UNIVERSITÀ POLITECNICA DELLE MARCHE
Repository ISTITUZIONALE

Fault Diagnosis of Rotating Machinery Based on Wasserstein Distance and Feature Selection

note finali coverpage

(Article begins on next page)

22 September 2024

Fault Diagnosis of Rotating Machinery based on Wasserstein Distance and Feature Selection

Francesco Ferracuti, *Member, IEEE,* Alessandro Freddi, *Member, IEEE,* Andrea Monteriù, *Member, IEEE,* and Luca Romeo

Abstract—This paper presents a fault diagnosis algorithm for rotating machinery based on Wasserstein distance. Recently, Wasserstein distance has been proposed as a new research direction to find better distribution mapping when compared with other popular statistical distances and divergences. In this work, firstly, frequency and time-based features are extracted by vibration signals and, secondly, the Wasserstein distance is considered for the learning phase to discriminate the different machine operating conditions. Specifically, the 1-dimensional (1D) Wasserstein distance is taken into account thanks to its low computational burden because it can be evaluated directly by the order statistics of the extracted features. Furthermore, a distance weighting stage based on neighborhood component features selection (NCFS) is exploited to achieve robust fault diagnosis at low signal-to-noise ratio (SNR) conditions and with high-dimensional features. In detail, the NCFS framework is here adapted to weight 1D Wasserstein distances evaluated from time/frequency features. Experiments are conducted on two benchmark datasets to verify the effectiveness of the proposed fault diagnosis method at different SNR conditions. The comparison with state-of-the-art fault diagnosis algorithms shows promising results.

Note to Practitioners—This article was motivated by the problem of fault diagnosis of rotating machinery under low SNR and different machine operating conditions. The algorithm employs a statistical distance-based fault diagnosis technique, which permits to obtain an estimation of the fault signature without the need for training a classifier. The algorithm is computationally efficient during the training and testing stages, and thus it can be used in embedded hardware. Finally, the proposed methodology can be applied to other application domains such as system monitoring and prognostics which can help to schedule the maintenance of rotating machinery.

Index Terms—Fault diagnosis, rotating machines, Wasserstein Distance, statistical distances, neighborhood component features selection.

I. INTRODUCTION

AS a consequence of the developments in the fields of technology and materials science, industrial equipment is increasing its functionality and complexity. Among them, rotating machinery plays a fundamental role in modern industrial applications, and fault diagnosis assumes utmost importance to ensure both availability and safety, prevent system downtime, and save economic losses to the customers [1], [2]. According to the types of data and how the data are processed, fault diagnosis methods can be divided into three main classes: *model-based* (or *online data driven*), *signal-based* and *knowledge-based* (or *historical data driven*) [3],

F. Ferracuti, A. Freddi, A. Monteriù, L. Romeo are with the Department of Information Engineering, Università Politecnica delle Marche, 60131 Ancona, Italy e-mail: (e-mail: f.ferracuti@univpm.it; a.freddi@univpm.it; a.monteriu@univpm.it; l.romeo@univpm.it).

[4]. The main trend in fault diagnosis for rotating machinery is to employ signal-based or knowledge-based methods, with the adoption of model-based methods limited to specific classes of rotating machinery and faults [5], [6]. Knowledge-based methods mainly rely on artificial intelligent approaches, while signal-based methods perform data processing for waveform and multidimensional data. In this work, we will focus our attention on the latter.

The most common waveform data in condition monitoring of mechanical systems are vibration signals and acoustic emissions. Moreover, motor currents and partial discharge are used as well in the literature for electrical machines [7]. The waveform data can then be analyzed in different domains, namely the *time-domain*, *frequency-domain* and *time-frequency-domain*. In the first case, typically, different statistical features of the signals in the time-domain are extracted and then analyzed, whether in a 1D (e.g., cross-correlation analysis) or 2D domain (e.g., by mapping the time signals into images) [8]. In the frequency-domain, instead, the most used methods rely on the fast Fourier transform (FFT); however, the FFT analyzes the signal within a specific time-window, thus returning an averaged frequency signal over time, which makes transient features difficult to examine [9].

Time-frequency methods try to overcome the problem of effective transient analysis. The most common time-frequency methods in the literature are the short-time Fourier transform, the empirical mode decomposition, the continuous wavelet transform, and the discrete wavelet transform [10]–[13]. These techniques, however, do not address the problem of quasi-stationary processes which present different spectral profiles during time. This is evident in condition monitoring of bearings in rotating machines, where the presence of noise, the quasi-stationary nature of bearing vibrations and the variation of the operating conditions make the single-time-segment spectrum of the healthy and faulty condition appear diverse for different time-segments [14]. The same applies to many common waveform data, e.g., motor currents [15].

In the last years, both frequency-based and time-frequency-based methods have been investigated, in order to cope with the above mentioned problem. Kurtogram methods are often employed to diagnose bearing faults by using vibration data in poor signal-to-noise ratio (SNR) conditions. The research efforts to produce enhanced kurtograms mainly follow two directions: to enhance the impulse signals produced by faults or to find the frequency band which contains the strongest impulse signals produced by faults [16], [17]. Recently, [14] developed a frequency-based approach that uses several-time-segment spectra of vibration signals to build a spectral image

for the purpose of fault diagnosis. A similar approach was investigated by the authors in [18] as well. Similarly, a solution based on power spectrum density (PSD)-images in combination with deep convolutional autoencoder for high-level feature extraction is proposed in [19]. Feature extraction of the spectrum and deep learning-based algorithms are proposed extensively for fault classification and degradation prediction in the last years [20]–[24], whereas, in this context, Wasserstein distance-based solutions for fault diagnosis are still at the beginning of the investigation in the literature.

The Wasserstein distance is a true metric [25] and can be traced back to the mass transport problem [26]. A more detailed account of the history and description of this distance can be found in [27], [28]. In computer science, the Wasserstein distance, in a special case, is better known as the earth mover's distance (EMD) [29]. Recently, Wasserstein distance has taken much attention in deep learning, in particular in [30], in which the authors introduced a new algorithm named Wasserstein-generative adversarial networks (WGAN), an alternative to traditional generative adversarial networks (GAN) training, allowing to prevent problems with vanishing gradient. The exploitation of statistical distances/divergences in the fault diagnosis framework has been widely investigated in the literature [31]–[33]. The distances/divergences allow to detect the fault, to identify the fault and, finally, to obtain an estimation of its severity without the need for training a classifier [34], [35].

Currently, in [30], [36], the authors demonstrated that Wasserstein distance was a viable approach to find better distribution mapping. Moreover, in [30], the authors demonstrated that Wasserstein distance is a more sensible cost function compared with other popular probability distances and divergences, such as Kullback-Leibler (KL) divergence and Jensen-Shannon (JS) divergence when learning distributions supported by low dimensional manifolds. Currently, in [18], the authors proposed the exploitation of Wasserstein distance for fault diagnosis in the frequency domain and in [37], the authors expanded this topic by proposing the Wasserstein-Fourier distance to measure the dissimilarity between time series by quantifying the displacement of their energy across frequencies. In this work, the Wasserstein distance is considered in the learning phase to discriminate the different machine operating conditions. Specifically, the 1-dimensional (1D) Wasserstein distance is taken into account thanks to its low computational burden since it can be evaluated directly by the order statistics of the extracted features. These preliminary considerations supported our motivation to include the Wasserstein distance as loss function to be optimized in the neighborhood component features selection (NCFS) procedure, with the aim to improve the generalization performances in the presence of low SNR conditions and high dimensional feature set. NCFS is a supervised method that allows improving the interpretation of the diagnosis stage thanks to the weighting of the relevant frequencies involved in the fault process. In the present work, the standard NCFS framework is adapted to weight 1D Wasserstein distances evaluated from time/frequency features.

The rest of this paper is organized as follows. In Section II the related methods are reviewed, and the proposed fault diagnosis framework presented. The experimental setups of two bearing benchmarks are described in Section III. Two bearing benchmarks are separately analyzed using the proposed method in Section IV. Finally, conclusions are drawn in Section V.

II. MATERIAL AND METHODS

This section introduces the Wasserstein distance, the distance weighting method based on NCFS, and the proposed fault diagnosis algorithm which combines both 1D Wasserstein distance and NCFS.

A. Wasserstein distance

Wasserstein distance has been broadly used in the field of statistics and probability theory as a distance measure between probability distributions. Because Wasserstein distance can measure the similarity between probability measures, we take it as a measure of ability to classify different categories.

Let $(\mathcal{M}, \mathcal{D})$ be a Polish metric space and let $p \in [1 \, \infty)$, for any two probability measures μ and ν on $\mathcal{M}$ the p-Wasserstein distance $W_p(\mu, \nu)$ is defined as follows:

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Gamma(\mu, \nu)} \int_{\mathcal{M} \times \mathcal{M}} \mathcal{D}(x, y)^p d\pi(x, y) \right)^{1/p} \quad (1)$$

where $\Gamma(\mu, \nu)$ denotes the set of all joint probability measures on $\mathcal{M} \times \mathcal{M}$ whose marginals are μ and ν, $\mathcal{D}(x, y)$ is a distance on a Polish space $\mathcal{M}$, and $X \sim \mu$ and $Y \sim \nu$. Normally, the L^p-norm is considered as a distance $\mathcal{D}(\cdot)$ on the Polish space, thus in the one-dimensional case it is simply $\mathcal{D}(\cdot) = \sum |\cdot|^p$. In the one-dimensional case, some works define the 1D p-Wasserstein distance in terms of Mallows' distance, i.e., as $W_p^p(\mu, \nu)$ [28], [38], [39]. In detail, the Mallows' L_p distance (equivalently the 1D p-Wasserstein distance) can be expressed with a simple formula in terms of the inverse of the cumulative distribution functions $F_\mu^{-1}(q)$ and $G_\nu^{-1}(q)$ and often referred as quantile functions [39], where $F_\mu^{-1}(q) = \inf\{x : F_\mu(x) \geq q\}$, and $F_\mu(x)$ is the cumulative distribution function (CDF).

In particular, the 1D p-Wasserstein distance (Mallows' L_p distance) between two probability measures μ and ν on $\mathbb{R}$ with p-finite moments is:

$$W_p^p(\mu, \nu) = \int_0^1 \mathcal{D}\left(F_\mu^{-1}(q), G_\nu^{-1}(q)\right)^p dq \quad (2)$$

In the special case of $p = 1$ and $\mathcal{D}(\cdot) = \sum |\cdot|$, the 1D 1-Wasserstein (Mallows' L_1) distance can be evaluated as the area between the two CDFs; in particular, the following equivalence holds:

$$W_1^1(\mu, \nu) = \int_0^1 |F_\mu^{-1}(q) - G_\nu^{-1}(q)| dq = \quad (3)$$

$$= \int_{\mathbb{R}} |F_\mu(x) - G_\nu(x)| dx \quad (4)$$

To formalize the discrete 1D p-Wasserstein distance, let define

the empirical cumulative distribution function (ECDF) and the empirical quantile function.

The empirical cumulative distribution function is defined as:

$$F_n(x) = \frac{1}{n} \sum_{k=1}^{n} I_{(-\infty, x]}(x_k) \tag{5}$$

where $I_{(-\infty, x]}(x_k)$ is an indicator function assuming the value 1 if $x_k \leq x$ and 0 otherwise, $\sum_{k=1}^{n} I_{(-\infty, x]}(x_k)$, which is the number of $x_k \leq x$, has a binomial distribution with parameters n and $F_\mu(x)$. It is possible to prove that $F_n(x)$ is an unbiased estimator for the cumulative distribution function $F_\mu(x)$ [40]. Given a time series of n real numbers, the ECDF is a step function $F_n(x)$ defined by increasing by $1/n$ at each data point. The ECDF, which is obtained after sorting the data, is parameter free by definition. In the special case of $p = 1$, the discrete 1D 1-Wasserstein distance can be evaluated as the area between the two ECDFs. Note that in the case of $p > 1$, the previous remark is not true. In the literature, the parameter $p = 2$ and then the corresponding 1D 2-Wasserstein distance is considered to evaluate the distance from normality [41], [42].

The empirical quantile function is defined as follows:

$$F_n^{-1}(q) = \inf\{x : F_n(x) \geq q\} = x_k^{(s)} \tag{6}$$

where $\frac{k-1}{n} \leq q \leq \frac{k}{n}$ and $x_1^{(s)} \leq, \ldots, \leq x_n^{(s)}$ is a sequence of order statistics.

The 1D p-Wasserstein distance can be defined also in terms of the empirical quantile function and then through its order statistics.

Remark 1. *Consider the i.i.d. (independent and identically distributed) random variables $X \sim \mu$ and $Y \sim \nu$ with equal sample size and the corresponding sequence of order statistics $x_k^{(s)}$ and $y_k^{(s)}$, in the discrete and one-dimensional case, the 1D p-Wasserstein distance can be evaluated as:*

$$W_p^p(\mu, \nu) = \frac{1}{n} \sum_{k=1}^{n} \mathcal{D}\left(x_k^{(s)}, y_k^{(s)}\right)^p \tag{7}$$

Thus, for the special case related to i.i.d. random variables, the 1D p-Wasserstein distance can be approximated into order statistics calculation (i.e., sorting problem) that can be solved efficiently ($\mathcal{O}(n \log n)$ in the worst case and $\mathcal{O}(n)$ in the best case [43]) and calculating $\mathcal{D}\left(x_k^{(s)}, y_k^{(s)}\right)^p$.

The equivalence between the 1D p-Wasserstein distance and the 1D Earth Mover's distance $EMD(x_k, y_k)$ [29], [44] can be summarized as follows.

Remark 2. *If $\mu = (x_k, w_x) \in \mathbb{R}$ and $\nu = (y_k, w_y) \in \mathbb{R}$ have equal weight $w_x = w_y$ and are i.i.d., then the optimization problem of the Earth Mover's distance can be solved explicitly [38]:*

$$EMD(x_k, y_k) = W_p^p(\mu, \nu) = \frac{1}{n} \sum_{k=1}^{n} \mathcal{D}\left(x_k^{(s)}, y_k^{(s)}\right)^p \tag{8}$$

The previous remarks show as the discrete 1D p-Wasserstein distance can be calculated explicitly by order statistics and how, in the specific case of an equal sample size of the two random variables, the distance is equivalent to the 1D Earth Mover's distance. The discrete 1D p-Wasserstein distance calculation by order statistics will be considered in the proposed fault diagnosis algorithm due to its efficient calculation.

B. Neighborhood Component Features Selection

In the Sections II-B and II-C, the following notation is considered in order to formulate the classification problem:

- Let $T = (\mathcal{X}, \boldsymbol{y})$ be a supervised training sample.
- Let $\mathcal{X}$ be a finite set of N observations of training features $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N\}$, where $\boldsymbol{X}_i = \{\boldsymbol{x}_1^{[i]}, \ldots, \boldsymbol{x}_d^{[i]}\}$ is a matrix of d-dimensional feature vectors of n values, $\boldsymbol{x}_l^{[i]} = \{x_{1l}^{[i]}, \ldots, x_{nl}^{[i]}\}$, $l \in \{1, \ldots, d\}$, $i \in \{1, \ldots, N\}$.
- $x_l^{(s)[i]}$ denotes the order statistics of $x_l^{[i]}$.
- Let $\boldsymbol{y}$ be a finite set of the corresponding class labels $\{y_1, \ldots, y_N\}$, with $y_i \in \{1, \ldots, C\}$ where C is the number of classes.

Feature weighting methods aim to weight features that not only have maximum relevancy between each other, but also have a strong ability to recognizing different classes or categories. Neighborhood component feature selection is a non-parametric and embedded method to select relevant features for high-dimensional data in order to maximize the expected leave-one-out classification accuracy [45]. We denote the weighting vector $\boldsymbol{\omega} = \{\omega_1^2, \ldots, \omega_d^2\}$, and a general weighted distance between two samples matrices $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ by:

$$\Xi_{\boldsymbol{\omega}}(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sum_{l=1}^{d} \omega_l^2 \mathcal{D}(\boldsymbol{x}_l^{[i]}, \boldsymbol{x}_l^{[j]}) \tag{9}$$

where, in the standard NCFS algorithm, $\mathcal{D}(\cdot)$ is the L_1-norm distance that is considered a sparsity measure [46], [47].

Defining the reference point as the number of nearest neighbors to be selected on a k-NN classifier [45], the probability that $\boldsymbol{X}_i$ selects $\boldsymbol{X}_j$ as its reference point is:

$$p_{ij}(\boldsymbol{\omega}) = \frac{\kappa(\Xi_{\boldsymbol{\omega}}(\boldsymbol{X}_i, \boldsymbol{X}_j))}{\sum_{k \neq i} \kappa(\Xi_{\boldsymbol{\omega}}(\boldsymbol{X}_i, \boldsymbol{X}_k))} \tag{10}$$

where

$$\kappa(z) = e^{-z/\sigma} \tag{11}$$

The probability that the query point $\boldsymbol{X}_i$ is correctly classified can be computed as:

$$p_i(\boldsymbol{\omega}) = \sum_{j=1}^{N} t_{ij} p_{ij}(\boldsymbol{\omega}) \tag{12}$$

where $t_{ij} = 1$ if and only if $y_i = y_j$, and $t_{ij} = 0$ otherwise. In order to compute the weighting vector $\boldsymbol{\omega}$, the approximate leave-one-out classification accuracy is considered, namely:

$$\psi(\boldsymbol{\omega}) = \frac{1}{N} \sum_{i=1}^{N} p_i(\boldsymbol{\omega}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} t_{ij} p_{ij}(\boldsymbol{\omega}) \tag{13}$$

Adding a regularization term $\lambda \geq 0$ the following objective function is obtained:

$$\xi(\boldsymbol{\omega}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} t_{ij} p_{ij}(\boldsymbol{\omega}) - \lambda \sum_{l=1}^{d} \omega_l^2 \tag{14}$$

The optimization problem is:

$$\boldsymbol{\omega}^* = \arg\max_{\boldsymbol{\omega}} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} t_{ij} p_{ij}(\boldsymbol{\omega}) \tag{15}$$

$$\text{subject to } \sum_{l=1}^{d} \omega_l^2 \leq \epsilon \tag{16}$$

where $\epsilon \geq 0$. The objective function can be optimized using the gradient descent method. The function $\xi(\boldsymbol{\omega})$ is differentiable and then its derivative with respect to ω_l is:

$$\frac{\partial \xi(\boldsymbol{\omega})}{\partial \omega_l} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} t_{ij} \frac{\partial p_{ij}(\boldsymbol{\omega})}{\partial \omega_l} - 2\lambda \omega_l \tag{17}$$

Defining:

- $S_{ij}(\boldsymbol{\omega}) = \kappa(\Xi_{\boldsymbol{w}}(\boldsymbol{X}_i, \boldsymbol{X}_j)) = e^{-\Xi_{\boldsymbol{w}}(\boldsymbol{X}_i, \boldsymbol{X}_j)/\sigma}$
- $p_{ij}(\boldsymbol{\omega}) = \frac{S_{ij}(\boldsymbol{\omega})}{\sum_{k \neq i} S_{ik}(\boldsymbol{\omega})}$
- $\frac{\partial S_{ij}(\boldsymbol{\omega})}{\partial \omega_l} = \frac{-S_{ij}(\boldsymbol{\omega})}{\sigma} \frac{\partial \Xi_{\boldsymbol{w}}(\boldsymbol{X}_i, \boldsymbol{X}_j)}{\partial \omega_l}$

the derivative with respect to ω_l is calculated as:

$$\frac{\partial \xi(\boldsymbol{\omega})}{\partial \omega_l} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} t_{ij} \cdot$$
$$\cdot \left[\frac{p_{ij}(\boldsymbol{\omega})}{\sigma} \left(\sum_{k \neq i} p_{ik}(\boldsymbol{\omega}) \frac{\partial \Xi_{\boldsymbol{w}}(\boldsymbol{X}_i, \boldsymbol{X}_k)}{\partial \omega_l} - \frac{\partial \Xi_{\boldsymbol{w}}(\boldsymbol{X}_i, \boldsymbol{X}_j)}{\partial \omega_l} \right) \right] - 2\lambda \omega_l \tag{18}$$

C. Fault Diagnosis via Wasserstein distance

The proposed methodology is based on two main steps, the 1D p-Wasserstein distance calculation, and the distance weighting. The first step consists of creating a random set of time-segments from the signal to classify (i.e., diagnose). The signal is divided randomly (i.e., i.i.d. data) into time-segments using a fixed periodic window. In this work, the Hamming window is considered as its periodic extension is useful for DFT/FFT purposes. For each time-segment, a feature is extracted, then the obtained feature values are sorted. Note that the specified order, i.e., ascending or descending order, does not affect the dignostic accuracy as long as the specified order is the same both for training and testing. Then, the 1D p-Wasserstein distance is calculated directly from order statistics (see Eq. (7)).

Without the distance weighting strategy, the distance between a training/reference experiment can be evaluated as the sum of all the distances of the domain (e.g., time and/or frequency). This leads to consider irrelevant features which decrease the generalization performance of the algorithm. Considering the feature matrix $\boldsymbol{X}_i$ composed of features, then, the distance between two experiments is given by:

$$\Xi(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sum_{l=1}^{d} \mathcal{D}(\boldsymbol{x}_l^{(s)[i]}, \boldsymbol{x}_l^{(s)[j]}) \tag{19}$$

where $\mathcal{D}$ represents the statistical distance or metric used for fault diagnosis such as Euclidean distance, Kullback-Leibler divergence, Jensen-Shannon divergence, Bhattacharyya distance, Chernoff distance, Hellinger distance, Total variation distance, Kolmogorov distance, and more [18]. Given a testing matrix $\boldsymbol{X}_{test}$, the prediction is given by y_{test}, where $y_{test} = y_j$ and:

$$j^* = \arg\min_{j} \sum_{l=1}^{d} \mathcal{D}(\boldsymbol{x}_l^{(s)[test]}, \boldsymbol{x}_l^{(s)[j]}) \tag{20}$$

$\boldsymbol{x}_l^{(s)[test]}$ is the l-th vector of order statistics of the testing experiment and $\boldsymbol{x}_l^{(s)[j]}$ is the l-th vector of order statistics related to each j-th observation (i.e., faulty or faultless machines), $j \in \{1, \ldots, N\}$.

Considering the distance weighting strategy, once 1D p-Wasserstein distances have been calculated, the distances are weighted in order to maximize the detection of different classes or categories. In this work, differently from the standard NCFS, features are described by matrices of distances instead of vectors, so the NCFS has been adapted to deal with matrices and to exploit the 1D p-Wasserstein distances as discriminative information. Thus, the distance considered in Eq. (9) is:

$$\mathcal{D}(\boldsymbol{x}_l^{(s)[i]}, \boldsymbol{x}_l^{(s)[j]}) = \frac{1}{n} \sum_{k=1}^{n} |x_{kl}^{(s)[i]} - x_{kl}^{(s)[j]}|^p \tag{21}$$

Considering the 1D p-Wasserstein distance in the NCFS framework, the gradient of Eq. (9) with respect to ω_l, that is used for the optimization strategy through gradient descent, is:

$$\frac{\partial \Xi_{\boldsymbol{\omega}}(\boldsymbol{X}_i, \boldsymbol{X}_j)}{\partial \omega_l} = \frac{2\omega_l}{n} \sum_{k=1}^{n} |x_{kl}^{(s)[i]} - x_{kl}^{(s)[j]}|^p \tag{22}$$

Differently from the standard NCFS algorithm, we considered only one nearest neighbor as a reference point (i.e., 1-NN) for each faulty condition as highlighted in Eq. (23). Although this assumption may reduce the generalization performance, it appears to be realistic since in fault diagnosis different observations of faulty cases are difficult to obtain, and often only one observation is available for each faulty condition, thus leading to a small number of faulty training points. Thus, the selection of only one nearest neighbor in our task represents a good compromise in terms of bias-variance trade-off.

The distance weighting algorithm is applied in order to weigh the most discriminative 1D p-Wasserstein distances. Given a testing observation $\boldsymbol{X}_{test}$, the prediction is given by y_{test}, where $y_{test} = y_j$ and:

$$j^* = \arg\min_{j} \sum_{l=1}^{d} \omega_l^{*2} \mathcal{D}(\boldsymbol{x}_l^{(s)[test]}, \boldsymbol{x}_l^{(s)[j]}) =$$
$$= \arg\min_{j} \frac{1}{n} \sum_{l=1}^{d} \omega_l^{*2} \sum_{k=1}^{n} |x_{kl}^{(s)[test]} - x_{kl}^{(s)[j]}|^p \tag{23}$$

where the optimal weight vector $\boldsymbol{\omega}^* = \{\omega_1^{*2}, \ldots, \omega_d^{*2}\}$ is found by the distance weighting algorithm as previously described in Section II-B.

III. EXPERIMENTAL SETUPS

This section describes the experimental results carried out on two benchmarks proposed in the literature. The first benchmark regards the Case Western Reserve University (CWRU) Bearing Data Center [48], whereas the second benchmark regards the Prognostic Health Management (PHM) 2009 Data Challenge [49].

A. Datasets

1) Case Western Reserve University: The CWRU dataset has become a standard reference used to test different diagnostic algorithms, allowing to make an objective and fair comparison of our solution with respect to others [50]. A detailed description of the benchmark can be found in [48]. The acquisition time of the waveforms is 10 s and in this work: the first 5 seconds are used to extract randomly the time-segments during the training stage, and the last 5 seconds are used for the testing stage. In this way no overlapped windows are used both for training and testing. The considered time window durations are $T_W = [0.02, 0.05, 0.0854, 0.1, 0.5]$ s (i.e., 240, 600, 1024, 1200 and 6000 samples, and 256, 1024, 2048, 4096 and 8192 DFT points), whereas the number of time-segments are $n = [10, 20, 50, 100, 200]$. In order to test the proposed fault diagnosis algorithm over different conditions and compare it with related works, two different datasets are considered. The first dataset has a total of 10 faults (i.e., 11 classes are considered as shown in Table I), and all fault classes refer to a fault severity of 0.007. The vibration signals of the second dataset were collected from the drive end of the motor in the test rig for a total of 9 faults (i.e., 10 classes are considered as shown in Table I). For each experiment, the algorithm performances are obtained varying the number of time-segments and the length of the time windows. Moreover, in order to evaluate the robustness of the algorithm, different low SNR values are considered: $+\infty$ dB, -5 dB, -10 dB, -15 dB, -20 dB and -25 dB. SNR is defined as:

$$\text{SNR dB} = 10 \log_{10}(P_{signal}/P_{noise}) \tag{24}$$

where P_{signal} and P_{noise} denote the power of the original signal and the power of the additive white Gaussian noise (AWGN), respectively. Considering Eq. (24), $+\infty$ dB means the original vibration signals is not corrupted by additional Gaussian noise.

TABLE I
CLASSES DEFINITION, WHERE 3OC, 6OC AND 12OC MEAN 3 O'CLOCK, 6 O'CLOCK AND 12 O'CLOCK, RESPECTIVELY.

Class	Label dataset 1	Label dataset 2
1	Faultless	Faultless
2	Ball_007_DE	Ball_007_DE
3	Ball_007_FE	Ball_014_DE
4	InnerRaceway_007_DE	Ball_021_DE
5	InnerRaceway_007_FE	InnerRaceway_007_DE
6	OuterRaceway_6oc_007_DE	InnerRaceway_014_DE
7	OuterRaceway_6oc_007_FE	InnerRaceway_021_DE
8	OuterRaceway_3oc_007_DE	OuterRaceway_6oc_007_DE
9	OuterRaceway_3oc_007_FE	OuterRaceway_6oc_014_DE
10	OuterRaceway_12oc_007_DE	OuterRaceway_6oc_021_DE
11	OuterRaceway_12oc_007_FE	–

2) Prognostic Health Management Data Challenge 2009: In this benchmark, data were acquired from three measuring points of a gearbox; in particular, channel 1 is the input side accelerometer, channel 2 is the output side accelerometer, and channel 3 is the tachometer signal. Two geometries are used, one using spur gears, the other using spiral cut (i.e., helical) gears. In this paper, only the accelerometer signals and data related to helical gears are considered. The accelerations are measured by Endevco sensors with the following specs: 10 mv/g, $\pm 1\%$ error, resonance > 45 kHz and sample rate of $200/3$ kHz. Data were collected at 30, 35, 40, 45 and 50 Hz shaft speed, under high and low loading. In order to test the proposed fault diagnosis algorithm over different conditions and compare it with related works, a total of 5 faults are considered in this benchmark (i.e., 6 classes are considered as shown in Table II). PHM 2009 dataset provides two acquisitions for each condition, i.e., shaft speeds and loading conditions. The first dataset is considered for the training stage and the second dataset is used for the testing stage. For each experiment, the algorithm performances are obtained varying the number of time-segments, $n = [10, 20, 50, 100, 200]$, and the length of the time windows, $T_W = [0.05, 0.1, 0.15, 0.2]$ s (i.e., 3333, 6666, 10000 and 13333 samples, and 8192, 16384, 32768 DFT points). Moreover, in order to evaluate the robustness of the algorithm, different low SNR conditions are considered: $+\infty$ dB, -5 dB, -10 dB, -15 dB, -20 dB, and -25 dB.

TABLE II
CLASSES DEFINITION OF PHM 2009 DATASET.

Class	Faults (location)
1	Healthy
2	Chipped (gear 24T)
3	Broken (gear 24T), Combination (bearing IS:OS), Inner (bearing ID:OS) and Bent Shaft (shaft IS)
4	Combination (bearing IS:OS), ball (ID:OS), Imbalance (shaft IS)
5	Broken (gear 24T) and Inner (bearing ID:OS)
6	Bent Shaft (shaft IS)

B. Performance criteria and hyperparameters setting

The proposed method is compared with related works using both classification accuracy and *macro F1-score*; the latter is the unweighted mean of the F1-score for each label, and then, it does not take label imbalance into account. F1-score is a commonly used criterion measuring the performance of a classification method [51]. The optimization of the algorithm hyperparameters (i.e., kernel scale σ and regularization term λ) was performed by the implementation of a grid search and the optimization of the macro F1-score in nested 10-fold cross-validation on the training dataset. In Section IV-A, the optimization of the hyperparameters returns the optimal regularization term to be 0.05 and the best kernel width 1, whereas, in Section IV-B, the optimization of the hyperparameters returns the optimal regularization term to be 0.05 and the best kernel width 0.01. In the Section IV-C, all the experimental test was performed setting the regularization term λ to 0.05 and the kernel width σ to 1. The step size of the gradient descent is computed iteratively by a line

search strategy based on weak Wolfe condition. The results reported in Sections IV-A and IV-B are carried out by 220 and 120 labeled observations for CWRU and PHM 2009 bearing dataset, respectively, where 50% of the observations are used for training and the remaining for testing. The results reported in Section IV-C are carried out by 40 labeled observations for the training stage, and 300 labeled observations for the testing stage. In order to evaluate the effectiveness of the improvements in terms of macro F1-score caused by the distance weighting strategy with respect to the case without distance weighting, the independent two-sample t-test is considered. In particular, the alternative hypothesis that the population mean without distance weighting is less than the population mean with distance weighting is tested, and it was assumed that the null and alternative hypothesis come from normal distributions with unknown and unequal variances. The hyperparameter p of the 1D p-Wasserstein distance has been set empirically by the optimization of the macro F1-score, namely $p = 1$ for SNR≤ -15, and $p = 2$ for SNR> -15.

IV. RESULTS

In this section, the experimental results carried out with the CWRU and PHM 2009 benchmarks and the comparison with the state-of-the-art are described. The Fault Diagnosis via Wasserstein distance algorithm has been applied as described in Section II-C, with the following considerations:

1) The spectral contents of each time-segment are calculated by using FFT, then adapted into a 2D matrix, which reports the frequencies along one dimension and the amplitudes on the other dimension.
2) Each feature vector has a dimension of n, namely the number of time-segments.
3) The amplitudes of each frequency are sorted: the sort operation is needed to obtain the statistical information of the amplitude distribution of each frequency; indeed, this operation implicitly allows to obtain the order statistics of the amplitude.

A. CWRU results

Fig. 1 shows the comparison of fault classification by 1D p-Wasserstein distance without distance weighting (label W) and with distance weighting (label WNCFS) in terms of macro F1-score at different n, time window T_W and SNR values for the CWRU benchmark. In particular, Fig. 1(a) shows the macro F1-scores of all experiments based on 1D p-Wasserstein distance without distance weighting, whereas Fig. 1(b) shows the macro F1-scores of all experiments based on 1D p-Wasserstein distance with distance weighting. For CWRU benchmark, it is worth noting that the distance weighting improves significantly in terms of macro F1-score the performance of classification at high level of noise as also confirmed by Figs. 3(a) and 4(a). In particular, Fig. 3(a) shows the macro F1-score at different SNR conditions evaluated considering all experimental conditions, whereas, Fig. 4(a) shows the boxplot of all macro F1-scores at different SNR conditions evaluated at different n, time window T_W and speed values. As results, WNCFS improvement in terms of macro F1-score reached statistical significance for

SNR$=-20$ dB ($t_{185.5739} = -3.3445$, $p_{value} < .001$) and SNR$=-25$ dB ($t_{197.7441} = -5.1233$, $p_{value} < .001$) as can be seen in the Figs. 1, 3(a) and 4(a).
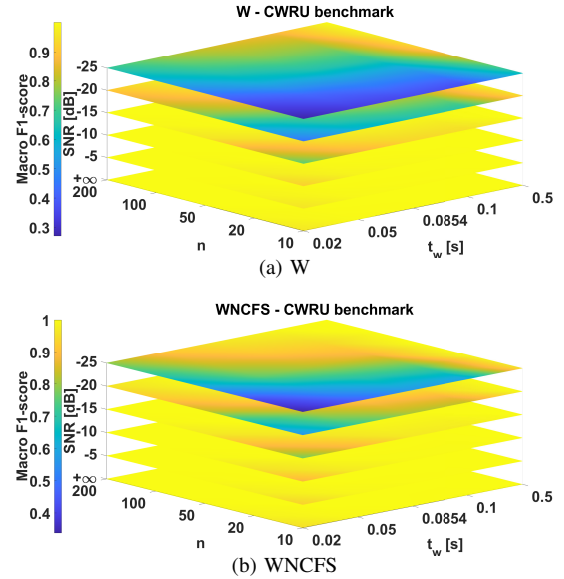

Fig. 1. Macro F1-scores of the CWRU bearing benchmark evaluated at different n, time window T_W and SNR conditions.

B. PHM 2009 results

Fig. 2 shows the comparison of fault classification by 1D p-Wasserstein distance without distance weighting (label W) and with distance weighting (label WNCFS) in terms of macro F1-score at different n, time window T_W and SNR values for the PHM 2009 benchmark. In particular, Figs. 2(a) and 2(c) show the macro F1-scores of all experiments based on 1D p-Wasserstein distance without distance weighting in the cases of low and high loading, respectively, whereas Figs. 2(b) and 2(d) show the macro F1-scores of all experiments based on 1D p-Wasserstein distance with distance weighting. For PHM 2009 benchmark, it is worth to note that the distance weighting improves significantly the performances of classification from medium-low level of noise as also confirmed by Figs. 3(b), 3(c), 4(b) and 4(c). In particular, Figs. 3(b) and 3(c) show the macro F1-score at different SNR conditions evaluated considering all experimental conditions, whereas, Figs. 4(b) and 4(c) show the boxplots of all macro F1-scores at different SNR conditions evaluated at different n, time window T_W and speed values. WNCFS improvement in terms of macro F1-score reached statistical significance for low loading at SNR$=-10$ dB ($t_{162.6453} = -5.4978$, $p_{value} < .001$), -15 dB ($t_{196.2169} = -9.4108$, $p_{value} < .001$), -20 dB ($t_{135.1572} = -10.3941$, $p_{value} < .001$) and -25 dB ($t_{103.2548} = -5.7559$, $p_{value} < .001$), as can be seen in the Figs. 2(a), 2(b), 3(b) and 4(b). Whereas, WNCFS improvement in terms of macro F1-score reached statistical significance for high loading at SNR$=-5$ dB ($t_{148.7976} = -3.7931$, $p_{value} < .001$), SNR$=-10$ dB ($t_{179.3559} = -7.5835$, $p_{value} < .001$), -15 dB ($t_{181.1723} = -10.2911$, $p_{value} < .001$), -20 dB ($t_{112.5063} = -9.3165$, $p_{value} < .001$) and -25 dB
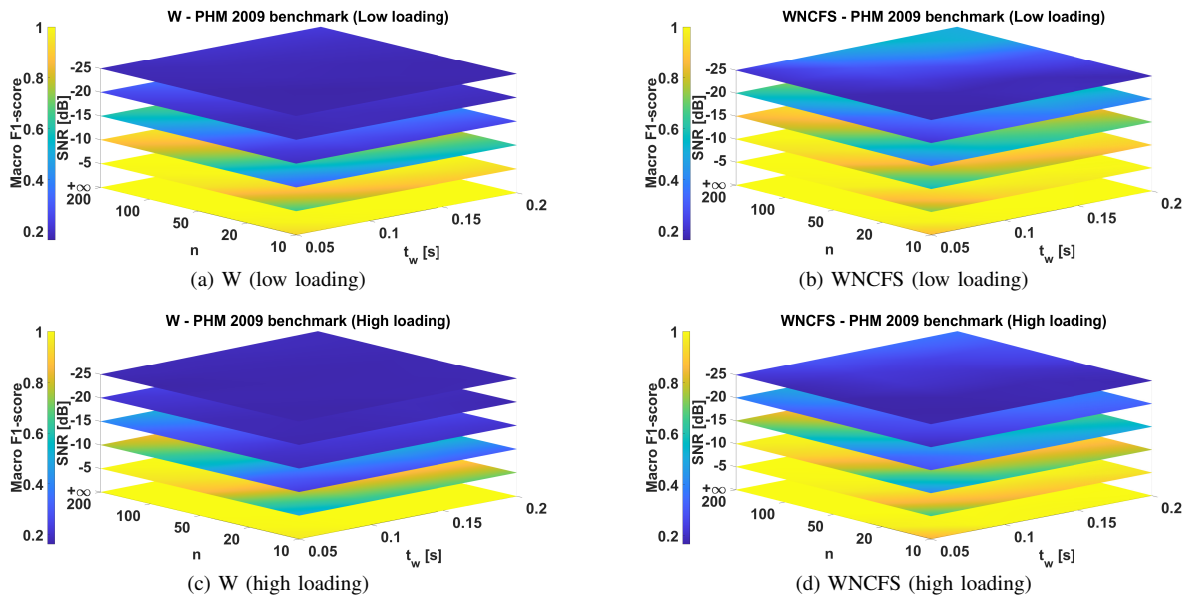
Fig. 2. Macro F1-scores of the PHM 2009 bearing benchmark evaluated at different n, time window T_W and SNR conditions.

$(t_{102.4642} = -5.1762, p_{value} < .001)$, as can be seen in the Figs. 2(c), 2(d), 3(c) and 4(c). A noteworthy aspect is that the algorithm is able to run with 32768 features in the case of $T_W = 0.2$ s.

C. Comparison with state-of-the-art

In order to compare the performance of the proposed algorithm with others presented in the literature, in this section, the results are reported in terms of classification accuracy. In [52], the authors have reported the classification accuracy at an SNR of -10 dB, as the worst case scenario, whereas in [14], the authors have reported the classification accuracy at a SNR of -15 dB. In [22], the authors have reported the classification accuracy at an SNR of -8 dB as the worst case scenario, whereas in [53], the authors have reported the classification accuracy at a SNR of -4 dB. The proposed method is compared using the same experimental settings with those proposed in [14], [22], [53]–[59]; the settings, considered to compare different approaches, provides for a window length of 1024 samples, same machine conditions for the cases of 10 labels and 11 labels and additive noise of Gaussian distribution. The detailed comparison results with the related works are presented in Table III. The table shows the results of diagnosis bearing health condition using the training and testing data from the same domain (i.e., 0,1,2,3→0,1,2,3) and the results with the variation of working condition (e.g., 0→3). It is seen that the proposed method outperforms the other approaches in different scenarios of about 10% - 15%. In addition, only the proposed solution is tested up to SNR=-25 dB and only a few works tested their algorithms over SNR=-8 dB.

D. Time- and frequency-domain based features

The proposed methodology can be generalized to all kinds of features, such as frequency-domain, time-domain or both time- and frequency-domain features. In this section, the experimental results carried out by using the frequency-domain features, defined in the previous experiments, and 12 time-domain features are shown. In detail, the time-domain features considered are: standard deviation, skewness, kurtosis, peak, peak-to-peak value, root mean square, square mean root, crest factor, clearance factor/margin factor, shape factor, impact factor/impulse factor, kurtosis factor [60]. In this case, the algorithm has to process features of different scales, so a feature scaling operation is needed to handle the features of different scales. The feature scaling to unit length is considered, and the features are scaled by their L_1-norms, then $\boldsymbol{x} = \boldsymbol{x}/\|\boldsymbol{x}\|_1$. The experimental results for PHM 2009 (high loading) benchmark are shown in Fig. 5. As shown in Figs. 5(a) and 5(b), in this test case, similar results are obtained in terms of macro F1-score. Considering a significance level of 0.05, the two-sample t-test discloses that the macro F1-score related to frequency-domain features is statistically lower than the macro F1-score related to both time- and frequency-domain features for the case SNR=-25 dB $(t_{189.4309} = -1.8386, p_{value} < .05)$, whereas the macro F1-score related to frequency-domain features, is not statistically greater than the macro F1-score mean related to both time- and frequency-domain features, this means that in this experimental results, the addition of time-domain features does not make the classification accuracy worse. In conclusion, the proposed algorithm based on 1D p-Wasserstein distance can handle features of different scales.

E. Analysis of computational time

The settings considered to analyze the algorithm computational time are the following: CWRU benchmark, 11 labels, 220 labeled observations for training, $\sigma = 1$, $\lambda = 0.05$, SNR=-15 dB, 20 Monte Carlo runs. The training time is the average of the Monte Carlo runs. The simulations are reported varying $T_W = [0.02, 0.05, 0.0854, 0.1, 0.5]$ s (i.e., 240, 600, 1024, 1200 and 6000 samples, and 256, 1024, 2048, 4096
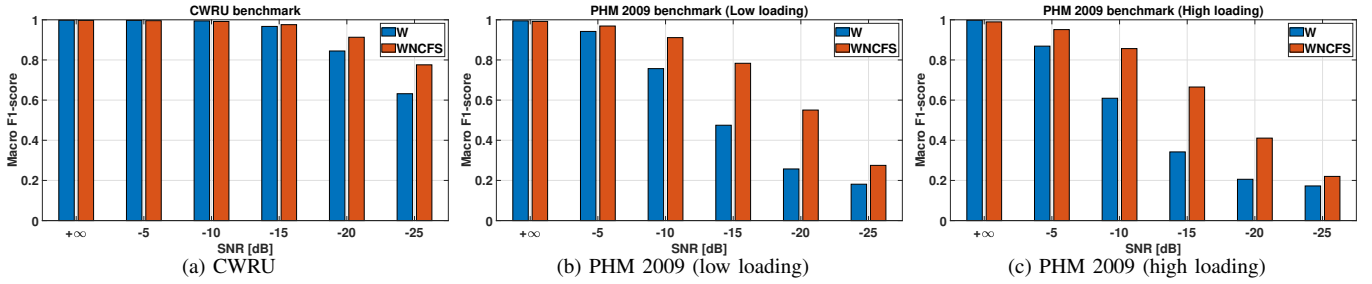
Fig. 3. Fault classification by 1D p-Wasserstein distance without (label W) and with (label WNCFS) distance weighting: macro F1-scores of CWRU and PHM 2009 bearing benchmark.
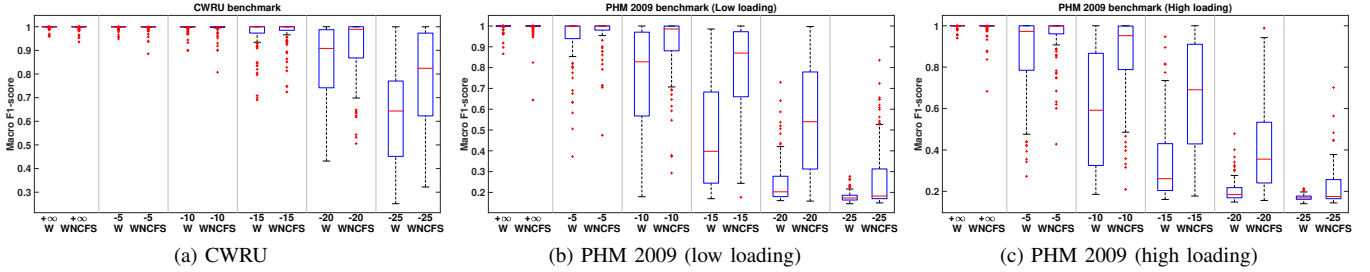

Fig. 4. Fault classification by 1D p-Wasserstein distance without (label W) and with (label WNCFS) distance weighting: boxplots of macro F1-scores of CWRU and PHM 2009 bearing benchmark.
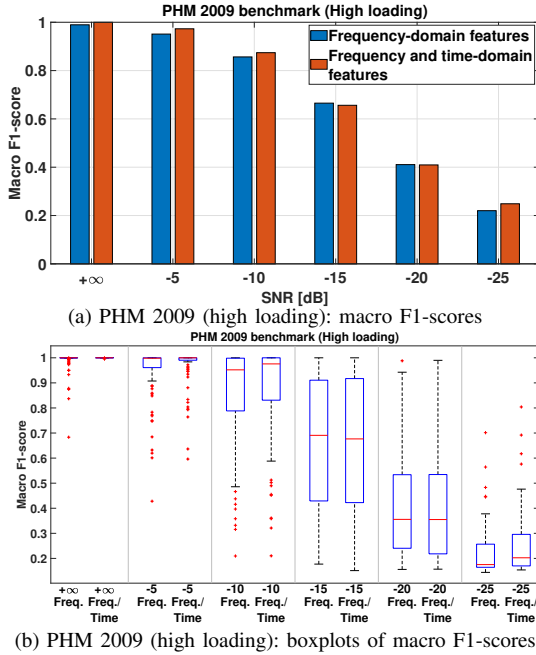

Fig. 5. Fault classification using WNCFS and frequency-domain and both time- and frequency-domain features: mean of macro F1-scores and boxplots of macro F1-scores.

and 8192 DFT points) and $n = [10, 20, 50, 100, 200]$. The analysis includes also the 10-fold CV for the hyperparameters optimization. The platform used to compute the training time is a laptop with CPU Intel 7700HQ, 16GB RAM, Matlab 2019a.

Fig. 6 shows the training time in seconds varying T_W and n. In the case of CWRU benchmark, fairly high testing diagnosis accuracy is achieved by the proposed method with 8192 samples and $n = 200$, and the average training time is around
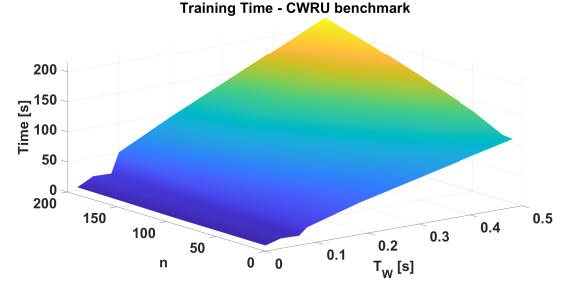

Fig. 6. Training time related to the CWRU bearing benchmark evaluated at different n and time window T_W

216 s. The computing burden of the training algorithm is considered low, at the same time, outperforming deep learning approaches in terms of classification accuracy [22].

For the sake of completeness, in Fig. 7 the sparsity percentage considering a threshold of 10^{-4} is reported. The sparsity measure was evaluated according to [46] by counting the number of zero weights of the model (l^0 measure). It is worth noting as the sparsity depends on the number of features related to T_W and it is almost independent of the number of time segments n.

Finally, considering the aforementioned settings, the testing time varies from 10^{-5} s in the best scenario to 0.03 s in the worst scenario with $n = 200$ and $T_W = 0.5$ s as shown in Fig. 8.

V. CONCLUSION

In this work, a fault diagnosis algorithm for rotating machinery based on Wasserstein distance is proposed. The Wasserstein distance is considered for the learning phase to discriminate the different machine operating conditions. Specifically, the 1D Wasserstein distance is taken into account

TABLE III
COMPARISONS OF RELATED RESEARCHES ON THE CWRU ROLLING BEARING DATASET. THE COMPARISON CONSIDERS THE DEFAULT EXPERIMENTAL SETTING WITH SAMPLE LENGTH OF 1024 SAMPLES ($T_W = 0.0854$ s)

| Method	# Classes	Testing accuracy (%)	Motor loads (Training→Testing)	SNR (dB)
[57]	4	95.8	0,1,2,3→0,1,2,3	$+\infty$
[59]	10	88.9	0,1,2,3→0,1,2,3	$+\infty$
[56]	10	92.5	0,1,2,3→0,1,2,3	$+\infty$
[58]	11	97.91	0,1,2,3→0,1,2,3	$+\infty$
[55]	10	99.66	0,1,2,3→0,1,2,3	$+\infty$
[22]	10	100.0	0,1,2,3→0,1,2,3	$+\infty$
Proposed	10	100.0	0,1,2,3→0,1,2,3	$+\infty$
Proposed	11	100.0	0,1,2,3→0,1,2,3	$+\infty$
[53]	10	82.05	1,2,3→1,2,3	-4
[22]	10	96.53	0,1,2,3→0,1,2,3	-4
[22]	10	74.90	0,1,2,3→0,1,2,3	-8
[14]	4	85.15	2→2	-15
Proposed	10	100.0, 100.0, 99.69, 83.21, 50.20	0,1,2,3→0,1,2,3	-5, -10, -15, -20, -25
Proposed	11	100.0, 100.0, 99.76, 87.29, 56.83	0,1,2,3→0,1,2,3	-5, -10, -15, -20, -25
[54]	4	94.73	0→3	$+\infty$
[53]	10	91.10	1→3	$+\infty$
[53]	10	90.20	3→1	$+\infty$
[22]	10	99.43	0→3	$+\infty$
[22]	10	97.82	3→0	$+\infty$
[22]	10	84.82	0→3	-4
[22]	10	84.45	3→0	-4
Proposed	10	99.87, 99.77, 96.20, 71.78, 44.64	0→3	-5, -10, -15, -20, -25
Proposed	10	99.78, 99.41, 99.66, 85.18, 47.47	0→2	-5, -10, -15, -20, -25
Proposed	10	100.0, 100.0, 99.82, 71.53, 46.08	0→1	-5, -10, -15, -20, -25
Proposed	10	99.66, 99.93, 99.30, 83.97, 46.24	1→0	-5, -10, -15, -20, -25
Proposed	10	100.0, 100.0, 99.90, 78.90, 47.68	1→2	-5, -10, -15, -20, -25
Proposed	10	99.93, 99.80, 99.01, 79.53, 46.67	1→3	-5, -10, -15, -20, -25
Proposed	10	100.0, 100.0, 99.44, 71.77, 45.44	2→0	-5, -10, -15, -20, -25
Proposed	10	99.83, 99.55, 97.85, 71.13, 45.40	2→1	-5, -10, -15, -20, -25
Proposed	10	100.0, 100.0, 99.87, 78.55, 45.79	2→3	-5, -10, -15, -20, -25
Proposed	10	99.93, 99.41, 96.57, 72.43, 43.09	3→0	-5, -10, -15, -20, -25
Proposed	10	100.0, 100.0, 99.98, 83.25, 47.36	3→1	-5, -10, -15, -20, -25
Proposed	10	100.0, 100.0, 99.44, 68.97, 45.97	3→2	-5, -10, -15, -20, -25
Proposed	11	100.0, 100.0, 98.55, 74.04, 40.78	0→1	-5, -10, -15, -20, -25
Proposed	11	100.0, 99.86, 97.08, 75.50, 44.57	1→0	-5, -10, -15, -20, -25
Proposed	11	99.97, 100.0, 98.93, 77.18, 40.87	0→2	-5, -10, -15, -20, -25
Proposed	11	99.99, 99.74, 97.67, 74.91, 43.13	2→0	-5, -10, -15, -20, -25
Proposed	11	99.78, 100.0, 99.59, 85.71, 44.10	0→3	-5, -10, -15, -20, -25
Proposed	11	99.96, 99.79, 97.82, 74.61, 42.74	3→0	-5, -10, -15, -20, -25
Proposed	11	100.0, 99.98, 99.52, 79.53, 43.80	1→2	-5, -10, -15, -20, -25
Proposed	11	100.0, 100.0, 98.59, 73.56, 41.85	2→1	-5, -10, -15, -20, -25
Proposed	11	100.0, 100.0, 99.73, 86.28, 44.09	1→3	-5, -10, -15, -20, -25
Proposed	11	100.0, 100.0, 98.89, 72.97, 42.03	3→1	-5, -10, -15, -20, -25
Proposed	11	99.86, 100.0, 99.97, 87.66, 47.88	2→3	-5, -10, -15, -20, -25
Proposed	11	100.0, 100.0, 99.55, 74.97, 43.56	3→2	-5, -10, -15, -20, -25
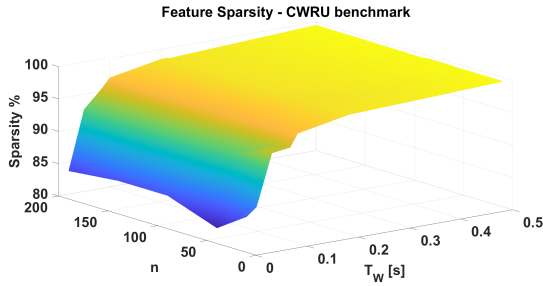
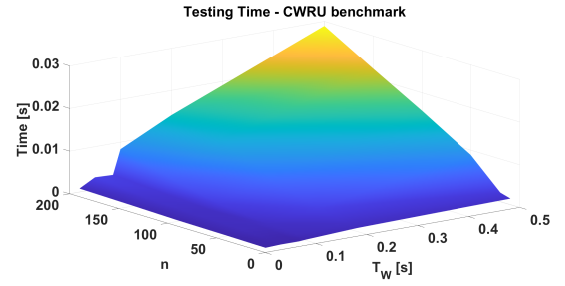Fig. 7. Sparsity related to the CWRU bearing benchmark evaluated at different n and time window T_W

Fig. 8. Testing time related to the CWRU bearing benchmark evaluated at different n and time window T_W

thanks to its low computational burden, due to the fact that it can be evaluated directly by the order statistics. The 1D Wasserstein distance has been exploited as the loss function to be optimized in the NCFS framework, to improve the generalization performances in the presence of low SNR conditions and high dimensional features set. Experiments are conducted on two benchmark datasets, CWRU bearing dataset and PHM 2009 dataset, to verify the effectiveness of the

proposed fault diagnosis method at different SNR conditions. Results have shown that the proposed fault diagnosis method is effective to learn the complex known and unknown patterns with low SNR conditions, many classes and different operating conditions. The authors are currently considering two possible future developments for the proposed fault diagnosis strategy. The former is related to the extension of the proposed approach to deal with nonstationary conditions. Since the assumption of

the algorithm is to process quasi-stationary vibration signals, the performance of the algorithm in the case of nonstationary conditions could be not satisfying. The latter is related to the setting of the parameter p of the 1D p-Wasserstein distance. The parameter could also be regulated by the gradient descent optimization method.

REFERENCES

[1] Z. Gao, C. Cecati, and S. X. Ding, "A Survey of Fault Diagnosis and Fault-Tolerant Techniques—Part I: Fault Diagnosis With Model-Based and Signal-Based Approaches," *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3757–3767, 2015.

[2] F. Ferracuti, A. Giantomassi, S. Iarlori, G. Ippoliti, and S. Longhi, "Electric motor defects diagnosis based on kernel density estimation and Kullback–Leibler divergence in quality control scenario," *Engineering Applications of Artificial Intelligence*, vol. 44, pp. 25–32, 2015.

[3] X. Dai and Z. Gao, "From Model, Signal to Knowledge: A Data-Driven Perspective of Fault Detection and Diagnosis," *IEEE Trans. Ind. Informat.*, vol. 9, no. 4, pp. 2226–2238, 2013.

[4] A. Giantomassi, F. Ferracuti, S. Iarlori, G. Ippoliti, and S. Longhi, "Electric Motor Fault Detection and Diagnosis by Kernel Density Estimation and Kullback-Leibler Divergence Based on Stator Current Measurements," *IEEE Trans. Ind. Electron.*, vol. 62, no. 3, pp. 1770–1780, 2015.

[5] Y. Wang, J. Xiang, R. Markert, and M. Liang, "Spectral kurtosis for fault detection, diagnosis and prognostics of rotating machines: A review with applications," *Mech. Syst. Signal Process.*, vol. 66, pp. 679–698, 2016.

[6] R. Liu, B. Yang, E. Zio, and X. Chen, "Artificial intelligence for fault diagnosis of rotating machinery: A review," *Mech. Syst. Signal Process.*, vol. 108, pp. 33–47, 2018.

[7] A. K. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mech. Syst. Signal Process.*, vol. 20, no. 7, pp. 1483 – 1510, 2006.

[8] V. Do and U.-P. Chong, "Signal Model-Based Fault Detection and Diagnosis for Induction Motors Using Features of Vibration Signal in Two-Dimension Domain," *Strojniški vestnik - Journal of Mechanical Engineering*, vol. 57, no. 9, 2011.

[9] G. G. Yen and K. C. Lin, "Wavelet packet feature extraction for vibration monitoring," *IEEE Trans. Ind. Electron.*, vol. 47, no. 3, pp. 650–667, 2000.

[10] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, and D. Siegel, "Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications," *Mech. Syst. Signal Process.*, vol. 42, no. 1, pp. 314 – 334, 2014.

[11] R. Yan, R. X. Gao, and X. Chen, "Wavelets for fault diagnosis of rotary machines: A review with applications," *Signal Processing*, vol. 96, pp. 1 – 15, 2014.

[12] Z. Feng, M. Liang, and F. Chu, "Recent advances in time-frequency analysis methods for machinery fault diagnosis: A review with application examples," *Mech. Syst. Signal Process.*, vol. 38, no. 1, pp. 165 – 205, 2013.

[13] W. Fan, Q. Zhou, J. Li, and Z. Zhu, "A Wavelet-Based Statistical Approach for Monitoring and Diagnosis of Compound Faults With Application to Rolling Bearings," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 4, pp. 1563–1572, 2018.

[14] M. Amar, I. Gondal, and C. Wilson, "Vibration Spectrum Imaging: A Novel Bearing Fault Classification Approach," *IEEE Trans. Ind. Electron.*, vol. 62, no. 1, pp. 494–502, 2015.

[15] S. H. Kia, H. Henao, and G. A. Capolino, "A High-Resolution Frequency Estimation Method for Three-Phase Induction Machine Fault Detection," *IEEE Trans. Ind. Electron.*, vol. 54, no. 4, pp. 2305–2314, 2007.

[16] X. Zhang, J. Kang, L. Xiao, J. Zhao, and H. Teng, "A new improved kurtogram and its application to bearing fault diagnosis," *Shock and Vibration*, vol. 2015, 2015.

[17] L. Wang, Z. Liu, Q. Miao, and X. Zhang, "Time-frequency analysis based on ensemble local mean decomposition and fast kurtogram for rotating machinery fault diagnosis," *Mech. Syst. Signal Process.*, vol. 103, pp. 60–75, 2018.

[18] L. Ciabattoni, F. Ferracuti, A. Freddi, and A. Monteriù, "Statistical Spectral Analysis for Fault Diagnosis of Rotating Machines," *IEEE Trans. Ind. Electron.*, vol. 65, no. 5, pp. 4301–4310, 2018.

[19] C. Vununu, K.-S. Moon, S.-H. Lee, and K.-R. Kwon, "A Deep Feature Learning Method for Drill Bits Monitoring Using the Spectral Analysis of the Acoustic Signals," *Sensors*, vol. 18, no. 8, 2018.

[20] B. Wang, T. Fujinaka, S. Omatu, and T. Abe, "Automatic Inspection of Transmission Devices Using Acoustic Data," *IEEE Trans. Autom. Sci. Eng.*, vol. 5, no. 2, pp. 361–367, 2008.

[21] S. Langarica, C. Rüffelmacher, and F. Núñez, "An Industrial Internet Application for Real-Time Fault Diagnosis in Industrial Motors," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 1, pp. 284–295, 2020.

[22] X. Li, W. Zhang, and Q. Ding, "A robust intelligent fault diagnosis method for rolling element bearings based on deep distance metric learning," *Neurocomputing*, vol. 310, pp. 77 – 95, 2018.

[23] W. Deng, H. Liu, J. Xu, H. Zhao, and Y. Song, "An Improved Quantum-Inspired Differential Evolution Algorithm for Deep Belief Network," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 10, pp. 7319–7327, 2020.

[24] H. Zhao, H. Liu, J. Xu, and W. Deng, "Performance Prediction Using High-Order Differential Mathematical Morphology Gradient Spectrum Entropy and Extreme Learning Machine," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 7, pp. 4165–4172, 2020.

[25] S. T. Rachev, "The Monge-Kantorovich Mass Transference Problem and Its Stochastic Applications," *Theory of Probability & Its Applications*, vol. 29, no. 4, pp. 647–676, 1985.

[26] C. Villani, *Topics in Optimal Transportation*, ser. Graduate studies in mathematics. American Mathematical Society, 2003.

[27] ——, *Optimal Transport: Old and New*, ser. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.

[28] C. L. Mallows, "A Note on Asymptotic Joint Normality," *The Annals of Mathematical Statistics*, vol. 43, no. 2, pp. 508–515, 1972.

[29] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *Int. J. Comput. Vision*, vol. 40, no. 2, p. 99–121, 2000.

[30] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 2017, pp. 214–223.

[31] M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, no. 4, pp. 349 – 369, 1989.

[32] ——, "Divergence measures for statistical data processing—An annotated bibliography," *Signal Processing*, vol. 93, no. 4, pp. 621 – 633, 2013.

[33] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. USA: Prentice-Hall, Inc., 1993.

[34] J. Harmouche, C. Delpha, D. Diallo, and Y. Le Bihan, "Statistical Approach for Nondestructive Incipient Crack Detection and Characterization Using Kullback-Leibler Divergence," *IEEE Trans. Rel.*, vol. 65, no. 3, pp. 1360–1368, 2016.

[35] X. Zhang, C. Delpha, and D. Diallo, "Incipient fault detection and estimation based on Jensen–Shannon divergence in a data-driven approach," *Signal Processing*, vol. 169, p. 107410, 2020.

[36] C. Cheng, B. Zhou, G. Ma, D. Wu, and Y. Yuan, "Wasserstein distance based deep adversarial transfer learning for intelligent fault diagnosis with unlabeled or insufficient labeled data," *Neurocomputing*, vol. 409, pp. 35 – 45, 2020.

[37] E. Cazelles, A. Robert, and F. Tobar, "The wasserstein-fourier distance for stationary time series," *IEEE Trans. Signal Process.*, vol. 69, pp. 709–721, 2021.

[38] E. Levina and P. Bickel, "The Earth Mover's distance is the Mallows distance: some insights from statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, 2001, pp. 251–256.

[39] A. Ramdas, N. Trillos, and M. Cuturi, "On wasserstein two-sample testing and related families of nonparametric tests," *Entropy*, vol. 19, no. 2, pp. 1–15, 2017.

[40] A. M. Mood and F. A. Graybill, *Introduction to the theory of statistics*, ser. International student edition. New York: McGraw-Hill, 1963.

[41] D. He, X. Xu, and J. Zhao, "A new procedure for testing normality based on the L2 Wasserstein distance," *Journal of Systems Science and Complexity*, vol. 26, pp. 572–582, 2013.

[42] E. del Barrio, J. A. Cuesta-Albertos, C. Matran, and J. M. Rodriguez-Rodriguez, "Tests of Goodness of Fit Based on the L2-Wasserstein Distance," *The Annals of Statistics*, vol. 27, no. 4, pp. 1230–1239, 1999.

[43] S. Kolouri, P. E. Pope, C. E. Martin, and G. K. Rohde, "Sliced-Wasserstein Autoencoder: An Embarrassingly Simple Generative Model," *preprint arXiv:1804.01947*, 2018.

[44] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, 1998, pp. 59–66.

[45] W. Yang, K. Wang, and W. Zuo, "Neighborhood Component Feature Selection for High-Dimensional Data," *Journal of Computers*, vol. 7, no. 1, 2012.

[46] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4723–4741, 2009.

[47] Z. Zhao, S. Wu, B. Qiao, S. Wang, and X. Chen, "Enhanced Sparse Period-Group Lasso for Bearing Fault Diagnosis," *IEEE Trans. Ind. Electron.*, vol. 66, no. 3, pp. 2143–2153, 2019.

[48] Case Western Reserve University Bearing Data Center, "Bearing Data Center," (http://csegroups.case.edu/bearingdatacenter/home), 2018.

[49] PHM09 Data Challenge, "Prognostic Health Management Challenge," (http://www.phmsociety.org/references/datasets), 2018.

[50] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study," *Mech. Syst. Signal Process.*, vol. 64, pp. 100 – 131, 2015.

[51] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *International Journal of Machine Learning Technology*, vol. 2, no. 1, pp. 37–63, 2011.

[52] M. F. Yaqub, I. Gondal, and J. Kamruzzaman, "Inchoate Fault Detection Framework: Adaptive Selection of Wavelet Nodes and Cumulant Orders," *IEEE Trans. Instrum. Meas.*, vol. 61, no. 3, pp. 685–695, 2012.

[53] W. Zhang, C. Li, G. Peng, Y. Chen, and Z. Zhang, "A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load," *Mech. Syst. Signal Process.*, vol. 100, pp. 439 – 453, 2018.

[54] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, and T. Zhang, "Deep Model Based Domain Adaptation for Fault Diagnosis," *IEEE Trans. Ind. Electron.*, vol. 64, no. 3, pp. 2296–2305, 2017.

[55] 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, 2016.

[56] X. Jin, M. Zhao, T. W. S. Chow, and M. Pecht, "Motor Bearing Fault Diagnosis Using Trace Ratio Linear Discriminant Analysis," *IEEE Trans. Ind. Electron.*, vol. 61, no. 5, pp. 2441–2451, 2014.

[57] W. Li, S. Zhang, and G. He, "Semisupervised Distance-Preserving Self-Organizing Map for Machine-Defect Detection and Classification," *IEEE Trans. Instrum. Meas.*, vol. 62, no. 5, pp. 869–879, 2013.

[58] X. Zhang, Y. Liang, J. Zhou, and Y. Zang, "A novel bearing fault diagnosis model integrated permutation entropy, ensemble empirical mode decomposition and optimized SVM," *Measurement*, vol. 69, pp. 164–179, 2015.

[59] W. Du, J. Tao, Y. Li, and C. Liu, "Wavelet leaders multifractal features based fault diagnosis of rotating mechanism," *Mech. Syst. Signal Process.*, vol. 43, no. 1-2, pp. 57–75, 2014.

[60] L. Ciabattoni, G. Cimini, F. Ferracuti, M. Grisostomi, G. Ippoliti, and M. Pirro, "Bayes error based feature selection: An electric motors fault detection case study," in *IECON 2015 - 41st Annual Conference of the IEEE Industrial Electronics Society*, 2015, pp. 003 893–003 898.

Francesco Ferracuti received the Ph.D. degree in Automation, Information and Management Engineering from Università Politecnica delle Marche, Ancona, Italy, in 2014. He is Assistant Professor at Department of Information Engineering, Università Politecnica delle Marche. His research interests include model-based and data-driven fault diagnosis, signal processing, statistical pattern recognition, data-driven model identification and their applications in industry.

Alessandro Freddi is Assistant Professor at Università Politecnica delle Marche, Ancona, Italy, where he received the Ph.D. in Automation, Information and Management Engineering in 2012. His main research activities cover fault diagnosis and fault-tolerant control with applications to robotics and industrial systems, and development and application of assistive technologies. He published more than 80 papers in international journals and conferences, was co-editor of 6 books and is involved both in national and international research projects.

Andrea Monteriù (S'04-M'06) received the M.Sc. degree in Electronic Engineering and the Ph.D. degree in Artificial Intelligence Systems from Università Politecnica delle Marche, Italy, in 2003 and 2006. He is now an associate professor at Università Politecnica delle Marche. He serves as Vice-Chair for the IEEE Italy Section CE Soc Chapter, and as Chair of the CTSoc Technical Stream on Consumer Systems for Healthcare and Wellbeing. Monteriù's research interests mainly focus on the areas of fault diagnosis, fault tolerant control, nonlinear dynamics and control, periodic and stochastic system control, applied in different fields including aerospace, marine, robotic and artificial intelligent systems.

Luca Romeo received the Ph.D. degree in Automation, Information and Management Engineering from Università Politecnica delle Marche, in 2018. He is currently a Postdoctoral Researcher with the Department of Information Engineering, Università Politecnica delle Marche. He is also affiliated with the Unit of Cognition, Motion and Neuroscience and Computational Statistics and Machine Learning, Fondazione Istituto Italiano di Tecnologia, Genova. His research interests include machine learning applied to biomedical applications and affective computing and motion analysis.