# Unsupervised Representation Learning for Monitoring Rail Infrastructures With High-Frequency Moving Vibration Sensors

Phusakulkajorn, Wassamon; Zeng, Yuanchen; Li, Zili; Nunez, Alfredo

# Unsupervised Representation Learning for Monitoring Rail Infrastructures With High-Frequency Moving Vibration Sensors

Wassamon Phusakulkajorn, Yuanchen Zeng, *Member, IEEE*, Zili Li, and Alfredo Núñez, *Senior Member, IEEE*

*Abstract*—Nowadays, rolling stock can be equipped with high-frequency vibration sensors to continuously monitor rail infrastructures and detect defects. These moving sensors measure at high speeds and sampling frequencies, generating a massive amount of data that covers each track position with very short signal durations. These data contain a variety of dynamic and transient responses that vary significantly along the track and are affected by noise. This leads to a large amount of unlabeled and noisy data, complicating the extraction of dynamic responses for effective anomaly detection. To address these challenges, this paper proposes an unsupervised representation learning methodology to automatically capture and extract characteristic features of dynamic responses that reflect the conditions of rail infrastructures. The unsupervised nature allows exploratory analysis of high-frequency vibration signals when prior knowledge or reference information about infrastructure conditions is unavailable or very limited. A collaborative optimization process that synchronizes empirical mode decomposition (EMD) with a convolutional autoencoder (CAE) is presented. The EMD level is tuned to remove noise while preserving effective vibration responses. The CAE is trained using demodulated signals that are considered normal to generate representations that ensure reconstruction quality and differentiate between normal and abnormal conditions. Furthermore, a Gaussian mixture model is used to showcase the effectiveness of the learned representations for rail infrastructures. Applied to validated axle box acceleration data for rail defect detection and train-borne laser Doppler vibrometer data for rail fastener monitoring, our method outperforms other variants of autoencoder-based models and the wavelet-based CAE in accurately identifying the conditions. It achieves an average improvement of 16% with the axle box acceleration data and 21% with the laser Doppler vibrometer data.

*Index Terms*—Unsupervised learning, axle box acceleration, laser Doppler vibrometer, autoencoders, empirical mode decomposition, high-frequency data.

The authors are with the Section of Railway Engineering, Department of Engineering Structures, Delft University of Technology, 2628 CN Delft, The Netherlands (e-mail: W.Phusakulkajorn@tudelft.nl; Y.Zeng-2@tudelft.nl; Z.Li@tudelft.nl; A.A.Nunezvicencio@tudelft.nl).

## I. INTRODUCTION

**S**TRUCTURAL health monitoring plays a pivotal role in ensuring the safety and integrity of rail infrastructures [1]. Through the development of various sensing technologies and data analytics techniques, defects can be detected timely, thus allowing corrective and predictive maintenance to prevent catastrophic accidents. The collection and analysis of data further enable the digitalization of railway transportation systems. Among various monitoring technologies, vibration-based monitoring is an effective approach to characterize a wide range of dynamic behaviors and properties of rail infrastructures [2], [3], [4], [5], [6].

Vibration-based monitoring can be implemented by distributing sensors on rail infrastructures and measuring their vibrations induced by moving train loads, such as in [7] and [8]. Distributed sensors can record structural vibrations at different locations over time, thus providing rich data for parameter estimation and health assessment. Numerous methods have been developed to analyze such signals, including signal processing methods [9], [10] and machine learning methods [11]. However, it is cost-prohibitive to apply distributed sensors to large-scale transportation systems, such as railway lines spanning thousands of kilometers.

Vibration-based monitoring of rail infrastructures with sensors on trains in operation is gaining increasing prominence. In [12], [13], [14], [15], and [16], accelerometers are utilized, while in [17], [18], and [19], vibrometers are employed to monitor various components and properties of railway track structures. In [20] and [21], smartphones are used for evaluating the quality of railway tracks. These technologies enable the dynamic behaviors at different locations of an infrastructure to be measured in a single run, which is highly preferred for large-scale monitoring. These technologies generally pursue monitoring under the operational speed and load of the rail network to avoid disturbance to train traffic and capture structural response under realistic loading conditions.

Compared to the use of distributed sensors, vibration-based monitoring with moving sensors poses several challenges for data analysis and anomaly detection. The first challenge is significant variability in dynamic behaviors at different locations,
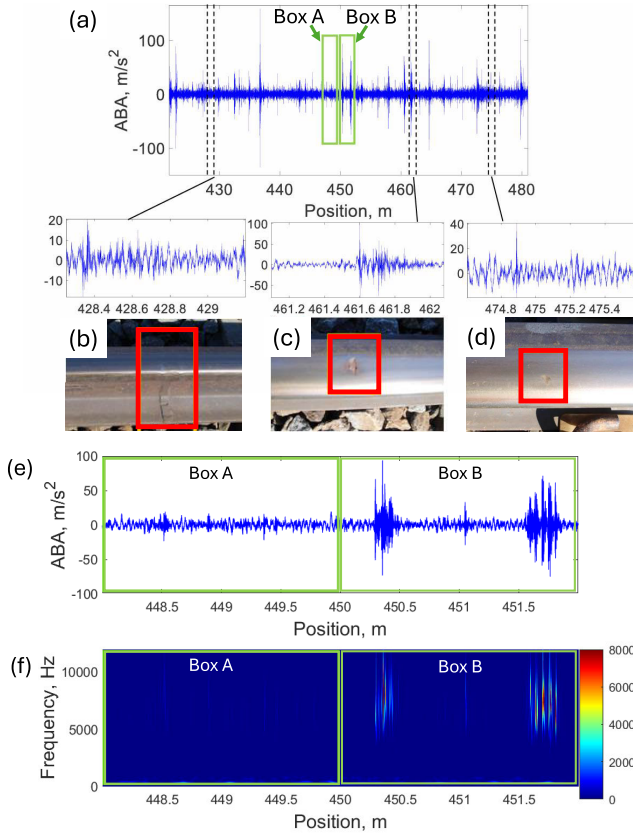
Fig. 1. (a) Example of a raw ABA signal containing various rail dynamics: (b) weld, (c) squat, (d) small defect, and (e) the ABA signals within the green boxes from (a), and (f) the corresponding wavelet power spectrum.
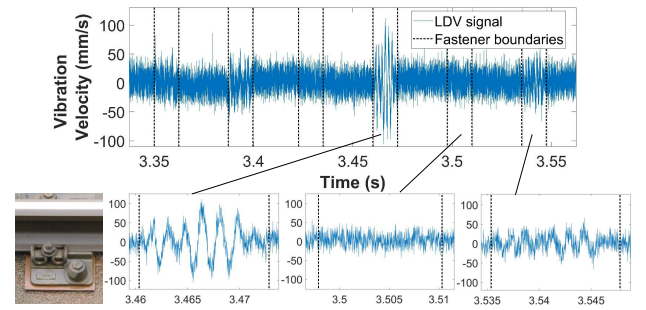


Fig. 2. Example of a raw LDV signal containing severe noise.

102.4 kHz in [17]. This generates large volumes of data with short durations and high frequencies, requiring effective and efficient methods to process them.

The third challenge concerns disturbances and noise from multiple sources, including vehicle vibrations, track irregularities, and measurement noise. Fig. 2 shows an example of the train-borne LDV signal measured on rail fasteners where severe noise is observed. The major source is the speckle noise due to the drastic change of speckle patterns as the laser spot scans the rough surfaces [23]. The speckle noise deteriorates the quality and usability of LDV signals. It obscures the actual vibration patterns of defects and anomalies, making it difficult to detect accurately. The mitigation of noise and disturbances is achieved with specialized filtering algorithms in [18] and [23].

The last challenge involves fuzziness in distinguishing between different dynamic behaviors. For instance, when dealing with defects at the early stage of their development, their responses in ABA signals are subtle and slightly different from those of healthy rails, as seen from the dynamic responses of the small defect in Fig. 1. These responses may not be apparent to human inspectors or traditional detection methods. As a result, many samples fall within the ambiguous boundaries between normal and abnormal conditions. These samples complicate the labeling process and yield a substantial amount of unlabeled data, hindering the use of supervised learning.

Conventionally, the above challenges are addressed by crafting data analysis methods based on the knowledge and experience of experts in the physics of targeted structures and defects, such as [18] and [24]. However, with the increasing volume of data, diversity of objects, and complexity of features, manual judgment becomes less efficient, robust, and reliable. Alternatively, the advancement of unsupervised learning has the potential to fill this gap [25], [26], [27], [28].

Several unsupervised learning approaches have been developed to analyze vibration data, such as t-distributed stochastic neighbor embedding (t-SNE) [29], K-means clustering [30], principal component analysis (PCA) [31], tensor clustering [32], and one-class support vector machine (OC-SVM) [33]. Deep learning has gained attention in addressing such problems in recent years, and various auto-encoder (AE) variants have been developed for feature learning. In [34], [35], [36], and [37], convolutional autoencoder (CAE) is exploited in feature extraction and fault detection based on vibration signals. In [38], an expectation-maximization algorithm with

which is affected by local properties of rail infrastructures and changing operational conditions of trains. For example, Fig. 1 depicts a portion of a vibration signal measured by an axle box accelerometers (ABA). This signal contains rail dynamics that vary along the track, which are evident by the vibration patterns of the weld and surface defects, highlighted as examples in the figure. Additionally, the green boxes in Fig. 1(e) represent the ABA signals from nearby locations that exhibit significantly different amplitudes in their frequency contents, shown in Fig. 1(f). This variability across large-scale infrastructures complicates the identification of anomalies, necessitating advanced techniques to isolate them.

The second challenge arises from the need to segment signals for high-resolution localization of defects. This segmentation process results in short-duration signal fragments. For instance, when a train travels at a speed of 100 km/h, it covers a distance of 1 meter in 36 milliseconds. As train speed increases, the time duration over a specific distance decreases accordingly. A shorter duration of the signals makes it more difficult to achieve accurate and reliable defect detection because less information is available within the segment. Therefore, a high sampling frequency is necessary to capture the variations of dynamic behaviors within these signals, resulting in an increased number of data points within a given distance. For example, ABA data are collected with a high sampling frequency of 25.6 kHz in [22], while the sampling frequency of a train-borne laser Doppler vibrometer (LDV) is

an adversarial autoencoder was used in feature extraction for rotating machinery fault diagnosis. In [39], an unsupervised feature learning method for machinery health monitoring was proposed using a generative adversarial networks model. In [40], [41], and [42], a method based on sparse filtering, an unsupervised two-layer neural network, was proposed for fault feature learning from mechanical vibration signals. In [43], an AR model was used to extract features obtained from accelerometers installed on the rails during the passage of traffic loads. Then, an unsupervised methodology using outlier and clustering analysis was developed to identify wheel flats. In [44], an unsupervised degradation state evaluation method based on a deep tensor autoencoder network was proposed to automatically extract representation and assess the health conditions of metro wheels. In [45], a second-generation wavelet deep autoencoder (SWDAE) network was developed in an unsupervised manner to evaluate the degradation state of metro wheels.

Table I summarizes the selected unsupervised learning research focused on vibration-based monitoring in the railway domain. Given the challenges mentioned above, a very limited number of unsupervised learning methods have been developed to handle vibration data from moving sensors. Therefore, this paper develops an unsupervised learning methodology to automatically extract features characterizing the dynamic behaviors of rail infrastructures measured by high-frequency moving vibration sensors. The methodology employs a collaborative optimization process that synchronizes empirical mode decomposition (EMD) with the parameters of a CAE. EMD is chosen for its ability to adaptively separate different frequency characteristics without the need for predefining each frequency range [13], [46]. EMD offers the possibility to explore dynamic behaviors over a broad frequency band while reducing the disturbance of measurement noise. The CAE is selected for its ability to extract informative features from vibration signals to characterize dynamic behaviors automatically. Furthermore, the CAE's simple yet effective structure is well-suited for limited labeled data common in railway applications. This minimizes the risk of overfitting while capturing meaningful representations. The latent features learned from the collaborative process of EMD and CAE ensure reconstruction quality and distinguish between normal and abnormal behaviors. This methodology facilitates various analyses, including dimensionality reduction, classification, clustering, and supports anomaly detection by identifying deviations from the normal behavior pattern. The key contributions of this work are summarized as follows:

- An unsupervised representation learning methodology is proposed to automatically extract features that characterize dynamic behaviors of rail infrastructures from vibration data that is noisy and short-duration obtained from high-frequency moving sensors at different locations.
- A collaborative optimization process between EMD level and the parameters of CAE is proposed for a generation of representations that demonstrate reconstruction quality and can differentiate between rail infrastructures under normal and abnormal conditions.

- Two field measurements with different targeted components, sensor types, and operational conditions are used to demonstrate the applicability and performance of the proposed methodology for monitoring rail defects using an ABA and rail fasteners using a train-borne LDV.

The rest of the paper is as follows. Section II presents fundamental knowledge of the methodology used. Section III presents problem formulation and the proposed framework. In Section IV, the proposed unsupervised representation learning methodology is elaborated. Section V describes the real-world applications used to showcase the methodology. The comparison of the proposed methodology with different models from the literature and discussions are presented in Section VI. The paper is concluded in Section VII.

## II. FUNDAMENTAL KNOWLEDGE

### A. Empirical Mode Decomposition

The EMD has been pioneered by Huang et al. [46] for adaptively representing nonlinear and non-stationary signals. For a given signal $x(t)$, EMD decomposes $x(t)$ into a series of intrinsic mode functions (IMFs), denoted as $c_i(t)$, where $i = 1, 2, \ldots, I$ and $I$ is the total number of IMFs, and a residual $r_i(t)$ [47]. The EMD algorithm, referred to as the sifting process, iteratively extracts IMFs based on the local maxima and minima of the signal. At the end of the EMD process, the original signal $x(t)$ can be expressed by a sum of IMFs and a residual component as [48]:

$$x(t) = \sum_{i=1}^{I} c_i(t) + r_I(t). \tag{1}$$

### B. Convolutional Autoencoder

A CAE is a type of autoencoder architecture that combines convolutional layers with an autoencoder model [34]. It aims at unsupervised learning of a lower-dimensional representation from higher-dimensional data. The CAE network comprises three main parts: encoder, latent representation, and decoder. The encoder and decoder are designed using convolutional layers and can have several hidden layers, making a deep CAE.

Given an input signal $x(t)$ and let $z_l$ denote the output of the $l^{th}$ convolutional layer, where $l$ ranges from 1 to $L$, in which $L$ is the number of convolutional layers designed in the encoder. The output of the encoder $z_L$ is a latent representation of the input data $x(t)$, typically with reduced dimensions. The decoder takes the latent representation $z_L$ and aims to reconstruct the original input $x(t)$. The decoder consists of a stack of transposed convolutional layers. Let $u_k$ denote the output of the $k^{th}$ transposed convolutional layer, where $k$ ranges from 1 to the number of transposed convolutional layers $K$. The output of the decoder $u_K = \tilde{x}(t)$ is the reconstructed version of the input data $x(t)$. Mathematically, the process of encoding and decoding in a CAE is given as follows:

$$z_l = \sigma\left(w_l * z_{l-1} + b_l\right), \tag{2}$$

$$u_k = \sigma\left(v_k * u_{k-1} + e_k\right), \tag{3}$$

where $*$ denotes the convolution operation, $\sigma$ denotes an activation function, $w_l$ and $b_l$ are the weights and biases of the

TABLE I

SUMMARY OF THE SELECTED UNSUPERVISED LEARNING RESEARCH FOCUSED ON VIBRATION-BASED MONITORING IN THE RAILWAY DOMAIN. NOTE THAT SENSOR TYPE IS DEFINED WITH RESPECT TO THE SPECIFIC COMPONENTS BEING TARGETED

| Ref. | Sensor Type | Component | Methodology | Key features |
|---|---|---|---|---|
| [31] | Distributed sensor (Accelerometers) ($f_s$ = N/A) | Railway bridge | Robust PCA | Achieve good damage detection accuracy based on simulated damage; Mitigate environmental effects; No automatic representations learning and dependent on domain knowledge; |
| [32] | Distributed sensor (Accelerometers) ($f_s$ = 25.6 kHz) | Train bearing | Density-based affining propagation tensor | Achieve good bearing fault detection based on laboratory tests; Handle the complexities associated with variable working conditions; No assumption about data distribution under varying conditions; No automatic representations learning and dependent on domain knowledge; Dimensionality increase; Lack interpretability; Computationally intensive. |
| [33] | Moving sensor (Accelerometers) ($f_s$ = 400 Hz) | Track | OC-SVM | Effectively identify track geometrical defects based on real field tests; Limited robustness to noise and data quality issues; No automatic learning of representations and dependent on domain knowledge; Curse of dimensionality. |
| [35] | Distributed sensor (ultra-weak FBG) ($f_s$ = 1 kHz) | Fasteners | CAE with Pseudo-Hilbert scan | Effectively extract features for fastener with complete looseness based on real field tests; Automatic representation learning; Lack robustness to noise; Computational intensive for signal transformation into images; Lack interpretability. |
| [43] | Moving sensor (Accelerometers) ($f_s$ = 20 kHz) | Wheels | AR model; Outlier analysis; K-means clustering | Effectively detect railway wheel flats based on simulated data; Lack robustness to noise; No automatic learning of representations; Difficulty in defining a unique solution for the AR model order; Dependency on domain knowledge for outlier threshold and the number of clusters. |
| [44], [45] | Distributed sensor (Accelerometers) ($f_s$ = 2 kHz) | Wheels | Deep tensor AE; Second-generation wavelet deep AE | Effectively characterize wheel degradation based on real field tests; Automatic representation learning; Mitigate noise; Handle missing values in [44]; Lack interpretability in [44]; Meaningful representations in [45]; Need predefined frequency bands in [45]. |
| This paper | Moving sensor (ABA, LDV) ($f_s$ = 25.6 kHz) ($f_s$ = 102.4 kHz) | Rail; Rail fasteners | CAE with EMD | Effectively detect rail defects and invalid LDV measurements on fasteners based on real field tests; Automatic representation learning; Mitigate noise; Meaningful representations; Handle vagueness and imprecision in distinguishing between different behaviors; No predefined basis functions needed for adaptively decomposing signals. |

$l^{th}$ convolutional layer, $v_k$ and $e_k$ are the weights and biases of the $k^{th}$ transposed convolutional layer.

## III. PROBLEM FORMULATION AND THE PROPOSED FRAMEWORK

Let $x_a(t) = a(p(t))$ denote acceleration signals from ABA and $x_v(t) = v(p(t))$ denote velocity signals from LDV obtained from a positioning system at track position $p(t)$ and at time instance $t$. For simplicity, we will represent these signals collectively as $x(t)$ throughout the rest of the paper.

For a vibration signal $x(t)$ obtained from a moving sensor collected with sampling frequency $f$ Hz and measuring speed $s(t)$ m/s, this work assumes segmenting the signal $x(t)$ into a set $\mathcal{D}$ of $M$ smaller segments containing $\beta$ datapoints. The set $\mathcal{D}$ is mathematically expressed as:

$$\mathcal{D} = \left\{ x_m(t) \mid t = t_{(m-1)\phi+1}, \dots, t_{(m-1)\phi+\beta}, \right\}_{m=1}^{M}, \quad (4)$$

where $x_m(t)$ denote the $m^{th}$ segment corresponding to the time instances between $t_{(m-1)\phi+1}$ and $t_{(m-1)\phi+\beta}$ with time duration $\zeta_m = t_{(m-1)\phi+\beta} - t_{(m-1)\phi+1}$ and $\phi$ is the number of datapoints used as a step size for moving the segment.

The signal segmentation can be done to allow overlapping, meaning $\phi < \beta$. A local optimization approach can also be considered to identify the appropriate time duration suitable for the data analysis task. This is due to a need to balance between accommodating the characteristic response length and capturing sufficient detail for the detection and localization of the targeted rail infrastructure. Given a targeted rail infrastructure $r$, the process begins by identifying the types of

defects under consideration, in which a combination of defect types is possible for the analysis. Using field data and expert knowledge, we analyze the dynamic responses associated with each defect type. The response durations, which vary significantly depending on the type of defect, are estimated to provide initial guidelines for optimization. The signal is then segmented to cover the range of dynamic responses for the identified defects. Let $t_{b_r}$ and $t_{e_r}$ be the starting and ending time of its dynamic response, and $\tau_r = t_{e_r} - t_{b_r}$ be the duration of this response. Additionally, let $t_{\alpha_r}$ represent an additional time duration used to capture extra responses. Then, the initial time duration is defined as $\zeta_m = \tau_r + t_{\alpha_r}$. An iterative optimization process is then performed to refine $\zeta_m$, identifying the duration that best captures the relevant characteristics of the defects. This evaluation determines the segment length that yields local optimal performance for detecting or analyzing defects, which is subsequently selected for further analysis. The appropriate segment length can be determined by minimizing the datapoints $\beta$ contained within each segment while finding an optimal $t_{\alpha_r}$ that ensures the segments cover the critical dynamic response duration $\tau_r$. This optimization process can be formalized into the objective function and constraints as follows.

$$\min_{\beta, t_{\alpha_r}} \zeta_m(\beta) \quad (5)$$

subject to, $\forall r \in \mathbb{Z}^+$,

$$2t_{\alpha_r} + \tau_r \leq \zeta_m(\beta), \forall m = 1, \dots, M, \quad (6)$$

$$\left[ t_{b_r} - t_{\alpha_r}, t_{e_r} + t_{\alpha_r} \right] \subseteq \left[ t_{(m-1)\phi+1}, t_{(m-1)\phi+\beta} \right], \exists m. \quad (7)$$

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

PHUSAKULKAJORN et al.: UNSUPERVISED REPRESENTATION LEARNING FOR MONITORING RAIL INFRASTRUCTURES 5
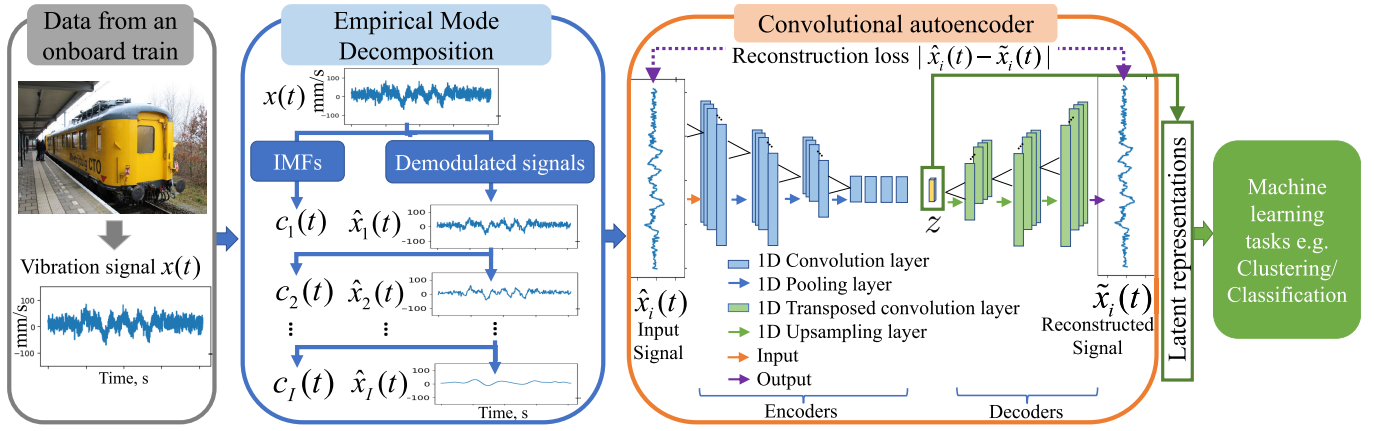


Fig. 3. Framework for unsupervised representation learning.

Unsupervised representation learning aims at autonomously extracting useful features that capture key characteristics of data without labels. This technique generally involves a reconstruction process relying on learning representations from data with normal behaviors/conditions to capture their common and essential features. In alignment with this principle, our paper introduces a hard threshold for identifying some normal samples from the unlabeled data. This hard threshold allows the reconstruction process to function without requiring a complete and pure dataset of normal data. Instead, it allows part of normal data to be utilized and even some ambiguous abnormalities to be included. This approach eliminates the need for extensive collections of purely normal data. The flexibility in data acquisition aligns with the unsupervised learning paradigm, which does not rely on prior input from experts or extensive fieldwork.

For a given hard threshold $\lambda$, this paper considers a set $\mathcal{N} \subseteq \mathcal{D}$ containing samples most likely with normal conditions for representation learning. The set $\mathcal{N}$ is obtained as:

$$\mathcal{N} = \left\{ x_n(t) \in \mathcal{D} \mid C\big(x_n(t), \lambda\big) \right\}_{n=1}^{N}, N < M, \qquad (8)$$

where $C$ is a specified condition for a selection of normal samples.

As vibration signals from moving sensors contain various disturbances and noise, this work employs EMD to decompose $x_n(t) \in \mathcal{N}$ into several IMFs. Each IMF represents an oscillatory mode embedded within the signal, with different levels of IMFs capturing different high-frequency components and noise inherent in the signal. The residual signal represents the part of the signal that cannot be effectively decomposed into the IMFs of the respective level. It typically consists of low-frequency components and trends of the signal. This paper considers the residual signal at various levels $r_i(t)$, $i = 1, 2, \ldots, I$, to extract representations inherent in the normal data. This approach resembles multi-level denoising, as the noise components are removed at various levels. In this context, these residual signals are referred to as demodulated signals as they are the signals from which the cumulative sum of IMFs up to level $i$ has been extracted from the original signal. Let $\widehat{N}_i$ be a set of demodulated signals at level $i$, then the demodulated signal $\hat{x}_{n,i}(t)$ at level $i$ within this set

is expressed as:

$$\hat{x}_{n,i}(t) = x_n(t) - \sum_{j=1}^{i} c_j(t). \qquad (9)$$

The EMD method allows for a separation of noise and extraction of meaningful representations from the vibration signals. Therefore, this paper considers learning a representation of $\mathcal{N}$ through the demodulated signals in $\widehat{\mathcal{N}}$.

For the reconstruction process at each EMD level $i$, an autoencoder model $F_i = D_i \circ E_i$ is developed such that the encoder $E_i$ encodes the demodulated signals $\hat{x}_{n,i}(t) \in \widehat{\mathcal{N}}_i$ into a lower-dimensional latent representation space $z_{n,i}(t)$. Then, the decoder $D_i$ decodes it back to the original space, yielding the reconstructed data $\tilde{x}_{n,i}(t)$ from its representation. The reconstruction process is mathematically expressed as:

$$\tilde{x}_{n,i}(t) = F_i(\hat{x}_{n,i}(t)) = D_i(E_i(\hat{x}_{n,i}(t))), \qquad (10)$$

in which

$$z_{n,i}(t) = E_i(\hat{x}_{n,i}(t)), \qquad (11)$$
$$\tilde{x}_{n,i}(t) = D_i(z_{n,i}(t)). \qquad (12)$$

This paper proposes a methodology to collaboratively train the EMD and reconstruction for generating discriminative and informative representations. Fig. 3 illustrates the proposed unsupervised representation learning framework, consisting of three main parts. The first part utilizes EMD to remove the disturbance and noise and preserve the vibration pattern in the learning data, resulting in the demodulated signals. The second part uses a CAE model to implement representation learning from the demodulated signals based on the reconstruction technique. The final part applies the learned representations to various data analysis tasks. This step demonstrates the practical utility of the extracted features for further analysis, such as anomaly detection.

## IV. PROPOSED METHODOLOGY

Fig. 4 illustrates the workflow of the collaborative method proposed in this paper. The method involves, firstly, signal denoising via EMD described in (9), secondly, identification of a corresponding autoencoder model defined in (10), (11),
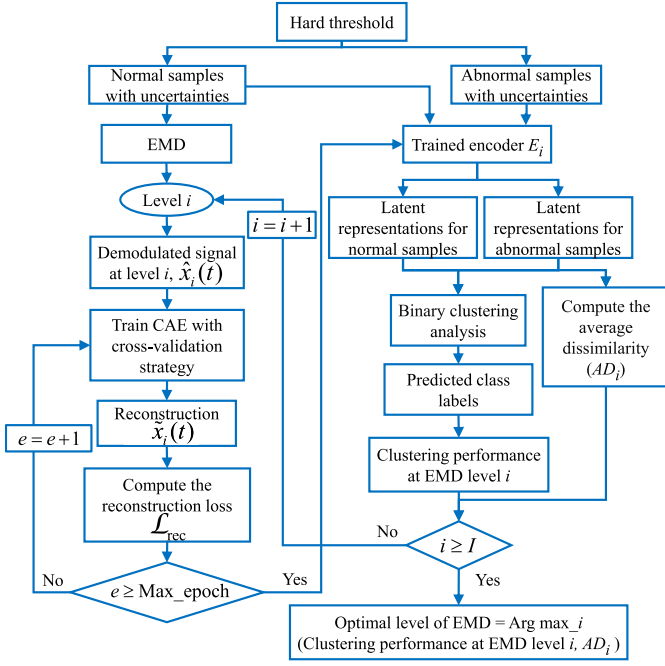
Fig. 4.   Workflow of the proposed methodology.

$$+ \frac{1}{2} \sum_{j} \left( \log(\sigma_{\text{prior}}^2) - 1 - \frac{\sigma_j^2}{\sigma_{\text{prior}}^2} + \frac{\mu_j^2}{\sigma_{\text{prior}}^2} \right), \quad (15)$$

where $N$ is the number of normal samples, $\mu_i$ and $\sigma_i$ are, respectively, the mean and standard deviation of the distribution of the latent representation $z$ along the $j^{th}$ dimension, $\mu_{\text{prior}}$ and $\sigma_{\text{prior}}$ are the mean and standard deviation of the prior distribution, in which its value can be obtained from a standard Gaussian distribution, i.e., $\mu_{\text{prior}} = 0$ and $\sigma_{\text{prior}} = 1$.

The trained CAE allows us to obtain an encoder $E_i$ developed with respect to the EMD level $i$, which is used to generate the respective embedded representation $z_{n,i}(t)$. This paper considers a binary clustering task to identify the optimal EMD level that provides the discriminative representation between normal and abnormal conditions of rail infrastructures. The Gaussian Mixture Models (GMM) are employed to showcase clustering, chosen for their robustness to initialization compared to k-means [49]. In this paper, the GMM is trained to provide two clusters using a dataset containing normal and abnormal samples. By following Fig. 4, a trained encoder $E_i$ encodes normal and abnormal samples into the respective representations. Then, class labels are predicted and compared with those from the specified hard threshold to assess clustering performance. The training process is repeated according to the total number of IMFs. Based on the clustering performance achieved for all the EMD levels, we identify the optimal level as the one that yields the highest clustering performance. Algorithm 1 presents a pseudo-algorithm detailing the steps involved in the proposed collaborative optimization.

The effectiveness of the proposed methodology is evaluated based on its performance in representation learning and clustering. The performance of representation learning is measured via a signal reconstruction and a generation of discriminative features to distinguish between normal and abnormal conditions. The former aspect is quantitatively assessed through the reconstruction loss, $\mathcal{L}_{\text{rec}}$, defined in (13). To assess the effectiveness of the method in generating discriminative features, we measure the dissimilarity between normal and abnormal samples obtained from clustering analysis, in which an average dissimilarity ($AD$) is exploited for this purpose. It is calculated from the average distance across all pairs of normal and abnormal samples. In this paper, an Euclidean distance metric is considered. A greater dissimilarity indicates a greater distinction between the normal and abnormal samples. The average dissimilarity based on the Euclidean distance is expressed as:

$$AD_i(z_{n,i}(t), z'_{k,i}(t))$$

$$= \frac{1}{N \cdot N'} \sum_{n=1}^{N} \sum_{k=1}^{N'} \sqrt{\sum_{j=1}^{\beta'} (z_{n,i}(t_j) - z'_{k,i}(t_j))^2}, \quad (16)$$

where $z_{n,i}(t_j)$ and $z'_{k,i}(t_j)$ denote the $j^{\text{th}}$ dimension of the latent representation at the EMD level $i$ of the $n^{\text{th}}$ normal and $k^{\text{th}}$ abnormal sample. The total number of normal and abnormal samples are denoted by $N$ and $N'$, respectively, and $\beta'$ is the number of dimensions of the latent representation.

and (12) that allows mapping input signals onto the latent representation such that the reconstruction loss is minimized, and, lastly, an identification of the optimal EMD level of (9) for anomaly detection of rail infrastructures via binary clustering.

For a given hard threshold, consider a set of vibration data $\mathcal{D} = \{x_m(t)\}_{m=1}^{M}$ and a set of normal data $\mathcal{N} = \{x_n(t)\}_{n=1}^{N}$. We define a set of abnormal data $\mathcal{N}'$ as a subset of $\mathcal{D}$ that remains after removing the data present in $\mathcal{N}$. This set $\mathcal{N}'$ contains abnormal data, including data with uncertain abnormalities, in which it is defined as

$$\mathcal{N}' = \mathcal{D} - \mathcal{N} = \left\{ x'_k(t) \in \mathcal{D} \mid x'_k(t) \notin \mathcal{N} \right\}_{k=1}^{N'}.$$

For a given EMD level $i$ and the corresponding set of demodulated signals $\widehat{\mathcal{N}}_i = \{\hat{x}_{n,i}(t)\}_{n=1}^{N}$, the reconstruction technique is considered to train CAE relying on learning representations from $\widehat{\mathcal{N}}_i$ to capture its common and essential features. Then, the CAE is trained on $\widehat{\mathcal{N}}_i$ to obtain the latent representation by minimizing a loss function that measures the discrepancy between the input data and its reconstruction. In this work, a loss function is defined as a combination of mean squared error (MSE), $\mathcal{L}_{\text{MSE}}$, and KL divergence, $\mathcal{L}_{\text{KL}}$, as our reconstruction loss, $\mathcal{L}_{\text{rec}}$, and it is expressed as:

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{KL}}, \quad (13)$$

where

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{n=1}^{N} \left( \hat{x}_{n,i}(t) - \tilde{x}_{n,i}(t) \right)^2, \quad (14)$$

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2} \sum_{j} \left( 1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2 \right)$$

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

PHUSAKULKAJORN et al.: UNSUPERVISED REPRESENTATION LEARNING FOR MONITORING RAIL INFRASTRUCTURES

7

**Algorithm 1** Collaborative Optimization Procedure

1: **Input:** Normal set ($x(t) \in \mathcal{N}$) and its labels ($Y$) obtained from a hard threshold, abnormal set ($x'(t) \in \mathcal{N}'$) and its labels ($Y'$) obtained from a hard threshold, the maximum number of IMFs ($I$)

2: **Output:** Optimal level of EMD ($i_{opt}$), clustering performance, dissimilarity between $\mathcal{N}$ and $\mathcal{N}'$

3: Initialize EMD level $i \leftarrow 1$

4: **repeat**

5:    // Step 1: EMD on $x(t) \in \mathcal{N}$

6:    $\hat{x}_i(t) \leftarrow H_i(x(t))$ %IMF at level $i$

7:    Initialize number of epoch $e \leftarrow 1$

8:    **repeat**

9:      // Step 2: CAE Training on $\hat{x}_i(t) \in \widehat{\mathcal{N}}_i$

10:      Train CAE using cross-validation on $\hat{x}_i(t)$

11:      $\tilde{x}_i(t) \leftarrow D_i(E_i(\hat{x}_i(t)))$ % Reconstruction at level $i$

12:      Compute reconstruction loss $\mathcal{L}_{rec}(\hat{x}_i(t), \tilde{x}_i(t))$

13:      $e \leftarrow e + 1$

14:    **until** $e \geq$ Maximum epoch

15:    // Step 3: Obtain Latent Representations

16:    $E_i \leftarrow$ Trained encoder at level $i$

17:    $\hat{x}'_i(t) \leftarrow H_i(x'(t))$ %EMD on $x'(t) \in \mathcal{N}'$

18:    Latent representations for normal samples: $z_i(t) \leftarrow E_i(\hat{x}_i(t))$

19:    Latent representations for abnormal samples: $z'_i(t) \leftarrow E_i(\hat{x}'_i(t))$

20:    // Step 4: Binary Clustering Analysis

21:    Perform binary clustering on $z_i(t)$ and $z'_i(t)$

22:    Obtain predicted class labels $\widehat{Y}_i$ and $\widehat{Y}'_i$

23:    Compute clustering performance at EMD level $i$

24:    // Step 5: Compute Average Dissimilarity

25:    Compute average dissimilarity $AD_i(z_i(t), z'_i(t))$

26:    $i \leftarrow i + 1$

27: **until** $i > I$

28: // Step 6: Determine Optimal EMD Level

29: $i_{opt} \leftarrow \arg\max_i \big($clustering performance at EMD level $i$, average dissimilarity at EMD level $i\big)$

30: **return** Optimal level of EMD $i_{opt}$, clustering performance, average dissimilarity

---

Compared against labels obtained from the hard threshold, the clustering performance is assessed in terms of precision, recall, and F1 score, which are expressed as:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{17}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{18}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{19}$$

where, in the context of anomaly detection, true positives (TP) are the number of correctly classified abnormalities, false positives (FP) are the number of normal samples classified as abnormalities, and False Negatives (FN) are the number of abnormalities classified as normal.
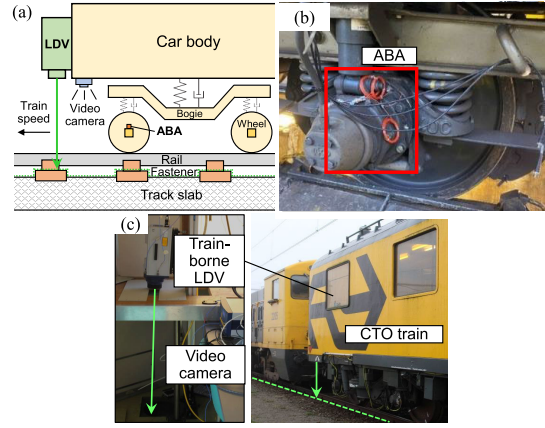


Fig. 5. Train-borne sensing technology. (a) Illustration; Implementation of (b) ABA measurement in Sweden and (c) train-borne LDV on the TU Delft CTO measurement train in the Netherlands.

## V. CASE STUDIES

To demonstrate the proposed methodology, we present two case studies from field measurements conducted on operational rail lines in Sweden and the Netherlands. In the two case studies, different track components are monitored using different sensing technologies applied on different operational conditions.

### A. Case 1: Monitoring Rail Surfaces With ABA

ABA is a measurement technology that utilizes accelerometers attached to the axle boxes of moving trains to measure dynamic responses due to wheel-rail contact. Rail component conditions and defects can be assessed by examining deviations in their responses. ABA has been successfully tested in various countries to assess the conditions of various railway components, e.g., fasteners, rails, insulated joints, transition zones, and crossings [14], [22], [24], [50]. This paper uses ABA technology to monitor rail surfaces and applies the proposed unsupervised representation learning for anomaly detection, where the method learns the normal behavior from the ABA data and helps detect abnormalities for further inspection.

Fig. 5 illustrates a setup of the ABA measurement system. The ABA data used in this paper are collected from the Iron Ore line between Luleå, Sweden, and Narvik, Norway. It is a single-track line with passenger-freight mixed traffic and heavy axle load (up to 31 t). This paper acquires information from accelerometers installed in the longitudinal direction as signals obtained from the longitudinal direction have proven effective in capturing early-stage characteristics of defects [22]. Additionally, the information is collected from both the left and right wheels of all axles. The obtained measurements contain various rail dynamics, including healthy rails, welds, insulated joints, switches, and rail surface defects. This paper assumes that the locations of insulated joints and switches are known. Hence, the signals at these locations are excluded from the analysis.

As we primarily focus on rail surface defects, the sensitivity analysis for segmentation is conducted to ensure the signals
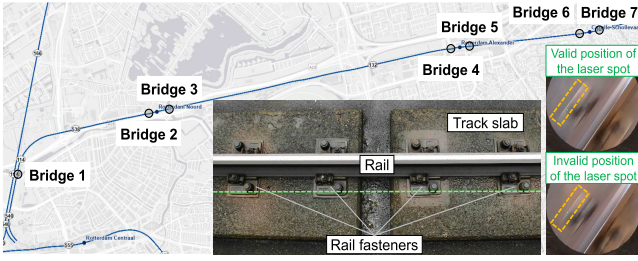
Fig. 6. Targeted railway track sections on the Rotterdam-Gouda line in the Netherlands (Map: from ProRail).

capture their dynamic responses. In this work, the ABA signals are aggregated into smaller segments, each containing 2000 datapoints, resulting in a total of 583 rail segments. With a hard threshold set to 60 m/s$^2$, we identify rail segments with ABA responses lower than the specified threshold for representation learning. These segments are assigned into a class of samples representing normal conditions, referred to as Class 0, containing 362 samples. Using information from fieldwork, we identify the defective samples containing rails with visible defects. These comprise 85 defective rail segments, referred to as Class 1. This leaves 136 ABA samples whose classification is ambiguous under the hard threshold, and they are grouped into Class U. A summary of the ABA dataset is presented in Table II.

### B. Case 2: Monitoring Rail Fasteners With Train-Borne LDV

While ABA monitors rails indirectly through train vibrations, in contrast, train-borne LDV is an innovative technology that offers the ability to measure track vibrations from a moving train directly by emitting a laser beam downward onto the track [17], [18], [19]. As illustrated in Fig. 5(a), an LDV is mounted on a train emitting a laser beam downward onto the track. As the train moves, the laser spot scans the track surface and measures its vibration velocity contactlessly. Different track components can be targeted, such as rails [18], [19] and sleepers [17]. In this case study, LDV is exploited to target rail fasteners, which are critical components in railway tracks to provide connection, constraint, and vibration reduction between rails and sleepers or track slabs.

The open-path scanning of a train-borne LDV may provide invalid measurements on some fasteners due to the lateral shift of the laser spot out of the fastener surface (see Fig. 6) or due to poor surface quality. An example of such an invalid measurement is shown in the second example of Fig. 2. Considering the large number of rail fasteners in a rail network, it is impractical to manually check the signal segment of each fastener and label it as valid or invalid. Meanwhile, the diversity of fastener vibrations and the presence of speckle noise induce significant fuzziness in distinguishing between valid and invalid signal segments, especially for those with small vibration amplitude, such as the third example in Fig. 2. Therefore, the developed unsupervised learning method is applied in this case study to extract useful features that characterize the rail fastener vibrations autonomously.

The LDV data considered in this case study are obtained from the rail fasteners of the same type on the slab tracks of seven bridges in the Rotterdam-Gouda line in the Netherlands, as shown in Fig. 6. On each bridge, the train-borne LDV scans these rail fasteners in a row and measures their vertical vibration individually. In total, 610 rail fasteners are scanned at the train speed of 45-75 km/h. The sampling frequency of the LDV is 102.4 kHz. Then, the LDV signal is cut into segments according to (5)-(7), in which $t_{b_r}$ and $t_{e_r}$ are the starting and ending time of the laser spot scanning each fastener (recorded by the camera). As a result, 610 segments with 2048 datapoints per segment are obtained.

The mean of the FFT spectrum within the frequency range between 200 and 800 Hz is considered for class assignment. For LDV, two hard thresholds are used. We identify LDV data at rail fasteners as valid if the mean exceeds one threshold. These result in 211 samples that are assigned to Class 0. The other 258 samples are identified as invalid measurements, referred to as Class 1 if the mean is below the other threshold. The remaining 141 LDV samples are labeled to Class U due to uncertainties and fuzziness, as their mean is between the two thresholds. A summary of the LDV dataset is presented in Table II.

## VI. RESULTS

### A. Implementation Details

Following the procedure depicted in Fig. 4, the vibration signals are first decomposed into IMFs by the EMD algorithm. In this paper, various EMD levels up to a maximum of five are explored. After decomposing the signal into different IMFs, the demodulated signal at each decomposition level is obtained by (9). The CAE structure considered in this paper is symmetric, meaning that the same number of convolutional layers is designed for the encoder and decoder. We experiment with the number of convolutional layers used within the structures in which up to four layers are considered. Four different numbers of filters are used in the trial: 64, 32, 16, and 8. We also experiment with different filter sizes: $3 \times 3$, $5 \times 5$, $7 \times 7$, $9 \times 9$, and $11 \times 11$. Each convolutional layer in the encoder is followed by an activation layer in which the rectifier or ReLU activation function is used in the convolutional layers. A max-pooling layer with stride two is defined for downsampling at the end of the layer. Similarly, the upsampling steps use transposed convolutions with the ReLU function. At each upsampling step, the number of filters is doubled.

The demodulated signals from Class 0 are divided into training and test sets with a ratio of 75:25. Then, we employ signals from the training set to train the CAE while the test set is held out to validate the generalization of the proposed methodology. Five-fold cross-validation is performed in which 90% of the training set is used to train the models, and the other 10% is used for validating the trained model. The Adam with Nesterov momentum optimizer is exploited for the training, and the maximum number of epochs considered is 100. The early stopping is also applied when $\mathcal{L}_{rec}$ has stopped improving for more than five epochs.

To account for fuzziness and disturbance presented in data, all samples are included in binary clustering analysis, with

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

PHUSAKULKAJORN et al.: UNSUPERVISED REPRESENTATION LEARNING FOR MONITORING RAIL INFRASTRUCTURES 9

TABLE II
SUMMARY OF THE DATASETS FROM THE TWO CASE STUDIES

| Cases | Target components | Segment length (datapoints) | Number of segments | Definition of (Number of samples) | | |
|---|---|---|---|---|---|---|
| | | | | Class 0 | Class 1 | Class U |
| ABA | Rail defects | 2000 | 583 | Normal rail (362) | Defective rail (85) | Uncertain condition (136) |
| LDV | Rail fasteners | 2048 | 610 | Valid measurement (211) | Invalid measurement (285) | Uncertain measurement (141) |

the trained CAE generating representations serving as input to the GMM. The GMM is executed multiple times with varying initializations, and the optimal result is selected based on within-cluster variance. Cluster separability is visualized to evaluate clustering results, including samples from Class U that exhibit more uncertainties and fuzziness. The report of clustering performance is based on the consideration of Class 0 and Class 1.

### B. Results of Different EMD Levels

This section investigates the impact of different EMD levels on the proposed method for representation learning, measured by reconstruction loss and average dissimilarity, and clustering performance, measured by F1-score. We showcase using LDV data as they contain more severe and complex noise components (speckle noise) than ABA data. In this analysis, the CAE architecture remains consistent as it learns representations from the demodulated signals at various EMD levels.

Fig. 7 shows the demodulated signals corresponding to various EMD levels obtained from rail fasteners with valid and invalid measurements. Specifically, analysis is conducted on demodulated signals for EMD levels up to 5. As the level of EMD increases, the demodulated signals become smoother, and the prominence of noise and high-frequency components is reduced. This enhances the signal reconstruction capability of the CAE, resulting in representations that can be used for a more accurate reconstruction for Class 0, as evidenced by the decreasing trend of the blue line in Fig. 8(a). This improvement is further shown by the consistency observed between the input signal and its corresponding reconstruction, as illustrated in Fig. 7(a). In contrast, Fig. 7(b) exhibits that a sample from Class 1 is not as effectively reconstructed compared to the sample from Class 0 at the respective EMD level.

Despite the enhancement in signal reconstruction at higher EMD levels, the disparity between the representations of normal and abnormal samples does not proportionally improve. This becomes apparent when evaluating the learned representations at each EMD level through clustering analysis, as shown in Fig. 8(b). It is suggested that the demodulated signal at EMD level 2 demonstrates optimal clustering efficacy as its embedded features are effective at distinguishing between normal and abnormal samples, evident by a high value of F1-score for both Class 0 and 1. The high value of F1-score is critical in real-world scenarios as this ensures reliable identification of abnormalities while minimizing the operational burden of false positives. Additionally, the average dissimilarity at EMD level 2 is promising, ranking second among all five levels, as depicted by the red line in Fig. 8(a). This observation underscores a tradeoff. While increasing the
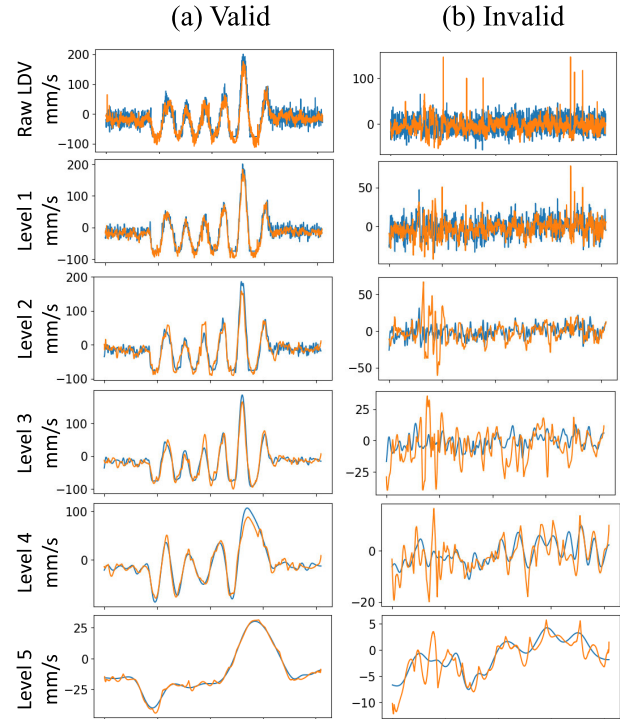


Fig. 7. Examples of LDV signals from two different rail fasteners and their respective demodulated signal at different EMD levels.
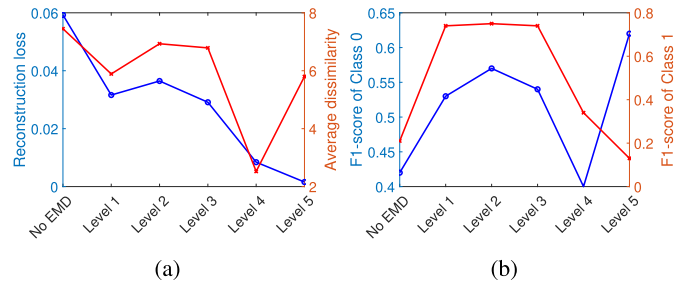


Fig. 8. (a) Effect of the EMD levels on the proposed method for the reconstruction loss and the ability to generate the discriminative features. (b) Effect of the EMD levels on the clustering results measured via F1-score.

level of EMD offers signal denoising, it simplifies vibration patterns, thus affecting the capability of CAE to discriminate between normal and abnormal samples. Achieving an optimal balance between denoising and representation learning capability is thus crucial for accurate anomaly detection. This demonstrates the ability of the proposed method to generate features that improve clustering accuracy under challenging conditions, such as severe speckle noise in LDV data.

In the subsequent sequels, we present the proposed methodology for monitoring rail surfaces with ABA and monitoring
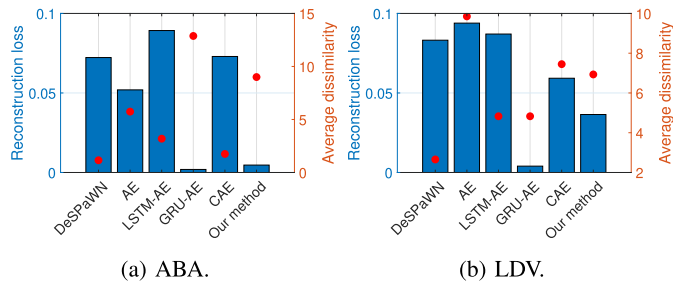
This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

10                                                                                                          IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS

(a) ABA.                                        (b) LDV.

Fig. 9.   Reconstruction loss and average dissimilarity between Class 0 and Class 1 obtained from different methods.



(a) DeSpaWN.                                    (b) AE.



(c) LSTM-AE.                                    (d) GRU-AE.



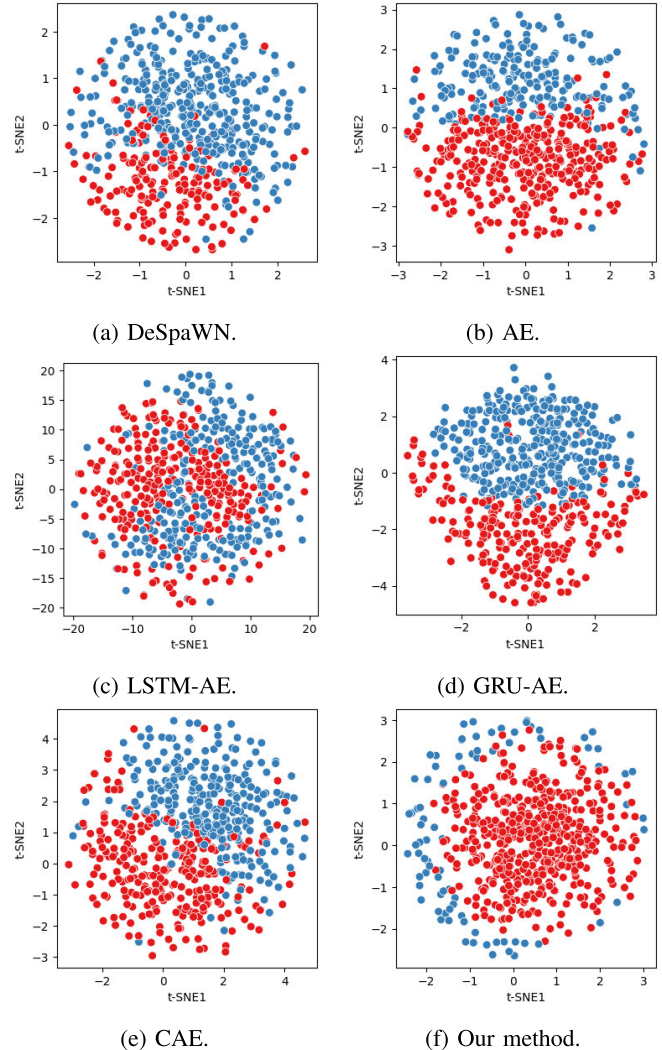(e) CAE.                                        (f) Our method.

Fig. 10.   Clustering results using latent features extracted from different models. The t-SNE with a perplexity of 200 is employed for visualizing clusters of two colors, from which red represents a cluster of Class 0 and blue represents a cluster of Class 1.

rail fasteners with train-borne LDV. For each case study, we provide a comparative analysis for representation learning considering four variants of autoencoder-based models: autoencoder (AE), long short-term memory (LSTM)-AE, gated recurrent unit (GRU)-AE, and CAE, in which suitable experimental configurations are adopted for their development. We also compare with the denoising sparse wavelet network (DeSpaWN) [34], in which the parameter setting and loss are followed from [34]. To be consistent with the development of our model, four maximum encoder and decoder layers are considered for all comparative models.

### C. Results of Rail Surface Defect Detection Using ABA

This section presents results from the representation learning considering the ABA data. It can be seen from Fig. 9(a) that different methods perform differently in signal reconstruction and differentiation of embedded representations between clusters. The GRU-AE method yields the lowest reconstruction loss. It also achieves the highest average dissimilarity, indicating that it generates embedded representations that discriminate between normal and abnormal samples better than the other methods. Our method provides the second-lowest reconstruction loss and average dissimilarity, reflecting its competitive capability for representation learning.

Fig. 10 presents clustering results obtained from using latent features extracted by different methods to train the GMM algorithm to provide two different clusters, considering all 583 ABA samples. The t-SNE [29], [30] using the perplexity of 200 is employed for a visualization of the 2D representations. Note that perplexity is related to the number of nearest neighbors each point considers in the t-SNE algorithm during the dimensionality reduction process. Lower values of perplexity make the algorithm focus on a very local structure, while higher values take into account a broader neighborhood. It is noticeable that the GRU-AE and our method provide better separability between two clusters with fewer overlaps than the others, reflecting their high average dissimilarity obtained. In contrast, the LSTM-AE exhibits more noticeable overlap, corresponding to its highest reconstruction loss and low average dissimilarity. The improved cluster separability provided by our method can assist infrastructure managers by pinpointing defect locations, enabling timely maintenance and reducing reliance on manual inspections.

Next, the clustering results are evaluated using labeled data from Class 0 and Class 1 to investigate the informativeness of the learned representations obtained from each method. Table III presents a comparative study on the clustering tasks using different sets of latent features obtained from different methods. The results show that our method using the demodulated signals at EMD level 2 provides very competitive results. It correctly assigns 96% of Class 0 and correctly assigns 41% of Class 1. While the DeSpaWN exhibits cluster separability with more overlap than our method, it demonstrates the highest accuracy in correctly identifying samples of Class 1, yielding a 38% higher accuracy for this class. However, the DeSpaWN detects Class 0 with 63% lower accuracy compared to our method. Additionally, the GRU-AE method demonstrates the lowest reconstruction loss and exhibits good cluster separability owing to the high average dissimilarity obtained. Nevertheless, our method outperforms GRU-AE in terms of cluster performance. Furthermore, GRU-AE shows a lower precision for Class 1, indicating a higher misclassification rate where samples from Class 0 are incorrectly labeled as Class 1. In contrast, the higher F1-scores achieved by our proposed
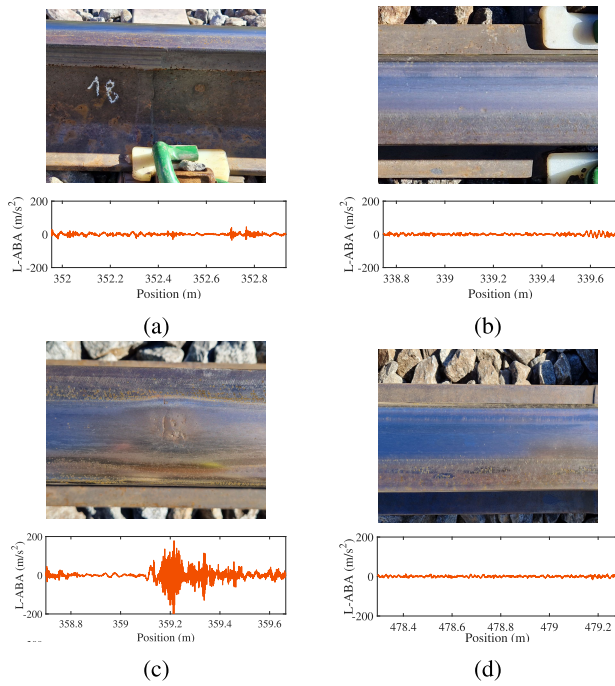
Fig. 11. Examples of validated rail segments at (a) a weld, (b) a small defect, (c) a squat, and (d) a rail without visible defects.

method for both Class 0 and Class 1 highlight its ability to accurately classify rail segments under normal and abnormal conditions, even in the presence of ambiguous ABA responses, as shown in Fig. 11. The superior clustering performance of the proposed method underscores its efficacy in generating discriminative and informative representations of normal and abnormal rail surface conditions using ABA, attributed to the utilization of EMD.

It is noteworthy that the parameter configuration for the DeSpaWN can be further optimized to align with the characteristics of the problem, potentially resulting in improved performance. Furthermore, as the GRU-AE has a complex network (8 hidden layers are exploited), the model might have overfitted on the normal samples during training. Using a larger dataset can be considered to train the GRU-AE, potentially resulting in improved performance.

Next, we examine the effectiveness of our proposed method in handling uncertainties and fuzziness. Figs. 11(a) and 11(b) show two examples from Class U whose ABA responses display ambiguous abnormalities. Consequently, both examples are labeled as Class 1 by the hard threshold for clustering. According to the hard threshold labels, the method correctly identifies Fig. 11(a) to Class 1 and misidentifies Fig. 11(b) to Class 0. Validation against fieldwork information reveals that Fig. 11(a) represents a rail segment at a weld, while Fig. 11(b) is a rail segment at a small defect. However, it can be seen that the ABA signal at the weld in Fig. 11(a) shares similar characteristics with the ABA signal at squat in Fig. 11(c). This suggests that the weld at this rail segment is in poor condition. The weld affects rail vibration, which ABAs capture. Further experiments can be conducted to confirm the condition. Similarly, the ABA signal of the rail with a small defect in Fig. 11(b) resembles that of the normal rails shown in
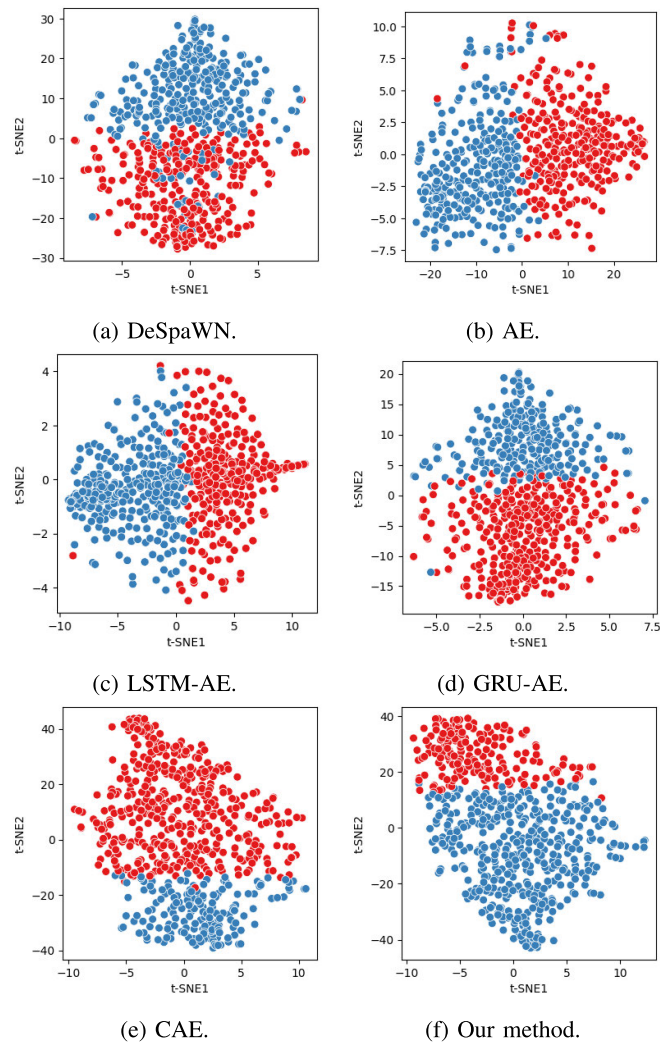


Fig. 12. Clustering results using latent features extracted from different models. The t-SNE with a perplexity of 30 is employed for visualizing clusters of two colors, from which red represents a cluster of Class 0 and blue represents a cluster of Class 1.

Fig. 11(d). These findings demonstrate that the latent features obtained from our method are informative for grouping rail dynamic responses with similar characteristics.

### D. Results of Monitoring Rail Fasteners With LDV

This section presents results from the representation learning considering the LDV data. Similar to ABA, Fig. 9(a) shows that different methods perform differently in signal reconstruction and generation of embedded features. The GRU-AE method yields the lowest reconstruction loss and generates the embedded representations that provide a good average dissimilarity between normal and abnormal samples. Our method provides the second-lowest reconstruction loss, reflecting its competitive capability for representation learning. Meanwhile, it has higher dissimilarity than the GRU-AE, reflecting its capability for clustering using the learned features. Similar to the ABA case study, this highlights the use of EMD in the proposed method.

Fig. 12 presents clustering results obtained from using latent features extracted by different methods. The t-SNE using the

TABLE III

COMPARISON RESULTS OF CLUSTERING PERFORMANCE FOR ABA DATA

| Method | Class 0 | | | Class 1 | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1-score | Precision | Recall | F1-score |
| DeSpaWN [34] | 0.87 | 0.33 | 0.48 | 0.22 | **0.79** | 0.34 |
| AE | 0.85 | 0.43 | 0.57 | 0.22 | 0.68 | 0.33 |
| LSTM-AE | 0.82 | 0.52 | 0.64 | 0.20 | 0.49 | 0.28 |
| GRU-AE | 0.84 | 0.52 | 0.65 | 0.22 | 0.56 | 0.31 |
| CAE | 0.81 | 0.62 | 0.70 | 0.19 | 0.36 | 0.25 |
| **Our method** | **0.87** | **0.96** | **0.91** | **0.69** | 0.41 | **0.51** |

NB: Better performances are highlighted in bold for each clustering algorithm.

TABLE IV

COMPARISON RESULTS OF CLUSTERING PERFORMANCE FOR LDV DATA

| Method | Class 0 | | | Class 1 | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1-score | Precision | Recall | F1-score |
| DeSpaWN [34] | 0.61 | 0.64 | 0.62 | 0.69 | 0.66 | 0.67 |
| AE | 0.37 | 0.40 | 0.38 | 0.47 | 0.44 | 0.46 |
| LSTM-AE | 0.39 | 0.41 | 0.40 | 0.49 | 0.47 | 0.48 |
| GRU-AE | 0.50 | **0.65** | **0.57** | 0.62 | 0.47 | 0.54 |
| CAE | 0.34 | 0.53 | 0.42 | 0.30 | 0.16 | 0.21 |
| **Our method** | **0.71** | 0.47 | **0.57** | **0.65** | **0.90** | **0.75** |

NB: Better performances are highlighted in bold for each clustering algorithm.



Fig. 13. Examples of LDV measurements that are identified as valid measurements by our methodology but identified as invalid by the prediction using the handcrafted features.
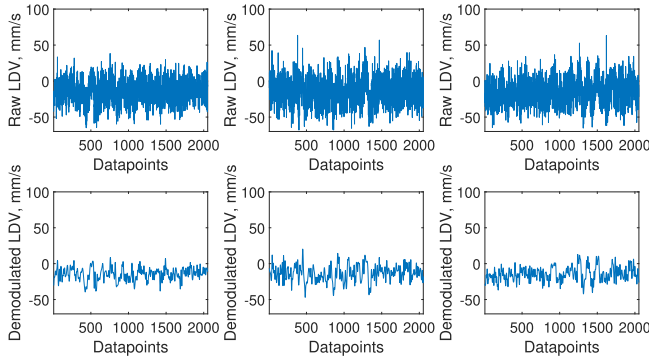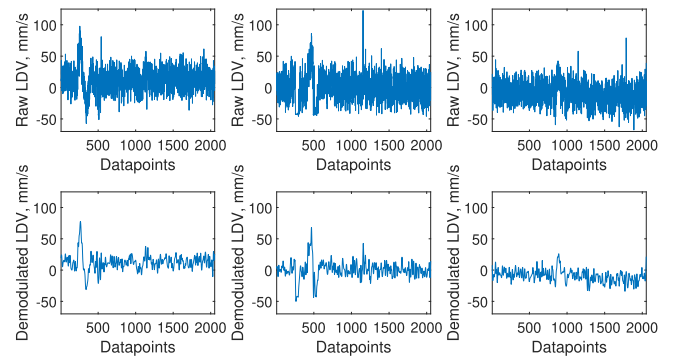


Fig. 14. Examples of LDV measurements that are identified as invalid measurements by our methodology but identified as valid by the prediction using the handcrafted features.

perplexity of 30 is employed for a visualization of the 2D representation of the encoded features obtained from each method. Despite different reconstruction loss and dissimilarity obtained, the latent features obtained from all methods provide a clear separation between a cluster of two colors representing a cluster of Class 0 and Class 1, with some minor overlaps seen more from the DeSpaWN than the other methods.

Evaluating against labeled data, Table IV shows that our proposed methodology using the demodulated signals from the EMD level 2 demonstrates competitive results for both Class 0 and Class 1. It correctly assigns 47% of Class 0 and correctly assigns 90% of Class 1. Although the GRU-AE method yields the least reconstruction loss and provides good cluster separability due to high average dissimilarity, it demonstrates a 43% lower accuracy in correctly identifying samples of Class 1 compared to our methodology. However, the GRU-AE achieves 18% higher detection accuracy for Class 0 than our method. The GRU-AE method also exhibits a higher rate of incorrectly identifying an LDV signal as invalid

when it is valid. Mitigating these false positives is crucial as false positives impair the learning of actual trends and patterns in rail fastener health. By reducing false alarms, operators can identify issues early and proactively address them before they escalate.

To examine the effectiveness of our proposed method in handling samples of Class U from LDV data, we compare the clustering results obtained from the latent features learned by our method with those obtained from a hand-crafted method. The hand-crafted method, as outlined in [51], involves the manual design of three features: the variations in the time and frequency domains as well as the power spectrum density entropy of each raw signal segment (without denoising). Various clustering algorithms are implemented, including k-means, k-medoids, and fuzzy c-means. For each algorithm, clustering with multiple clusters is used to handle the diverse patterns observed in the valid measurements and then merge them into a single group. The optimal number of clusters is

tuned for each algorithm. More details of the feature design and clustering analysis can be found in [51].

Out of the 141 samples of Class U, 10 (or 20) samples are labeled as valid measurements, while 131 (or 121) samples are labeled as invalid measurements when using the hand-crafted method in [51] (or the proposed method). The ratio between the two labels is similar between the two methods despite their significant differences in the feature design as well as the clustering algorithm. The two methods consistently label 115 out of 141 samples (81.56%), whereas the remaining is labeled differently. Figs. 13 and 14 show several samples with inconsistent clustering results. The LDV samples shown in Fig. 13 are labeled as invalid using the hand-crafted method in [51] but as valid using the proposed method. It can be seen that these samples on top indeed carry vibration patterns (evident by the demodulated LDV), while the severe speckle noise makes them less pronounced. Owing to the use of EMD for denoising, the proposed method captures these vibration patterns more effectively, thus providing more reasonable labels compared to the hand-crafted method. The samples shown in Fig. 14 are labeled as valid by the hand-crafted method, most likely due to the local and sudden variations in the signals, which are actually different from the targeted vibration patterns that are more stationary and continuous. The proposed method identifies such differences and avoids such signals being labeled as valid. These typical examples showcase the effectiveness of the developed representation learning framework in capturing the targeted vibration pattern while reducing the disturbance of the noise.

## VII. Conclusion

Current railway infrastructure management relies on manual inspections and automated methods, both limited by human error and the scarcity of labeled data, which hinders effective representation learning. Our methodology addresses this by employing unsupervised representation learning using high-frequency data from moving sensors. The methodology synchronizes the empirical mode decomposition (EMD) with a convolutional autoencoder (CAE). By testing with the ABA measurements from the Swedish rail network and the LDV measurements from the Dutch rail network, the proposed methodology demonstrates a promising performance for unsupervised rail defect and rail fastener analysis. Evaluating the obtained representations using the Gaussian mixture model clustering, cluster separability with minor overlaps is achieved for both application cases. This proves the effectiveness of the method in generating features that differentiate between normal and abnormal samples, even the inherent fuzziness and disturbance present in the ABA and LDV data. As representations are learned in an unsupervised manner, the methodology reduces the dependency on labeled data. Verified against labels from a hard threshold, it demonstrates an improvement in detection accuracy compared to other variants of autoencoder-based models and the wavelet-based CAE, achieving a 16% increase with ABA data and a 21% increase with LDV data. Furthermore, the latent features obtained from the proposed method have been proven to be informative. In the case of ABA data, clusters of rail segments with similar

characteristics can be used to guide the inframanager about the locations of defects. For LDV data, clusters of rail segments with similar characteristics can be used to learn trends and patterns in rail fastener health. The success of the proposed method highlights the importance of EMD for denoising, enhancing representation learning of rail infrastructure characteristics and reducing noise interference. Consequently, our methodology enables continuous and autonomous monitoring over long distances during train operations, which, in turn, minimizes inspection time, service disruptions, and maintenance costs, while enhancing detection accuracy.

Future research includes improving cluster separability through advanced clustering techniques and hybrid approaches to better distinguish between normal and abnormal samples. Incorporating additional data sources and measurements can be considered with a development of data fusion techniques to provide a comprehensive analysis and enhance accuracy and reliability of rail infrastructure conditions. Examining scalability and real-time processing capabilities can be considered for enabling real-time analysis of large-scale rail network data. Validating the methodology across different rail networks with varying environmental conditions and operational patterns is also essential to ensure its robustness and applicability.

## References

[1] Y.-W. Wang, Y.-Q. Ni, and S.-M. Wang, "Structural health monitoring of railway bridges using innovative sensing technologies and machine learning algorithms: A concise review," *Intell. Transp. Infrastruct.*, vol. 1, p. 009, Sep. 2022.

[2] J. M. W. Brownjohn, A. De Stefano, Y.-L. Xu, H. Wenzel, and A. E. Aktan, "Vibration-based monitoring of civil infrastructure: Challenges and successes," *J. Civil Struct. Health Monitor.*, vol. 1, nos. 3–4, pp. 79–95, Dec. 2011.

[3] D. Goyal and B. S. Pabla, "The vibration monitoring methods and signal processing techniques for structural health monitoring: A review," *Arch. Comput. Methods Eng.*, vol. 23, no. 4, pp. 585–594, Dec. 2016.

[4] S. M. Khan, S. Atamturktur, M. Chowdhury, and M. Rahman, "Integration of structural health monitoring and intelligent transportation systems for bridge condition assessment: Current status and future direction," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 8, pp. 2107–2122, Aug. 2016.

[5] C.-H. Ho, M. Snyder, and D. Zhang, "Application of vehicle-based sensing technology in monitoring vibration response of pavement conditions," *J. Transp. Eng., B, Pavements*, vol. 146, no. 3, Sep. 2020, Art. no. 04020053.

[6] I. Celiński, R. Burdzik, J. Młyńczak, and M. Kłaczyński, "Research on the applicability of vibration signals for real-time train and track condition monitoring," *Sensors*, vol. 22, no. 6, p. 2368, Mar. 2022.

[7] J. M. Castillo-Mingorance, M. Sol-Sánchez, F. Moreno-Navarro, and M. C. Rubio-Gámez, "A critical review of sensors for the continuous monitoring of smart and sustainable railway infrastructures," *Sustainability*, vol. 12, no. 22, p. 9428, Nov. 2020.

[8] V. J. Hodge, S. O'Keefe, M. Weeks, and A. Moulds, "Wireless sensor networks for condition monitoring in the railway industry: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1088–1106, Jun. 2015.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

14

IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS

[9] Y. Zeng, C. Shen, A. Núñez, R. Dollevoet, W. Zhang, and Z. Li, "An interpretable method for operational modal analysis in time-frequency representation and its applications to railway sleepers," *Struct. Control Health Monitor.*, vol. 2023, pp. 1–26, Jul. 2023.

[10] J. Zhang, W. Huang, W. Zhang, F. Li, and Y. Du, "Train-induced vibration monitoring of track slab under long-term temperature load using fiber-optic accelerometers," *Sensors*, vol. 21, no. 3, p. 787, Jan. 2021.

[11] G. Guo, X. Cui, and B. Du, "Random-forest machine learning approach for high-speed railway track slab deformation identification using track-side vibration monitoring," *Appl. Sci.*, vol. 11, no. 11, p. 4756, May 2021.

[12] I. La Paglia, M. Carnevale, R. Corradi, E. Di Gialleonardo, A. Facchinetti, and S. Lisi, "Condition monitoring of vertical track alignment by bogie acceleration measurements on commercial high-speed vehicles," *Mech. Syst. Signal Process.*, vol. 186, Mar. 2023, Art. no. 109869.

[13] H. Jiang and J. Lin, "Fault diagnosis of wheel flat using empirical mode decomposition-Hilbert envelope spectrum," *Math. Problems Eng.*, vol. 2018, pp. 1–16, Dec. 2018.

[14] C. Hoelzl et al., "Fusing expert knowledge with monitoring data for condition assessment of railway welds," *Sensors*, vol. 23, no. 5, p. 2672, Feb. 2023.

[15] N. Traquinho et al., "Damage identification for railway tracks using onboard monitoring systems in in-service vehicles and data science," *Machines*, vol. 11, no. 10, p. 981, Oct. 2023.

[16] A. De Rosa, S. Alfi, and S. Bruni, "Estimation of lateral and cross alignment in a railway track based on vehicle dynamics measurements," *Mech. Syst. Signal Process.*, vol. 116, pp. 606–623, Feb. 2019.

[17] Y. Zeng, A. Núñez, and Z. Li, "Railway sleeper vibration measurement by train-borne laser Doppler vibrometer and its speed-dependent characteristics," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 39, no. 16, pp. 2408–2426, Aug. 2024.

[18] C. Yang, K. Kaynardag, and S. Salamone, "Missing rail fastener detection based on laser Doppler vibrometer measurements," *J. Nondestruct. Eval.*, vol. 42, no. 3, p. 68, Sep. 2023.

[19] K. Kaynardag, C. Yang, and S. Salamone, "A rail defect detection system based on laser Doppler vibrometer measurements," *NDT E Int.*, vol. 137, Jul. 2023, Art. no. 102858.

[20] A. Rodríguez, R. Sañudo, M. Miranda, A. Gómez, and J. Benavente, "Smartphones and tablets applications in railways, ride comfort and track quality. Transition zones analysis," *Measurement*, vol. 182, Sep. 2021, Art. no. 109644.

[21] A. Azzoug and S. Kaewunruen, "RideComfort: A development of crowdsourcing smartphones in measuring train ride quality," *Frontiers Built Environ.*, vol. 3, p. 3, Feb. 2017.

[22] M. Molodova, Z. Li, A. Núñez, and R. Dollevoet, "Automatic detection of squats in railway infrastructure," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 1980–1990, Oct. 2014.

[23] Y. Zeng, A. Núñez, and Z. Li, "Speckle noise reduction for structural vibration measurement with laser Doppler vibrometer on moving platform," *Mech. Syst. Signal Process.*, vol. 178, Oct. 2022, Art. no. 109196.

[24] S. Unsiwilai, L. Wang, A. Núñez, and Z. Li, "Multiple-axle box acceleration measurements at railway transition zones," *Measurement*, vol. 213, May 2023, Art. no. 112688.

[25] W. Phusakulkajorn et al., "Artificial intelligence in railway infrastructure: Current research, challenges, and future opportunities," *Intell. Transp. Infrastruct.*, vol. 2, pp. 1–24, May 2023.

[26] A. Lasisi and N. Attoh-Okine, "An unsupervised learning framework for track quality index and safety," *Transp. Infrastruct. Geotechnol.*, vol. 7, no. 1, pp. 1–12, Mar. 2020.

[27] Q. He, Z. Liu, Q. Wang, M. Zhang, and P. Wang, "Anomaly detection of high-speed railway inspection images based on improved skip-GAN," *Tiedao Xuebao/J. China Railway Soc.*, vol. 46, no. 9, pp. 121–128, Jan. 2024.

[28] Y. Zhang, Y. Cheng, T. Xu, G. Wang, C. Chen, and T. Yang, "Fault prediction of railway turnout systems based on improved sparse auto encoder and gated recurrent unit network," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 12711–12723, Aug. 2022.

[29] S. Dong, W. Wu, K. He, and X. Mou, "Rolling bearing performance degradation assessment based on improved convolutional neural network with anti-interference," *Measurement*, vol. 151, Feb. 2020, Art. no. 107219.

[30] F. Jia, Y. Lei, S. Xing, and J. Lin, "A method of automatic feature extraction from massive vibration signals of machines," in *IEEE Int. Instrum. Meas. Technol. Conf. Proc.*, May 2016, pp. 1–6.

[31] K. Maes, L. Van Meerbeeck, E. P. B. Reynders, and G. Lombaert, "Validation of vibration-based structural health monitoring on retrofitted railway bridge KW51," *Mech. Syst. Signal Process.*, vol. 165, Feb. 2022, Art. no. 108380.

[32] Z. Wei et al., "Density-based affinity propagation tensor clustering for intelligent fault diagnosis of train bogie bearing," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 6, pp. 6053–6064, Jun. 2023.

[33] R. Ghiasi, M. A. Khan, D. Sorrentino, C. Diaine, and A. Malekjafarian, "An unsupervised anomaly detection framework for onboard monitoring of railway track geometrical defects using one-class support vector machine," *Eng. Appl. Artif. Intell.*, vol. 133, Jul. 2024, Art. no. 108167.

[34] G. Michau, G. Frusque, and O. Fink, "Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series," *Proc. Nat. Acad. Sci. USA*, vol. 119, no. 8, pp. 1–10, Feb. 2022.

[35] S. Li, L. Jin, J. Jiang, H. Wang, Q. Nan, and L. Sun, "Looseness identification of track fasteners based on ultra-weak FBG sensing technology and convolutional autoencoder network," *Sensors*, vol. 22, no. 15, p. 5653, Jul. 2022.

[36] Z. Ye, S. Yue, P. Yang, R. Zhou, and J. Yu, "Deep morphological shrinkage convolutional autoencoder-based feature learning of vibration signals for gearbox fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–12, 2024.

[37] J. Yu and X. Zhou, "One-dimensional residual convolutional autoencoder based feature learning for gearbox fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 10, pp. 6347–6358, Oct. 2020.

[38] T. Kim and S. Lee, "A novel unsupervised clustering and domain adaptation framework for rotating machinery fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 19, no. 9, pp. 9404–9412, Sep. 2023.

[39] J. Dai, J. Wang, W. Huang, J. Shi, and Z. Zhu, "Machinery health monitoring based on unsupervised feature learning via generative adversarial networks," *IEEE/ASME Trans. Mechatronics*, vol. 25, no. 5, pp. 2252–2263, Oct. 2020.

[40] Y. Lei, F. Jia, J. Lin, S. Xing, and S. X. Ding, "An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data," *IEEE Trans. Ind. Electron.*, vol. 63, no. 5, pp. 3137–3147, May 2016.

[41] R. Wang, F. Liu, X. Hu, and J. Chen, "Unsupervised mechanical fault feature learning based on consistency inference-constrained sparse filtering," *IEEE Access*, vol. 8, pp. 172021–172033, 2020.

[42] Q. He, J. Zhao, G. Jiang, and P. Xie, "An unsupervised multiview sparse filtering approach for current-based wind turbine gearbox fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 8, pp. 5569–5578, Aug. 2020.

[43] A. Mosleh, A. Meixedo, D. Ribeiro, P. Montenegro, and R. Calçada, "Automatic clustering-based approach for train wheels condition monitoring," *Int. J. Rail Transp.*, vol. 11, no. 5, pp. 639–664, Sep. 2023.

[44] W. Mao, Y. Wang, L. Kou, and X. Liang, "A new deep tensor autoencoder network for unsupervised health indicator construction and degradation state evaluation of metro wheel," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–15, 2023.

[45] W. Mao, Y. Wang, K. Feng, L. Kou, and Y. Zhang, "SWDAE: A new degradation state evaluation method for metro wheels with interpretable health indicator construction based on unsupervised deep learning," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–13, 2024.

[46] N. E. Huang et al., "The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London A*, vol. 454, pp. 903–995, Mar. 1998.

[47] P. Flandrin, G. Rilling, and P. Goncalves, "Empirical mode decomposition as a filter bank," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 112–114, Feb. 2004.

[48] A. Ayenu-Prah, N. Attoh-Okine, and N. E. Huang, "Empirical mode decomposition and the Hilbert–Huang transform," in *Transforms and Applications Handbook*, 3rd ed., Boca Raton, FL, USA: CRC Press, 2010, pp. 20-1–20-11.

[49] R. Wang et al., "Transfer-learning-based Gaussian mixture model for distributed clustering," *IEEE Trans. Cybern.*, vol. 53, no. 11, pp. 7058–7070, Nov. 2023.

[50] Z. Wei, A. Núñez, Z. Li, and R. Dollevoet, "Evaluating degradation at railway crossings using axle box acceleration measurements," *Sensors*, vol. 17, no. 10, p. 2236, Sep. 2017.

[51] Y. Zeng, A. Núñez, A. Zoeteman, R. Dollevoet, and Z. Li, "Direct monitoring of in-situ rail fastener vibrations from a moving train with laser Doppler vibrometer," in *Proc. IEEE 27th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2024, pp. 2725–2730.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

PHUSAKULKAJORN et al.: UNSUPERVISED REPRESENTATION LEARNING FOR MONITORING RAIL INFRASTRUCTURES
15

**Wassamon Phusakulkajorn** received the B.Sc. degree (Hons.) in mathematics from the Prince of Songkla University, Songkhla, Thailand, in 2007, and the Ph.D. degree in railway engineering from Delft University of Technology in 2024.

She was a Post-Doctoral Researcher with Delft University of Technology until April 2025. Currently, she is with the National Metal and Materials Technology Center (MTEC), the National Science and Technology Development Agency (NSTDA), Thailand. Her research interests lie in machine learning, big data analytics, data fusion, predictive analytics, and their applications in the condition monitoring and maintenance of railway infrastructure and machinery. She received the Best Paper Award from Oxford University Press in 2023 for her comprehensive review of AI in railway infrastructure.

**Zili Li** received the B.Sc. and M.Sc. degrees in mechanical engineering from Southwest Jiaotong University, Chengdu, China, in 1988 and 1991, respectively, and the Ph.D. degree in computational mechanics on numerical solution of rolling contact from Delft University of Technology, Delft, The Netherlands, in 2002.

From 1999 to 2005, he was with the Institute of Road Transportation, TNO, Delft, where he was involved in research and software development of multibody dynamics and finite element method for crash safety. In 2005, he joined the Faculty of Civil Engineering and Geosciences, Delft University of Technology, where he taught and performed research on railway engineering. He is currently a Full Professor with the Section of Railway Engineering, Delft University of Technology. He is also the Head of the Railway Engineering Section and the Scientific Director of the TU Delft Rail Institute (DelftRail). His current research interests include health monitoring and asset management of railway infrastructure, numerical solution of frictional rolling contact and its applications to analyses of wear, rolling contact fatigue, vehicle dynamics, and train–track interaction, particularly in the high-frequency/short-wave range and at switches and crossings and friction adhesion between wheel and rail. He is a member of the Editorial Board of the *International Journal of Rail Transportation*, *International Journal of Railway Technology*, and *Advances in Mechanical Engineering*.

**Yuanchen Zeng** (Member, IEEE) received the B.Sc. degree in mechatronic engineering from Zhejiang University, Hangzhou, China, in 2016, the Ph.D. degree in vehicle operation engineering from Southwest Jiaotong University, Chengdu, China, in 2022, and the Ph.D. degree in railway engineering from Delft University of Technology, Delft, The Netherlands, in 2023.

Since 2023, he has been a Post-Doctoral Researcher with the Department of Engineering Structures, Delft University of Technology. His research interests include laser Doppler vibrometer, structural health monitoring, vehicle-track dynamics, signal processing, and prognostics and health management. He was a recipient of European Railway Research Advisory Council (ERRAC) Award for the best Ph.D. dissertation in 2024. This dissertation develops and validates a novel train-borne LDV technology for measuring the vibration and load-response relationship of railway track structures over a wide frequency range.

**Alfredo Núñez** (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Chile, Santiago, Chile, in 2010.

He was a Post-Doctoral Researcher with Delft Center for Systems and Control, Delft University of Technology, and a Visiting Scholar at universities in Slovenia, Italy, Spain, Chile, Colombia, China, and USA. He is currently an Associate Professor with Delft University of Technology, specializing in intelligent railway infrastructures. His research focuses on intelligent transportation systems, railway engineering, and computational intelligence, with over 150 publications. His work has been supported by the Dutch Research Council, ProRail, and European projects. He has held key roles in EU projects and served as an associate editor and a guest editor for leading journals. He has contributed to major transportation conferences as a member of local organizations and as a speaker, delivering over 50 presentations. He has mentored award-winning Ph.D., M.Sc., and Eng.D. students and actively contributes to education, industry collaborations, and academic initiatives in railway infrastructure. He belongs to the Editorial Board of the journals IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, *Applied Soft Computing* (Elsevier), and *Intelligent Transportation Infrastructure* (Oxford Academic); and has been a guest editor of special issues in the journal *Wear* (Elsevier) and *Control Engineering Practice*.