Journal of Petroleum Science and Engineering 181 (2019) 106223

Contents lists available at ScienceDirect

Journal of Petroleum Science and Engineering

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

A realistic and public dataset with rare undesirable real events in oil wells

Ricardo Emanuel Vaz Vargas[a,b,*], Celso José Munaro[a], Patrick Marques Ciarelli[a], André Gonçalves Medeiros[b], Bruno Guberfain do Amaral[c], Daniel Centurion Barrionuevo[d], Jean Carlos Dias de Araújo[b], Jorge Lins Ribeiro[b], Lucas Pierezan Magalhães[c]

[a] Departamento de Engenharia Elétrica, Universidade Federal do Espírito Santo, Av. Fernando Ferrari, 514, Goiabeiras, Vitória, ES, CEP: 29060-370, Brazil
[b] Petróleo Brasileiro S.A., Av. Nossa Sra. da Penha, 1688, Barro Vermelho, Vitória, ES, CEP: 29057-570, Brazil
[c] Petróleo Brasileiro S.A., Rua Ulysses Guimarães, 565, Cidade Nova, Rio de Janeiro, RJ, CEP: 20211-160, Brazil
[d] Petróleo Brasileiro S.A., Cidade Universitária, Rio de Janeiro, RJ, CEP: 21941-970, Brazil

ARTICLE INFO

Keywords:
Fault detection and diagnosis
Oil well monitoring
Abnormal event management
Multivariate time series classification

ABSTRACT

Detection of undesirable events in oil and gas wells can help prevent production losses, environmental accidents, and human casualties and reduce maintenance costs. The scarcity of measurements in such processes is a drawback due to the low reliability of instrumentation in such hostile environments. Another issue is the absence of adequately structured data related to events that should be detected. To contribute to providing a priori knowledge about undesirable events for diagnostic algorithms in offshore naturally flowing wells, this work presents an original and valuable dataset with instances of eight types of undesirable events characterized by eight process variables. Many hours of expert work were required to validate historical instances and to produce simulated and hand-drawn instances that can be useful to distinguish normal and abnormal actual events under different operating conditions. The choices made during this dataset's preparation are described and justified, and specific benchmarks that practitioners and researchers can use together with the published dataset are defined. This work has resulted in two relevant contributions. A challenging public dataset that can be used as a benchmark for the development of (i) machine learning techniques related to inherent difficulties of actual data, and (ii) methods for specific tasks associated with detecting and diagnosing undesirable events in offshore naturally flowing oil and gas wells. The other contribution is the proposal of the defined benchmarks.

1. Introduction

In the general industrial context, there have been increasing demands for greater operational safety, productivity, quality, and energy efficiency (Jämsä Jounela, 2007). Complexity, instrumentation, and automation have increased significantly to meet these demands (Venkatasubramanian et al., 2003). Control loops, whether manual or automated, are developed to maintain operations under normal conditions, but there are changes and disturbances which these control loops cannot handle satisfactorily. *Faults* occur in these situations (Russell et al., 2000).

The term “fault” is defined in (Russell et al., 2000) as an unpermitted deviation of at least one characteristic behavior or process variable. Aldrich and Auret in (Aldrich and Auret, 2013) define *fault* as anomalous behavior causing systems or processes to deviate unacceptably from their normal operating conditions or states. The review provided by Venkatasubramanian et al. (2003) defines *fault* as an abnormality or process symptom, such as high temperature in a reactor or low product quality. This review also defines the underlying cause(s) of a fault, such as a failed coolant pump or a controller, as the *root cause(s)* or *basic event(s),* which are also referred to as *malfunction(s)* or *failure (s)*. Since root cause analysis is not the focus of our work, for simplicity, all these terms are generalized as *undesirable events*.

Detection and classification of *rare* undesirable events are tasks that are relevant and in vogue in several activities carried out and/or monitored by human beings. Some examples are flow influx detection during drilling (Tang et al., 2019); leak detection and location in water and oil pipelines (Liu et al., 2019); fault detection in industrial plants (Arrudaet al., 2014), (Peter He and Wang, 2007), (Xavier and de Seixas, 2018), in oil wells (Liuet al., 2011), (Liuet al., 2010a), (Liuet al., 2010b), in Electrical Submersible Pumps (ESPs) (PatriAnand et al., 2014), and in gas compressor valves (Patri et al., 2015a), (Patriet al.,

* Corresponding author. Petróleo Brasileiro S.A., Av. Nossa Sra. da Penha, 1688, Barro Vermelho, Vitória, ES, CEP: 29057-570, Brazil
E-mail addresses: ricardo.vargas@petrobras.com.br (R.E.V. Vargas), munaro@ele.ufes.br (C.J. Munaro), patrick.ciarelli@ufes.br (P.M. Ciarelli), andremedeiros@petrobras.com.br (A.G. Medeiros), bruno.do.amaral@petrobras.com.br (B.G.d. Amaral), dcbarrionuevo@petrobras.com.br (D.C. Barrionuevo), jeanaraujo@petrobras.com.br (J.C.D.d. Araújo), jorge_lribeiro@petrobras.com.br (J.L. Ribeiro), lucas.magalhaes@petrobras.com.br (L.P. Magalhães).

https://doi.org/10.1016/j.petrol.2019.106223
Received 18 February 2019; Received in revised form 17 June 2019; Accepted 28 June 2019
Available online 01 July 2019
0920-4105/

2016); abnormal event detection in oil wells (Santoset al., 2018), (Vargas et al., 2017), in Electroencephalography (EEG) (Unget al., 2017), and in Electrocardiography (ECG) (Anthony et al., 2017); power-quality disturbance detection (Gray and Morsi, 2015), (Gaing, 2004); and handgun detection (Olmos et al., 2018).

The task of responding to abnormal events in a process involves the timely detection of an abnormal event, diagnosing its root causes, and then taking appropriate control decisions and actions to bring the process back to a normal, safe, and operational state. This entire activity has come to be called *Abnormal Event Management* (AEM). Diagnosis in automated AEM can be viewed as a classification problem, and classification algorithms can be categorized in terms of their knowledge and search strategies (Venkatasubramanian et al., 2003).

The application of machine learning algorithms to data obtained from processes is a search strategy that has been used successfully, especially in recent years. Machine learning methods have been applied in different tasks, such as prediction of downhole working conditions of the beam pumping unit (Li et al., 2018), well-testing model classification from pressure transient test data (Ahmadi et al., 2017), automatic analysis of real-time drilling data, and detection of flow influx events (Tang et al., 2019). In (Bhattacharya and Mishra, 2018), machine learning algorithms are applied for facies and fracture classification in conventional and unconventional reservoirs. In (Xavier and de Seixas, 2018), a method is proposed that allows fault diagnosis, as well as detection, without any tunable parameters and that can be used with large volumes of data with very little information. Methods for fault detection are proposed in (Arrudaet al., 2014), (Peter He and Wang, 2007). A method that can detect the accumulation of hydrate in production or injection lines of oil wells with at least 1 h in advance is presented in (Santoset al., 2018). A specific transformation on each monitored variable is performed in (Vargas et al., 2017) before classification itself by machine learning algorithms to avoid model recalibration even in dynamic systems. Classification of time series is empirically evaluated in 39 datasets in (Li et al., 2016). Methods for failure prediction in oil wells are proposed in (Liuet al., 2011), (Liuet al., 2010a), (Liuet al., 2010b). Other works related to algorithms for time series classification are published in (Patriet al., 2014), (Patri et al., 2015b), (Geurts, 2001), (Xi et al., 2006).

In oil and gas wells, AEM can help prevent production losses, environmental accidents, and human casualties and reduce maintenance costs. The catastrophic Macondo incident that occurred in 2010 due to the failure of safety equipment exemplifies the potential magnitude of losses and costs. The deaths of 11 workers, the sinking of the Deepwater Horizon rig, and the massive marine and coastal damage marked this incident as one of the largest environmental disasters in US history (Sutherland et al., 2016). In terms of maintenance, the cost of a maritime probe to repair a production line, for example, can exceed US $500,000 per day (Andreolli, 2016).

In the middle of 2017, Petróleo Brasileiro S.A., known as Petrobras, conceived its first project designed to evolve its existing AEM in oil and gas wells. This project, entitled “Monitoramento de Alarmes Especialistas (MAE)”, was conceived at the Petrobras Operational Unit located in the Brazilian state of Espírito Santo (UO-ES). The idea was to complement (and overlap with automatic suppressions) the current system based on parametric univariate alarms with machine learning algorithms applied to Multivariate Time Series (MTS) obtained from the processes. The formal definition of MTS considered in this work is presented in Section 2.5. The main goal of the MAE project was to develop a new automated AEM capable of detecting and classifying occurrences of eight specific types of undesirable events in offshore naturally flowing wells that are in a normal state in a shorter time with better performance. This scope was established with the following justifications: (i) MTS, used in a priori domain knowledge composition, are highly available; (ii) machine learning algorithms have been shown to be a suitable search strategy for this type of knowledge; (iii) naturally flowing wells are less complex and are therefore good candidates for technological innovation projects; (iv) gains in offshore wells are potentially higher given their higher average production compared to onshore wells (Andreolli, 2016); (v) the selected types of undesirable events had been responsible for most of the production loss in the UO-ES in the last years. It is estimated that during 2016 the production loss was 1,514,000 bbl (barrels), which corresponds to US$75.7 million if we consider an average value of US$50/bbl in this period.

Some authors of this article participated in the conception of the MAE project and are still working on its development. Even after extensive research, no public or private dataset with enough undesirable events in terms of quantity and diversity in oil and gas wells has been found so far.

One of the private datasets found is part of the Electric Submersible Pump – Reliability Information and Failure Tracking System (ESP-RIFTS) (FER Technologies, 2018a), developed by the ESP-RIFTS “Joint Industry Project (JIP)” and maintained by the C-FER Technologies (FER Technologies, 2018b). The goal of the ESP-RIFTS is to improve the run life of ESPs significantly. Its dataset is composed of information extracted from about 112,000 ESP installations from about 800 fields operated by 26 companies around the world. However, this dataset does not contain any time series associated with physical quantities of processes. Only two datasets with MTS acquired from processes were found, both owned by Petrobras and not public. The first one consists of 11 occurrences (four simulated and seven real) of four types of undesirable events and was used in (Vargas et al., 2017). The second contains 12 real occurrences of only a single type of undesirable event and was explored in (Santoset al., 2018). Even together, these last two datasets do not have sufficient undesirable events in terms of quantity or diversity.

In public repositories (Dua and Taniskidou, 2017), (Dau et al., 2018), (Anthony et al., 2018), (Chen, 2018), (Istituto Nazionale di Statistica), only datasets associated with other contexts were found. Besides, in general these datasets are pretreated and become unrealistic at some level. Some examples of pretreatment are elimination or replacement of *not a number* (NaN) values or of frozen variables (due to sensor, system configuration, or network communication issues); consideration of only time series with necessarily equal sizes or with only simulated data; and balancing in relation to quantities of occurrences per type of undesirable event.

As a result of this research, it was decided to generate a dataset to be used in the development of automated AEM with machine learning algorithms. That is the 3W[1] dataset's origin, to the best of its authors' knowledge, the first realistic and public dataset with rare undesirable real events in oil wells.

This work gave rise to two relevant contributions. The first one is the 3W dataset, which has been made available in the supporting repository (Vargas et al., 2019) for this paper and can be readily used as a benchmark dataset for development of machine learning techniques related to inherent difficulties of actual data. An enormous number of possibilities in terms of, for example, preprocessing (normalization, NaN values, missing values, frozen variables, outliers, etc.), filters (smoothing, resampling, etc.), transformations (multiscale, wavelet, etc.), family of classifiers (based on trees, artificial neural networks, distances, ensembles, etc.), hyperparameter optimization, feature engineering, and performance metrics can be investigated using this dataset. Furthermore, these possibilities can be explored in specific tasks associated with detecting and diagnosing undesirable events in offshore naturally flowing oil and gas wells, such as early classification (He et al., 2013), (Xing et al., 2011), (Xing et al., 2009); novel fault detection (Krawczyk et al., 2017); and one-class (Krawczyk et al., 2017), binary (Krawczyk et al., 2017), multi-class (Krawczyk et al., 2017), multi-label (Zhang and Zhou, 2007), online, and offline classification.

[1] The name 3W was chosen because this dataset is composed of instances from **3** different sources and which contain undesirable events that occur in oil **W**ells.

More information about each one of all these tasks can be obtained in (Aldrich and Auret, 2013), (James et al., 2013), (Witten et al., 2011), (Hastie et al., 2009), (Christopher, 2006), (Duda et al., 2001). The second contribution of this article is the specification of challenges (benchmarks) that practitioners and researchers can use together with the 3W dataset.

The remainder of this article is organized as follows. The next section addresses the background required for a good understanding of what is proposed in this paper. Section 3 describes how the 3W dataset was prepared. The proposed specific benchmarks are defined in Section 4. The last section is dedicated to the conclusions of this work.

2. Background

2.1. Offshore naturally flowing wells

An oil well refers to a set of sensors and mechanical, pneumatic, and hydraulic systems that may be partially or fully installed on the seabed, downhole (the well itself), or on the surface (Pierre, 2007). Nowadays, there are different artificial lift techniques (Andreolli, 2016), such as beam pumps, gas lift, and electric submersible pumps, but the scope of this work considers only offshore naturally flowing wells.

Naturally flowing wells are those whose reservoir pressure is sufficient to produce hydrocarbons at a commercial rate without requiring any additional energy. This situation can occur in both offshore and onshore wells. This type of well tends to have less equipment and therefore less instrumentation, control loops, and automation. Fig. 1 presents a schematic that corresponds to an offshore scenario. The oil and gas flow from a reservoir through production tubing and then through a production line to a platform. A subsea Christmas tree is a type of equipment installed on the seabed and is basically composed of valves and sensors operated remotely through an electro-hydraulic umbilical. A *Permanent Downhole Gauge* (PDG) and a *Temperature and Pressure Transducer* (TPT) are devices that contain pressure and temperature sensors, respectively. The PDG remains fixed in a certain position of the production tubing, and the TPT is part of the subsea Christmas tree. A *Downhole Safety Valve* (DHSV) and a *Production Choke* (PCK) are valves and are better explained in the next subsection.

It is important to note that naturally flowing wells can also be equipped to be operated with artificial lift methods under certain circumstances. That is, it is not uncommon for a well to be operated in an intercalated fashion between an artificial lift technique and the natural method.

As mentioned in the introduction section, it was decided that the MAE project would focus on offshore naturally flowing wells that are in a normal state. In this state, the well continues producing without significant occurrences of anomalies. Therefore, all the occurrences of undesirable events present in the 3W dataset started from a *normal* state. The two mutually exclusive states considered by Petrobras specialists are called *closed-in* (when there is no flow at all) and *starting up* (a transient state between closed-in and normal state).

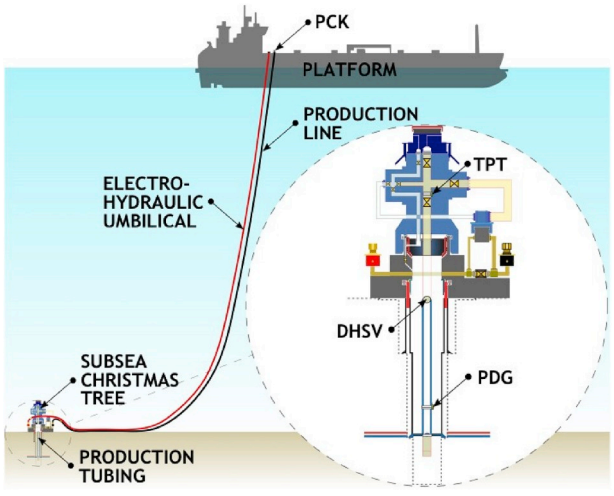

Fig. 1. Simplified schematic of a typical offshore naturally flowing well.

The most common monitored variables in Petrobras offshore naturally flowing wells are present in the 3W dataset and are listed below. This list takes into account the cost-benefit ratio of having instrumentation in certain positions and also whether the instrumentation is reliable enough despite the hostility of the environment.

- Pressure at the PDG;
- Pressure at the TPT;
- Temperature at the TPT;
- Pressure upstream of the PCK;
- Temperature downstream of the PCK.

2.2. Types of undesirable events in oil wells

The selected types of undesirable events in the MAE project and therefore present in the 3W dataset are described next. It is important to clarify that there is not always consensus regarding names and what these types of undesirable events mean, even among experts. That is why they are detailed next.

1 Abrupt Increase of BSW

Basic Sediment and Water (BSW) is defined as the ratio between the water and sediment flow rate and the liquid flow rate, both measured under *normal temperature and pressure* (NTP) (Andreolli, 2016), (Abass and Bass, 1988).

During the life cycle of a well, its BSW is expected to increase due to increased water production from either the natural reservoir aquifer or artificial injection to avoid declining production. However, a sudden increase of BSW can lead to several problems related to flow assurance, lower oil production, oil lifting, incrustation, industrial plant processing, and the recovery factor. Automatic identification of this type of undesirable event may permit actions such as administering production or artificial injection to avoid this sort of problem.

2 Spurious Closure of DHSV

DHSV, also referenced as just DSV, is a safety valve installed in the production tubing of wells. Its goal is to ensure closing of the well in case of a situation in which the production unit and well are physically disconnected or in the event of an emergency or catastrophic failure of surface equipment. It is set in a fail-safe mode such that any interruption or malfunction of the system will result in the safety valve closing to make the well safe (Schlumberger, 2018), (Standards Norway, 2013).

Eventually, this closure function fails in a spurious manner, often without any indication on the surface (e.g., pressure drop in the hydraulic actuator). Automatic identification of spurious closing of this valve in a timely manner may allow it to be reopened through corrective operational procedures, avoiding production losses and additional costs.

3 Severe Slugging

This is a critical type of instability. The two most striking features of this event are the well-defined periodicity (around 30, 45, or 60 min) and the intensity, which is generally sufficient to be detected by sensors along the entire production line (Meglio et al., 2012), (Schmidt et al., 1985).

Depending on the periodicity and intensity, this type of event can result in stress or even damage to equipment in the well and/or the

industrial plant. When it is detected in advance, specific actions in the operation of the well can be taken to reverse the situation.

4 Flow Instability

During a flow instability, at least one of the monitored variables undergoes relevant changes but with tolerable amplitudes. A characteristic that differentiates this type of undesirable event from severe slugging is the lack of periodicity between these changes (Theyab, 2018), (Takeiet al., 2010).

As instability can progress to severe slugging, its prognosis avoids all the negative aspects associated with this more severe anomaly.

5 Rapid Productivity Loss

The productivity of a naturally flowing well depends on several properties: static pressure reservoir, percentage of basic sediment and water, viscosity of the produced fluid, diameter of the production line, and so on (Hausler et al., 2015).

When these properties are changed so that the system's energy is no longer sufficient to overcome the losses, the flow slows or even stops. Automatic identification of this condition in a timely manner may allow the operation team to change the operating point of the well so that it does not lose its productivity.

6 Quick Restriction in PCK

The PCK is a control valve installed at the beginning of the production unit and is responsible for well control at the surface. The expression “quick restriction in PCK” is not well defined in the literature but is an English version of a term that is widely used internally at Petrobras. For this expression to be correctly used, the restriction must occur with an amplitude above a stipulated reference (e.g., 5%) and in a short time (e.g., less than 10 s).

When this type of valve is manually operated, unwanted quick restrictions may eventually occur due to operational problems. Identifying this event automatically is also desirable because unwanted restrictions can be reversed more quickly.

7 Scaling in PCK

Monitoring the production choke is important due to the susceptibility of inorganic deposits, which can dramatically reduce oil and gas production (Schlumberger, 2018).

Automatic identification of this condition in a timely manner is also desirable because appropriate actions, such as scale inhibitor injection, can be taken to avoid oil and gas production losses.

8 Hydrate in Production Line

Hydrate is one of the biggest problems in the oil industry. It is defined as a crystalline compound that is formed by water and natural gas and therefore resembles ice. As its formation requires the presence of water and natural gas, in addition to high pressures and low temperatures, oil pipelines in which dead oil flows do not suffer from this type of anomaly. Its occurrence is more frequent in gas pipelines and gas producing wells. However, hydrates can also be formed in oil-producing wells, to the point of totally interrupting their flow (Andreolli, 2016), (Ellison et al., 2000).

Avoiding occurrences of this type of undesirable event means avoiding production losses for days or even weeks. In certain cases, high costs of unblocking the production line are also avoided.

2.3. Response times

To confirm real occurrences of each considered type of undesirable event, professionals who perform well monitoring at Petrobras usually analyze time windows of different sizes whose estimates are presented in Table 1. These window sizes can be used as additional information to improve the performance of machine learning algorithms.

Table 1
Estimates of time window sizes used to confirm occurrences of undesirable events.

TYPE OF UNDESIRABLE EVENT	WINDOW SIZE
1 – ABRUPT INCREASE OF BSW	12 h
2 – SPURIOUS CLOSURE OF DHSV	5 min–20 min
3 – SEVERE SLUGGING	5 h
4 – FLOW INSTABILITY	15 min
5 – RAPID PRODUCTIVITY LOSS	12 h
6 – QUICK RESTRICTION IN PCK	15 min
7 – SCALING IN PCK	72 h
8 – HYDRATE IN PRODUCTION LINE	30 min–5 h

2.4. Machine learning

Machine learning is a subfield of artificial intelligence, directly associated with computer science. The main characteristic of an algorithm that performs machine learning is the existence of some systematized learning mechanism from examples (also called objects, occurrences, or instances). This type of algorithm works in two phases (Faceli et al., 2011):

- Training: an initial phase during which learning takes place. The approximation of the function, or model takes place at this stage;
- Use: the phase in which learning is actually used in new objects (not used in the training phase) to solve the task of interest. This stage represents the detection and classification part of automated AEM.

This work considers mainly *predictive* algorithms that perform *classification* (Faceli et al., 2011). The variables of the MTS acquired from the processes and the types of undesirable events are used respectively as *input* and *output attributes*.

The diversity of approaches used by these algorithms has grown considerably, and they give rise to families of algorithms. Some of the main ones are based on neural networks, support vector machines, trees, distances, Bayes' theorem, and boosting. There are so many possibilities that to explain them goes beyond the scope of this article.

2.5. Multivariate Time Series

In this work, the following definition of Multivariate Time Series (MTS) is adopted, which is similar to those used in (Zhou and Chan, 2015), (Weng, 2013), (He et al., 2013).

A *dataset* $\boldsymbol{DS}$ is a set of m MTS $(\boldsymbol{S}^i \mid i = \{1,\ 2,\ ...,\ m\},\ \forall\ m \in \mathbb{Z},\ \text{and}\ m > 1)$ and is defined as $\boldsymbol{DS} = \{\boldsymbol{S}^1, \boldsymbol{S}^2, ..., \boldsymbol{S}^m\}$. Each MTS i is an *instance* (also referenced in this paper as object or occurrence), that is composed of a set of n *univariate time series* $(x_j^i \mid j = \{1,\ 2, ..., n\},\ \forall\ n \in \mathbb{Z},\ \text{and}\ n > 1)$ (also referenced as *process variable* or just *variable*), and is defined as $\boldsymbol{S}^i = \{x_1^i,\ x_2^i, ..., x_n^i\}$. Each variable j that composes an MTS i is an ordered temporal sequence of p_i *observations* taken at the time t $(x_{j,t}^i \mid t = \{1,\ 2, ..., p_i\},\ \forall\ p_i \in \mathbb{Z},\ \text{and}\ p_i > 1)$. Therefore, each MTS i is viewed in this work as a matrix defined as $\boldsymbol{S}^i = \{x_{1,1}^i,\ x_{2,1}^i, ..., x_{n,1}^i;\ x_{1,2}^i,\ x_{2,2}^i, ..., x_{n,2}^i; ...; x_{1,p_i}^i,\ x_{2,p_i}^i, ..., x_{n,p_i}^i\}$.

Note that all instances have a fixed number of variables n, but each instance can be composed of any quantity of observations p_i. It is also important to note that all variables of an instance i have fixed number of observations p_i.

3. 3 W Dataset's preparation

This section explains the choices made during the 3W dataset's preparation and then describes its fundamental aspects. We do not claim that these choices are optimal, but we argue that they are reasonable enough and that they gave rise to a benchmark dataset for development of several kinds of techniques and methods for different tasks associated with undesirable events in oil and gas wells.

The 3W dataset is composed of three types of instances that are determined by their sources: real, simulated, and hand-drawn. Real instances are those that actually occurred in Petrobras' actual wells during oil production. The use of simulated and hand-drawn instances is fundamentally intended to decrease the imbalance of the dataset initially formed only by real instances, which is a common characteristic in industrial data (He et al., 2013), (Krawczyk et al., 2017). A possible approach to this is to accomplish re-sampling of real instances. In (Chawla et al., 2002), for example, a method is proposed that performs over-sampling of the minority class (rare events) and under-sampling of the majority class (normal condition). Our strategy is entirely different and seeks to enrich the a priori knowledge (dataset) with more instances obtained from different sources: simulations, and hand-drawn curves by specialists in the problem domain.

Two types of labeling were carried out by experts on each undesirable events. The first one occurred at the instances' level. Each instance, whether real, simulated, or hand-drawn, was necessarily labeled with a single code associated with normal operation or some code associated with the undesirable event existing at some point within the instance. Note, therefore, that no instance contains more than one undesirable event. Two benefits deriving from this type of labeling are that it provides a grouping of instances depending on the type of undesirable event they contain and that it allows the development of offline classifiers, those that do not aim to estimate when the event started or ended inside each instance. The second type of label was applied at the observation's level. Each observation of each instance of any type was labeled with a single code associated with normal operation or some code associated with the undesirable event existing at that instant. This type of label is essential for online classifier training.

Labeling at the observation's level was done so that each instance of any type has up to three periods *normal*, *faulty transient*, and *faulty steady state*. A period in this work means a continuous sequence of observations. In normal periods, there is no evidence of any type of anomaly. In faulty transient periods, the dynamics resulting from an undesirable event are still ongoing. When these dynamics cease, the faulty steady state period begins. The primary purpose of this strategy was to provide the possibility of early classification. That is, faulty transient periods can be learned, and their accurate detection predicts the period faulty steady state. In other words, the faulty transient period can be interpreted as an undesirable pre-event period. Fig. 2 and Fig. 3 bring examples of real instances labeled as normal and abnormal, respectively. In both cases, only the three most relevant variables for their types of events, according to specialists, are presented. Different periods are marked with different colors: green for normal, yellow for faulty transient, red for faulty steady state, and white for not labeled observations (due to the used tool limitations).

All real instances were extracted from the plant information system used to track the industrial processes of the UO-ES (PI System (OSIsoft, 2018)). This extraction was done without preprocessing to maintain their realistic aspects, such as NaN values, frozen variables (due to sensor or network communication issues), instances with different sizes, and outliers. This strategy allows evaluating which preprocessing techniques in raw data result in better performance in each task of interest to be developed. The processes used to generate simulated and hand-drawn instances (computer simulation and hand-drawn curves) gave rise to time series naturally free of such problems.

All simulated instances were obtained with OLGA (Schlumberger), a dynamic multiphase flow simulator adopted by several oil companies around the globe (Andreolli, 2016). This choice was done because OLGA is the standard tool used at Petrobras for the simulation of scenarios in oil wells and also because it is one of the few systems that simulate dynamic phenomena (faulty transients) (Ingebrigtsen Grødahl, 2014). Simulation of rare undesirable real events was prioritized. The threshold used to differentiate rare events was 1%, the same criterion as was used in (Zhang and Zhou, 2007). With this criterion, those types of events whose number of real occurrences represents less than 1% of all real instances, including normal ones, are considered rare.

A specific tool was developed for the 3W dataset to be enriched with

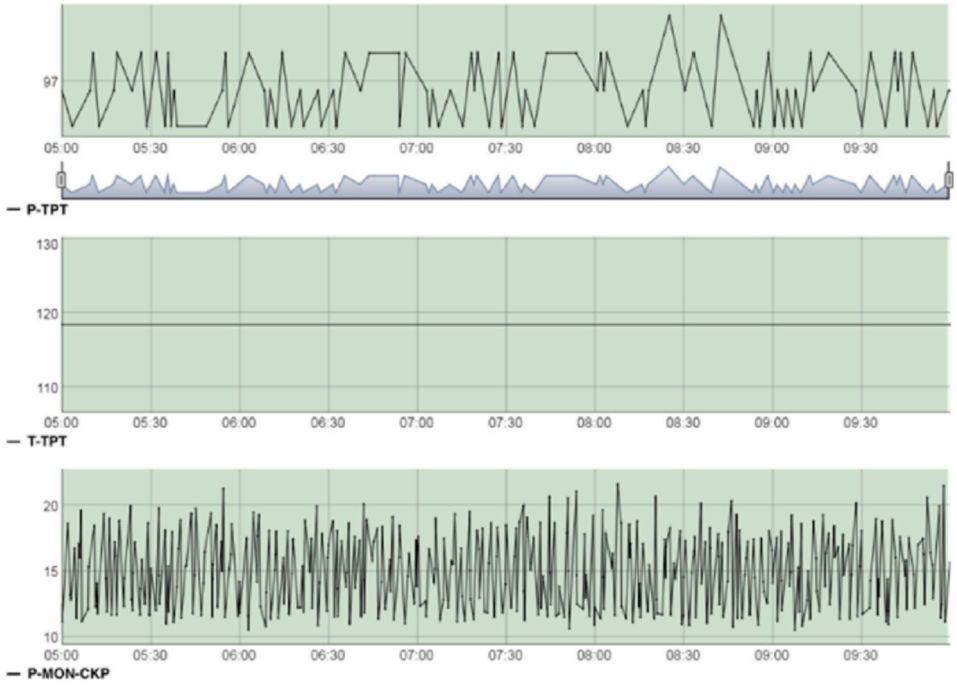

Fig. 2. Screenshot of a real instance labeled as normal. T-TPT is frozen.

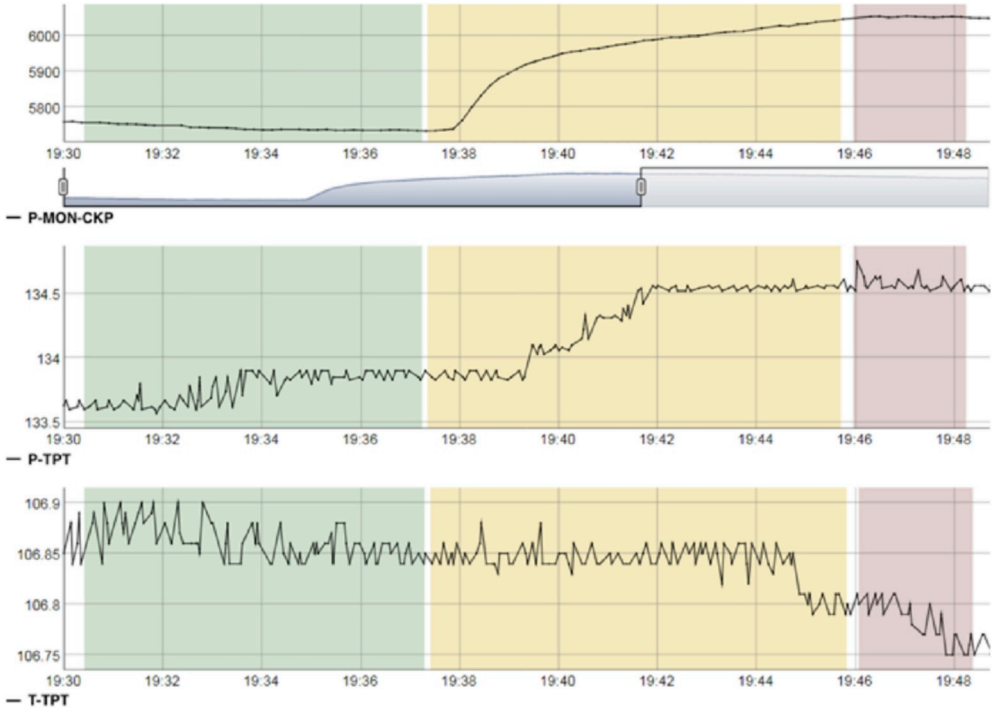

Fig. 3. Screenshot of a real instance labeled as abnormal (quick restriction in PCK).

hand-drawn instances, which naturally includes the tacit knowledge of experts regarding the format of the MTS that characterizes the types of undesirable events that were considered. Two rare undesirable real events were prioritized, one that was simulated using OLGA and another that not. The developed tool is composed of a chart template and a script for image processing. Fig. 4 illustrates the use of this chart template, in which an expert drew one curve by hand, and specified all its attributes: variable, type of event, instance identification, beginning of the normal, faulty transient, and faulty steady state periods, and scales.

The 3W dataset is available in the supporting repository (Vargas et al., 2019) for this paper with the following structure and general characteristics. Each instance, whether real, simulated, or hand-drawn, was saved in a standardized and dedicated *Comma-Separated Values* (CSV) file. All CSV files were grouped into directories based on the instance label. All instances were generated with observations obtained with a fixed sampling rate (1 Hz). Only the following units were used: Pascal [Pa], standard cubic meters per second [sm^3/s], and degrees Celsius [^oC]. The source of each instance was incorporated on the name of its CSV file. All actual well names were replaced by generic names as a requirement of Petrobras for the 3W dataset's publication.

Table 2 shows the quantities of instances that compose the 3W dataset by type of event and by knowledge source: real, simulated, and hand-drawn instances. When considering the threshold of 1% and only real instances, four types of undesirable events are rare: codes 1, 6, 7, and 8. If simulated instances are also considered, three of these types cease to be rare: codes 1, 6, and 8. Even after considering the hand-drawn instances, one type of undesirable event remains rare: code 7.

Fig. 5 shows a scatter map with all the real instances. The oldest one occurred in the middle of 2012 and the most recent one in the middle of 2018. In addition to the total number of considered wells, this map provides an overview of the occurrences distributions of each undesirable event over time and between wells.

The main 3W dataset's fundamental aspects related to inherent difficulties of actual data are:

- Missing variables: 4,947 (31.17% of all 15,872 variables of all 1,984 instances). When all observations of a variable in a particular instance have missing value due to sensor or network communication issues, this variable itself is considered missing. The more missing variables, the more sparse the dataset becomes, which can impose additional difficulties to the algorithms;
- Frozen variables: 1,535 (9.67% of all 15,872 variables of all 1,984 instances). When all observations of a variable in a particular instance have any single float or integer value due to any reason, this variable itself is considered frozen. A frozen variable does not always represent a problem, but this characteristic is a symptom of sensor, system configuration, or network communication issues. Therefore, the problematic frozen variables do not manifest the patterns associated with the undesirable events, which indeed impose additional difficulties to the algorithms;
- Unlabeled observations: 5,130 (0.01% of all 50,913,215 observations of all 15,872 variables of all 1,984 instances). Some observations were not labeled due to the used tool limitations. Even with this percentage, some technique should be used to treat the unlabeled observations.

4. Proposed benchmarks

Two specific benchmarks that practitioners and researchers can use together with the 3W dataset are proposed in this section. This proposal goal is to provide a standardized and appropriate means for algorithms, implemented by diverse participants with different techniques and approaches, to have their performances assessed and compared.

These benchmarks were designed for online binary classification[2] only. All observations from faulty transient and faulty steady state periods must be relabeled as positives and all observations from normal periods as negatives. In this operation, unlabeled observations must be

[2] In binary classification, only two labels are involved. In the context of this work, offline classification is the task whose goal is to estimate a single label for each MTS. A batch learning process that extracts features only of entire MTS is used. On the other hand, in online classification, label estimation and feature extraction are done for multiple subparts of each MTS.

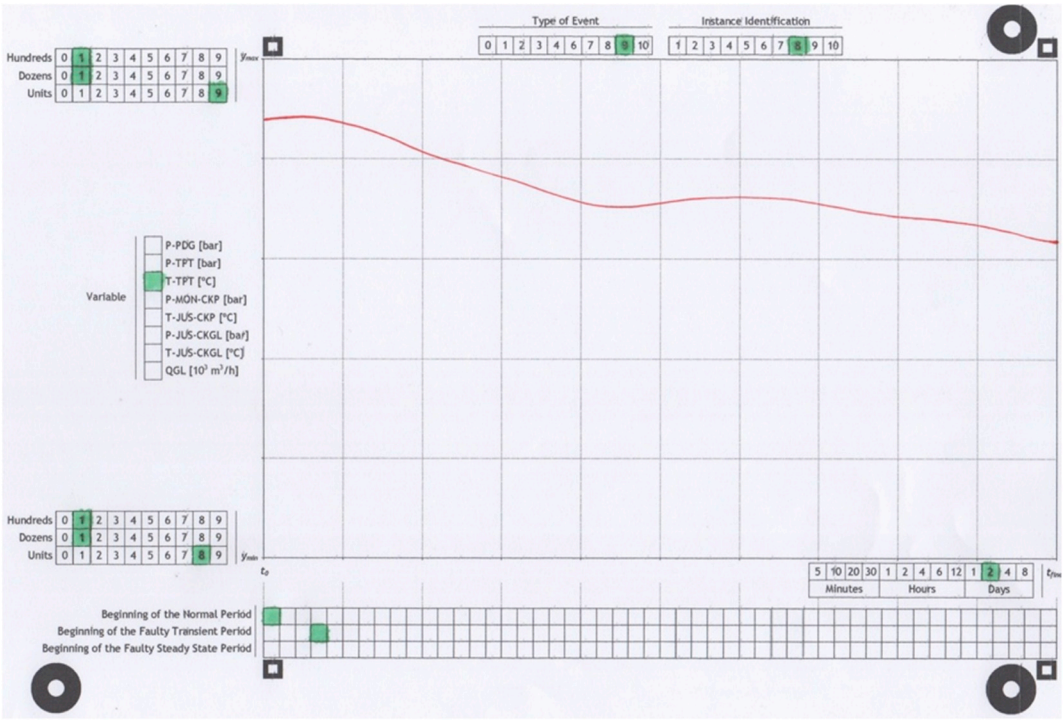

Fig. 4. Illustration of chart template filled out by a specialist.

Table 2
Quantities of instances that compose the 3W dataset.

TYPE OF EVENT	REAL	SIMULATED	HAND-DRAWN	TOTAL
	INSTANCES			
0 – NORMAL	597	–	–	597
1 – ABRUPT INCREASE OF BSW	5	114	10	129
2 – SPURIOUS CLOSURE OF DHSV	22	16	–	38
3 – SEVERE SLUGGING	32	74	–	106
4 – FLOW INSTABILITY	344	–	–	344
5 – RAPID PRODUCTIVITY LOSS	12	439	–	451
6 – QUICK RESTRICTION IN PCK	6	215	–	221
7 – SCALING IN PCK	4	–	10	14
8 – HYDRATE IN PRODUCTION LINE	3	81	–	84
TOTAL	1025	939	20	1984

kept as they are.

Aspects not included in benchmarks' rules can be evaluated and chosen freely. Some of these aspects are preprocessing techniques, window size (sample size), number of samples, approach to labeling samples, and feature engineering methods.

4.1. Impact of using simulated and hand-drawn instances

This benchmark was designed to investigate the impact of using simulated and hand-drawn instances in machine learning algorithm training for detection and classification of rare undesirable events in real instances. The following rules must be observed:

Rule 1: Only the event types 1 or 7 must be chosen. It is not allowed to merge them in the same classification. Just these event types have instances from the three sources (real, simulated, and hand-drawn). Therefore, this benchmark poses two distinct challenges.

Rule 2: Regardless of the source, only instances labeled as of the chosen type can be used. Those with different labels cannot be used. In other words, only CSV files saved in the directory whose name is the chosen type can be used. Besides, all these files must be used.

Rule 3: Multiple rounds of training and testing must be performed with the scheme that we named “leave one real instance out”. The number of rounds must be the number of real instances. In each round, seven pipelines must use different training sets, but precisely the same testing set composed by only one real instance left out of the training set. All other real instance must be used in the training set. Each real instance must be left out only once. From the instance in the testing set, the same number of samples of each label (positive and negative) must be extracted. Each pipeline must be implemented so that its training set is composed by any number of instances per source, as follows:

- Pipeline 1: only real instances;
- Pipeline 2: only *simulated* instances;
- Pipeline 3: only *hand-drawn* instances;
- Pipeline 4: only real and *simulated* instances;
- Pipeline 5: only real and *hand-drawn* instances;
- Pipeline 6: only *simulated* and *hand-drawn* instances;
- Pipeline 7: real, *simulated*, and *hand-drawn* instances.

Rule 4: In each round, precision, recall, and F1 score[3] must be computed, but others may also be considered. Mean value and standard deviation of each metric between all rounds must be presented. Mean value of the F1 score must be considered the main performance metric.

4.2. Anomaly detection

This benchmark intends to encourage the development, evaluation, and comparison of anomaly detecting algorithms. In this task, undesirable events (anomalies) must be distinguished from the normal condition. The following rules must be observed:

Rule 1: Only real instances with an undesirable event type that have a normal period (1, 2, 5, 6, 7, and 8) longer than or equal to 20 min

[3] The F1 score is the harmonic average of the precision and recall; reaches its best value at one (perfect precision and recall) and worst at zero (scikit-learn developers).

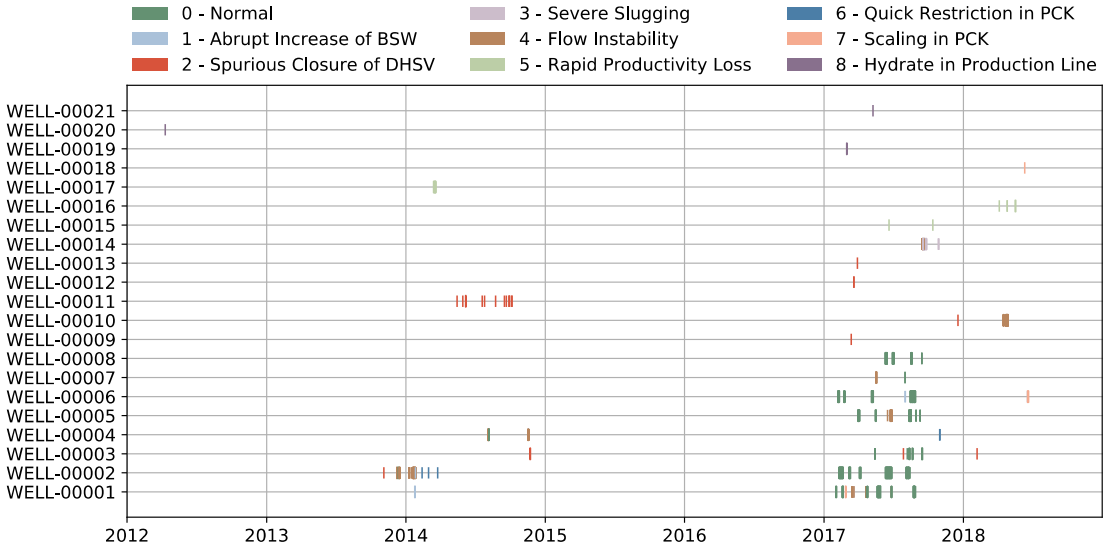

Fig. 5. Scatter map of real instances of the 3W dataset.

must be used. Those with different labels cannot be used. In other words, only CSV files saved in a directory whose name is one of these types can be used. Besides, all these files from all these directories must be used.

Rule 2: Multiple rounds of training and testing must be performed. The number of rounds must be the number of instances. In each round, the following pipeline must be implemented. Samples used for training or testing must be extracted from just one instance. Part of the negative samples must be used for training, and the other for testing. All positive samples must be used for testing only. Therefore, a technique of one-class learning (Krawczyk et al., 2017) must be used. The testing set must be composed of the same number of samples of each label (positive and negative).

Rule 3: In each round, precision, recall, and F1 score must be computed, but others may also be considered. Mean value and standard deviation of each metric between all rounds must be presented. Mean value of the F1 score must be considered the main performance metric.

5. Conclusion

The 3W dataset, an original and valuable resource with instances of eight types of undesirable events that may happen in offshore naturally flowing oil and gas wells, is proposed in this paper. The events are characterized by eight process variables, and the resulting dataset can be readily used as a benchmark for the development of machine learning techniques related to inherent difficulties of actual data. This resource can also be explored in tasks associated with detecting and diagnosing undesirable events in such wells.

A brief description of offshore naturally flowing wells and the undesirable events is followed by details about the dataset preparation, so that the reader may understand the benchmark and apply machine learning techniques.

Specific benchmarks that practitioners and researchers can use together with the published dataset are defined. Along with the proposed dataset, these challenges are expected to be a significant motivation to the community of engineers and scientist who develop machine learning and data analytics methods for the oil and gas field.

As future work, we intend to explorer methods already published in the literature at the context of the proposed benchmarks to conduct several investigations. For example, which types of undesirable events suffer from concept drift (i.e., changes in events signatures which occur over time and deteriorate the performance of the learned models (Krawczyk et al., 2017))?; which preprocessing techniques, transformations, and family of classifiers generate better performances? We also plan to evolve the 3W dataset in the aspects that prove necessary.

Acknowledgments

We thank Petróleo Brasileiro S.A. for encouragement, for providing the data and information necessary for the development of this work, and for permission to publish.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.petrol.2019.106223.

References

Abass, H., Bass, D., 1988. The critical production rate in water-coning system. In: Permian Basin Oil and Gas Recovery Conference, Texas, pp. 351–360.
Ahmadi, Rouhollah, Aminshahidy, Babak, Shahrabi, Jamal, January 2017. Well-testing model identification using time-series shapelets. J. Pet. Sci. Eng. 149, 292–305.
Aldrich, Chris, Auret, Lidia, 2013. Unsupervised Process Monitoring and Fault Diagnosis with Machine Learning Methods. Springer, London.
Andreolli, Ivanilto, 2016. Introdução à Elevação e Escoamento Monofásico e Multifásico de Petróleo. Interciência, Rio de Janeiro.
Anthony, Bagnall, Lines, Jason, Bostrom, Aaron, Large, James, Keogh, Eamonn, 2017. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Min. Knowl. Discov. 31 (3), 606–660.
Anthony, Bagnall, Lines, Jason, Vickers, William, Keogh, Eamonn, 2018. The UEA & UCR time series classification repository. [Online]. www.timeseriesclassification.com.
Arruda, Fellipe, et al., 2014. fault detection in industrial plant using k-nearest neighbors with random Subspace method. In: International Conference on Artificial Intelligence, Las Vegas.
Bhattacharya, Shuvajit, Mishra, Srikanta, November 2018. Applications of machine learning for facies and fracture prediction using Bayesian Network Theory and Random Forest: case studies from the Appalachian basin, USA. J. Pet. Sci. Eng. 170, 1005–1017.
Chawla, Nitesh V., Bowyer, Kevin W., Hall, Lawrence O., Philip Kegelmeyer, W., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.
Chen, Xiaming, 2018. Awesome public datasets. [Online]. https://github.com/awesomedata/awesome-public-datasets.
Christopher, M., 2006. Bishop, pattern recognition and machine learning. In: first ed. Michael Jordan, Jon Kleinberg, and Bernhard Scholkopf Springer-Verlag, New York, USA.
Dau, Hoang Anh, et al., 2018, October. The UCR time series classification archive. [Online]. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.
Dua, Dheeru, Taniskidou, Efi Karra, 2017. [UCI] machine learning repository. [Online]. http://archive.ics.uci.edu/ml.
Duda, Richard O., Hart, Peter E., Stork, David G., 2001. Pattern Classification, second ed. Wiley-Interscience, New York.
Ellison, B., Gallagher, C., Lorimer, S., 2000. The physical chemistry of wax, hydrates, and

asphaltene. In: Offshore Technology Conference, Houston.
Faceli, Katti, Lorena, Ana, Gama, João, Carvalho, André, 2011. Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina. LTC, Rio de Janeiro.
C-FER Technologies, 2018a. Electric Submersible Pump - Reliability Information and Failure Tracking System (ESP-RIFTS). [Online]. http://jip.esprifts.com/.
C-FER Technologies, 2018b. C-FER Technologies. [Online]. https://www.cfertech.com/.
Gaing, Zwe-Lee, 2004. Wavelet-based neural network for power disturbance recognition and classification. IEEE Trans. Power Deliv. 19 (4), 1560–1568.
Geurts, Pierre, 2001. Pattern extraction for time series classification. In: 5th European Conference on Principles of Data Mining and Knowledge Discovery, Freiburg, pp. 115–127.
Gray, M., Morsi, W.G., 2015. Application of wavelet-based classification in non-intrusive load monitoring. In: Canadian Conference on Electrical and Computer Engineering, Halifax, pp. 41–45.
Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome, 2009. The Elements of Statistical Learning, second ed. Springer-Verlag, New York.
Hausler, R.H., Krishnamurthy, R.M., Sherar, Brent, 2015. Observation of productivity loss in large oil wells due to scale formation without apparent production of formation brine. In: CORROSION 2015, Dallas.
He, Guoliang, Duan, Yong, Qian, Tieyun, Chen, Xu, 2013. Early prediction on imbalanced multivariate time series. In: International Conference on Information and Knowledge Management, San Francisco, pp. 1889–1892.
Ingebrigtsen Grødahl, Steinar, 2014. Small Scale Multiphase Flow Experiments on Surge Waves in Horizontal Pipes. Department of Energy and Process Engineering, Norwegian University of Science and Technology, Trondheim Master Thesis.
Istituto Nazionale di Statistica Time series. [Online]. http://seriestoriche.istat.it/index.php?id=1&no_cache=1&L=1&no_cache=1.
James, Gareth, Witten, Daniela, Hastie, Trevor, Tibshirani, Robert, 2013. An Introduction to Statistical Learning with Applications in R. Springer-Verlag, New York.
Jämsä Jounela, Sirkka-Liisa, 2007. Future trends in process automation. Annu. Rev. Contr. 31, 211–220.
Krawczyk, Bartosz, Minku, Leandro L., Gama, João, Stefanowski, Jerzy, Woźniake, Michał, September 2017. Ensemble learning for data stream analysis: a survey. Inf. Fusion 37, 132–156.
Li, Daoyuan, Bissyande, Tegawende, Klein, Jacques, Traon, Yves, 2016. Time series classification with discrete wavelet transformed data: insights from an empirical study. In: The 28th International Conference on Software Engineering and Knowledge Engineering, Redwood City.
Li, Kun, Han, Ying, Wang, Tong, January 2018. A novel prediction method for down-hole working conditions of the beam pumping unit based on 8-directions chain codes and online sequential extreme learning machine. J. Pet. Sci. Eng. 160, 285–301.
Liu, Cuiwei, Li, Yuxing, Xu, Minghai, April 2019. An integrated detection and location model for leakages in liquid pipelines. J. Pet. Sci. Eng. 175, 852–867.
Liu, Yintao, et al., 2010a. Failure prediction for artificial lift systems. In: SPE Western Regional Meeting, Anaheim.
Liu, Yintao, et al., 2010b. Failure prediction for rod pump artificial lift systems. In: SPE Western Regional Meeting, Anaheim.
Liu, Yintao, et al., 2011. Semi-supervised failure prediction for oil production wells. In: 11th International Conference on Data Mining Workshops, Vancouver, pp. 434–441.
Meglio, Florent, Petit, Nicolas, Alstad, Vidar, Kaasa, Glenn-Ole, April 2012. Stabilization of slugging in oil production facilities with or without upstream pressure sensors. J. Process Control 22 (4), 809–822.
OSIsoft, 2018. OSIsoft website. [Online]. https://www.osisoft.com/pi-system/.
Olmos, Roberto, Tabik, Siham, Herrera, Francisco, January 2018. Automatic handgun detection alarm in videos using deep learning. Neurocomputing 275 (C), 66–72.
Patri, Om, Reyna, Nabor, Anand, Panangadan, Prasanna, Viktor, 2015a. Predicting compressor valve failures from multi-sensor data. In: SPE Western Regional Meeting, Garden Grove.
Patri, Om, Kannan, Rajgopal, Anand, Panangadam, Prasanna, Viktor, 2015b. Multivariate time series classification using inter-leaved shapelets. In: NIPS Time Series Workshop, Montreal.
Patri, Om, Anand, Panangadan, Chelmis, Charalampos, McKee, Randall G., Prasanna, Viktor, 2014. Predicting failures from oilfield sensor data using time series shapelets. In: SPE Annual Technical Conference and Exhibition, Amsterdam.
Patri, Om, et al., 2014. Extracting discriminative shapelets from heterogeneous sensor data. In: International Conference on Big Data (Big Data), Washington, pp. 1095–1104.
Patri, Om, et al., 2016. Data mining with shapelets for predicting valve failures in gas compressors. In: SPE Western Regional Meeting, Anchorage.
Peter He, Q., Wang, Jin, Nov. 2007. Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes. IEEE Trans. Semicond. Manuf. 20 (4), 345–354.
Pierre, Donnez, 2007. Essentials of Reservoir Engineering. Editions Technip, Paris, France.
Russell, Evan L., Chiang, Leo H., Braatz, Richard D., 2000. Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis. Chemometr. Intell. Lab. Syst. 51 (1), 81–93.
Santos, Ismael H., et al., 2018. Hydrate failure detection in production and injection lines using model and data-driven approaches. In: Rio Oil&Gas Expo and Conference, Rio de Janeiro.
Schlumberger Schlumberger website. [Online]. https://www.software.slb.com/products/olga.
Schlumberger, 2018. The Oilfield Glossary: where the Oil Field Meets the Dictionary. [Online]. https://www.glossary.oilfield.slb.com.
Schmidt, Zelimir, Doty, Dale, Dutta-Roy, Kunal, Feb. 1985. Severe slugging in offshore pipeline riser-pipe systems. Soc. Petrol. Eng. J. 25 (1), 27–38.
scikit-learn developers scikit-learn website. [Online]. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html.
Standards Norway. NORSOK Standard D-010.
Sutherland, V.A., Ehrlich, M., Engler, R., Kulinowski, K., 2016. Executive Summary – Drilling Rig Explosion and Fire at the Macondo Well. Washington, Investigation Report.
Takei, Jamaludin, et al., 2010. Flow instability in deepwater flowlines and risers – a case study of subsea oil production from chinguetti field, Mauritania. In: SPE Asia Pacific Oil and Gas Conference and Exhibition, vol. 2 Brisbane.
Tang, Hewei, Zhang, Shang, Zhang, Feifei, Venugopal, Suresh, January 2019. Time series data analysis for automatic flow influx detection during drilling. J. Pet. Sci. Eng. 172, 1103–1111.
Theyab, Muhammad, March 2018. Severe slugging control: simulation of real case study. J. Environ. Res. 2 (1).
Ung, Hoameng, et al., September 2017. Intracranial EEG fluctuates over months after implanting electrodes in human brain. J. Neural Eng. 14 (5).
Vargas, Ricardo, Munaro, Celso, Ciarelli, Patrick, Jean, Araújo, 2017. Proposal for two classifiers of offshore naturally flowing wells events using K-nearest neighbors, sliding windows and time multiscale. In: 6th International Symposium on Advanced Control of Industrial Processes (AdCONIP), Taipei, pp. 209–214.
Vargas, Ricardo, et al., 2019. The first realistic and public dataset with rare undesirable real events in oil wells. [Online]. https://github.com/ricardovvargas/3w_dataset.
Venkatasubramanian, V., Rengaswamy, R., Yin, K., Kavuri, S.N., 2003. A review of process fault detection and diagnosis: Part I. Quantitative model-based methods. Comput. Chem. Eng. 27 (3), 293–311.
Weng, Xiaoqing, 2013. Classification of multivariate time series using supervised neighborhood preserving embedding. In: Chinese Control and Decision Conference (CCDC), Guiyang, pp. 957–961.
Witten, Ian H., Frank, Eibe, Hall, Mark A., 2011. Data Mining: Practical Machine Learning Tools and Techniques, third ed. Morgan Kaufmann, Burlington, USA.
Xi, Xiaopeng, Keogh, Eamonn, Shelton, Christian, Li, Wei, Ratanamahatana, Chotirat, 2006. Fast time series classification using numerosity reduction. In: 23rd International Conference on Machine Learning, Pittsburgh, pp. 1033–1040.
Xing, Zhengzheng, Pei, Jian, Yu, Philip, 2009. Early prediction on time series: a nearest neighbor approach. In: 21st International Joint Conference on Artificial Intelligence, Pasadena, pp. 1297–1302.
Xavier, Gilberto M., de Seixas, José Manoel, 2018. Fault detection and diagnosis in a chemical process using long short-term memory recurrent neural network. In: International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, pp. 1–8.
Xing, Zhengzheng, Pei, Jian, Yu, Philip, Wang, Ke, 2011. Extracting interpretable features for early classification on time series. In: International Conference on Data Mining, Mesa, pp. 247–258.
Zhang, Min-Ling, Zhou, Zhi-Hua, 2007. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recogn. 40 (7), 2038–2048.
Zhou, Pei-Yuan, Chan, Keith, 2015. A feature extraction method for multivariate time series classification using temporal patterns. In: 19th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Ho Chi Minh City, pp. 409–421.