Expert Systems With Applications 207 (2022) 117946

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Supervised contrastive learning-guided prototypes on axle-box accelerations for railway crossing inspections

Julio Silva-Rodríguez [a,*], Pablo Salvador [a], Valery Naranjo [b], Ricardo Insa [a]

[a] Institute of Transport and Territory, Universitat Politècnica de València, Valencia, Spain
[b] Institute of Research and Innovation in Bioengineering, Universitat Politècnica de València, Valencia, Spain

ARTICLE INFO

Keywords:
Dynamic railway surveying
Axle-box accelerations
Crossing wear detection
Deep learning
Supervised contrastive learning

ABSTRACT

Increasing demands on railway structures have led to a need for new cost-effective maintenance strategies in recent years. Current dynamic railway track monitoring systems are usually based on the analysis of axle-box accelerations to automatically detect track singularities and defects. These methods rely on hand-crafted feature extraction and classifiers for different tasks. However, the low performance shown in previous literature makes it necessary to complement these analyses with in-situ inspections. Very recent works have proposed the use of deep learning systems that allow extracting more generalizable features from time–frequency spectrograms. However, the lack of specific public domain datasets and the finite number of track singularities in a railway structure have limited the development of deep learning based systems. In this paper, we propose a method capable of outstanding in low-data scenarios. In particular, we explore the use of supervised contrastive learning to cluster class embeddings nearly in the encoder latent space, which is used during inference for prototypical distance-based class assignment. We provide comprehensive experiments demonstrating the performance of our method in comparison to previous literature for detecting worn-out crossings.

1. Introduction

Railway structures are one of the main components of any country's transportation system. Railway maintenance plays a key role in achieving a high-performance, safe and cost-effective system (Tzanakakis, 2013). The increase in demand for passenger and cargo rail transport services has led to an increase in the maintenance needs of the rail network in recent years. Specifically, European countries invest between 15 and 25 billion euros annually in the maintenance and renewal of these structures (Lidén, 2015). With the advent of the Industry 4.0 paradigm and the development of enabling technologies such as sensing devices and artificial intelligence systems, predictive maintenance has been projected as a promising tool for cost-effective maintenance strategies.

In this work, among the different challenges on railway maintenance, we focus on track surveying. Different technologies have been proposed to support the maintenance process: vision camera-based methods, acoustic recording, laser sensors, etc. (Kouroussis et al., 2015). Among these procedures, the use of axle-box accelerometers have proved to be versatile enough to sense different track irregularities of different wavelengths and occurrence (Chia et al., 2019; Jing et al., 2021; Salvador et al., 2016). Some of its advantages are that this technology is not limited to any field of view, and it is able to perform a dynamic surveying of the direct interaction between the track and the railway. The presence of characteristic track element patterns and their deterioration in axle-box acceleration on time–frequency domain has been extensively studied in previous literature (Salvador et al., 2016). In addition, some models based on hand-crafted feature extraction based on traditional image processing methods and machine learning models have been proposed and used on maintenance practice (Nadarajah et al., 2018). However, the low performance of these methods makes it necessary to supplement these predictions with on-site visual inspections by operators.

The emergence of deep learning has led to an increase of performance of different computer-vision based industrial applications. In particular, very recent works have shown the benefits of using convolutional neural networks (CNNs) for axle-box track surveying characterization (Chellaswamy et al., 2019; Niebling et al., 2020; Yang et al., 2021). Under the supervised learning paradigm, deep learning models have achieved remarkable performance in a wide range of applications. Nevertheless, a main limitation of these models is the large amount of labelled data required for training. These limitations are accentuated in track surveying applications. The absence of domain-specific datasets makes it difficult use pre-trained fine-tuned models

* Corresponding author.
E-mail addresses: jjsilva@upv.es (J. Silva-Rodríguez), pabsalzu@upv.es (P. Salvador), vnaranjo@dcom.upv.es (V. Naranjo), rinsa@tra.upv.es (R. Insa).

https://doi.org/10.1016/j.eswa.2022.117946
Received 26 January 2022; Received in revised form 11 June 2022; Accepted 20 June 2022
Available online 3 July 2022

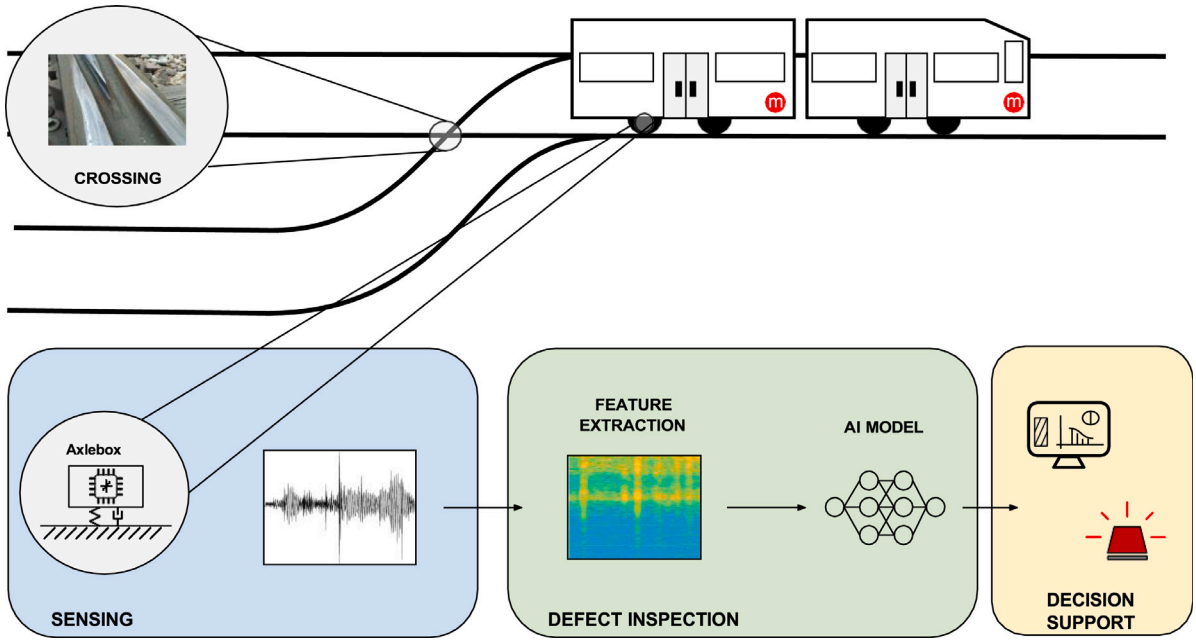

Fig. 1. System overview. In this work, we propose a deep-learning based system able to locate worn crossing on railway surveying maintenance. The sensing technology is based on axle-box vertical accelerations (Section 3.1). First, signals are transformed to time–frequency distributions (Section 3.2). Then, normalized features are used as input to an artificial intelligence model (Section 3.3) to detect worn crossings. The proposed model can be trained on scenarios with scarce training examples. This pipeline can be scaled to other analysis on dynamic railway surveying.

and the annotation process is costly, while the number of track elements is limited (Chenariyan Nakhaee et al., 2019). This encourages the development of novel strategies, capable of withstanding low data scenarios, to achieve robust and reliable automatic systems that may be used in decision making systems for dynamic track surveying.

Based on these observations, in this paper we propose a novel end-to-end system able to detect worn crossings using axle-box accelerations and deep-learning based features via convolutional neural networks (see Fig. 1). The key contributions of our work can be summarized as follows:

- We propose to deal with the scarcity of labelled training data inherent to track surveying applications by means of non-parametric prototypical inference over the feature encoding.
- Specifically, unlike previous work, class embeddings are distributed in the latent space indirectly, using a subspace guided by supervised contrastive losses.
- We compare the proposed system with previous methods in the literature. In-depth experiments demonstrate the superior performance of our approach, with accuracy gains of $\sim 8\%$.
- In addition, we report extensive ablation experiments to provide further insights into feature preprocessing, CNN architectures, and learning strategies in a deep learning-based analysis of axle-box accelerations.

2. Related work

2.1. Railway track surveying

Automatic track surveying is based on pattern analysis over sensed signals and images. Among sensing devices, different technologies such as thermal resistors (Bosso et al., 2018a), acoustic sensors (Chen et al., 2021; Zhang et al., 2017), video recording (Faghih-Roohi et al., 2016; Giben et al., 2015; Gibert et al., 2017; Hovad et al., 2021; James et al., 2019; Mittal & Rao, 2017; Wang et al., 2019; Zhang et al., 2018) or accelerators (Baasch et al., 2019; Bocz et al., 2018; Boogaard et al., 2018; Bosso et al., 2018b; Carrigan et al., 2019; Carrigan & Talbot, 2021; Chang et al., 2021; Chellaswamy et al., 2020, 2019; Ghosh et al., 2021; He et al., 2020; Hory et al., 2012; Li & Shi, 2019; Malekjafarian et al., 2019, 2021; Molodova et al., 2011; Ng et al., 2019; Niebling et al., 2020; Salvador et al., 2016; Song et al., 2020; Sysyn, Gerber, Nabochenko, Li, & Kovalchuk, 2019a; Sysyn, Gruen, Gerber, Nabochenko, & Kovalchuk, 2019b; Wei et al., 2017; Yang et al., 2021) have been proposed. In particular, the use of acceleration sensors on axle-box has become more popular for detecting track irregularities of different wavelengths and occurrence. Concretely, different applications include wheel flat (Bosso et al., 2018b; Sresakoolchai & Kaewunruen, 2021b), crossings monitoring (Sysyn, Gerber, Nabochenko, Li, & Kovalchuk, 2019a; Sysyn, Gruen, Gerber, Nabochenko, & Kovalchuk, 2019b), rail corrugation (Ghosh et al., 2021; Heusel et al., 2021; Li & Shi, 2019), roughness derivation (Carrigan et al., 2019; Carrigan & Talbot, 2021), rail joints (Yang et al., 2021), settlement and dipped joint (Sresakoolchai & Kaewunruen, 2021a) and other railway elements (Salvador et al., 2016, 2018). In the aim of predictive maintenance, first works focused on visual description of the patterns that elements and defects produce on time–frequency domain (Baasch et al., 2019; Carrigan et al., 2019; Carrigan & Talbot, 2021; He et al., 2020; Hory et al., 2012; Molodova et al., 2011; Ng et al., 2019; Salvador et al., 2016, 2018; Song et al., 2020). Among time–frequency distributions, both standard short-time Fourier transform and Wavelets have been used alike. Further on, some works described a set of features based on classic image processing such as peak intensity, frequency-band relative intensity, or other statistics. Then, first classifiers were used on these features, such as SVMs (Li & Shi, 2019) to predict rail corrugation, random forest for railway lifetime prediction (Sysyn, Gerber, Nabochenko, Li, & Kovalchuk, 2019a; Sysyn, Gruen, Gerber, Nabochenko, & Kovalchuk, 2019b), simple costume decision trees for fault detection (Ghosh et al., 2021), or recent neural networks classifiers (Sresakoolchai & Kaewunruen, 2021a, 2021b). Very recent works (Chellaswamy et al., 2020, 2019; Niebling et al., 2020; Yang et al., 2021) have proposed the use of deep learning models via CNNs to characterize acceleration spectrograms on predictive tasks. In line to recent advance on computer vision, these works have perform superior than previous approaches based on hand-crafted feature extraction (Chellaswamy et al., 2020, 2019; Sresakoolchai & Kaewunruen, 2021a;

Yang et al., 2021). Although these works have shown promising results, models are usually trained on small datasets, with scarce labelled data (Chenariyan Nakhaee et al., 2019). On vision camera-based methods, the vast amount of publicly available databases of natural images facilitates the use of previous knowledge for fine-tuning rich, pre-trained models (Mittal & Rao, 2017). Thus, camera-based surveying methods in the literature have been able to successfully train CNNs architecture such as UNets for track segmentation and fault classification (James et al., 2019) or YOLO networks for surface defect localization (Hovad et al., 2021). Nevertheless, time–frequency distribution of acceleration spectrograms are a too specific domain to apply such knowledge. To deal with this issue, different strategies have been proposed. For instance, some works use synthetic data to train CNNs directly on acceleration signals (Sresakoolchai & Kaewunruen, 2021a, 2021b). Still, the reliability of synthetic data is not clear in comparison with in situ data. Other works have resort to self-training strategies such us autoencoders (Niebling et al., 2020), which use unlabelled data to learn rich features. Regarding the CNNs training, the main strategy (Chellaswamy et al., 2019; Sresakoolchai & Kaewunruen, 2021a; Yang et al., 2021) is still the use of standard cross-entropy based supervised training of deep networks, which tend to generalize poorly when trained from scratch on small datasets.

2.2. Learning from limited data

In the context of deep learning, the branch that covers low-data training is few-shot learning. In this scenario the goal is to train a model capable of making predictions that can be generalized to new classes, of which few examples (K-shots) are given during inference. This model, instead of simply characterizing given classes on a standard supervised scenario, should be able to project a feature space from images, where samples from new, unknown concepts, behave similar. Although this setup has gained popularity on recent years, it is sometimes difficult to apply it in real applications, which need to prove its performance when all classes are used during both training and inference. Nevertheless, methods proposed on the few-shot learning paradigm tend also to generalize best on standard supervised scenarios train on very small data, as it is our case. Among different approaches in few-shot learning classification, metric-based methods aim to learn a good embedding space, where novel class samples can be nicely categorized. This categorization has been done learning a deep distance metric on matching (Vinyals et al., 2016) or relational networks (Sung et al., 2018), but also using memory-based nearest neighbour classifier (so-called prototypical networks) based on class-level prototypes via l2 (Euclidean) (Snell et al., 2017) or cosine distance (Chen et al., 2019; Zhang et al., 2020). These methods are trained on an episodic way, where training examples are divided between queries and support to simulate the few labelled examples encountered during inference. Nevertheless, recent works have demonstrated that such training strategy is data-inefficient, and produces detriments in model performance (Laenen & Bertinetto, 2020). Methods that learn to cluster samples in a non-episodic way resemble contrast-based learning methods, which have recently demonstrated leading results on classification tasks in self-training (Chen et al., 2020), and in standard supervised learning (Khosla et al., 2020). In the last case, clusters of points belonging to the same class are pulled together in a hyper-sphere subspace, while simultaneously pushing apart clusters of samples from different classes, in a mini-batch way. In this work, we investigate the use of contrastive learning on low-data scenarios for learning embeddings subsequently used via a prototypical-based inference.

3. Methods

3.1. Data acquisition

In this work, we study the dynamic train-track interaction as a system of masses, springs and dampers. In this model, any significant alteration in any of the elements will affect the rest of the system. Thus, it is possible to survey alterations on railway track status by recording the interaction on later elements of the system. The dynamic surveying of the railway status is performed by means of vertical accelerometers placed on the axle-boxes of the wheelsets for the left and right rails. From this interaction, we intend to train a classifier capable of recognizing whether a crossing is worn or not. Hereafter, we will refer to $x[n]$ as the signal acquired for any of the channels in a given window, which contains a crossing.

3.2. Feature extraction

The recorded signals $x[n]$ on time domain are transformed into the time–frequency spectrograms using the short-time Fourier transform, $X[m,\omega]$ such that:

$$X[m,\omega] = \sum_{n=-\infty}^{N-1} x[n]w[n-m]e^{-j\omega n} \tag{1}$$

where $w[n]$ is a hamming window, with length W samples. Each window, $w[n]$, get chunks of the original signal, overlapped by O to reduce artifacts. Note that, in the following, we refer to $X[m,\omega]$ as X for simplicity.

Then, spectrograms are scaled to improve model convergence and fasten training. Concretely, we propose to use a dynamic-margin normalization of the input spectrogram to ensure that $X \in [0,1]$, and use all the intensity range. This operation is parameterized by the desired dynamic margin in decibels, γ, such that:

$$X' = \frac{20\ \log_{10}(\frac{X}{W/2} + \epsilon) + \gamma}{\gamma} \tag{2}$$

where $\epsilon = 10^{(\frac{-\gamma}{20})}$. In the following, we refer to X' as X for notational simplicity

Feature extraction is applied to axle-box signals from both railways, and their features are concatenated into a two-channel tensor for both model training and inference.

3.3. Supervised contrastive feature learning

An overview of our algorithm for crossing wear detection is presented in Fig. 2. Below, we describe each component proposed for model training and inference.

Let us denote a set of I crossing features $\{X_i\}_{i=1}^{I}$, and their respective labels by $\{y_{i,k}\}_{i=1}^{I}$. Each individual label, $y_{i,k}$, is composed by a one-hot-encoding ground-truth that indicates if that crossing is worn, such that $y_{i,k} \in \{0,1\}$, with $k = \{0,1\}$. We also define an encoder, $f_{\boldsymbol{\theta}}(\cdot) : \mathcal{X} \rightarrow \mathcal{Z}$, parameterized by $\boldsymbol{\theta}$, that is trained to characterize each crossing into an embedding of lower dimensionality D_E, such that $\mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^{D_E}$. Then, we aim to train $f_{\boldsymbol{\theta}}(\cdot)$ such that the embedding representation of normal and anomalous crossings are discriminated. In this line, we propose to use a supervised contrastive strategy. Thus, we define a projection head, $f_{\boldsymbol{\phi}}(\cdot) : \mathcal{Z} \rightarrow \mathcal{R}$, parameterized by $\boldsymbol{\phi}$, which is composed by a two-layered perceptron with relu activations that maps the embedding space to a lower dimensionality, such that $\mathbf{z} \in \mathcal{R} \subset \mathbb{R}^{D_E/F_c}$, with F_c a system hyper-parameter. Then, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are trained via gradient descent to minimize the supervised contrastive loss (Khosla et al., 2020) defined as:

$$\mathcal{L}_c = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{r}_i \cdot \mathbf{r}_p/\tau)}{\sum_{a \in A(i)} \exp(\mathbf{r}_i \cdot \mathbf{r}_a/\tau)} \tag{3}$$

where $\cdot$ denotes the inner product, $\tau \in \mathbb{R}^+$ is a temperature parameter, $A(i) \equiv I \setminus \{i\}$ indicates all instances other than i, and $P(i) \equiv p \in A(i) : y_p = y_i$ refers to the set of instances positives, with $|P(i)|$ its cardinality.

It is noteworthy to mention that $\mathbf{r}$ are l2-normalized features, to apply the criterion on an unity hyper-sphere. Using supervised contrastive loss, points belonging to the same class (positives) are pulled together in the projected space, while simultaneously pushing apart clusters of samples from different classes (negatives).

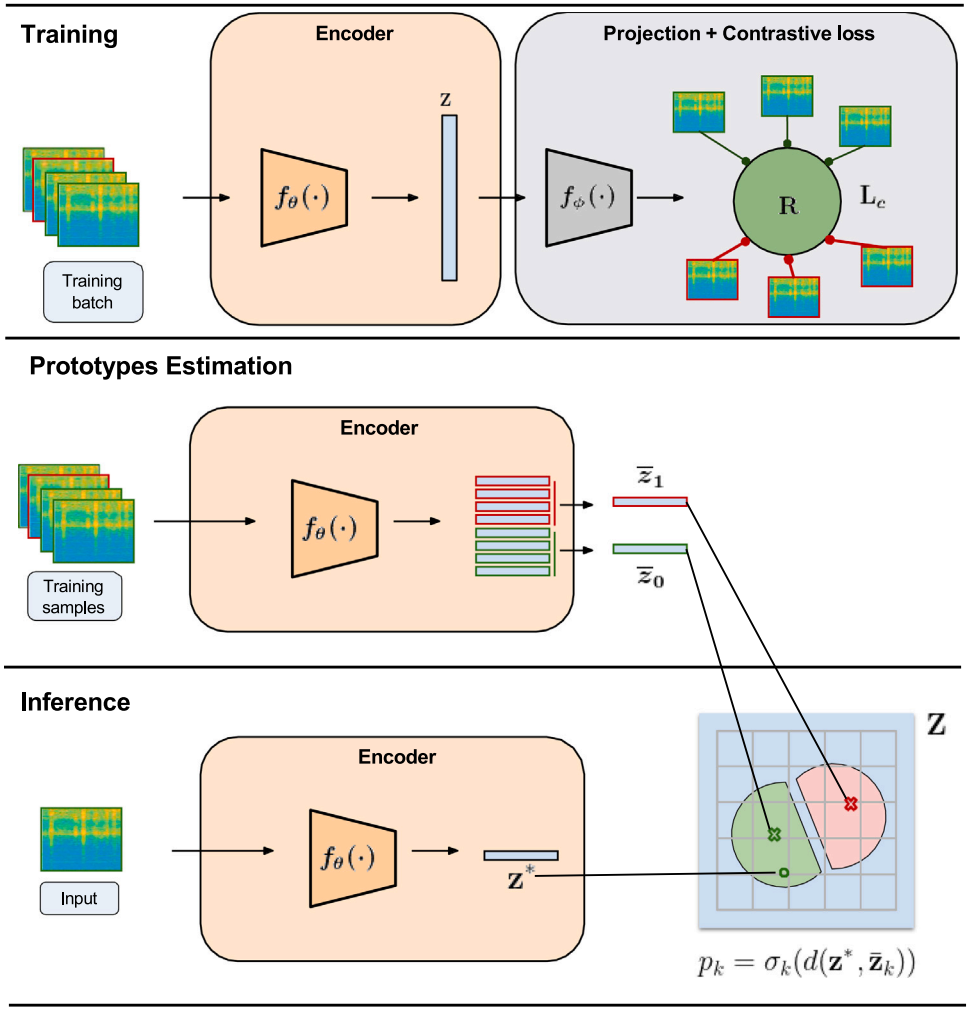

Fig. 2. Method overview. An encoder is trained to minimize supervised contrastive loss in Eq. (3) after projecting the produced embedding **z** into a subspace **r** that falls into an unit hyper-sphere. During inference, new queries are classified on the latent space projected by the encoder. Concretely, a non-parametric prototypical classifier is implemented using class-wise prototypes $\bar{\mathbf{z}}_k$ from the training set given by Eq. (4). In particular, the class of nearest prototype in terms of l2-distance is assigned to the new query sample.

3.4. Prototypical inference

For inference, contrastive-based methods usually train a linear classifier on top of the frozen representations **z** using a cross-entropy loss. In this work, we study the use of non-parametric inference strategy, to avoid overfitting on scenarios with limited data available. Concretely, we use prototypical-based inference (Snell et al., 2017), a memory-based approach that assigns predicted labels according to the distance in the latent space between new queries and precomputed representations of each class, called prototypes. This method creates softer decision boundaries compared to learned-based architectures. As we support later on our experiments, it generalizes better in the setting under study. Prototypes are calculated using all samples from training set such that:

$$\bar{\mathbf{z}}_k = \frac{1}{I}\sum_i \mathbf{z}_i \tag{4}$$

Given a new query sample, X^*, the wear prediction $\hat{y}_k$ is given by its relative distance to each prototype as follows:

$$\hat{y}_k = \sigma_k(d(f_{\boldsymbol{\theta}}(X^*), \bar{\mathbf{z}}_k)) \tag{5}$$

where σ_k indicates a softmax activation over classes, and $d(\cdot)$ indicates the Euclidean distance.

4. Experiments and results

4.1. Experimental setting

Dataset. The experiments described in this work were carried out using a private dataset of dynamic railway surveying on line 3 of Metrovalencia. 25 km of railway surveying were recorded using the data acquisition setup described in Section 3.1, with accelerometers of model KS76C100 manufactured by MMF and sampling frequencies of 3.2 KHz. The train used in the tests was an Electrical Multiple Unit (EMU 4300 series), which has four cars of two bogies each one, being motorized the wheelsets of the last car. The run tests had a maximum speed of 80 km/h, and included ballasted track with single-block concrete sleepers, and Stedef slab track. From the entire path, 33 crossing points were selected and manually on-site evaluated by experienced operators in terms of wear. Of this dataset, 17 crossings points showed damages that required follow-up and maintenance actions. Observed deterioration included spalls, burrs and squats. Examples of the deteriorated crossings are presented in Fig. 3. The acceleration signals recorded were windowed using 4 seconds around each crossing point.

Implementation details. The 4 seconds crossing signals acquired as detailed in Section 3.1 are transformed to time–frequency spectrograms as detailed in Section 3.2. Concretely, based upon the studies by Salvador et al. (2016), hamming windows of $W = 0.25$ seconds with

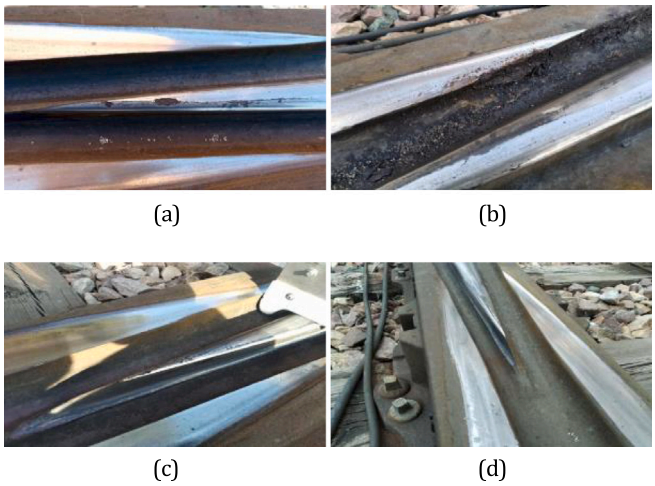

Fig. 3. Examples of deteriorated railway crossings included in the used dataset. Anomalies include squats (a–b–d), spalls (c), and burrs (d).

an overlap of $O = 95\%$ were used to compute the short-time Fourier transforms. Then, spectrograms were normalized using the dynamic-margin standardization with $\gamma = 20$, and resized to 256×320 pixels to reduce computational requirements. Using a 4-fold cross validation strategy, the encoder for crossing characterization was trained as described in Section 3.3. Concretely, ResNet-18 (He et al., 2016) was used as base architecture for the encoder. The architecture used included an initial convolutional layer to adapt the number of channels, and was composed of 2 residual blocks. The spatial features were reduced to a one-dimensional embedding $z \in \mathbb{R}^{64}$ via a global-average pooling. Regarding the projection head, a multi-layered perceptron that reduced the embedding size in an order of $F_c = 4$ with relu activation was used. The different modules were trained during 200 iterations, using ADAM optimizer with a learning rate of $1e{-}4$ and mini-batches of 8 samples. Finally, test samples form each fold are inferred as described in Section 3.4, using all samples from training subset to compute class-wise prototypes. The code and trained models are publicly available on (https://github.com/cvblab/contrastive_prototypes_railway).

Baselines. In order to compare our approach to state-of-the-art methods, we implemented proposals of prior works on accelerometer-based automatic railway maintenance, and validated them on the dataset used, under the same conditions. Due to the scarce literature on this field, we only differentiated three proposed approaches: hand-crafted feature-based methods, standard supervised learning using CNNs and cross entropy loss, and self-training ones via autoencoder features. ***Hand-crafted features methods***, aim to describe a series of features obtained by classic signal processing methods on time and frequency domains using human knowledge about the problem. Concretely, from the windowed crossing signal, we used as features the intensity peak amplitude, relative intensity at different bandwiths, entropy and other statistics such as skewness, similarly to Li and Shi (2019). Then, a support-vector machine (SVM) classifier with Gaussian kernel was trained to predict the wear crossing. ***Self-training methods***, aim to leverage knowledge on large amounts of unlabelled data from dynamic surveyings. Concretely, an autoencoder is trained to compress the spectrogram information into an embedding space, which is trained to minimize the reconstruction error using a trained decoder. Then, the resultant embedding space is used for clustering purposes. In our work, we implemented an autoencoder trained on the full dataset (including unlabelled data). Concretely, the same architecture with residual blocks used for our proposed method was used as encoder, and a symmetrical decoder was used to reconstruct the input spectrogram. The autoencoder architecture was pre-trained during 100 iterations using ADAM optimizer with a learning rate of $1e-4$ and mini-batches of 32 samples. Then, the non-parametric prototypical inference described in Section 3.4 was used for classification using the features extracted from the encoder. ***CNNs using cross-entropy loss***: Also, we include as an independent baseline the same CNN architecture trained using simply the binary cross entropy loss instead of the proposed learning method, as it has been used by Chellaswamy et al. (2019), Sresakoolchai and Kaewunruen (2021a) and Yang et al. (2021).

Evaluation metrics. We use standard metrics on classification tasks to evaluate the proposed system performance on crossing wear detection. In particular, accuracy, precision and recall are calculated using the expert and system labels. From precision and recall F1-score (FS) is calculated to summarize both figures of merit. For each experiment, the metrics shown are the mean of ten consecutive repetitions of the model training, to account for the variability of the stochastic factors involved in the process.

4.2. Results

4.2.1. Crossing wear detection

The quantitative results obtained by the proposed model and baselines on the cross-validation partitions are presented in Table 1. We can observe that the proposed methodology outperforms previous approaches by a large margin, with a substantial increase of $\sim 8\%$ in both accuracy and F1-score. Although the hand-crafted features baseline reached promising results (0.6124 accuracy), deep-learning methods outperformed this approach, which aligns to recent literature on railway surveying by Yang et al. (2021). Finally, the features learned by the autoencoder approach, even though it is trained on large quantities of data, obtained results inferior to those of the proposed method. This may be because the cross wear classification task requires specific features. In contrast, the autoencoder learns general features to reconstruct the original image that do not seem suitable for the supervised task.

4.2.2. Ablation studies

In the following, we provide comprehensive ablation experiments to validate several elements of our model, and motivate the choice of the values employed in our formulation, as well as our experimental setting.

Studies on model complexity. We first studied the configuration of the encoder used, ResNet-18, for the feature extraction stage. Concretely, we validated the proposed model using different number of residual blocks. Results are presented in Fig. 4(a), from which we can observe how the less residual blocks are used, the best the classification performance is. These results could be explained in two different ways: first, deep networks are over parameterized under scarce data conditions, and second, visual characterization on acceleration spectrograms are made up of by simple patterns, which are modelled on early layers of CNNs, together with intensity information.

Contrastive learning setup. Next, we study the multi-layered perceptron block used on the contrastive head. Concretely, ablation experiments are performed on the dimensionality of the unity hyper-sphere used to contrast samples, as a fraction of the dimension of the features extracted by the encoder. Concretely, the compression factor F_c is evaluated at $F_c = \{1, 2, 4, 8, 16\}$. Results are illustrated in Fig. 4(b). These show that reducing the dimension on the hyper-sphere used for contrastive losses produces slight benefits, with improvements around 3% on F1-score.

Table 1
Quantitative results on railway crossing wear detection for the proposed method and implemented baselines. Best results in bold.

Method	Metric ($\mu \pm \sigma$)			
	Accuracy	F1-Score	Precision	Recall
CNNs + BCE (Sresakoolchai & Kaewunruen, 2021a; Yang et al., 2021)	(0.5875 ± 0.0945)	0.6099 ± 0.1227	0.6111 ± 0.0863	0.6529 ± 0.2254
Hand-crafted Features + SVMs (Ghosh et al., 2021; Nadarajah et al., 2018)	(0.6124 ± 0.0619)	0.6529 ± 0.0147	0.6045 ± 0.0223	0.6512 ± 0.0543
Autoencoder Features (Niebling et al., 2020)	(0.6484 ± 0.0688)	0.6493 ± 0.0677	0.6771 ± 0.0767	0.6294 ± 0.0791
Proposed	$\mathbf{0.7156 \pm 0.7156}$	$\mathbf{0.7352 \pm 0.0508}$	$\mathbf{0.7264 \pm 0.049653}$	$\mathbf{0.7470 \pm 0.0647}$

Table 2
Ablation study on feature normalization methods. Best results in bold.

Normalization	Metric ($\mu \pm \sigma$)	
	Accuracy	F1- score
z-score	0.6124 (0.0619)	0.6045 (0.0223)
min–max	0.6484 (0.0259)	0.6529 (0.0147)
dynamic-margin	**0.7156 (0.0715)**	**0.7352 (0.0508)**

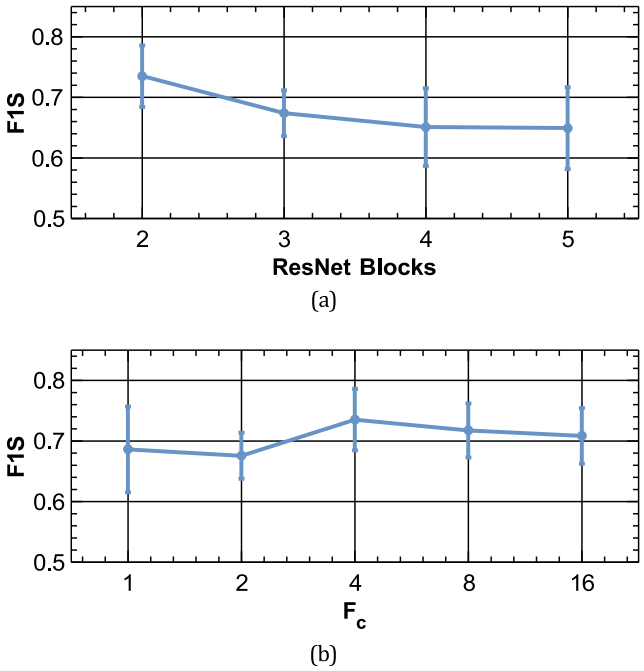

Fig. 4. Ablation studies on network architecture. Accuracy and F1-score are presented for each possible configuration. Best performance highlighted in bold. (a) Encoder complexity; (b) Contrastive head compression factor.

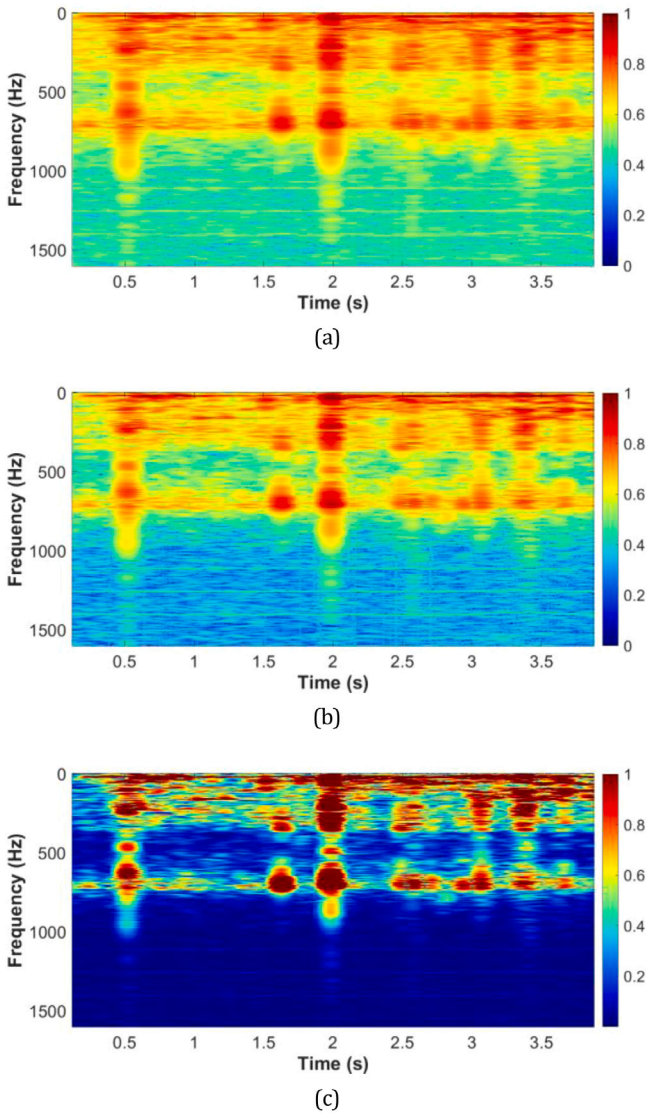

Fig. 5. Qualitative assessment of different normalization strategies. (a) min–max; (b) z-score; (c) dynamic margin.

Feature normalization. As previously mentioned, one of the main steps on deep learning systems is feature normalization. Concretely, the time–frequency spectrogram intensity should be constrained to small amplitudes, such that $x \in [0, 1]$. For this purpose, our method uses a dynamic-margin normalization described in Section 3.2. We now validate the proposed normalization, comparing both quantitatively and qualitatively with other well-known methods. In particular, we use minimum–maximum normalization, and z-score standardization on log-magnitude spectrograms. Results are presented in Table 2, while normalized spectrograms are presented in Fig. 5. Results demonstrate that benefits of dynamic-margin normalization, which outperforms other approaches by up to $\sim$ 8% in terms of F1-score. Qualitative evaluations show that the most large-intensity excited frequencies are contrasted from background on the spectrogram, the best the results are.

Learning strategies. In the following, we benchmark the proposed contrastive-based feature learning and prototypical inference with other common methods. Concretely, we train the proposed model using a linear classification layer and binary cross-entropy (BCE) loss to compare both contrastive and BCE-based training. For fair comparisons and to avoid the over-parametrization of densely classification, we also implement the prototypical inference on the BCE-trained model (BCE+Prototypes) as described in García et al. (2021). Finally, we also include a purely prototypical learning strategy (Prototypical), using episodic training and minimizing l2-distance between support and query samples as proposed in the original publication (Snell et al., 2017). Concretely, the number of query and support samples used during training was 4. The encoder architecture and hyper-parameters were the same to the ones optimized for our proposed method (see Section 4.1). Results for different methods are presented in Fig. 6 in terms of accuracy and F1-score. The proposed supervised contrastive

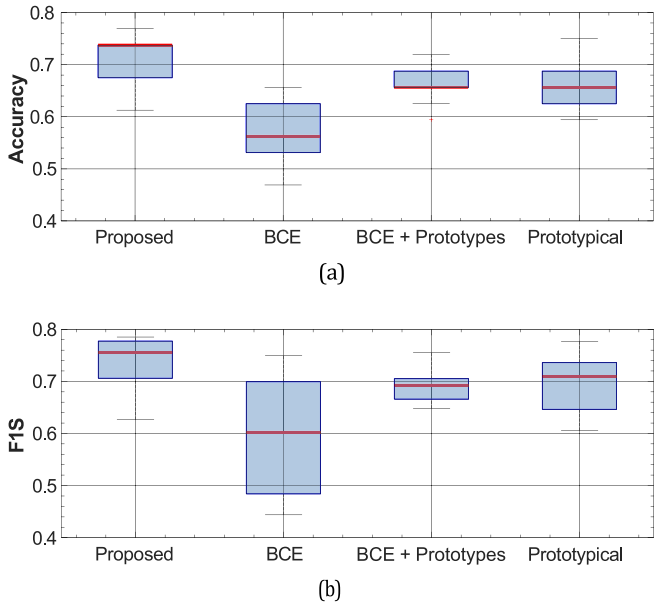

Fig. 6. Ablation studies on learning strategies. The illustrated metrics are accuracy (a) and F1-score (b).

learning model and prototypical inference outperforms by a large margin the BCE method, and shows greater stability in the results among experiment repetitions. Although results consistently improve using prototypical memory-based inference, our method reaches the best performance, which shows the benefits of contrastive learning strategies.

On the role of each element of the system. Different components have been presented to optimize the proposed method: dynamic margin normalization, prototypical inference, and contrastive feature learning have been the best performing settings. Nevertheless, it is still unclear the individual contribution of each element. For this reason, in the following, we discuss the incremental improvement of each module of the system. First, we focus on normalization methods, where dynamic margin normalization performed the best on the proposed setting (see Table 2). In addition, as shown in Fig. 7, this type of normalization is also indispensable to obtain promising results when we simply use a CNN with linear classifier, trained using cross-entropy (BCE). Thus, we consider this standardization to be an indispensable step for the operation of the system. Next, if we introduce an inference based on prototypes (BCE+Prototypes), improvements of $\sim 8\%$ are obtained (see Fig. 6). Finally, when we get rid of entropy-based objective functions, using the proposed contrastive learning setting, improvements of $\sim 5\%$ are obtained for both accuracy and F1-score figures of merit (see Fig. 6). Thus, we see that what most damages the model is the use of dense classifiers during inference, in the scenario studied with sparse data. Next, direct training of the model to generate prototypes based on contrastive learning also produces a substantial improvement.

4.2.3. On system explainability

Explainability on AI-based systems have become a relevant topic on the field that aims to prevent bias on learning systems and demonstrate the robustness of the model (Barredo Arrieta et al., 2020). In the following, we explore the explainability of the proposed model in order to provide confidence in its use during railway maintenance practice. Thus, we shed light into the features learned by the trained CNN to detect wear crossings using gradient-guided class activation maps (CAMs) (Selvaraju et al., 2020). For a given input image $\mathbf{x}$ its corresponding attention map is computed as: $a = \Sigma(\sum_k^K \alpha_k f_\theta^s(\mathbf{x})_k)$ where K is the total

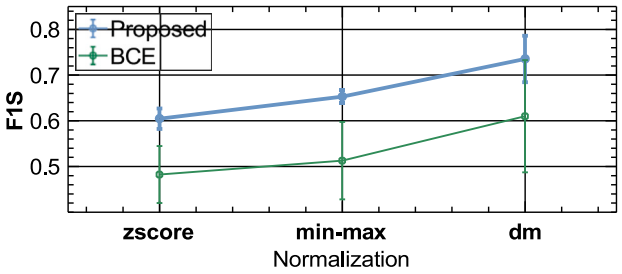

Fig. 7. Ablation study on feature normalization methods. In particular, performance using zscore, min–max, and the proposed dynamic margin (dm) normalization is compared for the proposed method and a CNN using linear classifier (BCE).

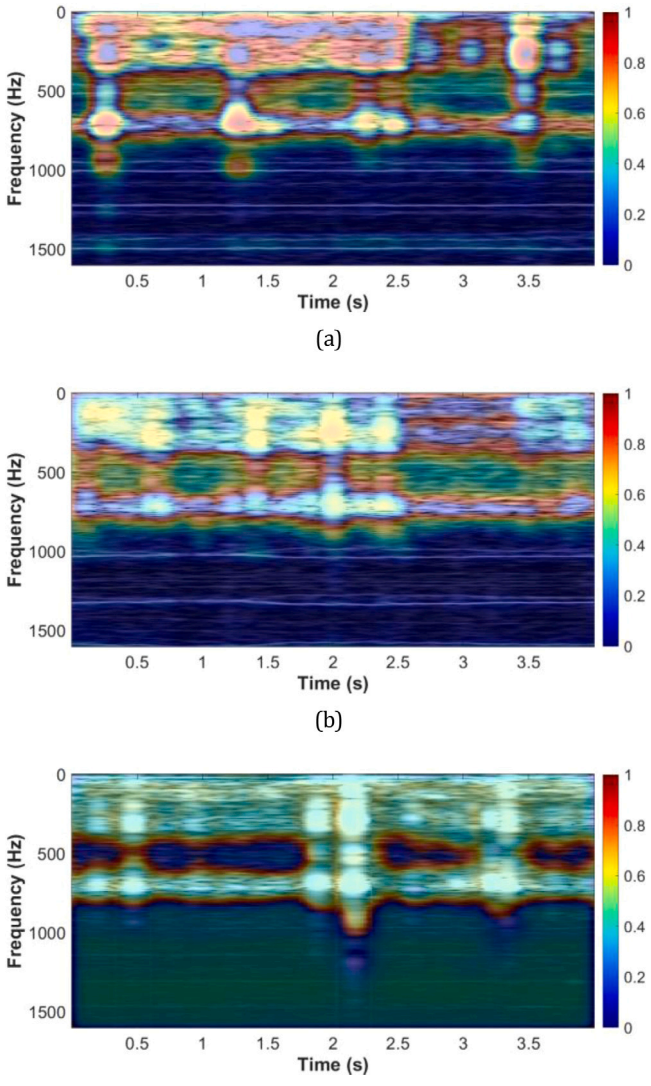

Fig. 8. Qualitative evaluation of the proposed model on wear crossing detection. For explainability, class-activation maps are obtained on true positive (a–b) and true negative (c) predictions, and overlaid over the input spectrogram.

number of filters of that encoder layer, Σ a sigmoid operation, and α_k are the generated gradients such that:

$$\alpha_k = \frac{1}{|\mathbf{a}|} \sum_{t \in \Omega_T} \frac{\partial \hat{y}_1}{\partial \mathbf{a}_{k,t}} \tag{6}$$

where Ω_T is the spatial features domain.

Generated CAMs of representative cases are visualized overlaid to the input spectrogram features on Fig. 8. These heat-maps highlight the important regions in the image for predicting a crossing as anomalous. Concretely, we can appreciate that CAMs focus on the band-with between 650 to 850 relaxation frequencies. These findings are consistent with previous literature in Salvador et al. (2016), that identified wider patterns and higher relative amplitude on this band related to crossings points on spectrograms.

5. Conclusions

A deep learning system capable of detecting worn crossings in dynamic railway inspections via axle-box accelerations sensing has been presented. Specifically, the system processes time–frequency spectrograms using convolutional neural networks through a novel combination of prototypical inference guided by supervised contrastive learning. The use of narrow CNNs showed the best results, as they extract mostly basic patterns, similar to those found in time–frequency spectrograms. Furthermore, normalization of these distributions using a dynamic margin scaling approach outperforms standard normalization in computer vision tasks. This method improves the contrast between the excited frequencies and the background, leading to better characterization. In addition, the supervised contrastive learning strategy has shown a promising performance for learning on small datasets. It outperforms standard cross-entropy based supervised learning by a wide margin, and improves other metric learning strategies from the few-shot learning domain, which resort to episodes-based training. The presented method achieves F1-score values of 0.7352 in a cross-validation, and outperforms previous literature by $\sim$ 8% for defect crossing classification. The presented system and its methods could be used to detect a wide range of singularities and defects in railway surveying.

CRediT authorship contribution statement

Julio Silva-Rodríguez: Conceptualization, Methodology, Software, Validation, Investigation, Writing – original draft. **Pablo Salvador:** Conceptualization, Resources, Data curation, Writing — review & editing, Supervision, Funding acquisition. **Valery Naranjo:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Funding acquisition. **Ricardo Insa:** Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

Acknowledgements

The authors would like to thank Metrovalencia and its staff for their collaboration and support provided during the data acquisition.

Funding

This work was supported by the Spanish Ministry of Economy and Competitiveness through project TRA2017-84317-R-AR. J. Silva-Rodríguez work was also supported by the Spanish Government under FPI Grant PRE2018-083443.

References

Baasch, B., Roth, M., Havrila, P., & Groos, J. C. (2019). Detecting singular track defects by time-frequency signal separation of axle-box acceleration data. In *12th world congress on railway research* (pp. 1–6).

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Bocz, P., Vinkó, A., & Posgay, Z. (2018). A practical approach to tramway track condition monitoring: Vertical track defects detection and identification using time-frequency processing technique. *Selected Scientific Papers - Journal of Civil Engineering*, *13*, 135–146.

Boogaard, M. A., Li, Z., & Dollevoet, R. P. B. J. (2018). In situ measurements of the crossing vibrations of a railway turnout. *Measurement: Journal of the International Measurement Confederation*, *125*, 313–324.

Bosso, N., Gugliotta, A., & Zampieri, N. (2018a). Design and testing of an innovative monitoring system for railway vehicles. *Proceedings of the Institution of Mechanical Engineers, Part F (Journal of Rail and Rapid Transit)*, *232*, 445–460.

Bosso, N., Gugliotta, A., & Zampieri, N. (2018b). Wheel flat detection algorithm for onboard diagnostic. *Measurement: Journal of the International Measurement Confederation*, *123*, 193–202.

Carrigan, T. D., Fidler, P. R. A., & Talbot, J. P. (2019). On the derivation of rail roughness spectra from axle-box vibration: Development of a new technique. In *International conference on smart infrastructure and construction 2019, ICSIC 2019: Driving data-informed decision-making. 2019* (pp. 549–557).

Carrigan, T. D., & Talbot, J. P. (2021). Extracting information from axle-box acceleration: On the derivation of rail roughness spectra in the presence of wheel roughness. *Notes on Numerical Fluid Mechanics and Multidisciplinary Design*, *150*, 286–294.

Chang, C., Ling, L., Chen, S., Zhai, W., Wang, K., & Wang, G. (2021). Dynamic performance evaluation of an inspection wagon for urban railway tracks. *Measurement: Journal of the International Measurement Confederation*, *170*, Article 108704.

Chellaswamy, C., Krishnasamy, M., Balaji, L., Dhanalakshmi, A., & Ramesh, R. (2020). Optimized railway track health monitoring system based on dynamic differential evolution algorithm. *Measurement: Journal of the International Measurement Confederation*, *152*, Article 107332.

Chellaswamy, C., Santhi, P., & Venkatachalam, K. (2019). Deep learning based intelligent rail track health monitoring system. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *9*, 5111–5122.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *37th international conference on machine learning* (pp. 1–20).

Chen, W. Y., Wang, Y. C. F., Liu, Y. C., Kira, Z., & Huang, J. B. (2019). A closer look at few-shot classification. In *7th international conference on learning representations* (pp. 1–17).

Chen, S. X., Zhou, L., Ni, Y. Q., & Liu, X. Z. (2021). An acoustic-homologous transfer learning approach for acoustic emission–based rail condition evaluation. *Structural Health Monitoring*, *20*, 2161–2181.

Chenariyan Nakhaee, M., Hiemstra, D., Stoelinga, M., & van Noort, M. (2019). The recent applications of machine learning in rail track maintenance: A survey. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics): Vol. 11495 LNCS*, (pp. 91–105).

Chia, L., Bhardwaj, B., Lu, P., Bridgelall, R., & Member, S. (2019). Railroad track condition monitoring using inertial sensors and digital signal processing: A review. *IEEE Sensors Journal*, *19*, 25–33.

Faghih-Roohi, S., Hajizadeh, S., Núñez, A., Babuska, R., De Schutter, B., Faghih-Roohi, S., Hajizadeh, S., Núñez, A., Babuska, R., & De Schutter, B. (2016). Deep convolutional neural networks for detection of rail surface defects. In *2016 international joint conference on neural networks: Vol. 19*, (pp. 2584–2589).

García, G., del Amor, R., Colomer, A., Verdú-Monedero, R., Morales-Sánchez, J., & Naranjo, V. (2021). Circumpapillary OCT-focused hybrid learning for glaucoma grading using tailored prototypical neural networks. *Artificial Intelligence in Medicine*, *118*.

Ghosh, C., Verma, A., & Verma, P. (2021). Real time fault detection in railway tracks using fast Fourier transformation and discrete wavelet transformation. *International Journal of Information Technology*.

Giben, X., Patel, V. M., & Chellappa, R. (2015). Material classification and semantic segmentation of railway track images with deep convolutional neural networks. In *Proceedings - international conference on image processing: Vol. 2015-Decem*, (pp. 621–625).

Gibert, X., Patel, V. M., & Chellappa, R. (2017). Deep multitask learning for railway track inspection. *IEEE Transactions on Intelligent Transportation Systems*, *18*, 153–164.

He, X., Yu, K., Cai, C., & Zou, Y. (2020). Dynamic responses of the metro train's bogie frames: Field tests and data analysis. *Shock and Vibration*, *2020*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the conference on computervision and pattern recognition* (pp. 1–12).

Heusel, J., Baasch, B., Riedler, W., Roth, M., Shankar, S., & Groos, J. C. (2021). Detecting corrugation defects in harbour railway networks using axle box acceleration data. In *International conference on condition monitoring and asset management*.

Hory, C., Bouillaut, L., & Aknin, P. (2012). Time-frequency characterization of rail corrugation under a combined auto-regressive and matched filter scheme. *Mechanical Systems and Signal Processing*, *29*, 174–186.

Hovad, E., Wix, T., Khomiakov, M., Vassos, G., Silva Rodrigues, A. F. D., Miguel Tejada, A. D., & Clemmensen, L. H. (2021). Deep learning for automatic railway maintenance. In *Intelligent quality assessment of railway switches and crossings* (pp. 207–228).

James, A., Jie, W., Xulei, Y., Chenghao, Y., Ngan, N. B., Yuxin, L., Yi, S., Chandrasekhar, V., & Zeng, Z. (2019). TrackNet - A deep learning based fault detection for railway track inspection. In *International conference on intelligent rail transportation*.

Jing, L., Wang, K., & Zhai, W. (2021). Impact vibration behavior of railway vehicles: A state-of-the-art overview. *Acta Mechanica Sinica*, *77*.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. In *NeurIPS* (pp. 1–23).

Kouroussis, G., Caucheteur, C., Kinet, D., Alexandrou, G., Verlinden, O., & Moeyaert, V. (2015). Review of trackside monitoring solutions: From strain gages to optical fibre sensors. *Sensors*, *15*, 20115–20139.

Laenen, S., & Bertinetto, L. (2020). On episodes, prototypical networks, and few-shot learning. In *NeurIPS 2020 meta-learning workshop* (pp. 1–19).

Li, J., & Shi, H. (2019). Rail corrugation detection of high-speed railway using wheel dynamic responses. *Shock and Vibration*, *2019*.

Lidén, T. (2015). Railway infrastructure maintenance - A survey of planning problems and conducted research. *Transportation Research Procedia*, *10*, 574–583.

Malekjafarian, A., OBrien, E., Quirke, P., & Bowe, C. (2019). Railway track monitoring using train measurements: An experimental case study. *Applied Sciences*, *9*.

Malekjafarian, A., Obrien, E. J., Quirke, P., Cantero, D., & Golpayegani, F. (2021). Railway track loss-of-stiffness detection using bogie filtered displacement data measured on a passing train. *Infrastructures*, *6*, 1–17.

Mittal, S., & Rao, D. (2017). Vision based railway track monitoring using deep learning. arXiv.

Molodova, M., Li, Z., & Dollevoet, R. (2011). Axle box acceleration: Measurement and simulation for detection of short track defects. *Wear*, *271*, 349–356.

Nadarajah, N., Shamdani, A., Hardie, G., Chiu, W. K., & Widyastuti, H. (2018). Prediction of railway vehicles' dynamic behavior with machine learning algorithms. *Electronic Journal of Structural Engineering*, *18*, 38–46.

Ng, A. K., Martua, L., & Sun, G. (2019). Dynamic modelling and acceleration signal analysis of rail surface defects for enhanced rail condition monitoring and diagnosis. In *4th international conference on intelligent transportation engineering* (pp. 69–73). IEEE.

Niebling, J., Baasch, B., & Kruspe, A. (2020). Analysis of railway track irregularities with convolutional autoencoders and clustering algorithms. *Communications in Computer and Information Science*, *1279 CCIS*, 78–89.

Salvador, P., Naranjo, V., Insa, R., & Teixeira, P. (2016). Axlebox accelerations: Their acquisition and time-frequency characterisation for railway track monitoring purposes. *Measurement: Journal of the International Measurement Confederation*, *82*, 301–312.

Salvador, P., Villalba, I., Martínez-fernández, P., & Insa, R. (2018). Application of time-frequency representations for the detection of railway track singularities. In *The 5th joint international conference on multibody system dynamics* (pp. 28–29).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, *128*, 336–359.

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, *2017-Decem*, 4078–4088.

Song, Y., Liang, L., Du, Y., & Sun, B. (2020). Railway polygonized wheel detection based on numerical time-frequency analysis of axle-box acceleration. *Applied Sciences*, *10*.

Sresakoolchai, J., & Kaewunruen, S. (2021a). Detection and severity evaluation of combined rail defects using deep learning. *Vibration*, *4*, 341–356.

Sresakoolchai, J., & Kaewunruen, S. (2021b). Wheel flat detection and severity classification using deep learning techniques. *Insight-Non-Destructive Testing and Condition Monitoring*, *63*, 393–402.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H. S., & Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning, *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 1199–1208).

Sysyn, M., Gerber, U., Nabochenko, O., Li, Y., & Kovalchuk, V. (2019a). Indicators for common crossing structural health monitoring with track-side inertial measurements. *Acta Polytechnica*, *59*, 170–181.

Sysyn, M., Gruen, D., Gerber, U., Nabochenko, O., & Kovalchuk, V. (2019b). Turnout monitoring with vehicle based inertial measurements of operational trains: A machine learning approach. *Communications - Scientific Letters of the University of Zilina*, *21*, 42–48.

Tzanakakis, K. (2013). *Springer tracts on transportation and traffic*: *Vol. 2*, *The railway track and its log term behaviour* (pp. 279–292). Springer.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. (2016). Matching networks for one shot learning. In *Advances in neural information processing systems* (pp. 3637–3645).

Wang, L., Zhuang, L., & Zhang, Z. (2019). Automatic detection of rail surface cracks with a superpixel-based data-driven framework. *Journal of Computing in Civil Engineering*, *33*, Article 04018053.

Wei, Z., Núñez, A., Li, Z., & Dollevoet, R. (2017). Evaluating degradation at railway crossings using axle box acceleration measurements. *Sensors*, *17*.

Yang, C., Sun, Y., Ladubec, C., & Liu, Y. (2021). Developing machine learning-based models for railway inspection. *Applied Sciences*, *11*, 1–15.

Zhang, H., Jin, X., Wu, J. Q. M., He, Z., & Wang, Y. (2018). Automatic visual detection method of railway surface defects based on curvature filtering and improved GMM. *Yi Qi Yi Biao Xue Bao/Chinese Journal of Scientific Instrument*, *39*, 181–194.

Zhang, B., Li, X., Ye, Y., Huang, Z., & Zhang, L. (2020). Prototype completion with primitive knowledge for few-shot learning. In *Accepted on 2021 IEEE conference on computer vision and pattern recognition* (pp. 1–10).

Zhang, X., Wang, K., Wang, Y., Shen, Y., & Hu, H. (2017). An improved method of rail health monitoring based on CNN and multiple acoustic emission events. In *I2MTC 2017 - 2017 IEEE international instrumentation and measurement technology conference, proceedings* (pp. 1–6). IEEE.