Mechanical Systems and Signal Processing 208 (2024) 111006

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

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

Modeling of the hysteretic behavior of nonlinear particle damping by Fourier neural network with transfer learning

Xin Ye[a,b], Yi-Qing Ni[a,b,*], Wai Kei Ao[a,b], Lei Yuan[a,b]

[a] Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[b] National Rail Transit Electrification and Automation Engineering Technology Research Center (Hong Kong Branch), Hung Hom, Kowloon, Hong Kong

ARTICLE INFO

Keywords:
Particle damping
Neural tangent kernel
Fourier neural network
Transfer learning

ABSTRACT

The particle damper (PD) filled with granular material exhibits hysteretic behavior under dynamic excitation, meaning that its response depends not only on the current excitation but also on its excitation history. The hysteresis loops of a PD vary with the excitation frequency due to its nonlinear nature. To model the particle damping hysteresis, this study proposes using neural networks (NN), which have a powerful ability to recognize such nonlinear relationships. However, NNs suffer from a long-standing issue called spectra bias, which means they tend to learn low-frequency components first and struggle to recognize high-frequency components. This is a problem for modeling PDs, which may involve high-frequency features in the target function. To address this issue, the recently developed theory of neural tangent kernel (NTK) revealed why NNs are perplexed by the spectra bias. Based on this theory, Fourier features embedding is proposed to expedite the learning of NNs on high-frequency features to extricate NNs from the shackle of spectra bias. After implementing the Fourier features embedding, an investigation on the use of transfer learning (TL), incorporated with the physics-informed neural network (PINN), is conducted to improve the proposed model's performance. The concatenation of Fourier features embedding and TL formulates the proposed method, the Fourier features-embedded, transfer learning-incorporated physics-informed neural network (*ff*-TLPINN).

The established surrogate model of the PD's hysteretic response force under steady-state excitation covers a wide frequency range of 100–2000 Hz. The proposed model is validated using a dataset generated from the sweep-sinusoidal excitation and is shown to be more effective than a plain NN model. The study's findings demonstrate the potential of using NNs to model the hysteresis of PDs and the effectiveness of using Fourier features embedding and TL to overcome the issue of spectra bias and improve the model's performance. Overall, the proposed model provides a promising approach to accurately modeling the behavior of granular material-dilled PDs under dynamic excitation.

---

[*] Corresponding author at: Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong.

E-mail addresses: nixey.ye@connect.polyu.hk (X. Ye), ceyqni@polyu.edu.hk (Y.-Q. Ni), waikei.ao@polyu.edu.hk (W.K. Ao), lei2021.yuan@connect.polyu.hk (L. Yuan).

https://doi.org/10.1016/j.ymssp.2023.111006
Received 1 June 2023; Received in revised form 26 September 2023; Accepted 2 December 2023

1. Introduction

As a passive vibration control strategy, particle damping technology keeps various pronouncing merits for field application. The particle damping is achieved by combining collision and friction damping effects within a geometrically constrained particle bed [1]. The motion-caused in-elastic interaction, including collision and friction, inside particle dampers (PDs) transfer the vibration energy into heat. As such, this robust damping technology is not sensitive to environmental conditions and is effective in a broad frequency band. Moreover, the straightforward mechanism brings cost-effective manufacturing of this kind of damper. These attractive factors harnessing the PDs went a recent surge in the application, for instance, railway noise reduction [2–4], seismic response reduction [5,6], mining truck improvement [7], wind turbine tower and pipeline vibration control [8,9]. Although PDs are functional, the complex particle motion that occurs during vibration makes modeling their behavior a continuous challenge. A simplified model has been utilized in certain pursuits where the particle bed is replaced with a single impact ball, as described in Refs. [8,10–12]. The discrete element method (DEM) is also popular for simulating the agglomeration of dry macroscopic particles [13–17]. However, these well-developed modeling approaches require rigorous preset conditions, and scenario like high-frequency vibration remains a hurdle to these conventional attempts to comprehend particle damping effect. The PDs are susceptible to external excitation conditions, including amplitude and frequency. Granular material in the damper can behave like different states of matter, such as glass-like, liquid, or gaseous [18,19]. It means PDs exhibit highly nonlinear character, and the modeling of which is deservedly relied on experiments. Under such circumstances, machine learning (ML) methods [20], especially neural networks (NNs) [21], can be powerful for exploring the underlying nonlinear behavior of PDs hidden in experimental results.

Veeramuthuvel et al. [22] introduced a framework for modeling PDs using NNs. In their research, they attached a particle damper capsule to a printed circuit board (PCB). They developed a backpropagation NN and a radial basis function NN based on the experimental database, and the established NN models were able to accurately predict the acceleration response of the PCB under simple harmonic excitation. Similarly, Wang and Wu [23] also used back-propagation NN and incorporated genetic algorithms to determine the optimal damper design parameters for attenuating the vibration of a cantilever beam.

The challenge of adopting NNs to model PDs lies in the conflict that arises when there is a shortage of high-fidelity experimental data, making it difficult to provide the necessary resources to train the data hungry NNs. To address this issue, deep transfer learning (TL) was introduced in a previous study by the authors, which aimed to establish a multi-fidelity modeling approach [24]. The TL utilizes a trained model's knowledge to nurture the learning of a new dataset [25]. The authors proposed a multi-fidelity modeling approach that involves training a low-fidelity NN model with an approximate database of governing/constitutive equations. The general knowledge gained from the low-fidelity model is then transferred to the next training step by transferring the NN's low-level layers. In the final stage of the training process, the transferred low-level layers are frozen, and the limited high-fidelity experimental data is only used to update the few outermost layers.

Despite the previously mentioned limitations of NNs, there is growing evidence of another problematic behavior exhibited by deep fully connected networks, which is known as spectral bias [26–29]. The NNs are hindered in recognizing solution that contains high-frequency components, which has profoundly restricted the performance of NNs. Since PD modeling involves a broad frequency band. Thus, NNs require a collaborative effort to address this limitation for PD modeling. Recent studies [30–32] have attempted to alleviate spectra bias by scaling input features to approximate high-frequency components. Wang et al. [33,34] introduced an alternative approach by analyzing the fundamental weakness of NNs using the theory of neural tangent kernel (NTK) [35]. The observation indicates that NNs have a bias toward learning the dominant frequencies of the most significant eigenvectors of the NTK. As a result, researchers have proposed the use of Fourier features embedding to customize the NTK of the fully connected networks [36,37]. The Fourier NN successfully elevated the performance of NNs.

Based on the findings of these recent studies, an attempt is made in this work to model the hysteretic response force of a PD under a shaker's excitation in a wide frequency bandwidth (100–2000 Hz). Two forms of excitation, i.e., steady-state simple harmonic excitation and sweep-sinusoidal excitation, are applied to a PD that filled with granular material (tungsten powder). The response force is investigated to reveal the hysteretic behavior of PD. To model PD, the Fourier NN is examined and discussed its potential to improve training accuracy. Additionally, the effectiveness of wavelet basis features embedding [34] was evaluated in this study. These approaches formulate the Fourier/Wavelet neural network (*ff*/*wf*NN) method. Afterwards, this study presents an investigation on the *ff*NN coping with the former proposed deep TL [24].

The rest of the paper is arranged as follows: Section 2 aims to introduce the proposed methodology. Spectra bias is analyzed from the view of NTK, and the Fourier/Wavelet features embedding is applied to remedy this pathology. Section 3 describes dynamic experiments on a cylindrical PD under steady-state and sweep-sinusoidal excitations. In Section 4, the establishment of the *ff*/*wf*NN model is presented, followed by the comparison with the TL-PINN method. Section 5 shows the model reconstruction of the hysteresis loops based on the steady-state dataset and the validation on the sweep-sinusoidal dataset. Finally, Sec. 6 presents conclusions and future works.

2. Fourier/Wavelet neural network

Spectral bias is a ubiquitous pathology observed in neural network training. The gradient flow of NN becomes increasingly stiff when the objective function exhibits a high-frequency character. Such a bias is also associated with a broad input feature domain. As the input feature domain expands, the normalization of the input variables effectively turns low-frequency features into high-frequency features. This bias mentioned has an impact on the output of NNs right from their initialization stage. The derivation given in [38] proved when the parameters of a fully connected network $\boldsymbol{\theta}$ are i.i.d. sampled from Gaussian distribution $\mathcal{N}(0,1)$, the

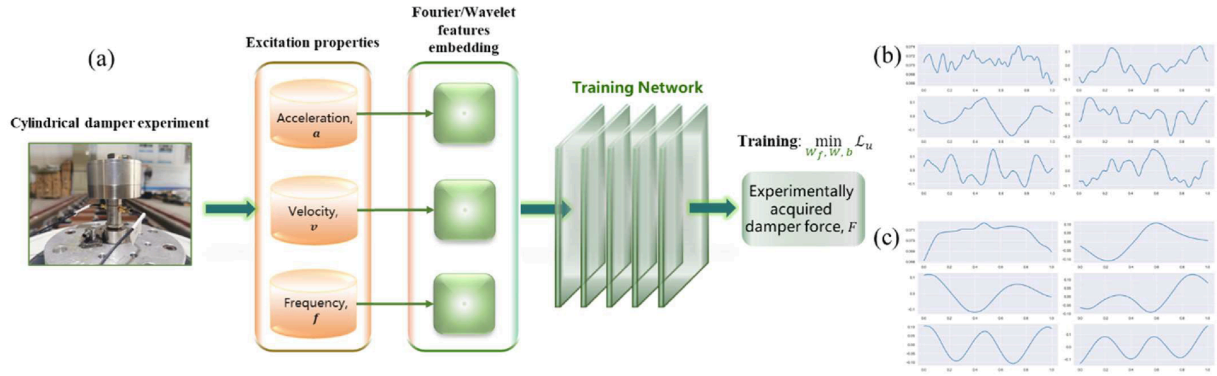

Fig. 1. (a) Flowchart of a Fourier/Wavelet network; The six leading NTK eigenvectors of a fully connected NN (5 layers, 100 units per hidden layer, *tanh* activations) (b) with embedded Fourier features that initialized by $\sigma = 10$; (c) without Fourier features embedding.

output of the initialized network is surely a flat function. This implies that an initialized NN's output function already exhibits characteristics of low frequency.

A novel insight is built on this limitation of NNs through the perspective by analyzing their NTK [37]. According to Jacot et al. [35], the continuous gradient descent flow is demonstrated with an infinitesimal learning rate η (gradient flow). Then, the gradient flow is defined by the neural tangent kernel operator $\boldsymbol{K}$ as follows:

$$\boldsymbol{K}_{ij} = \boldsymbol{K}(x_i, x_j) = \left\langle \frac{\partial f(x_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \frac{\partial f(x_j, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\rangle \tag{1}$$

where $f(\boldsymbol{x}, \boldsymbol{\theta})$ is the output of a fully connected NN that parameterized by $\boldsymbol{\theta}$, and $\boldsymbol{x} \in \mathbb{R}^d$ is the input to the NN. The training of the network is focused on the parameters $\boldsymbol{\theta}$ since $\boldsymbol{X}_{\boldsymbol{train}}$ is stable. In that case, with the asymptotic conditions [39], the dynamic of the $f(\boldsymbol{x}, \boldsymbol{\theta})$ can be derived by (see Appendix A):

$$\frac{df(\boldsymbol{X}_{\boldsymbol{train}}, \boldsymbol{\theta}(t))}{dt} \approx -\boldsymbol{K} \bullet (f(\boldsymbol{X}_{\boldsymbol{train}}, \boldsymbol{\theta}(t)) - \boldsymbol{Y}_{\boldsymbol{train}}) \tag{2}$$

It directly follows that:

$$f(\boldsymbol{X}_{\boldsymbol{train}}, \boldsymbol{\theta}(\boldsymbol{t})) \approx \left(I - e^{-\boldsymbol{K}t}\right) \bullet \boldsymbol{Y}_{\boldsymbol{train}} \tag{3}$$

where $\boldsymbol{X}_{\boldsymbol{train}} = (x_i)_{i=1}^{N}$ are inputs and $\boldsymbol{Y}_{\boldsymbol{train}} = (y_i)_{i=1}^{N}$ are the corresponding outputs, $\boldsymbol{N}$ is the dataset size. $\boldsymbol{K}$ is a semi-definite positive kernel, so $e^{-\boldsymbol{K}t}$ can be written into $\boldsymbol{Q}^{\boldsymbol{T}} e^{-\Lambda t} \boldsymbol{Q}$, where the matrix $\boldsymbol{Q}$ is orthogonal. The ith column $\boldsymbol{q}_{\boldsymbol{i}}$ of the matrix $\boldsymbol{Q}$ is related to the eigenvector of $\boldsymbol{K}$. The matrix $\boldsymbol{\Lambda}$ is diagonal, and its diagonal entries are eigenvalues λ_i of the NTK operator $\boldsymbol{K}$. Herein, Eq. (3) can be reformed as:

$$\begin{bmatrix} \boldsymbol{q}_1^{\boldsymbol{T}} \\ \boldsymbol{q}_2^{\boldsymbol{T}} \\ \vdots \\ \boldsymbol{q}_N^{\boldsymbol{T}} \end{bmatrix} (f(\boldsymbol{X}_{\boldsymbol{train}}, \boldsymbol{\theta}(\boldsymbol{t})) - \boldsymbol{Y}_{\boldsymbol{train}}) = \begin{bmatrix} e^{-\lambda_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{-\lambda_N t} \end{bmatrix} \begin{bmatrix} \boldsymbol{q}_1^{\boldsymbol{T}} \\ \boldsymbol{q}_2^{\boldsymbol{T}} \\ \vdots \\ \boldsymbol{q}_N^{\boldsymbol{T}} \end{bmatrix} \boldsymbol{Y}_{\boldsymbol{train}} \tag{4}$$

$$\Rightarrow f(\boldsymbol{X}_{\boldsymbol{train}}, \boldsymbol{\theta}(t)) - \boldsymbol{Y}_{\boldsymbol{train}} = \sum_{i=1}^{N} \left(e^{-\lambda_i t} \boldsymbol{q}_{\boldsymbol{i}}^{\boldsymbol{T}} \boldsymbol{Y}_{\boldsymbol{train}}\right) \boldsymbol{q}_{\boldsymbol{i}} \tag{5}$$

The left-hand side of this equation represents the error of the NN model. Obviously, along the NTK's eigendirections (eigenvectors) or eigenfunction of the network with a larger eigenvalue will earn an eminent convergence speed than others. However, in the case of a conventional fully connected network, the primary eigenfunctions tend to possess low frequency attributes. Meanwhile, as the frequency of these eigenfunctions increases, the corresponding eigenvalues will gradually decrease in a monotonic manner [37]. This theoretical analysis reveals the essence of the spectral bias and provides potential solutions to this underlying drawback of NNs. By manipulating the eigenspace of the NTK, it is possible to overcome the spectral bias that affects NNs.

The NTK theory suggests that when the NN width is infinitely large, the kernel $\boldsymbol{K}$ becomes deterministic [35]. Unfortunately, this condition typically does not align with the given content. Still, the analysis on the initialization of the NN to glimpse its ability to recognize high-frequency components can still be exploited. The corresponding proof can be found in Appendix B through an illustrative example utilizing a physics-informed neural network (PINN) [40].

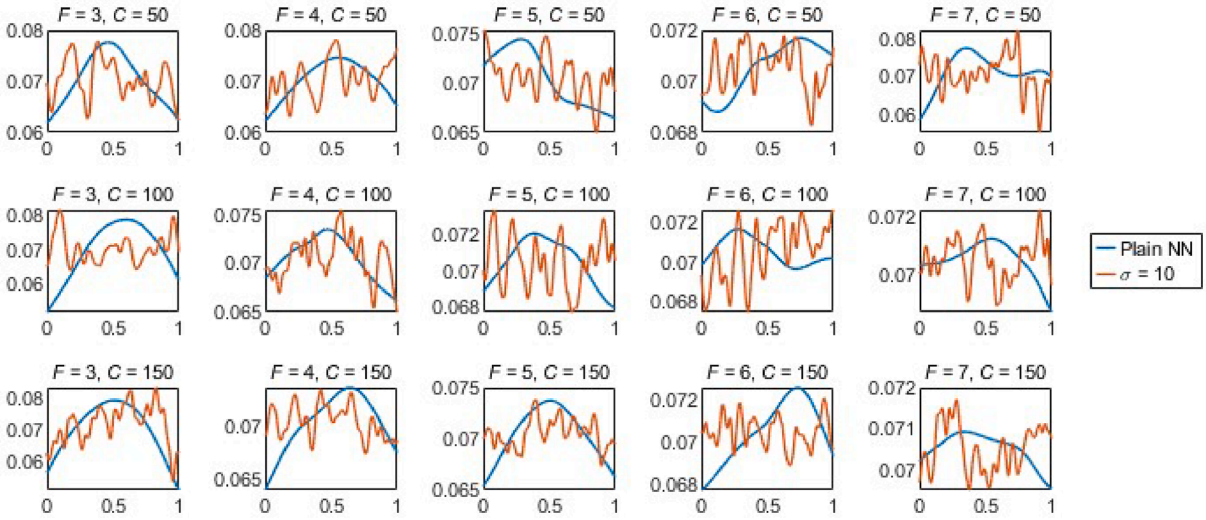

Fig. 2. The leading NTK eigenvector of each NN architecture with and without Fourier features embedding.

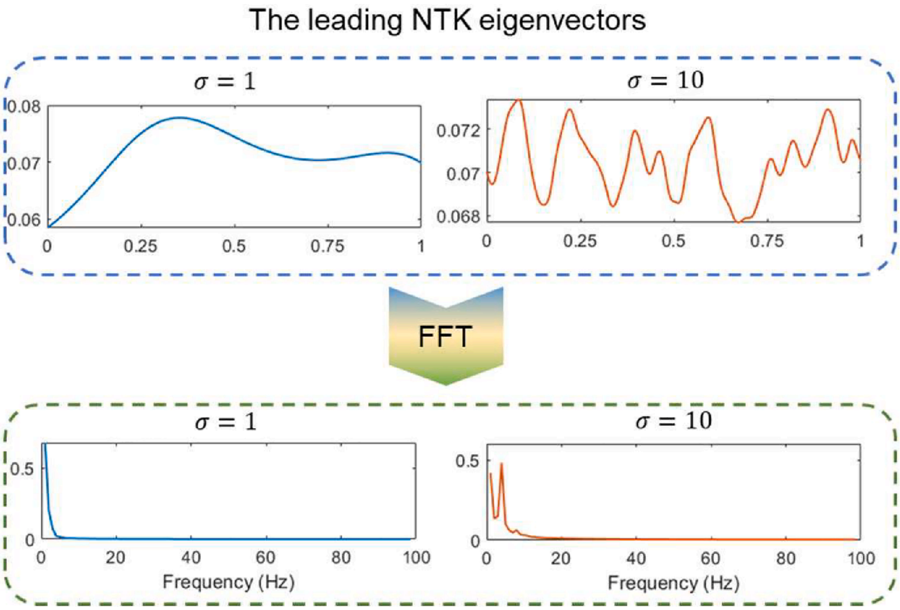

Fig. 3. Effect of different σ values on the leading eigenvector.

In this study, a Fourier NN (*ff*NN) is proposed, which incorporates a random Fourier mapping on the input features. This technique is shown to enhance the eigenspace of the NTK, resulting in superior performance when compared to a conventional NN. Further details regarding this can be found in Appendix B. Additionally, in this work, the wavelet mapping [41] will also be evaluated.

2.1. Fourier features embedding

To consider an *ff*NN with 100 hidden units per hidden layer, the random Fourier mapping is defined by $\gamma_f(x) = \begin{bmatrix} \cos(\boldsymbol{Bx}) \\ \sin(\boldsymbol{Bx}) \end{bmatrix}$, where $\boldsymbol{B} \in \mathbb{R}^{m \times d}$ is initialized by complying with the Gaussian distribution $\mathcal{N}(0,\sigma^2)$. Here, $m = 50$ is half of the layer width. The mapping γ_f in the first layer transfers the input data into Fourier feature space, which is subsequently fed into the fully connected NN (Fig. 1(a)). The introduction of the hyperparameter σ results in an increased likelihood of initializing a value $b_i \in \boldsymbol{B}$ with a higher value. Fig. 1(b) and (c) shows the efficacy of this simple method in adjusting the eigenspace of the NTK. The eigenvectors of NTK exhibit more high-frequency features when σ is larger. The frequency characteristics of the leading NTK eigenvectors (eigenvectors with the largest

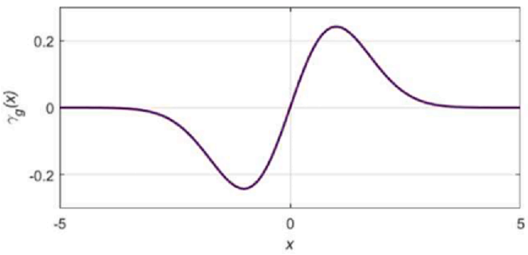

Fig. 4. Gaussian wavelet basis function ($b = 1, \tau = 0$).

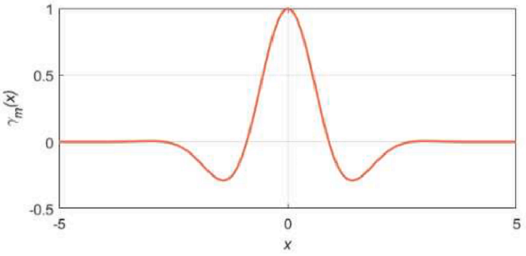

Fig. 5. Morlet wavelet basis function ($b = 1, \tau = 0$).

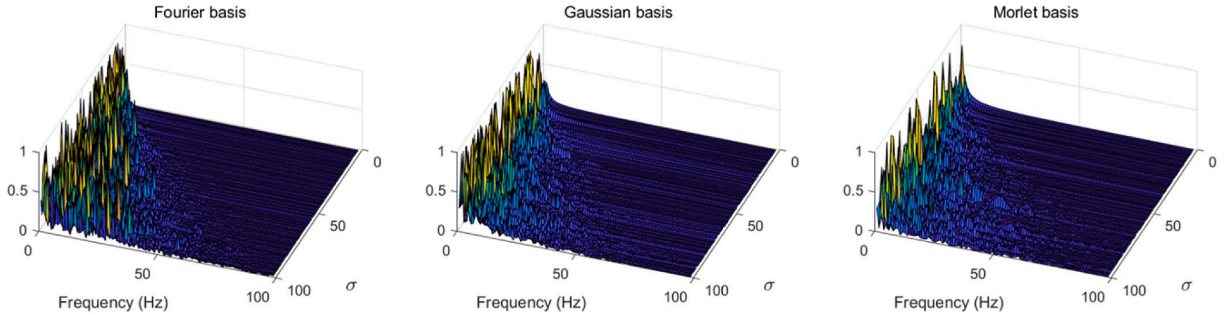

Fig. 6. Frequency analysis of the leading NTK eigenvector with the most significant eigenvalue for a NN (5 layers, 100 units per hidden layer, *tanh* activations) with (a) Fourier basis embedding; (b) Gaussian basis embedding, and (c) Morlet basis embedding that initialized by different $\sigma \in [1, 100]$.

eigenvalues) are the key indicator of the spectra bias of NN. This trick can enhance the NN's ability to recognize the high-frequency features, as it successfully modified the shape of the leading eigenvectors of a NN (Fig. 1(b) and (c)). This character is determined at the initialization stage.

In order to explore the effect of Fourier features on different NN architectures, the leading eigenvectors of different NN with the number of hidden layers $\mathscr{F}$ chosen from $\{3, 4, 5, 6, 7\}$, and the width of the hidden layer $\mathscr{C}$ chosen from $\{50, 100, 150\}$ are depicted in Fig. 2. It is observed that Fourier features can modify the eigenspace of the NN with different architectures. For the effect of the σ value, it can be seen that with a higher σ, the frequency components of the leading eigenvector also extend to higher frequency range (Fig. 3). A detailed discussion on the effect of σ value is given in Fig. 6.

2.2. Gaussian wavelet features embeddings

$$\gamma_g(x) = \frac{\boldsymbol{B}x - \boldsymbol{T}}{\sqrt{2\pi}} \bullet e^{-\frac{(\boldsymbol{B}x-\boldsymbol{T})^2}{2}}, \ \boldsymbol{B}\,\mathscr{N}\left(0, \sigma^2\right), \boldsymbol{T}\,\mathscr{N}(0, 1) \tag{6}$$

2.3. Morlet wavelet features embeddings

$$\gamma_m(x) = \cos[1.75 \bullet (\boldsymbol{B}x - \boldsymbol{T})] \bullet e^{-\frac{(\boldsymbol{B}x-\boldsymbol{T})^2}{2}}, \ \boldsymbol{B}\,\mathscr{N}\left(0, \sigma^2\right), \boldsymbol{T}\,\mathscr{N}(0, 1) \tag{7}$$

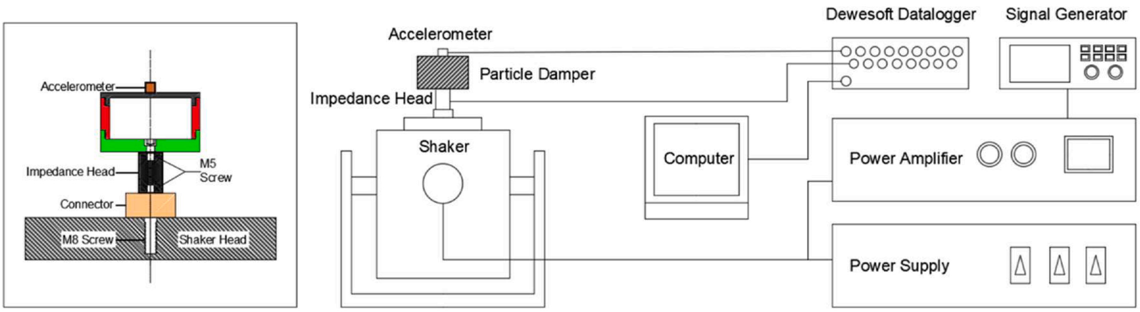

Fig. 7. Schematic diagram of experimental setup.

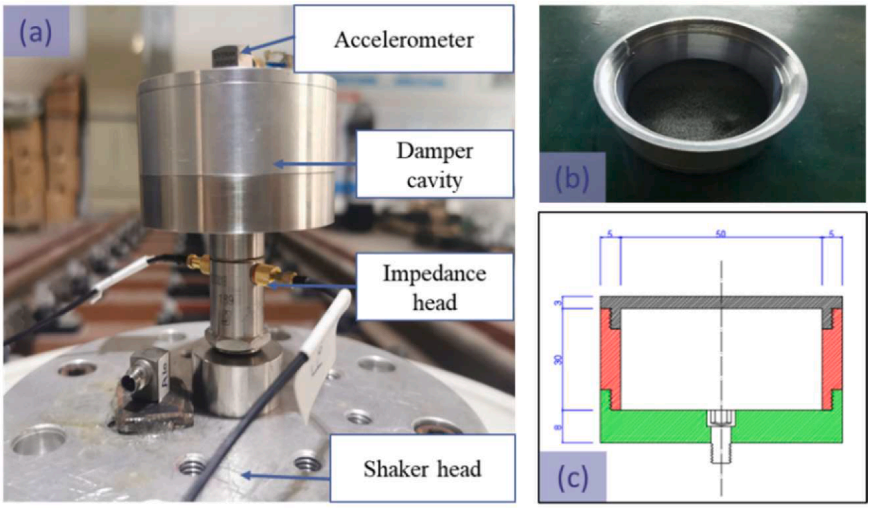

Fig. 8. (a) Setup of the cylindrical PD; (b) Tungsten powder filling in the damper cavity; (c) Schematic diagram of the damper cavity.

Similar to the mapping of Fourier features, the utilization of wavelet mapping (see Figs. 4 and 5) produces a comparable effect on the eigenspace of the NTK. For instance, a *wf*NN composed of 100 hidden units per hidden layer is considered. The continuous wavelet transform involves the Gaussian and Morlet wavelet functions [42] which serve as basis functions. These functions are scaled by a same vector $\boldsymbol{B} \in \mathbb{R}^{m\times d}$, which is sampled from a normal distribution with mean 0 and variance σ^2, $\mathscr{N}(0,\sigma^2)$. Additionally, the wavelet basis function can be shifted by a translation factor $\boldsymbol{T} \in \mathbb{R}^{m\times 1}$, where $\tau_i \in \boldsymbol{T}$ is initially sampled from $\mathscr{N}(0,1)$ and set as trainable in the *wf*NN. Here, $m = 100$ equals to the network width.

The frequency analysis of the eigenvector possessing the largest eigenvalue is depicted in Fig. 6. According to Eq. (5), this eigenvector is the most representative vector for the spectra character of NN. The frequency analysis result can roughly reflect the NN's ability to recognize high-frequency features according to different σ values. However, it should be noted that this frequency analysis does not represent the frequency range that NN can recognize. For both *ff*NN and *wf*NN, the spectrum is clearly extended with a larger σ, implying that *ff*NN and *wf*NN are excelling in accommodating high-frequency functions with the larger σ. It is noteworthy to mention that in our approach, the scaling vector $\boldsymbol{B}$ is set to be trainable, which makes the feature embedding layer a specific hidden layer of the NN. Moreover, when $\sigma = 1$, the resulting network is equivalent to a plain NN, with Fourier/Wavelet activation functions employed in the first hidden layer. The value of these hyper-parameters will significantly dictate the model's performance with different datasets. The selection of hyper-parameters will be discussed in Sec. 4, after obtaining the experiment dataset.

3. Dynamic test of a nonlinear PD

3.1. Experimental setup

At the current stage, the response-induced damping mechanism of the granular material bed remains an unsolved puzzle, especially in the high-frequency range. Therefore, a comprehensive investigation on the response of granular particle beds is required before applying PD. In this regard, a simple experiment was designed to obtain the interaction forces that the particle bed applies to the cavity during simple harmonic vibration. The recorded force signal consists of the interaction force between the particle bed and cavity, and the inertia force of the damper cavity. Since the inertia force of the damper cavity and the accelerometer is not of interest, it will be

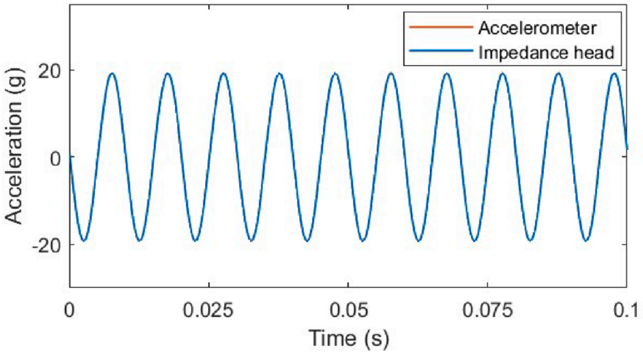

Fig. 9. Acceleration signal recorded by the accelerometer and the impedance head.

Table 1
Specifications of experimental apparatus.

Experimental apparatus	Parameter name	Parameter value
B&K LDS V-650 shaker	Sine force peak	2.2 kN
	Max acceleration sine peak	50 g
	Usable frequency range	5 Hz to 4 kHz
	Mass of moving element	4460 g
KD3001B impedance head	Max force peak	2.5 kN
	Max acceleration peak	100 g
	Working frequency range	0.5 Hz to6 kHz
	Weight	10.8 g
	Sensitivity	2 mV/N
Dytran 3273 accelerometer	Max acceleration peak	100 g
	Weight	3.0 g
Cylindrical damper cavity	Inner cavity diameter	5.0 cm
	Inner cavity height	3.0 cm
	Weight	275.4 g

deduced, and the response force mentioned afterward all refer to the interaction force from the particle bed.

In this experiment, the PD used was directly attached to the acceleration-controlled shaker (B&K LDS V-650) without being mounted onto any primary structure. The utilization of the shaker on a PD allows a straightforward examination of the damper mechanism [17]. The weight of the damper was effectively small compared to the mass of the shaker moving element. Hence, the PD was able to vibrate synchronously with the shaker head as an entity. It made the vibration of PD follow the demanded pattern. The PD was installed vertically on the impedance head (KD3001B) with the M5 screw. This sensor records the force and acceleration signal in the vertical direction. The impedance head was also bolted to a stainless-steel connector that connects to the shaker (see Fig. 7). An accelerometer (i.e., Dytran 3273) is attached to the top of the damper. This sensor records the vertical acceleration signal. This sensor's purpose was to ensure the PD vibrated in sync with the shaker head. As seen in Fig. 9, there was no relative displacement between the PD and the shaker during vibration. The sampling rate of the force and acceleration signal was 10 kHz. The Dewesoft SIRIUS data acquisition system was applied in the test to collect data, the signal-to-noise ratio of this system is higher than 130 dB.

The vibration was constrained in the vertical direction. A cylinder shapes the damper cavity to mitigate the influence of contact between the particle bed and cavity in the lateral direction. Aluminum alloy was chosen for the material of the damper cavity wall to reduce the inertia force generated by the cavity. However, the support base (marked in green in Fig. 8) was made of stainless steel to bear the high-frequency load. The filling particle is 0.2 mm tungsten powder, and the filling volume was 30% of the cavity volume, making the weight of the filled tungsten powder 128.5 g.

The detailed parameters of the experimental apparatus are given in Table 1. In this test, two types of excitations were adopted to excite the PD: steady-state simple harmonic excitation and sweep sinusoidal excitation.

3.2. Steady-state simple harmonic excitation

The steady-state simple harmonic excitation frequency ranges between 100 and 2000 Hz with an interval of 100 Hz. The excitation amplitude is controlled manually by switching the gain on the amplifier; five levels of amplitude are applied to the PD: The peak acceleration value exhibits a range of approximately 5–25 g, with intervals of 5 g. The excitation period is of 5 s for each condition, but only the middle 0.1 s of the signal starting at the π phase is selected for the investigation, with the aim of avoiding the transient effect. Each acceleration-based vibration signal contains 1000 temporal stamps as the signal sampling rate is 10 kHz.

Fig. 8 shows the acceleration and force signal that recorded at the frequency of 100 Hz and 1500 Hz, with an acceleration amplitude of approximately 20 g. It can be seen that there contains a certain degree of multi-frequency components (ultra-harmonics, e.g., 3000

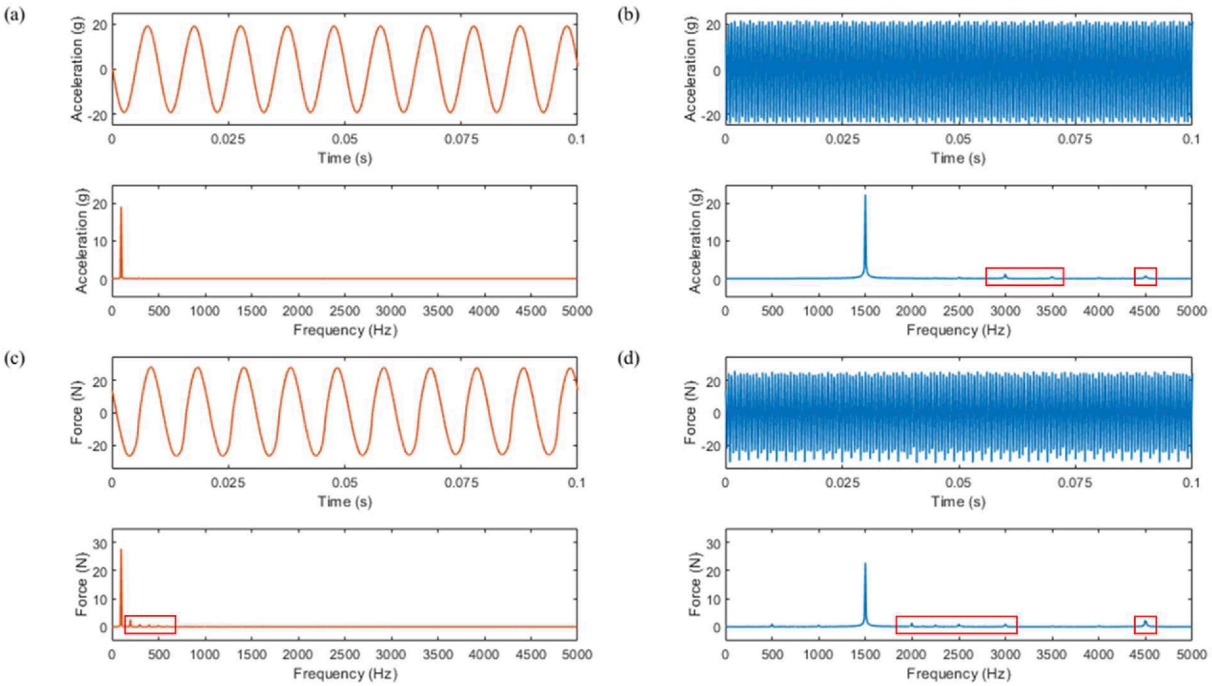

Fig. 10. Recorded signal in time and frequency domains of (a) acceleration at 100 Hz; (b) acceleration at 1500 Hz; (c) force at 100 Hz; (d) force at 1500 Hz.

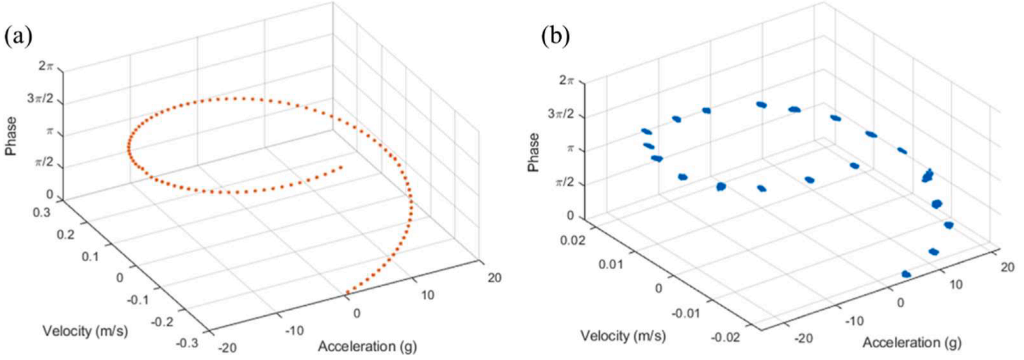

Fig. 11. Phase of the acceleration signal with respect to the value of acceleration and velocity at (a) 100 Hz; (b) 1500 Hz.

Hz and 4500 Hz) in a high frequency regime (see Fig. 10). The appearance of ultra-harmonics is due to the nonlinear nature of the particle damping, which was generated from the complex interaction force of the particle bed. The reactive force inevitably influenced the vibration of the shaker head. However, since the mass of PD is small, the influence is limited, and the vibration can still be roughly regarded as simple harmonic vibration according to the acceleration spectrum. The acceleration signal extracted for analyzation starts from the π phase, and its phase goes forward on the timeline. Because the signal is periodic, its phase is equivalent to the range of (0, 2π]. Fig. 11 illustrates the acceleration phase with respect to the value of acceleration and velocity at the frequency of 100 Hz and 1500 Hz. Similar as observed in Fig. 10. At the frequency of 1500 Hz, nonlinear effects are more significant than that at the frequency of 100 Hz. It is presented by a larger noise in the phase of the acceleration signal. This nonlinear effect-induced noise would influence the model reconstruction. For this uncertainty phase reconstruction, a Bayesian-based approach can be a solution, which will be discussed in future work. This work mainly focuses on the improvement brought by Fourier features embedding.

Fig. 12 gives the hysteresis loop of the PD under different excitation frequencies and levels, where the response force is dependent on the current excitation and the history of excitation. Relative displacement doesn’t exist in this experimental set. The velocity amplitude varies 20 times in the frequency range from 100 Hz to 2000 Hz. However, since the velocity amplitude is negligible compared to the acceleration in such high-frequency vibration. The amplitude of the response force is dependent on the equivalent inertia force of PD, which doesn’t change with frequency. It is noted that the acceleration amplitude is manually controlled and does not precisely meet the desired value in the test. Therefore, the term “excitation levels” (ELs) is chosen to represent the different

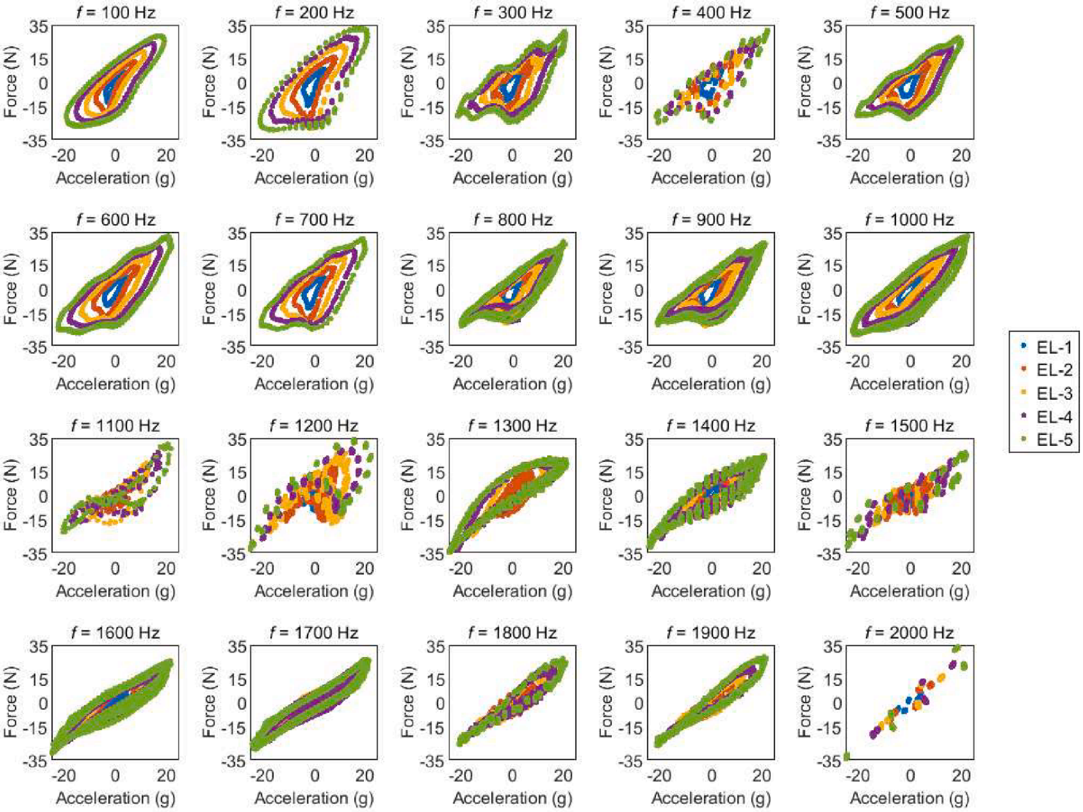

Fig. 12. Hysteresis loop (output force versus excitation acceleration) under different excitation frequencies and levels.

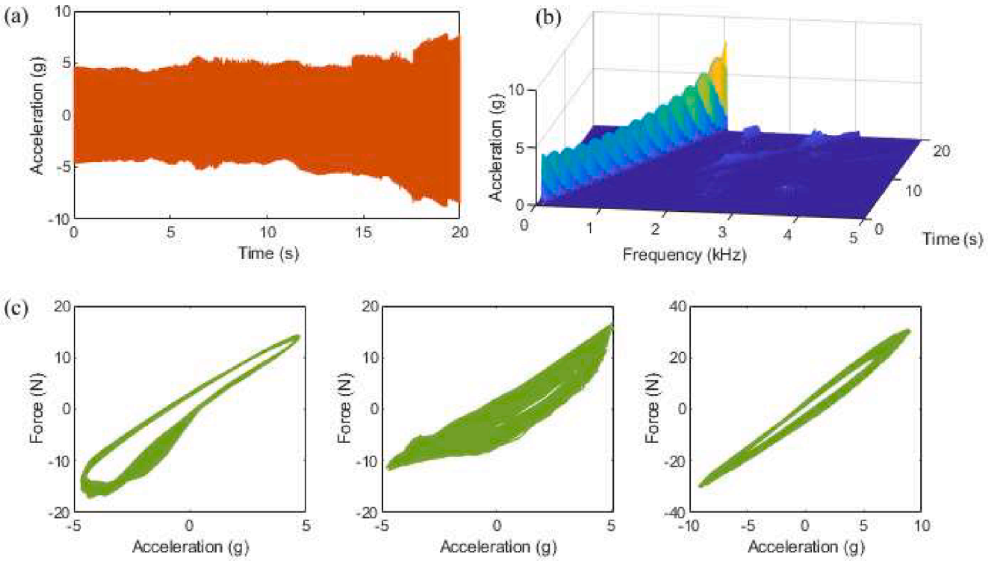

Fig. 13. (a) Sweep sine acceleration signal in time; (b) Time-frequency spectrum of the acceleration signal; (c) Diagram of force versus acceleration at frequency range of 100–150 Hz (left), 1000–1050 Hz (middle), and 1950–2000 Hz (right).

acceleration amplitudes. The nonlinear damping effect of the PD varies in the frequency range: it exhibits a viscous behavior at low frequencies, while at high frequencies, the damping effect is restrained by the hysteretic phenomenon. The underlying rationale for the reduced damping effect observed in the acceleration-controlled test scheme can be attributed to the scaling down of velocity and displacement with respect to frequency. Therefore, it can be conceptualized that the particle bed emulates a lump mass during micro-

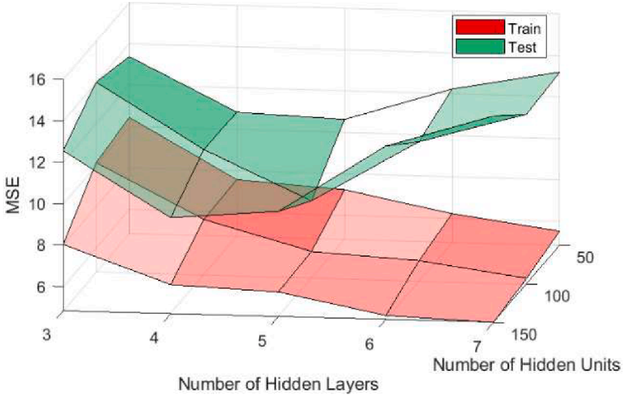

Fig. 14. Training and testing MSE of each NN architecture after 15,000 iterations.

displacement vibration, thereby contributing to the observed phenomenon of low damping effect. The shape of hysteresis loops in different frequencies varies a lot, which reflects the intricate motion of the particle bed inside the cavity. The particle bed's movement is more intense in lower frequencies with the higher displacement amplitude. Thereupon, a larger area of the hysteresis loop can be observed due to the elevated energy dissipation level by the particle bed. However, it is rather hard to dissect the shape of every hysteresis loop. Luckily, the data-driven deep learning approach offers an opportunity to avoid expressing the intricate particle damping explicitly. The application of the proposed deep learning method will be discussed in the next section.

3.3. *Sweep sinusoidal excitation*

A sweep-sinusoidal signal is also selected as an input signal of the shaker. It keeps the same frequency range as the steady-state excitation: the signal frequency linearly sweeps from 100 to 2000 Hz, the total sweeping time is 20 s, and the sweep-frequency rate is 95 Hz/s. The acceleration amplitude is not controllable during the sweeping. In the present situation, the voltage applied to the shaker, which governs the electrical signal, was maintained at a consistent level equivalent to the first level of excitation during the steady-state condition.

Nonlinear characteristics of PD can be observed through the detection of ultra-harmonic signals, which occur when the fundamental frequency surpasses 1000 Hz. However, the sweeping frequency rate is not significantly high according to the setting of 95 Hz/s. The resultant hysteresis loops in different frequency ranges are similar to those in the steady-state excitation case (see Fig. 13(c)), which are not the smooth incline ellipse shape.

4. Framework of exploiting the Fourier NN

4.1. *Selection of the NN architecture*

For selecting a network architecture, large networks (in regard to width and depth) exhibit greater robustness in terms of their regression capability. However, training such networks is inevitably more laborious due to the larger number of parameters that must be updated during each iteration. Meanwhile, expanding the network may result in only marginal improvements in performance, particularly when a smaller NN is already capable of accomplishing the given task. Moreover, a solid strategy to guide the network selection is currently absent to the best of the authors' knowledge. Therefore, the plain NNs on the steady-state dataset (total size of 1×10^5) are tested. This dataset consists of 20 sets of frequencies (100, 200, …, 2000 Hz); 5 sets of acceleration levels (approximately 5, 10, …, 25 g); each piece of the acceleration-controlled vibration data is 0.1 s long starting at the π phase with 1,000 temporal stamps. In consideration of representing the hysteretic effect, the excitation acceleration a, and its integration (i.e., velocity) v are chosen to be the input features. The purpose of inputting velocity is to provide the phase information of the signal. Calculating the phase directly and inputting it into the NN is unsuitable. Since the vibration in the test is not purely simple harmonic according to Fig. 10. The additional input, excitation frequency f, is intended to reflect the nonlinearity of PD in the frequency domain (see Fig. 1(a)). The network outputs the force response of the damper. The data corresponding to the third-level excitation (EL-3, 20% of the steady-state dataset) is extracted for testing the model. Training is conducted on the rest excitation levels (EL-1, EL-2, EL-4, EL-5, 80% of the steady-state dataset). The entire dataset is preprocessed by the min–max normalization [43].

Various architectures with different widths and depths are empirically selected from the candidate set. The number of hidden layers $\mathscr{T}$ is chosen from $\{3, 4, 5, 6, 7\}$, and the width of the hidden layer $\mathscr{C}$ is chosen from $\{50, 100, 150\}$. In this study, all hidden layers in one NN architecture are set to be the identical width. The *tanh* activation function is applied for these NNs. The training and testing mean squared error (MSE) after 15,000 iterations for each NN architecture is given in Fig. 14. The training of NN intuitively receives a reduced loss with a larger NN size. However, this merit also comes along with over-fitting, as demonstrated by the increasing testing loss after meeting the inflection point. The selection of the NN architecture should consider these two aspects simultaneously. By

Table 2
Average MSE on the entire dataset of each NN architecture after 15,000 iterations.

	$\mathscr{C} = 50$	$\mathscr{C} = 100$	$\mathscr{C} = 150$
$\mathscr{T} = 3$	11.02	10.90	8.93
$\mathscr{T} = 4$	8.23	8.19	6.88
$\mathscr{T} = 5$	7.91	**6.60**	6.80
$\mathscr{T} = 6$	7.42	6.96	6.67
$\mathscr{T} = 7$	7.05	6.69	6.83

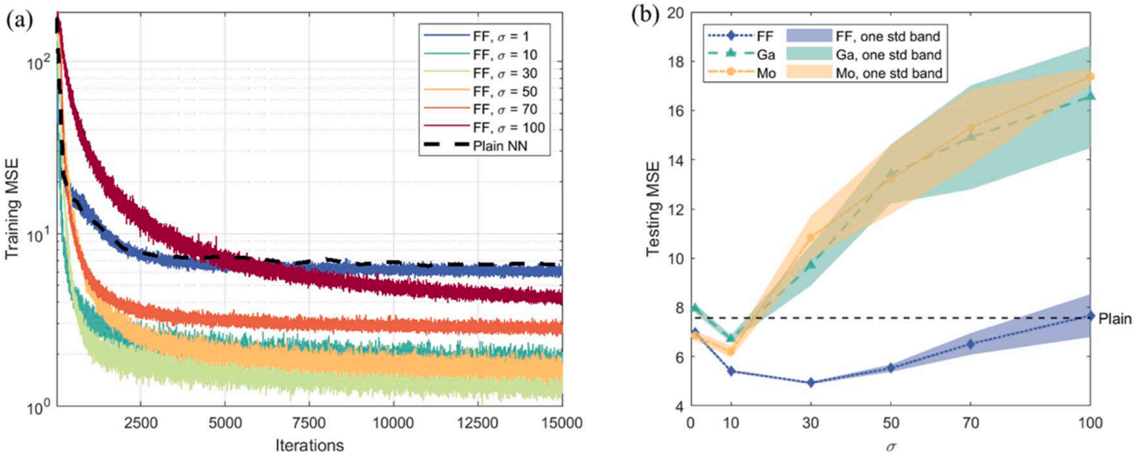

Fig. 15. (a) Evolution of the training MSE from one trial test of a NN embedded with Fourier features initialized by the different $\sigma \in [1, 100]$ and a plain network; (b) Testing MSE with the three feature functions by the different $\sigma \in [1, 100]$ after 15,000 iterations of gradient descent over 10 independent trials. The black dash line represents the final testing loss of the plain network after 15,000 iterations.

comparing the average MSE of different NN architectures on the entire dataset (see Table 2), it can be identified that NN with $\{\mathscr{T} = 5, \mathscr{C} = 100\}$ obtains the best model accuracy. Therefore, the network architecture of 5 hidden layers with 100 neurons in this study is chosen. This selection is based on the plain NNs. In addition, for a fair comparison, all methodologies will adopt the selected architecture afterwards, and the training and testing loss are presented in MSE.

4.2. Validation of the optimum features embedding strategy

As mentioned in Table 2, the NN architecture ($\mathscr{T} = 5, \mathscr{C} = 100$) is adopted in this section to explore the optimum features embedding strategy. The search is conducted on the dataset of steady-state excitation test (5 excitation levels, 20 excitation frequencies, and 1,000 temporal stamps for each set), in which the EL-3 dataset is assigned as the testing dataset while the rest is for training. Ten independent trial tests are conducted on the NNs embedded with Fourier (FF), Gaussian (Ga), and Morlet (Mo) features that are initialized by different $\sigma \in [1, 100]$.

The evolution of the training MSE from one trial test with various Fourier features embedding shown in Fig. 15(a) reveals that with a proper σ, the *ff*NN is faster in convergence and consequently achieves a better accuracy after a certain number of iterations. Testing results in Fig. 15(b) show that Fourier features embedding is more robust (thinner one standard band) than wavelet features embedding (i.e., Ga and Mo). The latter one even shows worse performance than the plain network in most cases. A possible explanation for this phenomenon is that when the wavelet basis function is scaled with a large-value factor $b_i \in \boldsymbol{B}$, the function becomes a "thin" basis containing a wide range of zero values in its domain. As a result, the function in question exhibits limited capacity to accommodate the objective function effectively. Another flaw of the *wf*NN is that, unlike the matrix operation of the scaling factor $\boldsymbol{B}$, the translation factor $\boldsymbol{T}$ can only be applied to each neuron in the embedding layer. It means that different dimensions of input $\boldsymbol{x}$ must share the same translation factor $\tau_i \in \boldsymbol{T}$ on one neuron, which also restricts the performance of the *wf*NN.

In alignment with the preceding discussion, it can be observed that for the case where $\sigma = 1$, the *ff*/*wf*NN is virtually a variant of plain networks, with the activation functions of the first hidden layer being the Fourier/Wavelet features mapping. Demonstrating the validity of this assertion, the training and testing results of the *ff*/*wf*NNs are in the vicinity of the plain NN. Among the various experimental trials conducted, the Fourier features initialized with $\sigma = 30$ yield the best performance.

4.3. Incorporating ffNN with TL

4.3.1. Framework of TL

The knowledge acquired from the source task may be leveraged to enhance performance on the target task through TL. Previous

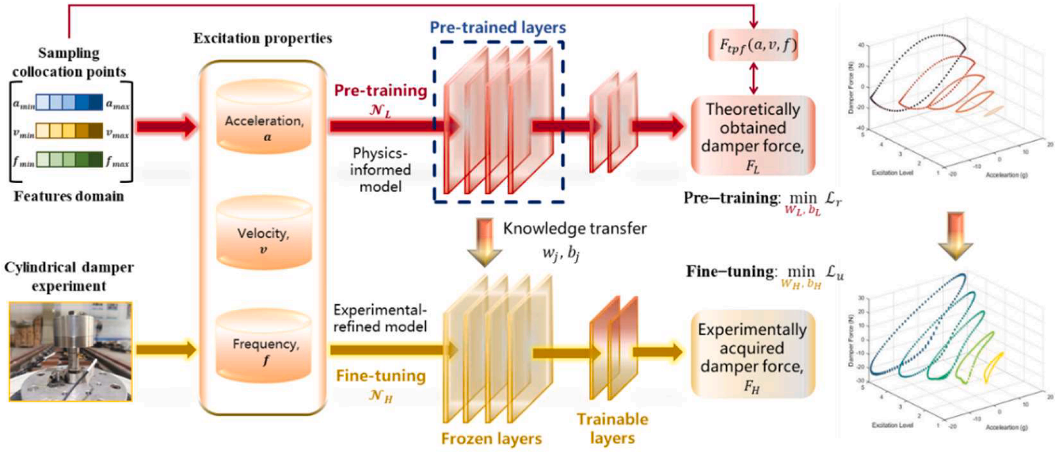

Fig. 16. Flowchart of the TL-incorporated PINN modeling approach for the PD.

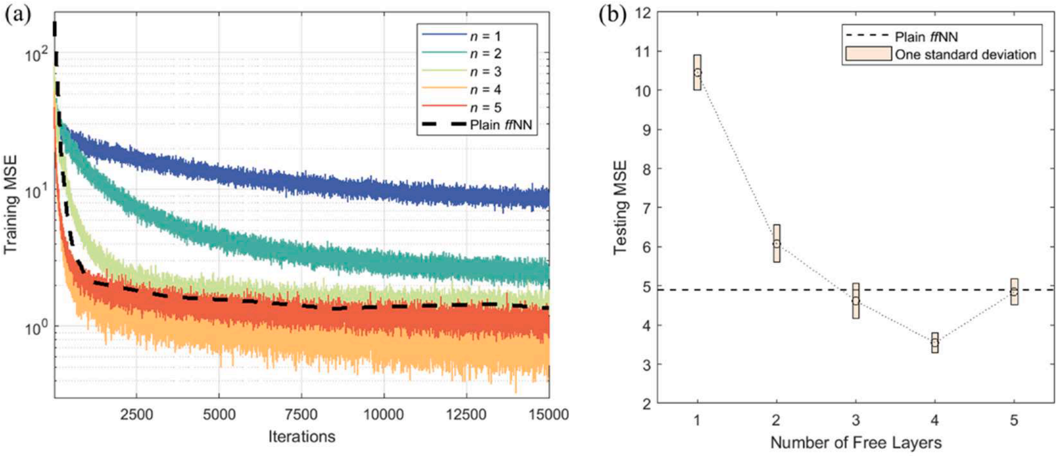

Fig. 17. (a) Evaluationolution of the training MSE from one trial test with the different free-layer number $n \in [1,5]$ and the plain *ff*NN; (b) Testing MSE with the different free-layer number $n \in [1,5]$ after 15,000 iterations of gradient descent over 10 independent trials. The black dash line represents the final testing loss of the plain *ff*NN after 15,000 iterations.

works [44,45] have revealed that the layers at the low level of a pre-trained NN model learn features with generality. The learned knowledge resides in the weights/biases of the hidden layers is transferrable across tasks. After transferring the layers from the pre-trained NN, the layers at the low levels are frozen. In the context of the target task, only fine-tuning the high-level layers are needed. Analogous to the framework introduced by the authors in a prior study [24], a low fidelity *ff*NN model is pre-trained with the approximate governing equation of PD. Subsequently, this model is refined using the high-fidelity dataset obtained from the experiment observations (see Fig. 12). Specifically, the training of the source task utilizing the approximate governing equation is carried out through the physics-informed neural networks (PINNs) [40], which obviate the need for generating a database through the approximate governing/constitutive equations. In contrast to conventional numerical methods, PINNs produce a solution across an entire domain without requiring a pre-specified grid of points. The methodology of utilizing PINNs to address differential equations is presented upon in Appendix B.

Following the discussion presented in the prior study [21], the approximate constitutive equation for two-phase flow equivalent viscosity model is proposed by Wu and Lei [46,47]. This equivalent model provides low-fidelity knowledge of the PD nature within the scope discussed in this study (high-frequency vibration). According to the experimental setup, the rigid connection between the PD and the shaker eliminates the relative displacement term in the dynamic force equation. Therefore, the force response of the PD is computed based on the two-phase flow model, which is given by:

$$F_{tpf} = ma + c_{eq}v \tag{8}$$

where m is the total mass of the damper, a and v are vibration acceleration and velocity, respectively, c_{eq} is the total equivalent viscous damping coefficient which relates to the damper response (velocity and frequency), it is written as:

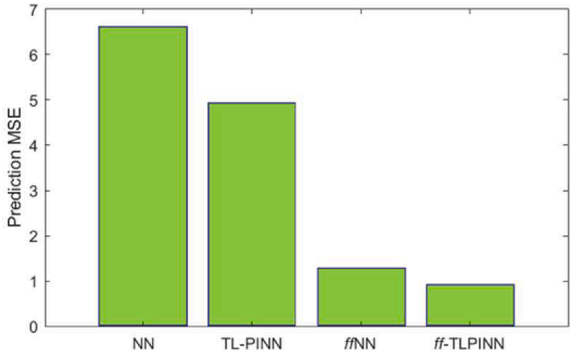

Fig. 18. Prediction MSE for different approaches with the same NN architecture.

$$c_{eq} = c_c + c_f \tag{9}$$

where c_c represents the equivalent viscosity relates to the inter-granular collision, and c_f is related to the inter-granular friction. These terms are calculated according to the two-phase flow theory [46,47]. In this PINN, the features domain is maintained identical to that of the experimental setup, i.e., acceleration ranges from 5 to 25 g, and frequency ranges from 100 to 2000 Hz. Consequently, the features domain is kept consistently across both the source and target tasks, making this study a prototypical example of inductive TL [48]. The domain adaption across tasks is no requirement. The PINN outputs the theoretical damper force F_{tpf} based on the two-phase flow model. Additionally, model testing is unnecessary for the source training task (see Fig. 16).

4.3.2. *Effect of the free layers number*

The selected *ff*NN architecture (5 layers, 100 units per hidden layer, *tanh* activations, $\sigma = 30$) is adopted in this subsection. Remarkably, the number of frozen layers transferred from the source task is a crucial factor for the performance of the target task. A reduced number of free layers in the target model can use the low-fidelity knowledge to a greater extent and expedite the training. Conversely, this approach impedes the learning capability of the high-fidelity dataset.

This section aims to explore the influence of the free-layer number and confirm the optimum number of free layers. The search is conducted on the same dataset from the steady-state excitation test (EL-3 for testing and the rest for training). The performance of the different free-layer number $n \in [1, 5]$ over 10 independent trials is tested. For the purpose of comparison, the mean squared error (MSE) evaluation of the plain *ff*NN is also provided.

The evolution history of the training MSE from one trial test in Fig. 17(a) shows that the transferred layers from the source task facilitate learning in the target task with a decreased initial error. Nevertheless, in the case of a complex and highly nonlinear issue such as PD, a reduction in the number of trainable layers leads to a decline in the accuracy of both training and testing using the high-fidelity dataset (see Fig. 17), even though the increased efficiency of training as fewer parameters are needed. The attractive advantage of TL is its potential in training data hungry NNs when high-fidelity data is limited [24]. It is noted that the present size of the steady-state dataset amounts to 1×10^5, which is abundant for NN training purpose. An intuitive explanation is that the plain *ff*NN reaches almost the identical MSE where $n = 5$, indicating that the plain *ff*NN is well-driven, eventually converging to the performance level of the TL approach, even though its higher initial MSE. It is reasonable to anticipate that in the case of a smaller training dataset, the TL approach would be more appealing. A further examination of the impact of training dataset size is presented in Appendix C.

In addition, the TL provides a slightly enhanced performance where $n = 4$. Therefore, despite the fact that the improvement is not as significant as that achieved through proper Fourier features embedding, the *ff*NN with TL-PINN (*ff*-TLPINN) is chosen to incorporate as it ultimately yields superior results.

4.4. *Comparison between the surrogate models*

After the discussion on *ff*NN and TL, the optimum strategy is achieved by the fusion of *ff*NN and TL-PINN (*ff*-TLPINN), in which the first layer is the Fourier-features mapping and followed by the TL-PINN approach. To be acquainted with these methods, all results from the different methods are compared in this subsection, including the NN, *ff*NN, and the *ff*-TLPINN (Fourier features embedding with $\sigma = 30$, transferring layers with the free-layer number $n = 4$). Furthermore, to reveal the effect of the TL-PINN approach on the plain NN, a similar discussion as in Section 4.3.2 is conducted. Finally, the optimum result of pure TL-PINN is also obtained with $n = 4$. Consequently, the results are presented for the sake of comparison. All surrogate models are trained using the steady-state dataset for 15,000 iterations, employing the same training/testing dataset and network architecture. The model evaluation is then conducted on the entire dataset.

Fig. 18 shows the model prediction of the MSE on the steady-state dataset of various methods. It can be seen that the TL-PINN (MSE 4.92) enhanced the model accuracy regarding a plain NN (MSE 6.60), but a more remarkable result is presented by the properly initialized *ff*NN (MSE 1.27). Furthermore, blending the TL approach with the *ff*NN brings a slight improvement in the model accuracy (MSE 0.90) regarding the plain *ff*NN, indicating the most critical challenge for this task is the spectral bias of the NN. As a result, the

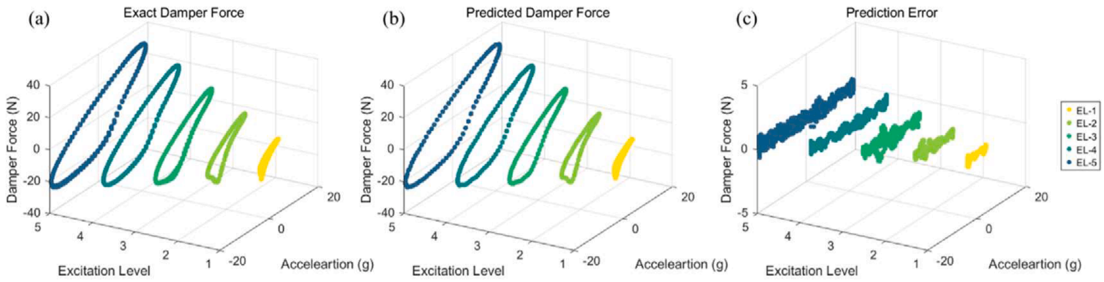

Fig. 19. Hysteresis loop of 100 Hz: (a) Exact damper response force recorded in the experiment; (b) The predicted damper force response from the proposed model; (c) The prediction error of the surrogate model.

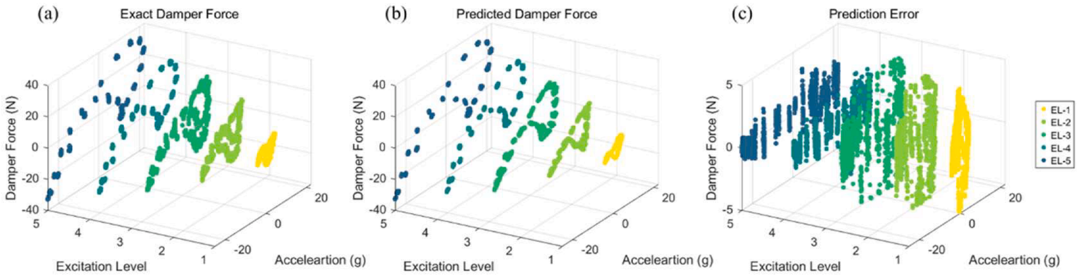

Fig. 20. Hysteresis loop of 1200 Hz: (a) Exact damper response force recorded in the experiment; (b) The predicted damper force response from the proposed model; (c) The prediction error of the surrogate model.

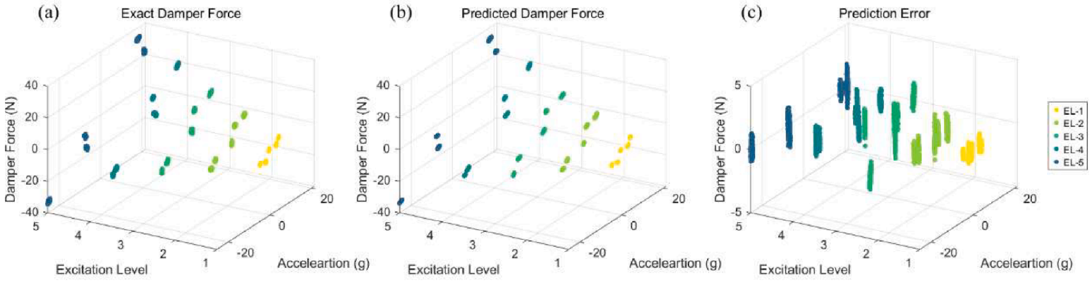

Fig. 21. Hysteresis loop of 2000 Hz: (a) Exact damper response force recorded in the experiment; (b) The predicted damper force response from the proposed model; (c) The prediction error of the surrogate model.

best accuracy is given by *ff*-TLPINN. This method is selected to reconstruct the hysteretic effect of PD.

5. Model reconstruction for the PD force response

The damper force response is reconstructed by implementing the proposed Fourier features-embedded, transfer learning-incorporated PINN (*ff*-TLPINN). An *ff*NN (5 layers, 100 units per hidden layer, *tanh* activations) with the Fourier feature embedding $\gamma_f(X)$ initialized by $\sigma = 30$ was firstly trained according to the low-fidelity physics model, i.e., the two-phase flow equivalent viscosity model. Then, the pre-trained layers were leveraged to the second stage of training, where the low-level layers were frozen. The number of free layers to be updated by the high-fidelity experimental data is 4. The input vector $\boldsymbol{X}$ of the steady-state excitation case consists of excitation frequency $\boldsymbol{f}$ (20 settings in total), excitation acceleration $\boldsymbol{a}$, and the corresponding velocity $\boldsymbol{v}$ (5 levels in total). Each set has 1000 temporal stamps. After training with the steady-state dataset for 15,000 iterations on the designated training dataset (excluding EL-3), the sweep-sinusoidal dataset is used to validate the established surrogate model.

5.1. Reconstruction of the steady-state excitation case

The proposed model manages to reconstruct the hysteresis loop of the PD. The hysteresis loop at 100 Hz shows a typical viscous effect (see Fig. 19), while the shape of the hysteresis loop at 1200 Hz (see Fig. 20) is significantly different from other frequencies due to the nonlinearity nature of the PD. The most considerable prediction error is found in this frequency case. As observed in Fig. 21, the

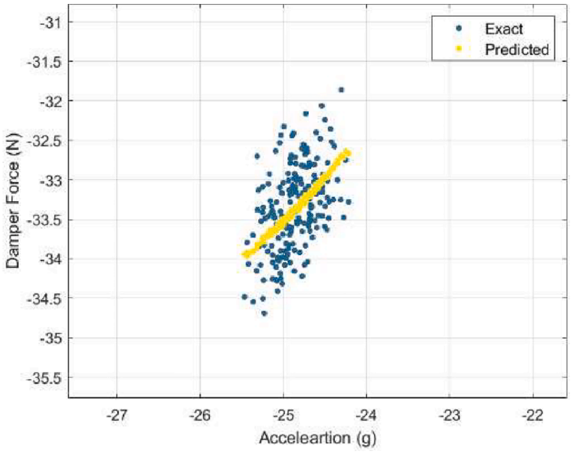

Fig. 22. Exact and predicted damper force under fifth excitation level at 2000 Hz (phase $-\pi$).

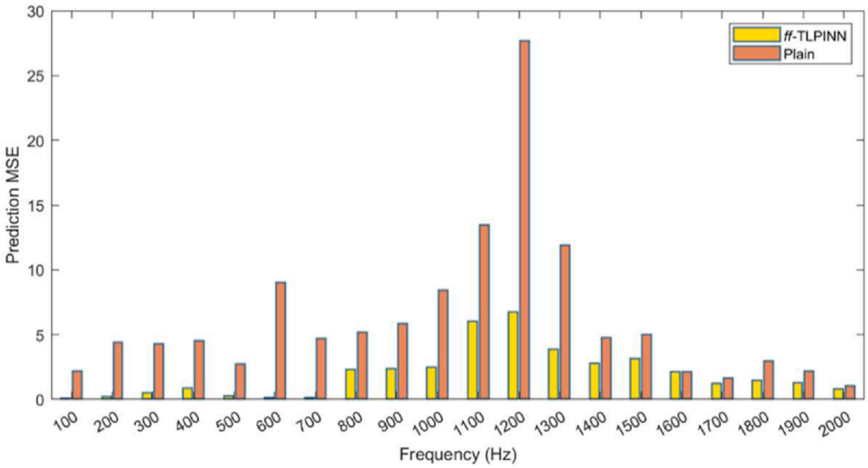

Fig. 23. Prediction MSE of the proposed model and the plain NN model with respect to each frequency.

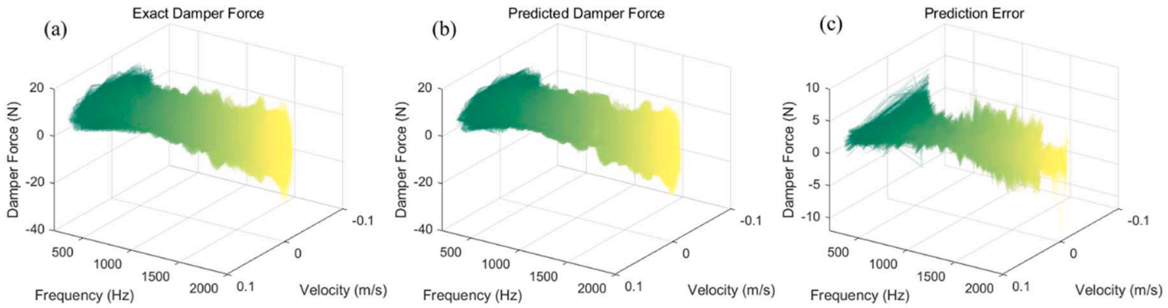

Fig. 24. (a) Exact damper force response recorded in the sweep-sinusoidal experiment. (b) The predicted damper force response from the proposed model. (c) The prediction error of the surrogate model.

hysteretic phenomenon is significantly restrained at 2000 Hz. Since the excitation is acceleration-controlled, the PD actually undergoes a micro-displacement vibration at this frequency. Most of the force response is contributed by the inertia force of the damper. In this case, since the sampling frequency (10 kHz) happens to be five times the primary frequency (2 kHz), there can only be a sample of five points in one loop. Although the dataset contains an enormous number of temporal stamps, there are a lot of duplicated data since the input features are periodic signals. The error resides in these reduplicative cases and results in the prediction error. For instance, the force response of PD under the fifth level excitation at 2000 Hz (phase $-\pi$) is shown in Fig. 22. The experimental error is the systematic

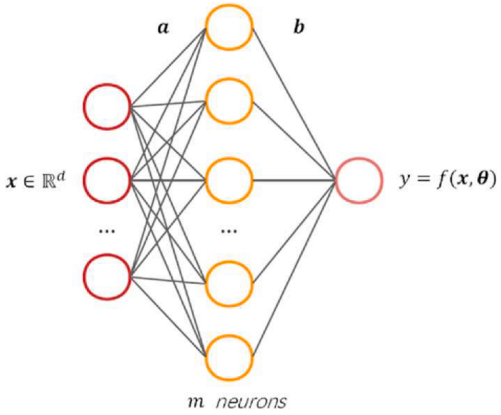

Fig. A1. One-hidden-layer NN with *m* hidden neurons.

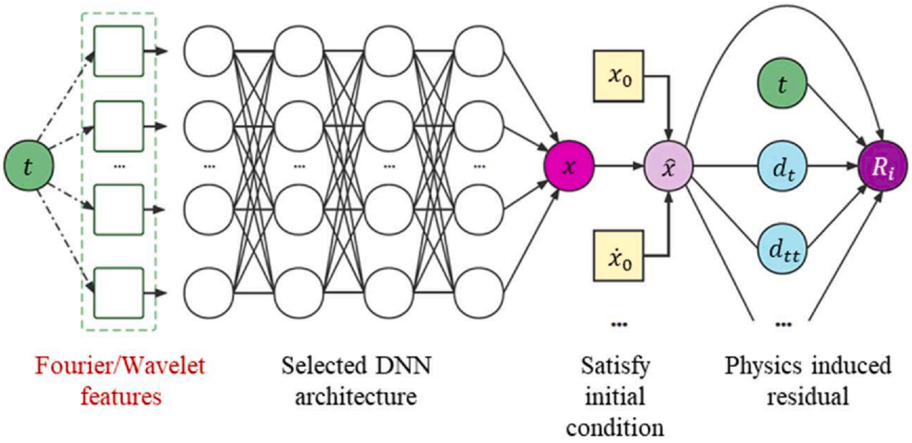

Fig. B1. Structure of Fourier/Wavelet features embedded PINN.

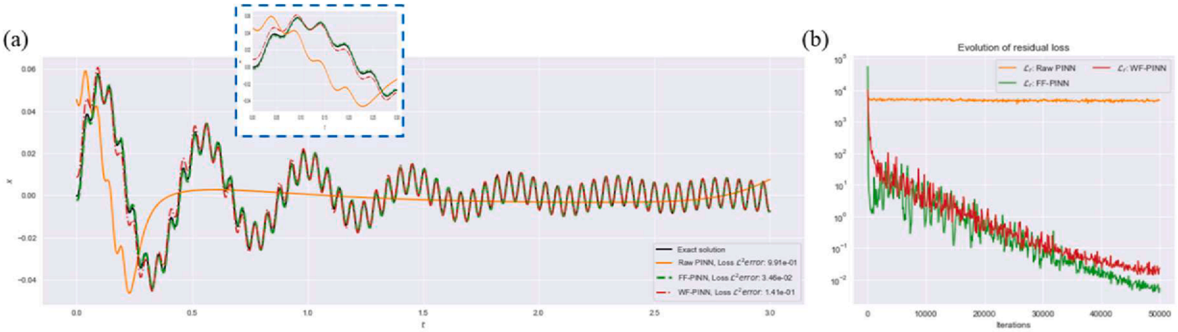

Fig. B2. (a) Calculation results of different PINNs after 50,000 iterations. (b) Evolution of the residual loss of different PINNs.

error in terms of the model training.

The above prediction error is related to the shape of the hysteresis loop and the experimental error. Fig. 23 presents the MSE prediction on the entire dataset of the proposed surrogate model and a plain NN model for each frequency. The error is high for the proposed model in the 800–1900 Hz range. The experimental result (see Fig. 12) shows that the experimental error is relatively high in this range (thick trajectory). However, the signal is sampled five times per period at the frequency of 2000 Hz, reducing the difficulty in regression. Both models find the most considerable error at 1200 Hz, reflecting the identical influence of the unequable hysteresis loop shape.

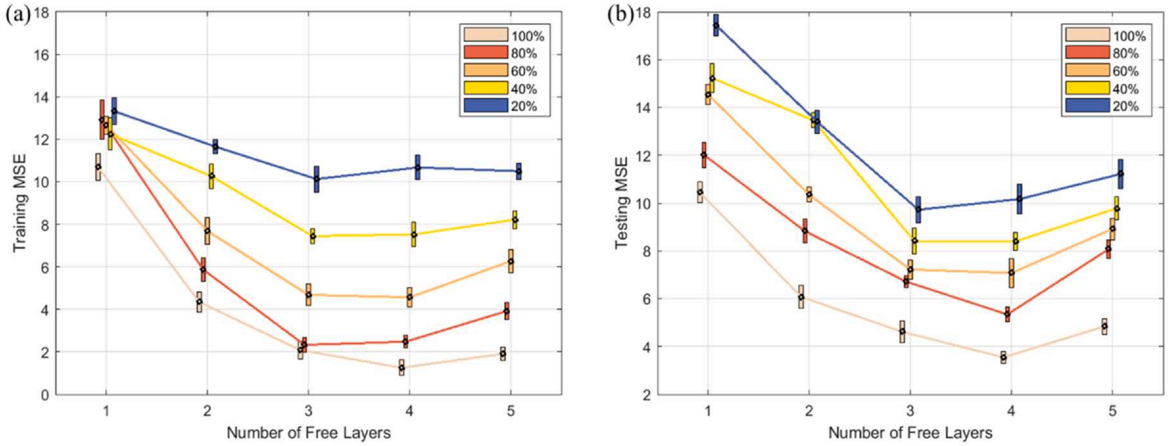

Fig. C1. Efficacy of free-layer numbers on various dataset sizes on (a) model training and (b) model testing. Evaluated with mean MSE and one standard deviation.

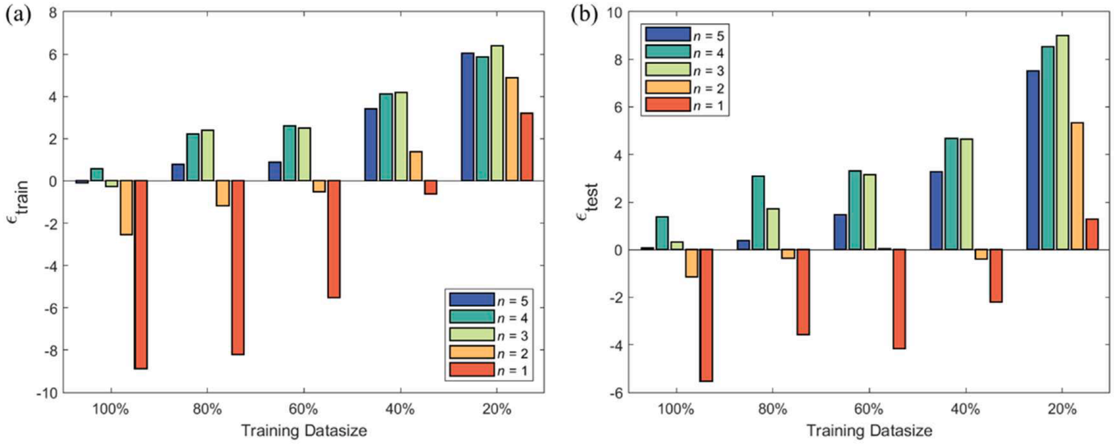

Fig. C2. Comparison between TL and non-TL approaches on various dataset sizes on (a) model training and (b) model testing.

5.2. Validation of the model with sweep-sinusoidal dataset

Section 3.3 mentioned that the hysteretic effect of PD under a slow sweep-sinusoidal excitation is similar to the steady-state case. Therefore, the proposed surrogate model can be validated by the sweep-sinusoidal dataset. After establishing the surrogate model based on the steady-state dataset, the excitation signal from the sweep-sinusoidal test is input to the surrogate model to predict the corresponding force response. The result presented in Fig. 24 is expressed by force response in terms of the excitation velocity in place of acceleration in order to have better visualization. However, the test is still acceleration controlled, as same as the steady-state test. In general, the model output captures the feature of the damper force. However, an enlarged prediction error can be observed compared to the steady-state case in which the model is trained. Specifically, the MSE prediction on the two datasets is 4.83 for the sweep-sinusoidal dataset and 0.90 for the steady-state dataset, respectively.

In this case, the prediction error in the high-frequency range is more significant with respect to a lower frequency. The error could come from the transient effect at this range when the damper response tends to be more complex in high-frequency vibration. The hysteresis loop recorded in Fig. 13 supports this conjecture, where the shape of the hysteresis loop shows a thick trajectory, indicating the instability from the transient response of the PD.

6. Conclusions and future works

The framework exploiting NN to establish the surrogate model for characterizing a particle damper (PD) was investigated in this paper. After in-depth exploration of the long-standing issue of spectra bias in neural networks (NNs), the proposed Fourier neural network (*ff*NN) is equipped with the capability to acquire high-frequency features. This simple technique is concatenated with the TL approach, which integrated PINN. The findings indicate that the efficacy of TL-PINN is limited when dealing with a task that has an abundant amount of high-fidelity experimental data. The conclusions can be summarized as follows:

(1) Compared to the wavelet features, the Fourier features embedding performs better in this case. Besides, the hyperparameter σ should keep in line with the frequency feature in the objective function. The proper imposition of this method can address the pathology of spectra bias of NNs.
(2) The number of free layers released in the second training step in TL-PINN influences the result. For a case with abundant high-fidelity learning resources, fewer free layers can speed up the training but hinder the accommodating ability, which worsens the performance of NNs. Therefore, more free layers are preferred in such situations.
(3) The concatenated method *ff*-TLPINN is reliable for reconstructing the hysteretic effect of PDs in a broad frequency band. The MSE arising from the *ff*-TLPINN model (0.90) is less than one-seventh of the plain network (6.60) with the same architecture. Correspondingly, the MSE is 4.92 for TL-PINN and 1.27 for *ff*NN.

This study investigates a PD with 30% of tungsten powder under various excitations. However, the performance of a PD is related to plentiful parameters, e.g., particle material, particle diameter, filling ratio, size, and shape of the damper cavity, etc. Therefore, a surrogate model considering all these parameters should be presented to design an optimum PD in noise/vibration control. However, merging all these factors into the input feature space will exponentially increase the size of the input dataset. Such a task will be too bulky for NNs as the input size on the temporal space is enormous (1000 stamps). Confronting the issue of dimension exploration, a redesign of the input–output scheme, which excludes the temporal space, is required. Moreover, the experimental error observed in Fig. 22 is another burden for the identification by the proposed approach. The experimental error should also be considered in NN training.

The future improvement of this study can be pursued by utilizing the phase of the periodic signal in place of the temporal stamps as the input-related factor. This approach enables the response obtained at the same phase to be viewed as a distribution (see Fig. 22). The experimental error can thus be integrated into model learning. This idea is naturally oriented to the Bayesian theory-supported neural networks, where probability is imported to regulate the parameters in a network that differs from a point-estimate network. In another line of work, a large class of hysteresis loops are described by fractional derivatives. PINN can pursue the equation discovery of particle damping in high-frequency range based on fractional derivation equations. The Fourier features embedding can also be investigated in facilitating that attempt. For particle damping, there is still much more to be explored in the intricate motion of the particle bed, especially in this high-frequency and high-nonlinear stage. In the next step, a direct study of the particle bed's exact movement is planned.

CRediT authorship contribution statement

Xin Ye: Conceptualization, Methodology, Formal analysis, Software, Writing – original draft. **Yi-Qing Ni:** Conceptualization, Supervision, Resources, Project administration, Funding acquisition, Writing – review & editing. **Wai Kei Ao:** Software, Validation, Writing – review & editing. **Lei Yuan:** Validation.

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.

Data availability

Data will be made available on request.

Acknowledgements

The authors would like to appreciate the funding support by the Innovation and Technology Commission (ITC) of Hong Kong SAR Government to the Hong Kong Branch of Chinese National Rail Transit Electrification and Automation Engineering Technology Research Center (Grant No. K-BBY1 and ITS/096/21), the authors would also like to thank Dr. Masoud Sajjadi for providing the prototype design of the particle damper.

Appendix A. Derivation of the training dynamics of NNs

For illustration, one-hidden-layer NN with m hidden neurons are considered (see Fig. A1), biases are included in the input vector a and output vector b. The single output of the NN is $y = f(\boldsymbol{x}, \boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$.

From this structure, the output of the network can be expressed as:

$$y = \sum_{i=1}^{m} b_i \delta\left(a_i^T x\right) \tag{A1}$$

where $\delta(\bullet)$ is the activation function. Training the network is to adjust the output to meet the labeled data $\hat{\boldsymbol{y}} \in \mathbb{R}^n$. Normally, the training is controlled by the L2 norm loss function, which can be flexibly written in (auxiliary coefficient $\frac{1}{2}$ is implemented to comfort the following derivation):

$$\mathscr{L}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}(f(x_i,\boldsymbol{\theta}) - \hat{y}_i)^2 \tag{A2}$$

Consider the gradient descend procedure of the network parameters, which in noted as:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \nabla_{\boldsymbol{\theta}}\mathscr{L}(\boldsymbol{\theta}_t) \tag{A3}$$

$$\Rightarrow \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t (f(\boldsymbol{x},\boldsymbol{\theta}) - \hat{\boldsymbol{y}})\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x},\boldsymbol{\theta}_t) \tag{A4}$$

During training of the network, the input $\boldsymbol{x}$ remains stable, and only the parameter space $\boldsymbol{\theta}_t$ is updated every iteration. If a gradient flow of the $f(\boldsymbol{x},\boldsymbol{\theta})$ with an infinitely small learning rate η_t is defined, it is in fact a flow of the $\boldsymbol{\theta}_t$. The rate of the $\boldsymbol{\theta}_t$ is given by:

$$\frac{d\boldsymbol{\theta}}{dt} = \frac{\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t}{\eta_t} = -(f(\boldsymbol{x},\boldsymbol{\theta}) - \hat{\boldsymbol{y}})\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x},\boldsymbol{\theta}_t) \tag{A5}$$

Also, the first order Taylor expansion of $f(\boldsymbol{x},\boldsymbol{\theta})$ around the initial parameter value $\boldsymbol{\theta}_0$ is considered:

$$f(\boldsymbol{x},\boldsymbol{\theta}) \approx f(\boldsymbol{x},\boldsymbol{\theta}_0) + \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x},\boldsymbol{\theta}_0) \bullet (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \tag{A6}$$

Therefore, the dynamic of this network output is:

$$\frac{df(\boldsymbol{x},\boldsymbol{\theta})}{dt} \approx \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x},\boldsymbol{\theta}_0) \bullet \frac{d\boldsymbol{\theta}}{dt} \tag{A7}$$

$$\Rightarrow \frac{df(\boldsymbol{x},\boldsymbol{\theta})}{dt} \approx -\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x},\boldsymbol{\theta}) \bullet \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x},\boldsymbol{\theta}_0)^T (f(\boldsymbol{x},\boldsymbol{\theta}) - \hat{\boldsymbol{y}}) \tag{A8}$$

As derived by Jacot et al. [35], the term $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x},\boldsymbol{\theta})$ becomes deterministic and remains static when the network is infinitely wide. Even not like this, the parameters also hardly vary for an over-parameterized network during training, which is the so-called "lazy training" [49]. Under such circumstances, $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x},\boldsymbol{\theta}_0) \approx \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x},\boldsymbol{\theta})$ can be assumed. Since the neural tangent kernel (NTK) operator $\boldsymbol{K}$ is defined as:

$$\boldsymbol{K}_{ij} = \boldsymbol{K}(x_i,x_j) = \left\langle \frac{\partial f(x_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \frac{\partial f(x_j,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\rangle \tag{A9}$$

The Eq. A. 8 can be reformed as:
$\frac{df(\boldsymbol{x},\boldsymbol{\theta})}{dt} \approx -\boldsymbol{K} \bullet (f(\boldsymbol{x},\boldsymbol{\theta}) - \hat{\boldsymbol{y}})$(A10).
This concludes the proof.

Appendix B. . Fourier/Wavelet feature embedded PINN on solving an SDOF vibrator

In this illustrative example, a damped SDOF vibrator under the forced vibration is considered. This simple case is driven by the following ordinary differential equation (ODE):

$$m\ddot{x} + c\dot{x} + kx = Fsin(wt) \tag{B1}$$

With $m = 1\,kg$, $k = 200N/m$, damping ratio $\xi = 0.1$, $F = 100\,N$, $w = 120\,rad/s$. The initial condition is $x_0 = 0\,m$, $\dot{x}_0 = 0\,m/s$. According to these settings, the nature frequency of this vibrator is approximately 2.3 Hz, and the excitation frequency is roughly 19.1 Hz. Here, the physics-informed neural network (PINN) to pursue the solution to this ODE is adopted. As an emerging paradigm that promises profound influence on computational physics, PINNs have shown seductive efficacy in various disciplines [50–53]. However, PINN also suffers from spectra bias in the framework of NNs.

To present a brief overview of the PINN, Eq. B. 1 can be reformed by the general form:

$$\mathscr{N}[\boldsymbol{x}](\boldsymbol{t}) = \boldsymbol{f}(\boldsymbol{t}), \boldsymbol{t} \in \Omega \tag{B2}$$

$$\mathscr{B}[\boldsymbol{x}](\boldsymbol{t}) = \boldsymbol{g}(\boldsymbol{t}), \boldsymbol{t} \in \partial\Omega \tag{B3}$$

where $\mathscr{N}[\bullet]$ is the differential operator and $\mathscr{B}[\bullet]$ denotes the boundary condition, or initial condition. To seek the latent solution $x(t)$, a network with the input of t and output of x (see Fig. B. 1) is built. The derivative of x with respect to t can be obtained through auto differentiation [54]. Then, the PINN can be trained by minimizing the physics-induced residual loss and boundary/initial condition loss.

The calculation of this ODE is in the range of $t \in [0,3]$. The *ff/wf*PINN to solve the ODE is adopted, where Fourier/Wavelet features are embedded in the PINN. A vanilla PINN with the same architecture is also employed for comparison. The Gaussian basis function is selected for the *wf*PINN, both *ff*PINN and *wf*PINN are initialized by $\sigma = 10$. The fully connected network contains 3 hidden layers, 100 units per hidden layer, and *tanh* activation function is chosen in the network.

The raw PINN fails in capturing the solution to this ODE, as its residual loss barely has no descending in the training. On the contrary, *ff*PINN and *wf*PINN are capable to reconstruct the relatively high-frequency components in the objective function. Besides, the *ff*PINN is also outperforms the *wf*PINN in this case (see Fig. B1 and Fig. B2).

Appendix C. . Influence of the dataset size on the TL approach

Following the discussion in Section 4.3.2, the size of the training dataset is 8×10^4 (EL-3 extracted for testing). The considerable dataset size goes against the TL approach showing its merits. Herein, the training dataset is randomly cut down to lessen the dataset size, namely 80%, 60%, 40%, and 20%. The efficacy of the TL under different number of free layers is dissected on these reduced datasets based on *ff*NN with $\sigma = 30$.

Firstly, the most intuitive conclusion is that reducing the dataset will raise the MSE in training and testing. However, such influence is decreased when training the model with just one free layer, indicating that only limited data is required for training in this condition. But even though the model is well-trained to capture the features in the training dataset, the lessened training dataset still fails to provide a complete view of the task. So that the rising testing MSE with one free layer can be found when reducing the training dataset.

It can also be observed that the effect of TL becomes more apparent with a lessened dataset. Compared with the intact training dataset, the optimum free-layer number is decreased on the lessened dataset. It means the source task knowledge plays a more vital role when the training dataset of the target task is reduced. To better present this conclusion, for each dataset size, the result on the plain model without the TL approach is compared with the TL approach. An index $\epsilon_{train/test}$ is defined, which is calculated by the training/testing MSE of the non-TL model deducting the training/testing MSE of the TL models. When $\epsilon > 0$, the TL outperforms the non-TL approach and vice versa (see Fig. C1).

The comparison in Fig. C2 provides compelling proof that $\epsilon_{train/test}$ is increasing with the decreasing dataset size. In short, the TL approach can facilitate establishing a better surrogate when the dataset size is small, no matter the number of free layers.

References

[1] S.S. Hsiau, M.H. Wu, C.H. Chen, Arching phenomena in a vibrated granular bed, Powder Technol. 99 (1998) 185–193, https://doi.org/10.1016/S0032-5910(98)00103-X.

[2] J. Jin, H.I. Koh, J. Park, Experimental investigation on the rolling noise and train interior noise reduction effect with tuned particle impact damper, in: INTER-NOISE 2019 MADRID - 48th International Congress and Exhibition on Noise Control Engineering, 2019, pp. 1–3.

[3] J. Jin, W. Yang, H.I. Koh, J. Park, Development of tuned particle impact damper for reduction of transient railway vibrations, Appl. Acoust. 169 (2020), 107487, https://doi.org/10.1016/j.apacoust.2020.107487.

[4] J. Jin, H. Kim, H.I. Koh, J. Park, Railway noise reduction by periodic tuned particle impact damper with bounce and pitch-coupled vibration modes, Compos. Struct. 284 (2022), 115230, https://doi.org/10.1016/J.COMPSTRUCT.2022.115230.

[5] B.o. Fu, H. Jiang, T. Wu, Experimental study of seismic response reduction effects of particle damper using substructure shake table testing method, Struct. Control Health Monit. 26 (2) (2019), https://doi.org/10.1002/STC.2295.

[6] S. Liu, Z. Lu, P. Li, W. Zhang, E. Taciroglu, Effectiveness of particle tuned mass damper devices for pile-supported multi-story frames under seismic excitations, Struct. Control Health Monit. 27 (2020) e2627.

[7] W. Xiao, D. Lu, L. Song, H. Guo, Z. Yang, Influence of particle damping on ride comfort of mining dump truck, Mech. Syst. Sig. Process. 136 (2020), 106509, https://doi.org/10.1016/J.YMSSP.2019.106509.

[8] C. Ma, Z. Lu, D. Wang, Z. Wang, Study on the damping mechanisms of a suspended particle damper attached to a wind turbine tower, Wind Struct. 33 (2021) 103–114, https://doi.org/10.12989/WAS.2021.33.1.103.

[9] G. Jin, Z. Zhao, B. Liu, W. Cun, Z. Zhao, M. Hou, G. Chen, Design of a particle damper and experimental study on vibration damping of the pipeline, Adv. Mech. Eng. 13 (2021) 1–14, https://doi.org/10.1177/16878140211044923/ASSET/IMAGES/LARGE/10.1177_16878140211044923-FIG2.

[10] S.F. Masri, A.M. Ibrahim, Response of the impact damper to stationary random excitation, J. Acoust. Soc. Am. 53 (1973) 200–211. https://doi.org/10.1121/1.1913319.

[11] Z. Lu, X. Lu, W. Lu, S.F. Masri, Shaking table test of the effects of multi-unit particle dampers attached to an MDOF system under earthquake excitation, Earthq. Eng. Struct. Dyn. 41 (2012) 987–1000, https://doi.org/10.1002/eqe.1170.

[12] Z. Lu, S.F. Masri, X. Lu, Studies of the performance of particle dampers attached to a two-degrees-of-freedom system under random excitation, J. Vib. Control 17 (2011) 1454–1471, https://doi.org/10.1177/1077546310370687.

[13] C. Gnanasambandham, A. Schönle, P. Eberhard, Investigating the dissipative effects of liquid-filled particle dampers using coupled DEM–SPH methods, Comput. Part. Mech. 6 (2019) 257–269, https://doi.org/10.1007/S40571-018-0212-9/FIGURES/17.

[14] N. Meyer, R. Seifried, Numerical and experimental investigations in the damping behavior of particle dampers attached to a vibrating structure, Comput. Struct. 238 (2020), 106281, https://doi.org/10.1016/J.COMPSTRUC.2020.106281.

[15] H. Guo, K. Ichikawa, H. Sakai, H. Zhang, X. Zhang, K. Tsuruta, K. Makihara, A. Takezawa, Numerical and experimental analysis of additively manufactured particle dampers at low frequencies, Powder Technol. 396 (2022) 696–709, https://doi.org/10.1016/J.POWTEC.2021.11.029.

[16] F. Biondani, M. Morandini, G.L. Ghiringhelli, M. Terraneo, P. Cordisco, Efficient discrete element modeling of particle dampers, Processes. 10 (2022) 1247, https://doi.org/10.3390/PR10071247.

[17] Y. Hu, H. Zan, Y. Guo, J. Jiang, Z. Xia, H. Wen, Z. Peng, Energy dissipation characteristics of particle dampers with obstacle grids, Mech. Syst. Sig. Process. 193 (2023) 110231.

[18] E. DeGiuli, J.N. McElwaine, M. Wyart, Phase diagram for inertial granular flows, Phys. Rev. E 94 (2015), https://doi.org/10.1103/PhysRevE.94.012904.

[19] K. Zhang, T. Chen, X. Wang, J. Fang, Motion mode of the optimal damping particle in particle dampers, J. Mech. Sci. Technol. 30 (4) (2016) 1527–1531.

[20] S. He, F. Chen, N. Xu, High-speed train localization algorithm via cooperative multi-classifier network using distributed heterogeneous signals, J. Frank. Inst. 360 (12) (2023) 8096–8117.

[21] S. He, F. Chen, H. Chen, A latent representation generalizing network for domain generalization in cross-scenario monitoring, IEEE Trans. Neural Netw. Learn Syst. (2023) 1–15, https://doi.org/10.1109/TNNLS.2023.3296942.

[22] P. Veeramuthuvel, K.K. Sairajan, K. Shankar, Vibration suppression of printed circuit boards using an external particle damper, J. Sound Vib. 366 (2016) 98–116, https://doi.org/10.1016/j.jsv.2015.12.034.

[23] D. Wang, C. Wu, Parameter estimation and arrangement optimization of particle dampers on the cantilever rectangular plate, J. Vibroeng. 17 (2015) 2503–2520.

[24] X. Ye, Y.Q. Ni, M. Sajjadi, Y.-W. Wang, C.-S. Lin, Physics-guided, data-refined modeling of granular material-filled particle dampers by deep transfer learning, Mech. Syst. Sig. Process. 180 (2022), 109437, https://doi.org/10.1016/J.YMSSP.2022.109437.

[25] R. Wang, F. Yan, L. Yu, C. Shen, X. Hu, J. Chen, A federated transfer learning method with low-quality knowledge filtering and dynamic model aggregation for rolling bearing fault diagnosis, Mech. Syst. Sig. Process. 198 (2023), 110413, https://doi.org/10.1016/J.YMSSP.2023.110413.

[26] Y. Cao, Z. Fang, Y. Wu, D.X. Zhou, Q. Gu, Towards understanding the spectral bias of deep learning, in: IJCAI International Joint Conference on Artificial Intelligence, 2019, pp. 2205–2211, https://doi.org/10.48550/arxiv.1912.01198.

[27] R. Basri, D. Jacobs, Y. Kasten, S. Kritchman, The convergence rate of neural networks for learned functions of different frequencies, Adv. Neural Inf. Proces. Syst. 32 (2019), https://doi.org/10.48550/arxiv.1906.00425.

[28] Z.Q.J. Xu, Y. Zhang, T. Luo, Y. Xiao, Z. Ma, Frequency principle: Fourier analysis sheds light on deep neural networks, Comm. Comput. Phys. 28 (2019) 1746–1767, https://doi.org/10.4208/cicp.OA-2020-0085.

[29] N. Rahaman, A. Baratin, D. Arpit, F. Draxlcr, M. Lin, F.A. Hamprecht, Y. Bengio, A. Courville, On the spectral bias of neural networks, in: 36th International Conference on Machine Learning, 2018, pp. 9230–9239, https://doi.org/10.48550/arxiv.1806.08734.

[30] X.A. Li, Z.Q.J. Xu, L. Zhang, A multi-scale DNN algorithm for nonlinear elliptic equations with multiple scales, Comm. Comput. Phys. 28 (2020) 1886–1906, https://doi.org/10.4208/cicp.OA-2020-0187.

[31] Z. Liu, W. Cai, Z.Q.J. Xu, Multi-scale deep neural network (MscaleDNN) for solving Poisson-Boltzmann equation in complex domains, Comm. Comput. Phys. 28 (2020) 1970–2001, https://doi.org/10.4208/cicp.OA-2020-0179.

[32] B. Moseley, A. Markham, T. Nissen-Meyer, Finite basis physics-informed neural networks (FBPINNs): a scalable domain decomposition approach for solving differential equations, arXiv preprint arXiv:2107.07871, 2021, https://doi.org/10.48550/arxiv.2107.07871.

[33] S. Wang, X. Yu, P. Perdikaris, When and why PINNs fail to train: a neural tangent kernel perspective, J. Comput. Phys. 449 (2022). https://arxiv.org/abs/2007.14527v1.

[34] S. Wang, Y. Teng, P. Perdikaris, Understanding and mitigating gradient flow pathologies in physics-informed neural networks, SIAM J. Sci. Comput. 43 (2021) 3055–3081, https://doi.org/10.1137/20M1318043.

[35] A. Jacot, F. Gabriel, C. Hongler, Neural tangent kernel: convergence and generalization in neural networks, Adv. Neural Inf. Proces. Syst. (2018) 8571–8580, https://doi.org/10.48550/arxiv.1806.07572.

[36] M. Tancik, P.P.P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J.T.T. Barron, R. Ng, Fourier features let networks learn high frequency functions in low dimensional domains, Adv. Neural Inform. Process. Syst. (2020), https://doi.org/10.48550/arxiv.2006.10739.

[37] S. Wang, H. Wang, P. Perdikaris, On the eigenvector bias of Fourier feature networks: from regression to solving multi-scale PDEs with physics-informed neural networks, Comput. Methods Appl. Mech. Eng. 384 (2021), 113938, https://doi.org/10.1016/J.CMA.2021.113938.

[38] G. Wang, L. Guo, H. Duan, Wavelet neural network using multiple wavelet functions in target threat assessment, Scient. World J. 2013 (2013) 1–7.

[39] J.C. Wong, C. Ooi, A. Gupta, Y.S. Ong, Learning in sinusoidal spaces with physics-informed neural networks, IEEE Trans. Artif. Intell. (2021), https://doi.org/10.48550/arxiv.2109.09338.

[40] J. Lee, L. Xiao, S.S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, J. Pennington, G. Brain, Wide neural networks of any depth evolve as linear models under gradient descent, Adv. Neural Inf. Process. Syst. (2019) 32. https://g.co/airesidency.

[41] M. Raissi, P. Perdikaris, G.E. Karniadakis, Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, J. Comput. Phys. 378 (2019) 686–707, https://doi.org/10.1016/j.jcp.2018.10.045.

[42] N. Jin, D. Liu, Z. Pang, T. Huang, Wavelet basis function neural networks, in: IEEE International Conference on Neural Networks, 2007, pp. 500–505. https://doi.org/10.1109/IJCNN.2007.4371007.

[43] H. Henderi, T. Wahyuningsih, E. Rahwanto, Comparison of min-max normalization and Z-score normalization in the k-nearest neighbor (knn) algorithm to test the accuracy of types of breast cancer, Int. J. Inf. Inform. Syst. 4 (2021) 13–20, https://doi.org/10.47738/IJIIS.V4I1.73.

[44] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, C. Liu, A survey on deep transfer learning, in: Artificial Neural Networks and Machine Learning – ICANN 2018, 11141 LNCS, 2018, pp. 270–279, https://doi.org/10.1007/978-3-030-01424-7_27.

[45] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proc. IEEE 109 (2021) 43–76, https://doi.org/10.1109/JPROC.2020.3004555.

[46] C.J. Wu, W.H. Liao, M.Y. Wang, Modeling of granular particle damping using multiphase flow theory of gas-particle, J. Vib. Acoust. 126 (2004) 196–201, https://doi.org/10.1115/1.1688763.

[47] X. Lei, C. Wu, Non-obstructive particle damping using principles of gas-solid flows, J. Mech. Sci. Technol. 31 (2017) 1057–1065, https://doi.org/10.1007/s12206-017-0204-3.

[48] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (2010) 1345–1359, https://doi.org/10.1109/TKDE.2009.191.

[49] L. Chizat, E. Oyallon, F. Bach, On lazy training in differentiable programming, Adv. Neural Inf. Proces. Syst. 32 (2018), https://doi.org/10.48550/arxiv.1812.07956.

[50] Y. Zhu, N. Zabaras, P.S. Koutsourelakis, P. Perdikaris, Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data, J. Comput. Phys. 394 (2019) 56–81, https://doi.org/10.1016/J.JCP.2019.05.024.

[51] Z. Fang, J. Zhan, Deep physical informed neural networks for metamaterial design, IEEE Access 8 (2020) 24506–24513, https://doi.org/10.1109/ACCESS.2019.2963375.

[52] F. Sahli Costabal, Y. Yang, P. Perdikaris, D.E. Hurtado, E. Kuhl, Physics-informed neural networks for cardiac activation mapping, Front. Phys. 8 (2020) 42, https://doi.org/10.3389/FPHY.2020.00042/BIBTEX.

[53] Y.A. Yucesan, F.A.C. Viana, A hybrid physics-informed neural network for main bearing fatigue prognosis under grease quality variation, Mech. Syst. Sig. Process. 171 (2022), 108875, https://doi.org/10.1016/J.YMSSP.2022.108875.

[54] A. Güne¸, G. Baydin, B.A. Pearlmutter, J.M. Siskind, Automatic differentiation in machine learning: a survey, J. Mach. Learn. Res. 18 (2018) 1–43, https://doi.org/10.48550/arXiv.1502.05767.