Computers in Industry 125 (2021) 103378

Contents lists available at ScienceDirect

Computers in Industry

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

Fault detection and diagnosis for rotating machinery: A model based on convolutional LSTM, Fast Fourier and continuous wavelet transforms

Masoud Jalayer*, Carlotta Orsenigo, Carlo Vercellis

Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Via Lambruschini 24/b, 20156, Milan, Italy

ARTICLE INFO

Article history:
Received 29 April 2020
Received in revised form 12 October 2020
Accepted 13 November 2020
Available online 25 December 2020

Keywords:
Fault detection and diagnosis
Convolutional long short-term memory
Deep learning
Continuous wavelet transform
Fast Fourier transform
Feature engineering

ABSTRACT

Fault Detection and Diagnosis (FDD) of rotating machinery plays a key role in reducing the maintenance costs of the manufacturing systems. How to improve the FDD accuracy is an open and challenging issue. To make full use of signals and reveal all the fault features, this paper proposes a new feature engineering model which combines Fast Fourier Transform (FFT), Continuous Wavelet Transform (CWT) and statistical features of raw signals. Then a novel Convolutional Long Short-Term Memory (CLSTM) is developed to understand and classify these multi-channel array inputs. In order to evaluate the effectiveness of the proposed model, three different datasets are used. The paper performs a sensitivity analysis on the input channels to evaluate the efficiency of the proposed multi-domain feature set in different DL architectures, where CLSTM shows its superiority in understanding the feature set. Secondly, a comprehensive review of the state-of-the-art models is conducted, and twelve algorithms are chosen for the comparison to evaluate the performance of the proposed FDD model. The paper also performs an input length sensitivity analysis, showing that the proposed model can achieve 100 % of accuracy with shorter inputs compared to other models, meaning that it causes less delay in an online condition monitoring system. The results demonstrate the superiority of the proposed model over the state-of-the-art models in terms of accuracy on different datasets.

1. Introduction and literature review

During the past few decades, the mechanical equipment technology and science have experienced a rapid growth and development. One of the most important mechanical equipment of our modern industries is rotating machinery. Such machineries are used for long time, usually under severe and rough conditions, leading to their components' breakdown during the operation, which threatens the safe working conditions and causes economic loss. The healthy running condition of the rotating machinery components is crucial, since a considerable portion of the faults of rotating machineries are directly related to these components, such as the rolling bearings (Wang et al., 2012). Rolling bearings get different types of faults, associated with the cracks in their different parts including outer ring, inner ring, rolling elements or cage (Yang et al., 2018), or chipping of the ball due to the working load and stresses caused by an unbalanced shaft and fatigue failure. Such faults affect and alter the patterns of the vibration and the noises transmitted from the machine.

The first appearance of the rotating machinery FDD in the literature dates back to 1969 in Boeing Co., when Balderston (1969) pointed out some effects and characteristics of the fault signs on the signals measured by an accelerometer in natural and high frequencies. Thereupon, in another study the rectified envelope signals with a synchronous averaging were employed to identify bearing local faults (Weichbrodt and Smith, 1970). The fundamental analysis of synchronous averaging and 'high frequency resonance technique' (HFRT), which was later named 'envelope analysis', has been used since then in many studies. Finding the locations of the peaks in the vibration signal spectrum is another classic example of the fault detection methods for the ball bearing faults (Li and Ma, 1997).

Nowadays, practices of artificial intelligence (AI) techniques such as *k*-nearest neighbors (kNN), Support Vector Machines (SVM) and Artificial Neural Networks (ANN) in fault detection and diagnosis systems are of great importance (Li et al., 2011). There are plenty of frameworks in which these techniques are used to detect faults in

* Corresponding author.
E-mail addresses: masoud.jalayer@polimi.it (M. Jalayer), carlotta.orsenigo@polimi.it (C. Orsenigo), carlo.vercellis@polimi.it (C. Vercellis).

https://doi.org/10.1016/j.compind.2020.103378
0166-3615/

various domains. Generally, in the case of rotating machinery FDD, there are two principal steps that most frameworks follow: first, extracting features via a signal processing method. and then feeding the selected features to a fault identification and classification processor to recognize the pattern (Shao et al., 2018).

Generally, the vibration or ultra-sound signals collected from machines are raw temporal signals which consist of both useful information of the machinery health condition and ineffectual noises (Zhang et al., 2017). Some different techniques were proposed in the literature to process these raw information, such as Fourier spectral analysis (Rai and Mohanty, 2007; Banerjee and Das, 2012), wavelet analysis (Li et al., 2013), wavelet transformation techniques (Seo et al., 2017), singular spectrum analysis (Muruganatham et al., 2013), stochastic resonance (Guo et al., 2017; Lu et al., 2016a), time-domain statistical analysis (Wang et al., 2015) and autoregressive modelling (Al-bugharbee and Trenda, 2016). These methods are among the most prominent solutions proposed to extract useful information from sound, vibration and temperature signals (Lu et al., 2016b). After the feature extraction step, a feature selection method is sometimes used to reduce the data size and improve the computational efficiency. Such methods include principal component analysis (PCA) (Misra et al., 2002), independent component analysis (ICA) (Widodo and Yang, 2007) and feature discriminant analysis.

After preprocessing and feature extraction the further step is represented by the employment of a classifier or a regressor to monitor the system's condition. Traditional AI techniques e.g. kNN, Naïve Bayes, Decision Trees, SVM and ANN are among the earliest classifiers used for this purpose. In the past few years, some ensemble learning methods such as random forest (RF) (Cerrada et al., 2016), Boosting and ensemble SVMs (Zheng et al., 2017) have been applied as well. These techniques were capable of great achievements in some fault detection fields (Lee and Kim, 2015), though, they still show troubles in detecting more complex fault patterns, specifically for rotating machinery FDD. According to Zhao et al. (2016) ML-based algorithms often have difficulties in representing complex functions due to their generalization abilities and poor performance.

With the exponential development of DL in different AI applications, its great capacity to overcome the inherent disadvantages of traditional intelligent methods, such as their hand-designed feature dependency and their difficulty in understanding the sequential data (Liu et al., 2017), has been shown by many scholars. Consequently, researchers replaced more traditional classifiers in its favor in many fault diagnosing tasks. The main difference between traditional AI models and DL models is that the former can automatically learn some valuable features from the raw data (LeCun et al., 2015). Since the emergence of DL, many architectures like Deep Belief Network (DBN) and Convolutional Neural Network (CNN) have gradually become the most common used in this field. In particular, CNN has several distinct advantages that make it popular, among which the most salient are its shift-invariance and weight sharing through convolutional connections. The successful training of the hierarchical layers is another advantage of CNN which makes it ideal for multi-channel data processing, such as images. On the other hand, DBN has excellent performance on account of its capability of features extracting due to its complex multi-layer architecture (Zhang and Zhao, 2017). It is a generative neural network which has exhibited a powerful unsupervised feature learning ability (Shao et al., 2018) and can show superiority when dealing with imbalance data (Liu et al., 2017). Therefore, Lee et al. combined these two architectures and developed a new model, called Convolutional Deep Belief Network (CDBN), which showed more promising performance on cloud fault detection compared to its standard parents (Lee et al., 2011). However, it still had some shortcomings when dealing with rolling bearing vibration data, as described in Huang et al. (2014). More recently, some researchers like Shao et al. (2018) tried to enhance CDBN to make it more suitable for rolling bearing applications.

Among DL techniques, autoencoders (AE) are similarly of the most potential tools for automatic feature extraction of mechanical signals. They have been adopted in many fault detection cases, e.g. in semiconductor industry (Lee et al., 2017a), foundry and iron-making processes (Zhang et al., 2016), gearboxes (Liu et al., 2018) and, likewise, in rotating machinery parts (Lu et al., 2017; Haidong et al., 2017). In another study, its "stacked" architecture has been employed to initialize the weights and offsets of a multi-layer neural network and to provide an expert knowledge for spacecraft conditions (Li and Wang, 2015). However, to cope with mechanical signal data its traditional form exhibits some drawbacks; for example, it sometimes learns only similar features in feature extraction and the learned features may have shift variant properties which potentially cause the misclassification. Some approaches were proposed to make this architecture appropriate for signal-based fault diagnosis tasks. Jia et al. used a local connection network on a normalized sparse AE, called NSAE-LCN, to overcome these shortcomings (Jia et al., 2018a). Lei et al. (2016) used stacked-AE to directly learn features of mechanical vibration signals on a motor bearing dataset and a locomotive bearing dataset: specifically, they first used a two-layer AE for sparse filtering and then applied a softmax regression to classify the motor condition. The combination of these two techniques let the method achieved high accuracy in bearing fault diagnosis.

In light of its astonishing performance at capturing long-term dependencies of a sequence, LSTM has been used for condition monitoring of industrial machineries in the last few years. Recently, it has been employed on raw time-series signals of a wind turbine to identify the faults associated to the acceleration, wind or rotor speed and displacement (Lei et al., 2019). Qian et al. (2019) proposed a model to analyze the residual signals obtained by deducting the actual values from the values predicted by an LSTM. They showed that the model had a better performance in the application of the wind turbine condition monitoring. Park et al. (2019) coupled it with an AE binary classifier to enhance the accuracy on imbalance data. In particular, they used the AE classifier for detecting the faults and the LSTM classifier for diagnosing the type of faults based on the raw sensory signals. (Zhao et al., 2017) presented a Convolutional Bi-directional LSTM on raw signal, where CNN extracts local features which are robust and informative from the sequential input and the bi-directional LSTM encodes temporal features. Sabir et al. (2019) employed LSTM on a combination of statistical features and wavelet packet decomposition (WPD) of bearing signals. To reduce the number of LSTM network parameters, (Tan et al., 2020) developed a simplified version of LSTM, a Single Gated Unite (SGU) Recurrent Neural Network which classifies the bearing signals based on their WPDs.

Table 1 summarizes and sorts chronologically some of the novel FDD models proposed in literature.

1.1. Research gap

In recent years, different ML and DL-based strategies have been proposed to cope with FDD problems. While DL techniques have shown promising abilities for FDD systems, their architecture design and the feature engineering are still challenging. The related literature has two main open issues: (1) the development of the feature engineering techniques and (2) the development of proper classifiers to learn the features with a higher performance. The most popular feature types used in the literature are as follows: the statistical features of the signals, time-domain diagram, grayscale diagram, Fourier transform-based diagrams and wavelet transform-based diagrams. Each of these feature types has its own

Table 1
Summary of some recently proposed models for FDD.

Source	*Publishing year*	*Proposed Classifier*	*Input Array*
(Mao et al., 2021)	2021	Stacked discriminant information-based AE (sdiAE)	FFT spectrum
(Shenfield and Howarth, 2020)	2020	RNN-wide first kernel CNN (RNNwdCNN)	Raw signal
(Tan et al., 2020)	2020	SGU-RNN	wavelet packet decomposition (WPD)
(Sabir et al., 2019)	2019	LSTM	WPD + Kurtosis + Impulse Factor
(Chen et al., 2019)	2019	CNN + Extreme Learning Machine (ELM)	CWT
(Liang et al., 2018)	2018	Convolutional Recurrent Neural Network (CRNN)	Ensemble Empirical Mode Decomposition (EEMD) + Autoregressive Spectrum Analysis
(Lu et al., 2017)	2018	Stacked-denoising AE (sdAE)	Raw signal
(Jia et al., 2018a)	2018	Normalized Sparse Autoencoders (NSAE)	Raw signal
(Zhang et al., 2018)	2018	CNN	Raw signal
(Nakazawa and Kulkarni, 2018)	2018	CNN	Raw image
(Shao et al., 2018)	2018	Convolutional Deep Belief Network (CDBN)	Compressed Sensing (CS)
(Lee et al., 2017b)	2017	CNN	Raw signal
(Wen et al., 2017)	2017	CNN (LeNet-5 version)	Raw signal
(Zheng et al., 2017)	2017	Ensemble-SVM	Composite Multiscale Fuzzy Entropy
(Lee et al., 2017a)	2017	Stacked-denoising AE (sdAE)	Raw signal
(Haidong et al., 2017)	2017	AE	Raw signal
(Jana and Abhinandan, 2017)	2017	ANN	Extended Kalman Filter
(Zhao et al., 2017)	2017	Convolutional Bi-directional LSTM (CBLSTM)	Raw signal
(Zhang et al., 2017)	2017	CNN	Raw signal + Statistical
(Gan et al., 2016)	2016	Deep Belief Network (DBN)	Wavelet Packet Transform (WPT)
(Jia et al., 2016)	2016	AE	Frequency spectra
(Yao et al., 2016)	2016	RF	Statistical
(Lei et al., 2016)	2016	Stacked-denoising AE (sdAE)	Raw signal
(Zhang et al., 2016)	2016	Stacked-denoising AE (sdAE)	Statistical
(Zhang et al., 2015)	2015	Ensemble-SVM	FFT + Statistical
(Recioui et al., 2015)	2015	kNN	WPD
(Unal et al., 2014)	2014	ANN	FFT + Hilbert transform + envelope analysis
(Fengqi and Meng, 2006)	2006	SVM	Full-Spectrum Cascade Analysis

advantages and disadvantages in representing the machine condition. Silva et al. compared the capability and robustness of Discrete Fourier Transform (DFT), FFT and CWT to represent rub detection, based on gas turbine acceleration signals (Silva et al., 2020). Recently, few papers used a combination of some feature types to harness the advantages of them simultaneously. (Zhang et al., 2015) and (Unal et al., 2014) used FFT and statistical features of the signal time-domain, whereas (Zhang et al., 2017) used statistical features and raw signals together. (Sabir et al., 2019) presented a framework based on WPD and the signal kurtosis. In (Džakmić et al., 2018), the authors presented a combination of a Fourier transform (FFT or STFFT) and Mexican-hat wavelet for transmission and distribution networks. It firstly finds the fault's harmonics via Fourier diagram and after a calculation on the scale factor, afterwards, it employs Mexican-hat wavelet. Another work used bagged decision trees to localize the fault in microgrids using the features extracted from FFT and Windowed Wavelet Transform (Netsanet et al., 2018). However, the two approaches that coupled Fourier transform with wavelet transform techniques only presented fault signature identification, and no DL architecture was employed to demonstrate if a DL architecture can efficiently use such feature spectra. On the other hand, to the best of our knowledge, the combination of Fourier transform-based diagrams, wavelet transform-based diagrams and time-domain features has not been discussed yet in the literature.

1.2. The paper's contributions

The key contributions of this paper are as follows:

1 To get a deeper insight into the fault type signatures, a novel feature engineering framework is designed, parallelizing a set of channels driven from three different popular feature types: FFT, CWT and time-domain spectrum. The framework adds some time-domain-based statistical features to the classifier, to make a full use of all possible fault signatures.
2 A hybrid Convolutional LSTM architecture is designed which firstly employs one-dimensional CNN to extract the local features from the multi-channel data. Secondly, it utilizes a LSTM for handling temporal correlations.
3 The efficiency of the feature engineering is evaluated using three different DL architectures.
4 A comprehensive review of the state-of-the-art FDD models is conducted, illustrating the results of different papers which used the Case Western Reserve University (CWRU) bearing dataset,their classifier and input descriptions.
5 Several state-of-the-art FDD models are selected to realize a comprehensive comparison of the performances on three rotating machinery datasets.
6 A sensitivity analysis is conducted on the input length, illustrating that the proposed model can reach 100 % of accuracy with shorter input lengths compared to the other models.

The rest of the paper is organized as follows. Section 2 briefly introduces Convolutional and LSTM layers. The proposed method of FDD is presented in Section 3. In Section 4 comprehensive experimental comparisons on different models and datasets are described. Section 5 contains some discussions and conclusions.

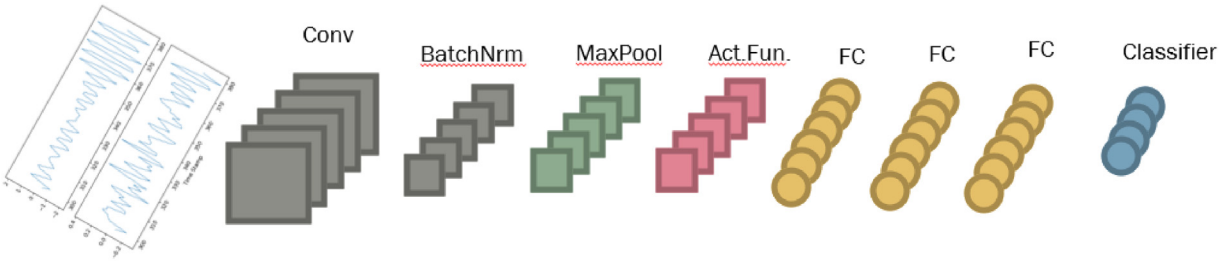

Fig. 1. The Architecture of CNN for Signal Fault Diagnosis.

2. Materials and methods

This section provides some insights about the convolutional layers and LSTM architecture.

2.1. Convolutional neural networks (CNN)

CNN is a multi-stage neural network which is composed of some filter stages and one classification stage. It was inspired by the visual system's structure and developed by LeCun and collaborators in 1990 (Lecun et al., 1990) for image processing, and still is widely used in computer vision applications and are among the best performing systems for pattern recognition purposes (Bengio, 2009). In general terms, CNN architectures contain 2 blocks: the filter block, which is designed to extract features from the inputs and comprises at least one convolutional layer and some different kinds of layers (e.g. pooling layer, batch normalization layer), and the classification block, which is basically a multi-layer perceptron composed of several fully connected layers. Fig. 1 depicts the general structure of a CNN architecture, having the layers as follows:

2.1.1. Convolutional layer

The convolutional layer convolves the input local regions with filter kernels followed by the activation unit to generate the output features. Each filter uses the same kernel to extract the local features of the input local region, which is usually referred to as weight-sharing in the literature. One filter corresponds to one frame in the next layer and the number of frames is called the depth of this layer. K_i^l is used to denote the weights and bias of the i^{th} filter kernel in layer l, respectively, and $x^{l(r^j)}$ to denote the j^{th} local region in layer l. Therefore, the convolution process is described as follows

$$y^{l(i,j)} = K_i^l * x^{l\left(r^j\right)} = \sum_{j'=0}^{W} K_i^l\left(j'\right) x^{l(j+j')}, \tag{1}$$

Where the notation $*$ indicates the dot product of the kernel and the local regions and $K_i^l(j')$ denotes the j'^{th} weights in frame i of layer $l+1$. W is the width of the kernel.

2.1.2. Pooling layer

It is prevailing to have a pooling layer after convolutional layer since it reduces the spatial size of the features and consequently makes the learning stage quicker. There are several types of pooling functions, among which *max-pooling* is the most popular. Max-pooling layer performs the local max operation over the input features with a certain kernel/pool size and gains location-invariant features. The max-pooling transformation is described as:

$$p^{l(i,j)} = \max_{(j-1)W+1 \le t \le jW} \{a^{l(i,t)}\}. \tag{2}$$

2.1.3. Batch normalization layer

To reduce the shift of interval covariance and make the learning faster by less computations load, a *batch normalization layer* is introduced. It is also usually added immediately after convolutional layer or before activation function layer. The batch normalization transformation can be described as

$$\hat{y}^{l(i,j)} = \frac{y^{l(i,j)} - \mu_B}{\sqrt{(\sigma_B^2 + \varepsilon)}} \tag{3}$$

$$z^{l(i,j)} = \gamma^{l(i)} \hat{y}^{l(i,j)} + \beta^{l(i)}, \tag{4}$$

Where $z^{l(i,j)}$ is the output of one neuron response, $\mu_B = E[y^{l(i,j)}]$, $\sigma_B^2 = Var[y^{l(i,j)}]$, ε is a small constant added for numerical stability, $\gamma^{l(i)}$ and $\beta^{l(i)}$ are the scale and shift parameters to be learned, respectively.

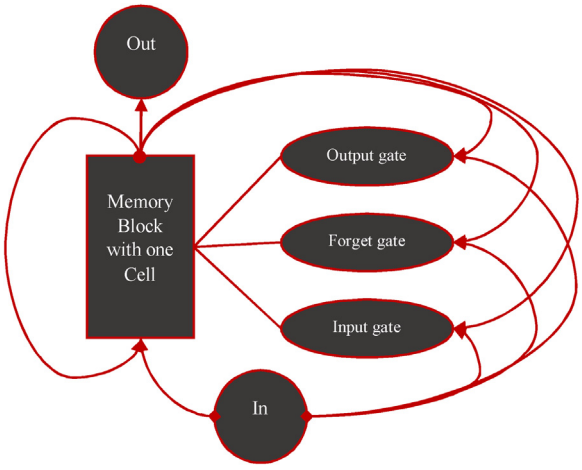

Fig. 2. A three-layer LSTM topology with one input and one output unit.

2.1.4. Activation function layer

The presence of activation function layer is vital in any convolutional block. It lets the network achieve a nonlinear expression of the input signal, which enhances the representation and makes the learned features more distinguishable. There is a variety of activation functions that can be adopted: among these, ReLU, Identity, Tanh and Sigmoid are broadly used.

2.2. Convolutional long short-term memory (CLSTM)

Recurrent Neural Networks (RNN) are widely used for sequence learnings, although the vanishing gradient problem of the training backpropagation step impedes its performance. To avoid this impediment and capture long-term dependencies of the data features, RNN have been improved to a new architecture by Hochreiter and Schmidhuber (1997) called Long Short-term Memory (LSTM). It has shown more efficient classifications and regressions performances on time series e.g. voices and natural language processing datasets than those of RNN. Fig. 2 illustrates a simple three-layered LSTM topology (Gers and Schraudolph, 2002). Despite some com-

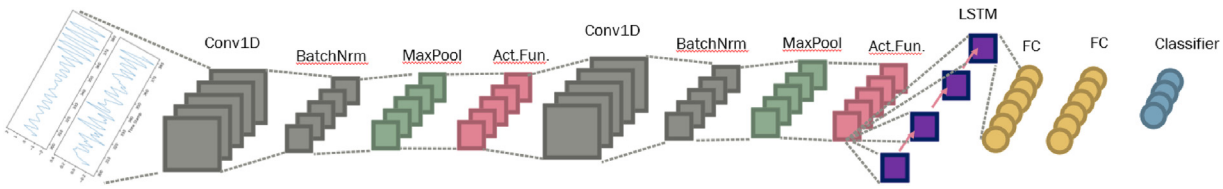

Fig. 3. Schematic architecture of the CLSTM architecture.

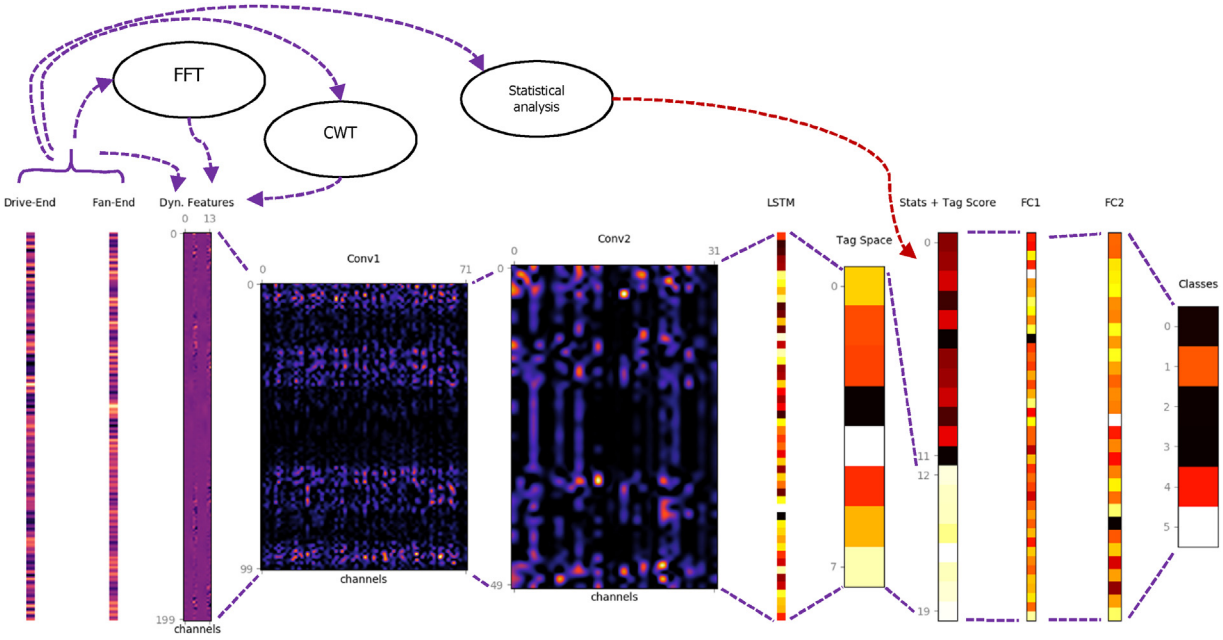

Fig. 4. Overall structure of the proposed model.

mon elements that exist between RNN architectures and LSTM, the main novelty of LSTM is the introduction of the *forget gate*, which regulates the use of information in the cell state to overcome the long-term dependency problem. LSTM showed its superiority compared to traditional RNNs when it was applied to tasks such as speech recognition, image captioning, genomic analysis and natural language processing, being able to account for the long-term dependencies and non-linear dynamics (Zhao et al., 2017).

By combining convolutional layers explained in Section 2.1 with LSTM architecture we obtain a Convolutional LSTM (CLSTM). Convolutional layers enable the architecture to extract different types of the waveform input arrays, whose abstract output is immediately processed by the LSTM layer to analyze its intrinsic sequential, periodic behavior. According to Shi et al. (2015), CLSTM can capture spatio-temporal correlations better than other architectures, like CNN or simple LSTM. Fig. 3 draws a diagram for a CLSTM architecture.

3. The proposed method

As mentioned in 1.1, there is a lack of attention and research regarding the capability of DL techniques to gain knowledge simultaneously from Fourier transform-based diagrams, wavelet transform-based diagrams and time-domain diagrams. Therefore, this paper presents a framework which embeds two novelties: on the one hand, the simultaneous employment of FFT, CWT and statistical features alongside with the time-domain diagram to provide the DL classifier a deeper understanding of the fault's identity. On the other hand, the use of novel CLSTM in rotatory machines FDD, illustrated in Fig. 4.

The CNN blocks (Conv1 and Conv2) consist of 1D-Convolutional layers accompanied by 1D-Batch Normalization, ReLU and Max Pooling layers, to extract invariant features. The pooling layers are added to reduce the number of parameters and ease the calculation. The proposed model uses a LSTM block augmented to the CNN blocks to learn the spatial and temporal dependencies characterizing the fault types. It is followed by SoftMax and concatenation layers to include the statistical features. Finally, three fully connected neural networks are added to classify the augmented features.

If the sample bursts are collected from D different coupled sensors, with the time length of L, the i^{th} signal burst can be represented as a matrix $X^i_{D\times L}$. The proposed model uses CWT and FFT functions to obtain another representation of it, given by $\hat{X}^i_{\Psi\times L} = f\left(X^i_{D\times L}\right)$. Therefore, the model runs the following steps:

- CWT technique is used to acquire the first K components of the raw signal

$$C^i_{(K\times D)\times L} = f_{CWT}\left(X^i_{D\times L}\right), \tag{5}$$

Where $C^i_{(K\times D)\times L}$ is a matrix composed by K vectors, represented by the first K mother wavelets for each sensor signal in i^{th} instance.

- Similarly, FFT technique is used to acquire a new representation of the raw signal in two "real" and "imaginary" vectors, $\{\alpha_R, \alpha_I\}$. For the i^{th} instance, therefore, the model will have

$$A^i_{(2\times D)\times L} = f_{FFT}\left(X^i_{D\times L}\right). \tag{6}$$

- The model concatenates X^i, C^i and A^i on their first dimension

$$\hat{X}^{i}_{\Psi\times L} = X^{i}_{D\times L} \oplus C^{i}_{(K\times D)\times L} \oplus A^{i}_{(2\times D)\times L}, \tag{7}$$

Where, $\Psi = D(K+3)$.

- After the previous steps, $\hat{X}^{i}$ has become a one-dimensional-multichannel matrix, having Ψ channels with the same length of L. As a consequence, it can be given in input to the filter block of a DL architecture:

$$\hat{\hat{X}}^{i}_{\Psi\times L} = f_{FB}\left(\hat{X}^{i}_{\Psi\times L}\right). \tag{8}$$

- $\hat{\hat{X}}^{i}_{\Psi\times L}$ is flattened to a vector:

$$v^{i}_{(\Psi\times L)\times 1} = vec\left(\hat{\hat{X}}^{i}_{\Psi\times L}\right). \tag{9}$$

- Some statistical features of each raw sensor signals are then calculated to combine the output of the filter block

$$x^{i}_{((\Psi\times L)+2D)\times 1} = v^{i}_{(\Psi\times L)\times 1} \oplus s^{i}_{D\times 1} \oplus r^{i}_{D\times 1}, \tag{10}$$

Where $s^{i}_{D\times 1}$ and $r^{i}_{D\times 1}$ are the vectors containing standard deviations and range of each D sensor signals, respectively.

- Finally, the x^i can be inputted to any classifier algorithm; in the present study, the model uses a three-layer fully connected ANN.

4. Experimental evaluation

4.1. Data sets introduction

4.1.1. Case Western Reserve University bearing dataset

This paper evaluates the models on a real bearing dataset[1] provided by Case Western Reserve University (CWRU), where the motor bearing has been seeded with faults using electro-discharge machining. The dataset consists of five different types of faults corresponding to inner race, balls and outer race in three different orientations: 3 o'clock (directly in the load zone), 6 o'clock (orthogonal to the load zone) and 12 o'clock (opposite to the load zone). Moreover, the faults are collected in a range of severity varying between 0.007 in. to 0.040 in. in diameter. The dataset is also recorded for different motor loads, from 0 to 3 horsepower. However, for the sake of simplicity this paper uses only one motor speed of 1797 RPM.

The vibration signals were collected with three accelerometer sensors, one mounted on the drive end (DE), another on the fan end (FE) and the last one on the base (BE). Nevertheless, due to the lack of BE signal for the healthy condition we only used the data collected from DE and FE (Fig. 5).

4.1.2. Gearbox fault diagnosis dataset

The Gearbox fault diagnosis dataset[2], which is here briefly called Gearbox, was collected by recording the vibration data through the SpectraQuest's Gearbox Simulator. It consists of raw vibration signals measured by means of sensors in four different directions. The sensors' frequency is fixed on 30 Hz and the motor load is set in a range between 0 and 90 percent. Gearbox dataset comprises originally two categories, represented by "Healthy" and "Broken tooth" conditions, respectively.

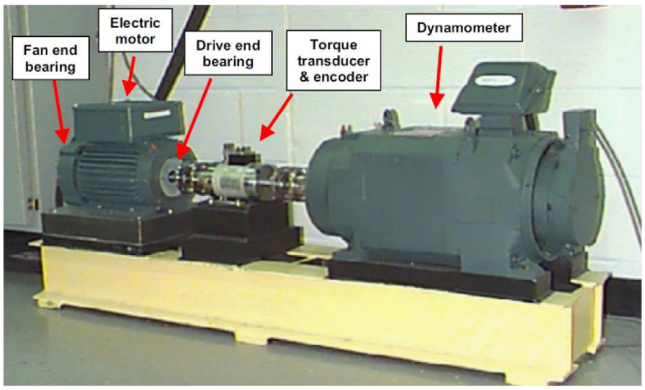

Fig. 5. A two-horsepower motor (left), a torque transducer/encoder (center) and a dynamometer (right) used to collect the dataset.

4.1.3. Wind turbine data

The Wind Turbine dataset, recently published by Zappala et al. (2019), was obtained by using a real wind turbine with a 30 kW induction generator. In particular, signals were collected in some different situations, under rotor electrical unbalance (REU) and healthy conditions at varying loads and fault levels. The dataset includes three motor speeds of 1530, 1560 and 1590 rpm. Authors introduced some additional phase resistances equal to 0.099Ω, 0.1485Ω and 0.198Ω into one rotor phase to get 150 %, 225 % and 300 % REU, respectively. In the present study one of the provided sensors measuring the mechanical torque was analyzed, which gathered the experimental data related to two conditions: the "healthy" data associated to the balanced state and the "faulty" data collected from 300 % REU signals. Therefore, similarly to the Gearbox dataset, the Wind Turbine dataset involves a binary classification problem.

4.2. Data pre-processing and initialization

The raw CWRU data corresponding to two types of faults are plotted in Fig. 6. As it is shown, the raw data comprises several long time-series for each bearing condition. To make sample bursts in order to build the training, the test and the validation sets, some procedures were applied as explained below. The labeling system for the Gearbox and the Wind Turbine datasets is simply a binary one, where the zero value is associated to the healthy samples. The labels for the CWRU dataset are, instead explained in Table 2.

For the sample-making process the study used different time-windows, called burst length. Specifically, for each evaluation a dataset corresponding to a specific burst length (ranging from 20 to 500 timestamps) was built. Two time horizons were considered to divide the training, the validation and the test sets. Specifically, the training samples were gathered from the lower range of the first time horizon point, the validation samples were taken from the interval between the two horizons and the test samples were collected from beyond the second horizon.

The parameter '*step*' shows the shift between two adjacent samples. In general, deep learning architectures need to be fed with abundant samples to reach a remarkable performance. Thus, to collect more samples we set the parameter *step* as small as possible. Notice that, by fixing *step = 1* the maximum number of samples

[1] http://csegroups.case.edu/bearingdatacenter/pages/download-data-file
[2] https://openei.org/datasets/dataset/gearbox-fault-diagnosis-data

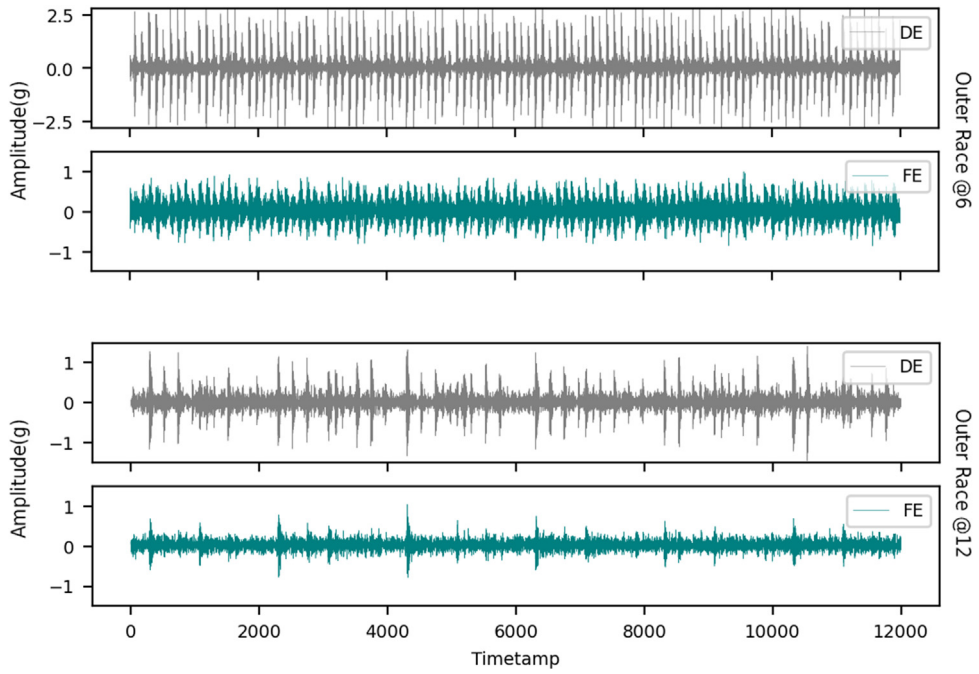

Fig. 6. CWRU Faulty signals corresponding to outer race @6 o'clock (upper dual signal) and outer race @12 o'clock (lower dual signal).

Table 2
CWRU Labeling system.

Full label	*Short label*	*Numerical label*
Healthy baseline condition	Baseline	0
Inner race fault	Inner	1
Ball fault	Ball	2
Outer race fault Centered (@6:00) position	Outer1	3
Outer race fault Orthogonal (@3:00) position	Outer2	4
Outer race fault Opposite (@12:00) position	Outer3	5

can be achieved. However, to avoid getting 'out-of-memory' problems, the *step* was here set to 5 for the Wind Turbine, 10 for the Gearbox and 20 for the CWRU dataset. By choosing the training time horizon as the 80 % of the original raw signal, with *steps = 1* and *length = 100,* we were able to collect ~1.5 million samples for the training set and more than 365,000 samples for the validation and the test sets on the CWRU dataset. Similarly, it was possible to collect at least 163,000 and 480,000 bursts for training and 37,000, 108,000 samples for validation and test sets for the Gearbox and Wind Turbine dataset, respectively. Fig. 7 is an example that illustrates the sample-making process and labeling for a small time-window of raw signals recorded with two sensors, where *step = 40* and *length=80*.

Fig. 8 shows a representation of some random samples from CWRU dataset with burst length of 100, after the proposed feature extraction step, ready to be fed to the filter block.

4.3. Evaluation

To evaluate the effectiveness of the proposed approach we resorted to common quality metrics represented by accuracy, f1-score, recall and precision. These are defined as follows and can be directly obtained by the so-called confusion matrix, depicted in Table 3.

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{11}$$

$$recall = \frac{tp}{tp + fn} \tag{12}$$

$$precision = \frac{tp}{tp + fp} \tag{13}$$

$$f_1 - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{14}$$

To have a better visualization of the true and false predictions, we employed normalized confusion matrices, where the element in each cell is defined as in Table 4.

Therefore, to calculate the aforementioned metrices for CRWU, which is a multi-class dataset, we resorted to a weighted scheme implemented by the sci-kit[3] python library.

Notice that, all the methods were implemented and run on a personal laptop with an i7 6500U processor, 8 GB of memory and a Python 3.6 language environment. Since specific choices of some initial parameters, such as bursts longer than 500 or smaller *step* values, aimed at collecting a greater number of samples would have caused out-of-memory errors, we set the experimental evaluation carefully to achieve a comprehensive understanding of the models' behaviors despite the computational limitations at hand. All the results which will be presented in this part are the classification performances of the trained models on the test sets, which were separated from both training and validation sets. The validation sets were used for the parameter tuning of the models.

4.3.1. The efficiency of the extracted feature combination

The performance of each architecture was assessed for different combinations of the input channels, to figure out whether inputting the proposed channels strengthens the classifiers or not. To this aim, CRWU dataset is used and the initial parameters to collect the samples were set to *step = 20* and *length = 50.* The confusion matrices and f1-scoresare shown in Fig. 9. These results confirm the feature combination efficiency in achieving higher f1-scores with all three DL architectures. In the figure, '*de*', '*fe*' and '*stat*' stand for 'drive-end raw sensory signal', 'fan-end raw sensory signal' and 'statistical features', respectively.

[3] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

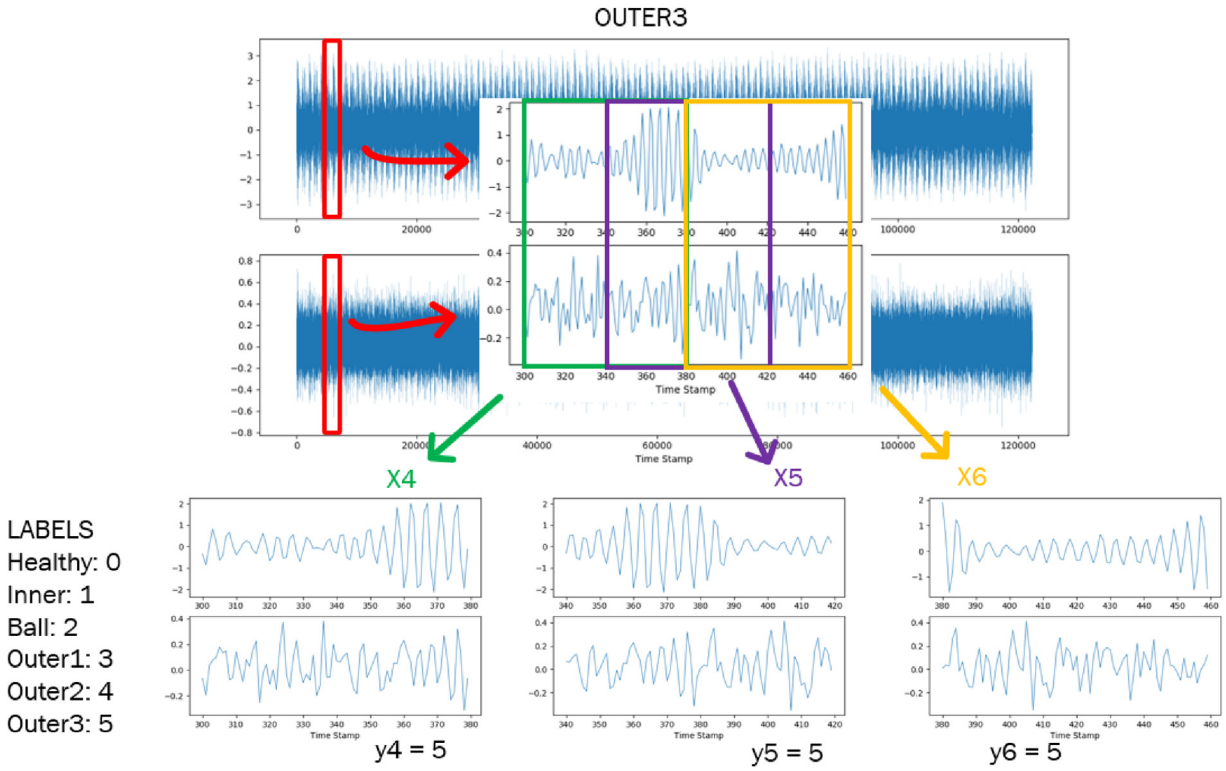

Fig. 7. The sample making and labeling process.

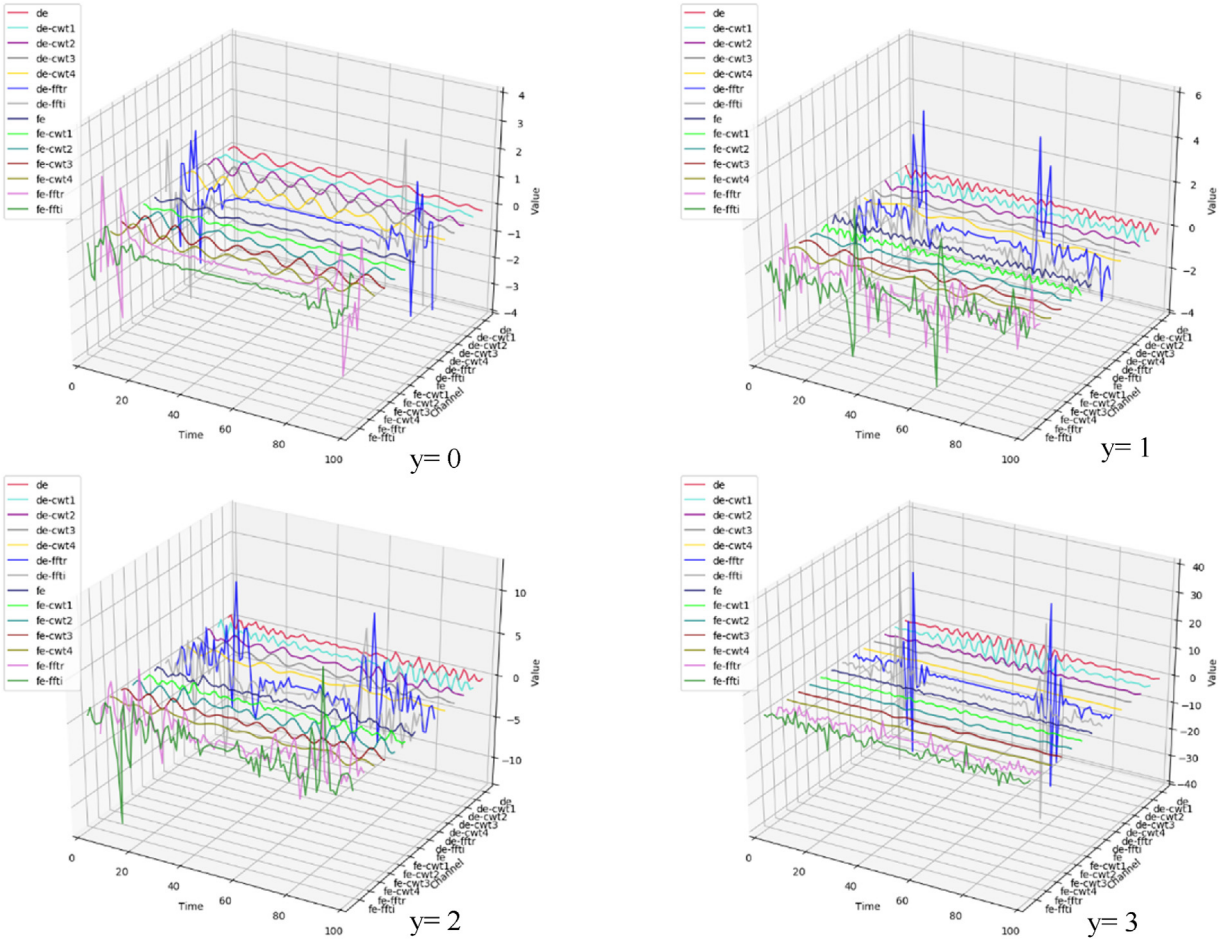

Fig. 8. Illustration of different channels for some random samples with different classes.

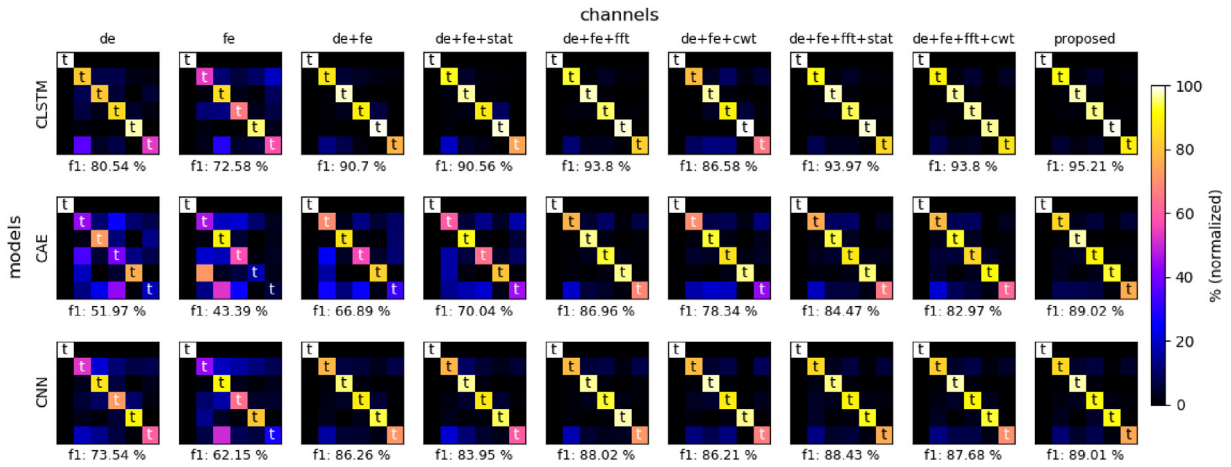

Fig. 9. Normalized confusion matrices for different channel combinations (t: true labeled classes).

Table 3
Confusion Matrix (tn: true negative, fn: false negative, fp: false positive, tp: true positive).

		Predicted class	
		negative	*positive*
Actual class	*negative*	*tn*	*fp*
	positive	*fn*	*tp*

Table 4
Normalized matrix elements.

		Predicted class	
		negative	*Positive*
Actual class	*negative*	*tn / (tn + fp)*	*fp / (tn + fp)*
	positive	*fn / (tp + fn)*	*tp / (tp + fn)*

In order to see whether Fourier-transform and wavelet-transform domains generate unique features, and add values to the time-domain spectra, we used three renown correlation methods: Pearson, Spearman and Kendall. To this end, we picked 1000 random samples from the training dataset of the gearbox dataset with burst length of 200 timestamps and performed a correlation analysis after normalizing the features. The 'distribution' column in Table 5 represents the percentage of the sample whose features exhibited a meaningful correlation with each other. Considering the coefficient average values and standard deviations, it can be concluded that there is not a significant correlation between FFT and CWT features, and that they can generate unique spectra. However, as CWT layers are intrinsically mother wavelets of the raw signals, it could be expected that raw signals and CWT levels show a meaningful correlation. It is worth mentioning that any possible repeated feature will be handled within the proposed CLSTM architecture by pooling layers which down-sample the feature map.

Some FDD approaches mentioned in Table 1 are similar to '*de + fe*' (raw signals), '*de + fe + cwt*', '*de + fe + fft*' or '*de + fe + stat*', all of which are outperformed by the proposed feature combination. Fig. 9 also demonstrates that CLSTM learns significantly better from each of these feature sets. At its best performance, it reaches an f1-score slightly greater than 95 % when it is coupled with the proposed feature set, while both CNN and CAE achieve approximately an 89 % score. When learning from the raw signals, instead, CLSTM performs almost 24 % and 7% better than CAE and CNN, respectively, in terms of f1-score.

4.3.2. *The performance comparison and the importance of burst length*

CWRU bearing dataset has been used in the literature by many researchers. Table 6 illustrates the performance and description of some of the proposed models using this dataset. As it can be clearly seen in the table, only few of these models reached 100 % of accuracy, using burst lengths of at least 2048 which is longer and needs more computational power compared to the burst lengths considered in the present study. In this section, we conduct a sensitivity analysis on the burst length and demonstrate that the proposed

Table 5
Correlation between the feature diagrams.

Features	*Method*	*Distribution*	*Coef. mean value*	*Coef. STD* [a] *value*
CWT[1], FFT[1]	Pearson	7.0 %	0.12041	0.32603
CWT[1], FFT[1]	Spearman	5.7 %	0.00819	0.3428
CWT[1], FFT[1]	Kendall	4.4 %	−0.05378	0.23999
CWT[1], FFT[2]	Pearson	9.2 %	−0.02332	0.35539
CWT[1], FFT[2]	Spearman	6.0 %	0.08769	0.36017
CWT[1], FFT[2]	Kendall	6.2 %	0.02625	0.25004
Raw, FFT[1]	Pearson	15.0 %	0.17025	0.29195
Raw, FFT[1]	Spearman	11.5 %	0.15831	0.3058
Raw, FFT[1]	Kendall	9.6 %	0.12174	0.2126
Raw, FFT[2]	Pearson	15.7 %	−0.11499	0.36115
Raw, FFT[2]	Spearman	13.7 %	−0.15255	0.3524
Raw, FFT[2]	Kendall	14.2 %	−0.11203	0.2412
Raw, CWT[2]	Pearson	83.9 %	0.59892	0.18095
Raw, CWT[2]	Spearman	81.1 %	0.59763	0.17222
Raw, CWT[2]	Kendall	83.6 %	0.43334	0.14705

[a] Standard deviation.

Table 6
Performance of different FDD models on CWRU.

Model	*Usage*	*Input*	*Length*	*LSA*	*Training size*	*Highest accuracy*	*Reference*	*Publication year*
sdiAE	Proposed	FFT	2048	NA	900	96.13%	(Mao et al., 2021)	2021
sdAE	Benchmark	FFT	2048	NA	900	95.01%	(Mao et al., 2021)	2021
MC-CNN	Benchmark	Raw signal	2048	NA	900	95.56%	(Mao et al., 2021)	2021
DBN	Benchmark	FFT	2048	NA	900	94.07%	(Mao et al., 2021)	2021
SAE	Benchmark	FFT	2048	NA	900	92.96%	(Mao et al., 2021)	2021
RNNwdCNN	Proposed	Raw signal	4096	NA	56250	100 %	(Shenfield and Howarth, 2020)	2020
wdCNN	Benchmark	Raw signal	4096	NA	56250	99.89 %	(Shenfield and Howarth, 2020)	2020
SVM	Benchmark	FFT	4096	NA	56250	94.87%	(Shenfield and Howarth, 2020)	2020
MLP	Benchmark	FFT	4096	NA	56250	90.00%	(Shenfield and Howarth, 2020)	2020
SGU-RNN	Proposed	WPD	2048	NA	2420	99.58%	(Tan et al., 2020)	2020
CNN-stft	Proposed	STFT	4096	256 to 4096	4000	100 %	(Zhang et al., 2020)	2020
sdAE	Proposed	Raw signal	200	50 to 200	488000	99.83%	(Lu et al., 2017)	2018
CNN-HMMS	Proposed	Raw signal	100	NA	9600	98 %	(Wang et al., 2018)	2018
PSO-SVM	Proposed	Statistical + FFT + VMD	2048	NA	300	100 %	(Yan and Jia, 2018)	2018
ELM	Benchmark	Statistical + FFT + VMD	2048	NA	300	98 %	(Yan and Jia, 2018)	2018
wdCNN	Proposed	Raw signal	1025	NA	19800	99.93 %	(Zhang et al., 2017)	2017
SVM	Benchmark	FFT	1025	NA	19800	70%	(Zhang et al., 2017)	2017
DBN	Proposed	WPT	4096	NA	500	99.03 %	(Gan et al., 2016)	2016
ANN	Benchmark	FFT	4096	NA	500	95.14 %	(Gan et al., 2016)	2016

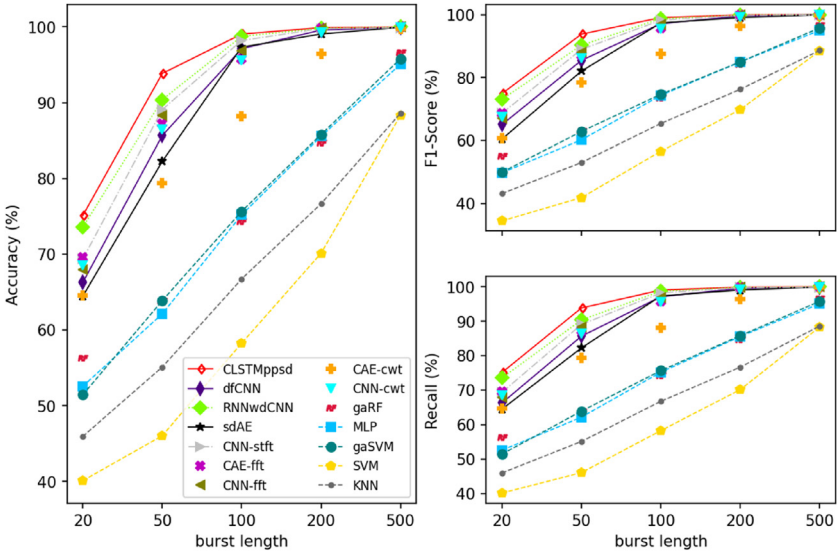

Fig. 10. Accuracy, f1-score and recall of each model for CWRU.

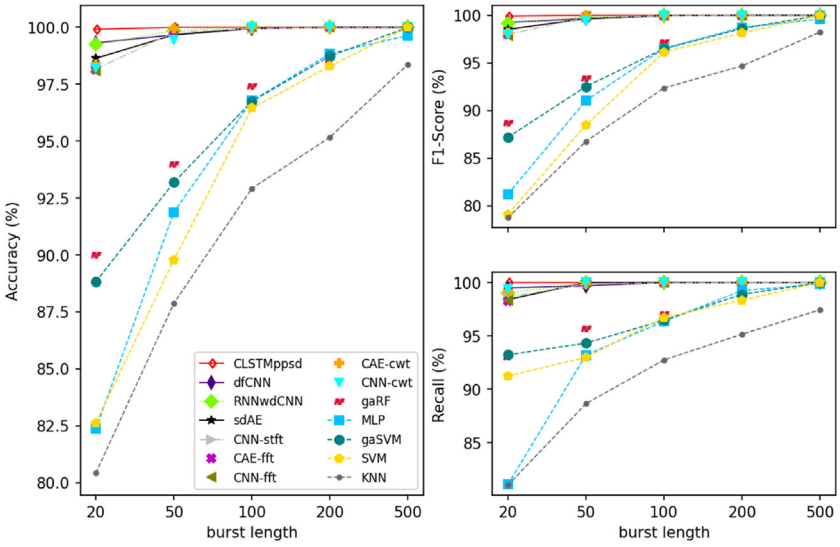

Fig. 11. Accuracy, f1-score and recall of each model for Gearbox.

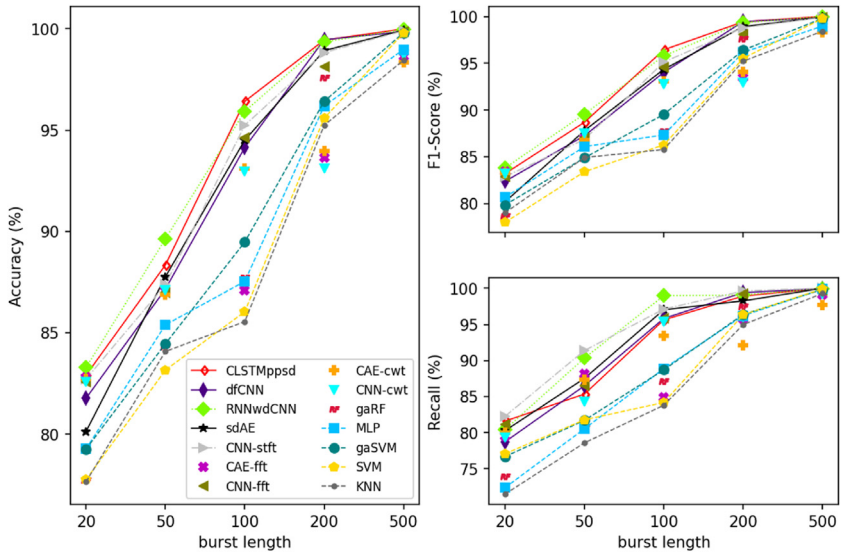

Fig. 12. Accuracy, f1-score and recall of each model for Wind Turbine.

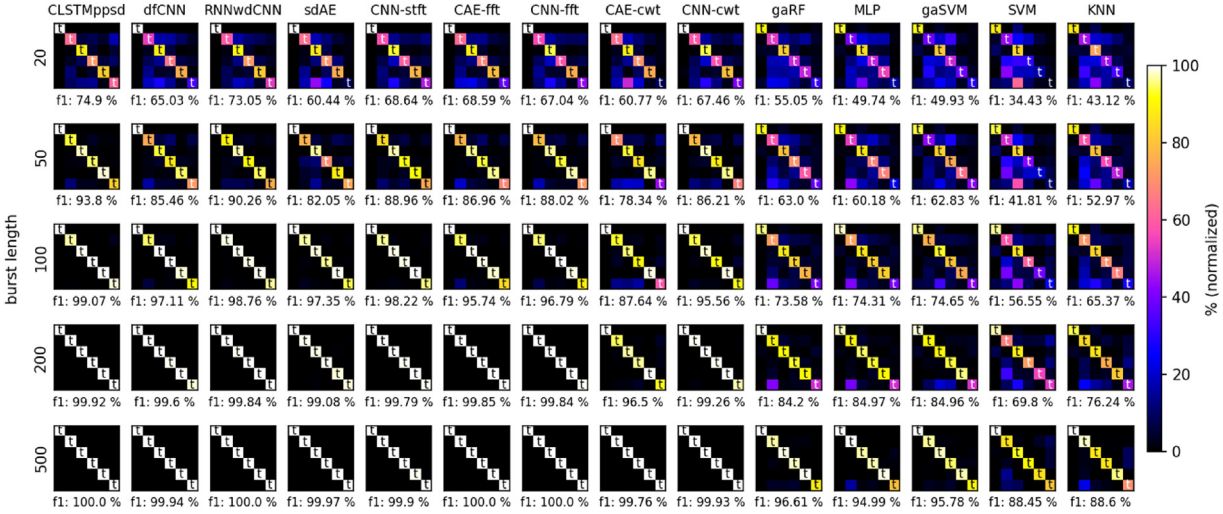

Fig. 13. Normalized confusion matrices for CWRU (t: true labeled classes).

model reaches 100 % of accuracy with a burst length of less than 500 timestamps.

In Zhang et al. (2017) the length was set approximately to 1025 timestamps and the architecture therein proposed reached a maximum accuracy of 99.93 %. Lu et al. (2017), instead, set the length to 200 timestamps and employed different models, from SVM and RF to stacked-denoising AE (sdAE), achieving accuracy values between 84.0 % and 95.5 %. On the same dataset, Gan et al. (2016) used three different types of models (i.e. DL-based, SVM-based and back-propagation ANN-based models), with burst lengths of 4096 timestamps; the highest accuracies obtained were 99.03 %, 96.48 % and 95.14 %, for the three models, respectively.

In order to have a clearer insight about the performance of the proposed model and validate its superiority, this study performs an experimental comparison among several different state-of-the-art and classical models on the aforementioned datasets. The specifications of the selected models are displayed in Table 7. These models have shown promising performances on different rotating machinery datasets; therefore, we included them in the comparison panel. Most of the classifier parameters were selected by either a grid or a random search to find a near-optimal setup for each competing algorithm. In the case of DL-based methods the number of epochs was set to 10, with the batch-size of 32 and a learning rate equal to 1×10^{-4}.

As mentioned earlier, to collect different datasets and have a broader perspective about the advantages and disadvantages of the proposed model a set of burst lengths ranging from 20 to 500 timestamps was considered. In particular, the *step* was set to 20, 10 and 5 timestamps for the CWRU, Gearbox and the Wind Turbine dataset, respectively. The metric values and the confusion matrices of the models are shown from Figs. 10–15.

On the three datasets the proposed model (CLSTMppsd) outperformed the competing methods and proved its superiority in terms of classification accuracy and f1-score, as indicated in Figs. 14–16. Nonetheless, similar to dfCNN, RNNwdCNN and sdAE, it needed more time to be trained compared to CAEfft, CNNfft, CAEcwt and CAEcwt, and rather longer time to classify new samples (Fig. 17); a couple of milli-seconds per sample at its most. On the burst lengths shorter than 200 timestamps, gaRF and gaSVM were the slowest models to train.

As the results perceptibly display, in almost all cases the greater the burst length was, the higher the accuracies and f1 scores were. Notice that, the burst length impacts dramatically on most of the classifiers' training speeds, either positively or negatively, as shown

Table 7
The model specifications of the comparison panel.

Model	*Input*	*Description*	*Reference*
CLSTMppsd	Proposed features	Its architecture comprises two CNN blocks (containing 1D-Convolutional layers, Batch Normalization, ReLU and Max Pooling), a LSTM block, a Logarithmic SoftMax, a concatenation which adds statistical features and three fully connected neural networks for the classification.	Ours
dfCNN	Raw signal	The raw signal is reshaped to a 2D image, two CNN block blocks (containing 2D-Convolutional layers, Batch Normalization, Max Pooling), and 3 fully connected layers are used to classify the results.	(Mao et al., 2021)
RNNwdCNN	Raw signal	Its architecture comprises five CNN blocks (containing 1D-Convolutional layers, Batch Normalization, ReLU and Max Pooling) and a 1D-wide kernel Convolutional layer coupled with a RNN block and a SoftMax layer.	(Shenfield and Howarth, 2020)
sdAE	Raw signal	sdAE is a multilayered architecture composed of four auto-associative neural network layers, which represent one input layer and three AEs.	(Lu et al., 2017)
CNN-stft	STFT	The architecture consists of three CNN blocks (containing one 1D-Convolutional layer and a Pooling layer), two fully connected layers, and a SoftMax classification layer.	(Zhang et al., 2020)
CAE-fft	FFT	Its architecture comprises four CNN blocks (containing 1D-Convolutional, ReLU and Max Pooling layers) such that the first two CNN blocks make an encoder and the last two make the decoder. Three fully connected layers are added to the end of the AE.	(Shen et al., 2018)
CNN-fft	FFT	Its architecture comprises two CNN blocks (containing 1D-Convolutional and Pooling layers) followed by three fully connected layers.	(Jia et al., 2018b)
CAE-cwt	CWT	Similar to CAE-fft architecture	–
CNN-cwt	CWT	Similar to CNN-fft architecture	(Chen et al., 2019)
gaRF	Statistical features of WPD and raw signal	A Genetic Algorithm for feature selection and optimizing the RF number of estimators, and criteria (between 'gini' and 'entropy') parameters	(Cerrada et al., 2016)
MLP	Statistical features of raw signal	Multi-layer perceptrol with three layers and ReLU activation function	Classical
gaSVM	Statistical features of raw signal + FFT	A Genetic Algorithm for feature selection and optimizing the SVM kernel and degree parameters	(Yan and Jia, 2018)
SVM	Statistical features of raw signal	SVM with polynomial kernel and degree of 2	Classical
KNN	Statistical features of raw signal	kNN with 10 number of neighbors to calculate, uniform weights, Euclidean distance and leaf size of 30.	Classical

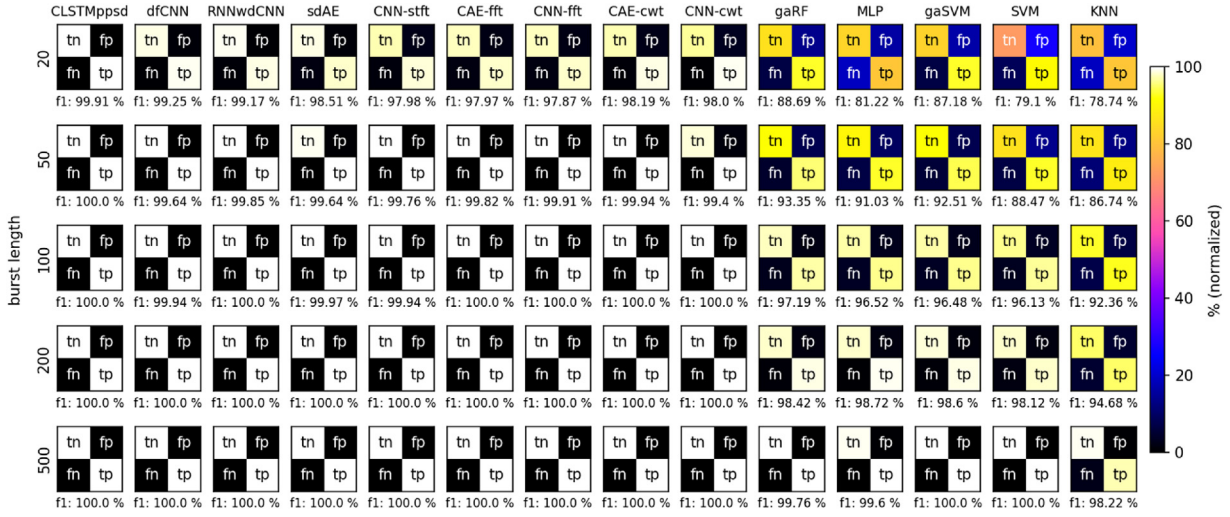

Fig. 14. Normalized confusion matrices for Gearbox.

in Fig. 16. For DL-based models shorter burst-length corresponds to shorter training and test time, which compensates the models' intrinsic slower speed. It is worthwhile to observe that, the proposed model needs shorter burst-length to achieve 100 % of accuracy, and is the only model reaching 100 % accuracy when burst length is equal to 500 timestamps on the Wind Turbine dataset. On the contrary, CNNcwt emerged as the worst DL-based model in terms of misclassification.

Interestingly, the results also showed that some classical ML algorithms in some FDD cases deserve more attention, especially if the collected bursts are long enough. For instance, despite providing lackluster performances with small bursts on the three

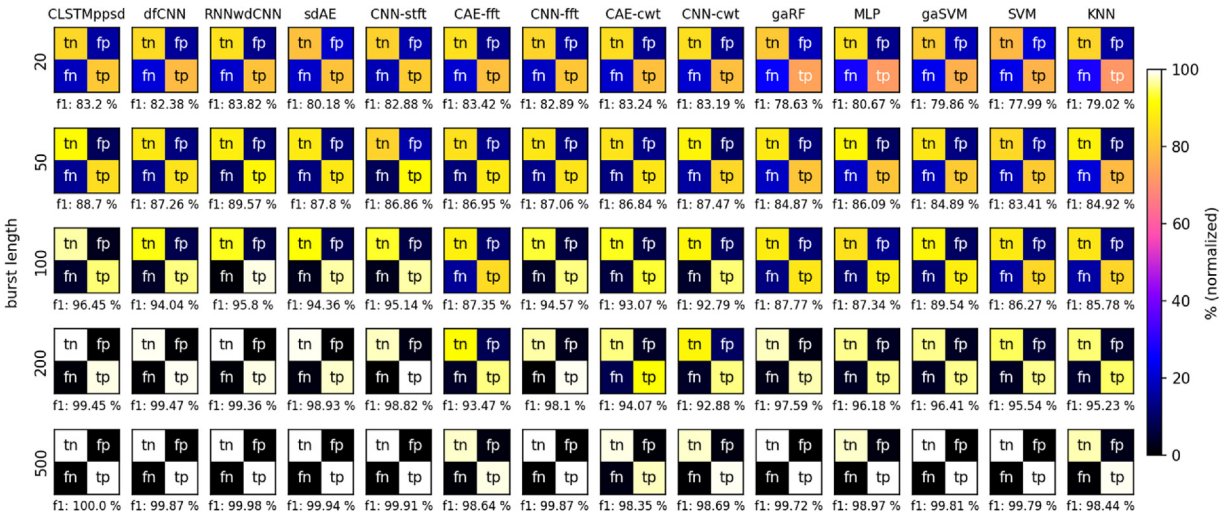

Fig. 15. Normalized confusion matrices for Wind Turbine.

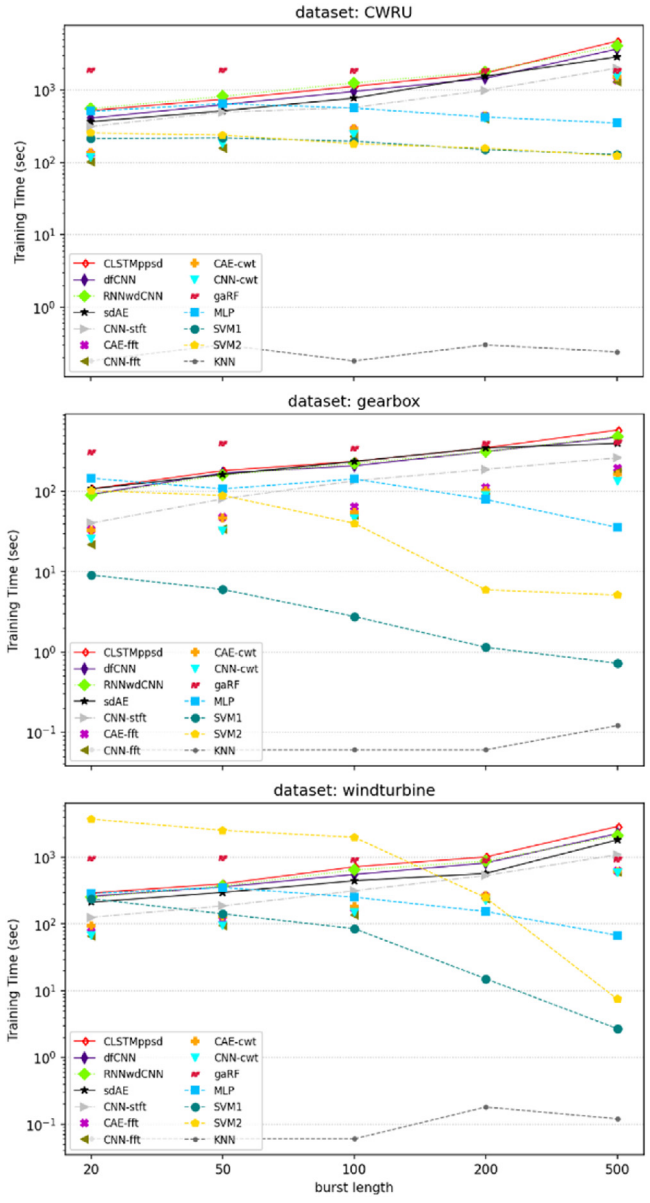

Fig. 16. Training times on different datasets.

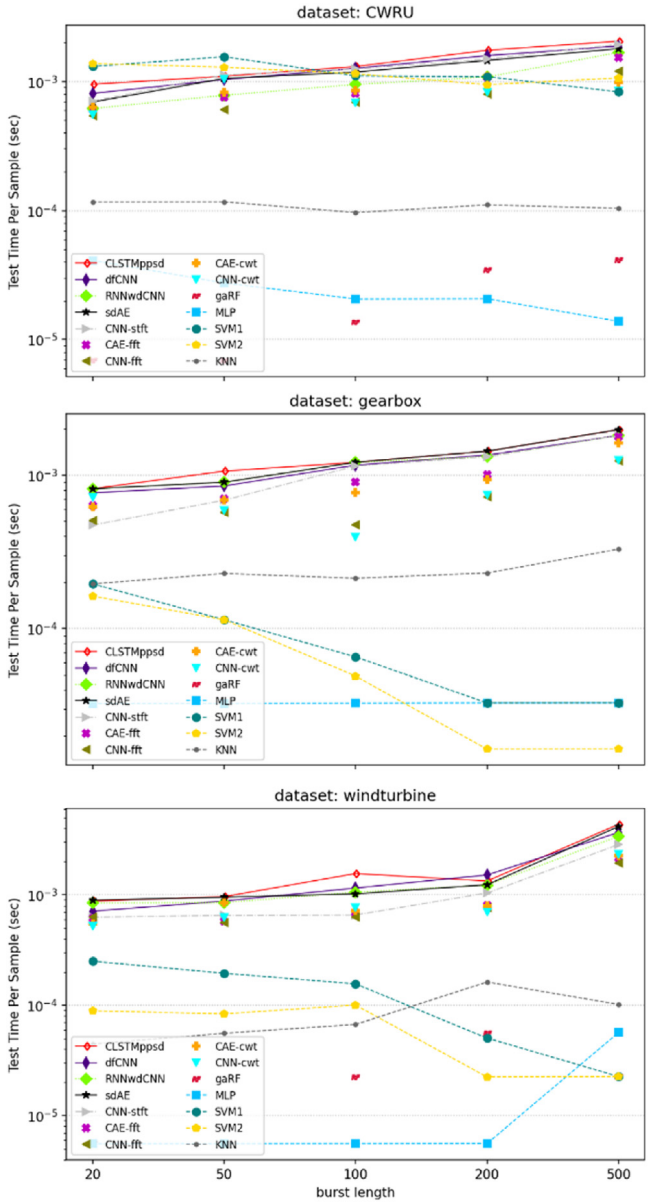

Fig. 17. Test times per sample on different datasets.

datasets, ANN, SVM and kNN produced convincing results on the Gearbox and the Wind Turbine datasets by achieving f1-scores higher than 98 % with *length=500*. Although they did not beat DL-based methods, they required meaningfully less training times. This places them among the fastest algorithms when dealing with long bursts. Compared to other ML-based models both gaRF and gaSVM models showed longer training times due to their genetic algorithm runtime bottleneck. In terms of test time, however, gaRF was faster than DL-based algorithms when diagnosing the bursts of *length=500*. The results for both Gearbox and Wind Turbine datasets had a similar trend: the longer the bursts were, the faster the ML algorithms got and the slower the DL-based approaches became.

The results described in this section and in Section 4.3.1 demonstrate how, at the cost of a slight increase in the training time, CLSTMppsd outperforms other DL algorithms, such as RNNwdCNN, sdAE, CAEfft and dfCNN, in terms of classification accuracy.

5. Discussion and conclusion

The paper presented a novel approach to cope with signal-based FDD for rotating machinery. For the feature engineering, the proposed model combines two different signal transform techniques, CWT and FFT, as well as some statistical features of the raw signal to extract all the fault signatures. The proposed framework uses a CLSTM architecture to handle the multi-channel input data and learns its spatio-temporal features more efficiently. Conducting a sensitivity analysis on the input channels, the paper shows that the combination of these multi-domain features intensifies the classifier's accuracy. On the other hand, by using three different DL architectures it is demonstrated that the use of the new CLSTM meaningfully helps the FDD system to understand the data structural characteristics and get higher accuracy. Compared to thirteen state-of-the-art and benchmark models, the proposed model achieved better classification performances on three different datasets. Besides the novel approach, the paper also illustrates a sensitivity analysis on the burst length parameter in order to highlight its prominent role in FDD. It also compares the results of different papers on the CWRU bearing dataset, to achieve a benchmark and portray the trend of classifiers and input types being proposed in the literature. To the best of our knowledge, such a comprehensive analysis has not been carried out previously in literature. This analysis proved that the proposed model needs shorter burst lengths to reach 100 % of accuracy on each dataset, meaning that in an online FDD application it can reach an error-free condition monitoring with shorter delay compared to other models (see Table 6).

In contrary to what related literature emphasize about the dominance of DL models over ML methods, the experimental evaluations performed in the study demonstrate that ML algorithms can still get promising performances in some situations, especially when the computational power is limited and a machinery needs rather fast algorithms to classify the emitting signals. However, the test time of all the compared models are relatively short, being less than 0.01 s per sample.

Future extensions of the present work will focus on the feature engineering for FDD on severely imbalance datasets, finding also an optimal hyper-parameter tuning to train the classifier faster without jeopardizing its accuracy. Also, it could be of attraction and importance to evaluate the feasibility and efficiency of applying the proposed FDD framework to Fault Prediction and Prognosis problems.

CRediT authorship contribution statement

Masoud Jalayer: Methodology, Conceptualization, Software, Data curation, Investigation, Visualization, Writing - original draft. **Carlotta Orsenigo:** Validation, Writing - review & editing, Supervision. **Carlo Vercellis:** Supervision.

Declaration of Competing Interest

The authors report no declarations of interest.

References

Al-bugharbee, H., Trenda, I., 2016]. A fault diagnosis methodology for rolling element bearings based on advanced signal pretreatment and autoregressive modelling. J. Sound Vib. 369, 246–265, http://dx.doi.org/10.1016/j.jsv.2015.12.052.

Balderston, H.L., 1969]. Incipient Failure Detection: Incipient Failure Detection in Ball Bearings.

Banerjee, T.P., Das, S., 2012]. Multi-sensor data fusion using support vector machine for motor fault detection. Inf. Sci. (N.Y.) 217, 96–107, http://dx.doi.org/10.1016/j.ins.2012.06.016.

Bengio, Y., 2009. Learning Deep Architectures for AI., http://dx.doi.org/10.1561/2200000006.

Cerrada, M., Zurita, G., Cabrera, D., Sánchez, R.V., Artés, M., Li, C., 2016]. Fault diagnosis in spur gears based on genetic algorithm and random forest. Mech. Syst. Signal Process. 70–71, 87–103, http://dx.doi.org/10.1016/j.ymssp.2015.08.030.

Chen, Z., Gryllias, K., Li, W., 2019]. Mechanical fault diagnosis using convolutional neural networks and extreme learning machine. Mech. Syst. Signal Process. 133, 106272, http://dx.doi.org/10.1016/j.ymssp.2019.106272.

Džakmić, Š., Namas, T., Husagić-Selman, A., 2018]. Combined fourier transform and mexican hat wavelet for fault detection in distribution networks. 2017 9th IEEE-GCC Conf. Exhib. GCCCE 2017, http://dx.doi.org/10.1109/IEEEGCC.2017.8447905.

Fengqi, W., Meng, G., 2006]. Compound rub malfunctions feature extraction based on full-spectrum cascade analysis and SVM. Mech. Syst. Signal Process. 20, 2007–2021, http://dx.doi.org/10.1016/j.ymssp.2005.10.004.

Gan, M., Wang, C., Zhu, C., 2016. Construction of hierarchical diagnosis network based on deep learning and its application in the fault pattern recognition of rolling element bearings. Mech. Syst. Signal Process. 72–73, 92–104, http://dx.doi.org/10.1016/j.ymssp.2015.11.014.

Gers, F.A., Schraudolph, N.N., 2002]. Learning precise timing with LSTM recurrent networks. J. Mach. Learn. Res. 3, 115–143.

Guo, W., Zhou, Z., Chen, C., Li, X., 2017]. Multi-frequency weak signal detection based on multi-segment cascaded stochastic resonance for rolling bearings. Microelectron. Reliab. 75, 239–252, http://dx.doi.org/10.1016/j.microrel.2017.03.018.

Haidong, S., Hongkai, J., Huiwei, Z., Fuan, W., 2017. A novel deep autoencoder feature learning method for rotating machinery fault diagnosis. Mech. Syst. Signal Process. 95, 187–204, http://dx.doi.org/10.1016/j.ymssp.2017.03.034.

Hochreiter, S., Schmidhuber, J., 1997]. Long short-term memory. Neural Comput. 9, 1735–1780.

Huang, W., Song, G., Hong, H., Xie, K., 2014]. Deep architecture for traffic flow prediction: deep belief networks with multitask learning. IEEE Trans. Intell. Transp. Syst. 15, 2191–2201, http://dx.doi.org/10.1109/TITS.2014.2311123.

Jana, S., Abhinandan, D., 2017]. A novel zone division approach for Power system fault detection using ANN-based pattern recognition technique détection de défaut des réseaux électriques utilisant la technique de reconnaissance de formes basée sur RNA. Can. J. Electr. Comput. Eng. 40, 275–283.

Jia, F., Lei, Y., Lin, J., Zhou, X., Lu, N., 2016]. Deep neural networks: a promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. Mech. Syst. Signal Process. 72–73, 303–315, http://dx.doi.org/10.1016/j.ymssp.2015.10.025.

Jia, F., Lei, Y., Guo, L., Lin, J., Xing, S., 2018a]. A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines. Neurocomputing 272, 619–628, http://dx.doi.org/10.1016/j.neucom.2017.07.032.

Jia, F., Lei, Y., Lu, N., Xing, S., 2018b. Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization. Mech. Syst. Signal Process. 110, 349–367, http://dx.doi.org/10.1016/j.ymssp.2018.03.025.

Lecun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D., 1990]. Handwritten digit recognition with a back-propagation network. In: Adv. Neural Inf. Process. Syst. Morgan Kaufmann, Denver, CO, pp. 396–404 http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.32.5076%5Cnhttp://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.32.5076&rep=rep1&type=pdf%0Ahttp://www.ncbi.nlm.nih.gov/pubmed/23301817.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444, http://dx.doi.org/10.1038/nature14539.

Lee, T., Kim, C.O., 2015]. Statistical comparison of fault detection models for semiconductor manufacturing processes. IEEE Trans. Semicond. Manuf. 28, 80–91, http://dx.doi.org/10.1109/TSM.2014.2378796.

Lee, H., Grosse, R., Ranganath, R., Ng, A.Y., 2011]. Unsupervised learning of hierarchical representations with convolutional deep belief networks. Commun. ACM 54, 95, http://dx.doi.org/10.1145/2001269.2001295.
Lee, H., Kim, Y., Kim, C.O., 2017a. A deep learning model for robust wafer fault monitoring with sensor measurement noise. IEEE Trans. Semicond. Manuf. 30, 23–31, http://dx.doi.org/10.1109/TSM.2016.2628865.
Lee, K.B., Cheon, S., Kim, C.O., 2017b. A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes. IEEE Trans. Semicond. Manuf. 30, 135–142, http://dx.doi.org/10.1109/TSM.2017.2676245.
Lei, Y., Jia, F., Lin, J., Xing, S., Ding, S.X., 2016]. An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data. IEEE Trans. Ind. Electron. 63, 3137–3147, http://dx.doi.org/10.1109/TIE.2016.2519325.
Lei, J., Liu, C., Jiang, D., 2019]. Fault diagnosis of wind turbine based on Long Short-term memory networks. Renew. Energy 133, 422–432, http://dx.doi.org/10.1016/j.renene.2018.10.031.
Li, C.J., Ma, J., 1997]. Wavelet decomposition of vibrations for detection of bearing-localized defects. NDT E Int. 30, 143–149, http://dx.doi.org/10.1016/S0963-8695(96)00052-7.
Li, Ke, Wang, Quanxin, 2015. Study on signal recognition and diagnosis for spacecraft based on deep learning method. In: 2015 Progn. Syst. Heal. Manag. Conf., IEEE, pp. 1–5, http://dx.doi.org/10.1109/PHM.2015.7380040.
Li, H., Wang, C., Chen, C., Yan, G., 2011]. Review of vibration signals trend forecasting methods. Procedia Environ. Sci. 10, 837–842, http://dx.doi.org/10.1016/j.proenv.2011.09.135.
Li, P., Kong, F., He, Q., Liu, Y., 2013]. Multiscale slope feature extraction for rotating machinery fault diagnosis using wavelet analysis. Measurement 46, 497–505, http://dx.doi.org/10.1016/j.measurement.2012.08.007.
Liang, K., Qin, N., Huang, D., Fu, Y., 2018]. Convolutional recurrent neural network for fault diagnosis of high-speed train bogie. Complexity 2018, 13.
Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E., 2017. A survey of deep neural network architectures and their applications. Neurocomputing 234, 11–26, http://dx.doi.org/10.1016/j.neucom.2016.12.038.
Liu, G., Bao, H., Han, B., 2018]. A stacked autoencoder-based deep neural network for achieving gearbox fault diagnosis. Math. Probl. Eng. 2018, 1–10, http://dx.doi.org/10.1155/2018/5105709.
Lu, S., He, Q., Yuan, T., Kong, F., 2016a]. Online fault diagnosis of motor bearing via stochastic-resonance-Based adaptive filter in an embedded system. IEEE Trans. Syst. Man Cybern. Syst. 47, 1111–1122, http://dx.doi.org/10.1109/TSMC.2016.2531692.
Lu, S., Wang, X., He, Q., Liu, F., Liu, Y., 2016b. Fault diagnosis of motor bearing with speed fluctuation via angular resampling of transient sound signals. J. Sound Vib. 385, 16–32, http://dx.doi.org/10.1016/j.jsv.2016.09.012.
Lu, C., Wang, Z.Y., Qin, W.L., Ma, J., 2017]. Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification. Signal Process. 130, 377–388, http://dx.doi.org/10.1016/j.sigpro.2016.07.028.
Mao, W., Feng, W., Liu, Y., Zhang, D., Liang, X., 2021]. A new deep auto-encoder method with fusing discriminant information for bearing fault diagnosis. Mech. Syst. Signal Process. 150, 107233, http://dx.doi.org/10.1016/j.ymssp.2020.107233.
Misra, M., Yue, H.H., Qin, S.J., Ling, C., 2002. Multivariate process monitoring and fault diagnosis by multi-scale PCA. Comput. Chem. Eng. 26, 1281–1293, http://dx.doi.org/10.1016/S0098-1354(02)00093-5.
Muruganatham, B., Sanjith, M.A., Krishnakumar, B., Murty, S.A.V.S., 2013. Roller element bearing fault diagnosis using singular spectrum analysis. Mech. Syst. Signal Process. 35, 150–166, http://dx.doi.org/10.1016/j.ymssp.2012.08.019.
Nakazawa, T., Kulkarni, D.V., 2018]. Wafer map defect pattern classification and image retrieval using convolutional neural network. IEEE Trans. Semicond. Manuf. 6507, http://dx.doi.org/10.1109/TSM.2018.2795466.
Netsanet, S., Zhang, J., Zheng, D., 2018]. Bagged decision trees based scheme of microgrid protection using windowed fast Fourier and wavelet transforms. Electron 7, http://dx.doi.org/10.3390/electronics7050061.
Park, P., Di Marco, P., Shin, H., Bang, J., 2019]. Fault detection and diagnosis using combined autoencoder and long short-term memory network. Sensors 19, 1–17, http://dx.doi.org/10.3390/s19214612.
Qian, P., Tian, X., Kanfoud, J., Lee, J.L.Y., Gan, T.-H., 2019]. A novel condition monitoring method of wind turbines based on long short-term memory neural network. Energies 12, 1–15.
Rai, V.K., Mohanty, A.R., 2007]. Bearing fault diagnosis using FFT of intrinsic mode functions in Hilbert-Huang transform. Mech. Syst. Signal Process. 21, 2607–2615, http://dx.doi.org/10.1016/j.ymssp.2006.12.004.
Recioui, A., Benseghier, B., Khalfallah, H., 2015]. Power system fault detection, classification and location using the K-nearest neighbors. In: 2015 4th Int. Conf. Electr. Eng., IEEE, pp. 1–6, http://dx.doi.org/10.1109/INTEE.2015.7416832.
Sabir, R., Rosato, D., Hartmann, S., Guehmann, C., 2019. LSTM based bearing fault diagnosis of electrical machines using motor current signal. Proc. - 18th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2019, 613–618, http://dx.doi.org/10.1109/ICMLA.2019.00113.
Seo, Y., Choi, Y., Choi, J., 2017. River stage modeling by combining maximal overlap discrete wavelet transform, support vector machines and genetic algorithm. Water (Switzerland). 9, http://dx.doi.org/10.3390/w9070525.
Shao, H., Jiang, H., Zhang, H., Duan, W., Liang, T., Wu, S., 2018. Rolling bearing fault feature learning using improved convolutional deep belief network with compressed sensing. Mech. Syst. Signal Process. 100, 743–765, http://dx.doi.org/10.1016/j.ymssp.2017.08.002.
Shen, C., Qi, Y., Wang, J., Cai, G., Zhu, Z., 2018. An automatic and robust features learning method for rotating machinery fault diagnosis based on contractive autoencoder. Eng. Appl. Artif. Intell. 76, 170–184, http://dx.doi.org/10.1016/j.engappai.2018.09.010.
Shenfield, A., Howarth, M., 2020]. A novel deep learning model for the detection and identification of rolling element-bearing faults. Sensors (Switz.) 20, 1–24, http://dx.doi.org/10.3390/s20185112.
Shi, X., Chen, Z., Wang, H., 2015]. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. NIPS'15 Proc. 28th Int. Conf. Neural Inf. Process. Syst., 1–12.
Silva, A., Zarzo, A., Machuca González, J.M., Munoz-Guijosa, J.M., 2020. Early fault detection of single-point rub in gas turbines with accelerometers on the casing based on continuous wavelet transform. J. Sound Vib. 487, 115628, http://dx.doi.org/10.1016/j.jsv.2020.115628.
Tan, W., Sun, Y., Qiu, D., An, Y., Ren, P., 2020]. Rolling bearing fault diagnosis based on single gated unite recurrent neural networks. J. Phys. Conf. Ser. 1601, 042017, http://dx.doi.org/10.1088/1742-6596/1601/4/042017.
Unal, M., Onat, M., Demetgul, M., Kucuk, H., 2014]. Fault diagnosis of rolling bearings using a genetic algorithm optimized neural network. Measurement, http://dx.doi.org/10.1016/j.measurement.2014.08.041.
Wang, Y., Kang, S., Jiang, Y., Yang, G., Song, L., Mikulovich, V.I., 2012]. Classification of fault location and the degree of performance degradation of a rolling bearing based on an improved hyper-sphere-structured multi-class support vector machine. Mech. Syst. Signal Process. 29, 404–414, http://dx.doi.org/10.1016/j.ymssp.2011.11.015.
Wang, X., Zheng, Y., Zhao, Z., Wang, J., 2015]. Bearing fault diagnosis based on statistical locally linear embedding. Sensors 15, 16225–16247, http://dx.doi.org/10.3390/s150716225.
Wang, S., Xiang, J., Zhong, Y., Zhou, Y., 2018]. Convolutional neural network-based hidden Markov models for rolling element bearing fault identification. Knowl.-Based Syst. 144, 65–76, http://dx.doi.org/10.1016/j.knosys.2017.12.027.
Weichbrodt, B., Smith, K.A., 1970]. Signature analaysis. Non-intrusive techniques for incipient failure identification. In: NBS Spec Publ 336, Proc 5th Sp. Simul. Symp Conf, Elsevier B.V., Gaithersburg, MD, USA.
Wen, L., Li, X., Gao, L., Zhang, Y., 2017]. A new convolutional neural network based data-driven fault diagnosis method. IEEE Trans. Ind. Electron. 65, http://dx.doi.org/10.1109/TIE.2017.2774777, 1–1.
Widodo, A., Yang, B.-S., 2007]. Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors. Expert Syst. Appl. 33, 241–250, http://dx.doi.org/10.1016/j.eswa.2006.04.020.
Yan, X., Jia, M., 2018. A novel optimized SVM classification algorithm with multi-domain feature and its application to fault diagnosis of rolling bearing. Neurocomputing 313, 47–64, http://dx.doi.org/10.1016/j.neucom.2018.05.002.
Yang, H., Lin, H., Ding, K., 2018]. Sliding window denoising K-singular value decomposition and its application on rolling bearing impact fault diagnosis. J. Sound Vib. 421, 205–219, http://dx.doi.org/10.1016/j.jsv.2018.01.051.
Yao, Q., Wang, J., Yang, L., Su, H., Zhang, G., 2016]. A fault diagnosis method of engine rotor based on random forests. In: 2016 IEEE Int. Conf. Progn. Heal. Manag., IEEE, pp. 1–4, http://dx.doi.org/10.1109/ICPHM.2016.7542838.
Zappala, D., Sarma, N., Djurovi, S., Crabtree, C.J., Mohammad, A., Tavner, P.J., 2019]. Electrical & mechanical diagnostic indicators of wind turbine induction generator rotor faults. Renew. Energy 131, 14–24, http://dx.doi.org/10.1016/j.renene.2018.06.098.
Zhang, X., Wang, B., Chen, X., 2015]. Intelligent fault diagnosis of roller bearings with multivariable ensemble-based incremental support vector machine. Knowl.-Based Syst. 89, 56–85, http://dx.doi.org/10.1016/j.knosys.2015.06.017.
Zhang, T., Wang, W., Ye, H., Huang, D., Zhang, H., Li, M., 2016. Fault detection for ironmaking process based on stacked denoising autoencoders. Proc. Am. Control Conf. 2016-July, 3261–3267, http://dx.doi.org/10.1109/ACC.2016.7525420.
Zhang, W., Peng, G., Li, C., Chen, Y., Zhang, Z., 2017]. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. Sensors (Switz.) 17, http://dx.doi.org/10.3390/s17020425.
Zhang, W., Li, C., Peng, G., Chen, Y., Zhang, Z., 2018]. A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. Mech. Syst. Signal Process. 100, 439–453, http://dx.doi.org/10.1016/j.ymssp.2017.06.022.
Zhang, Y., Xing, K., Bai, R., Sun, D., Meng, Z., 2020. An enhanced convolutional neural network for bearing fault diagnosis based on time–frequency image. Meas. J. Int. Meas. Confed. 157, 107667, http://dx.doi.org/10.1016/j.measurement.2020.107667.
Zhang, Z., Zhao, J., 2017]. A deep belief network based fault diagnosis model for complex chemical processes. Comput. Chem. Eng. 107, 395–407, http://dx.doi.org/10.1016/j.compchemeng.2017.02.041.
Zhao, G., Zhang, G., Ge, Q., Liu, X., 2016]. Research advances in fault diagnosis and prognostic based on deep learning. Progn. Syst. Heal. Manag. Conf., 1–6, http://dx.doi.org/10.1109/PHM.2016.7819786.
Zhao, R., Yan, R., Wang, J., Mao, K., 2017]. Learning to monitor machine health with convolutional Bi-Directional LSTM networks. Sensors 17, 1–18, http://dx.doi.org/10.3390/s17020273.
Zheng, J., Pan, H., Cheng, J., 2017. Rolling bearing fault detection and diagnosis based on composite multiscale fuzzy entropy and ensemble support vector machines. Mech. Syst. Signal Process. 85, 746–759, http://dx.doi.org/10.1016/j.ymssp.2016.09.010.