Contents lists available at ScienceDirect

Applied Soft Computing

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

Black-box adversarial examples via frequency distortion against fault diagnosis systems

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

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

ARTICLE INFO

Keywords:
Adversarial attack
Black-box setting
Fourier transform
Rotating machinery
Fault diagnosis

ABSTRACT

Deep learning has significantly impacted prognostic and health management, but its susceptibility to adversarial attacks raises security risks for fault diagnosis systems. Previous research on the adversarial robustness of these systems is limited by unrealistic assumptions about prior model knowledge, which is often unobtainable in the real world, and by a lack of integration of domain-specific knowledge, particularly frequency information crucial for identifying unique characteristics for machinery states. To address these limitations and enhance robustness assessments, we propose a novel adversarial attack method that exploits frequency distortion. Our approach corrupts both frequency components and waveforms of vibration signals from rotating machinery, enabling a more thorough evaluation of system vulnerability without requiring access to model information. Through extensive experiments on two bearing datasets, including a self-collected dataset, we demonstrate the effectiveness of the proposed method in generating malicious yet imperceptible examples that remarkably degrade model performance, even without access to model information. In realistic attack scenarios for fault diagnosis systems, our approach produces adversarial examples that mimic unique frequency components associated with the deceived machinery states, leading to average performance drops of approximately 13 and 19 percentage points higher than existing methods on the two datasets, respectively. These results reveal potential risks for deep learning models embedded in fault diagnosis systems, highlighting the need for enhanced robustness against adversarial attacks.

1. Introduction

As the complexity of industrial equipment increases due to diverse machinery and its advanced functions, safety issues in the industry can occur even with minor faults on the part of the equipment [1,2]. Thus, prognostics and health management (PHM), in which fault diagnosis is one of the important tasks, has emerged as an essential research field to maintain the stability and reliability of industrial systems [3,4].

With the growing availability of industrial sensor data, data-driven approaches have been widely used in PHM [5,6]. In particular, recent advancements in deep learning have shown remarkable promising performance in fault diagnosis [7–9]. However, one potential drawback of deep learning models embedded in fault diagnosis systems is their vulnerability to subtle noise, particularly maliciously perturbed examples known as adversarial examples, which are generated by adversarial attacks [10–12]. These attacks typically exploit the high-dimensional and nonlinear nature of deep learning models, leading to substantial distortions in the output even from minor input perturbations [13–15]. In other words, malicious and undetectable perturbations to vibration signals from rotating machinery can cause deep learning-based fault diagnosis models to incorrectly predict machinery states, leading to safety accidents or opportunity costs [16,17]. Therefore, given the increasing use of deep learning-based fault diagnosis models, it is essential to evaluate their practical robustness against adversarial attacks.

Nevertheless, to our knowledge, the robustness evaluation of deep learning-based fault diagnosis models against adversarial attacks has been underexplored. Although some studies assessed the potential risk of adversarial attacks against the models [18–20], these evaluations have two notable limitations. First, previous studies have primarily focused on adversarial attacks under an unrealistic white-box setting, which assumes full access to model information, such as its parameters and structure. However, in real-world scenarios, such direct access is often unattainable, making these methods less effective and practical.

* Corresponding author.
** Corresponding author at: Department of Industrial and Systems Engineering, Dongguk University-Seoul, Pildong-ro 1-gil, Seoul, Republic of Korea.
E-mail addresses: wj926@dgu.ac.kr (W. Lee), youngdoo@dongguk.edu (Y. Son).

https://doi.org/10.1016/j.asoc.2025.112828
Received 9 August 2023; Received in revised form 9 November 2024; Accepted 22 January 2025
Available online 10 February 2025

Second, these studies relied on adversarial attack techniques originally developed for the computer vision domain, such as the fast gradient sign method (FGSM) [14] and projected gradient descent (PGD) [21], without considering the domain-specific knowledge that is beneficial for identifying unique model-independent characteristics for each machinery state; hence, they can overestimate the actual robustness of the models.

Therefore, we propose a novel adversarial attack that leverages frequency information to generate adversarial examples. Since vibration signals from rotating machinery are typically created by the interference of specific frequency components [22], their frequency information enables capturing unique model-independent characteristics of each machinery state [23,24]. Unlike existing approaches that only disrupt waveforms in the time domain, our method explicitly manipulates the frequency components of the vibration signals, which are fundamental to identifying machinery states. By distorting these frequency components, we create transferable adversarial examples capable of successfully deceiving fault diagnosis models for rotating machinery even under a *black-box setting*, which is a realistic yet strong scenario that is inaccessible to any prior model information. This approach yields adversarial examples that are effective for various fault diagnosis models, significantly enhancing the evaluation of the practical robustness of these systems against adversarial attacks. In Section 3.2, we further discuss the benefits of frequency distortion in terms of adversarial attacks.

To demonstrate the superiority of the proposed method compared to the existing attack methods, we performed extensive experiments on a self-collected Dongguk University (DU) bearing dataset, first unveiled in this paper, in addition to the Case Western Reserve University (CWRU) bearing dataset [25], a widely used benchmark. Consequently, we validated the overwhelming attack performance of our approach on various fault diagnosis models compared to other existing attacks, uncovering the practical risks of fault diagnosis systems against adversarial attacks.

The main contributions of this study are as follows:

- We propose a novel adversarial attack method that generates transferable adversarial examples to thoroughly assess the practical robustness of fault diagnosis models for rotating machinery.
- To obtain malicious yet imperceptible examples in both time and frequency domains, we manipulate the frequency components, which contain the model-independent characteristics, along with the waveforms.
- Experimental results on two bearing datasets, including our DU bearing dataset, highlight the security vulnerabilities of fault diagnosis systems by showing that the proposed method notably degrades model performance in the black-box setting.

The remainder of this paper is organized as follows. In Section 2, we briefly review the fault diagnosis systems and adversarial attacks against them. Next, we introduce the motivation and detailed algorithm of our approach in Section 3. In Section 4, we present extensive experimental settings and results on two bearing datasets, including the self-collected dataset, demonstrating the superiority of our approach. Finally, concluding remarks are provided in Section 5.

2. Related work

2.1. Deep learning-based fault diagnosis

Fault diagnosis is one of the essential tasks in PHM [5]. To maintain the stability and reliability of industrial systems, traditional fault diagnosis approaches typically leverage signal processing, statistical analysis, and physics-based models. However, they heavily depend on domain knowledge and predefined rules, making them often time-consuming, overlooking subtle data patterns, and lacking scalability and generalization [5,26,27]. Therefore, with the growing availability of industrial sensor data, data-driven fault diagnosis approaches based on machine learning techniques, such as support vector machines, k-nearest neighbor, and restricted Boltzmann machines, have been widely used rather than traditional approaches [5,6]. In particular, recent advances in deep learning have shown significant promising performance in fault diagnosis [7–9].

Let $\boldsymbol{x}$ be a vibration signal collected from the industrial monitoring system of rotating machinery, and y be the corresponding machinery state. Since a deep learning-based fault diagnosis model is treated as a classifier, the fault diagnosis is eventually a task that classifies the machinery state into the correct one, allowing us to know the reason for faults and then fix it on a timely basis. This process can be formulated as

$$f(\boldsymbol{x}) = \arg\max_{y\in\mathcal{Y}} P(y|\boldsymbol{x}), \tag{1}$$

where $\mathcal{Y}$ denotes a set of unique machinery states.

With increasing attention to fault diagnosis in PHM, various deep learning-based methods have been actively studied [28]. Some studies employed the existing deep learning architectures specialized for time-series analysis to handle the vibration signals collected from fault diagnosis systems. For example, Zheng et al. [29] adopted a temporal convolutional network (TCN), introduced in Bai et al. [30], for the fault diagnosis of bearings, and Li et al. [31] utilized a fully convolutional network with long short-term memory (LSTM-FCN), introduced in Karim et al. [32], to improve fault diagnosis performance for rotating machinery. Other studies presented deep learning architectures tailored to fault diagnosis. For example, Zhang et al. [33] introduced a deep convolutional neural network with wide first-layer kernels (WDCNN) for accurately classifying machinery states, even with the presence of noise in the signals. In addition, Gao et al. [34] improved WDCNN by combining it with LSTM to learn informative features of the signals for fault diagnosis, and Chen et al. [35] proposed a variant of one-dimensional CNN (CNN1D) architecture suitable for fault diagnosis by utilizing a residual network.

2.2. Adversarial attack

Previous research has shown that recent deep learning models for fault diagnosis are vulnerable to subtle noise, particularly maliciously perturbed examples known as adversarial examples [13]. These examples can be generated by adversarial attacks, which perturb original vibration signals to exploit weaknesses in the model [14]. Such attacks often stem from external malicious actors and system vulnerabilities, and their effectiveness is influenced by several factors, including model architectures, data distribution, and attack strategies. As a result, adversarial attacks can cause misdiagnosis of fault diagnosis models, leading to safety accidents and substantial financial losses and undermining the reliability of fault diagnosis systems [16,17]. Therefore, to ensure their practicability and reliability, it is crucial to evaluate the robustness of deep learning-based fault diagnosis models against adversarial attacks.

Let $\boldsymbol{x}$ be a vibration signal and $\tilde{\boldsymbol{x}}$ be a maliciously corrupted signal that deceives the model to predict the wrong machinery state. Given a target model in the fault diagnosis system, $f_{\mathcal{T}}$, the adversarial attack aims to generate the adversarial example $\tilde{\boldsymbol{x}}$ satisfying $f_{\mathcal{T}}(\tilde{\boldsymbol{x}}) \neq f_{\mathcal{T}}(\boldsymbol{x})$ by optimizing the following objective:

$$\max_{\|\boldsymbol{\delta}\|\leq\epsilon} \mathcal{L}(f_{\mathcal{T}}(\boldsymbol{x}+\boldsymbol{\delta}), y), \tag{2}$$

where ϵ is the maximum perturbation size, and $\boldsymbol{\delta}$ is the perturbation for creating $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\delta}$ that deceives the target model $f_{\mathcal{T}}$.

To effectively learn the perturbation $\boldsymbol{\delta}$ by maximizing Eq. (2), Goodfellow et al. [14] proposed FGSM that directly uses the gradient accent approach as follows:

$$\boldsymbol{\delta} = \epsilon \cdot \mathtt{sign}(\nabla_{\boldsymbol{\delta}}\mathcal{L}(f_{\mathcal{T}}(\boldsymbol{x}+\boldsymbol{\delta}), y)), \tag{3}$$

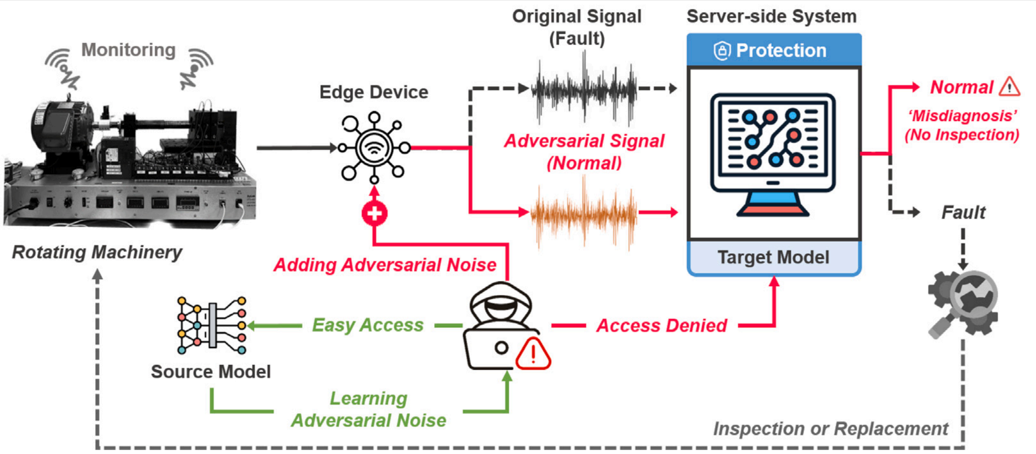


Fig. 1. Illustration of the practical vulnerability to adversarial attacks against fault diagnosis systems for rotating machinery. These systems are generally highly secured, preventing attackers from directly accessing the target model used for fault diagnosis. Consequently, adversarial attacks under the white-box setting are often unsubstantial. Instead, attackers can attempt to deceive the target model by generating adversarial noise using a more easily accessible source model.

where ∇ is the vector differential operator and sign is the element-wise sign operator. To further improve FGSM, Madry et al. [21] proposed PGD leveraging the multi-step optimization as follows:

$$\boldsymbol{\delta}^s = \Pi_\epsilon\left[\boldsymbol{\delta}^{s-1} + \alpha \cdot \mathtt{sign}(\nabla_{\boldsymbol{\delta}}\mathcal{L}(f_\mathcal{T}(\boldsymbol{x}+\boldsymbol{\delta}^{s-1}), y))\right], \tag{4}$$

where Π_ϵ denotes the projection function onto the ϵ-ball, α is the step size, and $s \in [1, S]$ is each optimization step in the predefined number of steps, S. In general, the initial perturbation $\boldsymbol{\delta}^0$ is defined as a random vector, and the final perturbation $\boldsymbol{\delta}^S$ is regarded as $\boldsymbol{\delta}$.

Recently, some studies observed the susceptibility of fault diagnosis models by merely applying the existing adversarial attacks, such as FGSM and PGD, under the white-box setting, which assumes that the attacker has complete prior knowledge of the target model, including the model information [11,18–20]. However, the white-box setting is often unrealistic because fault diagnosis models in most industrial monitoring systems are typically worked on private servers. In other words, the target model $f_\mathcal{T}$ is securely protected so that the attacker has difficulty accessing the gradients of the target model. Thus, the black-box setting, which assumes the target model $f_\mathcal{T}$ is unknown, is more practical and adequate for evaluating the actual danger of adversarial attacks for fault diagnosis models.

Under the black-box setting, a transfer attack, which learns the perturbation by a source model that is easy for attackers to access and injects it into the target model, is known to be the most effective approach [36–39]. That is, the transfer attack uses the source model $f_\mathcal{S}$ to generate adversarial examples to deceive the unknown target model $f_\mathcal{T}$. During this process, instead of using the gradients of $f_\mathcal{T}$, the attacker leverages the gradients of $f_\mathcal{S}$ for obtaining $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\delta}$.

However, vanilla gradient-based attacks, such as FGSM and PGD, may overestimate the practical robustness of the models because they cannot consider domain-specific knowledge that enables identifying model-independent characteristics for machinery states. In the computer vision domain, several studies have utilized domain-specific knowledge to generate transferable adversarial examples [40–42]. For example, inspired by the effect of the vision transformations against model performance drop, Xie et al. [40] introduced the diverse inputs iterative fast gradient sign method (DIM) that leverages the vision transformations to generate malicious examples. DIM learns the perturbation $\boldsymbol{\delta}$ for $\tilde{\boldsymbol{x}}$ using the vision transformation ψ as follows:

$$\boldsymbol{\delta}^s = \Pi_\epsilon\left[\boldsymbol{\delta}^{s-1} + \alpha \cdot \mathtt{sign}(\nabla_{\boldsymbol{\delta}}\mathcal{L}(f_\mathcal{S}(\psi(\boldsymbol{x}+\boldsymbol{\delta}^{s-1})), y))\right]. \tag{5}$$

Similarly, the translation-invariant method (TIM) [41] applied the kernel convolution method during the gradient manipulation to reflect domain-specific knowledge that is beneficial for assessing the adversarial robustness of the target model under the black-box setting. Nevertheless, since these methods were designed for the computer vision domain, they do not reflect domain-specific knowledge associated with fault diagnosis for rotating machinery, so it is difficult to expect a sufficient performance drop in fault diagnosis.

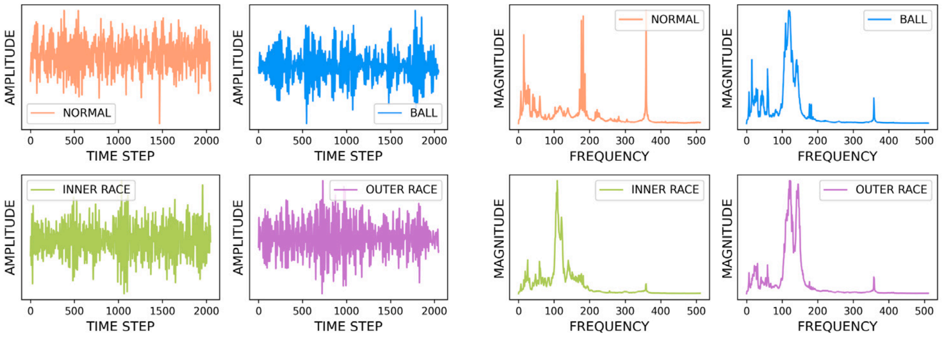
In contrast, our approach remarkably reduces fault diagnosis performance compared to the existing attack methods under the black-box setting by manipulating frequency information of vibration signals, which contains model-independent characteristics for each machinery state, successfully assessing the practical robustness of fault diagnosis systems.

3. Proposed method

As mentioned in Section 1, frequency information regarding the vibration signals, one of the domain-specific knowledge of fault diagnosis for rotating machinery, can capture distinct characteristics for each machinery state, regardless of fault diagnosis models. Thus, under the black-box setting, where attackers are unaware of the target model, this frequency information can be beneficial for verifying the practical robustness of the model [40–43]. Here, we first describe the main problem of fault diagnosis systems addressed in this study. Then, we analyze the adversarial examples obtained from conventional attack methods in the frequency domain and propose a novel adversarial attack method called Frequency Information-based Transfer Attack (FITA).

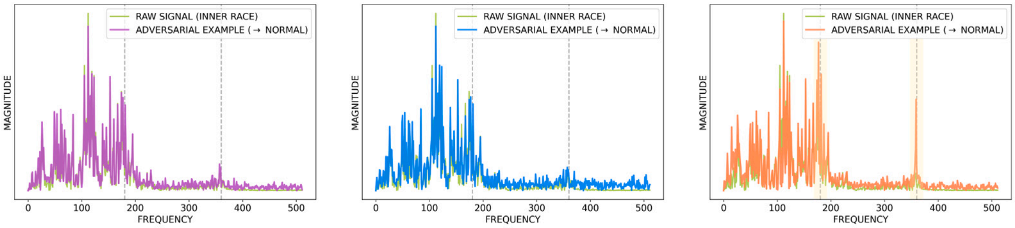
3.1. Problem statement

In Fig. 1, we illustrate the practical vulnerability to adversarial attacks against a fault diagnosis system for rotating machinery. In this system, vibration signals from rotating machinery, such as bearings, are monitored by multiple sensors. These signals are transmitted to edge devices, which collect sensor data and process it in real time. Then, the edge devices transfer the preprocessed signals to a highly secured server-side system to ensure its reliability. In the server-side system, a target model analyzes the received signals and diagnoses their current states to notify engineers to inspect them on time. Since the target model is also protected with high-level security, attackers with malicious intent often struggle to access this model directly. In other words, the attackers cannot perform adversarial attacks under the white-box setting, which assumes that the target model is publicly accessible [11,18]. Thus, they should consider the black-box setting, which is inaccessible to prior knowledge of the target model. In this scenario, the attackers generally learn adversarial noise from a source model, which is easily accessible, and inject the noise into the original signals to deceive the target model. As a result, the engineers



(a) Time domain (b) Frequency domain

Fig. 2. Averages of (a) waveforms and (b) frequency components obtained from vibration signals for each state. Each color indicates each state of bearing.



(a) PGD (b) FGSM (c) FITA (ours)

Fig. 3. Frequency components of a signal corrupted by (a) PGD, (b) FGSM, and (c) FITA. Only FITA exhibits increased magnitude at unique frequencies of the normal state.

may receive incorrect reports for the current state of the rotating machinery, leading to erroneous decisions that result in safety accidents due to insufficient inspections or substantial opportunity costs arising from unnecessary shutdowns for the maintenance of the rotating machinery [16,17].

Therefore, to investigate the practical risk of adversarial attacks against fault diagnosis systems for rotating machinery, we propose a novel transfer attack that incorporates domain-specific knowledge, which captures model-independent unique data characteristics that are beneficial in the black-box setting. Let $\boldsymbol{x}$ be a vibration signal and y be its corresponding state. Our objective is to learn a malicious perturbation $\boldsymbol{\delta}$ using a source model $f_{\mathcal{S}}$ to generate a transferable adversarial example $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\delta}$, which cause a target model $f_{\mathcal{T}}$ to misdiagnose y as the incorrect state $\tilde{y}$ ($\tilde{y} \neq y$).

3.2. Adversarial examples in frequency domain

In general, the vibration signals have sophisticated waveforms created by the interference of the diverse frequency components, making it difficult to distinguish their machinery states [22]. Thus, instead of the waveforms, their frequency information is often used to capture unique model-independent characteristics of each machinery state [23,24]. Here, we analyzed the adversarial examples created by the existing attack methods in the frequency domain using the CWRU bearing dataset to ensure the importance of frequency information.

Figs. 2(a) and 2(b) show waveforms and frequency components of the vibration signals for each state, respectively. Here, to derive the frequency information of the signals, we employed the Fast Fourier Transform (FFT) [44], which is an efficient way to compute the discrete Fourier transform of a signal. For each state, we observed that the frequency information is notably different, whereas the waveform is indistinguishable. Specifically, in this experiment, the signals in the normal state had fundamental and harmonic frequencies of about 180 (4.2 kHz) and 360 (8.4 kHz), respectively.

Even though the frequency information regarding the signals is important to identify each machinery state, the previous studies have only focused on corrupting their waveforms. Thus, adversarial examples generated by the existing attacks do not form unique frequency components for the signals in target states. For example, Figs. 3(a) and 3(b) show the frequency information for adversarial signals deceived the state from the inner race fault to the normal using popular conventional methods, PGD and FGSM. The corrupted signals were classified as the normal state but did not form the fundamental and harmonic frequencies of the normal state. In other words, the adversarial examples generated by only corrupting their waveforms cannot form the frequency components corresponding to the target state, making it easy to identify whether they have been attacked. Moreover, these adversarial examples tend to be specialized only to deceive the source model used to learn perturbations rather than generating genuinely confusing samples.

By contrast, as shown in Fig. 3(c), when we intentionally distorted the frequency information of the signals in addition to corrupting their waveforms, a notable difference in the frequency domain arose. Specifically, the frequency components of the signals distorted by our approach increased magnitude at frequencies of around 180 and 360, which are unique fundamental and harmonic frequency components of the normal state, respectively. This result implies that we can generate adversarial examples indistinguishable in both time and frequency domains by incorporating frequency distortion. In addition, as the frequency information contains model-independent characteristics for each machinery state, manipulating this information is even more crucial to corrupt the signals under the black-box setting. In Section 4.2.2, we demonstrate the efficacy of the frequency distortion in our method compared to the existing attack methods.

3.3. Frequency information-based transfer attack

We propose a novel adversarial attack method, *frequency information-based transfer attack (FITA)*, that generates transferable adversarial examples fatal for fault diagnosis systems by reflecting domain-specific knowledge. FITA explicitly distorts the frequency information of the vibration signals, which indicates inherent characteristics for each

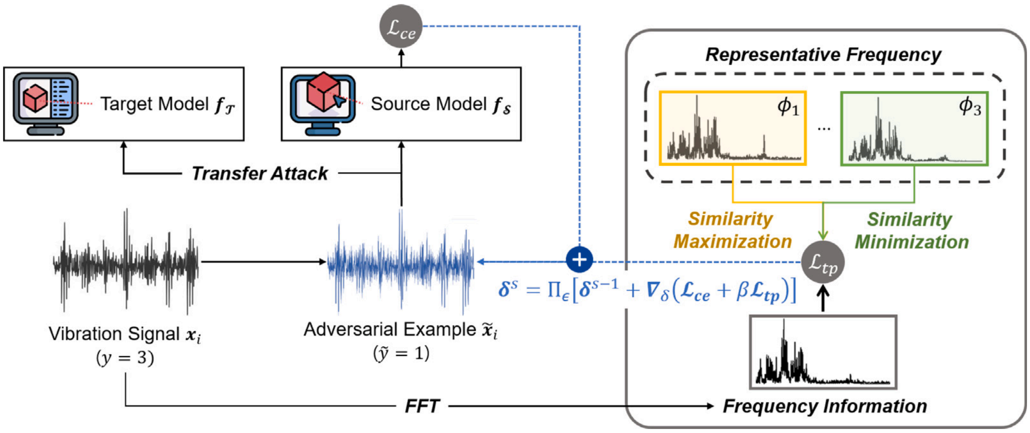


Fig. 4. Overview of the proposed method, FITA.

machinery state, along with their waveforms to achieve a significantly high attack performance under the black-box setting. Fig. 4 is an overview of our method, FITA, and its overall procedure is summarized in Algorithm 1.

Algorithm 1 Frequency Information-based Transfer Attack

```
Input: source model $f_\mathcal{S}$, vibration signal $\boldsymbol{x} \in \mathbb{R}^L$ and its label $y \in \mathbb{N}$,
  sets of signals $\mathcal{X}$ and labels $\mathcal{Y}$, number of classes $C$, number of steps
  $S$, maximum perturbation size $\epsilon$, step size $\alpha$, and frequency distortion
  parameter $\beta$
Output: adversarial example $\tilde{\boldsymbol{x}} \in \mathbb{R}^L$
  # Representative Frequency Set
  $\Phi \leftarrow \{\phi_c = \mathbb{E}_{x' \in \mathcal{X}_c}[\mathrm{FFT}(x')] \mid c \in \{1, ..., C\}\}$
  $\delta^0 \leftarrow 0$
  for $s \in \{1, ..., S\}$ do
      $\mathcal{L}_{ce} \leftarrow -y \log(f_\mathcal{S}(x + \delta^{s-1}))$    # Cross-entropy Loss
      $F_x \leftarrow \mathrm{FFT}(x + \delta^{s-1})$
      $\mathcal{L}_{tp} \leftarrow \max(\|F_x - \phi_{c=y}\|_2^2 - \|F_x - \phi_{c' \neq y}\|_2^2 + \xi, 0)$    # Triplet Loss
      $\delta^s \leftarrow \Pi_\epsilon[\delta^{s-1} + \alpha \cdot \nabla_\delta(\mathcal{L}_{ce} + \beta\mathcal{L}_{tp})]$    # Perturbation
  end for
  $\delta \leftarrow \delta^S$
  $\tilde{x} \leftarrow x + \delta$
```

Representative frequency set. Let $\mathcal{X}$ be a set of vibration signals with L observations, and $\mathcal{Y}$ be a set of labels indicating the machinery states of the signals. We first derive a *representative frequency set*, $\Phi = \{\phi_c\}_{c=1}^{C}$, where C is the number of classes. In specific, we apply FFT to each signal $\boldsymbol{x} \in \mathcal{X}$ to extract its frequency information as follows:

$$\mathrm{FFT}(\boldsymbol{x})[k] = \sum_{\ell=1}^{L} \boldsymbol{x}_\ell \cdot \exp(-2\pi i \cdot k\ell/L), \tag{6}$$

where $\mathrm{FFT}(\boldsymbol{x})[k]$ is the kth frequency component of $\boldsymbol{x}$, and i denotes the imaginary unit. In addition, $\exp(-2\pi i \cdot k\ell/L)$ is the twiddle factor representing the phase shift and amplitude of the kth complex exponential in $\boldsymbol{x} \in \mathbb{R}^L$. Then, for each state c, the representative frequency ϕ_c is calculated by averaging the magnitudes of the frequency components in $\mathcal{X}_c$, which consists of signals belonging to the state c, as follows:

$$\phi_c = \mathbb{E}_{x' \in \mathcal{X}_c}[\mathrm{FFT}(\boldsymbol{x}')]. \tag{7}$$

The representative frequency set characterizes the frequency information for each machinery state, thereby being utilized to corrupt an arbitrary signal $\boldsymbol{x} \in \mathcal{X}$ in the frequency domain when generating an adversarial example $\tilde{\boldsymbol{x}}$.

Subsequently, our approach learns a perturbation δ through *wavelet manipulation* and *frequency distortion* to create the adversarial example $\tilde{\boldsymbol{x}}$, which is corrupted from $\boldsymbol{x} \in \mathcal{X}$ in both the time and frequency domains.

Wavelet manipulation. Given a source model $f_\mathcal{S}$, we calculate the *cross-entropy* loss $\mathcal{L}_{ce}$ between a true machinery state y of $\boldsymbol{x}$ and the predicted state $f_\mathcal{S}(\boldsymbol{x} + \delta)$ as follows:

$$\mathcal{L}_{ce} = -y \log(f_\mathcal{S}(\boldsymbol{x} + \delta)). \tag{8}$$

By maximizing this loss, the waveform of $\boldsymbol{x}$ is perturbed to have a different pattern from the signals belonging to the state y in the time domain.

However, under the black-box setting, focusing solely on the waveform manipulation tends to generate adversarial examples that overfit the source model $f_\mathcal{S}$ because the unique model-independent characteristics for each machinery state are disregarded. Consequently, the practical robustness of a target model $f_\mathcal{T}$ can be overestimated. Therefore, we incorporate the frequency distortion with this waveform manipulation.

Frequency distortion. If we manipulate the frequency components to be solely distant from the original ones by using loss functions, such as mean squared error or Kullback–Leibler divergence, awkward frequency components that do not correspond to any meaningful machinery state can be created. Thus, we exploit the *triplet* loss function, which minimizes the distance between an anchor and a positive sample while maximizing the distance between the anchor and a negative one, to corrupt the frequency information of the signal $\boldsymbol{x}$. Here, the anchor is defined the frequency components $F_{\boldsymbol{x}}$ of signal $\boldsymbol{x}$ derived by $\mathrm{FFT}(\boldsymbol{x})$; the positive sample is the representative frequency corresponding to the label y of $\boldsymbol{x}$ ($\phi_{c=y} \in \Phi$); and the negative sample is $\phi_{c' \neq y} \in \Phi$, where c' is a randomly selected state from $\mathcal{Y}$ except y. Note that we can deliberately select the state c', which we desire to deceive. To measure the distances between the frequency components, we employ Euclidean distance; thereby, the triplet loss, $\mathcal{L}_{tp}$, is calculated as follows:

$$\mathcal{L}_{tp} = \max(\|F_{\boldsymbol{x}} - \phi_{c=y}\|_2^2 - \|F_{\boldsymbol{x}} - \phi_{c' \neq y}\|_2^2 + \xi, 0), \tag{9}$$

where ξ is a non-negative margin representing the minimum acceptable difference between the positive and negative distances. By maximizing this loss, we intentionally make $F_{\boldsymbol{x}}$ distant from its positive representative frequency, $\phi_{c=y}$, and similar to its negative one, $\phi_{c' \neq y}$. In other words, the frequency information of $\boldsymbol{x}$ becomes similar to the representative frequency of another state, not the frequency components of its original state.

Perturbation optimization. Through the multi-step optimization [21] with these two losses, $\mathcal{L}_{ce}$ and $\mathcal{L}_{tp}$, we obtain the malicious perturbation δ within maximum perturbation size ϵ. The perturbation δ^s corresponding to each optimization step $s \in \{1, \ldots, S\}$ is learned as follows:

$$\delta^s = \Pi_\epsilon[\delta^{s-1} + \alpha \cdot \nabla_\delta(\mathcal{L}_{ce} + \beta\mathcal{L}_{tp})], \tag{10}$$

where Π_ϵ is the projection function onto the ϵ-ball, α is the step size, and β is the frequency distortion parameter determining the influence of the triplet loss term. We set the initial perturbation δ^0 to a random vector, and the perturbation δ^S in the last step S is used as the resulting perturbation $\boldsymbol{\delta}$.

Finally, we generate the adversarial example $\tilde{\boldsymbol{x}}$ by adding the resulting perturbation $\boldsymbol{\delta}$, which is manipulated in both time and frequency domains, to the benign vibration signal $\boldsymbol{x}$ ($\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\delta}$).

4. Experiments

We considered two practical attack scenarios, *random* and *targeted*, which mirror real-world fault diagnosis systems for rotating machinery, as follows:

- *Random attack scenario*: We generate adversarial examples so that the machinery state assigned to each signal is perceived as another arbitrary state. This scenario mirrors the actual situation in which an attacker manipulates a signal with the normal state to be an arbitrary fault state, leading to opportunity costs due to unnecessary shutdowns for maintenance.
- *Targeted attack scenario*: Adversarial examples are obtained by attacking so that the original state of each signal is deliberately deceived to a specific state. This scenario can be regarded as a realistic situation in which a signal in one of the fault states is attacked to become the normal state, causing severe industrial safety accidents that endanger employees due to inadequate inspections and breakdowns of the rotating machinery.

In this section, we conducted a series of experiments to confirm that the proposed method generates sufficient adversarial examples to hinder fault diagnosis systems in both two attack scenarios.

4.1. Experimental settings

Datasets. To validate the superiority of our method in evaluating adversarial robustness, we used the CWRU and self-collected DU bearing datasets.

- **CWRU** bearing dataset is the most popular public bearing dataset and is widely used for research on fault diagnosis. The signals in the CWRU are collected at 12 kHz and 48 kHz under four types of motor loads (unit: hp): 0, 1, 2, and 3. The signals have three fault states: inner race, outer race, and ball faults. Each fault state has fault diameters (unit: inch) of 0.007, 0.014, and 0.021, respectively. Here, we used the signals collected from the drive-end accelerometer at a sampling rate of 48 kHz. In addition, we regard each state with different fault diameters and motor loads as the same state; thereby, there are four bearing states, including three fault states and a normal state.
- **DU** bearing dataset comprises the vibration signals acquired from a self-constructed test rig with three bearings, as shown in Fig. 5. This dataset assumes that noises are mixed in the signals from a bearing due to its adjacent bearings. In addition, the DU dataset comprises the signals collected under various experimental conditions, including different rotating speeds, types of bearings, and other factors. In this experiment, we set the rotating speed to 900 RPM and the vertical load to 200 kgf. Also, we used bearings with inner and outer diameters of 35 mm and 80 mm, respectively. There are three fault states, including ball, inner race, and outer race faults, along with a normal state. Since the vibration signals were collected with a sampling rate of 10 kHz, the dataset has 10,000 observations per second. Specifically, we collected data for five minutes, setting the preheating duration to the front and back of one minute; thereby, 1,800,000 observations were used. Note that we only used the signals obtained from the first bearing among the three bearings of the test rig.

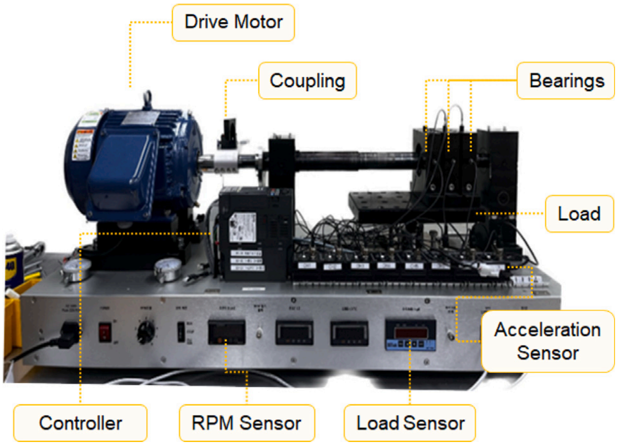


Fig. 5. DU Bearing Test Rig.

Data preprocessing. For all datasets, we set the length L of all signals $x \in \mathcal{X}$ to 2,048 and scaled them in a range from −1 to 1 by

$$\bar{x} = 2 \times \frac{x - x^{min}}{x^{max} - x^{min}} - 1,$$

where $\bar{x}$ is the scaled signal, and x^{max} and x^{min} denote the maximum and minimum observations in x, respectively.

Fault diagnosis models. We employed six deep learning model architectures popularly used as backbone models for bearing fault diagnosis: MLP [45], TCN [30], LSTM-FCN [32], GRU-FCN [46], WDCNN [33], and CNN1D [35].

Baseline attack methods. Transfer attacks have been primarily studied in the computer vision domain by considering its characteristics, making it challenging to apply these methods to fault diagnosis systems [47]. Moreover, transfer attack methods specialized for fault diagnosis systems have been underexplored. Therefore, we employed DIM [40] and TIM [41]—transfer attack methods that have demonstrated high attack performance across various domains [48,49]—along with two popular adversarial attack methods, PGD [21] and FGSM [14], to evaluate the attack performance of the proposed method for fault diagnosis systems.

Hyperparameters. For our approach, we set the number of steps S, maximum perturbation size ϵ, and step size α to 10^1, 10^{-1}, and $2\epsilon/S$, respectively. The frequency distortion parameter β is set to 10^2, and the margin, ξ, for the triplet loss term is set to 1.

Evaluation metric. We used the macro-averaged F1 score to evaluate the fault diagnosis performance for considering the class imbalance problem. The macro-averaged F1 score is calculated by averaging the class-wise F1 scores for C classes as follows:

$$\text{Macro-averaged F1 score} = \frac{1}{C}\sum_{c=1}^{C} \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c},$$

where C is the number of machinery states. For each state c, TP_c, FP_c, and FN_c denote the numbers of true positives, false positives, and false negatives, respectively.

Computational resources. All experiments were executed using the Pytorch platform on a system with an Intel Core i9-10900X CPU clocked at 3.70 GHz, 256 GB RAM, and GeForce RTX 3090 24 GB GPU.

4.2. Experimental results

Here, we first compared the attack performance of the proposed method against fault diagnosis models with that of the baselines under the black-box setting. Then, the advantages of the frequency distortion

Table 1
Macro-averaged F1 score for each model on the adversarial examples generated by each attack method using the CWRU and DU datasets in the random attack scenario.

Model		CWRU					DU				
Source	Target	PGD	FGSM	DIM	TIM	FITA	PGD	FGSM	DIM	TIM	FITA
MLP	TCN	0.6961	**0.6661**	0.7777	0.8428	0.6816	0.9761	0.9819	0.9810	0.9915	**0.7318**
	LSTM	0.5905	0.5847	0.6280	0.6666	**0.5593**	0.9722	0.9495	0.9860	0.9980	**0.8507**
	GRU	0.5501	0.5478	0.5605	0.5983	**0.5139**	0.9606	0.9388	0.9759	0.9980	**0.8697**
	WDCNN	0.8401	0.8613	0.8681	0.9292	**0.8027**	0.9980	1.0000	0.9980	0.9980	**0.9496**
	CNN1D	**0.6279**	0.6906	0.6605	0.8637	0.6425	0.9719	0.9607	0.9830	0.9940	**0.9142**
TCN	MLP	0.8291	0.8273	0.8363	0.8335	**0.7000**	0.7635	0.7540	0.7552	0.7818	**0.6190**
	LSTM	0.5800	0.5848	0.5970	0.5615	**0.4805**	**0.1779**	0.2963	0.1991	0.9223	0.2210
	GRU	0.5231	0.5202	0.5377	0.5459	**0.4363**	0.2151	0.3706	**0.2144**	0.9389	0.2876
	WDCNN	0.8192	0.8235	0.8268	0.8999	**0.7774**	0.9611	0.9211	0.8804	0.9960	**0.8204**
	CNN1D	0.5261	0.5547	0.5235	0.8112	**0.5126**	**0.2130**	0.3689	0.2225	0.9460	0.3126
LSTM	MLP	0.8213	0.8294	0.8175	0.8513	**0.7559**	0.7812	0.7758	0.7840	0.7779	**0.6138**
	TCN	0.8474	**0.7045**	0.8334	0.8418	0.7961	**0.2383**	0.3760	0.3562	0.8061	0.4822
	GRU	**0.0875**	0.2874	0.1046	0.3277	0.1924	**0.0000**	0.1249	0.0012	0.9514	0.3415
	WDCNN	0.9582	0.9332	0.9561	0.9453	**0.9325**	0.9271	0.7861	**0.6752**	0.9970	0.8247
	CNN1D	0.5741	0.6040	**0.5615**	0.8736	0.5721	0.1717	0.2960	**0.1331**	0.9377	0.4401
GRU	MLP	0.8198	0.8290	0.8199	0.8488	**0.7479**	0.7807	0.7719	0.7896	0.7678	**0.6100**
	TCN	0.8151	**0.7104**	0.8122	0.8756	0.7524	**0.2391**	0.4535	0.3717	0.8556	0.4965
	LSTM	0.1279	0.1961	**0.0569**	0.3374	0.2409	**0.0000**	0.0926	0.0000	0.9580	0.3793
	WDCNN	0.9595	0.9365	0.9540	0.9525	**0.9124**	0.8978	0.8153	**0.6926**	0.9950	0.7947
	CNN1D	0.6172	**0.5881**	0.6243	0.8877	0.6062	0.1493	0.2152	**0.0690**	0.9330	0.4246
WDCNN	MLP	0.8084	0.8142	0.8083	0.8033	**0.6528**	0.7683	0.7573	0.7875	0.7520	**0.5837**
	TCN	0.7844	0.7517	0.7707	0.7941	**0.6481**	0.8046	0.8185	0.7605	0.9302	**0.4636**
	LSTM	0.5803	0.5613	0.5769	0.6383	**0.4955**	0.7299	0.7762	0.7455	0.9822	**0.6662**
	GRU	0.5149	0.5106	0.5223	0.5694	**0.4527**	0.7573	0.8100	0.8090	0.9842	**0.7227**
	CNN1D	0.6943	0.6943	0.6427	0.7827	**0.6197**	**0.5740**	0.8085	0.5756	0.9529	0.6987
CNN1D	MLP	0.8224	0.8169	0.8339	0.8368	**0.7001**	0.7845	0.7804	0.7667	0.7701	**0.5792**
	TCN	0.7158	**0.5949**	0.6363	0.8469	0.6573	0.4644	0.6080	**0.3476**	0.9453	0.3716
	LSTM	0.4600	0.4385	**0.4359**	0.6172	0.4398	**0.2616**	0.5151	0.4531	0.9761	0.4338
	GRU	0.4490	0.4518	**0.4043**	0.5571	0.4095	**0.2752**	0.5538	0.4827	0.9781	0.4710
	WDCNN	0.8936	0.8587	0.8613	0.9302	**0.8348**	0.9393	0.9467	0.7950	0.9970	**0.7279**

used in our approach were discussed. Finally, we performed sensitivity analyses for the hyperparameters in FITA.

4.2.1. Transfer attack performance

Table 1 presents the macro-averaged F1 scores for each fault diagnosis model using adversarial examples generated by each baseline using two datasets, CWRU and DU, in the *random attack scenario*. For the CWRU dataset, FITA achieved an overwhelming performance drop on average compared to other attack methods. Specifically, when we used TCN and WDCNN as the source model, the performance of all target models decreased by about 10 percentage points, at least, compared to the baselines. For the DU dataset, FITA was also superior to the baselines. Especially, when we used MLP as the source model, the attack success rate of FITA was about 15 percentage points higher than the others.

Subsequently, we compared the transfer attack performance of FITA to that of the baseline attacks using the two bearing datasets in the *targeted attack scenario*, and the results are provided in Table 2. In this scenario, for both datasets, FITA also showed outstanding attack performance compared to the baselines, no matter what the source model is. In contrast, PGD and DIM exhibited a relatively high attack performance only when LSTM-FCN and GRU-FCN were used as the source model. In addition, FGSM and TIM failed to attack in most cases.

Furthermore, we verified the practicability of our method by comparing the transfer attack success rates of FITA with those of the baseline methods in the targeted attack scenario. We obtained adversarial examples using WDCNN, which shows the highest fault diagnosis performance in both CWRU and DU datasets, as the source model and then evaluated them using other fault diagnosis models as the target models. Here, the averaged success rates across the five target models for each attack method were provided. Consequently, as shown in Fig. 6, we observed that FITA showed overwhelming transfer attack success rates compared to the baseline attack methods, especially in deceiving from three fault states to the normal one.

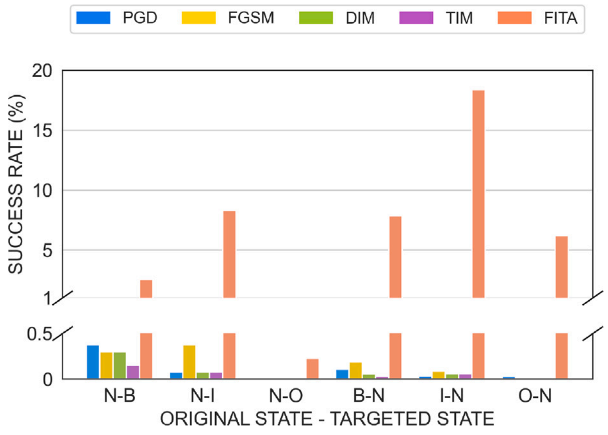


Fig. 6. Transfer attack success rates of FITA and the baselines in the targeted attack scenario (N: normal, B: ball, I: inner race, O: outer race).

These results in both scenarios support that the proposed method is effective in assessing the practical robustness of fault diagnosis systems by leveraging frequency information to generate adversarial examples.

4.2.2. Effect of frequency distortion

To discuss the effectiveness of our approach, we compared the change of frequency components of FITA to that of the other baseline attack methods when the signal was attacked from one of the fault states to the normal state in the targeted attack scenario. We used WDCNN as the source model to generate adversarial examples from each baseline.

Table 2
Macro-averaged F1 score for each model on the adversarial examples generated by each attack method using the CWRU and DU datasets in the targeted attack scenario.

Model		CWRU					DU				
Source	Target	PGD	FGSM	DIM	TIM	FITA	PGD	FGSM	DIM	TIM	FITA
MLP	TCN	0.9917	0.9906	0.9917	1.0000	**0.9462**	0.9929	0.9929	0.9929	0.9965	**0.7700**
	LSTM	0.7429	0.8984	0.8258	1.0000	**0.6518**	1.0000	1.0000	1.0000	1.0000	**0.7840**
	GRU	0.7096	0.9064	0.7992	1.0000	**0.5936**	1.0000	1.0000	1.0000	1.0000	**0.7709**
	WDCNN	0.9990	1.0000	0.9979	0.9990	**0.9746**	1.0000	1.0000	0.9988	1.0000	**0.7630**
	CNN1D	0.8221	0.9917	0.9462	1.0000	**0.6573**	1.0000	1.0000	1.0000	1.0000	**0.6182**
TCN	MLP	0.9958	0.9958	0.9969	0.9979	**0.8926**	0.8742	0.8713	0.8710	0.8830	**0.7288**
	LSTM	0.9990	0.9979	1.0000	0.9979	**0.7200**	0.3295	0.3312	0.3514	0.9988	**0.2629**
	GRU	0.9979	0.9948	0.9969	1.0000	**0.6304**	0.3312	0.3354	0.3437	0.9976	**0.2535**
	WDCNN	0.9990	0.9979	0.9979	1.0000	**0.9744**	0.9905	0.9771	0.9941	1.0000	**0.8058**
	CNN1D	1.0000	0.9894	0.9990	1.0000	**0.8437**	0.5130	0.3625	0.4615	0.9976	**0.2809**
LSTM	MLP	0.9791	0.9215	0.9705	0.9968	**0.8267**	0.8879	0.8855	0.8814	0.8769	**0.7177**
	TCN	0.9818	0.9852	0.9808	0.9979	**0.9222**	0.3979	0.4186	**0.3518**	0.8712	0.4323
	GRU	0.3291	0.5630	**0.2166**	0.9826	0.2465	0.2040	0.3651	**0.1343**	0.9976	0.2582
	WDCNN	1.0000	0.9979	0.9969	1.0000	**0.9765**	0.9435	0.9267	0.8985	1.0000	**0.7184**
	CNN1D	0.9270	0.7687	0.9050	1.0000	**0.6747**	0.5206	0.5009	**0.3718**	0.9976	0.3891
GRU	MLP	0.9783	0.8982	0.9782	0.9958	**0.8324**	0.8767	0.8816	0.8748	0.8777	**0.7155**
	TCN	0.9671	0.9884	0.9762	0.9979	**0.9222**	0.3703	0.4219	**0.2935**	0.8846	0.4203
	LSTM	0.1032	0.1712	**0.0941**	0.9938	0.2812	0.2122	0.3593	0.1715	0.9988	**0.1295**
	WDCNN	0.9979	0.9990	0.9979	0.9990	**0.9755**	0.9359	0.9411	0.9103	1.0000	**0.7231**
	CNN1D	0.9644	0.7809	0.9340	1.0000	**0.6623**	0.3757	0.4773	**0.3163**	0.9965	0.3359
WDCNN	MLP	0.9969	0.9968	0.9979	0.9979	**0.8891**	0.8956	0.8953	0.8833	0.8794	**0.6846**
	TCN	0.9979	0.9958	0.9958	0.9979	**0.9256**	0.8417	0.8712	0.8469	0.9747	**0.6313**
	LSTM	1.0000	0.9916	1.0000	1.0000	**0.7281**	0.8955	0.9429	0.8831	1.0000	**0.6814**
	GRU	0.9969	0.9969	0.9979	1.0000	**0.6841**	0.9044	0.9705	0.9184	1.0000	**0.6848**
	CNN1D	0.9990	0.9969	1.0000	1.0000	**0.7823**	0.8137	0.9746	0.7990	1.0000	**0.5303**
CNN1D	MLP	0.9937	0.9937	0.9937	0.9992	**0.8589**	0.8747	0.8722	0.8684	0.8802	**0.6852**
	TCN	0.9907	0.9477	0.9927	1.0000	**0.9330**	0.8303	0.8882	0.8307	0.9722	**0.6165**
	LSTM	0.8213	0.7424	0.7977	1.0000	**0.6681**	0.7426	0.7274	0.7147	1.0000	**0.5778**
	GRU	0.8573	0.7990	0.8116	1.0000	**0.6063**	0.7285	0.7201	0.6844	1.0000	**0.5693**
	WDCNN	0.9979	0.9958	0.9990	1.0000	**0.9756**	0.9665	0.9589	0.9733	1.0000	**0.6851**

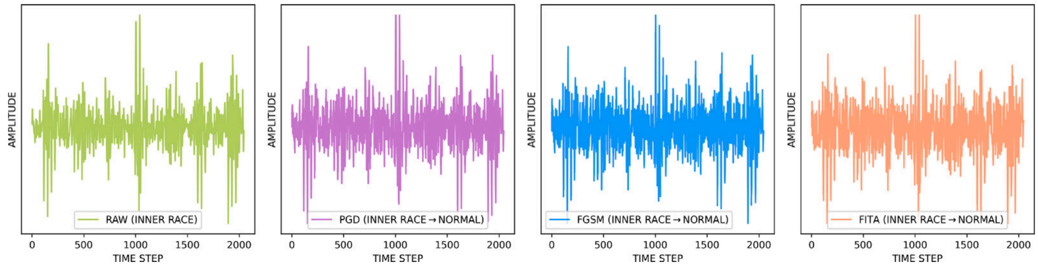


Fig. 7. Comparison of the waveforms for a raw signal and the signals corrupted by PGD, FGSM, and FITA in the time domain.

As shown in Fig. 7, in the time domain, there is no visible difference between the benign and corrupted signals. By contrast, as shown in Fig. 3, the adversarial example of FITA shows a notable difference in the frequency domain compared to those of PGD and FGSM. Interestingly, the frequency information for the signal attacked by FITA has an increased magnitude at the frequency components of around 180 (4.2 kHz) and 360 (8.4 kHz), mimicking the normal state (see Fig. 2(b)), whereas PGD and FGSM do not.

4.2.3. Sensitivity analysis

We analyzed the sensitivity of each hyperparameter used in FITA to provide a comprehensive understanding of its impact on the robustness evaluation of fault diagnosis models. Specifically, we examined three key hyperparameters: the maximum perturbation size ϵ, the number of steps S, and the frequency distortion parameter β. Here, we utilized WDCNN as the source model and the others as the target model.

Fig. 8(a) illustrates the macro-averaged F1 scores of the target models when exposed to the adversarial examples corrupted by FITA with different ϵ values. In all cases, the fault diagnosis performance declined substantially with larger values of ϵ, as it allows larger perturbation sizes.

In addition, in Fig. 8(b), we analyzed the impact of the number of steps S on the attack performance of our method. FITA exhibited a stable performance in response to variations in the number of steps, particularly when S exceeds approximately ten.

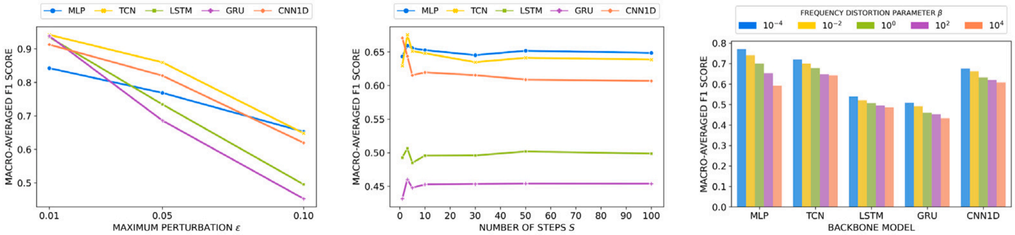
Fig. 8(c) compares model performance when using different β values to corrupt the signals. We observed a gradual performance drop across all target models as the β value increased. This finding reaffirms the effectiveness of FITA that incorporates the frequency distortion, as the frequency information is manipulated more strongly as the value of β increases.

5. Conclusion

In this study, we introduce FITA, a novel transfer attack method designed to assess the practical robustness of fault diagnosis systems against adversarial attacks. Here, we summarize the key findings and future research directions.

5.1. Summary of findings

FITA is developed to thoroughly evaluate the robustness of fault diagnosis systems for rotating machinery under adversarial conditions.



(a) Maximum Perturbation ϵ (b) Number of Steps S (c) Frequency Distortion Parameter β

Fig. 8. Macro-averaged F1 scores for five target models attacked by FITA with different values of (a) ϵ, (b) S, and (c) β.

By explicitly manipulating both frequency components and waveforms of the vibration signals, we can generate malicious yet imperceptible adversarial examples that sufficiently hinder fault diagnosis models even under the black-box setting, which is similar to real-world environments of fault diagnosis systems but has been underexplored.

Our experimental results on two bearing datasets, including the self-collected DU bearing dataset, highlight the benefit of frequency distortion, which creates imperceptible adversarial examples in the frequency domain. In addition, we also demonstrate that FITA, which incorporates frequency distortion with waveform manipulation, shows superior performance under the black-box setting. For example, in the random attack scenario using MLP as the source model on the DU dataset, FITA achieved a 15 percentage point higher attack success rate than the baselines. Similarly, the proposed method showed an overwhelming attack success rate compared to the baselines under the targeted attack scenario, regardless of the source models. These results validate FITA's effectiveness in evaluating practical robustness in fault diagnosis systems and underscore the need for further research to uncover and mitigate potential security risks against adversarial attacks.

5.2. Future research directions

In future work, we can devise defense mechanisms that counter adversarial attacks through waveform or frequency manipulation, enhancing the practical robustness of fault diagnosis systems and promoting the development of reliable ones. Given the recent attention to real-world industrial environments with limited data [50,51], this exploration can extend to adversarial attack and defense mechanisms specialized for such constrained data scenarios. Furthermore, exploring adversarial robustness to the traditional fault diagnosis models, which is often regarded as a separate research topic due to their different model structures and attack strategies from deep learning models [42,52], is also worthy of study. Another future research direction is developing an adversarial attack method applicable to industrial monitoring systems for different types of machinery beyond rotating machinery.

CRediT authorship contribution statement

Sangho Lee: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Hoki Kim:** Writing – review & editing, Validation, Software, Investigation, Formal analysis. **Woojin Lee:** Writing – review & editing, Supervision, Funding acquisition. **Youngdoo Son:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (MSIT) of Korea (Nos. RS-2023-00208412 and 2022R1F1A1074393), by the Ministry of Education (No. RS-2023-00271054), 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), by the MSIT of Korea under the Information Technology Research Center (ITRC) support program (IITP-2025-2020-0-01789 and IITP-2025-RS-2024-00438056), by the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2025-RS-2023-00254592), and by the artificial intelligence semiconductor support program to nurture the best talents (IITP-2024(2025)-RS-2023-00266605).

Data availability

Data will be made available on request.

References

[1] J. Miao, C. Deng, H. Zhang, Q. Miao, Interactive channel attention for rotating component fault detection with strong noise and limited data, Appl. Soft Comput. 138 (2023) 110171.

[2] J. Yang, T. Gao, G. Yan, C. Yang, G. Li, A fault location method based on ensemble complex spatio-temporal attention network for complex systems under fluctuating operating conditions, Appl. Soft Comput. (2023) 110489.

[3] Z. Zhu, Y. Lei, G. Qi, Y. Chai, N. Mazur, Y. An, X. Huang, A review of the application of deep learning in intelligent fault diagnosis of rotating machinery, Measurement 206 (2023) 112346.

[4] R. Lin, H. Wang, M. Xiong, Z. Hou, C. Che, Attention-based gate recurrent unit for remaining useful life prediction in prognostics, Appl. Soft Comput. 143 (2023) 110419.

[5] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, A.K. Nandi, Applications of machine learning to machine fault diagnosis: A review and roadmap, Mech. Syst. Signal Process. 138 (2020) 106587.

[6] B. Wang, W. Qiu, X. Hu, W. Wang, A rolling bearing fault diagnosis technique based on recurrence quantification analysis and Bayesian optimization SVM, Appl. Soft Comput. 156 (2024) 111506.

[7] S. Lee, J. Choi, Y. Son, Efficient visibility algorithm for high-frequency time-series: application to fault diagnosis with graph convolutional network, Ann. Oper. Res. (2023) 1–21.

[8] B.A. Tama, M. Vania, S. Lee, S. Lim, Recent advances in the application of deep learning for fault diagnosis of rotating machinery using vibration signals, Artif. Intell. Rev. 56 (5) (2023) 4667–4709.

[9] Q. Guo, J. Li, F. Zhou, G. Li, J. Lin, An open-set fault diagnosis framework for MMCs based on optimized temporal convolutional network, Appl. Soft Comput. 133 (2023) 109959.

[10] Y. Zhuo, Z. Ge, Data guardian: A data protection scheme for industrial monitoring systems, IEEE Trans. Ind. Informatics 18 (4) (2021) 2550–2559.

[11] Y. Zhuo, Z. Yin, Z. Ge, Attack and defense: Adversarial security of data-driven FDC systems, IEEE Trans. Ind. Informatics 19 (1) (2022) 5–19.

[12] O. Gungor, T. Rosing, B. Aksanli, Stewart: Stacking ensemble for white-box adversarial attacks towards more resilient data-driven predictive maintenance, Comput. Ind. 140 (2022) 103660.

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

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

[15] N. Papernot, P. McDaniel, X. Wu, S. Jha, A. Swami, Distillation as a defense to adversarial perturbations against deep neural networks, in: 2016 IEEE Symposium on Security and Privacy, SP, IEEE, 2016, pp. 582–597.

[16] J. Chen, D. Yan, Adversarial attacks on machinery fault diagnosis, 2021, arXiv preprint arXiv:2110.02498.

[17] S. Ntalampiras, Adversarial attacks against acoustic monitoring of industrial machines, IEEE Internet Things J. 10 (4) (2022) 2832–2839.

[18] Y. Ge, H. Wang, Z. Liu, 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, 2021, pp. 757–761.

[19] M.S. Ayas, S. Ayas, S.M. Djouadi, Projected gradient descent adversarial attack and its defense on a fault diagnosis system, in: 2022 45th International Conference on Telecommunications and Signal Processing, TSP, IEEE, 2022, pp. 36–39.

[20] Z. Zhao, T. Li, B. An, S. Wang, B. Ding, R. Yan, X. Chen, Model-driven deep unrolling: Towards interpretable deep learning against noise attacks for intelligent fault diagnosis, ISA Trans. 129 (2022) 644–662.

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

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

[23] Z. Feng, X. Yu, D. Zhang, M. Liang, Generalized adaptive mode decomposition for nonstationary signal analysis of rotating machinery: Principle and applications, Mech. Syst. Signal Process. 136 (2020) 106530.

[24] D. Zhang, Z. Feng, Enhancement of time-frequency post-processing readability for nonstationary signal analysis of rotating machinery: Principle and validation, Mech. Syst. Signal Process. 163 (2022) 108145.

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

[26] A. Widodo, B.-S. Yang, Support vector machine in machine condition monitoring and fault diagnosis, Mech. Syst. Signal Process. 21 (6) (2007) 2560–2574.

[27] L. Pan, D. Zhu, S. She, A. Song, X. Shi, S. Duan, Gear fault diagnosis method based on wavelet-packet independent component analysis and support vector machine with kernel function fusion, Adv. Mech. Eng. 10 (11) (2018) 1687814018811036.

[28] S. Zhang, S. Zhang, B. Wang, T.G. Habetler, Deep learning algorithms for bearing fault diagnostics—A comprehensive review, IEEE Access 8 (2020) 29857–29881.

[29] H. Zheng, Z. Wu, S. Duan, Y. Chen, Research on fault diagnosis method of rolling bearing based on TCN, in: 2021 12th International Conference on Mechanical and Aerospace Engineering, ICMAE, IEEE, 2021, pp. 489–493.

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

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

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

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

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

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

[36] Y. Liu, X. Chen, C. Liu, D. Song, Delving into transferable adversarial examples and black-box attacks, in: International Conference on Learning Representations, 2016.

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

[38] W. Zhou, X. Hou, Y. Chen, M. Tang, X. Huang, X. Gan, Y. Yang, Transferable adversarial perturbations, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 452–467.

[39] M. Andriushchenko, F. Croce, N. Flammarion, M. Hein, Square attack: a query-efficient black-box adversarial attack via random search, in: European Conference on Computer Vision, Springer, 2020, pp. 484–501.

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

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

[42] H. Kim, J. Park, J. Lee, Generating transferable adversarial examples for speech classification, Pattern Recognit. 137 (2023) 109286.

[43] X. Peng, H. Xian, Q. Lu, X. Lu, Semantics aware adversarial malware examples generation for black-box attacks, Appl. Soft Comput. 109 (2021) 107506.

[44] J.W. Cooley, J.W. Tukey, An algorithm for the machine calculation of complex Fourier series, Math. Comp. 19 (90) (1965) 297–301.

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

[46] N. Elsayed, A.S. Maida, M. Bayoumi, Deep gated recurrent and convolutional network hybrid model for univariate time series classification, 2018, arXiv preprint arXiv:1812.07683.

[47] K. Wang, X. He, W. Wang, X. Wang, Boosting adversarial transferability by block shuffle and rotation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 24336–24346.

[48] X. Wang, X. He, J. Wang, K. He, Admix: Enhancing the transferability of adversarial attacks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16158–16167.

[49] X. Wang, Z. Zhang, J. Zhang, Structure invariant transformation for better adversarial transferability, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4607–4619.

[50] J. Yang, C. Wang, et al., A novel Brownian correlation metric prototypical network for rotating machinery fault diagnosis with few and zero shot learners, Adv. Eng. Informatics 54 (2022) 101815.

[51] C. Wang, J. Yang, B. Zhang, A fault diagnosis method using improved prototypical network and weighting similarity-Manhattan distance with insufficient noisy data, Measurement (2024) 114171.

[52] Z. Chen, C. Xu, H. Lv, S. Liu, Y. Ji, Understanding and improving adversarial transferability of vision transformers and convolutional neural networks, Inform. Sci. 648 (2023) 119474.