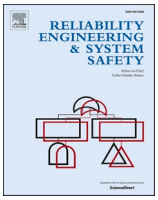
Reliability Engineering and System Safety 233 (2023) 109107

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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



Failure diagnosis of a compressor subjected to surge events: A data-driven framework

Leonardo Leoni [a], Filippo De Carlo [a,*], Mohammad Mahdi Abaei [b], Ahmad BahooToroody [c], Mario Tucci [a]

[a] Department of Industrial Engineering (DIEF), University of Florence, Florence, Italy
[b] Department of Geography and Geology, University of Turku, Turku, Finland
[c] Marine and Arctic Technology Group, Department of Mechanical Engineering, Aalto University, Finland

ARTICLE INFO

Keywords:
Condition monitoring
Failure diagnosis
Empirical mode decomposition
Neighborhood component analysis
Supervised classification

ABSTRACT

Due to higher reliability and safety requirements, the importance of condition monitoring and failure diagnosis has progressively cleared up. In this context, being able to properly deal with noise and data reduction is fundamental for improving failure diagnosis and assuring safe operations. These tasks are particularly difficult in presence of many non-stationary and non-linear signals. Accordingly, this paper proposes a failure diagnosis methodology that integrates Empirical Mode Decomposition (EMD) and Neighborhood Component Analysis (NCA) for noise removal and data reduction. While noise detection and reduction techniques are established to reduce the uncertainties integrated with data acquisition, traditional approaches that cannot capture the non-stationary and non-linear nature of data might result in higher uncertainty. As a validated denoising method, EMD is applied to cope with the previous limitations. The NCA overcomes typical limitations such as imposing class distributions. After data pre-processing, the diagnosis is performed through a Random Forest. The methodology is tested on real data of a compressor subjected to surge, showing an accuracy higher than 97%. Moreover, the surge accuracy is close to 95%, while the regime accuracy is higher than 97%. The developed framework could assist practitioners in evaluating the condition of assets and, accordingly, planning maintenance.

1. Introduction

During recent decades, Condition Monitoring (CM) and the related failure diagnosis have seen widespread adoption in many engineering fields such as wind turbines [1,2], induction motors [3,4], and railways [5,6]. This trend is related to the relevance of CM, which allows the early detection of industrial equipment failures [7]. This feature is aligned with the safety and reliability requirements, that are becoming more stringent for process industries [8]. CM, continuous or periodic [9], could be defined as monitoring the working condition of a given system to evaluate its health status and, accordingly, define maintenance tasks [10]. CM approaches could be divided into three main phases, respectively known as data acquisition, data preprocessing, and data processing. During the first stage, data related to relevant Process Variables (PVs) are acquired. The data preprocessing stage consists of noise reduction and feature selection. Finally, data processing aims at analyzing data with appropriate tools that enable diagnosis or even prognosis. Adopting a proper CM approach is pivotal to assure that the monitored equipment could fulfill its mission while guaranteeing the safety of the operations. Indeed, the health state of a machine is strongly related to reliable and safe operations, thus being able to determine its operating condition with a high degree of confidence could be helpful to intervene whenever the operations are considered unacceptable from a safety perspective. To this end, noise removal and data reduction are of prominent importance to improve the accuracy and reduce the calculation time of the subsequent data processing, especially if a component is monitored by a high number of non-stationary and dynamic PVs. As a result, a CM framework must include proper noise removal and data reduction techniques to accurately evaluate the health of a system and perform failure diagnosis.

Despite the advances in sensors and related technologies, most actual signals contain noise, defined as an undesired component that alters the true signal. Accordingly, to obtain a better understanding of the true signal, noise should be detected and removed. In this sense, the main

* Correspondant author.
E-mail address: filippo.decarlo@unifi.it (F. De Carlo).

https://doi.org/10.1016/j.ress.2023.109107
Received 25 May 2022; Received in revised form 13 October 2022; Accepted 19 January 2023
Available online 21 January 2023

objective of denoising approaches is to extract the noise while preserving all relevant information hidden within the signal itself [11]. Signal denoising techniques could be classified based on the working domain, either time, frequency, or time frequency. Among the time-domain techniques, there are the filter-based methods, which exploit appropriate filters to extract the noise from the acquired signal [12]. Although filter-based methodologies are easy to implement, they present two significant drawbacks [13]: (i) they require prior knowledge of the spectrum, and (ii) the signal must be stationary. On the other side, frequency-domain techniques are more suited compared to time-domain approaches to deal with fault detection since several machines are characterized by different frequencies in the normal and faulty states [14]. Within frequency domain methodologies, the Fast Fourier Transform (FFT) has attracted significant attention for CM, fault detection, and failure diagnosis purposes [15–18]. Despite their low computational complexity, frequency domain techniques have a significant limitation related to the very dynamic nature of noise [19], making them unable to deal with non-stationary signals. To overcome this problem, time-frequency analyses, such as Short-Time Fourier Transform (STFT) and Wavelet Transform (WT), are adopted [14]. As a result, there is an ongoing effort on STFT and WT within signal denoising, health condition assessment, and CM applications [20–24].

It is worth mentioning that non-stationarity means that the sine wave, frequency, and spectrum content of a signal are changing over time. STFT and WT can face non-stationarity signals; however, STFT is only employable under linear conditions of the acquired data [25], while WT is usable only under local nonlinearity. Furthermore, WT requires the specification of a basis function, which could be a challenging task, while the STFT needs piecewise stationarity whose scale is equal to the length of the adopted sliding window [13]. To overcome the aforementioned limitations, Huang et al. [26] developed the Empirical Mode Decomposition (EMD), which is very suitable for dealing with the non-stationarity and nonlinearity of time series, therefore, it is adopted in different fields such as bearing fault diagnosis [27]. Also, EMD does not need the indication of a basic function such as most WTs [28]. Due to its advantages, EMD and its derivative approaches have become popular tools to perform CM and failure diagnoses [29–34]. A recent study by Yan et al. [35] proposed a methodology to predict the temperature of a train axle. Specifically, the authors employed Complementary EMD to decompose the signal into a set of Intrinsic Mode Functions (IMFs), which were fed to a Long Short-Term Memory Neural Network (LSTMNN), tasked with the prediction. Then, they adopted a Particle Swarm Optimization and Gravitational Search Algorithm (PSOGSA) to improve the forecast accuracy. Another recent work by Gao et al. [36] presented a methodology to predict bearing failure. The authors exploited Ensemble EMD to decompose the signal into its IMFs, and subsequently, they retained only the most relevant ones. Next, the most informative IMFs were inserted as input in an LSTMNN to learn the failure behavior. In a very recent work by Li et al. [37], a new method called synthetical modal parameters identification is proposed to control the oscillation of power system over time. The method integrates EMD for noise removal with Prony and stochastic subspace identification for modal parameters identification, and it could be useful to avoid oscillations, showing health management capabilities.

CM applications could be characterized by several data sources, leading to large datasets. Although having a lot of data could generate better results; a greater amount of data will result in a higher impact of the curse of dimensionality [38]. Consequently, selecting a subset of relevant features or PVs is crucial to improve the subsequent calculation steps. Several techniques have been adopted to deal with data reduction problems, among which Principal Component Analysis (PCA) [39], Linear Discriminant Analysis (LDA) [40], Sequential Feature Selection (SFS) [41], and entropy-based feature extraction [42] are worth mentioning. For instance, the framework presented by Tian et al. [43] proposes an entropy-based method to extract entropies from an acoustic signal and subsequently perform failure diagnosis through a Random Forest (RF). However, the techniques mentioned above present critical drawbacks. In fact, both PCA and entropy-based feature extraction transform the original features into a new space, losing track of the original features, while LDA performs optimally when data are normally distributed. Finally, SFS techniques are unable to either determine whether a feature has become useless when a new feature is added or if a feature is valid after it has been discarded [44]. Meanwhile, Neighborhood Component Analysis (NCA) as a linear nonparametric feature selection approach has been introduced by Goldberger et al. [45], overcoming the limitations related to imposing a class distribution or decision boundaries. Moreover, NCA does not lose any information within the data reduction process [46], preserving the original features or PVs. Thanks to its advantages, NCA has been successfully applied within CM, failure diagnosis, and fault detection frameworks [47–49]. Yaman [47] used NCA for extracting the most relevant features, which are subsequently fed to classification techniques for performing the diagnosis of an induction motor. A similar work has been proposed by Zhou et al. [48], who presented a methodology to evaluate bearing failure through the integration of NCA and Couple Hidden Markov Model (CHMM).

After data reduction and denoising, a CM process requires data processing, which analyzes the obtained data to determine the health state of the monitored system. This last step allows for detecting possible anomalies or abnormal states, and subsequently, making decisions to restore safe and reliable conditions. Within this context, there is a fundamental distinction between classification and regression. The first identifies the state of the asset and is characterized by a categorical response variable, while the second aims to predict the evolution of a given response variable (e.g., a safety or reliability indicator), which is real-valued [50]. In a CM or failure diagnosis problem, Machine Learning (ML) and related techniques such as Deep Learning (DL) are among the most common approaches. Examples of ML algorithms used for this purpose are Support Vector Machine (SVM) [51], Neural Network (NN) [52], Decision Tree (DT) [53], and RF [54]. Due to the relevance of the topic, there is an ongoing effort on ML-based or DL-based CM, failure diagnosis, anomaly detection, and Remaining Useful Life (RUL) prediction frameworks [55–58]. A relevant example is a work presented by Zhu et al. [59], who exploited at first t-SNE-DBSCAN to reduce the dimension of data and, in particular, aggregate the data coming from different sensors and extract a health indicator. Finally, they employed an LSTMNN to predict the RUL. In another recent study by Xu et al. [60], the authors proposed an advanced methodology to predict the life cycle of lithium-ion batteries. In their work, clustering by fast search is first exploited for feature selection and, subsequently, they adopted a stacked denoising autoencoder for prediction purposes.

Despite all the ongoing efforts, there is still space to develop a methodology capable of determining in real-time the health of a system characterized by highly fluctuating PVs, allowing to identify dangerous operations and determine the actions required to repristinate safety conditions. To this end, this paper aims to present a novel failure diagnosis methodology based on the integration of EMD and NCA. EMD is adopted for its capability of dealing with nonlinear and non-stationary signals. The noisy IMFs are detected through Statistical Significance Test (SST). On the other side, NCA is exploited for its ability to preserve information. Finally, the denoised most relevant signals are fed to a RF to classify the state of the system. To demonstrate the applicability of the methodology, a compressor operating in a geothermal plant is chosen as a case study. The developed framework could assist asset managers in performing diagnosis for equipment characterized by many highly dynamic PVs, determining the operating condition of the system without the need to extract a new set of features from the acquired signals. Moreover, this task is conducted without requiring external opinions, allowing the detection of undesired states and accordingly adopting proper countermeasures. To the best of the authors' knowledge, up to now, EMD and NCA were used to determine the most relevant features of a signal rather than identifying the most relevant PVs that affect the

health of a given system. Furthermore, it is worth mentioning that many works on failure diagnosis are related to experimental or laboratory studies. Finally, few PVs are usually considered (generally one or two and mainly vibration-related measurements), and subsequently processed through feature extraction. By contrary, the proposed approach is applied to a real case study, considering many PVs monitored through different sensors. Within this context, the method directly acts on the measured PVs. Accordingly, it could be possible to determine the most important PVs and the associated sensors in addition to the condition of the equipment. This is of prominent importance to shed light on which PVs and sensors the condition monitoring should be focused. For instance, the sensors that are related to the detected most relevant PVs should be very reliable in terms of measurement accuracy and failure behavior. Indeed, the malfunctioning of these sensors could lead to failure diagnosis inaccuracy or impossibility.

The remainder of this paper is organized as follows; Section 2 introduces the material and methods, which are EMD, NCA, and RF. EMD and NCA are used for signal denoising and PV selection respectively, while the RF is adopted for the processing phase. Section 3 describes the developed framework, with a specific focus on the steps and tools required to perform the diagnosis, along with providing information on the inputs and outputs of each step. Section 4 illustrates the application of the novel approach to a case study, which is a centrifugal compressor characterised by 27 monitored PVs, among which some can be regarded as non-stationary and non-linear. In Section 5, the results are discussed, with an emphasis on the accuracy obtained through the approach. Finally, in Section 6, the conclusions, limitations, and possible future developments are presented.

2. Materials and methods

2.1. Empirical mode decomposition

Data acquired from sensors are characterized by two main parts, usually denoted as true signal and noise. The last one is a disturbing component that must be identified and removed during the preprocessing phase to improve the succeeding analysis. The EMD is a data-driven filtering approach whose introduction is based on the Hilbert-Huang transform [26]. The EMD decomposes the acquired signal into a series of components named IMFs and a residual term [61], as shown by Eq. (1).

$$x(t) = \sum_{i=1}^{n} c_i(t) + r(t) \tag{1}$$

where n is the number of IMFs, while $c_i(t)$ is the i th IMF. Finally, $r(t)$ is the residual term.

An IMF could either belong to the noise component or the true signal component, therefore the IMFs which determine the true signal are distinguished from the IMFs related to the random noise, as illustrated by Eq. (2) [29]:

$$x(t) = \sum_{i=1}^{n} c_{i,TS}(t) + \sum_{i=1}^{m} c_{i,N}(t) + r(t) \tag{2}$$

where $c_{i,TS}(t)$ and $c_{i,N}(t)$ identify a true signal IMF and a noise IMF respectively, while $r(t)$ denotes the residual term.

2.2. Neighborhood component analysis

Feature selection reduces the starting set of features by discarding the irrelevant or redundant ones, leading to an increase in accuracy, comprehensibility, and execution speed [62]. The NCA was introduced by Goldberger et al. [45], considering as a reference the well-known K-Nearest Neighbors (KNN) algorithm. NCA is a nonparametric feature selection approach whose objective is to find the weight denoting the importance of every feature [63]. This task is accomplished through the maximization of the leave-one-out classification accuracy. A detailed description of the NCA can be found in Goldberger et al. [45], Yang et al. [63], and Raghu and Sriraam [46].

2.3. Random forest

Several algorithms and techniques could be adopted for classification purposes. A RF is a well-known ML approach based on DT. Specifically, a RF is an ensemble classifier that combines a set of DTs through a bagging process [64]. Specifically, each DT is obtained by drawing with replacement a random sample from the original dataset, meaning that some observations can be considered more than once, while others could not be considered at all [65]. Also, each DT could consider different sets of predictors. Two relevant user-selected parameters are the number of DTs and the number of splits for each DT. Each DT assigns a class to an observation, for both the training and the test phase. During the test phase, the predicted class is the one that has been determined by most of the DTs. The RF is known for its easiness of implementation, explainability, and reliability in classification [66]. Furthermore, the joining of multiple individual classifiers, such as the RF, improves performance [67]. Finally, in some studies, RF resulted to be more accurate than other ML approaches [68,69]. The previous considerations led to the adoption of the RF. However, it is worth mentioning that trying different ML approaches could be useful to determine the most accurate one.

3. Methodology

The structure of the proposed methodology is illustrated in Fig. 1. Given a step represented by a rectangle, the tool used to conduct the step is reported on the right side, while the output of the step is represented below the step's rectangle.

3.1. Stage 1: Data acquisition

The starting stage consists of acquiring the data required to perform the failure diagnosis. First, a set of PVs is selected, and the respective sensors are considered (Step 1). Accordingly, a list of PVs to monitor and consider for the analysis is defined. Then, data are extracted from the sensors during operations (Step 2), providing as output the variation of each PV for a given time window. Finally, the acquired signals are classified into different operating conditions (Step 3), identifying the good working and the faulty state for each time window.

3.2. Stage 2: Data preprocessing

The second stage is devoted to noise removal and data reduction. The signals acquired through the previous step are decomposed into a set of IMFs and a trend through the adoption of EMD (Step 4). Therefore, each signal will be associated with a set of IMFs, among which some would be considered as noise and the remaining would be regarded as true signal. Next, each IMF is processed through a SST (Step 5). This step defined which are the noisy IMFs, which are removed from the original signal (Step 6). To conclude this stage, the denoised signal is processed through NCA (Step 7). As an output, this step provides a ranking of the monitored PVs, pointing out the most relevant ones, which will be considered during the processing phase. Accordingly, the classification is performed by considering only a subset of the total number of monitored PVs.

3.3. Stage 3: Data processing

The final stage is required to develop a model to perform diagnosis based on the monitored parameters. First, the reduced and denoised set of signals is processed through an ML classification tool (step 8). Specifically, a RF is exploited for this stage. Finally, the trained RF arising from the previous step is tested on data discarded during the training

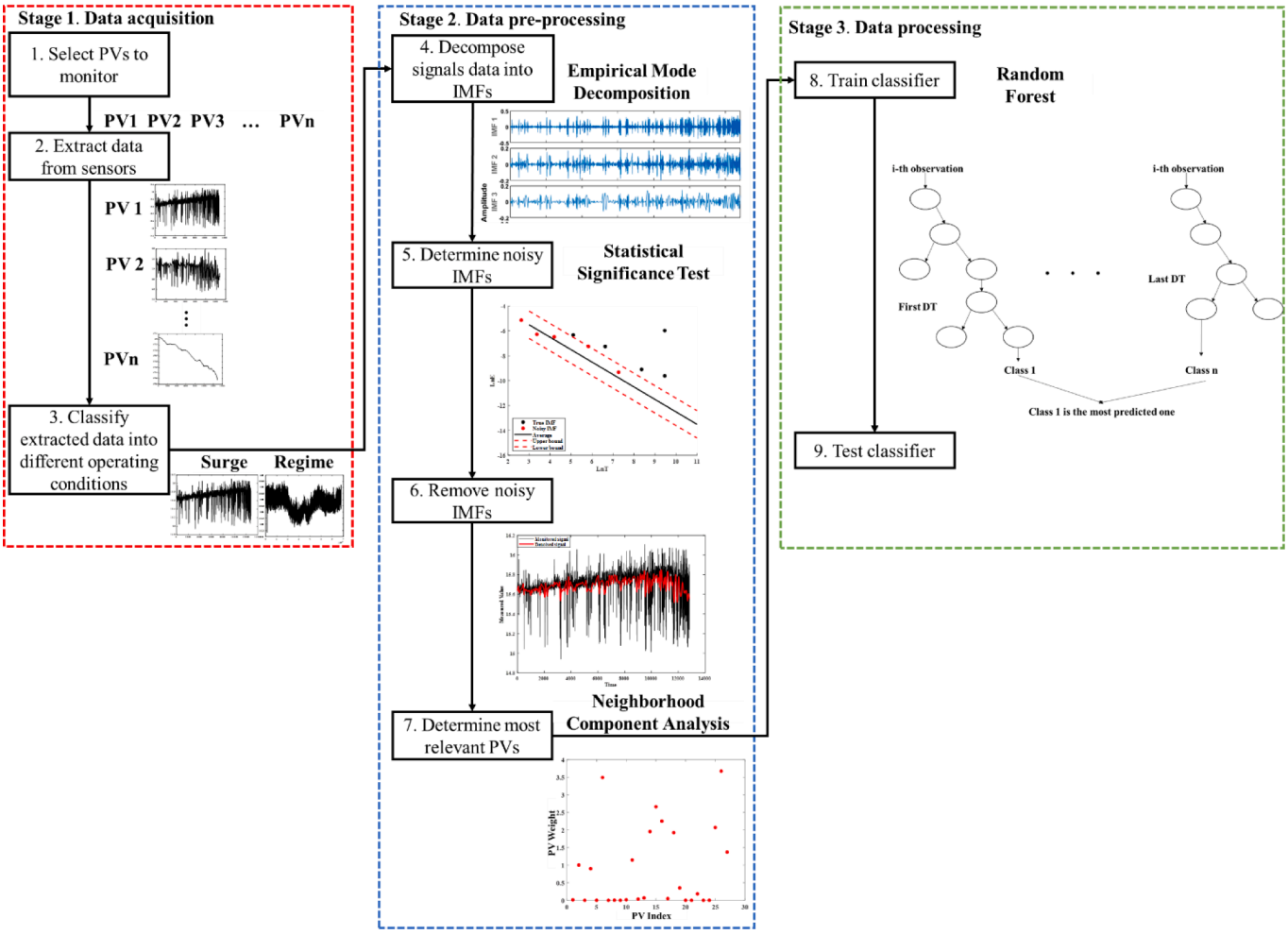


Fig. 1. Schematic representation of the steps required to perform the developed framework.

phase (step 9). Based on the results of the test phase, the generalization capability of the RF is evaluated. In case the test accuracy is evaluated as appropriate, the classification tool can be used to perform diagnosis.

4. Results: Application of the methodology

To demonstrate the applicability of the methodology, we considered a case study consisting in a compressor operating in a geothermal plant, capable of developing about 20 MW through a single flash condensation cycle. The system is a three-stage centrifugal compressor devoted to extracting non-condensable gasses. The first stage is characterized by two opposing impellers splined to the same shaft, while the second stage has a single impeller that is on the same shaft of the first stage. Finally, the last stage is composed of two opposing impellers splined to a different shaft compared to the first two stages. The compressor is linked to a turbine through a gear multiplier, which allows the compressor to rotate five times faster than the turbine. The mass flow of the system is between 10,000 kg/s and 22,000 kg/s, while the temperature and pressure of the gas flow at the outlet are 170 °C and 1.013 bar. A schematic representation of the considered system is shown in Fig. 2. It is worth mentioning that the compressor is deeply affected by surge phenomena, which undermine the compressor's performance. Furthermore, the surge could cause extensive damage to the machine if prolonged over time.

4.1. Stage 1: Data extraction and classification

Due to the importance of the plant, there are several sensors, each of which monitors a distinct PV. For this work, 27 different sensors (i.e., 27 PVs) monitoring the compressor operating condition are considered (Step 1) and listed in Table 1. The acquired signals are mostly continuous and analog, and they were sampled at a specific time interval. The initial selection of 27 PVs was conducted based on the available knowledge and the trend of the monitored signals. For instance, all the signals with low information content (e.g., almost constant signals) were discarded. However, it is worth mentioning that every PV can be included in the framework, but it could lead to longer processing time. The selected sensors measure either thermodynamic PVs of the elaborated fluid or relevant physical variables. Specifically, the considered PVs are mostly pressure and temperature values acquired in different positions of the considered equipment. Moreover, physical PVs such as the position of valves and net active power were included in the analysis. After this selection, data related to different periods are extracted (Step 2). A total of 11,195,120 data points, belonging to eleven distinct time series, were collected. It is worth mentioning that the PVs are characterized by a distinct nature, and the sampling frequency could be slightly different as well. Thus, a synchronization process is applied to align the data coming from different sensors. The extracted data are classified by expert judgments in two distinct operating conditions by analyzing the inlet pressure of the first stage of the compressor (Step 3). Specifically, the two operating conditions are denoted as follows: I) regime or good working, II) surge. The last operating condition could be considered a

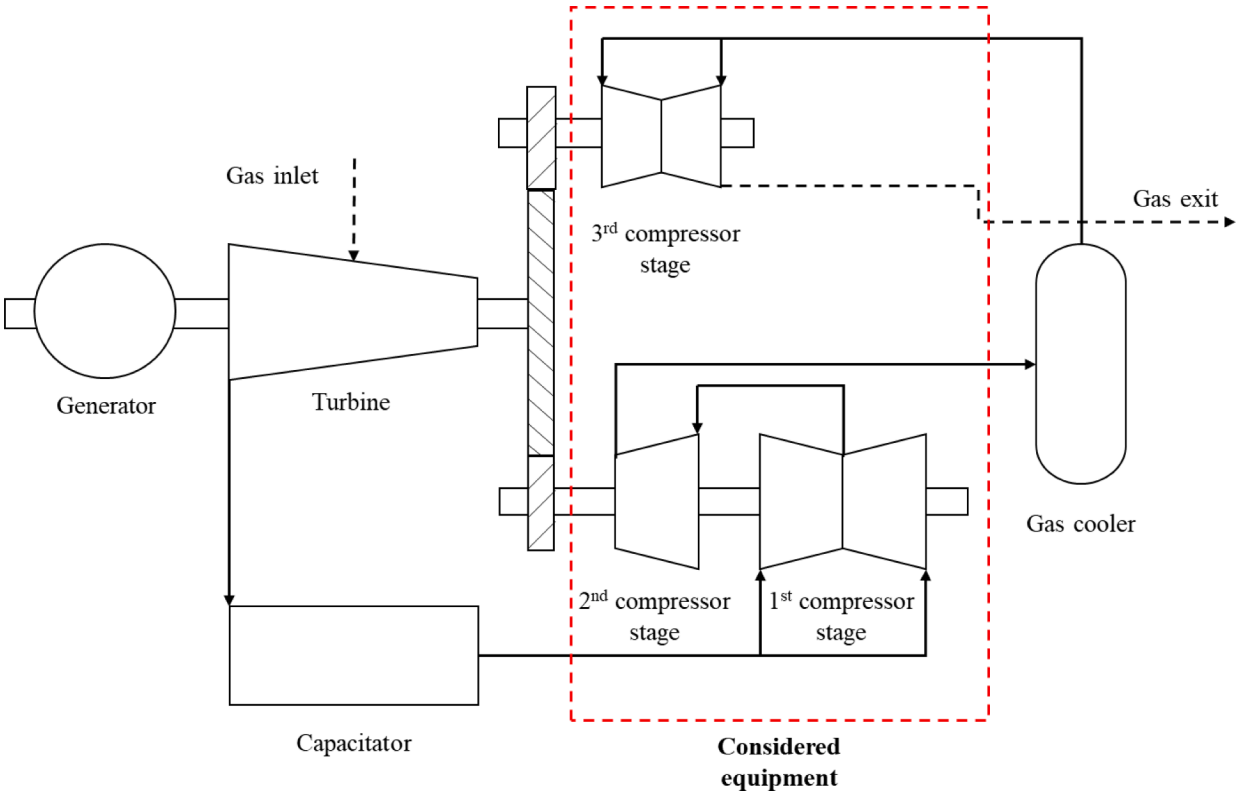


Fig. 2. Representation of the analyzed compressor within its operating system.

Table 1
Selected process variables.

#	Monitored process variable	Unit of measurement
1	Net active power	MW
2	Wet bulb temperature	°C
3	Flow rate - low pressure stage	mm_{H_2O}
4	Flow rate - high pressure stage	mm_{H_2O}
5	Suction gas pressure - low pressure stage	mbar
6	Suction gas pressure - medium pressure stage	mbar
7	Suction gas pressure - high pressure stage	mbar
8	Outlet high pressure stage gas pressure	mbar
9	Exhaust gas pressure	mbar
10	Interstage pressure gas extractor	mbar
11	Interstage pressure gas extractor	mbar
12	Interstage pressure gas extractor	mbar
13	Suction gas temperature - low pressure stage	°C
14	Suction gas temperature - low pressure stage	°C
15	Suction gas temperature - high pressure stage	°C
16	First stage temperature	°C
17	Second stage temperature	°C
18	Third stage temperature	°C
19	Outlet capacitator temperature	°C
20	Outlet third stage temperature	°C
21	Interstage gas temperature	°C
22	Interstage gas temperature	°C
23	Interstage gas temperature	°C
24	Interstage gas temperature	°C
25	Position of the first anti-surge valve	%
26	Position of the second anti-surge valve	%
27	Capacitator absolute pressure	mbar

failure mode since it is an undesired state that could lead to the failure of the entire compressor if it is prolonged over time. Among the 11,195,120, observations, only a total of 391,893 points were defined as surge observations, while the remaining 10,803,227 points were identified as the regime. To gain a better insight into the available dataset, Fig. 3 shows some of the collected signals for the 11 surge events and the eleven regime events. It is a reduced example due to the limited space and company policies.

4.2. Stage 2: Noise removal and data reduction

4.2.1. EMD application to detect noisy IMFs

Most of the acquired signals include a strong noise component, especially for the surge operating condition with highly dynamic and fluctuating PVs, presenting non-stationary and non-linear behavior. Consequently, removing random noise is a fundamental step in improving the accuracy of the methodology. This task is performed for each sensor through the EMD (Step 4) by setting a maximum number of IMFs equal to 20. An SST is conducted to distinguish noisy IMFs from the true signal IMFs (Step 6). First, the mean period of each extracted IMF is estimated according to Eq. (3) [29].

$$T_i = \frac{n}{P_i} \tag{3}$$

where n and P_i denote the number of acquired data points and the number of peaks of the i th IMF, respectively. Next, the energy density of each IMF is estimated through Eq. (4) [70].

$$E_i = \frac{1}{n}\sum_{t=1}^{n}\left|c_i^2(t)\right| \tag{4}$$

The mean period and the energy density could be seen as the mean and the variance of the IMFs, respectively. The first IMF is characterized by the highest order of fluctuations, and it is chosen as a reference for the hypothesis test. The hypothesis test used to identify the noisy IMF is based on Eq. (5), whose null hypothesis is that every IMF is a noisy IMF.

$$ln\left(\frac{1}{3}E_1\right) + lnT_1 < lnE_i + lnT_i < ln(3E_1) + lnT_1 \;\; i = 2, 3, \ldots, m \tag{5}$$

where m is the number of IMFs. Consequently, the first IMF is consistently recognized as noise. Furthermore, all IMFs, for which the null

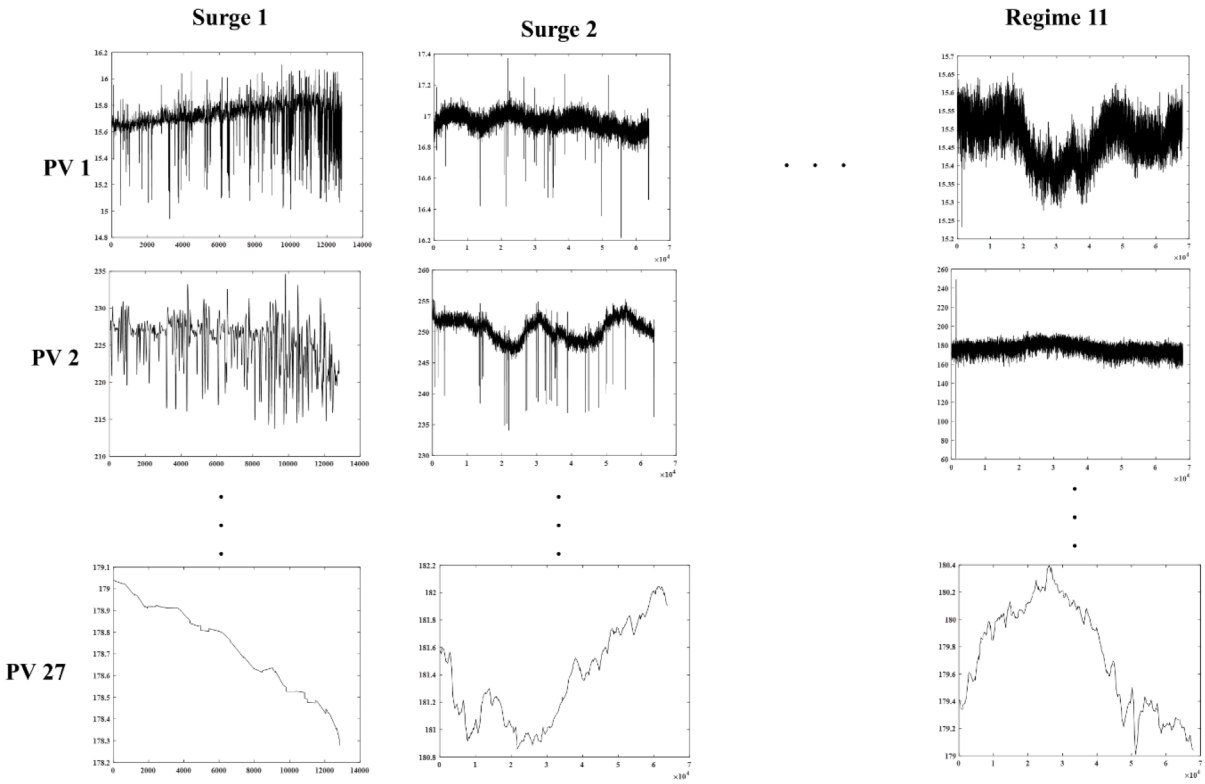


Fig. 3. Example of collected signals for distinct surge and regime events.

hypothesis is accepted, are defined as noisy IMFs.

As an example, the EMD of one of the sensors related to a surge event is considered. First, the monitored signal is decomposed into its corresponding IMFs and a residual through the sifting process. The sifting ends as soon as either the maximum number of IMFs is obtained, or the computed residual is monotonic. For the signal considered, ten IMFs are extracted, as depicted in Fig. 4.

For each IMF the mean period and energy density are calculated according to Eq. 9 and Eq. 10. Subsequently, based on the computed values, the null hypothesis of Eq. 11 is tested for each IMF to detect noisy IMFs. Among the ten IMFs, the first, the second, the third, the fifth, and the seventh resulted as noisy, while the remaining IMFs belong to the true signal (see Fig. 5). Finally, the denoised signal is reconstructed as the sum of the true signal IMFs and the residual, as illustrated by Eq. (6).

$$DS(t) = \sum_{i=1}^{n} c_{i,TS}(t) + r(t) \tag{6}$$

where $c_{i,TS}(t)$ and $r(t)$ denote the i th true signal IMF and the residual, respectively, while $DS(t)$ identifies the denoised signal. The original

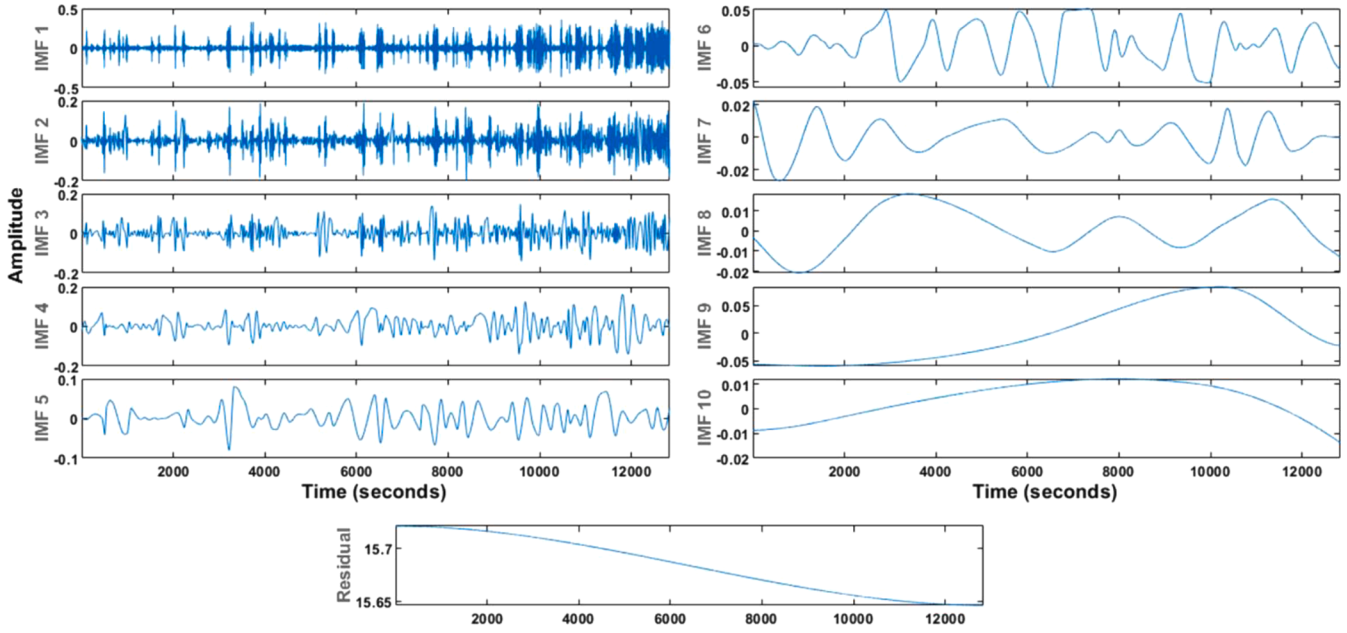


Fig. 4. Example of EMD for one of the PV monitored during a surge event.

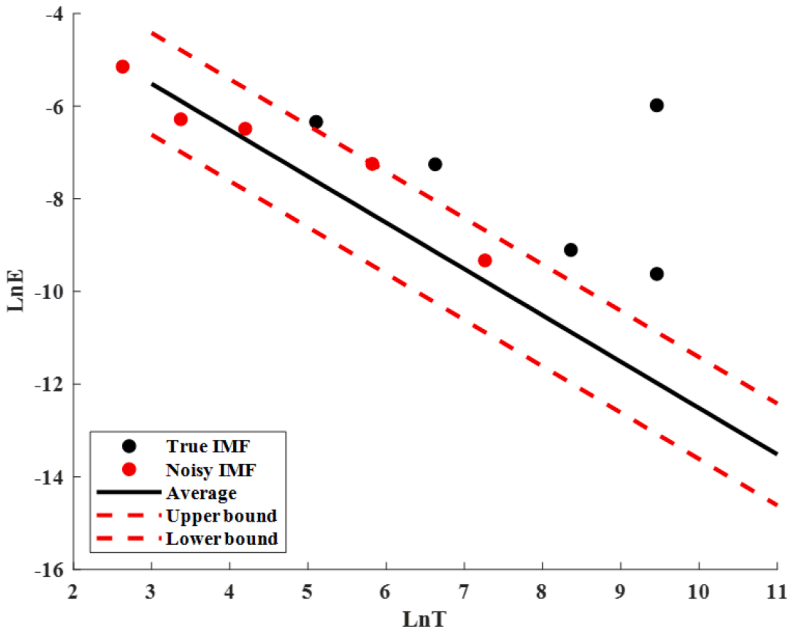


Fig. 5. Noisy and true signal IMFs for one of the PVs monitored during a surge event.

monitored signal and the denoised signal of the illustrated example are shown in Fig. 6.

The signal of the example is highly dynamic, with a high level of fluctuation. However, the filtering process can both capture the trend of the signal and reduce its peaks. It is worth mentioning that the combination of EMD and SST also performs well for less complex signals characterized by fewer fluctuations and variability. Indeed, for this kind of signal, the filter identifies a lower number of noisy IMFs, thus, the denoised signal could result very similar to the original one. As an example, the denoised signals and the original monitored signals for two less fluctuating PVs are shown in Fig. 8. The PVs of Fig. 7 show two temperatures: the left image represents the wet bulb temperature, while the right one depicts an interstage gas temperature. In fact, the temperature measurements are less fluctuating compared to the pressure measurements.

4.2.2. NCA application to determine the most relevant PVs

The collected data are highly unbalanced since 391,893 observations were collected for the surge operating condition, whereas the regime data points are 10,803,227. Thus, before applying the NCA, the dataset was balanced. Indeed, it is essential to adopt a well-balanced data set in a prediction model [71]. Nevertheless, it is worth mentioning that this was possible thanks to the large available dataset concerning regime observations. Based on the previous statements, 391,893 observations were randomly extracted from the regime dataset and fed to the NCA along with all surge data (step 7). The results arising from the

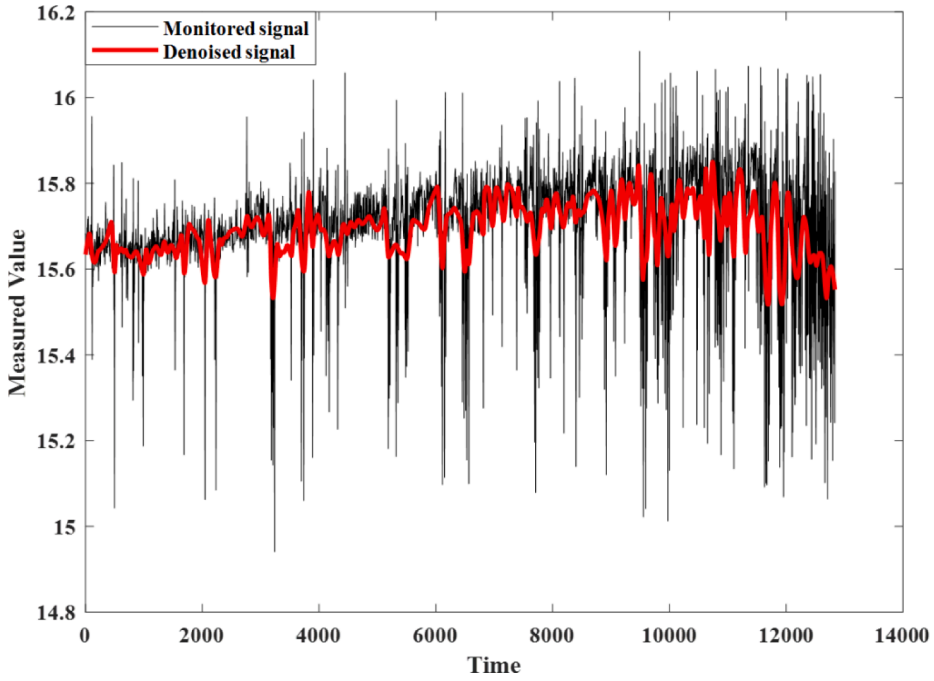


Fig. 6. Original and denoised signal of the considered PV during a surge event.

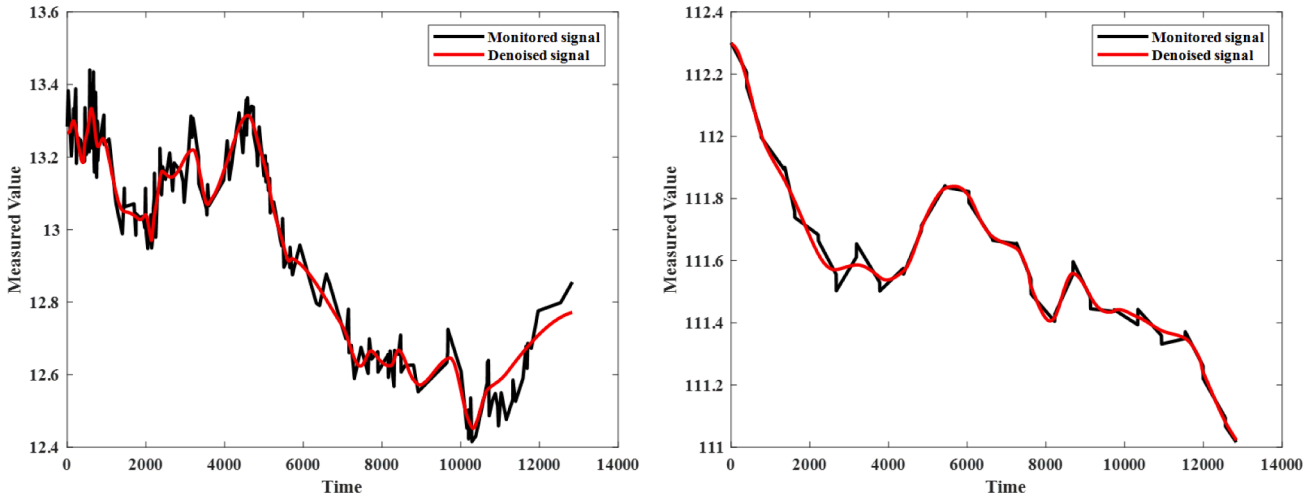


Fig. 7. Monitored and denoised signal of two PVs characterized by low fluctuations.

application of the NCA are depicted in Fig. 8 and Table 2, where the relative weight of the *i* th PV is obtained through the ratio of the absolute weight associated with the i th PV (W_i) and the sum of all the estimated absolute weights (see Eq. (12)).

$$RW_i = \frac{W_i}{\sum_{j=1}^{n} W_j} \tag{12}$$

It emerges that the most relevant PV is the suction gas temperature of the high-pressure stage, while the least important is the flow rate of the high-pressure stage. Furthermore, it could be seen that the contribution of the PVs after the thirteenth is almost equal to 0. Finally, the first four PVs explain more than 50% of the cumulative weight. Therefore, we decided to consider these PVs for the subsequent steps of analysis, to reduce the time required by the calculation, especially for online monitoring purposes. As a matter of fact, increasing the number of PVs could lead to higher accuracy, but it could also cause longer calculation time. Each PV after the fourth one accounts for less than 10% of the cumulative weight, thus, they are less relevant to identify the operating condition of the monitored equipment. Indeed, the first four PVs are associated with at least 10% of the cumulative weight. These considerations have led to select only the first four PVs as a first attempt to assess the performance of the method, while assuring short calculation time. Saying that, the impact of considering the fifth PV is presented in the discussion section.

4.3. Stage 3: Classification through machine learning

The initial data set was reduced to consider the first four most relevant PVs, which were identified as the suction gas temperature of the high-pressure stage, the gas temperature between stages, and the two interstage gas pressures. Moreover, the available data are split into a training and a test set to verify the generalization capability of the obtained model. To this end, 75% of the surge observations are randomly extracted as a training set (i.e., 293,920 observations). Furthermore, the same amount of data points were considered as a training set for the regime. Accordingly, 587,840 observations (equally divided between surge and regime conditions) were chosen and used as the training set. On the other hand, the remaining 10,607,280 observations (i.e., 97,973 surge observations and 10,509,307 regime observations) were used as a

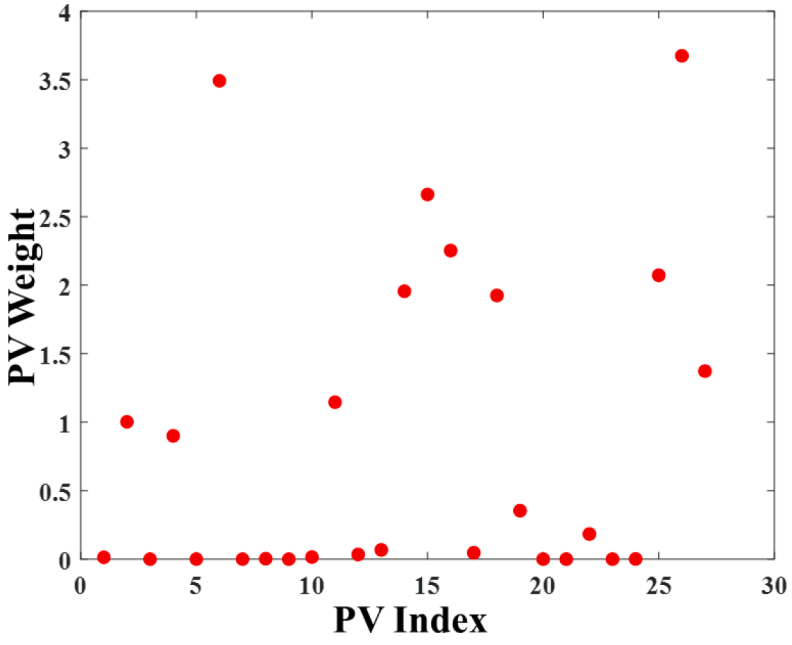


Fig. 8. Weight associated with the NCA to each PV.

Table 2
Ranking, weight, and relative weight of each PV.

Monitored process variable	Ranking	Weights	Relative Weight	Cumulative Weight
Suction gas temperature - high pressure stage	1	3.67	16%	16%
Interstage gas temperature	2	3.49	15%	31%
Interstage pressure gas extractor	3	2.66	11%	42%
Interstage pressure gas extractor	4	2.25	10%	52%
Third stage temperature	5	2.07	9%	61%
Interstage pressure gas extractor	6	1.96	8%	70%
Capacitator absolute pressure	7	1.93	8%	78%
Outlet third stage temperature	8	1.37	6%	84%
Suction gas pressure - high pressure stage	9	1.15	5%	89%
Flow rate - low pressure stage	10	1.00	4%	93%
Suction gas pressure - medium pressure stage	11	0.90	4%	97%
Position of the first anti-surge valve	12	0.35	2%	98%
First stage temperature	13	0.18	1%	99%
Exhaust gas pressure	14	0.07	0%	100%
Suction gas pressure - low pressure stage	15	0.05	0%	100%
Outlet high stage gas pressure	16	0.03	0%	100%
Wet bulb temperature	17	0.02	0%	100%
Net active power	18	0.01	0%	100%
Interstage gas temperature	19	0.00	0%	100%
Second stage temperature	20	0.00	0%	100%
Outlet capacitator temperature	21	0.00	0%	100%
Interstage gas temperature	22	0.00	0%	100%
Position of the second anti-surge valve	23	0.00	0%	100%
Interstage gas temperature	24	0.00	0%	100%
Suction gas temperature - low pressure stage	25	0.00	0%	100%
Suction gas temperature - low pressure stage	26	0.00	0%	100%
Flow rate - high pressure stage	27	0.00	0%	100%

test set. We decided to adopt 75% of the data as a training set since 75–25 is a common proportion for training and test set. As a matter of fact, exploiting a higher proportion for the training of the surge could lead to less generalization capabilities and overtraining. On the other hand, considering a lower training proportion could result in insufficient training data due to the low number of observations, and possibly worst accuracy. This fact is particularly critical for the surge condition since there could be high differences between one surge and another. Accordingly, it is important to consider enough data for the training phase to pursue sufficient generalization capabilities. By contrary, different regime states share more similarities, thus, even with few observations, it is possible to obtain high generalizability and reduce the training time. The previous consideration represents another reason that led to the adoption of a test set with more observations than the training set with regard to the regime state. This allows us to both better verify the generalization capability of the regime conditions and concurrently have a balanced dataset for the training phase, which is a strongly desired condition.

The optimization of a ML approach was out of the scope of this work since it is a well-known topic. Therefore, we adopted a RF with the characteristics highlighted in Table 3. However, it is worth mentioning that trying different RF's parameters and accordingly performing a sensitivity analysis could improve the accuracy of the classification. Based on this consideration, it is possible to test different combinations of number of learners-maximum number of splits and subsequently choose the best one.

The training was conducted through a 5-fold cross-validation (step 8), which resulted in the confusion matrix of Table 4. The calculation depicted that 13,585 surge observations were classified as regime, while only 2039 regime observations were misclassified as a surge. Defining the accuracy as the ratio between the number of correctly classified observations and the total number of observations, the training accuracy resulted equal to 97.34%. Based on this value, it is possible to state that the model is reliable for the classification purposes of the training set.

One of the main issues that could arise from ML approaches is the lack of generalization. In other words, a model could be very accurate for the training dataset but, in turn, it could not predict new observations accurately. This is a scenario that is related to an overlearning of the training dataset, which results in poor generalization. To avoid this issue, the algorithm is constantly tested on a new dataset called a test set. Consequently, the trained algorithm is adopted to predict the class of the test set (step 9), which was previously mentioned. The confusion matrix related to the test set is shown in Table 5.

The RF correctly predicted 97.35% of the observations, denoting a high degree of generalization.

5. Discussion

Based on the results illustrated in Section 4, it is possible to state that the proposed methodology is capable of removing noise from the monitored signal and, after selecting the most relevant PVs, it performs a diagnosis of the condition of the monitored equipment. Indeed, the model resulted to be very accurate and efficient since about 97% of the time the health of the system was correctly predicted. Moreover, the undesired operating condition (that is, the surge) was correctly classified 95% of the time, while the regime condition was incorrectly identified as a surge 3% of the time in the test set and only 1% of the time for the training set. This difference could be related to the nature of the surge events which could be very different. Despite that, these results look promising, since there is a high degree of generalization for the surge operating condition. Indeed, the misclassification cost related to the surge condition is higher compared to the regime operating state being classified as a surge. Indeed, the priority is detecting a dangerous operating state and subsequently activating appropriate procedures to restore a normal working condition. Accordingly, a false negative (i.e., classifying a surge state as a regime) could increase the time the system runs in an abnormal state, leading to a shorter useful life and simultaneously mining the safety of the operations. On the other hand, a false positive (i.e., classifying a regime as a surge) could result in performing unnecessary maneuvers or stopping the operations to reduce the amount of time that the system is spending in an unwanted operating state.

The developed approach is also quite practical since there is no need of specifying any opinion or information during the classification process. Indeed, the proposed model can classify on its own the observations based on the current monitored signals without any external interference. This peculiar feature allows to perform online diagnosis and accordingly define the actions to perform based on the detected

Table 3
Characteristics of the adopted RF.

Characteristic	Value
Ensemble Method	Bag
Split criterion	Gini index
Number of learners	30
Max. number of splits	20

Table 4
Confusion matrix of the training set. Dark cells represent correctly classified observations.

		Predicted class	
		Regime	Surge
True class	Regime	291,881	2039
	Surge	13,585	280,335

Table 5
Confusion matrix of the training set. Dark cells represent correctly classified observations.

		Predicted class	
		Regime	Surge
True class	Regime	10,233,530	275,777
	Surge	4914	93,059

state. The real-time evaluation of the operating condition is pivotal to further improve the safety of the operations since it could assist in reducing the time that the equipment is spending in a risky and undesired state.

To have a more in-depth insight into the obtained results, the scatter plots related to the considered most relevant variables are shown in Fig. 9.

As depicted in Fig. 9, there are some regions where the surge and regime conditions overlap, leading to a classification error. The overlapping could be related to the transition from a regime operating condition to a surge one. Another possible explanation is that the starting data were classified through expert judgments, thus, there is the possibility of including uncertainty and errors from the beginning. Anyway, the proposed model can distinguish a surge condition from a regime operating point even when they are very similar or there is a strong merge between the classes. This task is not easy, and it cannot be considered a normal routine. Therefore, the implementation of the model allows one to perform a tough diagnosis without considering any external input such as expert opinions or physical laws.

Finally, the number of PVs to consider was selected through the cumulative weight without considering any sensitivity analysis. Accordingly, varying the number of PVs adopted for the classification could be a viable option to improve the accuracy of the classification. Even though the selection of the best number of PVs was out of the scope of this work, as an example, the classification with the first five most relevant PVs is considered. The inclusion of the fifth PV resulted in the confusion matrices of Table 6 for the training set and Table 7 for the test set. Accordingly, the training and test accuracy are equal to 97.68% and 97.87%, respectively. Therefore, it is possible to state that the prediction accuracy of new observations is slightly increased; however, the complexity of the classification increases as well. A trade-off between accuracy and calculation time should be considered to determine the number of PVs to adopt for the prediction. Another important aspect is that the prediction accuracy of the surge event increases when adopting five PVs, while the prediction accuracy of the regime condition is slightly lower.

The number of PVs to consider could be determined through a sensitivity analysis based on the main purpose of the classification. For instance, the main objective could be correctly classifying the surge condition. Accordingly, it is possible to define the test accuracy as optimization parameter and repeat the training of the RF by adding one PV at a time, following the ranking presented in Table 2. The number of PV to consider for further analysis would be the one that maximizes the test accuracy of the surge condition (i.e., provide the higher generalizability performance for the surge). It is worth mentioning that it is not required to repeat the training for all 27 PVs, but it is possible to consider up to the thirteenth most relevant PV. Indeed, the PVs after the fourteenth are associated with a cumulative weight lower than 1%, thus, it is safe to assume that they are not relevant for the classification. As a matter of fact, they could even add uncertainty and lead to lower accuracy.

Table 6
Confusion matrix of the training set composed of five PVs.

		Predicted class	
		Regime	Surge
True class	Regime	287,507	6413
	Surge	7215	286,705

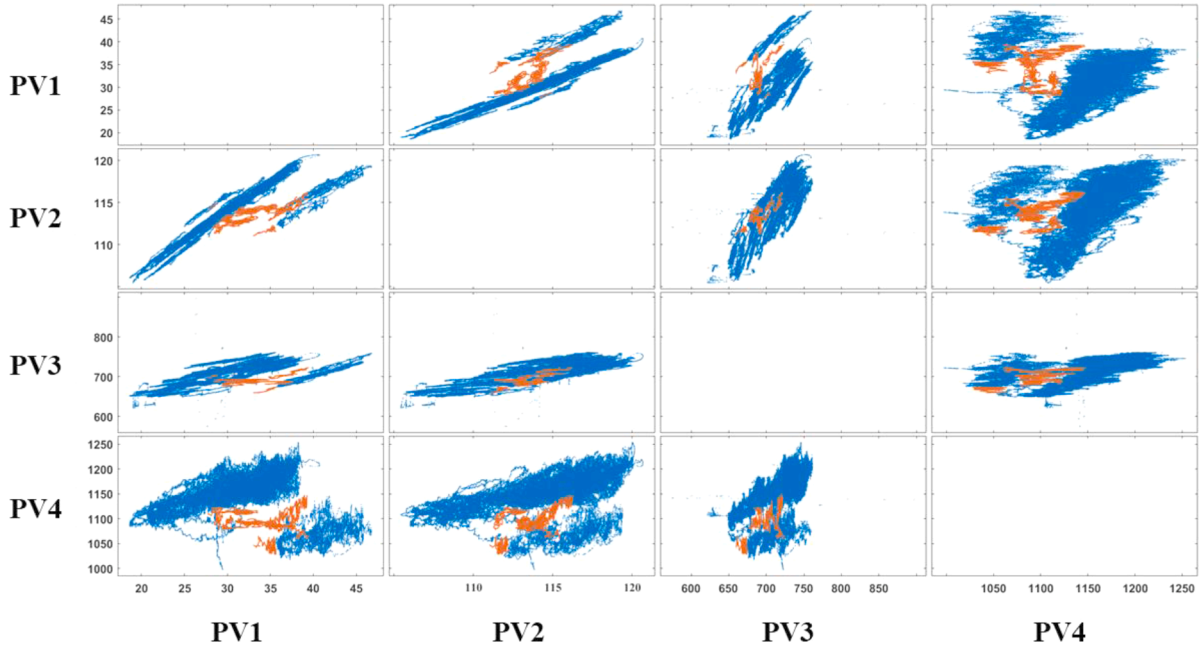


Fig. 9. Scatter plots for all four most relevant PVs. The blue and orange dots represent regime and surge observations respectively.

Table 7
Confusion matrix of the test set composed of five PVs.

		Predicted class	
		Regime	Surge
True class	Regime	10,285,150	224,157
	Surge	2219	95,754

6. Conclusions

This paper presents a novel methodology capable of performing failure diagnosis of a system based on a set of monitored PVs. In the proposed approach, a number of signals equal to the number of considered PVs are extracted from sensors, and their noise is filtered out through EMD. Next, the most relevant PVs are selected through NCA. Finally, the remaining PVs are exploited to implement a supervised RF classification model. The framework was tested on a real case study of a compressor operating in a geothermal plant. The obtained results are factual since the training and test accuracies were estimated as 97.34% and 97.35%, respectively.

The proposed approach could be used for online condition monitoring purposes of equipment with highly non-stationary and non-linear PVs, without the need to extract features from the acquired signals. Specifically, it could assist in the decision-making process related to maintenance planning. Indeed, the methodology facilitates online failure diagnosis, providing the current operating condition of the monitored equipment. In case the monitored equipment is identified in an undesired state, it is possible to intervene to repristinate the normal operating condition. This characteristic allows to assure the safety of the operations, limiting the time that the system spends in a dangerous state (e.g., the surge). Furthermore, the identification of the most relevant PVs is of prominent importance to pinpoint which are the sensors towards which directing condition monitoring efforts. Indeed, the malfunctioning of one sensor associated with one of the most relevant PV could lead to inaccuracy and, therefore, a higher level of uncertainty.

In this work, the optimization of the ML parameters and the selection of an optimum number of PVs was not considered. Accordingly, future works could include such aspects. For instance, the user could select a parameter to optimize (e.g., the test accuracy) and, subsequently, perform a sensitivity analysis by adding one PV at a time, following the ranking arising from the NCA. This process will allow determining the number of PVs that optimize the selected parameter. Moreover, the exploitation of distinct ML techniques could be taken into account. Once again different ML techniques could be tested for data processing and the ML algorithm that optimizes a target parameter should be chosen. It is also possible to combine a sensitivity analysis for the ML algorithm and the number of PVs, testing each classification technique with an increasing number of PVs, adding one PV at a time. Furthermore, it is worth mentioning that the comparison of the developed methodology with similar approaches could be useful to evaluate which is the better performing one. As a matter of fact, it could be interesting to try different combinations of noise removal (e.g., WT) and data reduction techniques (e.g., SFS). Finally, further developments could also be related to adopting the methodology for distinct case studies. Indeed, testing the framework on different applications could be helpful to analyze its strengths, capabilities, and limitations. Within this context, it could be useful to compare the presented methodology with similar frameworks for different case studies.

CRediT authorship contribution statement

Leonardo Leoni: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Filippo De Carlo:** Visualization, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing – review & editing. **Mohammad Mahdi Abaei:** Writing – review & editing, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ahmad BahooToroody:** Writing – review & editing, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mario Tucci:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

Dear Editor of Reliability Engineering & System Safety, there is no conflict of interest to declare.

Data availability

The authors do not have permission to share data.

References

[1] Liu Z, Zhang L. A review of failure modes, condition monitoring and fault diagnosis methods for large-scale wind turbine bearings. Measurement 2020;149:107002.
[2] Schlechtingen M, Santos IF. Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection. Mech Syst Signal Process 2011;25(5):1849–75.
[3] Wadibhasme J, Zaday S, Somalwar R. Review of various methods in improvement in speed, power & efficiency of induction motor. In: 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS; 2017. p. 3293–6.
[4] Gangsar P, Tiwari R. Signal based condition monitoring techniques for fault detection and diagnosis of induction motors: a state-of-the-art review. Mech Syst Signal Process 2020;144:106908.
[5] Toliyat HA, Abbaszadeh K, Rahimian MM, Olson LE. Rail defect diagnosis using wavelet packet decomposition. IEEE Trans Ind Appl 2003;39(5):1454–61.
[6] Márquez FPG, Schmid F, Collado JC. A reliability centered approach to remote condition monitoring. A railway points case study. Reliab Eng Syst Saf 2003;80(1):33–40.
[7] Zhang W, Jia M-P, Zhu L, Yan X-A. Comprehensive overview on computational intelligence techniques for machinery condition monitoring and fault diagnosis. Chinese J Mech Eng 2017;30(4):782–95.
[8] Soltanali H, Rohani A, Abbaspour-Fard MH, Parida A, Farinha JT. Development of a risk-based maintenance decision making approach for automotive production line. Int J Syst Assur Eng Manag 2020;11(1):236–51.
[9] Ferreira RJ, de Almeida AT, Cavalcante CA. A multi-criteria decision model to determine inspection intervals of condition monitoring based on delay time analysis. Reliab Eng Syst Saf 2009;94(5):905–12.
[10] Ilonen J, Kamarainen J-K, Lindh T, Ahola J, Kalviainen H, Partanen J. Diagnosis tool for motor condition monitoring. IEEE Trans Ind Appl 2005;41(4):963–71.
[11] Roy S, Sinha N, Sen AK. A new hybrid image denoising method. Int J Inf Technol Knowl Manag 2010;2(2):491–7.
[12] Vishwakarma M, Purohit R, Harshlata V, Rajput P. Vibration analysis & condition monitoring for rotating machines: a review. Mater Today: Proc 2017;4(2):2659–64.
[13] Tsolis G, Xenos TD. Signal denoising using empirical mode decomposition and higher order statistics. Int J Signal Process Image Process Pattern Recognit 2011;4(2):91–106.
[14] Saini K, Dhami SS. Predictive monitoring of incipient faults in rotating machinery: a systematic review from data acquisition to artificial intelligence. Arch Comput Meth Eng 2022:1–22.
[15] Morozov AL, et al. Microcontroller realization of an induction motors fault detection method based on FFT and statistics of fractional moments. In: 2021 29th Mediterranean Conference on Control and Automation (MED); 2021. p. 65–70.
[16] Gowid S, Dixon R, Ghani S. A novel robust automated FFT-based segmentation and features selection algorithm for acoustic emission condition based monitoring systems. Appl Acoust 2015;88:66–74.
[17] P. Sparis and G. Vachtsevanos, "Automatic diagnostic feature generation via the Fast Fourier Transform," Citeseer.
[18] Majali A, Mulay A, Iyengar V, Nayak A, Singru P. Fault identification and remaining useful life prediction of bearings using Poincare maps, fast Fourier transform and convolutional neural networks. Math Model Eng 2022;8(1):1–14.
[19] Hussein R, BashirShaban K, El-Hag AH. Denoising of acoustic partial discharge signals corrupted with random noise. IEEE Trans Dielectr Electr Insul 2016;23(3):1453–9.
[20] Zhang C, et al. Rolling element bearing fault diagnosis based on the wavelet packet transform and time-delay correlation demodulation analysis. Advances in asset management and condition monitoring. Springer; 2020. p. 1195–203.
[21] Bera A, Dutta A, Dhara AK. Deep learning based fault classification algorithm for roller bearings using time-frequency localized features. In: 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS); 2021. p. 419–24.

[22] Lopes WN, et al. An efficient short-time Fourier transform algorithm for grinding wheel condition monitoring through acoustic emission. Int J Adv Manuf Technol 2021;113(1):585–603.
[23] Bae SJ, Mun BM, Chang W, Vidakovic B. Condition monitoring of a steam turbine generator using wavelet spectrum based control chart. Reliab Eng Syst Saf 2019; 184:13–20.
[24] Jiménez AA, Muñoz CQG, Márquez FPG. Dirt and mud detection and diagnosis on a wind turbine blade employing guided waves and supervised learning classifiers. Reliab Eng Syst Saf 2019;184:2–12.
[25] Mousavi AA, Zhang C, Masri SF, Gholipour G. Structural damage detection method based on the complete ensemble empirical mode decomposition with adaptive noise: a model steel truss bridge case study. Struct Health Monitor 2022;21(3): 887–912.
[26] Huang NE, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc Math Phys Eng Sci 1998; 454(1971):903–95.
[27] Tian J, Wang S-G, Zhou J, Ai Y-T, Zhang Y-W, Fei C-W. Fault diagnosis of intershaft bearing using variational mode decomposition with TAGA optimization. Noise Control 2021;2021.
[28] Desavale RG, Jadhav PM, Dharwadkar NV. Dynamic response analysis of gearbox to improve fault detection using empirical mode decomposition and artificial neural network techniques. ASCE-ASME J Risk Uncert Engrg Sys Part B Mech Engrg 2021;7(3).
[29] BahooToroody A, Abaei MM, BahooToroody F, De Carlo F, Abbassi R, Khalaj S. A condition monitoring based signal filtering approach for dynamic time dependent safety assessment of natural gas distribution process. Process Saf Environ Prot 2019;123:335–43.
[30] Rafiq HJ, Rashed GI, Shafik MB. Application of multivariate signal analysis in vibration-based condition monitoring of wind turbine gearbox. Int Trans Electric Energy Syst 2021;31(2):e12762.
[31] Nishat Toma R, Kim C-H, Kim J-M. Bearing fault classification using ensemble empirical mode decomposition and convolutional neural network. Electronics (Basel) 2021;10(11):1248.
[32] Tang Y, Liu Q, Zhu Q. Fault simulation and forecast of helical cylindrical gear of reducer based on ADAMS. J Phys Conf Ser 2021;1983(1):012019.
[33] BahooToroody A, De Carlo F, Paltrinieri N, Tucci M, Van Gelder P. Bayesian regression based condition monitoring approach for effective reliability prediction of random processes in autonomous energy supply operation. Reliab Eng Syst Saf 2020;201:106966.
[34] Yu J. State of health prediction of lithium-ion batteries: multiscale logic regression and Gaussian process regression ensemble. Reliab Eng Syst Saf 2018;174:82–95.
[35] Yan G, Yu C, Bai Y. A new hybrid ensemble deep learning model for train axle temperature short term forecasting. Machines 2021;9(12):312.
[36] Gao Z, Liu Y, Wang Q, Wang J, Luo Y. Ensemble empirical mode decomposition energy moment entropy and enhanced long short-term memory for early fault prediction of bearing. Measurement 2022;188:110417.
[37] Li H, Bu S, Wen J-R, Fei C-W. Synthetical modal parameters identification method of damped oscillation signals in power system. Appl Sci 2022;12(9):4668.
[38] Adams S, et al. A comparison of feature selection and feature extraction techniques for condition monitoring of a hydraulic actuator. In: Annual Conference of the PHM society. 9; 2017.
[39] Caggiano A, Angelone R, Napolitano F, Nele L, Teti R. Dimensionality reduction of sensorial features by principal component analysis for ANN machine learning in tool condition monitoring of CFRP drilling. Procedia CIRP 2018;78:307–12.
[40] Ramirez-Chavez M, Saucedo-Dorantes JJ, Jaen-Cuellar AY, Rios RAO, de Jesus Romero-Troncoso R, Delgado-Prieto M. Condition monitoring strategy based on spectral energy estimation and linear discriminant analysis applied to an induction motor. In: 2018 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC); 2018. p. 1–6.
[41] Gierlak P, Burghardt A, Szybicki D, Szuster M, Muszyńska M. On-line manipulator tool condition monitoring based on vibration analysis. Mech Syst Signal Process 2017;89:14–26.
[42] Ai Y-T, Guan J-Y, Fei C-W, Tian J, Zhang F-L. Fusion information entropy method of rolling bearing fault diagnosis based on n-dimensional characteristic parameter distance. Mech Syst Signal Process 2017;88:123–36.
[43] Tian J, Liu L, Zhang F, Ai Y, Wang R, Fei C. Multi-domain entropy-random forest method for the fusion diagnosis of inter-shaft bearing faults with acoustic emission signals. Entropy 2019;22(1):57.
[44] Khalid S, Khalil T, Nasreen S. A survey of feature selection and feature extraction techniques in machine learning. In: 2014 science and information conference; 2014. p. 372–8.
[45] Goldberger J, Hinton GE, Roweis S, Salakhutdinov RR. Neighbourhood components analysis. Adv Neural Inf Process Syst, 17; 2004.
[46] Raghu S, Sriraam N. Classification of focal and non-focal EEG signals using neighborhood component analysis and machine learning algorithms. Expert Syst Appl 2018;113:18–32.
[47] Yaman O. An automated faults classification method based on binary pattern and neighborhood component analysis using induction motor. Measurement 2021;168: 108323.
[48] Zhou H, Chen J, Dong G, Wang H, Yuan H. Bearing fault recognition method based on neighbourhood component analysis and coupled hidden Markov model. Mech Syst Signal Process 2016;66:568–81.
[49] Dhiman HS, Deb D, Carroll J, Muresan V, Unguresan M-L. Wind turbine gearbox condition monitoring based on class of support vector regression models and residual analysis. Sensors 2020;20(23):6742.
[50] Murphy KP. Machine learning: a probabilistic perspective. MIT press; 2012.
[51] Islam MM, Kim J-M. Reliable multiple combined fault diagnosis of bearings using heterogeneous feature models and multiclass support vector Machines. Reliab Eng Syst Saf 2019;184:55–66.
[52] Tang T, Yuan H. A hybrid approach based on decomposition algorithm and neural network for remaining useful life prediction of lithium-ion battery. Reliab Eng Syst Saf 2022;217:108082.
[53] Lipinski P, Brzychczy E, Zimroz R. Decision tree-based classification for Planetary Gearboxes' condition monitoring with the use of vibration data in multidimensional symptom space. Sensors 2020;20(21):5979.
[54] Patel RK, Giri VK. Feature selection and classification of mechanical fault of an induction motor using random forest classifier. Perspect Sci 2016;8:334–7.
[55] Saeed U, Jan SU, Lee Y-D, Koo I. Fault diagnosis based on extremely randomized trees in wireless sensor networks. Reliab Eng Syst Saf, 205; 2021, 107284.
[56] Zhang C, Hu D, Yang T. Anomaly detection and diagnosis for wind turbines using long short-term memory-based stacked denoising autoencoders and XGBoost. Reliab Eng Syst Saf 2022;222:108445.
[57] Wan S, Li X, Zhang Y, Liu S, Hong J, Wang D. Bearing remaining useful life prediction with convolutional long short-term memory fusion networks. Reliab Eng Syst Saf 2022;224:108528.
[58] Azar K, Hajiakhondi-Meybodi Z, Naderkhani F. Semi-supervised clustering-based method for fault diagnosis and prognosis: a case study. Reliab Eng Syst Saf 2022; 222:108405.
[59] Zhu Y, Wu J, Wu J, Liu S. Dimensionality reduce-based for remaining useful life prediction of machining tools with multisensor fusion. Reliab Eng Syst Saf 2022; 218:108179.
[60] Xu F, Yang F, Fei Z, Huang Z, Tsui K-L. Life prediction of lithium-ion batteries based on stacked denoising autoencoders. Reliab Eng Syst Saf 2021;208:107396.
[61] Karatoprak E, Seker S. An improved empirical mode decomposition method using variable window median filter for early fault detection in electric motors. Math Probl Eng 2019;2019.
[62] Kumar V, Minz S. Feature selection: a literature review. SmartCR 2014;4(3): 211–29.
[63] Yang W, Wang K, Zuo W. Neighborhood component feature selection for high-dimensional data. J Comput 2012;7(1):161–8.
[64] Belgiu M, Drăguţ L. Random forest in remote sensing: a review of applications and future directions. ISPRS J Photogramm Remote Sens 2016;114:24–31.
[65] Bonissone P, Cadenas JM, Garrido MC, Díaz-Valladares RA. A fuzzy random forest. Int J Approx Reason 2010;51(7):729–47.
[66] Küppers F, Albers J, Haselhoff A. Random forest on an embedded device for real-time machine state classification. In: 2019 27th European signal processing conference (EUSIPCO); 2019. p. 1–5.
[67] Ahn H, Moon H, Fazzari MJ, Lim N, Chen JJ, Kodell RL. Classification by ensembles from random partitions of high-dimensional data. Comput Stat Data Anal 2007;51(12):6166–79.
[68] Meharie MG, Shaik N. Predicting highway construction costs: comparison of the performance of random forest, neural network and support vector machine models. J Soft Comput Civil Eng 2020;4(2):103–12.
[69] Shoar S, Chileshe N, Edwards JD. Machine learning-aided engineering services' cost overruns prediction in high-rise residential building projects: application of random forest regression. J Build Eng 2022;50:104102.
[70] Wu Z, Huang NE. A study of the characteristics of white noise using the empirical mode decomposition method. Proc Math Phys Eng Sci 2004;460(2046):1597–611.
[71] Rahman MM, Davis DN. Addressing the class imbalance problem in medical datasets. Int J Mach Learn Comput 2013;3(2):224.