Information Fusion 104 (2024) 102152

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

Multivariate multiscale dispersion Lempel–Ziv complexity for fault diagnosis of machinery with multiple channels

Shun Wang [a], Yongbo Li [a,*], Khandaker Noman [b], Zhixiong Li [c], Ke Feng [d], Zheng Liu [e], Zichen Deng [a]

[a] School of Aeronautics, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
[b] School of Civil Aviation, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
[c] Opole University of Technology, Faculty of Mechanical Engineering, Opole, 45-758, Poland
[d] Department of Industrial Systems Engineering and Management, National University of Singapore, 117576, Singapore
[e] School of Engineering, University of British Columbia, Kelowna V1V 1V7, Canada

ARTICLE INFO

Keywords:
Multichannel signal analysis
Lempel–Ziv complexity
Nonlinear dynamics
Feature extraction
Fault diagnosis

ABSTRACT

Lempel–Ziv complexity (LZC), as a nonlinear feature in information science, has shown great promise in detecting correlations and capturing dynamic changes in single-channel time series. However, its application to multichannel data has been largely unexplored, while the complexity of real-world systems demands the utilization of data collected from multiple sensors or channels so as to extract distinguishable fault features for fault diagnosis. This paper proposes a novel method called multivariate multiscale dispersion Lempel–Ziv complexity (mvMDLZC) to extract the fault features hidden in multi-source information. First, multivariate embedding theory is applied to obtain multivariate embedded vectors and multivariate dispersion patterns, which can reflect the inherent relationships in the multichannel series. Second, by assigning labels to these patterns, the original multichannel time series can be transformed into a symbolic sequence with multiple symbols instead of the original binary conversion, enabling the accurate recovery of the system dynamics. Finally, the complexity counter value and normalized LZC are calculated for the complexity measure. Experimental results using synthetic and real-world datasets demonstrate that mvMDLZC outperforms existing LZC-based methods and multivariate dispersion entropy in recognizing different states of mechanical systems. Additionally, mvMDLZC exhibits robustness in handling challenges such as small sample datasets and noise interference, making it suitable for real industrial applications. These findings highlight the potential of mvMDLZC as a valuable approach for dissecting multichannel systems across various real-world scenarios.

1. Introduction

Rotating machinery plays a critical role in various industries, such as transportation, power generation, aerospace, and automotive. However, the harsh operating environments and continuous usage make rotating machinery susceptible to failures and faults, which can lead to high maintenance costs and, in severe cases, accidents [1–3]. To address these challenges, the field of condition monitoring and fault diagnosis for rotating machinery has witnessed significant advancements [4,5].

In recent years, entropy and complexity measures have emerged as useful approaches in this domain [6–10]. These measures provide valuable insights into the dynamic behavior and health condition of rotating machinery. When faults or damages occur, they often manifest as changes in the amplitude and frequency modulation of signals. These changes introduce variations and complexities in the measured data. By leveraging entropy and complexity measures, it becomes possible to capture and analyze these dynamic changes [11,12].

Entropy-based metrics have gained significant attention since Shannon introduced the concept of entropy in 1948 [13]. Entropy provides a measure of the complexity or uncertainty of time series data, with higher entropy values indicating more complex signals. This makes entropy-based metrics effective for analyzing nonlinear and irregular signals. Over the years, researchers have developed various entropy methods [11], including Rényi entropy, conditional entropy, Kolmogorov–Sinai entropy, Eckmann–Ruelle entropy, approximate entropy, sample entropy, permutation entropy, fuzzy entropy, distribution entropy, dispersion entropy, diversity entropy [14] and so on [15–18].

* Corresponding author.
E-mail addresses: wangshun@mail.nwpu.edu.cn (S. Wang), yongbo@nwpu.edu.cn (Y. Li), khandakernoman93@nwpu.edu.cn (K. Noman), zhixiong.li@yonsei.ac.kr (Z. Li), ke.feng@outlook.com.au (K. Feng), zheng.liu@ubc.ca (Z. Liu), dweifan@nwpu.edu.cn (Z. Deng).

https://doi.org/10.1016/j.inffus.2023.102152
Received 30 June 2023; Received in revised form 3 November 2023; Accepted 20 November 2023
Available online 23 November 2023

Nomenclature

LZC	Lempel–Ziv complexity
mvDE	Multivariate dispersion entropy
mvDLZC	Multivariate dispersion Lempel–Ziv complexity
mvLZC	Multivariate Lempel–Ziv complexity
mvMDE	Multivariate multiscale dispersion entropy
mvMDLZC	Multivariate multiscale dispersion Lempel–Ziv complexity
mvMLZC	Multivariate multiscale Lempel–Ziv complexity
SD	Standard deviation
SNR	Signal-to-noise ratio
WGN	White Gaussian noise

Additionally, based on coarse-graining analysis [19], multiscale-based entropy methods have been developed for comprehensive complexity evaluation of time series under the different time scales [20–24].

In parallel to the concept of entropy, Lempel–Ziv complexity (LZC) has also gained popularity as a feature extraction technique due to its simplicity [25]. LZC is specifically designed to be parameter-free, eliminating the need for manual tuning and parameter optimization. This characteristic allows for a more streamlined and efficient application of LZC in various data analysis tasks. LZC quantifies the number of new patterns encountered in a time series, providing insights into its complexity [26]. It has found applications in various domains, including fault detection [27–30], biomedical signal processing [31–33], chaos analysis [34], and others [35]. A higher LZC value indicates the presence of more patterns in the time series, indicating greater complexity. LZC offers the advantage of being computationally straightforward and not requiring parameter settings. However, the process of converting the original signal into a 0–1 sequence for computation purposes results in a loss of time series information [35,36].

To address the limitations of LZC-based methods and to leverage the benefits of entropy analysis, researchers have attempted to combine some concepts from entropy measures with LZC. These efforts have led to the development of LZC approaches that aim to accurately assess the complexity of time-series and capture information content. For instance, Bai et al. introduced permutation entropy (PE) into LZC and proposed permutation LZC (PLZC) [31]. By replacing the binary mapping of LZC with permutation patterns from PE, PLZC demonstrated improved anti-interference ability and exhibited promising performance in detecting and analyzing EEG signals. In the field of fault diagnosis for railway vehicle systems, Li et al. combined the maximum entropy partitioning (MEP) with LZC [36]. This approach leveraged the synergies between MEP symbolization and LZC, leading to effective fault diagnosis outcomes. Mao et al. adopted the normal cumulative distribution function (NCDF) from dispersion entropy to replace the binary mapping of LZC [34]. This modification aimed to enhance the robustness of LZC against noise interference. Building upon this, Li et al. further improved the approach by integrating dispersion pattern and fluctuation-based dispersion pattern with LZC [8–10]. The results demonstrated the effectiveness of these methods in detecting dynamic changes in time series data.

Additionally, researchers have explored multiscale-based LZC methods [37,38] and hierarchical-based LZC methods [39,40] to capture complex patterns at different scales and hierarchical levels, respectively. These approaches provide more comprehensive insights into the complexity of time series data. These advancements in incorporating entropy measures, multiscale analysis, and hierarchical analysis into LZC have expanded its applicability and improved its performance in various domains, showcasing the continuous efforts to enhance the capabilities of LZC-based methods.

The LZC-based methods mentioned above are primarily designed for analyzing univariate time series. However, because of the complexity of the targets or systems, the data collected from a single sensor is not enough in decision-making process [41]. In contrast, multivariate signals obtained from multiple sensors or channels contain valuable information that can significantly enhance our understanding of the state of dynamical systems [42]. By considering the interactions and relationships among different channels, we can more effectively detect dynamic changes and achieve accurate fault diagnosis [43,44].

Although a multivariate LZC method (mvLZC) and its multiscale version (mvMLZC) have been proposed in the literature [45], they have limitations. These methods only average the complexity values of all channels, neglecting the inherent relationships in the data. Furthermore, their use of binary encoding fails to account for the underlying signal dynamics, potentially leading to inaccurate system representation.

To address these limitations, we propose a novel approach called multivariate dispersion Lempel–Ziv complexity (mvDLZC), inspired by a recent study [46]. The proposed mvDLZC extends the Lempel–Ziv complexity to the multivariate domain by incorporating the concept of multivariate dispersion pattern [46]. The construction of multivariate dispersion patterns involves applying multivariate embedding theory to obtain multivariate embedded vectors and multivariate dispersion patterns, which can reflect the inherent relationships in the multichannel series [47]. Additionally, by assigning labels to these patterns, the original multivariate time series can be transformed into a symbolic sequence with multiple symbols, enabling the accurate recovery of the system dynamics. Moreover, we extend mvDLZC to the multiscale space, termed as multivariate multiscale dispersion Lempel–Ziv complexity (mvMDLZC), to capture comprehensive feature information.

To validate the effectiveness of the proposed mvDLZC and mvMDLZC methods, we conduct systematic comparative studies using both synthetic and real-world datasets. Through statistical analysis and machine learning techniques, we demonstrate the superiority of our approach in detecting and differentiating complex signals, as well as its performance in fault diagnosis tasks. Experimental results highlight the superiority of mvMDLZC in detecting dynamic changes in time series and achieving the best performance in recognizing different fault states when compared to univariate MLZC, mvMLZC, and mvMDE. Overall, our proposed mvDLZC and mvMDLZC methods provide a comprehensive framework for analyzing multivariate time series data. The experimental results validate the superior performance of the proposed methods compared to existing LZC-based approaches, showcasing their potential in fault diagnosis applications.

The main contributions of this work can be summarized as follows:

(1) To capture the characteristics from multichannel or multivariate systems, multivariate dispersion Lempel–Ziv complexity is proposed to extend the original Lempel–Ziv complexity to multivariate form.

(2) The multivariate dispersion Lempel–Ziv complexity is further extended to multiple time scales, namely mvMDLZC, for comprehensive feature extraction.

(3) The effectiveness of the proposed mvMDLZC method is systematically validated through comparative studies using both synthetic and real-world multichannel signals.

The remainder of this paper is organized as follows: In Section 2, we provide a detailed explanation of the original LZC, the proposed mvDLZC, and its multiscale version, mvMDLZC. Section 3 presents the results of synthetic signal experiments to showcase the effectiveness of our proposed methods. In Section 4, we apply mvDLZC and mvMDLZC to analyze real-world mechanical signals, demonstrating their superiority over existing approaches in fault diagnosis applications. Finally, in Section 5, we summarize our findings and discuss future research directions.

2. Theory

2.1. Lempel–Ziv complexity

The LZC algorithm consists of two fundamental operations: copy and insert [25]. Here is a detailed description of the LZC algorithm:

Step 1 Mathematically, convert the finite sequence $x(t)$ into a symbolic sequence $S_N = \{s_1 s_2 ... s_N\}$ by comparing it with the threshold value (median value T_d) according to Eq. (1). By applying this process, the sequence $x(t)$ is converted into a symbol series represented by 0 and 1.

$$s_i = \begin{cases} 0, \text{if } x(i) < T_d \\ 1, \text{otherwise} \end{cases} \tag{1}$$

Step 2 From the resulting symbolic sequence S_N, the number of distinctive patterns is identified by parsing it from left to right. Set the initial value $S_{v,0} = \{\}$, $Q_0 = \{\}$, $C_N(0) = 0$, and $i = 1$. Note that S_v and Q represent the substrings of the symbol series S_N, and C_N represents complexity counter.

Step 3 Let $Q_i = \{Q_{i-1} s_i\}$ and check if Q_i is already present in the set of $S_{v,i-1} = \{S_{v,i-2} s_{i-1}\}$. If Q_i exists in $S_{v,i-1} = \{S_{v,i-2} s_{i-1}\}$, set $C_N(i) = C_N(i-1)$ and $i = i + 1$. Otherwise, set $Q_i = \{\}$, $C_N(i) = C_N(i-1) + 1$, and update $i = i + 1$.

Step 4 Repeat Step (3) until all symbols in the sequence have been processed, and then the $C_N(N)$ can be obtained. The resulting value of $C_N(N)$ represents the total number of distinct patterns identified, which corresponds to the Lempel–Ziv complexity of the sequence.

Step 5 The Lempel–Ziv complexity is normalized according to Eqs. (2) and (3).

$$C_{n,N} = \frac{C_N(N)}{C_{UL}} \tag{2}$$

$$C_{UL} = \lim_{N\to\infty} C_N(N) \approx \frac{N}{\log_2 N} \tag{3}$$

2.2. Multivariate dispersion Lempel–Ziv complexity

The original LZC method can be utilized for univariate time series analysis, but it is unsuitable to accurately reflect the complexity of multivariable time series in complex systems. Thus, in this section, we extend the Lempel–Ziv complexity to multivariate form and propose multivariate dispersion Lempel–Ziv complexity (mvDLZC), which extends the Lempel–Ziv complexity to multivariate form by introducing multivariate dispersion pattern into Lempel–Ziv complexity in this paper.

Assuming a multivariate time signal with channel p and length N: $\mathbf{X} = \{x_{k,i}\}_{k=1,2,\dots,p}^{i=1,2,\dots,N}$. In the mvDLZC algorithm, the detailed steps are as follows.

Step 1 The multivariate time signal $\mathbf{X} = \{x_{k,i}\}_{k=1,2,\dots,p}^{i=1,2,\dots,N}$ are mapped to $\mathbf{Z} = \{z_{k,i}^c\}_{k=1,2,\dots,p}^{i=1,2,\dots,N}$ with c classes from 1 to c using NCDF [46]. Firstly, the NCDF process maps $\mathbf{X}$ into $\mathbf{Y} = \{y_{k,i}\}_{k=1,2,\dots,p}^{i=1,2,\dots,N}$ from 0 to 1 as follows:

$$y_{k,i} = \frac{1}{\sigma_k\sqrt{2\pi}} \int_{-\infty}^{x_{k,i}} e^{\frac{-(t-\mu_k)^2}{2\sigma_k^2}} dt \tag{4}$$

where μ_k and σ_k are the mean and standard deviation of the time series x_k, respectively. Then a linear equation is used to assign each $y_{k,i}$ to an integer from 1 to c as follows:

$$z_{k,i}^c = round(c \cdot y_{k,i} + 0.5) \tag{5}$$

where $z_{k,i}^c$ denotes the ith member of the signal in the kth channel and the rounding process involves either increasing or decreasing a number to the next digit. Note that, although this part is linear, the whole mapping process is nonlinear because of the use of NCDF [46].

Step 2 Multivariate embedded vectors are generated according to the multivariate embedding theory [46,47]. Mathematically, the multivariate embedded reconstruction of $\mathbf{Z} = \{z_{k,i}^c\}_{k=1,2,\dots,p}^{i=1,2,\dots,N}$ is defined as:

$$\mathbf{Z}_m(j) = \begin{Bmatrix} z_{1,j}^c & z_{1,j+d_1}^c & \cdots & z_{1,j+(m_1-1)d_1}^c \\ z_{2,j}^c & z_{2,j+d_2}^c & \cdots & z_{2,j+(m_2-1)d_2}^c \\ \vdots & \vdots & \ddots & \vdots \\ z_{p,j}^c & z_{p,j+d_p}^c & \cdots & z_{p,j+(m_p-1)d_p}^c \end{Bmatrix} \tag{6}$$

where $\mathbf{m} = \{m_1, m_2, \dots, m_p\}$ and $\mathbf{d} = \{d_1, d_2, \dots, d_p\}$ denote the embedding dimension and the time lag vectors, respectively. Note that the length of $\mathbf{Z}_m(j)$ is $\sum_{k=1}^{p} m_k$. For simplicity, we assume $d_k = d$ and $m_k = m$, that is, all the embedding dimension values and all the delay values are equal.

Step 3 Each series $\mathbf{Z}_m(j)$ is mapped to a class of dispersion pattern, and each subsequence $\mathbf{Z}_m(j)$ is numbered by the class of these patterns. There are $c^{m\times p}$ classes in total, since the signal has $(m \times p)$ members and each member can be one of the integers from 1 to c. For example, assuming $m = 2, c = 2$, and $p = 2$, there are $c^{m\times p} = 2^{2\times 2} = 16$ potential dispersion patterns, as illustrated in Fig. 1.

Step 4 After conversion in Step (3), we can get a new sequence S_N by labeling $\mathbf{Z}_m(j)$ using multivariate dispersion patterns. S_N has a length of $N - (m-1)d$. All elements within S_N are integers falling within the range of 1 to $c^{m\times p}$.

Step 5 Calculate the complexity counter value $C_N(N)$ of pattern sequence S_N based on the definition of Lempel–Ziv complexity detailed in Section 2.1.

Step 6 Compute the normalized mvDLZC value according to:

$$mvDLZC = \frac{C_N(N)}{C_{UL}} \tag{7}$$

$$C_{UL} = \lim_{N\to\infty} C_N(N) \approx \frac{N-(m-1)d}{\log_{c^{m\times p}} N-(m-1)d} \tag{8}$$

Algorithm 1 Multivariate dispersion Lempel–Ziv complexity (mvDLZC)

Input: Multivariate time series $\mathbf{X} = \{x_{k,i}\}_{k=1,2,\dots p}^{i=1,2,\dots N}$ with channel p and length N, embedding dimension m and the time lag d, number of classes c

Output: The value of mvDLZC

1: Obtain the symbolic series $\mathbf{Z} = \{z_{k,i}^c\}_{k=1,2,\dots p}^{i=1,2,\dots N}$ according to Eq.(4) and Eq.(5).
2: Construct multivariate state space matrix $\mathbf{Z}_m(j)$ by multivariate embedding theory.
3: Define the multivariate dispersion patterns.
4: Get pattern sequence S_N by labelling $\mathbf{Z}_m(j)$ based on multivariate dispersion patterns.
5: Obtain complexity counter value $C_N(N)$ of S_N based on the definition of Lempel-Ziv complexity.
6: Compute the normalized mvDLZC value according to Eq.(7) and Eq.(8).

The pseudocode of multivariate dispersion Lempel–Ziv complexity is illustrated in Algorithm 1. Additionally, the schematic diagram of

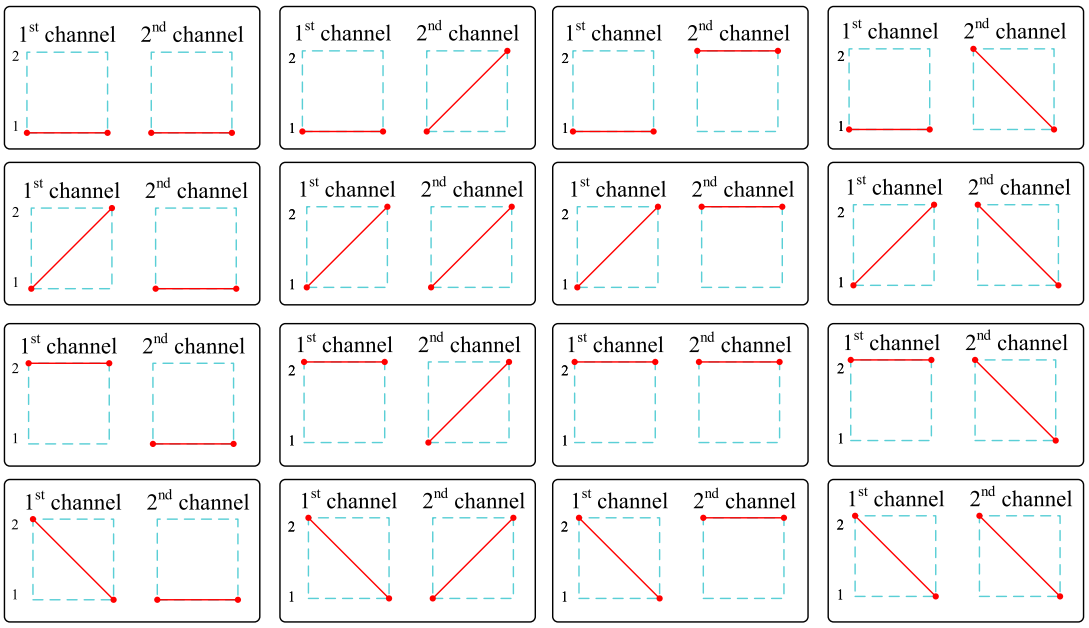

Fig. 1. An example for multivariate dispersion patterns with $m = 2$, $c = 2$, $p = 2$.

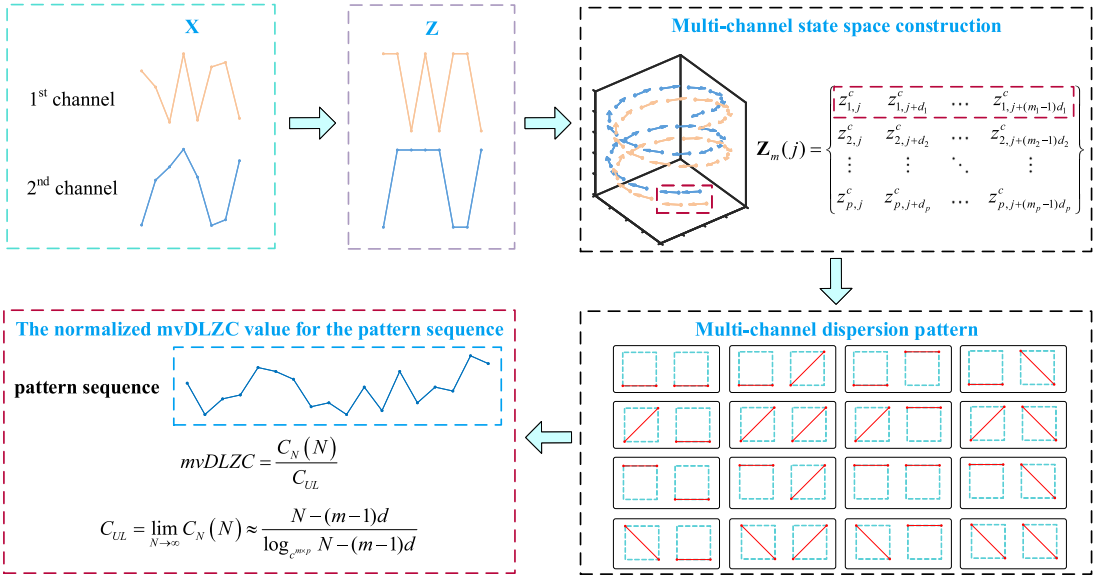

Fig. 2. The schematic diagram of the calculation steps for multivariate dispersion Lempel–Ziv complexity.

the calculation steps for two-channel data is illustrated in Fig. 2. In essence, this method transforms multivariate time series into a one-dimensional symbolic sequence through multivariate dispersion patterns and subsequently calculates the LZC value.

To provide a more detailed understanding of this method, let us consider a practical example using a two-dimensional signal with a length of 10 data points. In this example, we will calculate the mvDLZC value for a 2-channel time series, represented as $\mathbf{X} = \begin{Bmatrix} 8,6,5,0,3,2,2,2,8,0 \\ 4,6,1,6,8,2,3,8,6,4 \end{Bmatrix}$ with $d = 1$, $m = 2$, and $c = 2$. We first map $\mathbf{X}$ to two classes and obtain $\mathbf{Z} = \begin{Bmatrix} 2,2,2,1,1,1,1,1,2,1 \\ 1,2,1,2,2,1,1,2,2,1 \end{Bmatrix}$ according to Step (1) in Section 2.2.

In this example, there are $c^{m\times p} = 2^{2\times 2} = 16$ potential dispersion patterns, as illustrated in Fig. 1, and $10-(2-1) = 9$ multivariate embedding vectors with the length of two and their associated dispersion patterns are as follows:

$$\mathbf{Z}_2(1) = \begin{Bmatrix} 2,2 \\ 1,2 \end{Bmatrix} = \pi_{2212}, \mathbf{Z}_2(2) = \begin{Bmatrix} 2,2 \\ 2,1 \end{Bmatrix} = \pi_{2221},$$

$$\mathbf{Z}_2(3) = \begin{Bmatrix} 2,1 \\ 1,2 \end{Bmatrix} = \pi_{2112}, \mathbf{Z}_2(4) = \begin{Bmatrix} 1,1 \\ 2,2 \end{Bmatrix} = \pi_{1122},$$

$$\mathbf{Z}_2(5) = \begin{Bmatrix} 1,1 \\ 2,1 \end{Bmatrix} = \pi_{1121}, \mathbf{Z}_2(6) = \begin{Bmatrix} 1,1 \\ 1,1 \end{Bmatrix} = \pi_{1111},$$

$$\mathbf{Z}_2(7) = \begin{Bmatrix} 1,1 \\ 1,2 \end{Bmatrix} = \pi_{1112}, \mathbf{Z}_2(8) = \begin{Bmatrix} 1,2 \\ 2,2 \end{Bmatrix} = \pi_{1222},$$

$$\mathbf{Z}_2(9) = \begin{Bmatrix} 2,1 \\ 2,1 \end{Bmatrix} = \pi_{2121}.$$

By labeling $\mathbf{Z}$ using multivariate dispersion patterns, we can get a pattern sequence as follows:

$$\{\pi_{2212}, \pi_{2221}, \pi_{2112}, \pi_{1122}, \pi_{1121}, \pi_{1111}, \pi_{1112}, \pi_{1222}, \pi_{2121}\}$$
$$\downarrow$$
$$\{\text{'10', '12', '14', '3', '4', '1', '2', '7', '16'}\}$$

According to Steps (2)-(4) in Section 2.1, we can get $C_N(N) = 9$, $C_{UL} = \frac{N-(m-1)d}{\log_{c^{m\times p}} N-(m-1)d} = \frac{9}{\log_{16} 9} = 11.3567$. Finally, $mvDLZC = \frac{C_N(N)}{C_{UL}} = \frac{9}{11.3567} = 0.7925$ is calculated.

It is important to note that the order of channels in a multichannel time series does not affect the resulting LZC value. While the assignment of dispersion patterns may change, the overall LZC value remains unchanged. This is because the primary objective of the mvDLZC computation is to quantify the complexity based on the number of dispersion patterns, rather than relying on the precise numerical values or amplitudes within the pattern sequence.

2.3. Multivariate multiscale dispersion Lempel–Ziv complexity

The mvDLZC is a single-scale analysis method, which describes the characteristics at only one scale, limiting its ability to capture comprehensive feature information. To enhance the feature representation capability and capture the dynamical properties of signals at various time scales comprehensively, we employ multiscale analysis through the coarse-grained method [19]. The procedure involves decomposing the original time series into multiple scaled series through a coarse-graining process. Each scaled series is then processed individually using mvDLZC, which we refer to as multivariate multiscale dispersion Lempel–Ziv complexity (mvMDLZC). This approach allows us to examine the signal characteristics at different scales, providing a more comprehensive understanding of the underlying dynamics. The detailed calculation procedure of the proposed mvMDLZC is as follows:

Step 1 Given a multivariate time signal with channel p and length N: $\mathbf{X} = \{x_{k,i}\}_{k=1,2,\ldots,p}^{i=1,2,\ldots,N}$, divide it into coarse-grained series $\mathbf{T}^{(s)} = \{\mathbf{T}_{k,1}^{(s)} \ldots \mathbf{T}_{k,n_s}^{(s)}\}$, $n_s = \left[\frac{N}{s}\right]$, $1 \le s \le \tau$ according to Eq. (9).

$$\mathbf{T}_{k,j}^{(s)} = \frac{1}{s} \sum_{i=(j-1)s+1}^{js} x_{k,i}, 1 \le j \le n_s \tag{9}$$

where $1 \le k \le p$, τ represents the scale factor. To get the coarse-grained time series at the scale factor of s, the original time series is divided into non-overlapping windows of length s. Within each window, the data points from the original multivariate time series are averaged. The original multivariate time series can be considered to have a scale factor of $\tau = 1$, and it can be represented by $\mathbf{T}^{(1)}$.

Step 2 Calculate the mvDLZC value to quantify the stochasticity or irregularity of the coarse-grained multivariate time series $\mathbf{T}^{(s)}$. Mathematically, it can be expressed as:

$$mvMDLZC(\mathbf{X}, \tau, m, d, c) = \{mvDLZC(\mathbf{T}^{(s)}, m, d, c)\} \tag{10}$$

where $1 \le s \le \tau$.

The pseudocode of mvMDLZC can be seen in Algorithm 2. Regarding the scale factor τ, it represents the dimension of LZC features. A smaller τ may not adequately capture the critical fault features, leading to ineffective fault diagnosis. On the other hand, a very large τ can result in dimensionality issues, making it challenging to extract discriminative fault information and potentially leading to suboptimal recognition results. Additionally, a larger τ can increase the computational time required for feature extraction, potentially affecting the efficiency of the process. In this study, we have set τ at 20. By employing multiscale analysis and mvDLZC, we can analyze the dynamics of signal at various resolutions and capture its complexity across multiple time scales and channels.

Algorithm 2 Multivariate multiscale dispersion Lempel–Ziv complexity (mvMDLZC)

Input: Multivariate time series $\mathbf{X} = \{x_{k,i}\}_{k=1,2,\ldots,p}^{i=1,2,\ldots N}$ with channel p and length N, embedding dimension m and the time lag d, number of classes c, and scale factor τ
Output: The value of mvMDLZC
1: **for** $s = 1$ to τ **do**
2: Obtain the coarse-grained multivariate time series $\mathbf{T}^{(s)} = \{\mathbf{T}_{k,1}^{(s)} \ldots \mathbf{T}_{k,n_s}^{(s)}\}$ according to Eq.(9).
3: Compute the normalized mvDLZC value of $\mathbf{T}^{(s)}$.
4: Augment the LZC value $mvMDLZC_{1:s} = \{mvMDLZC_{1:s-1}; mvDLZC(\mathbf{T}^{(s)}, m, d, c)\}$.
5: **end for**

3. Performance verification using synthetic signals

In this section, we conducted the performance verifications on the proposed mvMDLZC method using synthetic signals and compared it to the original mvMLZC method. To assess the effectiveness of the proposed method, we utilized combinations of uncorrelated white Gaussian noise (WGN) and $1/f$ noise, which are known for their differences in complexity and irregularity.

3.1. White Gaussian noise and $1/f$ noise

Previous research has employed combinations of WGN and $1/f$ noise in multivariate time series to evaluate multivariate multiscale entropy algorithms [48,49]. To validate the performance of mvMDLZC, we formulated four distinct combinations of WGN and $1/f$ noise for the three-channel time series. These configurations resulted in four experimental setups for validation:

(1) Three channels of WGN (WGN WGN WGN).
(2) Two channels of WGN and one channel of $1/f$ noise (WGN WGN $1/f$).
(3) One channel of WGN and two channels of $1/f$ noise (WGN $1/f$ $1/f$).
(4) Three channels of $1/f$ noise ($1/f$ $1/f$ $1/f$).

The time-domain signals corresponding to the four combinations of WGN and 1/f noise are illustrated in Fig. 3(a)–(d), respectively. Each experimental setup was independently repeated 100 times, and the mean and standard deviation were calculated for each scale factor (τ) ranging from 1 to 20. Additionally, all experimental setups were replicated for samples with channel lengths of 15,000 and 10,000 to investigate any potential differences resulting from different data lengths.

It is worth noting that existing mvMLZC does not involve parameter selection. The parameter values used for mvMDLZC were chosen based on mvMDE and matched those used in the original mvMDE study to facilitate easy comparison between the two studies [46]. Therefore, for simplicity, we utilized $c = 4$, $d = 1$, and $m = 2$ for all signals employed in this study.

3.2. Synthetic time-series results

In this experimental analysis, we investigated the performance of the proposed mvMDLZC method and compared it to the traditional mvMLZC method using 100 independent realizations of uncorrelated trivariate WGN and $1/f$ noise, as described in Section 3.1. Each combination of the $1/f$ noise and WGN signals had 15,000 sample points. To evaluate the performance of both methods, we computed the average and standard deviation (SD) of the results obtained from the 100 realizations. The outcomes for the proposed mvMDLZC method are

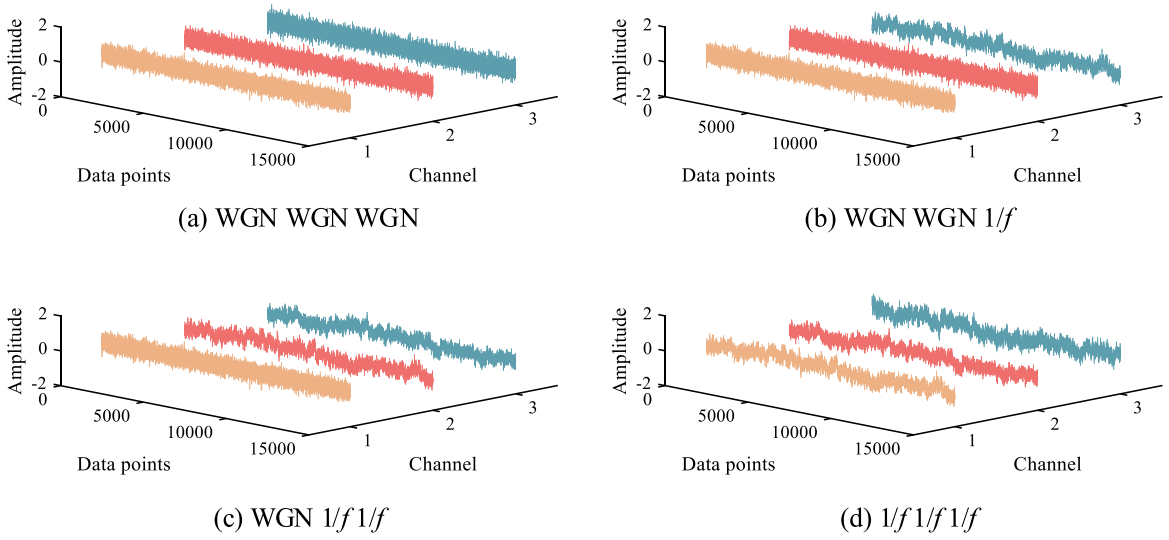

Fig. 3. The time-domain signals for four combinations of WGN and $1/f$ noise with data length of 15 000.

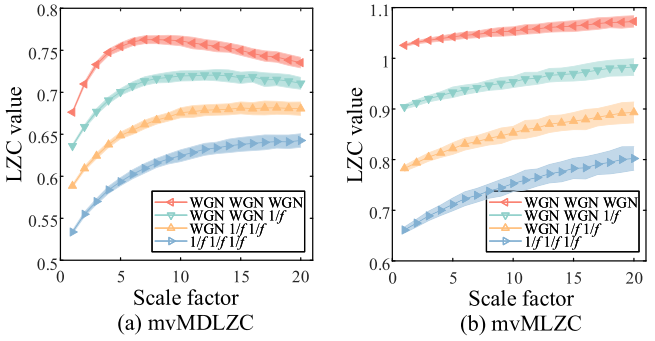

Fig. 4. Lempel–Ziv complexity curves for four groups of noise signals with data length of 15 000.

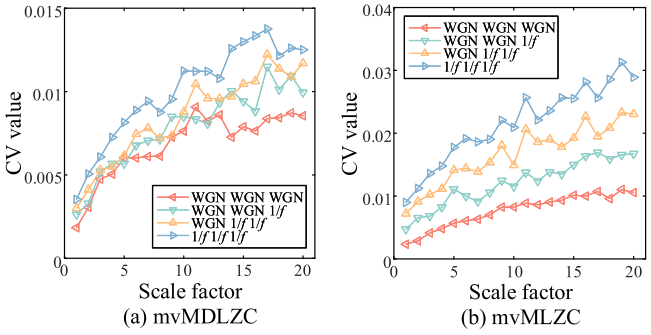

Fig. 5. Coefficient of variation (CV) curves of mvMDLZC and mvMLZC for four combinations of WGN and $1/f$ noise with data length of 15 000.

shown in Fig. 4(a), while those for the traditional mvMLZC method are presented in Fig. 4(b). By analyzing the average and SD values, we can assess the effectiveness and consistency of each method in capturing the characteristics and distinguishing between the different noise types in the trivariate time series data.

On the one aspect, as can be seen from Fig. 4, the SD values for proposed mvMDLZC and traditional mvMLZC show a clear decreasing trend. That is because multiscale analysis would reduce the length of signals. When the length of trivariate signals, obtained by the coarse-graining process, decreases (i.e., the scale factor increases), the SD becomes larger. On the other aspect, it can be observed that mvMLZC exhibit more unstable behavior at large scale factors, as indicated by larger error bars in Fig. 4(b). By contrast, the proposed mvMDLZC is much more stable, as shown in Fig. 4(a).

To further compare the performance of the mvMLZC and proposed mvMDLZC methods, we used the coefficient of variation (CV) as a measure of relative variability. The CV value was calculated as the standard deviation divided by the mean of a time series, allowing us to compare the degree of variation between different data series, even if their means are significantly different. We investigated the results obtained by uncorrelated noise signals for each τ value from 1 to 20, as illustrated in Fig. 5.

From Fig. 5, it can be observed that both mvMDLZC and mvMLZC methods exhibit a clear decreasing trend in CV values. However, the proposed mvMDLZC method outperforms the existing mvMLZC method in terms of stability of results, as evidenced by the smaller CV values for the four possible combinations of WGN and $1/f$ noise achieved by the mvMDLZC method.

To evaluate the influence of signal length on the performance of the proposed mvMDLZC method, we conducted experiments employing trivariate $1/f$ noise and WGN signals with a length of 10 000 sample points. The results for both the mvMDLZC and mvMLZC methods across scales 1 to 20 are depicted in Fig. 6(a) and (b), respectively. Furthermore, we computed and visualized the CV values achieved with a signal length of 10 000 sample points in Fig. 7.

The results obtained with a signal length of 10 000 points align with those obtained with a signal length of 15 000 points. Moreover, an examination of Fig. 7 reveals that the CV values attained by the proposed mvMDLZC method consistently remain below 0.02 across all scales. In contrast, the CV range achieved by the mvMLZC method can extend to approximately 0.032. This observation suggests that the performance of the proposed mvMDLZC method is not significantly impacted by signal length variations, indicating its robust and stable results.

For ease of comparison, Table 1 displays the CV values at a scale factor of 10 for both data lengths of 10 000 and 15 000. As depicted in Table 1, for both data lengths, the CV values obtained using the proposed mvMDLZC method are consistently smaller than those obtained with the existing mvMLZC method across all four types of multichannel signals. These findings emphasize the superior and stable performance of the proposed mvMDLZC method in the analysis of multichannel time series.

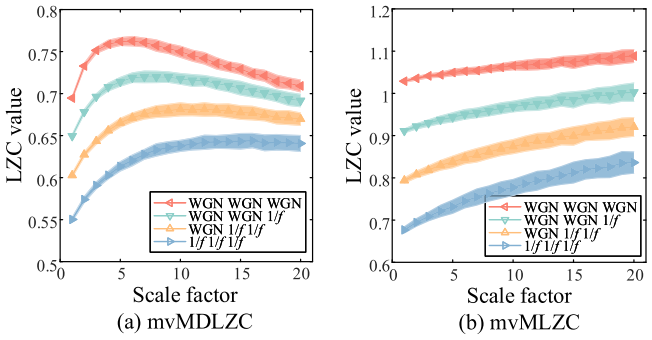

Fig. 6. Lempel–Ziv complexity curves for four groups of noise signals with data length of 10 000.

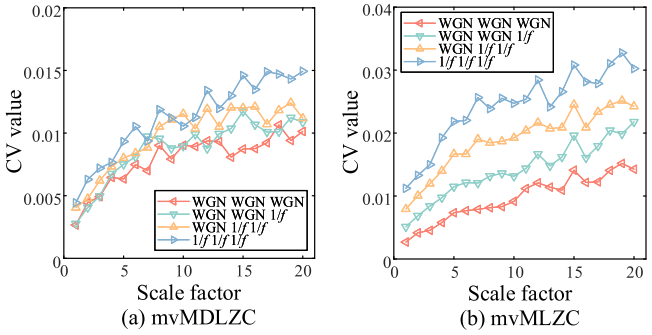

Fig. 7. Coefficient of variation (CV) curves of mvMDLZC and mvMLZC for four combinations of WGN and $1/f$ noise with data length of 10 000.

Table 1
The coefficient of variation (CV) of proposed mvMDLZC and mvMLZC at scale factor 10 for four combinations of WGN and $1/f$ noise.

Data length	Multichannel signals	mvMDLZC	mvMLZC
15 000	WGN WGN WGN	0.0076	0.0082
	WGN WGN $1/f$	0.0084	0.0115
	WGN $1/f$ $1/f$	0.0087	0.0149
	$1/f$ $1/f$ $1/f$	0.0112	0.0209
10 000	WGN WGN WGN	0.0090	0.0091
	WGN WGN $1/f$	0.0089	0.0131
	WGN $1/f$ $1/f$	0.0115	0.0192
	$1/f$ $1/f$ $1/f$	0.0105	0.0247

4. Applications for experimental signals

In this section, two different experimental case studies, involving a rotor system and a planetary gearbox, were conducted to showcase the effectiveness of mvMDLZC in practical fault diagnosis applications.

To assess the performance of mvDLZC and mvMDLZC, we conducted a comparative analysis against several established methods, including univariate LZC, mvLZC, mvDE, and their multiscale variants. Firstly, the comparison was made between univariate LZC and MLZC against the proposed methods to validate the benefits of multichannel data analysis. Secondly, the evaluation involved a comparison between mvLZC and mvMLZC, which are existing multivariate methods based on LZC. Thirdly, we compared the proposed methods with mvDE and mvMDE, as our proposed approaches are based on the concept of multivariate dispersion patterns, and mvMDE represents one of the commonly utilized multivariate entropy algorithms [49,50].

4.1. Case study I: Fault diagnosis of rotor system

4.1.1. Description of rotor system

In this study, a rotor test rig system manufactured by WuXi HouDe Automation Meter was utilized to simulate rubbing faults and rotor

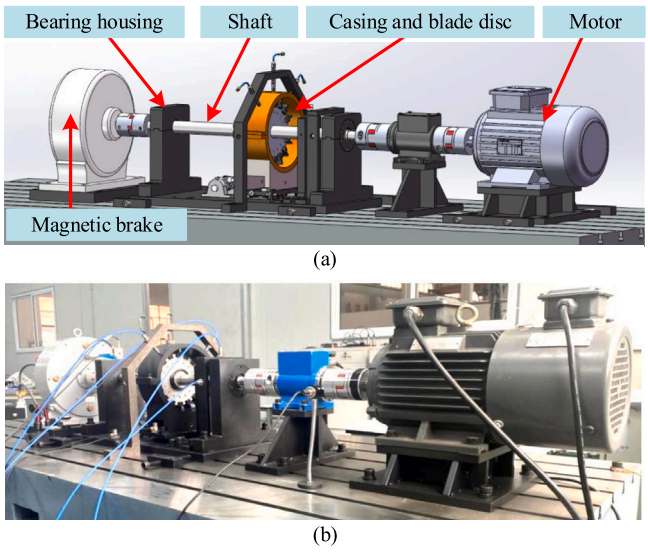

Fig. 8. The sketch of the rotor test rig: (a) three-dimensional model of rotor system, (b) real rotor test rig.

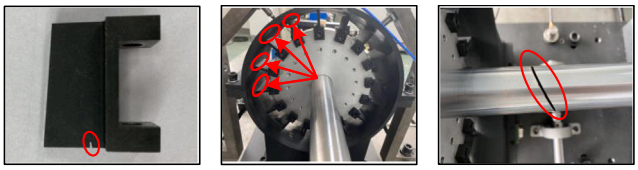

Fig. 9. The fault types of the experimental rotor system: (a) blade crack, (b) full annular rubbing, (c) shaft crack.

faults under different working conditions. The photograph of the experimental rotor rig, consisting of a rotating shaft, blade disk, casings, bearings, and sensors, is depicted in Fig. 8. The experimental procedure encompasses four distinct working conditions: normal condition (NOR) and three fault conditions. The three fault conditions are full annular rubbing (FAR), blade crack (BC), and shaft crack (SC), as illustrated in Fig. 9. It is noted that in the case of full annular rubbing, the rotor remains in constant contact with a fixed point.

To collect the vibration signals, an accelerometer was mounted on the top of the bearing casing. It should be noted that vibration signals were recorded in both the vertical and horizontal directions for this case study. The data acquisition system used had a sampling frequency of 10 kHz, and the rotation speed of the system was kept constant at 1000 RPM. Fig. 10 showcases the normalized two-channel time-domain waveforms corresponding to the four different states. Notably, there are 100 samples with two-channel time series data available for each health condition, resulting in a total of 400 samples for this case study.

4.1.2. Results and analysis

In the first experimental application, five single-scale methods were conducted for comparison to validate the effectiveness of multichannel signal analysis. The comparison results are presented as violin plots in Fig. 11(a)–(e) for proposed mvDLZC, univariate LZC (LZC_V and LZC_H), mvLZC and mvDE, respectively. The violin plots offer a visual representation of the full feature distribution for the five methods across four different states.

From Fig. 11(a), it is evident that the feature distribution of the healthy state and fault states in mvDLZC exhibit significant differences, enabling a clear differentiation between normal and faulty conditions. Additionally, the significant differences in median values for the three

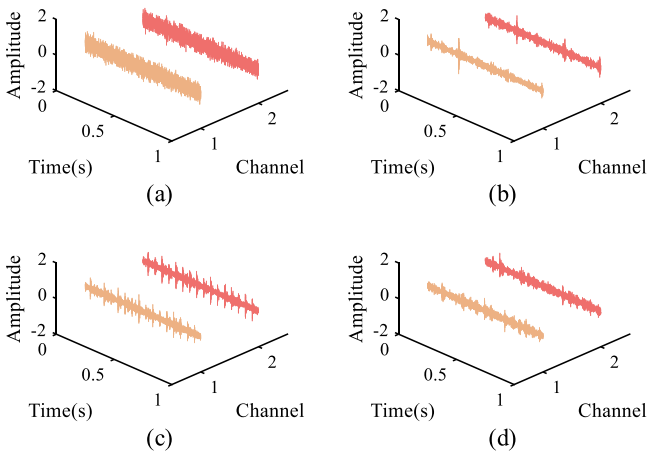

Fig. 10. The time-domain signals of four conditions for the rotor system: (a) normal condition, (b) blade crack, (c) full annular rubbing, (d) shaft crack.

Table 2
Differences between LZC values or entropy values using Mann–Whitney U-test.

p-value	Method				
	mvDLZC	mvLZC	mvDE	LZC_V	LZC_H
NOR vs. BC	2.5E−34	2.5E−34	2.6E−34	2.4E−34	2.5E−34
NOR vs. FAR	2.5E−34	2.5E−34	2.3E−22	2.4E−34	2.4E−34
NOR vs. SC	2.5E−34	2.5E−34	7.5E−29	2.4E−34	2.5E−34
BC vs. FAR	6.8E−23	3.1E−34	2.6E−34	2.4E−34	2.8E−26
BC vs. SC	1.2E−22	2.9E−29	0.8709	2.6E−34	0.0147
FAR vs. SC	9.9E−33	4.5E−16	8.0E−34	0.6586	1.9E−16

fault states signify that mvDLZC can not only detect faults but also effectively discriminate between different fault types. However, as shown in Fig. 11(d), LZC_V exhibits similar median entropy values for SC and FAR states, making it challenging to differentiate between these two states. Similarly, Fig. 11(c) and (e) demonstrates a similar phenomenon for mvDE and LZC_H, where the median values for BC state closely resemble those of the SC state, posing difficulties in distinguishing between them.

Moreover, in order to quantitatively assess the differences between feature values across different states, for each method, all feature values were statistically analyzed by Mann–Whitney U test to obtain p-values. The obtained p-values are used to assess the significance of the differences, with a significance level of $p \ll$ 1E-3 indicating statistically significant differences and $p \ll$ 1E-4 denoting more significant differences. The statistical results are summarized in Table 2.

From the results presented in Table 2, it is evident that both mvDLZC and mvLZC methods demonstrate highly significant differences ($p \ll$ 1E-4) in distinguishing between any two states, indicating the highest level of differentiation among all the methods. This highlights the effectiveness of multichannel data analysis in fault detection. On the other hand, LZC_V, LZC_H, and mvDE methods show poor performance in the analysis. Specifically, LZC_V fails to distinguish significant differences between the FAR and SC states, while LZC_H and mvDE are both unable to distinguish between the BC and SC states. Based on the experimental results and the statistical analysis, it can be concluded that the proposed mvDLZC method exhibits a strong capability to detect dynamic changes in mechanical signals.

Moreover, machine learning was applied to assess the feature discrimination capacities of different LZC-based methods and mvDE. Specifically, the support vector machine (SVM) algorithm [51] was implemented for classification tasks. The dataset was partitioned into a training set, which consisted of 75% of randomly selected samples (300 samples), and a test set, which included the remaining 25% of samples (100 samples).

Fig. 12 illustrates the results of the machine learning analysis. These results are consistent with the trends observed in the violin plots. Specifically, the performance of the mvDLZC and mvLZC methods surpasses that of the single-source LZC_V and LZC_H methods, highlighting the advantage of using multichannel data for capturing fault-related information and achieving improved classification rates. In contrast, the performance of the mvDE method is mediocre. On the other hand, the multiscale version, the mvMDLZC method with scale factor of 20, exhibits superior performance compared to mvDLZC, indicating the effectiveness of the multiscale analysis in fault diagnosis of rotating machinery. These results highlight the potential of the mvMDLZC method as a useful approach for accurate and reliable fault diagnosis of the rotor system.

4.2. Case study II: Fault diagnosis of planetary gearbox

4.2.1. Description of planetary gearbox

The second experimental case study focused on a planetary gearbox system, as depicted in Fig. 13. The system mainly consists of a motor, a planetary gearbox, a tachometer, and a magnetic damping component. The vibration signals used in this study were collected by employing an accelerometer that was positioned on the planetary gearbox casing. The data acquisition was performed with a sampling frequency of 16 kHz. Throughout the experiment, the load was set at 5 N m, and the motor's rotation speed was maintained at a constant value of 1000 RPM.

During the experimental process, a total of six conditions were investigated, including the normal condition (NOR) and five fault conditions. The fault conditions included planet gear fault (PGF), sun gear fault (SGF), bearing fault (BF), ring gear fault (RGF), and planetary carrier fault (PCF), as illustrated in Fig. 14. These conditions were simulated to evaluate the performance of the proposed methods in diagnosing faults in the planetary gearbox system.

The vibration signals in the radial, tangential, and axial directions were collected, respectively. The collected signals were then normalized, and the corresponding three-channel time-domain waveforms under different states are presented in Fig. 15. In line with the previous case study, there are 100 samples available for each health condition, amounting to a total of 600 samples.

4.2.2. Results and analysis

Similar to Case Study I, in this case study, we firstly compared the performance of mvDLZC, mvLZC, mvDE, and univariate LZC methods (LZC_R, LZC_T, LZC_A) for signal analysis. The feature distribution of these six single-scale methods across six different states is depicted as violin plots in Fig. 16. Consistently, we subjected the feature values to the Mann–Whitney U-test to obtain p-values for each method, as illustrated in Table 3.

These plots offer a comprehensive view of the distribution of LZC or entropy features. From Fig. 16(a), it is evident that the proposed mvDLZC method effectively distinguishes different working states of planetary gearbox, exhibiting significant differences among the states. On the other hand, as can be seen from Fig. 16(c), the mvDE method fails to differentiate between the RGF and PCF conditions. Similarly, from Fig. 16(d)–(f), it can be observed that the three univariate LZC methods also face challenges in distinguishing certain conditions. Specifically, LZC_R and LZC_A show limited differentiation between the SGF and BF conditions, and LZC_T struggles to distinguish the SGF and PGF conditions. The associated p-values, detailed in Table 3, confirm that proposed mvMDLZC exhibits significant distinctions between any two states, with p-values significantly below 1E-4. This underscores a high degree of differentiation between any pair of conditions.

Similarly, these features were employed for classification and fault diagnosis tasks using the SVM algorithm. To assess the performance of each method, 75% of samples from each of the six states were randomly selected as the training set, resulting in a total of 6×75 samples. The remaining 25% of samples were designated as the test set, amounting

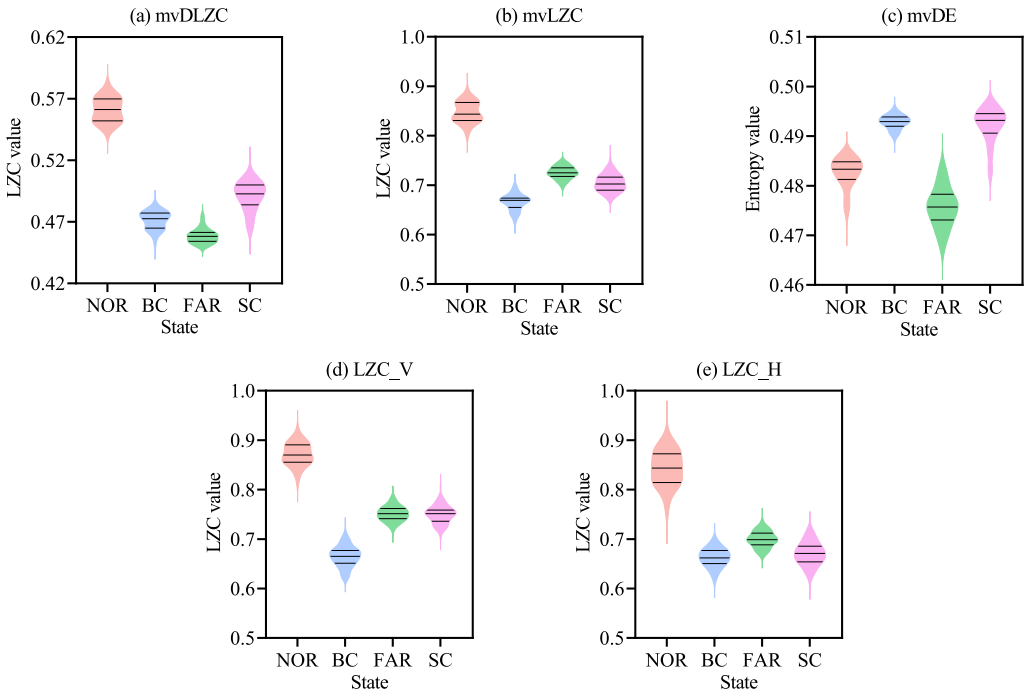

Fig. 11. Violin plots for features obtained by (a) mvDLZC, (b) mvLZC, (c) mvDE, (d) LZC_V, and (e) LZC_H for rotor system across four states.

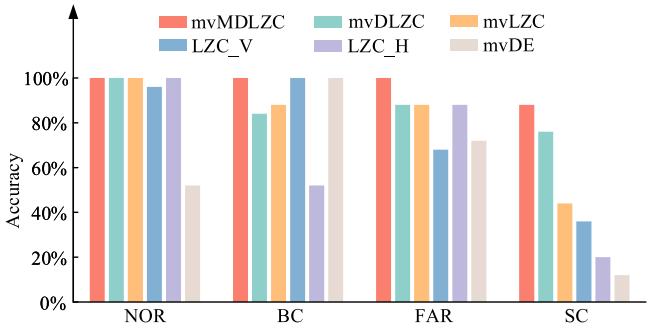

Fig. 12. Diagnostic accuracies of different methods for the rotor system.

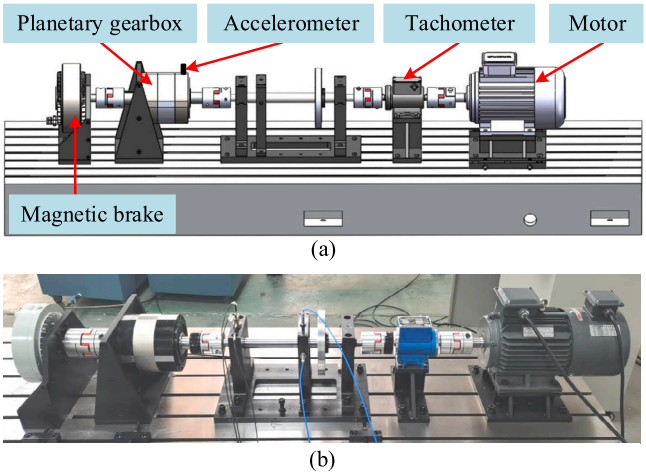

Fig. 13. The sketch of the planetary gearbox system: (a) three-dimensional model of planetary gearbox system, (b) real gearbox test rig.

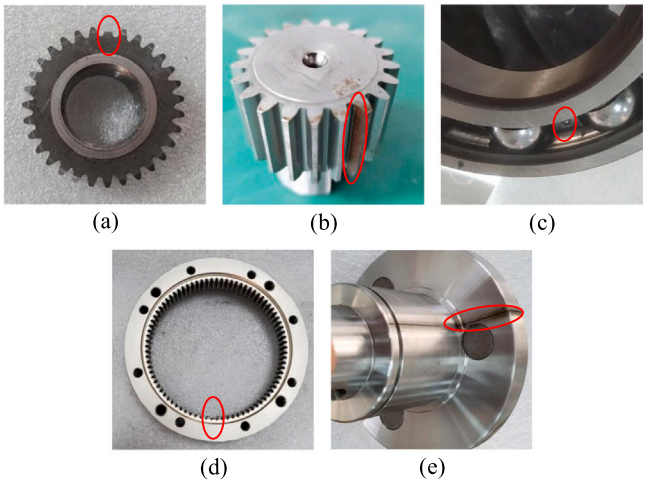

Fig. 14. The fault types of the experimental planetary gearbox: (a) planet gear fault, (b) sun gear fault, (c) bearing fault, (d) ring gear fault and (e) planetary carrier fault.

to a total of 6 × 25 samples. The comparison results are depicted in Fig. 17.

Among the eight methods compared, the proposed mvMDLZC consistently achieves the highest diagnostic performance across 20 repeated runs, with a mean identification accuracy of 98.43%. Additionally, mvMDLZC method exhibits the smallest error bars, indicating its high stability. These small error bars suggest that the results obtained from mvMDLZC are consistent and less affected by variations in experimental conditions or data samples. This stability is crucial in practical applications, ensuring the reliable and consistent performance of the mvMDLZC method in different scenarios and datasets.

The mvMLZC method, another multivariate approach, demonstrates the second-highest recognition accuracy of 95.93%, significantly surpassing the other single-measurement-based methods. In contrast, the mvDE method shows mediocre performance in comparison, while the univariate LZC methods exhibit lower recognition rates below 70%.

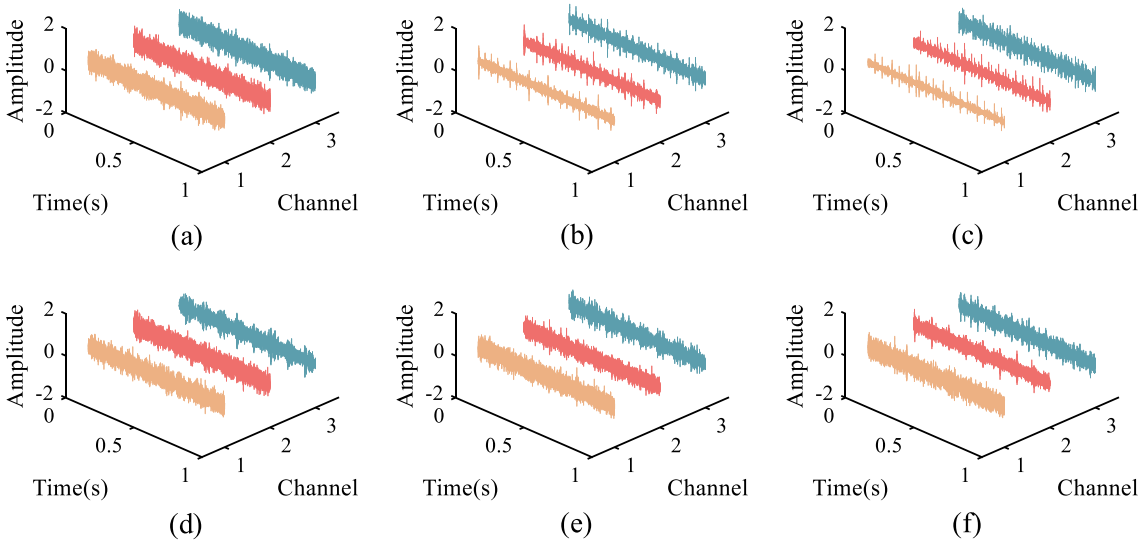

Fig. 15. The time-domain signals of six conditions for the rotor system: (a) normal condition (NOR), (b) planet gear fault (PGF), (c) sun gear fault (SGF), (d) bearing fault (BF), (e) ring gear fault (RGF) and (f) planetary carrier fault (PCF).

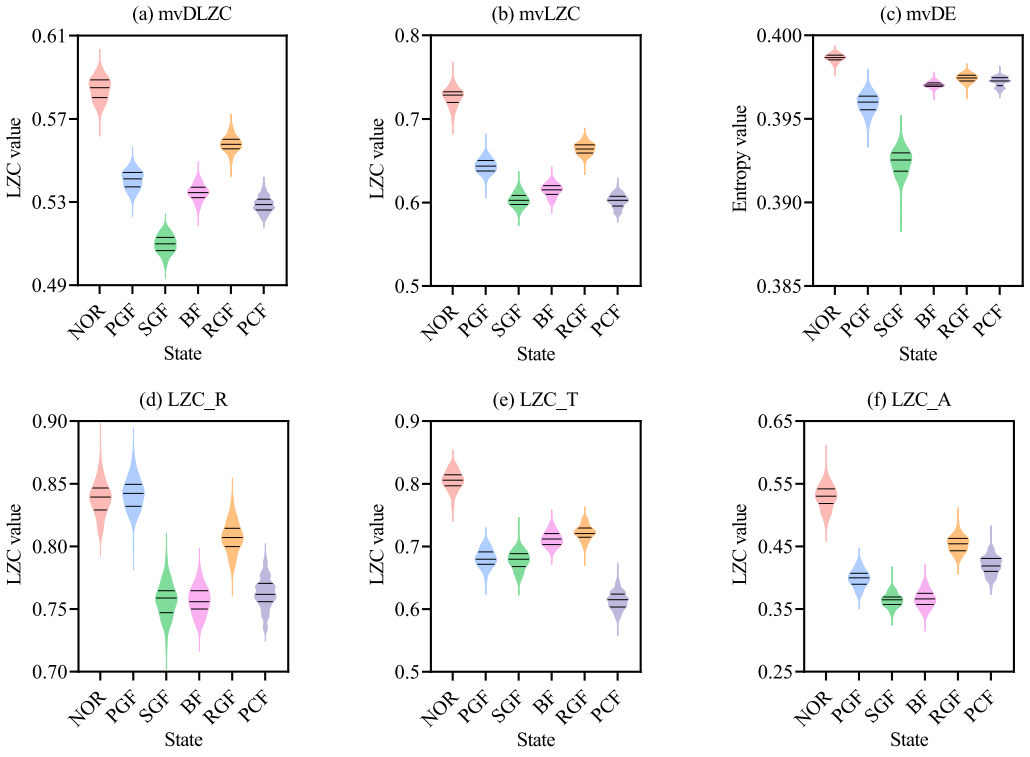

Fig. 16. Violin plots for features obtained by (a) mvDLZC, (b) mvLZC, (c) mvDE, (d) LZC_R, (e) LZC_T, and (f) LZC_A for the planetary gearbox system across six states.

These lower recognition rates of the univariate LZC methods indicate their limitations in accurately identifying and classifying faults in the given dataset. Additionally, the multiscale-based methods outperform the single-scale methods, highlighting the effectiveness and necessity of the coarse-graining process. These findings emphasize the importance of utilizing multivariate approaches, such as mvMDLZC and mvMLZC, which outperform both mvDE and univariate LZC methods in achieving higher accuracy and reliability in fault diagnosis tasks.

In order to provide a comprehensive illustration and comparison of the diagnostic performance among various multiscale-based approaches, we also conducted the confusion matrices and analyzed their quantitative diagnostic metrics, which include precision, recall, and F1-score [52]. The results are presented in both Fig. 18 and Table 4.

Fig. 18 visually presents the confusion matrix for planetary gearbox health diagnostics using different approaches. The confusion matrix offers a detailed overview of diagnostic outcomes for six health states, including specific classification numbers and an assessment of overall accuracy. Notably, our proposed mvMDLZC approach exhibits exceptional performance, achieving an impressive overall diagnostic accuracy of 99.3%. It is worth mentioning that the MLZC_T method misclassifies only a minimal number of test samples, resulting in an accuracy rate of 96%. In contrast, the existing multivariate method, mvMLZC, does not demonstrate a significant enhancement in multichannel diagnostics compared to MLZC_R and MLZC_A, and it even falls below the accuracy of MLZC_T. Additionally, mvMDE performs mediocrely, with an overall recognition rate of 86%.

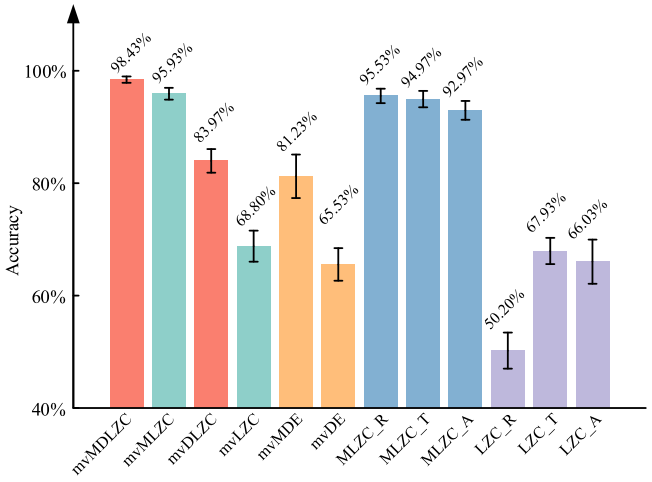

Fig. 17. Diagnostic accuracies of different methods for the planetary gearbox system.

Table 3
Differences between LZC values or entropy values using Mann–Whitney U-test for planetary gearbox system.

p-value	Method					
	mvDLZC	mvLZC	mvDE	LZC_R	LZC_T	LZC_A
NOR vs. PGF	2.5E−34	2.5E−34	2.6E−34	0.075	2.4E−34	2.4E−34
NOR vs. SGF	2.5E−34	2.5E−34	2.6E−34	2.2E−34	2.4E−34	2.3E−34
NOR vs. BF	2.5E−34	2.5E−34	2.6E−34	2.2E−34	2.3E−34	2.4E−34
NOR vs. RGF	2.5E−34	2.5E−34	2.6E−34	8.9E−29	2.3E−34	2.8E−34
NOR vs. PCF	2.5E−34	2.5E−34	2.6E−34	2.3E−34	2.4E−34	2.4E−34
PGF vs. SGF	2.5E−34	2.8E−34	2.6E−34	2.2E−34	0.1931	4.7E−31
PGF vs. BF	5.4E−17	1.2E−33	1.5E−31	2.2E−34	4.8E−29	9.9E−29
PGF vs. RGF	3.2E−34	2.7E−30	6.7E−34	8.4E−31	1.1E−32	2.7E−34
PGF vs. PCF	5.9E−31	2.6E−34	3.1E−32	2.2E−34	3.6E−34	1.7E−16
SGF vs. BF	2.5E−34	4.0E−18	2.6E−34	0.8340	1.5E−30	0.4321
SGF vs. RGF	2.5E−34	2.5E−34	2.6E−34	4.6E−34	1.1E−32	2.2E−34
SGF vs. PCF	2.5E−34	0.6405	2.6E−34	0.0039	5.3E−34	4.7E−34
BF vs. RGF	2.5E−34	2.4E−34	1.3E−20	2.6E−34	4.1E−08	2.3E−34
BF vs. PCF	1.1E−17	1.1E−19	6.5E−06	0.0010	2.3E−34	8.6E−34
RGF vs. PCF	2.5E−34	2.5E−34	8.5E−07	9.9E−34	2.3E−34	2.5E−28

Table 4
Classification results: precision, recall, and f1-score for the planetary gearbox system across six states.

Metric	Method	Classes						Average
		NOR	PGF	SGF	BF	RGF	PCF	
Precision	mvMDLZC	1	1	1	0.96	1	1	0.99
	mvMLZC	1	1	1	0.89	1	0.78	0.95
	mvMDE	1	0.88	1	0.80	0.79	0.75	0.87
	MLZC_R	1	0.91	0.88	1	0.83	0.96	0.93
	MLZC_T	1	0.92	0.96	0.89	1	1	0.96
	MLZC_A	1	1	0.81	0.87	0.96	0.93	0.93
Recall	mvMDLZC	1	1	0.96	1	1	1	0.99
	mvMLZC	1	0.96	0.64	1	1	1	0.93
	mvMDE	1	0.56	0.92	0.96	0.88	0.84	0.86
	MLZC_R	1	0.84	0.92	0.84	0.96	1	0.93
	MLZC_T	1	0.96	0.92	1	0.88	1	0.96
	MLZC_A	1	0.92	0.84	0.80	1	1	0.93
F1-score	mvMDLZC	1	1	0.98	0.98	1	1	0.99
	mvMLZC	1	0.98	0.78	0.94	1	0.88	0.93
	mvMDE	1	0.68	0.96	0.87	0.83	0.79	0.86
	MLZC_R	1	0.88	0.90	0.91	0.89	0.98	0.93
	MLZC_T	1	0.94	0.94	0.94	0.94	1	0.96
	MLZC_A	1	0.96	0.82	0.83	0.98	0.96	0.93

Table 4 provides diagnostic metric results for planetary gearbox health diagnostics using different approaches. Clearly, mvMDLZC stands out with the highest precision, recall, and F1-score, each achieving a score of 0.99. The overall average F1-scores for mvMDLZC, mvMLZC, mvMDE, MLZC_R, MLZC_T, and MLZC_A are 0.99, 0.93, 0.86, 0.93, 0.96, and 0.93, respectively. These metrics align with the recognition performance observed in Fig. 18. The existing multi-source method, mvMLZC, may struggle to extract information effectively.

Moreover, the t-distributed stochastic neighbor embedding algorithm (t-SNE) was utilized to visualize the extracted features of six multiscale-based methods in a two-dimensional space, as depicted in Fig. 19. Fig. 19(a) demonstrates the effectiveness of the mvMDLZC method, where each class exhibits a distinct class center and no overlapping between classes. This clear separation makes it easier for the classifier to classify the different fault conditions accurately. On the other hand, Fig. 19(b) shows the visualization of features extracted by mvMLZC, indicating different class centers for each fault condition. However, there is a drawback of overlapping between two classes, which may pose challenges for classification. In contrast, Fig. 19(d), (e), and (f) depict the visualization results of the univariate LZC methods, revealing uncertain class centers and difficulty for SVM to classify different faults accurately. These findings highlight the effectiveness of multichannel data analysis in achieving better fault classification results.

Overall, these findings emphasize the effectiveness of the proposed mvMDLZC in integrating information from multiple channels, leading to enhanced recognition performance in the context of health diagnostics with multichannel data.

Table 5
Accuracy and standard deviation (%) of the diagnostic results with different training percentages.

Methods	Training percentage			
	20%	30%	40%	50%
LZC_R	47.56 ± 2.23	48.36 ± 2.10	48.53 ± 2.72	49.62 ± 3.00
LZC_T	66.99 ± 1.28	67.90 ± 1.66	67.72 ± 1.49	67.88 ± 1.68
LZC_A	66.04 ± 1.50	65.12 ± 1.42	65.10 ± 2.46	65.62 ± 2.10
mvDE	68.13 ± 2.35	66.23 ± 3.67	66.00 ± 2.49	65.28 ± 2.45
mvLZC	70.35 ± 1.33	70.31 ± 1.53	70.03 ± 1.46	69.65 ± 1.76
mvDLZC	83.15 ± 1.02	83.30 ± 1.19	83.60 ± 1.24	83.87 ± 1.21
MLZC_R	91.58 ± 1.08	92.80 ± 1.05	93.68 ± 1.26	94.23 ± 1.33
MLZC_T	93.74 ± 1.32	94.27 ± 0.97	94.74 ± 0.79	94.67 ± 1.00
MLZC_A	92.00 ± 0.68	92.27 ± 0.70	92.18 ± 1.06	92.73 ± 1.34
mvMDE	84.27 ± 3.31	83.56 ± 4.48	82.56 ± 3.79	81.10 ± 2.86
mvMLZC	93.16 ± 1.94	94.30 ± 1.05	95.00 ± 0.59	96.07 ± 0.79
mvMDLZC	98.16 ± 0.73	98.27 ± 0.62	98.36 ± 0.57	98.42 ± 0.79

4.3. Diagnosis performance under challenges

4.3.1. Fault diagnosis with small sample

To evaluate the performance of the proposed method under the constraint of a small sample size, a comparative analysis was conducted using different proportions of samples for training. The case study II data was utilized for this purpose. The selected proportions included 20%, 30%, 40%, and 50% of the samples for training, while the remaining samples were used for testing. This process was repeated 20 times to account for randomness, and the average recognition rates were calculated. It is noted that SVM was used for classification, following the same procedure as in previous case studies. The classification results for different methods and varying proportions of the training set are presented in Table 5. Furthermore, to provide a visual representation of the diagnostic accuracies of the multiscale-based methods at different training percentages, a graph (Fig. 20) was plotted, which showcases the performance trends of different methods.

Table 5 and Fig. 20 clearly demonstrate that the proposed mvMDLZC method outperforms the other methods in fault diagnosis, consistently achieving recognition rates above 98%. Even with only 20% of training samples, the proposed mvMDLZC method achieves

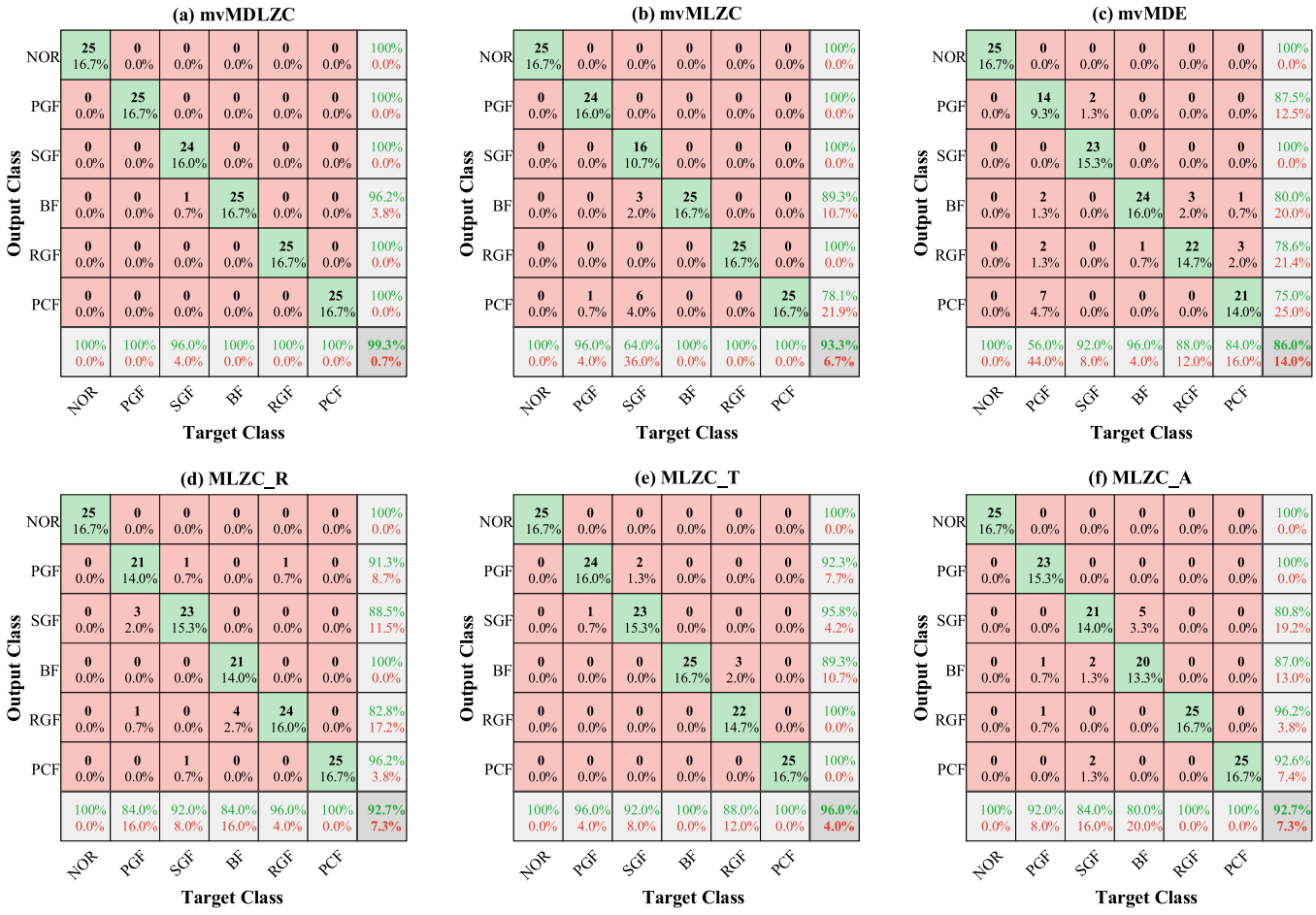

Fig. 18. Confusion matrix for (a) mvMDLZC, (b) mvMLZC, (c) mvMDE, (d) MLZC_R, (e) MLZC_T, and (f) MLZC_A.

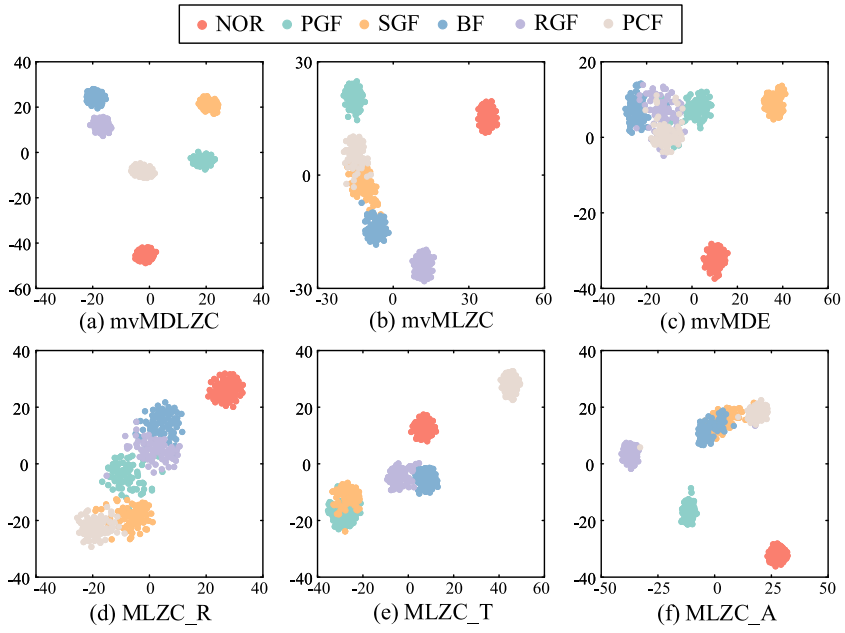

Fig. 19. Feature visualization via t-SNE for six multiscale-based methods: (a) mvMDLZC, (b) mvMLZC, (c) mvMDE, (d) MLZC_R, (e) MLZC_T, and (f) MLZC_A.

a remarkable recognition rate of 98% or higher. Furthermore, the low standard deviation of the diagnostic results, remaining below 1%, emphasizes the exceptional stability of the mvMDLZC approach, even under challenging conditions with limited training data.

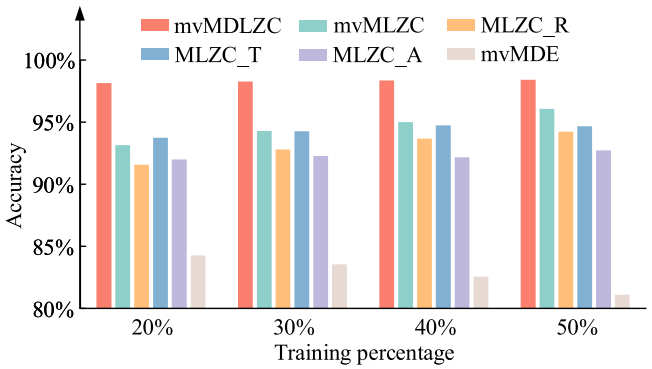

Fig. 20. Diagnostic accuracies for multiscale methods with different training percentage.

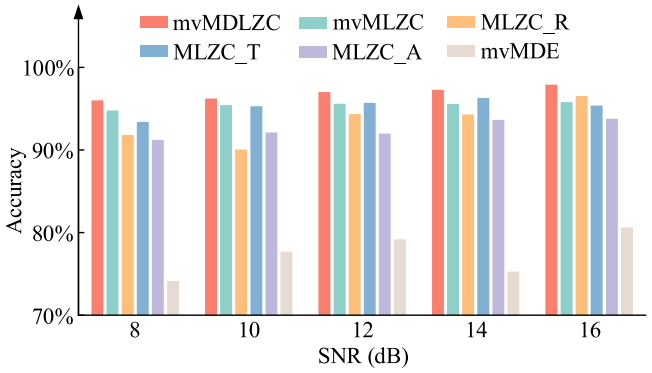

Fig. 21. Diagnostic accuracies for multiscale methods under different degrees of noise in signals.

From Fig. 20, overall, the multivariate methods, such as mvMDLZC and the existing mvMLZC, clearly outperform the single-source methods (MLZC_R, MLZC_T, and MLZC_A), in terms of diagnostic accuracy. This further highlights the advantage of utilizing multiple data sources for fault diagnosis. The superior performance of the multivariate methods reinforces the effectiveness of considering multiple channels or sources of information in fault diagnosis tasks. By incorporating information from different sources, such as the radial, tangential, and axial directions, the multivariate methods demonstrate enhanced discriminatory power and improved fault classification accuracy. These results emphasize the importance of leveraging the synergistic information present in multichannel data for more accurate and reliable fault diagnosis.

4.3.2. Robustness against noises

In practical industrial applications, the influence of noise on diagnosis performance is substantial. To evaluate the robustness of the proposed mvMDLZC method against noise, we introduced varying degrees of white Gaussian noise into the signals. The diagnostic outcomes under different noise levels are depicted in Fig. 21.

Fig. 21 clearly demonstrates the robustness of the proposed mvMDLZC method against noise in real industrial applications.

The decrease in SNR naturally leads to reduced diagnostic accuracy. However, the mvMDLZC consistently outperforms other methods, including mvMLZC, MLZC_R, MLZC_T, MLZC_A, and mvMDE, in terms of diagnosis accuracy across different levels of white Gaussian noise. Remarkably, even as SNR diminishes, the average identification accuracies of mvMDLZC remain high, reaching 96% at an SNR of 8 dB. This resilience can be attributed to the fusion of multichannel data, which allows mvMDLZC to leverage comprehensive information for precise fault diagnosis, even under challenging low SNR conditions.

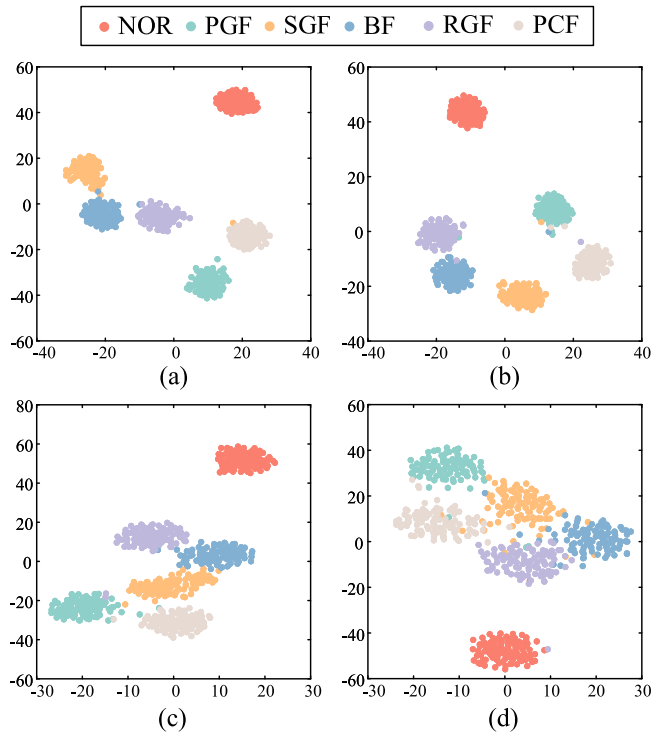

Fig. 22. Visualization of mvMDLZC features under different levels of noise: (a)–(d) represent vibration signals under 14 dB, 12 dB, 10 dB, 8 dB noise, respectively.

To gain deeper insights into the robustness of the mvMDLZC method against noise, Fig. 22 offers visualizations of the features at different SNR levels (14 dB, 12 dB, 10 dB, and 8 dB). Each color represents a specific health state of the gearbox.

From Fig. 22(a), at 14 dB SNR, the feature points corresponding to different states exhibit minimal overlap, signifying a high degree of discrimination between the classes. Nonetheless, the SGF and BF states display some similarity in their features, suggesting an inherent resemblance in their vibration characteristics. Even at 8 dB, each class maintains a distinct class center without significant overlap between feature points. This demonstrates that the extracted features from the mvMDLZC method remain discriminative and informative, enabling effective differentiation between different health states even in the presence of low SNR.

In summary, these results emphasize the robustness and effectiveness of the proposed mvMDLZC method in addressing noise in real industrial applications. It consistently delivers superior diagnostic performance, even under challenging noisy conditions. Moreover, the feature visualizations further validate the discriminative power of the extracted features, highlighting the method's capacity to effectively distinguish different health states.

5. Conclusion

This paper introduces the concept of multivariate multiscale dispersion Lempel–Ziv complexity (mvMDLZC) as an extension of the univariate LZC method for multichannel systems to extract the fault features hidden in multi-source information. Through comprehensive comparative studies employing synthetic and real-world multichannel datasets, the proposed mvMDLZC approach showcases its superiority over existing methods. It exhibits excellent performance in fault diagnosis of mechanical systems, even in challenging scenarios with noise and limited sample sizes. The results validate the effectiveness and robustness of mvMDLZC in recognizing various fault types. The proposed method contributes to the advancement of LZC-based methods and

opens up new possibilities for analyzing the complexity of multichannel systems.

Overall, this work contributes to the advancement of LZC-based methods by extending their applicability to multichannel data analysis and addressing the limitations of existing approaches. Future research will explore its application in other domains and further investigate its capabilities in signal analysis, potentially incorporating advanced multiscale methods and neural networks, aiming to contribute to advancements in multi-source information fusion and fault detection.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgment

The research is supported by the National Natural Science Foundation of China under Grant 12172290 and 52250410345. This research is also funded by the Opole University of Technology as part of the GRAS project under grant no. 270/23.

References

[1] H. Shao, J. Lin, L. Zhang, D. Galar, U. Kumar, A novel approach of multisensory fusion to collaborative fault diagnosis in maintenance, Inf. Fusion 74 (2021) 65–76.
[2] J. Huang, L. Cui, Tensor singular spectrum decomposition: Multisensor denoising algorithm and application, IEEE Trans. Instrum. Meas. 72 (2023) 1–15.
[3] D. Sun, Y. Li, S. Jia, K. Feng, Z. Liu, Non-contact diagnosis for gearbox based on the fusion of multi-sensor heterogeneous data, Inf. Fusion 94 (2023) 112–125.
[4] Y. Xu, K. Feng, X. Yan, R. Yan, Q. Ni, B. Sun, Z. Lei, Y. Zhang, Z. Liu, CFCNN: A novel convolutional fusion framework for collaborative fault identification of rotating machinery, Inf. Fusion 95 (2023) 1–16.
[5] J. Huang, L. Cui, J. Zhang, Novel morphological scale difference filter with application in localization diagnosis of outer raceway defect in rolling bearings, Mech. Mach. Theory 184 (2023) 105288.
[6] Z. Huo, M. Martínez-García, Y. Zhang, R. Yan, L. Shu, Entropy measures in machine fault diagnosis: Insights and applications, IEEE Trans. Instrum. Meas. 69 (6) (2020) 2607–2620.
[7] R. Yan, R.X. Gao, Complexity as a measure for machine health evaluation, IEEE Trans. Instrum. Meas. 53 (4) (2004) 1327–1334.
[8] Y. Li, S. Jiao, B. Geng, Refined composite multiscale fluctuation-based dispersion Lempel–Ziv complexity for signal analysis, ISA Trans. 133 (2023) 273–284.
[9] Y. Li, B. Geng, S. Jiao, Dispersion entropy-based Lempel-Ziv complexity: a new metric for signal analysis, Chaos Solitons Fractals 161 (2022) 112400.
[10] Y. Li, J. Wu, Y. Yi, Y. Gu, Unequal-step multiscale integrated mapping dispersion Lempel-Ziv complexity: A novel complexity metric for signal analysis, Chaos Solitons Fractals 175 (2023) 113945.
[11] C. Li, K. Noman, Z. Liu, K. Feng, Y. Li, Optimal symbolic entropy: An adaptive feature extraction algorithm for condition monitoring of bearings, Inf. Fusion (2023) 101831.
[12] K. Noman, Y. Li, S. Si, S. Wang, G. Mao, Oscillatory Lempel–Ziv complexity calculation as a nonlinear measure for continuous monitoring of bearing health, IEEE Trans. Reliab. (2022).
[13] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.
[14] X. Wang, S. Si, Y. Li, Multiscale diversity entropy: A novel dynamical measure for fault diagnosis of rotating machinery, IEEE Trans. Ind. Inform. 17 (8) (2020) 5419–5429.
[15] A. Omidvarnia, M. Mesbah, M. Pedersen, G. Jackson, Range entropy: A bridge between signal complexity and self-similarity, Entropy 20 (12) (2018) 962.
[16] T. Zhang, W. Chen, M. Li, Fuzzy distribution entropy and its application in automated seizure detection technique, Biomed. Signal Process. Control 39 (2018) 360–377.
[17] M. Rostaghi, M.M. Khatibi, M.R. Ashory, H. Azami, Fuzzy dispersion entropy: A nonlinear measure for signal analysis, IEEE Trans. Fuzzy Syst. 30 (9) (2021) 3785–3796.
[18] B. Zhang, P. Shang, Transition permutation entropy and transition dissimilarity measure: Efficient tools for fault detection of railway vehicle systems, IEEE Trans. Ind. Inform. 18 (3) (2021) 1654–1662.
[19] M. Costa, A.L. Goldberger, C.-K. Peng, Multiscale entropy analysis of complex physiologic time series, Phys. Rev. Lett. 89 (6) (2002) 068102.
[20] J. Zheng, H. Pan, S. Yang, J. Cheng, Generalized composite multiscale permutation entropy and Laplacian score based rolling bearing fault diagnosis, Mech. Syst. Signal Process. 99 (2018) 229–243.
[21] R. Zhou, X. Wang, J. Wan, N. Xiong, EDM-fuzzy: an euclidean distance based multiscale fuzzy entropy technology for diagnosing faults of industrial systems, IEEE Trans. Ind. Inform. 17 (6) (2020) 4046–4054.
[22] Z. Wang, L. Yao, G. Chen, J. Ding, Modified multiscale weighted permutation entropy and optimized support vector machine method for rolling bearing fault diagnosis with complex signals, ISA Trans. 114 (2021) 470–484.
[23] S. Wang, Y. Li, S. Si, K. Noman, Enhanced hierarchical symbolic sample entropy: Efficient tool for fault diagnosis of rotating machinery, Struct. Health Monit. 22 (3) (2023) 1927–1940.
[24] Y. Li, S. Wang, Y. Yang, Z. Deng, Multiscale symbolic fuzzy entropy: An entropy denoising method for weak feature extraction of rotating machinery, Mech. Syst. Signal Process. 162 (2022) 108052.
[25] H. Hong, M. Liang, Fault severity assessment for rolling element bearings using the Lempel–Ziv complexity and continuous wavelet transform, J. Sound Vib. 320 (1–2) (2009) 452–468.
[26] A. Lempel, J. Ziv, On the complexity of finite sequences, IEEE Trans. Inform. Theory 22 (1) (1976) 75–81.
[27] J. Yin, M. Xu, H. Zheng, Fault diagnosis of bearing based on symbolic aggregate approximation and Lempel-Ziv, Measurement 138 (2019) 206–216.
[28] Z. Su, J. Shi, Y. Luo, C. Shen, Z. Zhu, Fault severity assessment for rotating machinery via improved Lempel–Ziv complexity based on variable-step multiscale analysis and equiprobable space partitioning, Meas. Sci. Technol. 33 (5) (2022) 055018.
[29] Y. Li, S. Wang, Z. Deng, Intelligent fault identification of rotary machinery using refined composite multi-scale Lempel–Ziv complexity, J. Manuf. Syst. 61 (2021) 725–735.
[30] J. Yin, X. Zhuang, W. Sui, Y. Sheng, Manifold learning and Lempel-Ziv complexity-based fault severity recognition method for bearing, Measurement 213 (2023) 112714.
[31] Y. Bai, Z. Liang, X. Li, A permutation Lempel-Ziv complexity measure for EEG analysis, Biomed. Signal Process. Control 19 (2015) 102–114.
[32] C.-H. Yeh, W. Shi, Generalized multiscale Lempel–Ziv complexity of cyclic alternating pattern during sleep, Nonlinear Dynam. 93 (2018) 1899–1910.
[33] J.F. Restrepo, D.M. Mateos, G. Schlotthauer, Transfer entropy rate through Lempel-Ziv complexity, Phys. Rev. E 101 (5) (2020) 052117.
[34] X. Mao, P. Shang, M. Xu, C.-K. Peng, Measuring time series based on multiscale dispersion Lempel–Ziv complexity and dispersion entropy plane, Chaos Solitons Fractals 137 (2020) 109868.
[35] C. Barile, C. Casavola, G. Pappalettera, V.P. Kannan, Interpreting the Lempel–Ziv complexity of acoustic emission signals for identifying damage modes in composite materials, Struct. Health Monit. 22 (3) (2023) 1708–1720.
[36] Y. Li, F. Liu, S. Wang, J. Yin, Multiscale symbolic Lempel–Ziv: An effective feature extraction approach for fault diagnosis of railway vehicle systems, IEEE Trans. Ind. Inform. 17 (1) (2020) 199–208.
[37] J. Shi, Z. Su, H. Qin, C. Shen, W. Huang, Z. Zhu, Generalized variable-step multiscale lempel-ziv complexity: A feature extraction tool for bearing fault diagnosis, IEEE Sens. J. 22 (15) (2022) 15296–15305.
[38] A.J. Ibáñez-Molina, S. Iglesias-Parro, M.F. Soriano, J.I. Aznarte, Multiscale Lempel–Ziv complexity for EEG measures, Clin. Neurophysiol. 126 (3) (2015) 541–548.
[39] B. Han, S. Wang, Q. Zhu, X. Yang, Y. Li, Intelligent fault diagnosis of rotating machinery using hierarchical Lempel-Ziv complexity, Appl. Sci. 10 (12) (2020) 4221.
[40] Y. Li, L. Tan, M. Xiao, Q. Xiong, Hierarchical dispersion Lempel–Ziv complexity for fault diagnosis of rolling bearing, Meas. Sci. Technol. 34 (3) (2022) 035015.
[41] F. Xiao, Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy, Inf. Fusion 46 (2019) 23–32.
[42] Z. Xu, M. Bashir, W. Zhang, Y. Yang, X. Wang, C. Li, An intelligent fault diagnosis for machine maintenance using weighted soft-voting rule based multi-attention module with multi-scale information fusion, Inf. Fusion (ISSN: 1566-2535) 86–87 (2022) 17–29.
[43] M. Safizadeh, S. Latifi, Using multi-sensor data fusion for vibration fault diagnosis of rolling element bearings by accelerometer and load cell, Inf. Fusion 18 (2014) 1–8.
[44] P. Zhang, T. Li, G. Wang, C. Luo, H. Chen, J. Zhang, D. Wang, Z. Yu, Multi-source information fusion based on rough set theory: A review, Inf. Fusion 68 (2021) 85–117.
[45] D. Labate, F. La Foresta, G. Morabito, I. Palamara, F.C. Morabito, Entropic measures of EEG complexity in Alzheimer's disease through a multivariate multiscale approach, IEEE Sens. J. 13 (9) (2013) 3284–3292.
[46] H. Azami, A. Fernández, J. Escudero, Multivariate multiscale dispersion entropy of biomedical times series, Entropy 21 (9) (2019) 913.

[47] L. Cao, A. Mees, K. Judd, Dynamics from multivariate time series, Physica D 121 (1–2) (1998) 75–88.
[48] H. Azami, J. Escudero, Refined composite multivariate generalized multiscale fuzzy entropy: A tool for complexity analysis of multichannel signals, Physica A 465 (2017) 261–276.
[49] Z. Wang, P. Shang, Generalized entropy plane based on multiscale weighted multivariate dispersion entropy for financial time series, Chaos Solitons Fractals 142 (2021) 110473.
[50] B. Zhang, P. Shang, Q. Zhou, The identification of fractional order systems by multiscale multivariate analysis, Chaos Solitons Fractals 144 (2021) 110735.
[51] Z. Wang, L. Yao, Y. Cai, J. Zhang, Mahalanobis semi-supervised mapping and beetle antennae search based support vector machine for wind turbine rolling bearings fault diagnosis, Renew. Energy 155 (2020) 1312–1327.
[52] D. Liciotti, M. Bernardini, L. Romeo, E. Frontoni, A sequential deep learning application for recognising human activities in smart homes, Neurocomputing 396 (2020) 501–513.