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

Failure diagnosis with multiple non-stationary signals: a data-driven framework

Abstract

Due to higher reliability and safety requirements, the importance of condition monitoring and failure diagnosis has progressively cleared up. In this context, to be able to properly deal with noise and data reduction is fundamental to improve failure diagnosis and to assure safe operations. Accordingly, this paper aims to develop a failure diagnosis methodology that integrates Empirical Mode Decomposition (EMD) and Neighborhood Component Analysis (NCA) to separate the noise from the monitored signals and to determine the most relevant features. While noise detection and reduction techniques are established to reduce the uncertainties integrated with data acquisition and collection, traditional estimation approaches that cannot capture the non-stationary and nonlinear nature of data might result in higher uncertainty. As a validated denoising method, EMD is applied in this study to cope with the aforementioned limitations. The NCA overcomes typical limitations such as imposing class distributions. After data pre-processing, the diagnosis is performed through a Random Forest, one of the most renowned Machine Learning algorithms. The methodology is tested on real data coming from a compressor, showing an accuracy higher than 97% for both the training and test set. The developed framework could assist practitioners in evaluating the condition of assets and, accordingly, planning maintenance.

Keywords: Condition monitoring, Failure diagnosis, Empirical Mode Decomposition, Neighborhood Component Analysis, Supervised Classification

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 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, are worth mentioning 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 nonstationary 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].

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. Also, EMD does not need the indication of a basic function such as most WTs [27].

Due to its advantages, EMD and its derivative approaches have become popular tools to perform CM and failure diagnoses [28-33]. A recent study by Yan et al. [34] 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. [35] 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.

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 [36]. 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) [37], Linear Discriminant Analysis (LDA) [38], and Sequential Feature Selection (SFS) [39] are worth mentioning. The techniques mentioned above present critical drawbacks. In fact, PCA could produce information loss, and it does not provide labeled data, 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 [40]. Meanwhile, Neighborhood Component Analysis (NCA) as a linear nonparametric feature selection approach has been introduced by Goldberger et al. [41], overcoming the limitations related to imposing a class distribution or decision boundaries. Moreover, NCA does not lose any information within the data reduction process [42]. Thanks to its advantages, NCA has been successfully applied within CM, failure diagnosis, and fault detection frameworks [43-45]. Yaman [43] used NCA for extracting the most relevant features, which are subsequently fed to classification techniques for performing diagnosis of an induction motor. A similar work has been proposed by Zhou et al. [44], 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 [46]. 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) [47], Neural Network (NN) [48], Decision Tree (DT) [49], and Random Forest (RF) [50]. 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 [51-54]. A relevant example is a work presented by Zhu et al. [55], 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. [56], the authors proposed an advanced methodology to predict the life cycle of lithium-ion batteries. In their work, a 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 requires 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 Significant Testing (SST). On the other side, NCA is exploited for its ability to preserve information. Finally, the denoised most relevant signals are fed to an RF to classify the state of the system. The RF was chosen for its ease of implementation, explainability, and reliability in classification [57]. Furthermore, the joining of multiple individual classifiers, such as the RF, improves performance [58]. To demonstrate the applicability of the methodology, a compressor operating in a geothermal plant is chosen as a case study. 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. Moreover, EMD was used for feature extraction instead of noise removal. 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. Moreover, EMD was used for feature extraction instead of noise removal.

The remainder of this paper is organized as follows; **Section 2** introduces the material and methods, while **Section 3** describes the developed framework. **Section 4** describes the application of the novel approach to a case study. Finally, in **Section 5**, the results are discussed, while in **Section 6**, the conclusions are drawn.

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 [59], as shown by Eq. 1.

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

where n is the number of IMFs, while $c_i(t)$ is the i-th IMF. Finally, $r(t)$ is the residual term. The process of generating the IMFs is called sifting. It allows us to obtain a set of IMFs which fulfills the following requirements [60]: i) the difference between the number of extrema and zero-up crossings is zero or equal to one; ii) the mean value defined through the local minima envelope and local maxima envelope is zero in every point.

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 [28]:

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

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 [61]. The NCA was introduced by Goldberger et al. [41], 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 [62]. This task is accomplished through the maximization of the leave-one-out classification accuracy.

The following paragraphs summarize the procedure for performing NCA, which is widely described by Goldberger et al. [41], Yang et al. [62], and Raghu and Sriraam [42]. Given a training dataset denoted by $D = \{(\boldsymbol{X}_i, y_i), i = 1,2, \dots n\}$, where $\boldsymbol{X}_i$ and $y_i \in \{1,2, \dots C\}$ represent the m-dimensional feature matrix and the class label of the *i*-th observations respectively, the weighting distance in terms of weighting vector can be found through Eq. 3 [42].

$$WD_w(x_i, x_j) = \sum_{k=1}^{m} w_k^2 \, |x_{i,k} - x_{j,k}| \tag{3}$$

where x_i and x_j are two observations, while w_k is the weight associated with the *k*-th feature. Finally, m identifies the number of features. To maximize the classification accuracy through the leave-one-out technique, an observation is randomly extracted from D as a reference point. Specifically, the probability distributions that are used to select the reference point are illustrated in Eq. 4 [42].

$$p_{i,j} = \begin{cases} \frac{ker\left(WD_w(x_i,x_j)\right)}{\sum_{j=1}^{n} ker\left(WD_w(x_i,x_j)\right)} & if\ i \neq j \\ 0 & if\ i = j \end{cases} \tag{4}$$

where $ker(z) = \exp\,(-z/\sigma)$ is the kernel function with width denoted by σ. According to Eq. 4, the probability of the reference point x_i to be correctly classified is found through Eq. 5 [42].

$$p_i = \sum_{j=1}^{n} p_{i,j} y_{i,j} \quad i \neq j \tag{5}$$

where $y_{i,j} = 0$ for every i but $i = j$ which is characterized by $y_{i,j} = 1$. Thus, as reported by Yang et al. [62], the leave-one-out classification accuracy is expressed by Eq. 6 and it can be maximized after the introduction of a regularization term as denoted by Eq. 7:

$$CA(w) = \frac{1}{n}\sum_{i=1}^{n} p_i = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} p_{i,j} y_{i,j} \quad i \neq j \tag{6}$$

$$CA(w) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_{i,j} y_{i,j} - \lambda \sum_{k=1}^{m} w_k^2 \quad i \neq j \tag{7}$$

where $\lambda > 0$ is the regularization parameter. After taking the derivative of Eq. 7 and reordering some terms, Eq. 8 is obtained [42]:

$$\frac{\partial CA(w)}{\partial w} = 2\left(\frac{1}{\sigma}\sum_i \left(p_i \sum_{j \neq i} p_{i,j} \, |x_{i,k} - x_{j,k}| - \sum_j p_{i,j} y_{i,j\,|x_{i,k}-x_{j,k}|}\right) - \lambda\right) w_k \tag{8}$$

2.3 Random Forest

Several algorithms and techniques could be adopted for classification purposes. An RF is a well-known ML approach based on DT. Specifically, an RF is an ensemble classifier that combines a set of DTs through a bagging process [63]. 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 [64]. Also, each DT could consider different sets of features. 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 training phase, the final class is obtained through an arithmetic mean of each result arising from a single DT, while for the testing, the predicted class is the one which has been determined by most of the DTs. An example of RF is shown in Fig. 1.

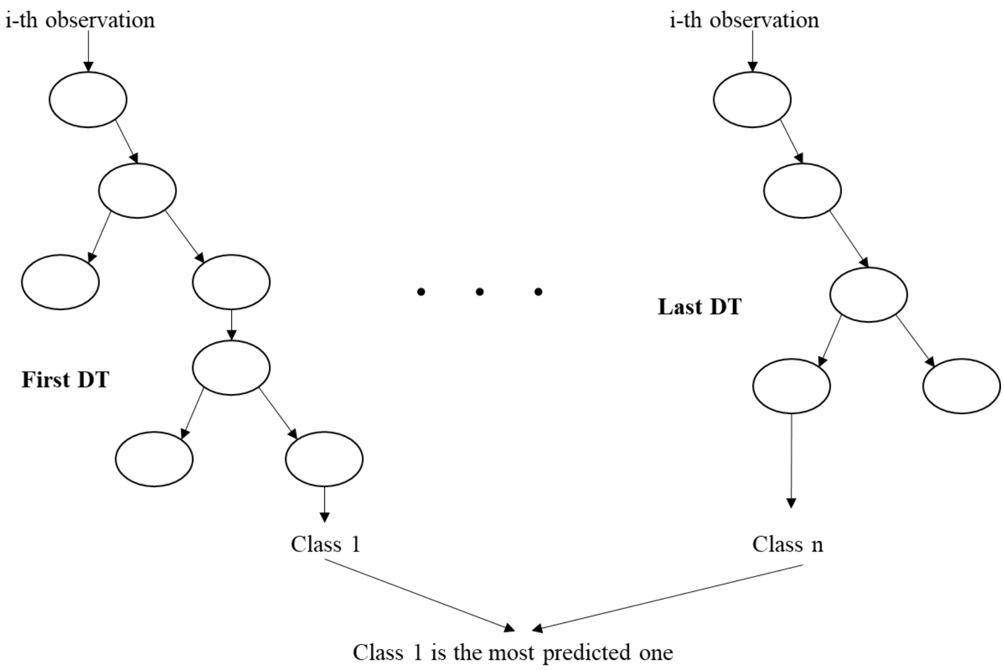

Fig. 1 Schematic example of RF prediction with *n* classes

3. Developed Methodology

The structure of the proposed methodology is illustrated in Fig. 2.

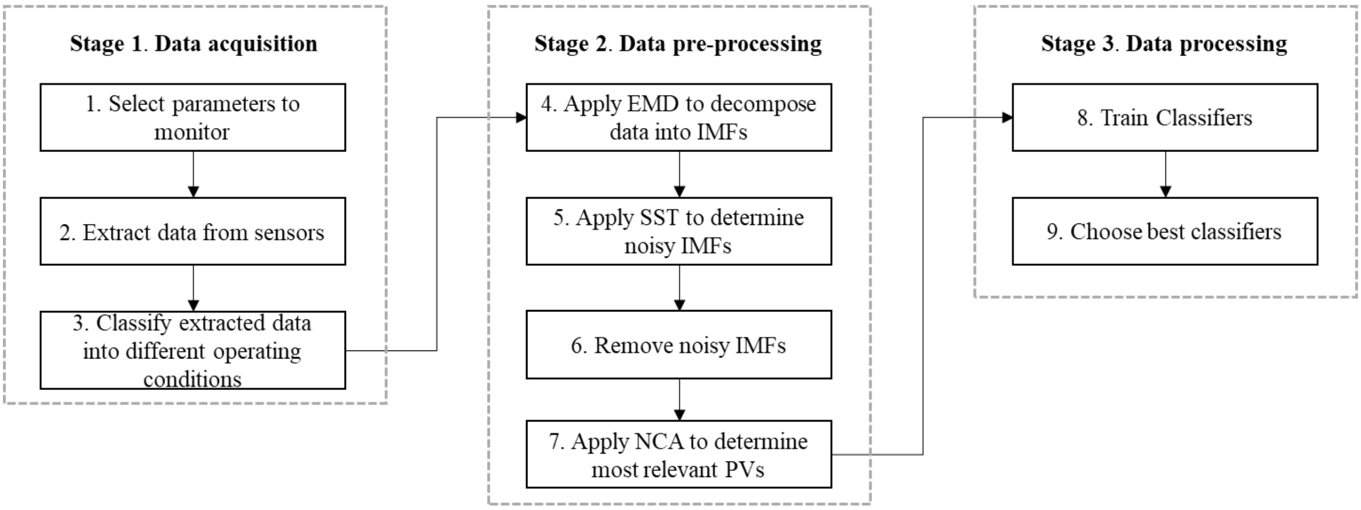

Fig. 2 Schematic representation of the steps required to perform the developed framework

3.1 Stage 1: Data acquisition

The starting stage consists of acquiring the data required to perform the failure diagnosis. First, a set of parameters is selected, and the respective sensors are installed (Step 1). Then, data are extracted from the sensors during operations (Step 2), and, finally, are classified into different operating conditions (Step 3).

3.2 Stage 2: Data preprocessing

The second stage is devoted to noise removal and data reduction. Each acquired signal is decomposed into its IMFs through EMD (Step 4). Next, each IMF goes through an SST to point out the noisy IMFs (Step 5), which are removed from the original signal (Step 6). To conclude this stage, NCA is exploited to depict the most relevant PVs of the denoised signal (Step 7).

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). Finally, the ML classification approach is tested on data not used for the training (step 9).

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 in Italy. The system is a three-stage centrifugal compressor devoted to extracting non-condensable gases. 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. 3.

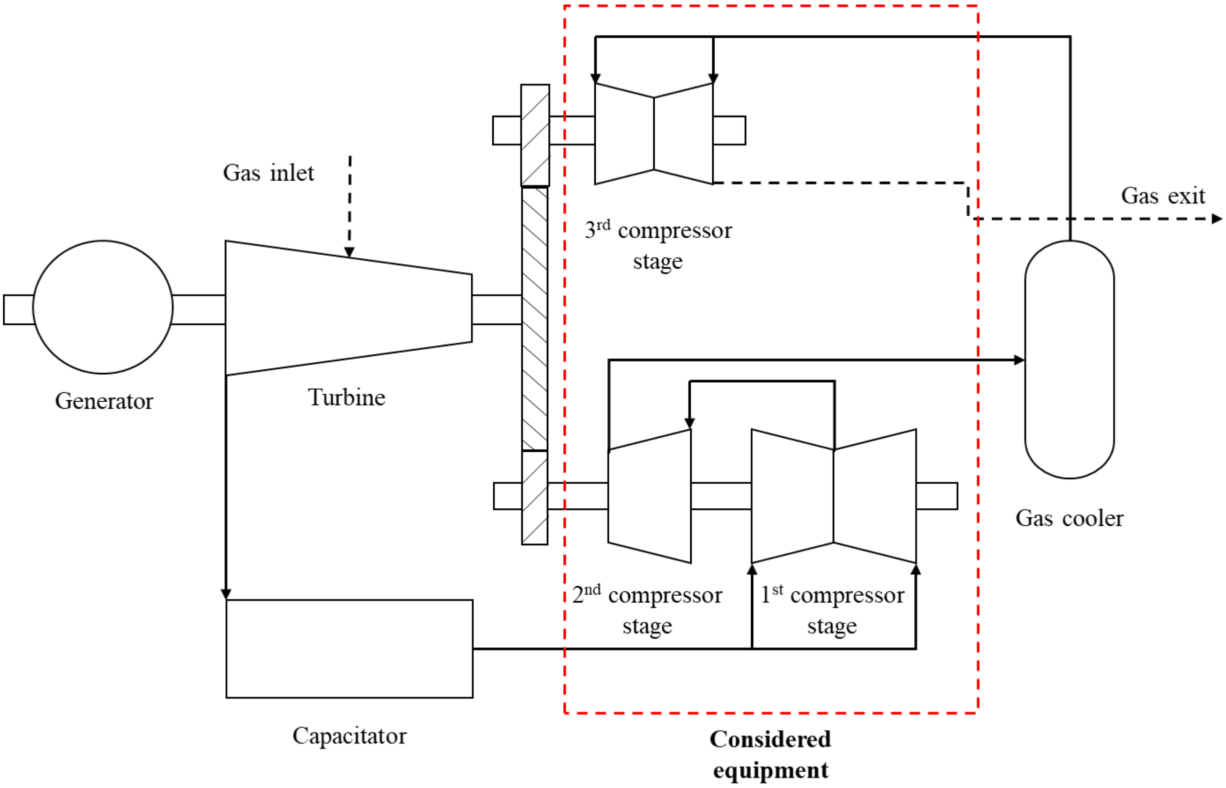

Fig. 3 Representation of the analyzed compressor within its operating system.

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 selected sensors measure either thermodynamic PVs of the elaborated fluid or relevant physical variables. 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 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 391,393 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. 4 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.

Table 1 Selected process variables

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

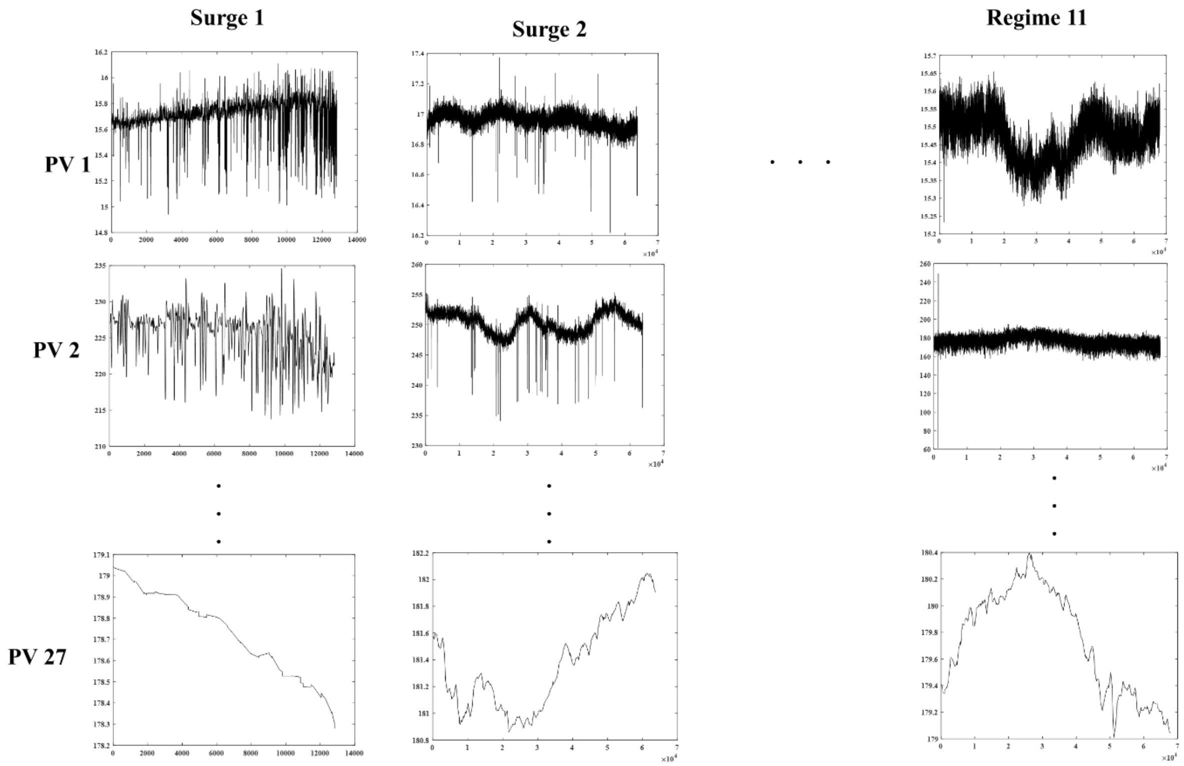

Fig. 4 Example of collected signals for distinct surge and regime events

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. The acquired data also have a strong nonstationary and nonlinear nature. 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. 9 [28].

$$T_i = \frac{n}{P_i} \tag{9}$$

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. 10 [65].

$$E_i = \frac{1}{n}\sum_{t=1}^{n} |c_i^2(t)| \tag{10}$$

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. 11, whose null hypothesis is that every IMF is a noisy IMF.

$$ln\left(\frac{1}{3}E_1\right) + lnT_1 < lnE_i + lnT_i < ln(3E_1) + lnT_1 \quad i = 2,3,\dots,m \tag{11}$$

where m is the number of IMFs. Consequently, the first IMF is consistently recognized as noise. Furthermore, all IMFs for which the null 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. 5.

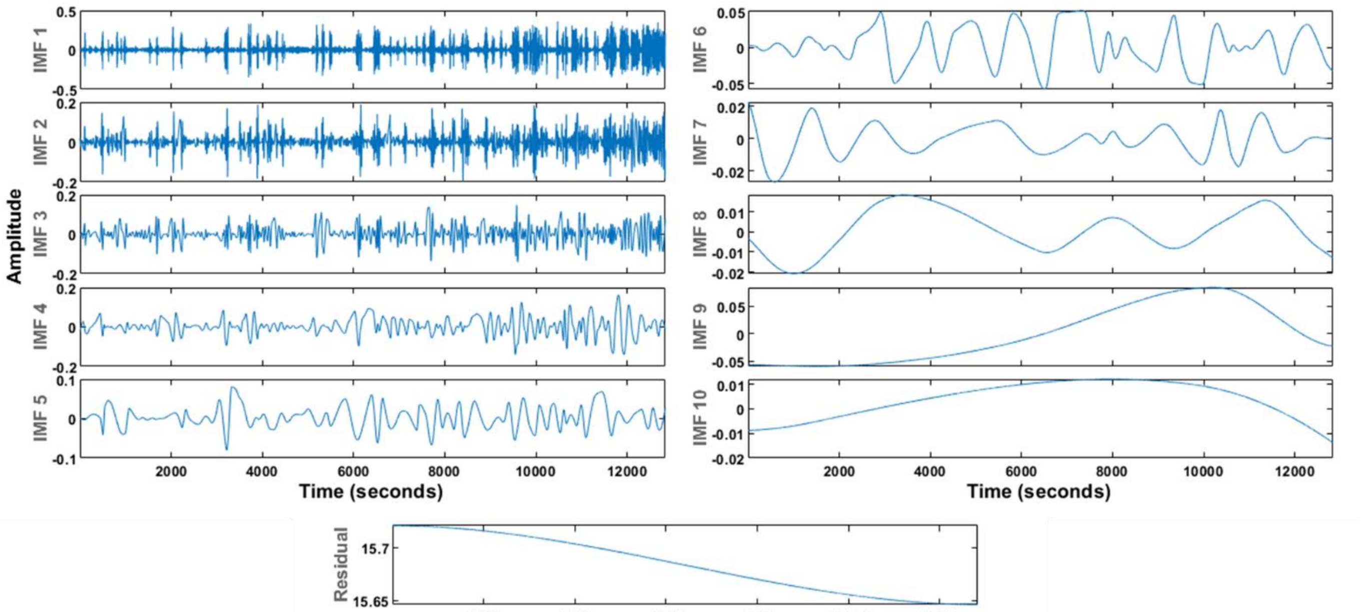

Fig. 5 Example of EMD for one of the PV monitored during a surge event.

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. 6). Finally, the denoised signal is reconstructed as the sum of the true signal IMFs and the residual, as illustrated by Eq. 12.

$$DS(t) = \sum_{i=1}^{n} c_{i,TS}(t) + r(t) \tag{12}$$

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 monitored signal and the denoised signal of the illustrated example are shown in Fig. 7.

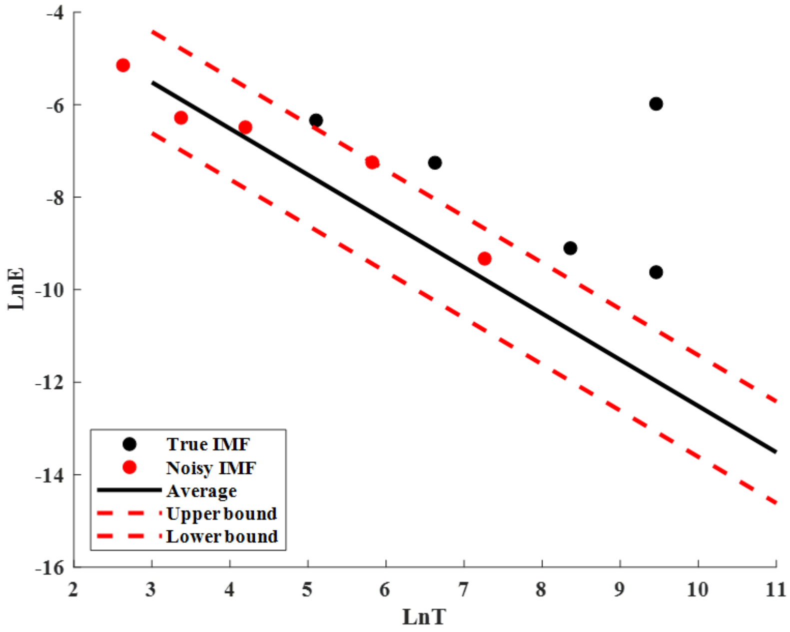

Fig. 6 Noisy and true signal IMFs for one of the PVs monitored during a surge event

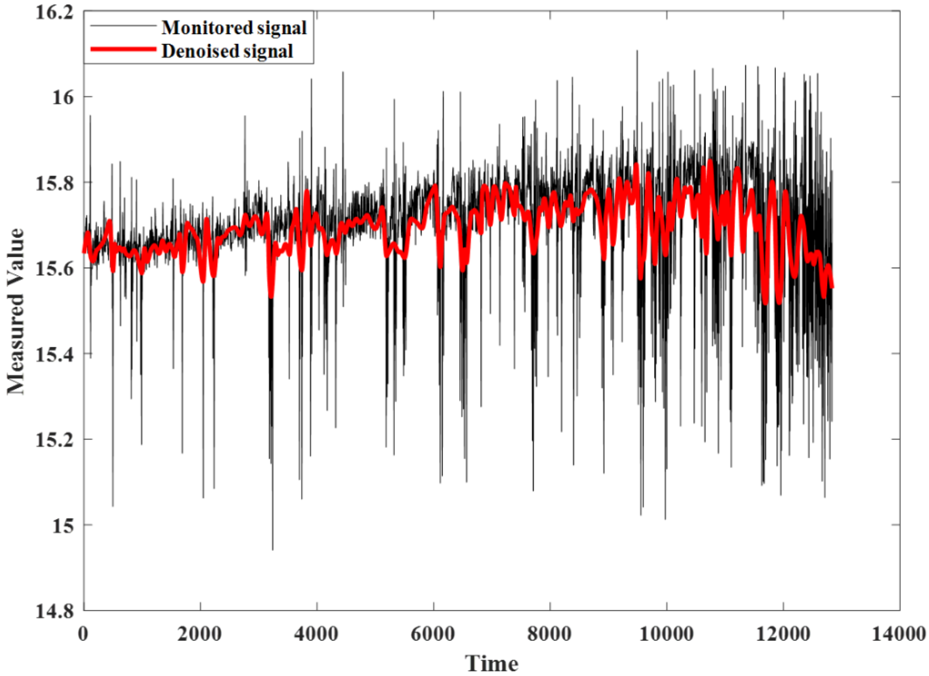

Fig. 7 Original and denoised signal of the considered PV during a surge event

The signal of the example is highly dynamic and nonstationary. 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.

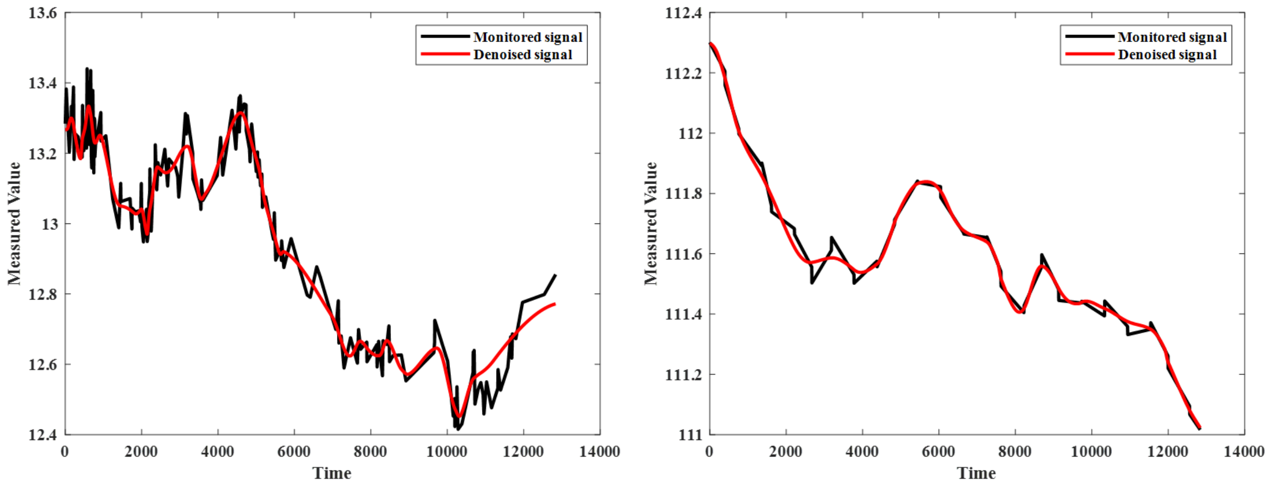

Fig. 8 Monitored and denoised signal of two PVs characterized by low fluctuations.

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 [66]. 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. The results arising from the application of the NCA are depicted in Fig. 9 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).

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

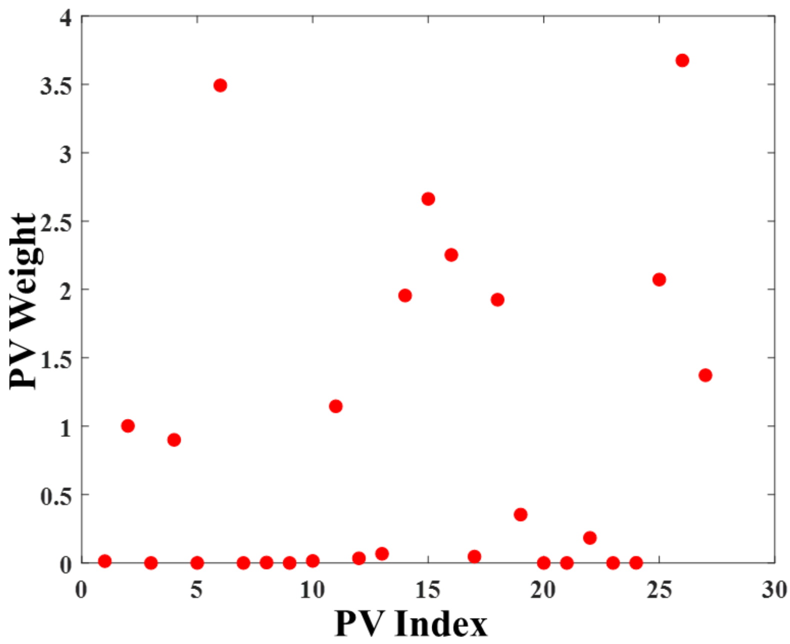

Fig. 9 Weight associated with the NCA to each PV.

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%

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 analysis steps, to reduce the time required by the calculation, especially for online monitoring purposes.

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. Furthermore, the same amount of data points was 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 were used as a 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. Moreover, since many data were available for the regime operating state, we decided to have a balanced training dataset, considering a small subset of the regime observations. This allows us to better verify the generalization capability of the regime conditions. On the other hand, since fewer data were available for the surge event, the standard proportion aforementioned between training and test was exploited.

The optimization of an ML approach was out of the scope of this work. Therefore we adopted an RF with the characteristics highlighted in Table 3.

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

The training was conducted through a 5-fold cross-validation, which resulted in the confusion matrix of Table 4. The calculation depicted that 13,585 surge observations were classified as regime, while only 2,039 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.

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

		Predicted class	
		Regime	Surge
True class	Regime	291,881	2,039
	Surge	13,585	280,335

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, which was previously mentioned. The confusion matrix related to the test set is shown in Table 5.

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	4,914	93,059

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, 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 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. 10.

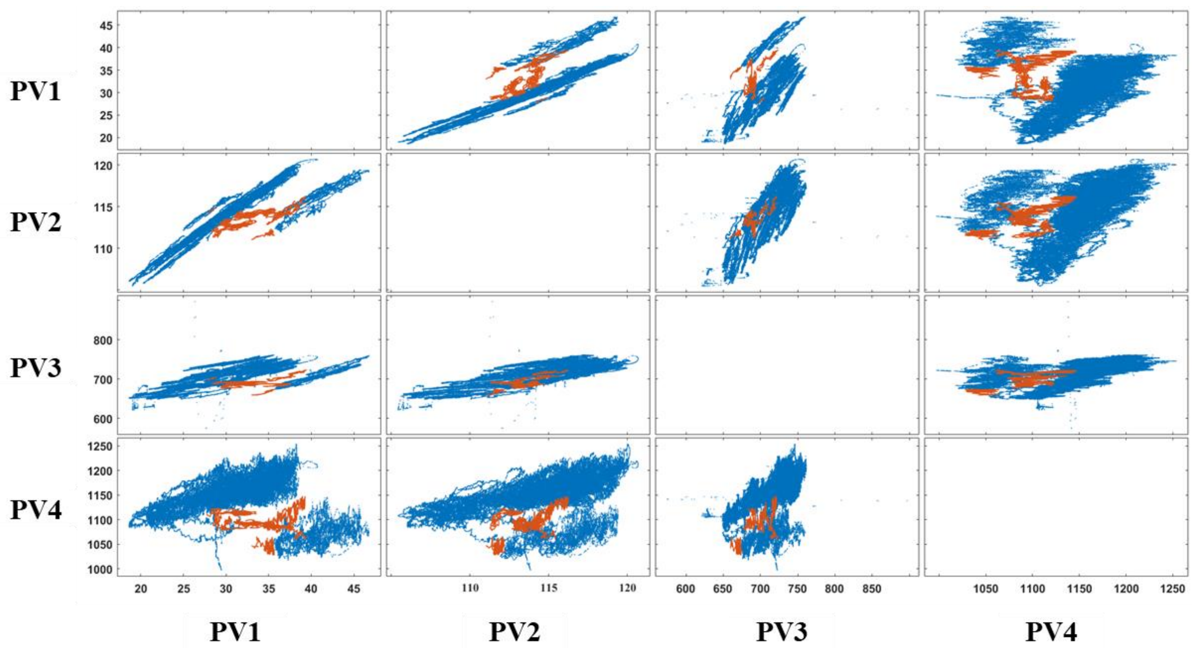

Fig. 10 Scatter plots for all four most relevant PVs. The blue and orange dots represent regime and surge observations respectively.

As depicted in Fig. 10, 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.

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

		Predicted class	
		Regime	Surge
True class	Regime	287,507	6,413
	Surge	7,215	286,705

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	2,219	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 accuracy 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 dynamic PVs. 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 assuring the safety of the operations, limiting the time that the system spends in a dangerous state (e.g., the surge).

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. Moreover, the exploitation of distinct ML techniques could be taken into account. 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.

7. References

1. Liu, Z.;L. Zhang A review of failure modes, condition monitoring and fault diagnosis methods for large-scale wind turbine bearings. *Measurement*. 2020. *149* p. 107002.
2. Schlechtingen, M.;I.F. Santos Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection. *Mechanical systems and signal processing*. 2011. *25* p. 1849-1875.
3. Wadibhasme, J.; S. Zaday;R. Somalwar. 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. IEEE.
4. Gangsar, P.;R. Tiwari Signal based condition monitoring techniques for fault detection and diagnosis of induction motors: A state-of-the-art review. *Mechanical systems and signal processing*. 2020. *144* p. 106908.
5. Toliyat, H.A.; K. Abbaszadeh; M.M. Rahimian;L.E. Olson Rail defect diagnosis using wavelet packet decomposition. *IEEE Transactions on Industry Applications*. 2003. *39* p. 1454-1461.
6. Márquez, F.P.G.a.; F. Schmid;J.C. Collado A reliability centered approach to remote condition monitoring. A railway points case study. *Reliability Engineering & System Safety*. 2003. *80* p. 33-40.
7. Zhang, W.; M.-P. Jia; L. Zhu;X.-A. Yan Comprehensive overview on computational intelligence techniques for machinery condition monitoring and fault diagnosis. *Chinese Journal of Mechanical Engineering*. 2017. *30* p. 782-795.
8. Soltanali, H.; A. Rohani; M.H. Abbaspour-Fard; A. Parida;J.T. Farinha Development of a risk-based maintenance decision making approach for automotive production line. *International Journal of System Assurance Engineering and Management*. 2020. *11* p. 236-251.

9. Ferreira, R.J.; A.T. de Almeida;C.A. Cavalcante A multi-criteria decision model to determine inspection intervals of condition monitoring based on delay time analysis. *Reliability Engineering & System Safety*. 2009. *94* p. 905-912.
10. Ilonen, J.; J.-K. Kamarainen; T. Lindh; J. Ahola; H. Kalviainen;J. Partanen Diagnosis tool for motor condition monitoring. *IEEE Transactions on Industry Applications*. 2005. *41* p. 963-971.
11. Roy, S.; N. Sinha;A.K. Sen A new hybrid image denoising method. *International Journal of Information Technology and Knowledge Management*. 2010. *2* p. 491-497.
12. Vishwakarma, M.; R. Purohit; V. Harshlata;P. Rajput Vibration analysis & condition monitoring for rotating machines: a review. *Materials Today: Proceedings*. 2017. *4* p. 2659-2664.
13. Tsolis, G.;T.D. Xenos Signal denoising using empirical mode decomposition and higher order statistics. *International Journal of Signal Processing, Image Processing and Pattern Recognition*. 2011. *4* p. 91-106.
14. Saini, K.;S. Dhami Predictive Monitoring of Incipient Faults in Rotating Machinery: A Systematic Review from Data Acquisition to Artificial Intelligence. *Archives of Computational Methods in Engineering*. 2022, p. 1-22.
15. Morozov, A.; R. Nigmatullin; G. Agrusti; P. Lino; G. Maione; Z. Kanovic;J. Martinez-Roman. 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. IEEE.
16. Gowid, S.; R. Dixon;S. Ghani A novel robust automated FFT-based segmentation and features selection algorithm for acoustic emission condition based monitoring systems. *Applied Acoustics*. 2015. *88* p. 66-74.
17. Sparis, P.;G. Vachtsevanos, *Automatic diagnostic feature generation via the Fast Fourier Transform*, Citeseer.
18. Majali, A.; A. Mulay; V. Iyengar; A. Nayak;P. Singru Fault identification and remaining useful life prediction of bearings using Poincare maps, fast Fourier transform and convolutional neural networks. *Mathematical Models in Engineering*. 2022. *8*.
19. Hussein, R.; K. BashirShaban;A.H. El-Hag Denoising of acoustic partial discharge signals corrupted with random noise. *IEEE Transactions on Dielectrics and Electrical Insulation*. 2016. *23* p. 1453-1459.
20. Zhang, C.; J. Guo; D. Zhen; H. Zhang; Z. Shi; F. Gu;A. Ball, *Rolling element bearing fault diagnosis based on the wavelet packet transform and time-delay correlation demodulation analysis*, in *Advances in Asset Management and Condition Monitoring*. 2020, Springer. p. 1195-1203.
21. Bera, A.; A. Dutta;A.K. Dhara. 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. IEEE.
22. Lopes, W.N.; P.O. Junior; P.R. Aguiar; F.A. Alexandre; F.R. Dotto; P.S. da Silva;E.C. Bianchi An efficient short-time Fourier transform algorithm for grinding wheel condition monitoring through acoustic emission. *The International Journal of Advanced Manufacturing Technology*. 2021. *113* p. 585-603.
23. Bae, S.J.; B.M. Mun; W. Chang;B. Vidakovic Condition monitoring of a steam turbine generator using wavelet spectrum based control chart. *Reliability Engineering & System Safety*. 2019. *184* p. 13-20.
24. Jiménez, A.A.; C.Q.G. Muñoz;F.P.G. Márquez Dirt and mud detection and diagnosis on a wind turbine blade employing guided waves and supervised learning classifiers. *Reliability Engineering & System Safety*. 2019. *184* p. 2-12.
25. Mousavi, A.A.; C. Zhang; S.F. Masri;G. Gholipour Structural damage detection method based on the complete ensemble empirical mode decomposition with adaptive noise: a model steel truss bridge case study. *Structural Health Monitoring*. 2021, p. 14759217211013535.
26. Huang, N.E.; Z. Shen; S.R. Long; M.C. Wu; H.H. Shih; Q. Zheng; N.-C. Yen; C.C. Tung;H.H. Liu The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*. 1998. *454* p. 903-995.
27. Desavale, R.G.; P.M. Jadhav;N.V. Dharwadkar Dynamic Response Analysis of Gearbox to Improve Fault Detection Using Empirical Mode Decomposition and Artificial Neural Network Techniques. *ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg*. 2021. *7*.

28. BahooToroody, A.; M.M. Abaei; F. BahooToroody; F. De Carlo; R. Abbassi;S. Khalaj A condition monitoring based signal filtering approach for dynamic time dependent safety assessment of natural gas distribution process. *Process Safety and Environmental Protection*. 2019. *123* p. 335-343.
29. Rafiq, H.J.; G.I. Rashed;M. Shafik Application of multivariate signal analysis in vibration-based condition monitoring of wind turbine gearbox. *International Transactions on Electrical Energy Systems*. 2021. *31* p. e12762.
30. Nishat Toma, R.; C.-H. Kim;J.-M. Kim Bearing fault classification using ensemble empirical mode decomposition and convolutional neural network. *Electronics*. 2021. *10* p. 1248.
31. Tang, Y.; Q. Liu;Q. Zhu. Fault simulation and forecast of helical cylindrical gear of reducer based on ADAMS. in Journal of Physics: Conference Series. 2021. IOP Publishing.
32. BahooToroody, A.; F. De Carlo; N. Paltrinieri; M. Tucci;P. Van Gelder Bayesian regression based condition monitoring approach for effective reliability prediction of random processes in autonomous energy supply operation. *Reliability Engineering & System Safety*. 2020. *201* p. 106966.
33. Yu, J. State of health prediction of lithium-ion batteries: Multiscale logic regression and Gaussian process regression ensemble. *Reliability Engineering & System Safety*. 2018. *174* p. 82-95.
34. Yan, G.; C. Yu;Y. Bai A New Hybrid Ensemble Deep Learning Model for Train Axle Temperature Short Term Forecasting. *Machines*. 2021. *9* p. 312.
35. Gao, Z.; Y. Liu; Q. Wang; J. Wang;Y. Luo Ensemble empirical mode decomposition energy moment entropy and enhanced long short-term memory for early fault prediction of bearing. *Measurement*. 2022. *188* p. 110417.
36. Adams, S.; R. Meekins; P.A. Beling; K. Farinholt; N. Brown; S. Polter;Q. Dong. A comparison of feature selection and feature extraction techniques for condition monitoring of a hydraulic actuator. in Annual Conference of the PHM Society. 2017.
37. Caggiano, A.; R. Angelone; F. Napolitano; L. Nele;R. Teti Dimensionality reduction of sensorial features by principal component analysis for ANN machine learning in tool condition monitoring of CFRP drilling. *Procedia CIRP*. 2018. *78* p. 307-312.
38. Ramirez-Chavez, M.; J.J. Saucedo-Dorantes; A.Y. Jaen-Cuellar; R.A.O. Rios; R. de Jesus Romero-Troncoso;M. Delgado-Prieto. 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. IEEE.
39. Gierlak, P.; A. Burghardt; D. Szybicki; M. Szuster;M. Muszyńska On-line manipulator tool condition monitoring based on vibration analysis. *Mechanical Systems and Signal Processing*. 2017. *89* p. 14-26.
40. Khalid, S.; T. Khalil;S. Nasreen. A survey of feature selection and feature extraction techniques in machine learning. in 2014 science and information conference. 2014. IEEE.
41. Goldberger, J.; G.E. Hinton; S. Roweis;R.R. Salakhutdinov Neighbourhood components analysis. *Advances in neural information processing systems*. 2004. *17*.
42. Raghu, S.;N. Sriraam Classification of focal and non-focal EEG signals using neighborhood component analysis and machine learning algorithms. *Expert Systems with Applications*. 2018. *113* p. 18-32.
43. Yaman, O. An automated faults classification method based on binary pattern and neighborhood component analysis using induction motor. *Measurement*. 2021. *168* p. 108323.
44. Zhou, H.; J. Chen; G. Dong; H. Wang;H. Yuan Bearing fault recognition method based on neighbourhood component analysis and coupled hidden Markov model. *Mechanical Systems and Signal Processing*. 2016. *66* p. 568-581.
45. Dhiman, H.S.; D. Deb; J. Carroll; V. Muresan;M.-L. Unguresan Wind turbine gearbox condition monitoring based on class of support vector regression models and residual analysis. *Sensors*. 2020. *20* p. 6742.
46. Murphy, K.P., *Machine learning: a probabilistic perspective*, MIT press,2012.
47. Islam, M.M.;J.-M. Kim Reliable multiple combined fault diagnosis of bearings using heterogeneous feature models and multiclass support vector Machines. *Reliability Engineering & System Safety*. 2019. *184* p. 55-66.
48. Tang, T.;H. Yuan A hybrid approach based on decomposition algorithm and neural network for remaining useful life prediction of lithium-ion battery. *Reliability Engineering & System Safety*. 2022. *217* p. 108082.

49. Lipinski, P.; E. Brzychczy;R. Zimroz Decision tree-based classification for Planetary Gearboxes' condition monitoring with the use of vibration data in multidimensional symptom space. *Sensors*. 2020. *20* p. 5979.
50. Patel, R.K.;V. Giri Feature selection and classification of mechanical fault of an induction motor using random forest classifier. *Perspectives in Science*. 2016. *8* p. 334-337.
51. Saeed, U.; S.U. Jan; Y.-D. Lee;I. Koo Fault diagnosis based on extremely randomized trees in wireless sensor networks. *Reliability Engineering & System Safety*. 2021. *205* p. 107284.
52. Zhang, C.; D. Hu;T. Yang Anomaly detection and diagnosis for wind turbines using long short-term memory-based stacked denoising autoencoders and XGBoost. *Reliability Engineering & System Safety*. 2022. *222* p. 108445.
53. Wan, S.; X. Li; Y. Zhang; S. Liu; J. Hong;D. Wang Bearing Remaining Useful Life Prediction with Convolutional Long Short-Term Memory Fusion Networks. *Reliability Engineering & System Safety*. 2022, p. 108528.
54. Azar, K.; Z. Hajiakhondi-Meybodi;F. Naderkhani Semi-supervised clustering-based method for fault diagnosis and prognosis: A case study. *Reliability Engineering & System Safety*. 2022. *222* p. 108405.
55. Zhu, Y.; J. Wu; J. Wu;S. Liu Dimensionality reduce-based for remaining useful life prediction of machining tools with multisensor fusion. *Reliability Engineering & System Safety*. 2022. *218* p. 108179.
56. Xu, F.; F. Yang; Z. Fei; Z. Huang;K.-L. Tsui Life prediction of lithium-ion batteries based on stacked denoising autoencoders. *Reliability Engineering & System Safety*. 2021. *208* p. 107396.
57. Küppers, F.; J. Albers;A. Haselhoff. Random Forest on an Embedded Device for Real-time Machine State Classification. in 2019 27th European Signal Processing Conference (EUSIPCO). 2019. IEEE.
58. Ahn, H.; H. Moon; M.J. Fazzari; N. Lim; J.J. Chen;R.L. Kodell Classification by ensembles from random partitions of high-dimensional data. *Computational Statistics & Data Analysis*. 2007. *51* p. 6166-6179.
59. Karatoprak, E.;S. Seker An improved empirical mode decomposition method using variable window median filter for early fault detection in electric motors. *Mathematical Problems in Engineering*. 2019. *2019*.
60. Boudraa, A.; J. Cexus;Z. Saidi EMD-based signal noise reduction. *International Journal of Signal Processing*. 2004. *1* p. 33-37.
61. Kumar, V.;S. Minz Feature selection: a literature review. *SmartCR*. 2014. *4* p. 211-229.
62. Yang, W.; K. Wang;W. Zuo Neighborhood component feature selection for high-dimensional data. *J. Comput.* 2012. *7* p. 161-168.
63. Belgiu, M.;L. Drăguţ Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*. 2016. *114* p. 24-31.
64. Bonissone, P.; J.M. Cadenas; M.C. Garrido;R.A. Díaz-Valladares A fuzzy random forest. *International Journal of Approximate Reasoning*. 2010. *51* p. 729-747.
65. Wu, Z.;N.E. Huang A study of the characteristics of white noise using the empirical mode decomposition method. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*. 2004. *460* p. 1597-1611.
66. Rahman, M.M.;D.N. Davis Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*. 2013. *3* p. 224.