Journal of Petroleum Science and Engineering 205 (2021) 108845

Contents lists available at ScienceDirect

Journal of Petroleum Science and Engineering

journal homepage: http://www.elsevier.com/locate/petrol

A data-driven pipeline pressure procedure for remote monitoring of centrifugal pumps

Riccardo Angelo Giro [a,*], Giancarlo Bernasconi [a], Giuseppe Giunta [b], Simone Cesari [b]

[a] Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133, Milano, Italy
[b] Eni S.p.A., San Donato, Milanese, Italy

ARTICLE INFO

Keywords:
Integrity monitoring
Pipeline integrity
Predictive maintenance
Pump failure diagnosis
Anomaly detection
Unsupervised learning

ABSTRACT

Pumping systems are a key component of oil and gas pipeline transportation assets: monitoring their integrity is a crucial operation from a safety and revenue point of view. The solutions currently employed in the industry apply supervised machine learning techniques to data collected by multi-domain sensors directly installed on several positions of the pump itself; however, such approaches are not applicable on older machines, in contexts where a direct access to the pump is not possible, or whenever labelled data are not at disposal. This paper, instead, presents a predictive maintenance strategy where the condition of a centrifugal pump is tracked by solely exploiting standard pressure measurements, recorded also on remote points along the pipeline, and using an unsupervised learning approach. The smart monitoring strategy is presented and validated on historical pressure signals collected by Eni for several years on a crude oil transportation pipeline, located in Italy. Pressure data, recorded along the fluid line, are used to compute several statistical indicators on appropriate window lengths. These indicators are then fed to an unsupervised clustering procedure, based on a Gaussian mixture model. The output is an index within four different pump operational regimes, and a clustering visualization that permits the interpretation of the automatic regime classification. In fact, the manual inspection of the clusters shows that three of them describe standard modes (regular pumping operation, pumps off, flow regulations). The fourth one corresponds to high amplitude peaks in the signals and indicators, and so it is tagged as "anomalous" mode: pump maintenance logs reveal that the peaks are associated to damaged roller bearing movements, which disappear after the activation of the pump backup system. Anomalies are reported several days before the pump switch, so that a preventive maintenance could have been triggered. The robustness of the clustering algorithm is assessed on a statistical basis, whereas the overall validity of the monitoring system is tested on an instantaneous basis by applying the proposed model on two independent datasets, collected on real transportation pipelines: the results demonstrate the reliability of the proposed monitoring strategy in predicting and detecting all the pump failure events reported by the available maintenance logs. With respect to the mostly employed approaches, our machine learning procedure does not require any previous supervision of the data. Moreover, input data are the pressure transients produced by the pumps and guided within the fluid in the pipeline for long distances: pump failure analysis can be run using sensors located at several kilometers of distance from the pump itself, making a remote control strategy feasible.

1. Introduction

Safety, efficiency and sustainability are key points in modern oil & gas pipeline transportation systems: in such a context, the development of effective, real-time integrity monitoring strategies acquires critical importance. However, pipeline networks are characterized by a high degree of complexity, making the satisfaction of the above-mentioned requirements a challenging task. The complexity increases when also considering the variability of the transported fluids, spanning from single products (oil, natural gas, refined products) to multiphase mixtures of gases, liquids and solids. Moreover, recently, big data technologies have entered the oil & gas industry (Koroteev and Tekic, 2021): this phenomenon brings the need of an efficient digital transformation process (Hajizadeh, 2019), coupled with new methods to manage ever-increasing databases.

Nowadays, modern assets typically employ a network of sensing

* Corresponding author.
E-mail address: riccardoangelo.giro@polimi.it (R.A. Giro).

https://doi.org/10.1016/j.petrol.2021.108845
Received 13 November 2020; Received in revised form 18 March 2021; Accepted 18 April 2021
Available online 22 April 2021

devices, that collect multi-domain data (e.g., pressure, acceleration, flow rate, density, etc.), in real-time, during the entire lifespan of the assets themselves. The large availability of such resources can be exploited for predictive analytics purposes. The synergic combination of machine learning (ML) methods and large, multi-domain datasets represents the current technological frontier: on one side, machine learning techniques allow to uncover hidden relations from the input data; on the other side, the complementarity of multi domain measurements permits to broaden the panorama of the observable operational regimes of the pipeline transportation system.

Focusing the attention to the pumps, which is the topic of this paper, a solid body of research has been conducted on monitoring their health using machine learning techniques. Certain solutions have been proved on controlled scenarios, such as experimental setups arranged in a laboratory, thus lacking a validation phase on real case studies (Alabied et al., 2019; Hailong et al., 2020; Soylemezoglu et al., 2011). Another set of approaches (Barrios Castellanos, Serpa, Biazussi, Monte Verde, & do Socorro Dias Arrifano Sassim, 2020; Chakravarthy et al., 2019; Orrù et al., 2020; Peng et al., 2020; Sharma and Pandey, 2016) has been successfully employed on real scenarios, yet they all require the use of multiple sensors (e.g., temperature, flow, current, etc.) directly installed on the pump itself: such techniques are then only viable for modern pumping systems, since older machines are typically equipped with just pressure sensors and would require additional instrumentation. Moreover, such direct monitoring approaches become unfeasible whenever a pump upgrade is cumbersome (e.g., offshore or submarine installations). Kalmár and Hegedűs (2019) proposed a pump monitoring technique based solely on pressure data: however, their method is devoted to the detection of cavitation, with a sensor setup directly arranged on the pump itself. Lastly, multiple data-driven approaches (Deng et al., 2019; Dutta et al., 2018; Van Rensburg, Kamin and Davis, 2019; Marins et al., 2020; Rauber et al., 2017; Panda et al., 2018) make use of supervised learning techniques, whose applicability is limited to instances in which manually labelled data are available: that is rarely the case in pipeline transportation systems (Lygren et al., 2019).

According to the authors' knowledge, unsupervised remote monitoring strategies using pressure signals have not been addressed in the literature yet. This paper presents a novel data-driven methodology for automated remote monitoring of centrifugal pumps by exploiting standard pressure measurements, recorded also on remote points along the pipeline and using an unsupervised learning approach, based on a Gaussian mixture model (GMM). In fact, pressure fluctuations produced by the pumps propagate as guided waves within the fluid itself and are measured by pressure sensors located in several positions on the line. These signals bring information about the source condition (i.e., the pumps) and can be processed with data-driven techniques to define a reference, healthy status. New vibroacoustic data are mapped in real-time within the statuses domain, and potential anomalies can be identified by measuring the distance between the observed operational point and the reference one.

The study falls within the framework of a research project conducted by Eni, whose main objective is the design of machine learning techniques to monitor the integrity of pipeline transportation assets. In such a context, accurately tracking the residual lifetime of pumps and/or promptly raising alerts when a failure is approaching represent key operations to preserve the integrity of the asset itself: this aspect is particularly true when transferring crude oil (whose high viscosity induces additional wear on pumping machinery) and whenever the fluids are conveyed in batches (as pumps are subject to additional fatigue).

The new monitoring strategy is presented and validated on a case history, using pressure data collected for more than 5 years at the pumping station of an oil pipeline, located in northern Italy: so, a second contribution of the work is the capacity of adding value to historical measurements, as they are reprocessed to tune the parameters and increase the performance of current integrity monitoring techniques (Giunta et al., 2020).

The paper is organized as follows. Section II presents the instrumental setup of the case history used in this work. Section III outlines the use of machine learning approaches to derive data-driven models, such as the one presented here. Section IV provides an overview on how the monitoring system has been designed, while Section V describes the related development phases in detail. Lastly, Section VI shows the applications of the method to the case history data. Section VII draws the conclusions.

2. Case study description

The data used to develop the automated pump monitoring procedure are collected by a proprietary multi-point vibroacoustic monitoring system, namely e-vpms® (Giunta et al., 2015). It has been installed on the oil transportation pipeline connecting the Eni R&M Logistic terminals of Chivasso (Turin) and Pollein (Aosta), both located in Northern Italy (satellite map in Fig. 1). The pipeline has a length of approximately 100 km, has an internal diameter of 16 inches. There are two pumping units, but only one at a time is operated, as the other serves as backup. The pressure of the transported fluid ranges from a maximum of 120 bar at the pumping station of Chivasso, to a minimum of 4 bar at the arriving terminal, with an average flow rate of 400 m^3/h.

The vibroacoustic monitoring setup consists in a set of five permanent stations located in different points along the pipeline: in particular, two of them are situated at the line terminals, Chivasso and Pollein, respectively A and E in Fig. 1. The remaining ones are distributed along the pipe path (B, C and D in Fig. 1). The distance between each station and the pumping terminal in Chivasso is reported in Table 1, while the technical specifications of the pump itself are listed in Table 2.

Each data recording station is equipped with the following instrumentation:

- A dynamic hydrophone, which measures small-scale pressure variations within the fluid. These transients propagate along the whole conduit in the form of acoustic waves;
- A static hydrophone, which provides the absolute pressure of the transported fluid.

Data sampling rate is 20 Hz. The technical specifications of the sensors are reported in Table 3.

The vibroacoustic signals propagating within the line are generated by any interaction with the pipeline and with the flowing fluid, such as the noise produced by flow regulation machinery (i.e., pumps, valves,

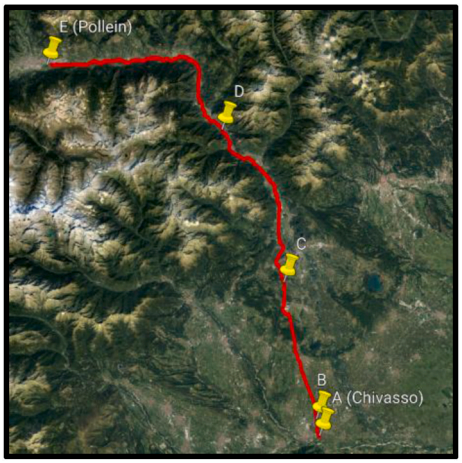

Fig. 1. Satellite map of Chivasso-Pollein crude oil pipeline route (red line) and position of the permanent stations (yellow pins). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 1
Distance between each e-vpms® station and the pumping terminal located in Chivasso.

Station	Distance with respect to station A (km)
A (Chivasso)	0
B	3.165
C	27.379
D	59.307
E (Pollein)	100.486

Table 2
Technical specifications of the centrifugal pump located in Chivasso.

Technical specification	Value	Measurement unit
Flow rate	430	m3/h
Operational density	0.7–0.86	kg/m3
Operational temperature	5–30	°C
Hydraulic head	875	mH2O
Absorbed power	1500	kW

Table 3
Technical specifications of the hydrophones installed on Chivasso-Pollein pipeline.

Measurement instrument	Technical specification	Value	Measurement unit
Dynamic hydrophone	Measurement range	344.8	kPa
	Measurement uncertainty	0.003	kPa
Static hydrophone	Measurement range	100	bar
	Measurement uncertainty	0.25% of full-scale range	bar

metering systems), turbulences within the transported product (i.e., bubbles, cavitation), human activities (i.e., pigging, spilling or maintenance operations) and natural sources (i.e., landslides, quakes, pipe expansions due to temperature increases). In this particular case, we focus the analysis to the pressure transients generated by the pumps, looking for characteristic features able to classify different pumping regimes, comprising anomalous conditions.

It should also be noted that vibrations generated during pumping operations are partly characterized by a periodic behavior in time, since they result from regular motion of mechanical elements: therefore, it is expected to observe repetitive patterns in pressure measurements related to faulty pumping machinery, too.

3. Machine learning approach for pump failure and operational status detection

3.1. Motivation

In recent years, the joint adoption of sensor data analytics and ML techniques for failure detection in the oil industry has increased significantly (Tejedor et al., 2017). This aspect is due to a number of reasons: firstly, such approaches have proven their greater effectiveness, compared to conventional, non-data driven solutions (Giunta et al., 2019a, 2019b); secondly, the development of novel technologies specific to oil and gas transportation systems (i.e., multi-point acoustic sensing) pushes towards the integration of multi-domain measurements and brings the necessity to find constitutive relations among large amounts of data; lastly, it is not always possible to devise physical, deterministic models that can accurately describe the behavior of pipeline networks, since the number of variables and parameters involved in such systems is typically very large.

Given all the above considerations, the necessity to design predictive maintenance and smart, automated detection methods becomes a critical matter: this way, one can overcome the limitations of typical solutions used to avoid machine breakdown, pertaining to reactive and preventive approaches. In the former case, a machine is used to its limit and it is repaired only after the failure has occurred, which can be very costly and also problematic from a safety point of view; the latter procedure tends to be very conservative, especially for critical equipment, and excessively frequent maintenance operations waste machine life that is actually still useable.

Lastly, the improvement introduced by ML techniques for pipeline integrity monitoring can be found both in enhanced production efficiency and reduced maintenance and repair costs (Bangari et al., 2019).

3.2. Theoretical framework

The basic principle of ML consists in the observation, for an adequate time interval, of a set of descriptors (features), which are directly extracted from the available measurements. Such features are then used to experimentally derive a set of constitutive relations between the observed data, in order to classify them into different categories and/or to uncover hidden regularities (patterns) that cannot be devised in a straightforward manner.

ML-based methods find a broad applicability range in pattern recognition problems. Generally speaking, the latter can be divided in two major categories, namely supervised and unsupervised learning, depending on the availability of labelled data: these are defined as a set of samples which have been manually tagged by a human expert with a meaningful value (label).

Whenever labelled data are at disposal, it is possible to apply any supervised learning technique. Regarding pipeline monitoring systems, this is a very sporadic case: as a matter of fact, the acquisition of properly labelled data in such instances is almost invariably a problematic task, partly due to the asset operability and partly to confidentiality concerns. For these reasons, one should resort to unsupervised learning techniques (Lygren et al., 2019).

The purpose of unsupervised learning is to identify undetected patterns within the available, unlabeled data, possibly allowing to model a set of probability densities based on the inputs. The latter procedure represents a critical step in clustering analysis, where the goal is to appropriately group common data based on a similarity measure (Giunta et al., 2019a, 2019b). Every time a new data sequence is given as input to an unsupervised learning algorithm, the latter indicates the presence/absence of commonalities for each data point, with respect to the historical ones previously analyzed: this way, one can detect which samples tend to fit in a certain group instead of another. Consequently, data pertaining to the same cluster will exhibit a higher similarity measure, compared to the ones falling into other clusters. As a last step, the results obtained are given a meaningful interpretation, based on how the various clusters have been associated to real instances of the problem taken into due account.

4. Data-driven methodology for integrity monitoring

The primary goal of this work consists in developing an automated procedure that can establish whether the pumping equipment of Chivasso-Pollein pipeline is operating adequately or not, exploiting the static and dynamic pressure measurements recorded at Chivasso station. Given the framework described in Section III.B, we are dealing with a data-driven modeling problem that can only be addressed using unsupervised learning methods, since no historical, labelled data are at disposal. In such a context, integrity monitoring and failure diagnosis can be globally framed in a three-step process:

1. Determination of the current operational status of the pump (i.e., on, off, etc.);

2. Detection of potentially anomalous instances, which deviate significantly from the expected behavior of the equipment in that particular status;
3. In affirmative case, real-time identification of the fault occurrence.

The approach presented here for monitoring pumping machinery consists in listening (by means of hydrophones) to the acoustic vibrations emitted by the pump itself and propagating within the fluid; then, a set of meaningful descriptors is extracted from the recorded signals (i.e., pressure standard deviation, kurtosis, etc.); lastly, a relation with the observed behaviors of the equipment is established. The latter operation is achieved by jointly conducting clustering analysis on historical data, using an unsupervised learning method, and by having the provided outcomes manually interpreted by an expert in the field.

4.1. Guidelines for system design

The designed pump monitoring system is based upon clustering the pressure measurements previously outlined, using a Gaussian mixture model (Reynolds, 2009). This unsupervised learning method allows to determine a set of K regions in an N -dimensional feature space (where N represents the number of descriptors extracted from the training data) and, for each sample, permits to compute the likelihood of that sample to belong to a specific cluster $\bar{k} \in \{1,\ldots,K\}$. The process is known as soft clustering, where each object is associated to every cluster according to a probability measure. The operative methodology is articulated as follows:

1. Removal of data points corresponding to not plausible measurements. In this case, static pressures lower than 0.5 bar and higher than 120 bar are discarded; likewise, dynamic pressure values below -200 kPa and above 200 kPa are eliminated from the training set. These outliers are due to rare electromagnetic disturbances affecting the power unit of the measuring station;
2. Selection of relevant features, evaluated over 1-min time windows, from a statistical analysis of the raw data. For this specific task, a detailed study of several indicators has been performed, arriving to the following set: variance and kurtosis of dynamic pressure; mean value, time gradient and variance of static pressure. Additional information about the choice of such features is provided in Section V;
3. Features normalization;
4. Clustering analysis using a GMM, providing the normalized features as input. The model is trained to learn the set of parameters $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{C}_k\}$ for each of the K Gaussian components ($\boldsymbol{\mu}_k$ and $\boldsymbol{C}_k$ respectively represent the first two statistical moments of the k -th, N -dimensional Gaussian probability density function, with $k \in \{1,\ldots, K\}$), such that the most appropriate label/cluster for every training data point can be determined. As a result, each input example $\boldsymbol{x}^{(i)}$ is given a posterior probability of belonging to any of the K clusters;
5. Association of each clustering region with a corresponding operational status of the pump;
6. Testing the trained GMM on another monitoring period of interest. For each data point $\boldsymbol{x}'(i)$ of the testing set, the model returns the likelihood of that sample to fall into one of the K clusters. As a result, the i -th example $\boldsymbol{x}'(i)$ is assigned to a specific cluster $\bar{k}$ if the corresponding a posteriori probability $p_{i\bar{k}} = P(\boldsymbol{x}'^{(i)} \in \bar{k} \,|\, \boldsymbol{x}'^{(i)})$ is the highest among all the others ($p_{i\bar{k}} > p_{ik}$, where $k \neq \bar{k}$ and $k \in \{1,\ldots,K\}$).

5. Smart pump monitoring system

This section presents the application of the pump monitoring system on the real dataset recorded along the Chivasso-Pollein pipeline transportation system, following the procedure described in Section IV.A.

5.1. Data preprocessing and analysis

Data cleansing represents an essential, preparatory operation in machine learning problems. Fig. 2 displays a subset of the dynamic and static pressure time series used to train the GMM (after the removal of faulty data points, as mentioned in Section IV. A). The entire training set corresponds to the vibroacoustic data recorded at Chivasso station from January 1st, 2010 to February 1st, 2015. A preliminary analysis of the data reveals four main scenarios, where the first three fall within the physiological behavior of the pump:

1. Steady state: the transportation system is at rest (pumps are off). This condition is characterized by small values of dynamic pressure (close to 0 kPa), while static pressure falls below 40 bar;
2. Flow regulations: pumps are turning on/off, pressure transients due to pumping fluctuations can be observed by the presence of sharp static pressure variations, coupled by dynamic pressure readings well over $\pm$ 20 kPa;
3. Pumps on, regular functioning: while in this status, the measured dynamic pressure usually ranges between $\pm$ 7 kPa, whereas the static is steadily above 50 bar. An example is displayed in Fig. 3, corresponding to the time frame between January 19th, 2015 and January 21st, 2015;
4. Pumps on, anomalous behavior: this instance is visually identified by observing the presence of periodic negative peaks in dynamic pressure measurements, accompanied by sporadic, positive ones; this aspect also affects static pressure readings (even though they are less noticeable), which are characterized by small, repetitive jumps: these signs are typically related to system part failure. As an example, Fig. 4 illustrates the fault of a roller bearing which occurred on January 7th, 2015.

5.2. Features selection

A set of relevant descriptors then needs to be extracted from the training data, in order to build a GMM that can recognize the previously mentioned scenarios accordingly. To this purpose, the following statistical indicators have been evaluated over 1-min intervals:

1) Dynamic pressure variance: this quantity can be interpreted as the amount of acoustic energy within the fluid. Low values ($<10^{-2}$ kPa^2) indicate that the transportation system is at rest, whereas a very high variance ($>10^2$ kPa^2) is related to flow regulation operations. Intermediate figures (between 1 and 10 kPa^2), instead, are referred to pumping equipment in operation;
2) Dynamic pressure kurtosis: defined as the fourth standardized moment of a random variable (in this case, raw dynamic pressure), it measures the spikiness of an input probability distribution and can be interpreted as to what degree such shape differs from that of a

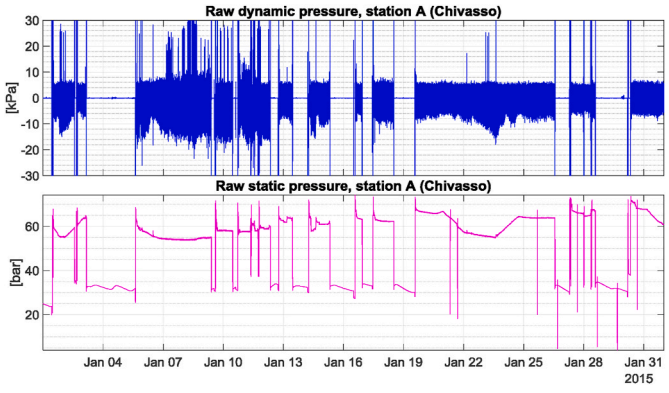

Fig. 2. Subset of the raw pressure time series used to train the GMM.

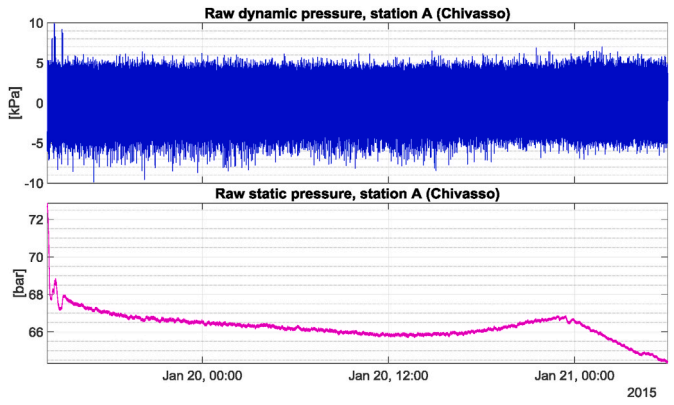

Fig. 3. Regular functioning of the pumping machinery between January 19th, 2015 and January 21st, 2015.

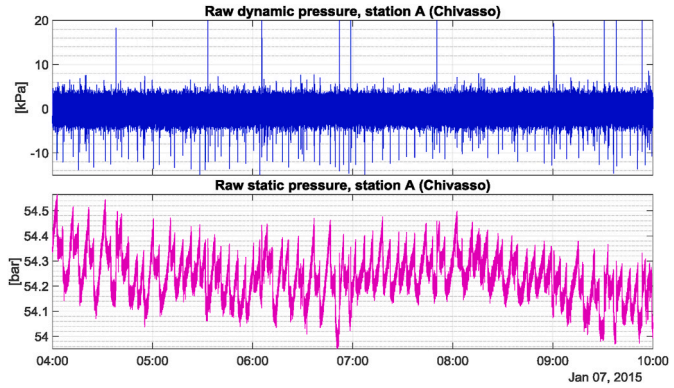

Fig. 4. Detail of a pump failure event occurred on January 7th, 2015.

normal distribution. It can be observed that, during normal pump functioning, the measured kurtosis is around 3 (Fig. 5, second plot from the top, red horizontal line), as for a normal distribution. Increases in kurtosis manifest the occurrence of undesired peaks in the measured signal which, in this case, are indicative of an anomalous pump behavior. Moreover, when kurtosis values are greater than 3, the trend of pressure fluctuations loses its typical symmetry, as the signal is no longer normally distributed. This aspect can be visually inferred by observing in Fig. 4 (top) the irregular evolution of raw dynamic pressure, as a function of time;

3) Average static pressure: tracking the absolute pressure within the fluid provides an overall information on the pipeline's operational regime, as explained in Section V.A;
4) Time gradient of static pressure: the presence of flow regulations corresponds to measured absolute values of such quantity greater than 1 bar; in the remaining cases, the system is either functioning or at rest;
5) Static pressure variance: the amount of energy related to absolute pressure measurements is tightly correlated to pumping fluctuations due to flow regulations. Increasing values (greater than 3 bar^2) indicate the presence of such events.

The time series of the raw features are shown in Fig. 5, extracted from the training subset illustrated in Fig. 2, whereas Fig. 6 displays a set of density plots between pairs of features (lower triangular part of the figure), along with the probability distribution of each descriptor (plotted on the main diagonal). At first sight, the presence of two distinct hotspots in several of the cross plots can be immediately noticed: for instance, by plotting the dynamic pressure variance against its kurtosis (Fig. 6, density plot located on the second row, first column) one can

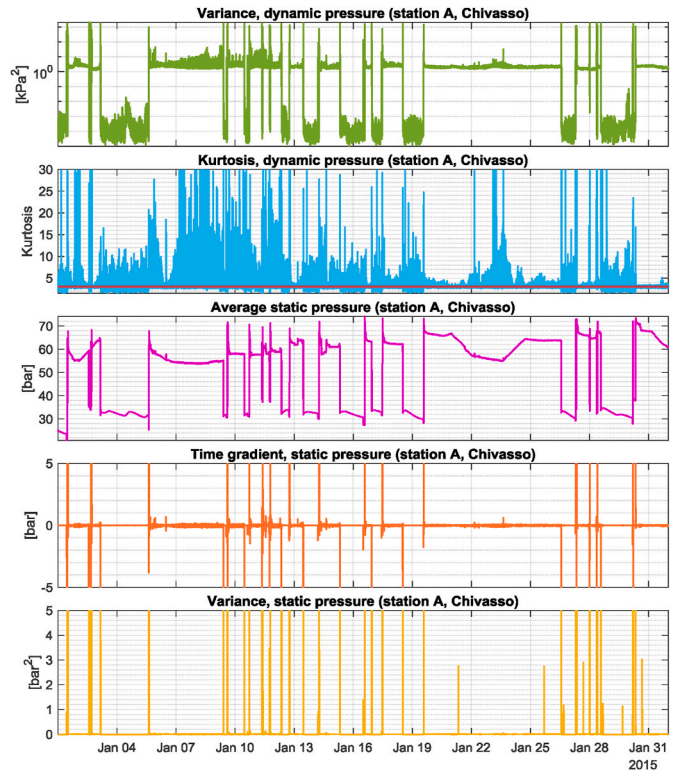

Fig. 5. Time series of the raw features, extracted from the training subset illustrated in Fig. 2.

observe both steady state (variance below 10^{-2} kPa^2, kurtosis between 2 and 6) and regular functioning regions (variance greater than 1 kPa^2, kurtosis around 3).

A more accurate observation, though, unveils additional information: considering the same chart, for variance and kurtosis values respectively larger than 1 kPa^2 and 3, another distinct cloud of points can be observed. The latter can be interpreted as the instances in which the pump is presenting an anomalous behavior, which correspond to the fault cases that need to be identified by the detection system. This last example shows that, even though certain events tend to occur rather infrequently in pipeline transportation systems (i.e., pump failure), a thorough data visualization process proves to be advantageous in revealing which features provide meaningful information for the problem considered and which do not.

5.3. Features normalization and clustering analysis

A typical requirement for several machine learning algorithms consists in standardizing the features given as input to the algorithm itself: in this way, data coming from multiple dimensions and having different scales and units are given a uniform reference. This operation proves to be valuable in cluster analysis problems, where similarities between different features are evaluated using a certain distance metric.

In our case, it was chosen to normalize the features such that each one of them has zero mean and unit variance, a process widely known as Z-score Normalization (Kreyszig, 1983). Considering an $m \times N$ feature matrix $\boldsymbol{X}$, where m and N respectively represent the number of training examples and features, the $m \times 1$ normalized vector z_n corresponding to the n -th feature is computed as:

$$\boldsymbol{z}_n = \frac{1}{\sigma_n}(\boldsymbol{x}_n - \mu_n), \tag{1}$$

where the $m \times 1$ vector $\boldsymbol{x}_n$ corresponds to the n -th column of $\boldsymbol{X}$, is characterized by mean value μ_n and has standard deviation σ_n.

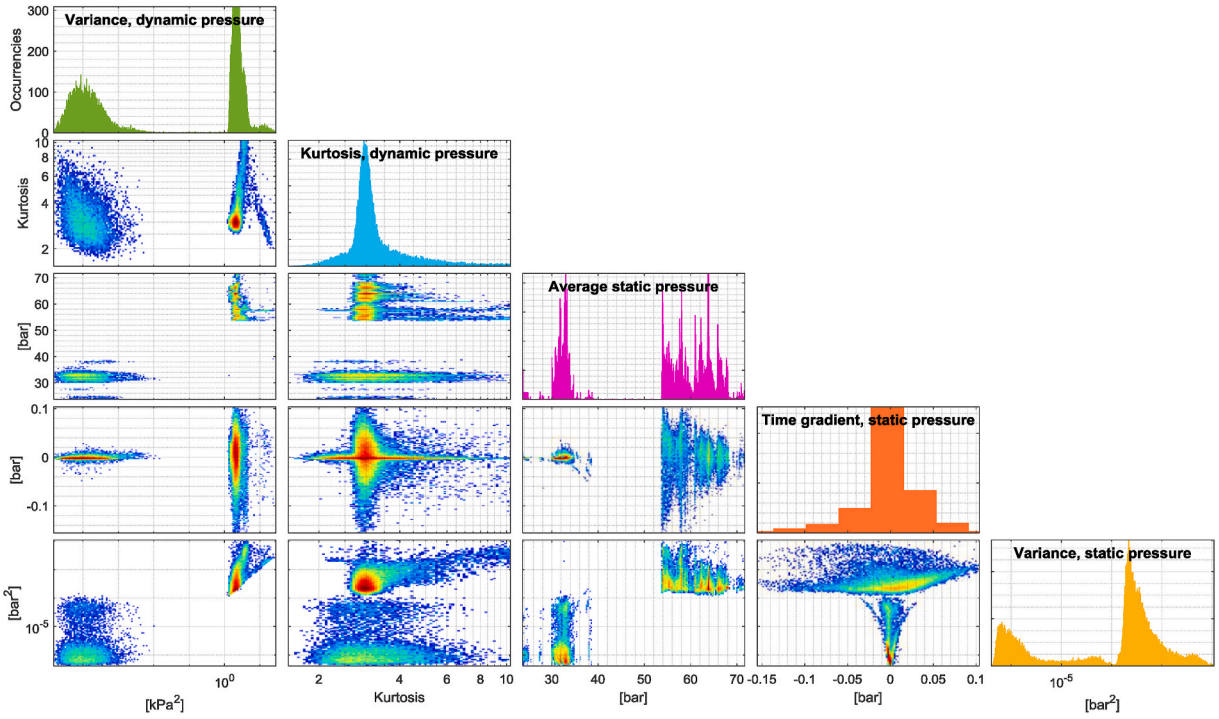

Fig. 6. Density plot between pairs of raw features (lower triangular part) and individual probability distributions (main diagonal).

The $m \times N$ normalized feature matrix $\boldsymbol{Z}$ is then used to train a GMM, which is manually specified to learn a number $K = 4$ of Gaussian components that can best identify and distinguish between the four scenarios described in Section IV. The optimization process is carried out using the iterative expectation-maximization algorithm (Moon, 1996), which returns the parameters $\boldsymbol{\theta}_k$ of the four Gaussian distributions ($k \in \{1, \ldots, 4\}$). Successively, the trained GMM is used to partition each training example (specified by any row of $\boldsymbol{Z}$) into one of the K clusters by selecting the $\bar{k}$ -th component with the largest a posteriori probability $p_{i\bar{k}}$.

5.4. Performance evaluation

The goodness of an unsupervised clustering algorithm cannot be determined through the typical metrics used in supervised learning problems (e.g., F1 score), since no ground truth labels are available. Therefore, the evaluation procedure requires using the model itself. A possible solution consists in establishing whether the clustering method is able to consistently separate data in well-defined regions, according to some distance metric: for instance, one could evaluate the mean Silhouette Coefficient $\mathscr{S}$, which measures the average degree of similarity existing between every data point and each of the K clusters. For a single sample $\boldsymbol{x}^{(i)}$, previously assigned to a specific region $\bar{k}$, the individual Silhouette index is defined as:

$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, & a(i) < b(i) \\ 0, \quad a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1, & a(i) > b(i) \end{cases} \tag{2}$$

where $a(i)$ represents the mean distance between $\boldsymbol{x}^{(i)}$ and all the other points assigned to the same cluster $\bar{k}$, whereas $b(i)$ corresponds to the mean distance between $\boldsymbol{x}^{(i)}$ and all the other examples belonging to the next nearest region $k \neq \bar{k}$. In practice, the Silhouette index quantifies to what extent an object is related to its assigned cluster, with respect to all the others.

By looking at (2), it can be inferred that

$$-1 \leq s(i) \leq 1, \tag{3}$$

where the upper bound indicates that $\boldsymbol{x}^{(i)}$ is well-matched to its own cluster $\bar{k}$ and weakly related to all the other regions $k \neq \bar{k}$; opposite considerations are true with regards to the lower bound.

A widely used similarity metric is the squared Euclidean distance. Considering two distinct points $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(j)}$ in the N -dimensional feature space, the squared Euclidean distance $d^2\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right)$ is expressed as:

$$d^2\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right) = \left(x_1^{(i)} - x_1^{(j)}\right)^2 + \left(x_2^{(i)} - x_2^{(j)}\right)^2 + \ldots + \left(x_N^{(i)} - x_N^{(j)}\right)^2 = \sum_{n=1}^{N} \left(x_n^{(i)} - x_n^{(j)}\right)^2. \tag{4}$$

Using (4), the quantities $a(i)$ and $b(i)$ can be rewritten:

$$a(i) = \frac{1}{C_{\bar{k}} - 1} \sum_{\boldsymbol{x}^{(j)} \in \bar{k},\ i \neq j} d^2\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right), \tag{5}$$

$$b(i) = \min_{k \neq \bar{k}} \left\{ \frac{1}{C_k} \sum_{\boldsymbol{x}^{(j)} \in k} d^2\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\right) \right\}, \tag{6}$$

where $C_{\bar{k}}$ and C_k respectively refer to the number of points assigned to clusters $\bar{k}$ and k.

For a set of m examples, the mean Silhouette Coefficient is computed as:

$$S = \frac{1}{m} \sum_{i=1}^{m} s(i) \tag{7}$$

As with s(i), $\mathscr{S}$ is bounded between -1 and 1: the former instance refers to an inappropriate clustering configuration, that is when K is either too large or too small; in the latter case, a strong structure has been found. If $\mathscr{S} = 0$, instead, overlapping issues between clusters are present.

The algorithm presented in this work achieves a mean Silhouette score equal to 0.83, indicative that the attained clustering configuration (with $K = 4$) refers to a model with well-defined and dense clusters: this aspect finds additional confirmation through a visual inspection of the labelled training data, using the tags obtained from the GMM described in Section V.C. As a reference, two examples are respectively reported in Fig. 7 and Fig. 8, where the assigned labels are given a meaningful interpretation based on the physical considerations listed in Section V.A. Considering Fig. 7, it can be observed that most of the anomalous dynamic pressure transients (red-colored lines) have been correctly detected and are clearly discernible in time from the values associated to a regular functioning regime (green lines). Such feature can be potentially advantageous for further development of a count-based detection system, where the occurrence of events is declared based on the number of times an instance is identified within a given time interval. Fig. 8 displays the typical behavior of healthy pumping equipment, where the three physiological regimes (steady state, flow regulation and regular functioning, respectively tagged with black, blue and green lines) have been properly recognized, as the measured pressure values are coherent with the individual characteristics of each state.

6. Automated pump health monitoring: application to historical data

This section presents how the trained GMM has been applied for automated pump health monitoring on the crude oil transportation system connecting Chivasso and Pollein. Moreover, it is shown that the model can be successfully adopted on other pipelines having similar pumping equipment.

6.1. Chivasso-Pollein (crude oil transportation pipeline)

A first, straightforward implementation of the model consists in the continuous analysis of pressure measurements recorded at Chivasso station, after the training phase (from February 1st, 2015 onwards). As previously outlined in step 6 of Section IV. A, each test data point is given as input to the GMM and is assigned to a specific cluster, after having evaluated and normalized the required set of features (Section V. B and V.C).

Maintenance logs report a pump fault event which occurred on February 15th, 2015. It can be inferred from Fig. 9 that the system is signaling a number of anomalies (represented by red-colored lines) on that date: interestingly, the abnormal behavior appears to show up several days before the actual failure occurrence.

Another potential application of the model is represented by remote pump monitoring, exploiting pressure sensors installed at long distances from the pumping terminal. Such a strategy becomes feasible as the pipeline behaves like a wave guide with respect to the pressure transients propagating in the fluid, and vibroacoustic signals generated by

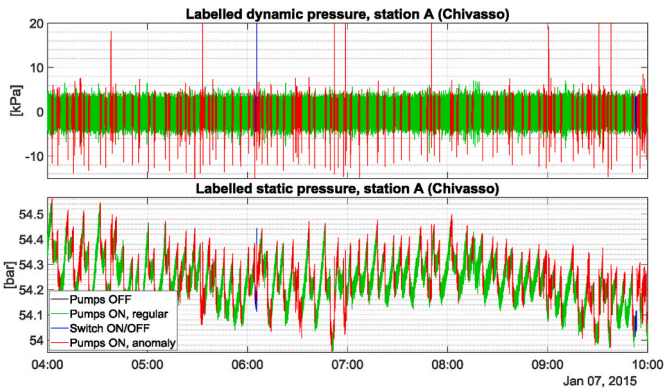

Fig. 7. Detection of the pump failure event occurred on January 7th, 2015.

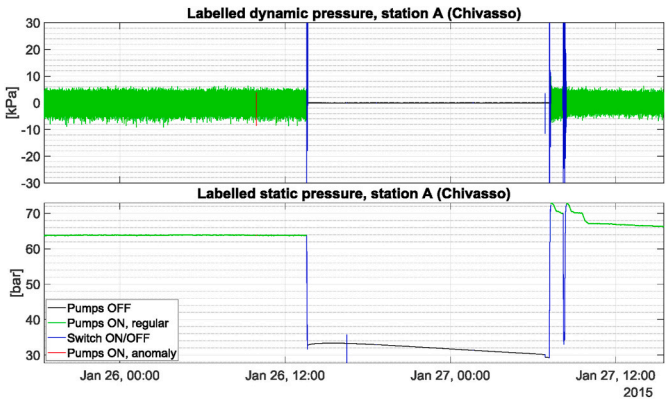

Fig. 8. Assessment of regular pump functioning between January 25th, 2015 and January 27th, 2015.

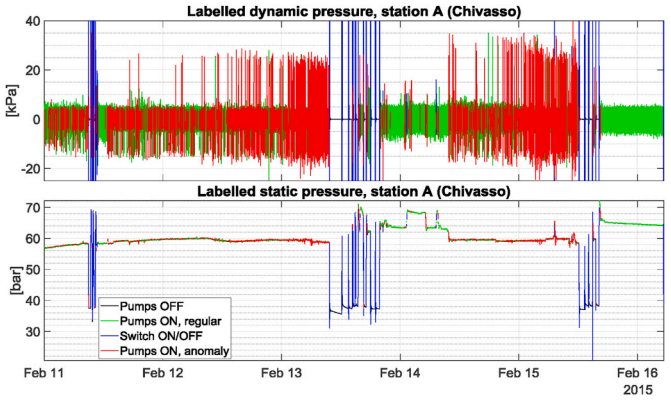

Fig. 9. Failure event occurred on February 15th, 2015.

pumping equipment can be sensed at tens of km from the source point, along the conduit. Moreover, this aspect can be advantageous for monitoring applications on offshore lines, where a direct access to electrical submersible pumps is typically resource demanding.

To illustrate the feasibility of such approach, two different solutions are proposed here. The choice between them is determined for each employed measurement station and according to the availability of sensor data:

1. If both static and dynamic hydrophones are at disposal, one can regularly exploit the trained GMM to automatically label the input data, as described at the beginning of Section VI. A;
2. Otherwise, in case one of the two sensors is unavailable, a combination of data obtained from different stations (e.g., static pressure measured at station A with dynamic pressure recorded at station B) is given as input to the GMM.

As a reference, Fig. 10 and Fig. 11 display the results obtained by applying the first strategy to each station on a failure event which occurred on May 10th, 2012, since all the sensors were operational during such time frame. Fig. 12, instead, shows the adoption of the second strategy to the monitoring period previously represented in Fig. 9. Concerning the latter instance, only two out of five measurement stations were fully operational (namely A and E), while station B could only provide dynamic pressure measurements, since the related static hydrophone was not functioning during the month of February 2015.

The results obtained so far are very promising and suggest that the adoption of multiple vibroacoustic sensors (deployed along the pipeline) may open up possibilities for advanced remote monitoring and control strategies: for instance, if anomalous pump behaviors are concurrently

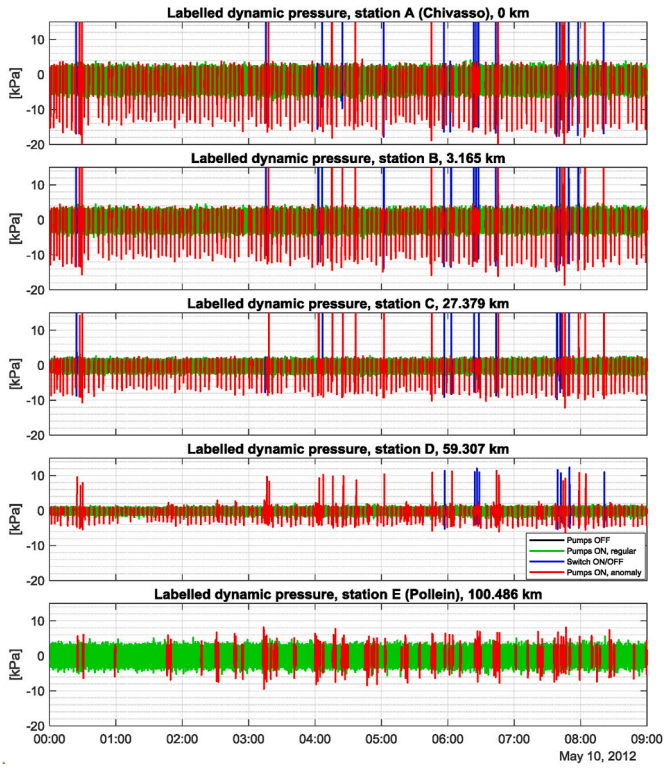

Fig. 10. Labelled dynamic pressure measurements (Chivasso-Pollein pipeline) during a pump failure event occurred on May 10th, 2012.

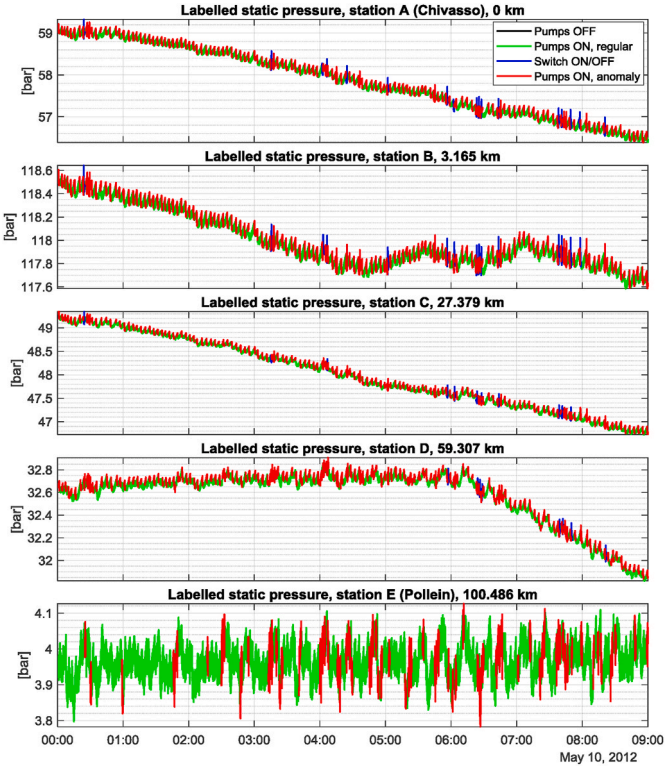

Fig. 11. Labelled static pressure measurements (Chivasso-Pollein pipeline) during a pump failure event occurred on May 10th, 2012.

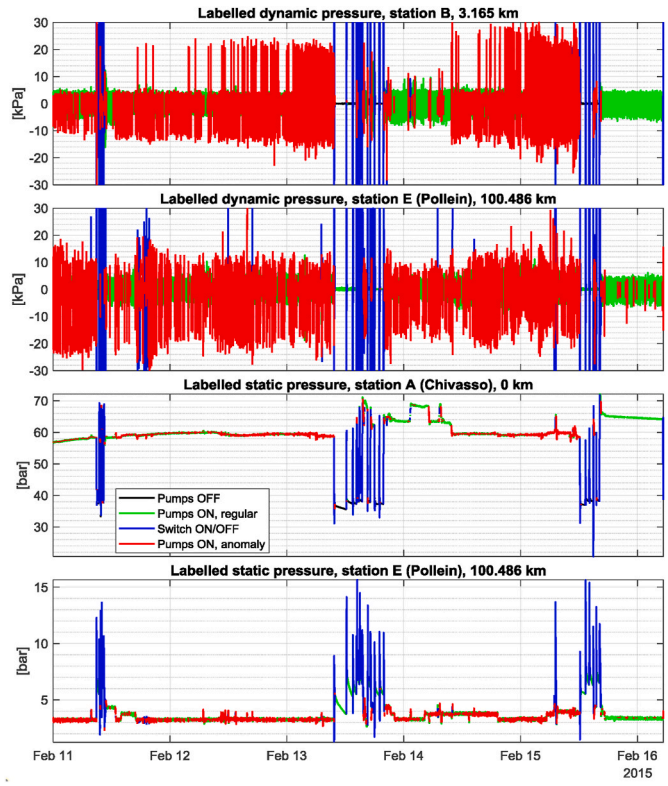

Fig. 12. Fault detection on February 15th, 2015, combining data obtained from different stations.

detected across multiple stations, this additional information can provide confirmation about the actual reliability of the fault itself. Moreover, multiple stations can be used to separate pressure transients coming from different locations (pumps in several positions, flow regulation equipment, third party interaction with the pipe), increasing the functionalities of the monitoring procedure.

6.2. *Gaeta-Pomezia (refined products transportation pipeline)*

The proposed model has also been applied on a refined products (diesel, kerosene) transportation system operated by Eni R&M, connecting the logistic terminals of Gaeta (Latina) and Pomezia (Rome). Such line was instrumented by e-vpms® technology in 2012, has a length of around 112 km and is used to transfer refined oil products through pipes having a nominal diameter of 16 inch. Compared to the crude oil Chivasso-Pollein pipeline, the service pressure of the transported fluid ranges from a maximum of about 40 bar at Gaeta pumping station, to a minimum of around 2 bar at the terminal in Pomezia: yet, it is worth noting that both pipelines are equipped with a single and similar pumping unit (the technical specifications of the pump located in Gaeta are reported in Table 5). Concerning the vibroacoustic monitoring setup (displayed in Fig. 13), a set of seven permanent stations has been

Table 4
Distance between each e-vpms® station and the pumping terminal located in Gaeta.

Station	Distance with respect to station A (km)
A (Gaeta)	0
B	17.409
C	34.770
D	55.886
E	77.344
F	94.751
G (Pomezia)	112.188

Table 5
Technical specifications of the centrifugal pump located in Gaeta.

Technical specification	Value	Measurement unit
Flow rate	525	m3/h
Operational density	0.7–0.86	kg/m3
Operational temperature	5–40	°C
Hydraulic head	455	mH2O
Absorbed power	725	kW

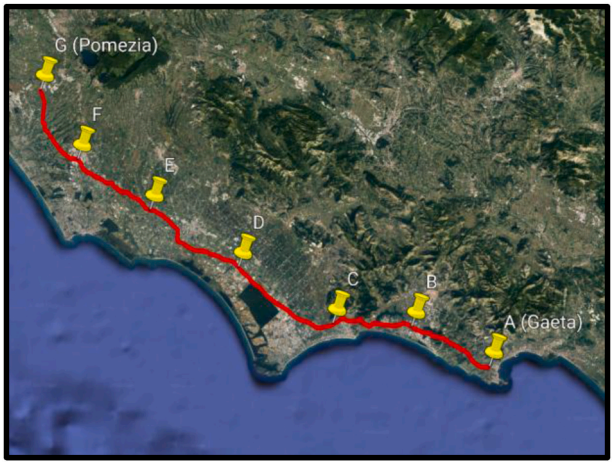

Fig. 13. Satellite map of Gaeta-Pomezia products pipeline route (red line) and position of the permanent stations (yellow pins). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

installed along the whole line; each one of them is equipped with the same measurement instrumentation and is given similar labelling as described in Section II. The distance between each station and the pumping terminal in Gaeta is reported in Table 4.

Local maintenance logs reported the occurrence of a fault event related to the pumping station located in Gaeta in the early afternoon of April 15th, 2019: we have therefore applied the GMM (previously trained using data recorded by station A of Chivasso-Pollein pipeline) to monitor the behavior of the pumps for the full month of April 2019.

Experimenting the validity of the data-driven model in such framework has three main purposes: firstly, verify if the proposed pump monitoring system is applicable on different oil transportation systems; in affirmative case, determine once again whether the anomalies can be observed at several kilometers of distance from the source; lastly, choosing a monitoring interval corresponding to a full month, rather than focusing on the individual failure event, allows to establish when the first signs of anomaly actually started occurring.

Given the above considerations, it was decided to exploit the static and dynamic pressure data recorded by the seven e-vpms® measurement stations. Fig. 14 and Fig. 15 respectively show a relevant detail of the time frame considered, where the trend of labelled dynamic and static pressure measurements (right before and after the corruption of the pump) is represented: in such a case, a clear separation between regular functioning (green-colored lines) and anomalous (red lines) regimes is observed at five out of seven stations and at approximately the same time. Compared to the event represented in Figs. 9 and 12, the typical signs of failure started manifesting right around the actual fault occurrence, rather than several days before.

From the same charts, the effect of attenuation on acoustic wave propagation can also be noticed, as the amplitude of dynamic pressure vibrations, generated by the pumps, decreases with increasing distances. Having multiple sensing stations at disposal in each pipeline, we have exploited this phenomenon to further validate the remote monitoring

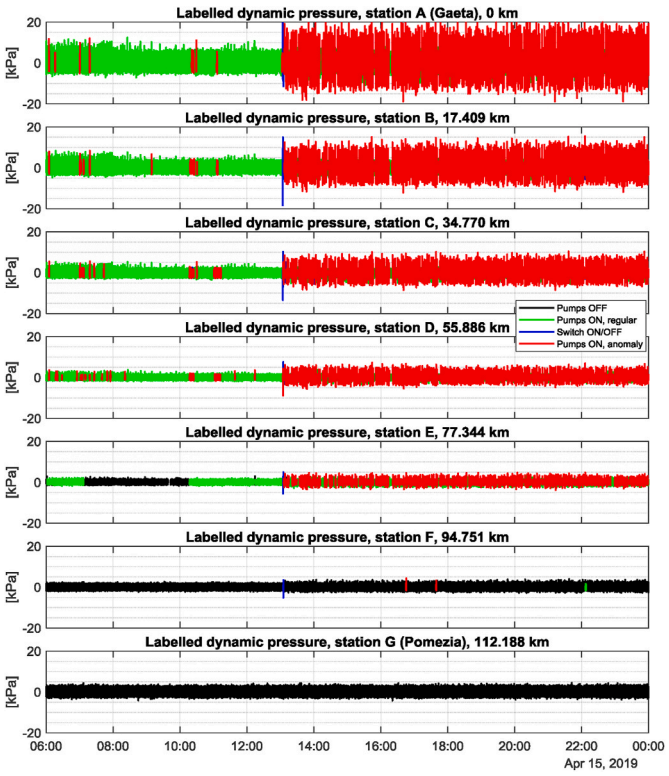

Fig. 14. Observed dynamic pressure at vlm3, vlm7 and vlm9 stations during the pump failure event of April 15th, 2019.

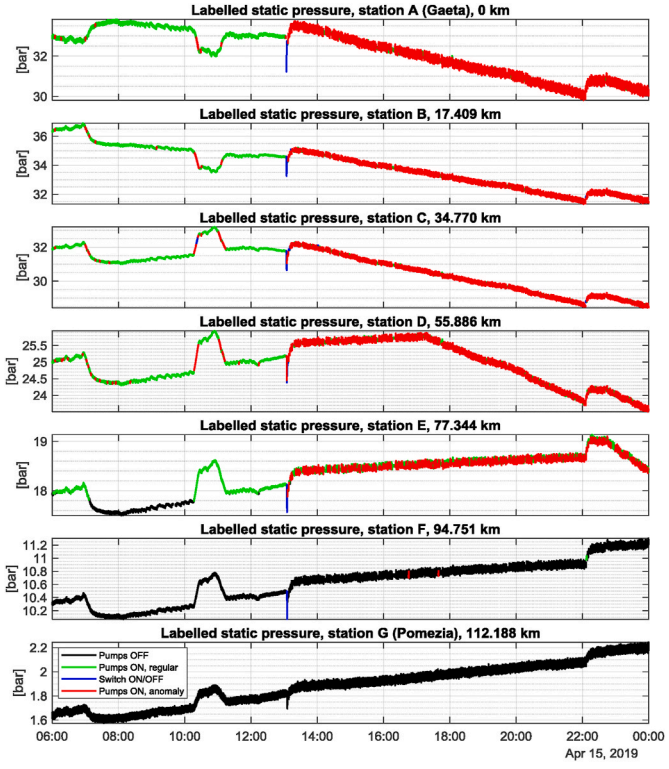

Fig. 15. Observed static pressure at vlm3, vlm7 and vlm9 stations during the pump failure event of April 15th, 2019.

capabilities of our system, thus determining the maximum distance at which pump health can be effectively tracked: considering Chivasso-Pollein pipeline, we managed to detect failure events using sensors located at up to 100 km of distance from the pump; the results for Gaeta-Pomezia are still very satisfactory, achieving a reliable monitoring range which extends up to 80 km.

Unsupervised clustering techniques based on Gaussian mixtures, such as the one presented here, have a relatively simple architecture, yet they can still be employed effectively for real-time monitoring of pumping machinery. Moreover, reprocessing of historical, untagged data allows to generate a set of labelled examples which can be successively employed to train supervised learning algorithms, allowing to design more advanced pipeline integrity methods. This may represent an important advantage, since labelled data are almost always unavailable in pipeline transportation systems.

7. Conclusion

The current frontier of pump monitoring in pipeline systems is the installation on the pump itself of a network of sensors and the processing of the data collected by these sensors with machine learning techniques. The objective of the monitoring strategy is to guarantee that the pump working regime stays between safe boundaries, no anomalies are reported (e.g., on vibrations, currents, temperature), and pumping performance is within specifications.

This paper presented an automated strategy to remotely monitor the status of centrifugal pumps in pipeline transportation systems, when the network of sensors is not available or not present. The procedure exploits pressure transients generated by the pumping equipment itself and measured within the fluid, also at several km from the originating point (pump). In practice, it "hears" the sound (pressure transients) generated by the pumps in the fluid environment. In a training phase the procedure builds a database of working regimes. By proper selecting the sound descriptors and the observation window, it is possible to classify and to distinguish the "normal" modes and the anomalies. In the detection phase, new data is mapped in the database to estimate the current pumping regime. We have presented the methodology and its validation on a real case history where, as interesting result, historical data have been repurposed and used to tune the monitoring strategy. The unsupervised clustering analysis has provided four distinct operational statuses of the pump and the definition of a reference model, parameterized with Gaussian mixtures. An expert analysis of such a model, produced automatically, reveals the ability of the system to correctly distinguish normal operation from scenarios where the pump is approaching failure. With respect to most approaches available in the literature, the proposed method does not require any labelled data, being of unsupervised nature: this becomes particularly advantageous whenever manual tagging of data points is unfeasible, impractical or excessively time consuming. Moreover, the monitoring procedure only requires standard pressure sensors, typically already installed both in modern and dated pumping systems. In addition, pump monitoring can be performed remotely (up to 100 km of distance), as the pipeline behaves like an efficient waveguide with respect to the acoustic signals generated by the pump itself: this guarantees additional freedom during a sensors setup phase, as the measurement instrumentation can be located at any point along the pipeline within the maximum distance. Moreover, it allows for remote monitoring solutions in contexts where a direct access to the equipment is resource demanding (e.g., electrical submersible pumps in offshore lines), and measurement instrumentation must be necessarily installed far from the pump.

Performance evaluation of the clustering algorithm was carried out on a statistical basis through the evaluation of the mean Silhouette Coefficient, achieving a score equal to 0.83. Additionally, the robustness of the monitoring strategy has been validated on two independent pressure datasets, collected from real pipeline transportation systems: we also managed to successfully predict and detect all the pump failure events reported by the available maintenance logs.

Future work consists in employing the labelled data, provided by the unsupervised clustering procedure, to develop integrity monitoring techniques based on supervised learning. In addition, further validation of the algorithm's performance needs to be carried out on untested scenarios, such as: monitoring multistage and/or multiphase pumps (for oil transportation purposes), automatic supervision of gas compressor stations, offshore pipelines. As a last development phase, we plan to work on an industrial deployment of the proposed remote monitoring and control strategy, to make it fully operational with streaming data collected in real time from pipeline assets. By taking advantage of these additional resources, one can potentially improve the design of advanced pipeline integrity systems, exploiting the availability of complex, big data assets. The ultimate goal consists in a digital transformation of the oil and gas industry processes to point towards Industry 4.0.

Credit author statement

Riccardo Angelo Giro: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization. **Giancarlo Bernasconi:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Supervision, Project administration. **Giuseppe Giunta:** Conceptualization, Methodology, Investigation, Resources, Writing – Original Draft, Writing – Review & Editing, Visualization, Supervision, Project administration. **Simone Cesari:** Investigation, Resources, Writing – Original Draft, Writing – Review & Editing, Visualization.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgment

This research was carried out in the framework of the R&D – DIONISIO project funded by Eni S. p.A. The authors are grateful to Eni R&M Logistic Department and SolAres JV teams for technical support during the field tests.

References

Alabied, S., Daraz, A., Rabeyee, K., Alqatawneh, I., Gu, F., Ball, A.D., 2019. Motor Current Signal Analysis Based on Machine Learning for Centrifugal Pump Fault Diagnosis. 2019 25th International Conference on Automation and Computing (ICAC). https://doi.org/10.23919/IConAC.2019.8895057.

Bangari, P., Nangare, K., Al Mazrouei, K.H., 2019. Improving Equipment Reliability and Availability through Real-Time Data. Society of Petroleum Engineers. https://doi.org/10.2118/197347-MS.

Barrios Castellanos, M., Serpa, A.L., Biazussi, J.L., Monte Verde, W., do Socorro Dias Arrifano Sassim, N., 2020. Fault identification using a chain of decision trees in an electrical submersible pump operating in a liquid-gas flow. J. Petrol. Sci. Eng. 184 https://doi.org/10.1016/j.petrol.2019.106490.

Chakravarthy, R., Bharadwaj, S.C., Subramaniam, U., Padmanaban, S., Dutta, N., Holm-Nielsen, J.B., 2019. Electrical fault detection using machine learning algorithm for centrifugal water pumps. In: 2019 IEEE International Conference on Environment and Electrical Engineering and 2019 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I CPS Europe). https://doi.org/10.1109/EEEIC.2019.8783841.

Deng, X., Cao, Y., Yang, M., Sun, Z., Cui, W., 2019. Fault diagnosis of sucker rod pumping system using modified extreme learning machine assisted by gravitational search algorithm. In: 2019 Chinese Control and Decision Conference (CCDC). https://doi.org/10.1109/CCDC.2019.8833177.

Dutta, N., Umashankar, S., Shankar, A., Padmanaban, S., Leonowicz, Z., Wheeler, P., 2018. Centrifugal pump cavitation detection using machine learning algorithm technique. In: 2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I CPS Europe). https://doi.org/10.1109/EEEIC.2018.8494594.

Giunta, G., Bernasconi, G., Bondi, L., 2019a. Pipeline digital monitoring based on vibroacoustic measurements. In: Ravenna: Offshore Mediterranean Conference. http://hdl.handle.net/11311/1120343.

Giunta, G., Cesari, S., Giro, R.A., Bernasconi, G., 2020. Digital Transformation of Historical Data for Advanced Predictive Maintenance. Abu Dhabi International Petroleum Exhibition & Conference (ADIPEC), Abu Dhabi. https://doi.org/10.2118/202906-MS.

Giunta, G., Nielsen, K.L., Bernasconi, G., Bondi, L., Korubo, B., 2019b. Data Driven Smart Monitoring for Pipeline Integrity Assessment. Society of Petroleum Engineers. https://doi.org/10.2118/197327-MS.

Giunta, G., Timossi, P., Borghi, G.-P., Schiavon, R., Bernasconi, G., Chiappa, F., 2015. Field deployment of Eni vibroacoustic pipeline monitoring system (e-vpms™): long term performance analysis. Ravenna: Offshore Mediterranean Conference and Exhibition. https://doi.org/10.13140/RG.2.1.4526.1522.

Hailong, J., Gonghui, L., Jun, L., Tao, Z., Chao, W., 2020. Drilling fault classification based on pressure and flowrate responses via ensemble classifier in Managed pressure drilling. J. Petrol. Sci. Eng. https://doi.org/10.1016/j.petrol.2020.107126.

Hajizadeh, Y., 2019. Machine learning in oil and gas; a SWOT analysis approach. J. Petrol. Sci. Eng. 176, 661–663. https://doi.org/10.1016/j.petrol.2019.01.113.

Kalmár, C., Hegedűs, F., 2019. Condition monitoring of centrifugal pumps based on pressure measurements. Period. Polytech. - Mech. Eng. 80–90 https://doi.org/10.3311/PPme.12140.

Koroteev, D., Tekic, Z., 2021. Artificial intelligence in oil and gas upstream: trends, challenges, and scenarios for the future. Energy and AI, p. 100041. https://doi.org/10.1016/j.egyai.2020.100041.

Kreyszig, E., 1983. Advanced Engineering Mathematics. Wiley, New York.

Lygren, S., Piantanida, M., Amendola, A., 2019. Unsupervised, Deep Learning-Based Detection of Failures in Industrial Equipments: the Future of Predictive Maintenance. Society of Petroleum Engineers. https://doi.org/10.2118/197629-MS.

Marins, M.A., Barros, B.D., Santos, I.H., Barrionuevo, D.C., Vargas, R.E., de, M., Prego, T., Netto, S.L., 2020. Fault detection and classification in oil wells and production/service lines using random forest. J. Petrol. Sci. Eng. https://doi.org/10.1016/j.petrol.2020.107879.

Moon, T.K., 1996. The expectation-maximization algorithm. IEEE Signal Process. Mag. 13 (6), 47–60.

Orrù, P.F., Zoccheddu, A., Sassu, L., Mattia, C., Cozza, R., Arena, S., 2020. Machine learning approach using MLP and SVM algorithms for the fault prediction of a centrifugal pump in the oil and gas industry. Sustainability. https://doi.org/10.3390/su12114776.

Panda, A., Rapur, S., Tiwari, R., 2018. Prediction of flow blockages and impending cavitation in centrifugal pumps using support vector machine (SVM) algorithms based on vibration measurements. Measurement 44–56. https://doi.org/10.1016/j.measurement.2018.07.092.

Peng, L., Han, G., Pagou, A.L., Shu, J., 2020. Electric submersible pump broken shaft fault diagnosis based on principal component analysis. J. Petrol. Sci. Eng. https://doi.org/10.1016/j.petrol.2020.107154.

Rauber, T.W., Oliveira-Santos, T., de Assis Boldt, F., Rodrigues, A., Varejao, F.M., Ribeiro, M.P., 2017. Kernel and random extreme learning machine applied to submersible motor pump fault diagnosis. In: 2017 International Joint Conference on Neural Networks (IJCNN). https://doi.org/10.1109/IJCNN.2017.7966276.

Reynolds, D.A., 2009. Gaussian mixture models. Encycl. Biometr. 741. https://doi.org/10.1007/978-1-4899-7488-4_196.

Sharma, R., Pandey, N., 2016. A neural network model for electric submersible pump surveillance. In: Melmaruvathur: 2016 International Conference on Communication and Signal Processing (ICCSP). https://doi.org/10.1109/ICCSP.2016.7754545.

Soylemezoglu, A., Jagannathan, S., Saygin, C., 2011. Mahalanobis-taguchi system as a multi-sensor based decision making prognostics tool for centrifugal pump failures. IEEE Trans. Reliab. 864–878. https://doi.org/10.1109/TR.2011.2170255.

Tejedor, J., Macias-Guarasa, J., Martins, H., Pastor-Graells, J., Corredera, P., Martin-Lopez, S., 2017. Machine Learning Methods for Pipeline Surveillance Systems Based on Distributed Acoustic Sensing: A Review. Applied Sciences. https://doi.org/10.3390/app7080841.

Van Rensburg, N.J., Kamin, L., Davis, S.R., 2019. Using Machine Learning-Based Predictive Models to Enable Preventative Maintenance and Prevent ESP Downtime. Society of Petroleum Engineers. https://doi.org/10.2118/197146-MS.