. Deep learning algorithms for bearing fault diagnostics—A comprehensive review. IEEE Access 8, 29857–29881.
Zhao, S., Gao, C., Shao, Y., Li, L., Yu, C., Ji, Z., Sang, N., 2020. Gtnet: Generative transfer network for zero-shot object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 07. pp. 12967–12974.
Zheng, H., Wu, Z., Duan, S., Chen, Y., 2021. Research on fault diagnosis method of rolling bearing based on TCN. In: 2021 12th International Conference on Mechanical and Aerospace Engineering. ICMAE, IEEE, pp. 489–493.
Zhou, W., Hou, X., Chen, Y., Tang, M., Huang, X., Gan, X., Yang, Y., 2018. Transferable adversarial perturbations. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 452–467.
Zhuo, Y., Ge, Z., 2021. Data guardian: A data protection scheme for industrial monitoring systems. IEEE Trans. Ind. Inform. 18 (4), 2550–2559.
Zhuo, Y., Yin, Z., Ge, Z., 2022. Attack and defense: Adversarial security of data-driven FDC systems. IEEE Trans. Ind. Inform. 19 (1), 5–19.