Studies in Systems, Decision and Control 255

Emil Pricop
Jaouhar Fattahi
Nitul Dutta
Mariam Ibrahim *Editors*

Recent Developments on Industrial Control Systems Resilience

Springer

Editors

Emil Pricop
Control Engineering, Computers
and Electronics Department
Petroleum-Gas University of Ploiesti
Ploiesti, Romania

Nitul Dutta
Computer Engineering Department
Marwadi University
Rajkot, Gujarat, India

Jaouhar Fattahi
Department of Computer Science
and Software Engineering
Laval University
Quebec City, QC, Canada

Mariam Ibrahim
Department of Mechatronics Engineering,
School of Applied Technical Sciences
German Jordanian University
Amman, Jordan

ISSN 2198-4182 ISSN 2198-4190 (electronic)
Studies in Systems, Decision and Control
ISBN 978-3-030-31327-2 ISBN 978-3-030-31328-9 (eBook)
https://doi.org/10.1007/978-3-030-31328-9

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Machine Learning Based Predictive Maintenance of Infrastructure Facilities in the Cryolithozone

Andrey V. Timofeev and Viktor M. Denisov

Abstract This chapter provides some practical aspects and peculiarities of the use of Machine Learning based Predictive Maintenance for the infrastructure facilities in the cryolithozone. Some mathematical models of Machine Learning based Predictive Maintenance are described, which have shown their practical effectiveness. The solutions of several important problems of Predictive Maintenance for pipelines located in cryolithozone are considered, including: problem of leak detection from pipelines taking into account the possible damage to the pipeline foundation due melting of permafrost; problem of automatic classifying of defects that led to leaks; problem of prompt corrosion spot detection in the pipelines as well as problem of identifying the current state of the corrosion process in the pipeline. The problem of optimizing the procedure for incident tickets processing in the Predictive Maintenance system for oil pipelines was also considered.

1 Introduction

Predictive maintenance (PdM) [1–4] is a technology to recognize the condition of an equipment to identify maintenance requirements to maximize its performance. The PdM system's output is probabilistic prediction of the future equipment state. This prediction is directed towards preventing equipment from future breakdowns, failures, or outages with usage of continuously monitoring diverse data related to performance and efficiency of a given asset. The use of PdM allows to predict when and which equipment needs maintenance. In industry [5], predictive maintenance serves to organize optimum utilization of industrial assets with conditional monitoring of data from different types of sensors. PdM also plays an important role to detect existing problems before scheduled inspection [4]. Apart from these, it plays a crucial role in Timely Repairing of Machines and Hazardous Outage Preventing.

A. V. Timofeev (✉)
LLP "EqualiZoom", Astana, Kazakhstan
e-mail: timofeev.andrey@gmail.com

V. M. Denisov
"Flagman Geo" Ltd., Saint-Petersburg, Russia

E. Pricop et al. (eds.), *Recent Developments on Industrial Control Systems Resilience*,
Studies in Systems, Decision and Control 255,
https://doi.org/10.1007/978-3-030-31328-9_3

In short words, the main goal of PdM is to predict at a particular time moment, using the various type data, which were collected up to this time moment, whether the equipment will fail in the close future [6]. According to PricewaterhouseCoopers [7], there are four levels of PdM:

- **Level 1**. Visual inspections: periodic physical inspections; conclusions are based solely on inspector's expertise.
- **Level 2**. Instrument inspections: periodic inspections; conclusions are based on a combination of inspector's expertise and instrument read-outs.
- **Level 3**. Real-time condition monitoring: continuous real-time monitoring of assets, with alerts given based on pre-established rules or critical levels.
- **Level 4**. PdM 4.0: continuous real-time monitoring of assets, with alerts sent based on statistical and AI-based predictive techniques including regression analysis and various automatic classification methods (ANN, SVM, XGBoost etc.).

In the following text, instead the name "PdM 4.0", we will use the name "**Machine Learning-based Predictive Maintenance** (**ML PdM**)" [8, 9], since it is greater extent corresponds to the essence of the present study. Namely ML PdM, which is a part of Industry 4.0 concept, will be in focus of this research.

To date, a number of industries are objectively ready for a large-scale implementation of ML PdM: thousands of sensors monitor equipment (phase: condition monitoring); data from these sensors are collected in special data banks (phase: data collection). This collected data can be leveraged for better maintenance practices, but it is not being fully leveraged or in many cases, it is ignored. The only way to exploit this high volume of data is using machine learning (ML) and other mathematical methods, which have been developed to solve various predictive problems. With special learning approaches, ML-algorithms are trained to detect abnormal and correlated patterns of abnormal sensor data. Based on this analysis, the ML-algorithms identify machine degradation or fault before they occur. Advanced ML-algorithms do not require rules or simplistic threshold setting, because it is looking at behavioral patterns. Vast amounts of data can be analyzed in real time without the need of human involvement, the more that people are simply not able to process such a huge amount of data. In simply words, ML PdM can be formulated in one of the four ways: (a) regression approach (predicts how much time is left before the next failure); (b) classification approach (predicts whether there is a possibility of failure in next n-steps); (c) flagging anomalous behavior; (d) survival models for the prediction of failure probability over time. In this study, which is very limited in scope, we will use only approaches (b) and (c). The application of ML PdM to objects located in the cryolithozone has certain specific features. First of all, these features are associated with the need to take into account the influence of instability of the foundations on the state of the controlled objects (buildings or equipment). Also, during the analysis of the state of objects of control it is necessary to take into account weather conditions that are extremely harsh in the cryolithozone. In addition, a critical feature of the practical operation of various metal structures (including pipelines) in the cryolithozone is the increased risk of the development of corrosion processes, which lead to accidents and loss of efficiency.

The paper discusses several very important practical problems related to the maintenance of the pipelines systems. It will also show how the ML-methods are used in the maintenance of pipelines systems in the cryolithozone. In particular, the following will be considered: new methods for reliable detection of leaks in pipelines; methods for the operational classification of the type of defect that led to a leak; new method for timely detection and evaluation of the stage of corrosion processes in pipeline, based on the joint use of high-precision methods for analyzing the pipeline vibroacoustic field (photon counting technology) and ML-methods for processing measurement data. There will also be considered a practically effective solution based on use of ML-methods and designed to optimize the incident tickets processing in the oil pipelines control systems.

2 Research Objectives

The aim of the study is to create a group of math-methods based on statistical and AI-approaches, which intended to solve following problems of the PdM for infrastructure facilities located in the cryolithozone:

- Reliable detection of leaks in pipelines located in cryolithozone using fiber optic monitoring systems. The solution to this problem is absolutely necessary for the early planning of the resources that are required to solve it.
- Operational classification of the type of defect that led to a leak. Solving this problem is necessary for optimal planning of the resources required to eliminate the leak.
- Timely detection and evaluation of the stage of corrosion processes in pipeline. This problem is relevant for the conditions of the cryolithozone, where corrosion processes are developing very intensively. Solving this problem will make it possible to predict and prevent the occurrence of leaks, which sooner or later will occur as a result of the development of local corrosion spots due to metal loss. It is necessary to monitor the rate of corrosion metal loss, which will allow implementing preventive measures in time.
- Optimizing the sequence of handling incident tickets in PdM systems for oil pipelines. This problem is extremely important for the Predictive maintenance of oil pipelines in cryolithozone conditions, since the promptness in servicing the most important incident tickets ensures the minimization of the risks of accidents, the elimination of which is extremely expensive in the permafrost zone.

3 General Concepts and Notations

- $o \in$ O is a maintenance object: pipeline, bridge, structure being in stress-strain state. Here O—denotes the entire set of service objects. Abbreviated: **MO**.

- $F(o)$. The volume of space actually occupied by the MO. For simplicity, we assume that $F(o)$ is a convex region.
- $p(o)$ is approximate value of the object perimeter $o \in \mathrm{O}$.
- $t_1 < t_2 < \ldots < t_k < \ldots$. The increasing sequence of time points, in each of which a decision is made on the state of the MO.
- $\Delta t_k = (t_{k-1}, t_k)$: ***k*-th time interval**: the time interval between the moments t_{k-1} and t_k. During this interval, information is accumulated that is needed to make a decision about the state of the MO at the time t_k. Otherwise, Δt_k is called the ***k-th* monitoring interval**.
- **Monitoring** the MO current state is one of the basic components of PdM system. Monitoring is carried out using a network of different types of sensors located on the object.
- **Monitoring Task**. For different types of MO, monitoring tasks are different. The work addresses the following tasks:

  - Leak detection in pipelines. Task code: “LD”
  - Automatic classification of the damage type through which pipeline leakage occurs. Task code: “CL”.
  - Detection of pipeline corrosion processes. Task code: “CP”.
  - Intelligent diagnosis of corrosion degradation state of a pipeline. Task code: “IDC”.

- $\Theta_{v_o}^{(B)}$ is a finite set of all possible values of status parameters characterizing MO ($\Theta_{v_o}^{(B)}$ is determined a priori). The identification of these parameters is the goal of the specific **monitoring task**. The composition of the set $\Theta_v^{(B)}$ depends on the object type (v) and on the monitoring task type (B—task code*)*. In this chapter, one type of MO is considered: “pipeline”. Object code: “PL”.
- $\Omega(x|o, x^c)$ is a **monitoring point**. $\Omega(x|o, x^c)$ defines a convex subdomain of $F(o)$ domain, the current state of which can be estimated within the existing technical capabilities of the monitoring system. $\Omega(x|o, x^c) \subseteq F(o)$. Here, $x^c \in F(o)$ is the geometric center of the $\Omega(x|o, x^c)$, x is the set of minimal projections of the point x^c on the object’s o surface (in general, x is a set). The coordinates of x^c are tied to the $F(o)$ in accordance with strict rules that are defined at the design stage of the monitoring system for the MO. The dimensions of the $\Omega(x|o, x^c)$ determine the spatial resolution of the monitoring system. By definition, for any real $o \in \mathrm{O}$, the number of possible monitoring points is limited and equal to the number of information channels of the monitoring system. In general, $\bigcup_k \Omega(x_k|o, x_k^c) \neq F(o)$. Further, for convenience (if this does not cause ambiguity), instead of $\Omega(x|o, x^c)$, we will simply write x. Example: MO is pipeline. Monitoring system: C-OTDR [10]. In this case, monitoring point $\Omega(x|o, x^c)$ is a cylinder “stretched” on a real pipeline, whose height is equal to the spatial resolution of the C-OTDR-system, x^c is a certain point in the pipeline center that coincides with the geometric center of the cylinder, x is a set of minimal x^c projections on the cylinder surface. In this case, the number of monitoring points is equal to the integral part from dividing the fiber optic sensor length by the value of spatial resolution of the monitoring

system. The number of monitoring points in this case is equal to the number of the F-OTDR system channels. The set of monitoring points is denoted by the symbol **x**.

- X is the space domain that the foundation of the object o occupies and which is monitored by the monitoring system. For example, for objects of type PL, the domain X is a straight parallelepiped with linear dimensions $M_L \times M_W \times M_D$. On the upper face of this parallelepiped is located a part of the object's structures (or the whole object together with its foundation) so that the boundaries of the object's basement recede from the boundaries by the values of $\pm M$ meters. Here M_L, M_W are the linear dimensions of the edges of the parallelepiped, bounding its upper face, which approximately coincides with the surface of the Earth, M_D is the linear size of the "vertical" edge of the parallelepiped, which determines the depth of its immersion into the ground.
- $\boldsymbol{\Delta}(X, o) = \{(\delta_i, t_i, x_i)_i | i = 1, \ldots N_\Delta\}$ are soil shifts that occurred within domain X and which were detected by the monitoring system. Here i—number of shift within the set $\boldsymbol{\Delta}(X, o)$; δ_i is the absolute value of the soil shift; $t_i \in \Delta t_k$ is the time moment of the shift (if the shift was gradual, this parameter is assumed to be t_k), $x_i^{(P)} \in X$ projection of coordinates of the shift center on the surface of the MO; N_Δ is the number of shifts in the monitoring domain X. In $\boldsymbol{\Delta}(X, o)$ not only the soil shifts recorded during the current monitoring period Δt_k are saved, but also those that occurred earlier, as they continue to affect the foundation of the MO. Over time, this effect decreases, which is taken into account by the coefficient $\gamma(x_i, t, t_i|\alpha)$, which is described in Sect. 6.1.
- $k(o)$ is seismic resistance coefficient MO $o \in \mathrm{O}$, $k(o) \in [0, 1]$.
- $\phi(x, t|\eta_\delta)$ is a generalized function of soil shift effect on the value of the probability of damage to the MO structure at a point $x \in X$ on its surface, at time t. Here $\eta_\delta = (\Delta, \alpha, a, b, A_0)$ is some parameters set (for more details, see in Sect. 6.1). The set $\{\alpha, a, b, A_0\}$ is determined as a result of a machine learning training or empirically.
- $E_o = \left\{e_i^{(o)}\right\}$ is the set of MO structural member (element) $e_i^{(o)}$, $o \in \mathrm{O}$. For extended objects (for example, for oil pipelines), the element $e_i^{(o)}$ is otherwise called a **section**; $F(e_i^{(o)})$ is the volume of space occupied by the $e_i^{(o)}$.
- $\lambda\left(e_i^{(o)}\right)$ is a parameter that, in the case of an extended object, indicates the frequency of accidents in section e_i.
- $S(t, x|\Delta t_k)$ is the so-called set of objective parameters (OP) of monitoring, defined for the subinterval of time $t \subseteq \Delta t_k, \cup t = \Delta t_k$, for a specific structural member $e_i^{(o)} \in E_o$, at the monitoring point $x \in \mathbf{x}, x \subseteq F(e_i^{(o)})$. This set consists of various parameters directly measured by the sensor system either once or repeatedly during the interval Δt_k. The specific set of parameters depends on the MO type. Examples of objective parameters: the angles of inclination of the structural elements, the natural frequencies of these elements, the parameters of the vibroacoustic field

of the structural elements, and so on. Based on the analysis of the OP values dynamics, the main monitoring tasks are effectively solved, the essence of which is to estimate the MO status parameters.

- **The MO vibroacoustic field parameters** are included in the set of objective parameters $S_{E_o}(t, x|\Delta t_k)$. In the process of monitoring with usage of a point microphone network, as well as data from fiber-optic monitoring systems, a group of parameters characterizing the frequency responses of the MO vibroacoustic field is calculated. These parameters are calculated at each monitoring point $x \in \mathbf{x}$ (channel), for the time interval $t \subseteq \Delta t_k$. This group includes:
  - $B_t(x, \omega_h)$ is the Fourier coefficients vector for the frequency range of $[0, \omega_h]$ (the value ω_h is determined by the frequency bandwidth of the monitoring system and lies within ranges from a few hundred hertz to several kilohertz);
  - $P_t(x|\omega_h)$ is the vector of cepstral coefficients;
  - $EV_t(x)$ is the vector of heuristic characteristics of the stochastic dynamics of the oscillatory process in each monitoring system channel, during the interval $t \subseteq \Delta t_k$, which takes into account the macrodynamics of these oscillations and the irregularity of their amplitude in time.

The parameters of the vibroacoustic field form a subset $s(x, t) = (EV_t(x), B_t(x|\omega_h), P_t(x|\omega_h))$ of the main set of objective parameters $S(t, x|\Delta t_k)$, which is used to solve the problems "LD", "CL" and "CP". Denote by $\mathbf{s}(x|\Delta t_k) = \{s(x, t)|t \in \Delta t_k\}$, $x \in \mathbf{x}$, the sequence of such parameter subsets repeatedly measured during each monitoring cycle by the subsystem of fiber-optic monitoring based on the count of single photons.

4 Specialties of ML PdM Processes in Cryolithozone Conditions

The PdM-processes for infrastructure facilities in the cryolithozone are significantly affected by the soil instability factor at the base of the foundations of these objects. This is due to the influence of freezing–thawing processes of foundation soils. These processes are intensified by the process of global warming. To take into account these circumstances, information of soil shifts in the infrastructure foundations location area must necessarily be included in the general information contour of PdM system. You should also consider the factor of low temperatures, changes in atmospheric pressure, as well as sharp daily fluctuations in air temperature, which are so characteristic of the cryolithozone. For example, the operability of trunk pipelines in conditions of low climatic temperatures is mainly determined by their cold resistance. Major damage to trunk pipelines in the cryolithozone occurs under static loading and leads to brittle, quasi-brittle and ductile fracture. In addition, corrosion processes are significantly accelerated in the cryolithozone. All of these factors can be taken into account by using of appropriate types of sensors, as well as by using of joint data processing peculiarities. In Fig. 1 schematically shows the principle of data processing in a ML PdM system of pipeline system management in the cryolithozone. This diagram displays the following levels of data processing: Sensors-Data Level, Sensors Data Transfer Level, Data Collection and Storage Level, ML PdM Decision Support System Level. To ensure the effective functioning of the ML PdM system for pipelines in the cryolithozone, it is necessary to collect and take into account a wide variety of data that fully describe both the state of the object itself and its environment. For this it is necessary to solve the following monitoring tasks:

- **The technological parameters monitoring of pipeline elements**. To solve this monitoring task the next sensor types are used: gauges, thermometers, vibrometers and etc. In Fig. 1 these sensors have number “1”.
- **Monitoring of soils and foundations**. To solve this problem, strings of inclinometers are used, which burrow into the ground near the foundations of objects. These

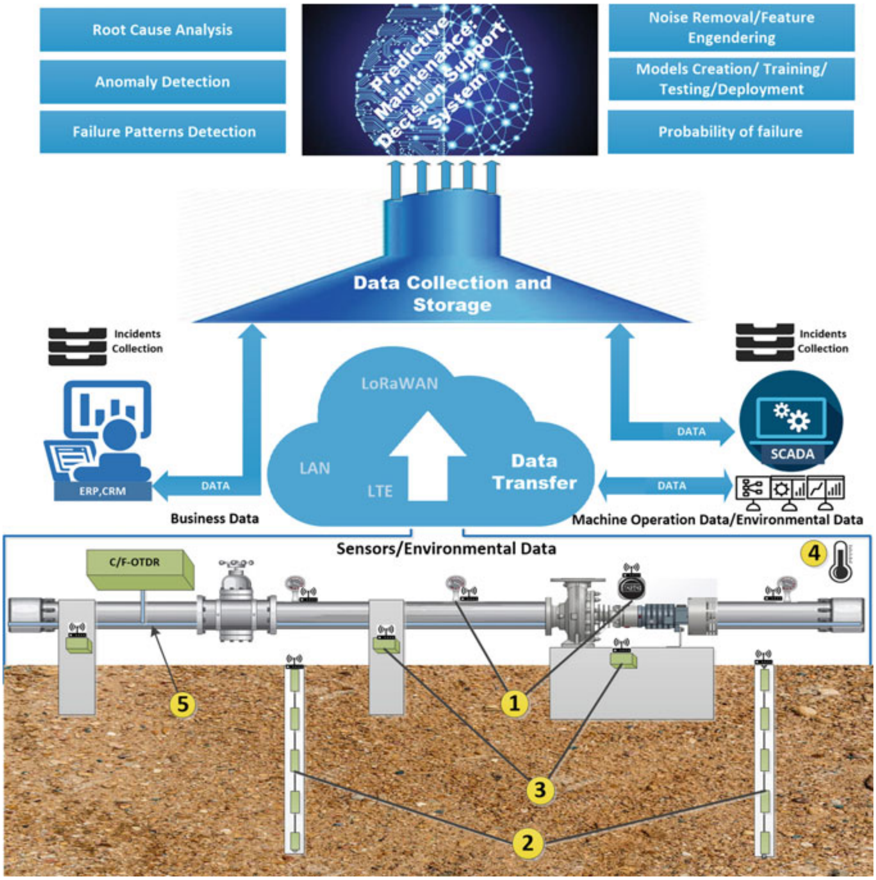
Root Cause Analysis
Anomaly Detection
Failure Patterns Detection
Predictive Maintenance Decision Support System
Noise Removal/Feature Engendering
Models Creation/ Training/ Testing/Deployment
Probability of failure
Data Collection and Storage
Incidents Collection
ERP,CRM
DATA
Business Data
LoRaWAN
LAN
LTE
Data Transfer
DATA
DATA
Incidents Collection
SCADA
Machine Operation Data/Environmental Data
Sensors/Environmental Data
C/F-OTDR
4
5
1
3
2

Fig. 1 Data processing in ML PdM systems

sensors make it possible to detect even minimal ground shifts in the foundation area. In Fig. 1 these sensors have number "2".

- **Monitoring of facilities and structures**. Networks of inclinometric sensors, as well as sensors for estimating the natural frequency of structural elements and the logarithmic decrement of these vibrations, are placed on the surface of objects, including its foundation and supports. The readings of these sensors must clearly correspond to the intervals of admissibility, determined at the stage of designing objects. In Fig. 1 these sensors have number "3".
- **Climate conditions are monitored** using temperature, humidity, pressure, and wind speed sensors. This data is used in the intellectual unit of ML PdM as a supplement to the feature set that characterize the targeted incidents occurring on the object. In the diagram, these sensors have number "4".
- **Monitoring of corrosion** processes can be done by various methods. This study will focus on vibroacoustic methods based on fiber-optic monitoring systems [10]. In Fig. 1 the fiber optic sensor has the number "5".

5 Models and Methods of ML-Based PdM Systems

Currently, many mathematical models and methods have been developed that are used in ML PdM processes. For example, models based on Markov processes and queuing theory are widely used. On the other hand, various variants of optimization strategies are used for optimal planning in PdM systems. Due to the limited scope of the chapter, we will focus only on some of these models. In particular, they will be briefly described:

- an approach based on analysis of targeted incidents;
- a method of interval estimation of the moment, at which a random process parameters changed (this approach has shown high efficiency in a number of practically important cases);
- some methods for multi-class classifying of stochastic objects that have proven themselves in solving PdM problems.

5.1 A Incident-Based Approach to the Intellectualization of PdM Processes

An **inciden**t is a collection of situations registered by the sensory system and bearing signs of a **pre-failure condition**. For example, an incident is a short-term increase above the normal temperature of a component of a system or a short-term increase above a normal pressure in a pipeline system or a short-term increase in the maximum allowable equipment voltage. Not every incident that carries part of the signs of a pre-failure condition actually corresponds to a failure. The fact is that real industrial systems are very complex and multi-connected. When trying to change the operating parameters of these systems, for example, in the process of repair or maintenance, some of the system's sensors may well give abnormal readings even if the system as a whole is in good condition. Predicting this kind of alerts is pretty hard. ML PdM systems provide special mechanisms for optimal response to these cases, which will be discussed in the following sections. The sequence of states of the incident life cycle is simple: (1) "new"; (2) "eliminated"; (3) "completed" and (4) "closed". In the process of managing an object, the corresponding process control systems constantly form a database of historical data of incidents. Each incident, in manual mode, is classified by the system operators upon completion of its life cycle. In Fig. 1 shows that incident data comes from both production control systems (SCADA) and business systems (ERP, CRM). As a result, the Incident Historical Data Base (IHDB) is formed, which underlies the intellectualization system of PdM processes. IHDB also retains all sorts of characteristics of incidents, which are otherwise called incident feature. The set of features, naturally corresponding to the incident of a particular type, will be called the patterns of this incident. Incident feature set include: sensor system data, time and place characteristics, climatic environment parameters etc. The IHDB also stores information about the feature dynamics which correspond to the time intervals preceding each specific incident. Thus, this database contains all the necessary information for training incident automatic classification system. This approach is the key to intelligently solving the following PdM problems: "Predicting Repairs" and "Predictive Maintenance to the Rescue". The PdM approach based on IHDB analysis using Machine Learning methods will be called the incident approach. In addition to this approach, there are other ways to PdM system organize. Especially common is the approach based on the use of semi-Markov degradation models, described below.

5.2 Semi-Markov Models of Parametric Degradation Description

Most of the degradation processes of MO target parameters are continuously monotonous. Despite this, the transition from one parametric status to another is quite adequately described by the semi-Markov model [11], the number of states of which is countable. A semi-Markov process is a random process that passes from one state to another in accordance with specified probability distributions. The residence time of the process in any state is a random variable. The distribution of this quantity depends both on this state and on the state to which the next process transition will take place. This model adequately approximates the processes of the MO structure components degradation. This adequacy is due to the fact that the procedure for identifying the current parametric status of MO is rather long. During the entire identification process, from the monitoring system view point, the current parametric status of MO remains unchanged. The parametric status of MO, with a certain probability, can be discretely changed only after the completion of the identification procedure. And status MO can remain the same. An example of the adequacy of the semi-Markov model in describing the degradation: the semi-Markov model well approximates the processes of monotonous metal loss during the development of corrosion processes in pipelines, as well as the processes of gradual changes in the natural frequencies of structures, which is due to its slow destruction.

Let $X(t)$ be a semi-Markov process with a finite set of states $N = \{\mu_1(t_1), \mu_2(t_2), \ldots.\}$, which has stepped trajectories with jumps at times $0 < t_1 < t_2 < t_2 < \ldots$. In most practically interesting cases, the set N is (not strictly) monotonic sequence of scalar quantities. For example, in the case of monitoring the development of a local corrosion center in a pipeline, we have:

$$N_{1,n}(t_1, t_n) = (\mu_1(t_1), \mu_2(t_2), \ldots, \mu_n(t_n)), \mu_1(t_1) > \mu_2(t_2) > \cdots > \mu_n(t_n),$$

here $\mu_1(t_1)$ is a pipeline wall thickness at the beginning time of the corrosion spots observation, $\mu_2(t_2)$ is the pipeline wall thickness which is correspond to the corrosion process next stage, that led to such the loss of metal $\delta(t_1, t_2)$ in the wall of the pipeline so that $\mu_2(t_2) = \mu_1(t_1) - \delta(t_1, t_2)$. Here $(\delta(t_1, t_2), \delta(t_3, t_4), \ldots \delta(t_{n-1}, t_n))$ is a deterministic sequence of scalar quantities that determine the quantitative difference between different pipe corrosion states. This sequence is determined a priori, at the PdM system setup stage, for each type of pipe. The values of the semi-Markov $X(t_n)$ process at the moments of "jumps" form a Markov chain with transition probabilities:

$$p(\mu_{n-1}(t_{n-1}), \mu_n(t_n)) = P(X(t_n) = \mu_n(t_n)|X(t_{n-1}) = \mu_{n-1}(t_{n-1})).$$

In turn, distributions of jumps moments $\{t_n\}$ are described as follows:

$$\begin{aligned}&P(t_n - t_{n-1} \le x, X(t_n) = \mu_n(t_n)|X(t_{n-1}) = \mu_{n-1}(t_{n-1})) =\\&p(\mu_{n-1}(t_{n-1}), \mu_n(t_n)) \cdot F_{\mu_{n-1}(t_{n-1})\mu_n(t_n)}(x)\end{aligned}$$

Here $F_{\mu_{n-1}(t_{n-1})\mu_n(t_n)}(x)$ is a distribution function. For most technical applications, the functions of $F_{\mu_{n-1}(t_{n-1})\mu_n(t_n)}(x)$ and $p(\mu_{n-1}(t_{n-1}), \mu_n(t_n))$ are a priori unknown, since they depend on many factors, some of which cannot be taken into account either at the analysis stage or at the stage of system operation. In this regard, the main task of the semi-Markov process state tracking system, is the timely evaluation of consecutive points of $T^{(n)} = \{t_2, t_2, \ldots, t_n\}$, at which the parametric status of this process changes. This problem can be solved by various methods. In practice, the process of $X(t)$ is almost never observed directly. Only the $z(t) = X(t) + \varepsilon(t)$ process is directly observed, where $\varepsilon(t)$ is a random process with known statistical characteristics that describes measurement errors. The method for this problem solution from the standpoint of interval estimation is briefly described in Sect. 5.3. An alternative to this approach is the use of methods for **automatic classification** of the MO current parametric status. For example, classical binary classifiers like SVM, RVM, logistic regression and others can be used to solve this problem. The classification decision is made according to the observations of the $X(t)$ -process during the time interval $\Delta t = [t_s, t_f] \subseteq \Delta t_k$. Let the MO be in the state μ_k at the time moment $t = t_s$. Then, during the interval Δt the observations of process $X(t)$ form a data set $\mathbf{X}(\Delta t) = \{X(t)|t \in \Delta t\}$. The hypothesis of which class corresponds to this set is tested: class μ_k or the class of the next state μ_{k+1}? At the stage of training the system, each of the classes $\mu_k \in N_{1,n}$, must be represented by the corresponding training set $\mathbf{X}^{(k)}(\Delta t)$. These data sets are collected either at the stage of pre-setting of the PdM system in a test site, or in actual operating conditions (if possible). Thus, for a sequence of parametric states $N_{1,n}(t_1, t_n)$, for each μ_k, the corresponding binary classifier $C_{k,k+1}: \mathbf{X}^{(k)}(\Delta t) \rightarrow \{\mu_k, \mu_{k+1}\}$ must be created and trained. To consistently solve classification problems $N_{1,n}(t_1, t_n)$ it is necessary to have a sequence of binary classifiers $(C_{1,2}, C_{2,3}, \ldots C_{n-1,n})$. Instead of a sequence of binary classifiers, we can use one multiclass classifier $C(N_{1,n})$ such that $\underset{k}{\forall} C(N_{1,n}): \mathbf{X}^{(k)}(\Delta t) \rightarrow N_{1,n}$. This approach is very convenient, but in case of insufficiently representative training sets, the multi-class classifier can be significantly inferior to the sequence of binary classifiers in reliability. This approach, as well as the incident one, makes it possible to effectively solve the “Predicting Repairs” and “Predictive Maintenance to the Rescue” tasks. Some examples of such solutions will be given below.

5.3 Interval Estimation of the Random Process Change-Point

Let the semi-Markov process $X(t)$ change its parametric status at a priori unknown moments of time $T^{(n)} = \{t_2, t_2, \ldots, t_n\}$. Observations are described by the following model $z(t) = X(t) + \varepsilon(t)$, where $\varepsilon(t)$ is a noise process with known statistical characteristics p_ε. In [10], a method was proposed for sequential interval estimation of the $t \in T^{(n)}$, which makes the decision according to the following rule:

$$\begin{cases} \text{If } Y(X(t)|\alpha, \beta) \geq b(p_\varepsilon) \text{ Then } t \in \Delta\mathbf{t} \\ \text{If } Y(X(t)|\alpha, \beta) < b(p_\varepsilon) \text{ Then } t \notin \Delta\mathbf{t} \end{cases}$$

Here $\Delta\mathbf{t}$ is the time interval with prescribed length $\boldsymbol{\delta} = |\Delta\mathbf{t}| > 0$, $Y(X(t)|\alpha, \beta)$ is some statistic, which depending on observations and predetermined values $0 < \alpha, \beta < 1$. Here α is a predetermined upper bound for the probability of making type I errors; β—is a predetermined upper bound for the probability of making type II errors; $b(p_\varepsilon)$—decision threshold. At the same time, the decision reliability is guaranteed in the following form:

$$\mathbf{P}(Y(z(t)|\alpha, \beta) \geq b|t \notin \Delta\mathbf{t}) \leq \alpha;$$
$$\mathbf{P}(Y(z(t)|\alpha, \beta) < b|t \in \Delta\mathbf{t}) \leq \beta.$$

Under certain conditions, which, as a rule, are carried out in practice, the properties of this procedure are strictly proved (Theorem 4.2 [10]). In general case, observations are not scalar. In this situation, the described procedure is implemented for each component of the observation vector independently. The decision that $t \in \Delta\mathbf{t}$ is made when, at least for one of the single components of the vector of observation, have place inequality $Y(\bullet|\alpha, \beta) > b$. For convenience, we denote the procedure of interval estimation of the moment t as follows: $\mathbf{IE}(\{z(t)\}|\mathbf{PR})$. Here $\mathbf{PR} = (z, \alpha, \beta)$ is the parameters set, that define the properties of this procedure. Also valid entry: $\mathbf{IE}(\{z(t)\}|\mathbf{PR}) = \Delta\mathbf{t}(z)$, where $\Delta\mathbf{t}(z)$ is the time interval of a given length z, for which the assertions of Theorem 4.2 will be satisfied [10]. Thus the interval $\Delta\mathbf{t}(z)$ is the confidence interval for the parameter t.

5.4 Some Methods for Classifying Objects

In context of the PdM, different approaches to classification problems solutions are in demand. Many classification algorithms are relevant for PdM due to the fact that in various practical cases we have significantly different conditions and restrictions. There are situations when we have only training sample of small volume, and vice versa: there are situations when training sample volume is large, but it is littered with false data or statistical outliers. In this connection, we must use different methods to solve the different classification problems in PdM. In this section we shortly describe some basic classification methods which provide the acceptable effectiveness in various PdM practical cases. In order to save space, we will not give a complete mathematical basis of the used classification methods, especially since the study of the mathematical details of various classifiers is not the goal of this work. We confine ourselves to the description of the main features of the application of some classification methods to the tasks of the PdM.

Multiclass SVM (Support Vectors Machine). This is a well-studied and proven in various practical applications, which has already become a classic method. The SVM mathematical basic is fully described in [12]. The positive features of SVM include the following properties: accuracy; SVM works well on smaller cleaner datasets; SVM works well with data of high (more than 100) dimensions; SVM has a minimal set of hyper-parameters (regularization parameters and others), the essence of which is clear and understandable; SVM is defined by a convex optimization problems (no local minima) for which we have many efficient methods. The main SVM disadvantage is that it's not very suited to larger datasets as in this case the SVM training time can be unacceptable high. Also SVM is less effective on noisier datasets with overlapping classes. We have good experience of usage multi-class SVM in next classification problems of the **PdM** context: class leak identification, corrosion process stage identification in frame of the RrM-procedures for pipelines, incident class identification.

XGBoost (Extreme Gradient Boosting). XGBoost is a very popular implementation of Gradient Boosting. Ever since its introduction in 2014, XGBoost has been lauded as one of the most efficient classification algorithms: various teams repeatedly became winners of machine learning competitions using XGBoost [13]. XGBoost is an ensemble learning method. This algorithm provides the best trade-off between bias-related errors and variance-related errors. In contrast to SVM, XGBoost has quite a bit of hyper-parameters, which must be identified very carefully in the learning process. If these seven hyper-parameters determined non-optimal the model will be overfitted. The main XGBoost advantages: extremely fast (due parallel computation); very effective even on large datasets; versatile (can be used for classification and regression); do not require feature engineering (missing values imputation, scaling and normalization). The main disadvantages: XGBoost only works with numeric features (XGBoost cannot handle categorical features by itself, it only accepts numerical values); if hyper-parameters are not tuned properly its leads to overfitting. Despite some shortcomings, XGBoost shows excellent results in analyzing big data, including in the incident type classification tasks (context PdM). The classical **Gradient Boosting** (GB) may be used too, but GB always will show worse performance in comparison with XGBoost.

ANN (**Artificial Neural Net**). We mention this widely discussed method simply because even those who have never practiced intellectual data processing have heard about it. The mention of this method in scientific and especially in popular science literature has become so frequent, although in many cases at the same time inappropriate, that it has generated a whole wave of unhealthy speculation around this technology. At some point it might even seem that the ANN is almost a panacea and is able to effectively solve all tasks in the field of intellectual data processing. In actual practice, this is not quite true. Indeed, in the field of image classification, as well as in some other applications, the ANN shows very good results. On the other hand, in those problems where it is necessary to work with training datasets of relatively small volume, ANN does not have an advantage over other classification methods, for example, over the XGBoost. But at the same time, ANN learns much more slowly than, for example, modern ensemble methods (XGBoost, CatBoost, LightGBM). Therefore, for each specific problem, in practice, the most convenient and adequate method of classification is chosen. And not always the model of ANN has advantages in this choice. In our practice of solving **PdM** problems, as a result of testing and comparing, ANN-model always lost to other methods of classification. Probably, this situation is due to the fact that we never had large data sets for training: in **PdM** applications, the size of the training sample rarely exceeds 300–500.

6 ML-Based PdM of Oil Pipelines in the Cryolithozone

This section describes solutions to some specific ML PdM tasks that are relevant to the maintenance of oil pipelines in the cryolithozone.

6.1 Leaks Detection Task in Cryolithozone Context

If PdM is successfully implemented on the pipeline, then one of its main objectives will be to minimize the leakage from the pipeline. In this regard, as part of the implementation of PdM procedures, a whole range of measures should be implemented to prevent the occurrence of a leak. However, the likelihood of a leak will remain non-zero even if PdM is used. For example, a leak may occur due to force majeure (sabotage, technological incident) or due to hidden factory defect of pipeline equipment. Therefore, pipelines should be provided with leak detection systems, which mainly play a controlling role. When deciding on leak detection, the likelihood of errors of the first and second kinds should be below the limits set by the relevant industry standards. Therefore, leak detectors will always be part of PdM systems for pipeline monitoring. The peculiarity of the solution of the leak detection problem from pipeline systems in the cryolithozone is the influence of the instability of the pipeline foundation. If in the area of the location of the pipeline foundation there was a significant soil shift due to thermokarst processes, this shift can significantly deform the pipeline design and conditions will be created for its destructive changes to occur. And this, in turn, may cause a leakage of the transported agent from the pipeline. In these circumstances, to correctly solve the leak detection problem, it is necessary to take into account information on the dynamics of soil shifts in vicinity of the pipeline construction foundation.

A system for detecting leaks will be referred to as a "leak detector" and will be denoted as follows: $D_L(\mathbf{s}(x|\Delta t_k))$. The source data for the leak detector $D_L(\mathbf{s}(x|\Delta t_k))$ is a sequence of subsets $\mathbf{s}(x|\Delta t_k)$ collected using a single-photon fiber optic monitoring system. Ordinary C/F-OTDR system can also be used, but the spatial resolution will be slightly worse (1–5 m vs. 2–3 cm).

Denote by τ_L—the leakage occurrence moment. The accuracy of the leak localization is determined by the size of the monitoring point $\Omega(x|o, x^c)$. For simplicity, we assume that the hole through which leakage of the transported agent is realized belongs to a single monitoring point, which actually determines the geometric dimensions of the monitoring system channel. Thus, the leak must be detected in one (or in several) channels $x \in \mathbf{x}$ (monitoring points) of the monitoring system. General requirements for systems of this type are set forth in the standard practice of the American Society for Testing and Materials [14], as well as in the recommendations of the European Committee for Standardization E1211-97 [15]. The proposed approach allows evaluating the status of the vibroacoustic field throughout the object. In fact, this is a continuous monitoring of whole physical body of the object. This provides spatial resolution in linear coordinates from 0.5 m to 5 m, depending on the system settings. This feature is a significant advantage due to the use of a fiber-optic monitoring system based on the single-photons counting technology. When deciding whether there is a leak, we work not with individual points in time, but with finite time intervals $u_z \subseteq \Delta t_k$, with length z. In this case, the decision is made according to the results of observations of the data measured during the interval $u_z \subseteq \Delta t_k$. According to Sect. 3, this data set is $\mathbf{s}(x|\Delta t_k)$. A priori, there are two hypotheses:

- **First hypothesis** (status: **no leak**): $L_0(u_z, x)$: **no leak** in the interval u_z, in monitoring point (channel) $x \in \mathbf{x}$: event ω_{0L}: $\tau_L \notin u_z$.
- **Second hypothesis** (status: **leak**): $L_1(u_z, x)$: **leak** in the interval u_z, in monitoring point (channel) $x \in \mathbf{x}$: event ω_{1L}: $\tau_L \in u_z$).

Here $\Theta_{PL}^{LD} = \{L_0(u_z, x), L_1(u_z, x)\}$ is a set of status parameter values. According to the results of the analysis of the set of objective parameters $\mathbf{s}(x|\Delta t_k)$, collected during the k-th monitoring interval Δt_k, at the time t_k of this interval completion, the leak detector, $D_L(\mathbf{s}(x|\Delta t_k))$, decides which of the hypotheses of the set Θ_{PL}^{LD} is most likely. In the general case $u_z \subseteq \Delta t_k$, but in practice we often have $u_z = \Delta t_k$. Anyway, $\max\{t'|t' \in u_z\} = t_k$. The leak detector $D_L(\mathbf{s}(x|\Delta t_k))$ generates a solution in two phases. The first phase is implemented on the principle of interval estimation, described in Sect. 5.3. This approach is very economical computationally. In the case when the signal-to-noise ratio is small (less than 2–3 dB), for this simplicity and efficiency we will have to accept a significant width of the confidence interval, which contains the moment τ_L. The length of this interval can be from several seconds to several minutes. Based on the readings of the sensors placed directly on the surface of the object, the probability of the event ω_{1L}: $\tau_L \in u_z$ is calculated at the monitoring point $x \in \mathbf{x}$ (actually, the LD task).

At the same time, a priori information about the state of the object, statistics of past malfunctions and leaks, as well as data on soil shifts in vicinity of the object's construction are **not taken into account**. The output of the detector $D_L(\mathbf{s}(x|\Delta t_k))$ at this stage is a two-dimensional vector $\mathrm{P}_{LD}^{(I)}(t_k, x) = (P(\omega_{0L}), P(\omega_{1L}))$, whose components are estimates of the probabilities (reliabilities) of hypotheses from the set Θ_{PL}^{LD} under the condition of observations $\mathbf{s}(x|\Delta t_k)$ (without taking into account various a priori information about the state of the object). Naturally, that $P(\omega_{0L}) + P(\omega_{1L}) = 1$.

Taking into account the notation introduced in Sect. 5.3, the interval estimate of the moment τ_L will be obtained in the form of a time interval $u_z = \mathbf{IE}(\mathbf{s}(x|\Delta t_k)|\mathbf{PR}) \subseteq \Delta t_k$ with length z. Here $\mathbf{IE}()$ is an interval estimation procedure, $\mathbf{PR} = (z, \alpha, \beta)$ is a set of quality parameters of the procedure $\mathbf{IE}()$.

In the case when the signal-to-noise ratio for each point $x \in \mathbf{x}$ significantly different (by more than 3–5 dB), for each $x \in \mathbf{x}$ you can choose different values of the width of the confidence interval z_x. The processing principles remain the same, but $\forall x1 \neq x2 \in \mathbf{x}$: $z_{x1} \neq z_{x2}$. In view of Theorem 4.2 [10], it is not difficult to see that $\mathbf{P}(\tau_L \in u_z) > 1 - \alpha$. Accordingly, $\mathbf{P}(\tau_L \notin u_z) \leq \alpha$. Therefore, taking into account the fact that $u_z \subseteq \Delta t_k$, the following entry is valid: $\mathrm{P}_{LD}^{(I)}(t_k, x) = (\alpha, 1-\alpha)$.

Thus, the **first phase of detection is completed** and the **second phase begins**, the purpose of which is to correctly account for a priori information within the framework of the **Bayesian paradigm**. In the absence of any priori information about the object state, the hypothesis $L_0(u_z, x)$ and $L_1(u_z, x)$ are equally probable.

Therefore, a priori probabilities of hypotheses from the set Θ_{PL}^{LD} are equal, that is: $P_A(L_0(u_z, x)) = P_A(L_1(u_z, x)) = 0.5$. A priori information, which can significantly affect leak detection performance, is of various types. Including:

- data on soil shifts that occurred near the base of the MO foundation;
- $\lambda_i = \lambda\left(e_i^{(o)}\right)$ is the frequency of accidents in the pipeline section $e_i^{(o)}$ to which point $x \in \mathbf{x}$ belongs, provided that for the whole object o the average intensity of accidents $\tilde{\lambda}_o$ is determined;
- When point $x \in \mathbf{x}$ (PL type object) belongs to the so-called risk zone. Examples of risk zones: gas pipeline sections after compressor stations (5 km, risk factor: non-stationary dynamic loads); sections of gas pipelines in the connection points; sections of underwater crossings; pipelines sections with high anthropogenic activity.

There may also be significant other a priori information that characterizes the $e_i^{(o)}$ section, for example, the frequency of passage through this section of the cleaning projectile (PIG). The passage of PIG, in some cases, can provoke the occurrence of a hole through which leakage can occur, for example, in the presence of a factory marriage or in the presence of a developed corrosion hearth. Also important is the life of the pipeline and life of its structural elements, as well as brand of the structural elements manufacturer.

As part of this work, data on soil shifts, which are so frequent in the cryolithozone, as well as data on accident statistics at pipeline sections, are especially important for us. It will be show how this information is taken into account when a leak is detected.

So, a priori information about the presence of soil shifts that occurred near point $x \in \mathbf{x}$ (PL type object) is taken into account using the value of influence function $\phi(x, t|\eta_\delta)$. The purpose of this function is to smoothly approximate the effect of soil shift on the likelihood of destructive processes in the object's body. The main requirements for this function are as follows: the closer to the point $x \in \mathbf{x}$ was a the soil shift (determined by the value of $\|x - x_i\|$, x_i—is the i-th shift coordinate), the larger the amount of this shift $\delta_i > 0$, and the less time occurred since its inception (determined by the value of $t - t_i > 0$, here t_i is the moment of the i-th shift), the higher should be the a priori probability of destruction of the object's structure at the point $x \in X$ at the moment of time t. And vice versa: the larger the values of $\|x - x_i\|$, $(t - t_i)$ and the smaller the magnitude of the shift δ_i, the lower the prior probability of destruction of the object's structure to the point $x \in \mathbf{x}$ at the moment of time t. The general view of the function $\phi(x, t|\eta_\delta)$ is shown in Fig. 2. Considering the fact that the soil shift sensors are located in the immediate vicinity of the foundation, the distance from the shift center to the foundation can be neglected. The function $\phi(x, t|\eta_\delta)$ must be differentiable in all its arguments, and $\underset{x,t}{Max}(\phi(x, t|\bullet)) = 1, \phi(x, t|\eta_\delta) > 0$. There are many functions that satisfy these requirements, including, for example, the following function:

$$\phi(x, t|\eta_\delta) = N^{-1} \sum\nolimits_{i=1}^{N} W(\delta_i|A_0, o)\mu(x - x_i|a, b)\gamma(x_i, t, t_i|\alpha),$$

$$\eta_\delta = (\Delta, \alpha, a, b, A_0), \mu(x - x_i^{(P)}|a, b) = \left(1 + \left(\left\|x - x_i^{(P)}\right\| \cdot a^{-1}\right)^{2b}\right)$$

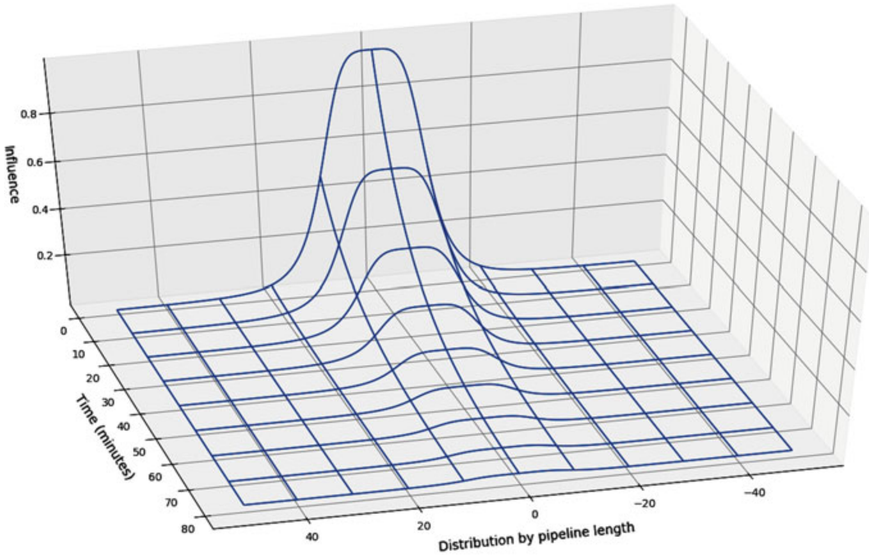

Fig. 2 The influence function $\phi(x, t|\eta_\delta)$

$$\gamma(x_i, t, t_i|\alpha) = \begin{cases} \exp(-\alpha|t_i - t|), & t \geq t_i \\ 0, & t < t_i \end{cases},$$

$$W(\delta_i|A_0, o) = 1 + \ln\left(1 + A_0 \cdot \delta_i^2\right)k^{-1}(o)$$

Here, the parameters α, a, b, A_0 are determined at the stage of the system setting based on the data of the field tests, expert estimates and the results of computational experiments.

Consider the matrix function:

$$\Gamma(x, t_k|\eta_\delta) = diag\left(\frac{P_A(L_0(u_z, x))}{L(x, u_z)}, \frac{P_A(L_1(u_z, x)) \cdot \phi(x, u_{\max}|\eta_\delta)}{L(x, u_z)}\right),$$

where $L(x, u_z) = P_A(L_0(u_z, x)) + P_A(L_1(u_z, x))\phi(x, u_{\max}|\eta_\delta), u_{\max} = \max\{x \in u_z\}$.

In the case when no additional information was involved, the prior probabilities of the hypotheses from the set Θ_{PL}^{LD} are equal, the form of this function is simplified:

$$\Gamma(x, t_k)|\eta_\delta) = diag\left((1 + \phi(x, u_{\max}|\eta_\delta))^{-1}, \phi(x, u_{\max}|\eta_\delta) \cdot (1 + \phi(x, u_{\max}|\eta_\delta))^{-1}\right).$$

Accounting for information on the soil shifts contained in the set $\mathbf{\Delta}(X, o)$, as follows: $\mathbf{P}_{LD}(t_k, x) = \mathrm{P}_{LD}^{(I)}(t_k, x)\Gamma(x, t_k|\eta_\delta)$. Components of vector $\mathbf{P}_{LD}(t_k, x)$ carry information about the effect of the soil shifts that occurred at the instants of time preceding the time t_k in vicinity of $x \in \mathbf{x}$ on the magnitude of the leakage probability.

Accounting for a priori information on accident statistics at the pipeline section, which at the monitoring point $x \subseteq e_i^{(o)}$ is set as a parameter λ_i, looks like this:

$$\Gamma_i(x, t_k)|\eta_\delta, \eta_\lambda) = diag\left(\varphi_\delta^{(1)} \cdot W\left(\tilde{\lambda}_o|\eta_\lambda\right), \varphi_\delta^{(2)} \cdot W(\lambda_i|\eta_\lambda)\right),$$

$$\varphi_\delta^{(1)} = \frac{P_A(L_0(u_z, x))}{P(\tilde{\lambda}_o, \lambda_i, \eta_\delta, \eta_\lambda)}, \varphi_\delta^{(2)} = \frac{P_A(L_1(u_z, x)) \cdot \phi(x, u_{\max}|\eta_\delta)}{P(\tilde{\lambda}_o, \lambda_i, \eta_\delta, \eta_\lambda)},$$

$$W(\lambda_i|\eta_\lambda) = 1 + A_2 \cdot \ln(1 + A_1 \cdot \lambda_i),$$

$$P(\tilde{\lambda}_o, \lambda_i, \eta_{\delta\lambda}) = P_A(L_0(u_z, x))W\left(\tilde{\lambda}_o|\eta_\lambda\right) + P_A(L_1(u_z, x))\phi(x, u_{\max}|\eta_\delta)W(\lambda_i|\eta_\lambda)$$

$$\eta_{\delta\lambda} = (\Delta, \alpha, a, b, \lambda_i, A_0, A_1, A_2, o), \eta_\lambda = (A_1, A_2).$$

The output of the detector $D_L(\mathbf{s}(x|\Delta t_k))$, in this case, is the following vector: $\mathbf{P}_{LD}(t_k, x) = \mathrm{P}_{LD}^{(I)}(t_k, x)\Gamma_i(x, t_k)|\eta_\delta, \eta_\lambda) = (P(L_0(u_z, x)), P(L_1(u_z, x)))$, whose components are estimates of the probabilities of hypotheses from the set Θ_{PL}^{LD} under the condition of observations of $\mathbf{s}(x|\Delta t_k)$ and taking into account various a priori information about the state of the object (in this case, taking into account information on the soil shifts and accidents frequency in vicinity of $x \subseteq e_i^{(o)}$). For simplicity, we assume that among the components $< \mathbf{P}_{LD}(t_k, x) >_k$ of vector $\mathbf{P}_{LD}(t_k, x)$ there is a single maximal component. The final solution is as follows: $\theta = Arg \underset{k}{Max} < \mathbf{P}_{LD}(t_k, x) >_k$—is the index of the maximum component. Accordingly, the hypothesis $L_\theta(u_z, x) \in \Theta_{PL}^{LD}$ will be chosen as true. A priori information of other types can be taken into account in a completely similar way.

6.2 Automatic Classification of the Damage Type Through Which Pipeline Leakage Occurs

The type of leakage and its intensity fundamentally determine the response of the PdM system to this incident. To optimize costs and properly plan mitigation actions for this incident, it is very important to determine the type of leak. For these purposes, a system is used which we will call the leakage classifier and denoted as follows: $C_L(\mathbf{s}(x|\Delta t_k))$. As a source of data, this system uses a set of objective parameters $\mathbf{s}(x|\Delta t_k)$. The set of status parameters for this task (CL code) is as follows: Θ_{PL}^{CL} = {“*H*”, “*S_Cr*”, “*M_Cr*”, “*G_R*”}. The set of status parameters for this task (CL code) is as follows: W. Here “*H*” –hole, “*S_Cr*” is a small crack, ”*M_Cr*” is a middle crack, “*G_R* ”- guillotine pipe rupture.

Total, we have four main classes of leakage. The validity of just such a composition of the set Θ_{PL}^{CL} is confirmed in a number of works, for example, in [16, 17]. Consider this issue in more detail. Let L be the characteristic linear size of the defect, m; D be the nominal diameter of the pipeline; S_0 be the cross-sectional area of the pipe, m^2; S_e be the equivalent area of the defective hole, m^2; f_L be the conditional probability of occurrence of a defective hole with a characteristic size L. Table 1 [16,

Table 1 Defective holes parameters

Defective hole parameters	Defective hole type			
	“*H*”	“*S_Cr*”	“*M_Cr*”	“*G_R*”
L/D	≪0.3	0.3	0.75	1.5
S_e/S_0	$\leq 10^{-4}$	0.0072	0.0448	0.179
f_L	0.7	0.165	0.105	0.030

17] shows the results of field studies that prove the existence of a stable statistical relationship between a set of values $(L/D, S_e/S_0)$ and the probability of occurrence of a certain type (class) defect. This is an extremely important result, which significantly simplifies the system training procedure, and also potentially improves the efficiency of the classifier, by providing the possibility of taking into account the value of a priori probabilities $\{f_L\}$ for each class of leakage. The very mechanism of the account of this type of a priori information may be various, for example, it may correspond to the scheme set out in Sect. 6.1.

Let the event “leak” be characterized by the following two features, forming a pair $(L/D, S_e/S_0)$. The information presented in the table, in fact, postulates that if in a given feature space an infinite stream of “leakage” events will be realized (corresponding to physically realizable situations on real pipelines), then three clusters with centers $d_{S\_Cr} = (0.3, 0.0072)$, $d_{M\_Cr} = (0.75, 0.0448)$ and $d_{G\_R} = (1.5, 0.179)$ in this space will be formed. These clusters will correspond to three types of defect, respectively: “*S_Cr*”, “*M_Cr*”, “*G_R*”: All other events will correspond to relatively low-power and more frequency type defects “*H*”.

This information provides opportunities for effective solution of the classification problem. In this case, the solution will be to simulate on the test site only four types of defects (all types from Θ_{PL}^{CL}), leading to a leak. For each of the above-mentioned clusters, a set of realizations of the defect $\{d_i^j | i \in \Theta_{PL}^{CL};\ j = 1, \ldots N_i\}$, is formed, with the powers $N_i = 1 \ldots 50$ for each class (the value of N_i depends on the type of class). For classes “*S_Cr*”, “*M_Cr*”, with which, usually, there are no problems with their physical modeling, $N_i = 50$ (determined by the capabilities of the test site: the more, the better). For class “*H*”, the cluster center of which is not defined due to the large number of variants of defects of this type, it is sufficient to provide $N_i = 5 \ldots 10$ realizations of the defect, for example, with sizes $0.1 \times 10^{-4}\ \text{m}^2$ and $0.01 \times 10^{-4}\ \text{m}^2$. For the class of powerful events “*G_R*” one size is sufficient, for example, coinciding with the center of the corresponding cluster. The fact is that the events of the “*H*” and “*G_R*” classes are radically different from the classes in terms of the power of vibroacoustic emission, therefore, they are classified quite reliably when the parameter “power of vibroacoustic emission per unit of time” is included in the feature space. All implementations of imitation defects created in this way at a test site differ by at least the value of one parameter.

The values of the characteristic parameters of imitating defects, first of all it concerns the classes “*S_Cr*”, “*M_Cr*”, must uniformly cover the area of cluster scattering, which is determined by the dispersions of the scattering of its components.

That is, $\forall i \forall j, k : ||d_i^j - d_i^k|| \geq \varepsilon$, for some $\varepsilon > 0$. For each realization of a defect, it is necessary to carry out the agent being transported under various pressures corresponding to the actual modes of pipeline operation. For reasons of economy, a measurement cycle, the purpose of which is to form a training set, can be carried out for the smallest size of the pipeline. As in the case of solving a problem of the LD type, for solving the CL problem, observations $\mathbf{s}(x|\Delta t_k)$ of the vibroacoustic field of the object are used. At the stage of training the system under the conditions of the test site, a training set of the following structure is formed:

$$S_L = \{([s(x,t), P_k]; d_i^j)|t \in \Delta t_L, P_k \in \mathbf{P} = \{P_1, P_2, \ldots, P_L\},\\ i \in \Theta_{PL}^{CL};\ j = 1, \ldots N_i\}.$$

Here $\mathbf{P} = \{P_1, P_2, \ldots, P_L\}$ is the map of the model values of pressure in the pipeline, Δt_L is the training interval of the system. In the $([s(x,t), P_k]; d_i^j)$, the d_i^j element is marker and $f_p(t) = [s(x,t), P_p]$ element are features. Consider the following partitioning of the set S_L: $S_L = \bigcup_p S_L^{(p)} = \bigcup_p \{f_p(t); d_i^j)\}$. Each of the subsets of $S_L^{(k)}$ corresponds to a specific value of P_p. Each of these subsets is used to train the classifiers described in Sect. 5.2. To increase the generalizing ability, the methodology of cross-validation is used. Thus, we have $|\mathbf{P}|$ classifiers C_p: $f_p(t) \to \Theta_{PL}^{CL}$, each of which corresponds to a certain value of $P_p \in \mathbf{P}$. In real conditions, there is always information about the current value of pressure P in the pipeline. Therefore, from $\mathbf{P}$ we choose two adjacent values $P_p, P_{p+1} \in \mathbf{P}$ such that $P_p \leq P \leq P_{p+1}$, and solve the classification problem simultaneously for the classifiers C_k and C_{k+1}. Based on the hypothesis of the monotony of the effect of pressure in the pipeline on the parameters of the vibroacoustic field created by the movement of the transported agent, the solution of the CL problem is the class of Θ_{PL}^{CL} that has the highest probability value assigned to it by these classifiers.

6.3 Intelligent Diagnosis of Corrosion Degradation State

Detecting and tracking the dynamics of corrosion processes in pipeline systems is one of the most relevant in the PdM tasks group. For example, according to "Rostekhnadzor" reports, pipe metal corrosion is the main cause of accidents on Russian pipelines [18]. In this section, we will consider a new method for detecting and controlling the development of corrosion spots in pipelines, based on the use of fiber-optic monitoring using the single-photon counting technology in a receiving unit. The basis of the proposed method is the effect of focal metal loss due to corrosion on the dynamics of parameters $\mathbf{s}(x|\Delta t_k)$. The problem is to restore the function of this influence as accurately as possible. To do this, it is necessary to estimate the parameters of the vibroacoustic field of the pipeline with a high degree of spatial resolution (2–3 cm along the linear coordinate), since the size of corrosion centers has a centimeter scale. This high resolution can be achieved using the C\F-OTDR system, the receiving unit of which is based on the principle of single photon counting. In addition, it is extremely important to create a training data sets that link models of corrosion foci of various shapes and depths with observations of $\mathbf{s}(x|\Delta t_k)$. From practical studies it is known [19] that the maximum size of corrosion areas does not exceed 15–20 cm in linear size and has a shape similar to an ellipse. Under the conditions of the test site, artificially create a series $\Pi = \{\pi_i\}$ of mechanical depressions of ellipsoidal shape in the pipeline metal (with a maximum major axis size of 20 cm, with a certain proportion of the axle length ratio), with different depths $H = \{\eta_k\}$. Here, for predefined values $\varepsilon_\Pi > 0$, $\varepsilon_H > 0$, $\pi_i, \pi_i \in \Pi$: $|\pi_i - \pi_i| \geq \varepsilon_\Pi$, $\eta_i, \eta_i \in \mathrm{H}$: $|\eta_i - \eta_i| \geq \varepsilon_\mathrm{H}$. The smaller the values ε_Π, ε_H, the more accurately reproduced many possible corrosion defects, but the more expensive the full-scale experiment becomes. Thus, a set of markers $\Pi \otimes H$ is formed, each element $(\pi, \eta) \in \Pi \otimes H$ of which corresponds to a certain configuration of a model defect. The pressure in the pipeline is taken into account in the same way as in Sect. 6.2. As a result of such field experiments, the learning set S_L is formed. Here

$$S_L = \{([s(x,t), P_p]_j; (\pi, \eta))|t \in \Delta t_L, P_p \in \mathbf{P}, (\pi, \eta) \in \Pi \otimes H, j = 1, \ldots N\},$$

$\mathbf{P} = \{P_1, P_2, \ldots, P_L\}$ is a map of the model values of pressure in the pipeline, Δt_L is the system learning interval, N is the sample size of measurements $[s(x,t), P_p]$ for each defect configuration (π, η). In the $([s(x,t), P_p]_j; (\pi, \eta))$ element, component (π, η) is the defect configuration marker, and $f_p(t) = [s(x,t), P_p]$ is the features vector. Consider the sets $\{f_p(t); (\pi, \eta)\}$, $p = 1, \ldots L$, each of which corresponds to a specific value of $P_p \in \mathbf{P}$. On each of these sets, using the cross-validation method, one of the classifiers described in Sect. 5.2 is trained. Thus, we have $|\mathbf{P}|$ classifiers C_p: $f_p(t) \to \Pi \otimes H$. The problem of the mismatch of the actual pressure in the pipeline P to the values of the map $\mathbf{P}$ is solved in the same way as this problem is solved in Sect. 6.2. Using this approach, in essence, creates the basis for the automatic solution of the "Predicting Repairs" and "Predictive Maintenance to the Rescue" tasks in relation to pipeline systems.

6.4 Optimizing the Sequence of Handling Incident Tickets in ML PdM Systems for Oil and Gas Pipeline

The features of the PdM processes for the pipeline infrastructure in the cryolithozone are due to both the influence of low ambient temperatures and possible soil shifts caused by the dynamics of thermokarst. The main tasks of PdM for pipeline infrastructure include monitoring the technical status of the equipment, forecasting the condition of the equipment, taking into account many factors, as well as planning timely maintenance and the maintenance procedures types. The following parameters are subject to control:

- pressure (hydrodynamic, static), measured at control points of the pipeline system;
- valves status (discrete parameter);
- temperature of the transported agent in the control points;
- density of the transported agent;
- the amount of transported agent passed through the pipeline for the period;
- technical parameters of pumping equipment (engine current, temperature, vector of vibration characteristics, working status, condition of fans, gas content parameter, heater condition, etc.);
- temperature and pressure in valves;
- vector of valves vibration characteristics;
- air temperature;
- precipitation in the area of infrastructure;
- shifts and temperature of the soil in the area of the foundations;
- angles of inclination of structural elements;
- vector of natural oscillations of structural elements;
- information on the condition of the anticorrosion coating of the pipeline;
- equipment failure statistics;
- emergency statistics (with technical details);
- statistics on the maintenance (with technical details);
- and others.

In addition, information about the state of the vibroacoustic field of the pipeline along its entire length, as well as information about automatically detected leaks and the places of origin of corrosion centers, are regularly received from the fiber-optic monitoring system. Thus, every moment of the pipeline system life is characterized by thousands of parameters, some of which are systematically stored in the **IHDB**. To collect these data, various types of sensors are used: pressure and temperature point sensors; fiber optic sensors on a single photon counting; point vibration sensors; underground inclinometric spits; inclinometers to control the angles of inclination of structures; infralow frequency sensors for monitoring of natural frequencies of supporting structures. Information about classified incidents, as well as maintenance activities data, comes from several of technological control systems. At the same time, on the basis of special predictive models (see Sect. 5) the equipment current status estimation problems are being solved. Basic values of the "equipment status" parameter are: "norm", "pre-fail" and "fail".

In this section, we will focus on the automatic classification procedures of the incident type, which is based on the analysis of its characteristics by AI methods, since these procedures play a significant role in the ML PdM process. As follows from the information in Sect. 5.1, the incident is not always associated with malfunctions. In pipeline systems, the cause of an incident can be maintenance procedures, urgent repair work on the pipeline or any other reasons that, at a purely technical level, induce the appearance of local signs of an incident in the sensor network. The pipeline control system receives information about several thousand incidents per day. The operators of these systems, in semi-automatic mode, are obliged to find out: is each specific incident related to a real pre-failure condition of the equipment? If the incident is not related to a pre-failure condition, it is transferred to the "closed" status. Crucially, a significant number of resources are spent on handling each incident. Due to limited resources, this is the reason for the inevitable omission of targeted incidents, which are really due to the pre-failure condition of the equipment. Thus, it is very important, first of all, to handle those incidents that are as close as possible to pre-incident ones. It is known that the importance of an incident is determined by the likelihood that the incident is due to a real pre-failure condition, as well as the consequences of the failure, for example: the economic effect.

In this section, we will focus only on the "likelihood of a pre-failure condition for a given incident" and build predictive models that predict the likelihood of a pre-failure state based on data from the IHDB and online information on pre-failure signs. So, we need to create such a classifier, which analyzing the group of incident features, in real time, gives an estimate: with what probability does this incident correspond to the real pre-fault condition? This problem is formulated as an ordinary binary classification problem, where:

- **class 1**: "the incident is associated with a pre-failure";
- **class 2**: "the incident is not associated with a pre-failure".

The highly efficient XGBoost algorithm was used to solve this classification problem [13].The predictive model is an ensemble of deep decision trees. In the process of learning, the XGBoost parameters were selected in such a way that 3400 decision trees of 12 levels were built. An example of a typical branch of a decision tree is shown in Fig. 3. When training was used IHDB base, collected in one of the oil companies. The system was trained on 750,000 incidents, of which only 45,800 corresponded to real pre-failure conditions. Размерность набора признаков инцидента равнялась 128. The dimension of the incident feature set was 128. As a result, the quality metrics [20] of the binary classification had the following values: precision: **0.98**; recall **0.98**: F1-score: **0.97**; Brier score: **0.018** (ideal value is 0), AUC: **0.92**. Each incident is assigned a rank equal to the estimated probability of belonging to class 1. The handling of incidents is carried out according to a decrease in their rank. The obtained results were recognized as practically effective, because of the **first 15% of processed incidents, more than 85% were indeed related to pre-failure conditions**.

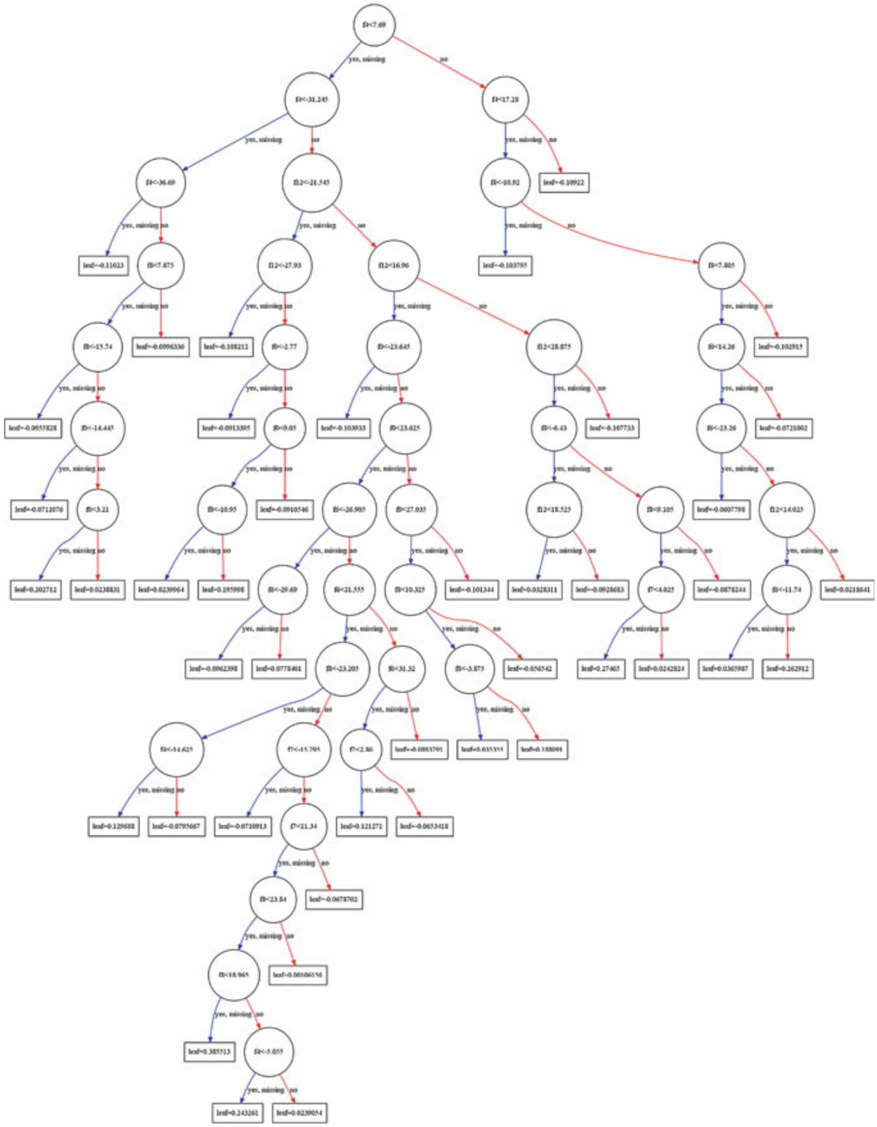

Fig. 3 Typical branch of a decision tree

7 Conclusions and Further Work

This chapter briefly were discussed the features of the use of Machine Learning-based Predictive Maintenance systems of infrastructure facilities in the cryolithozone. In particular, using the example of a pipeline system, solutions of a number of PdM problems are described, taking into account the influence of the cryolithozone factor. Due to the limited scope of the paragraph, some extremely important problems related to the operational control of the state of various types of structures (buildings, roads, bridges) with usage of intelligent monitoring methods set were beyond consideration. General approaches to solving these problems, in many respects, correspond to the approaches outlined in the chapter, but each of them has its own characteristics. In particular, when monitoring stress-strain structures in the cryolithozone, particular attention should be paid to the online assessment of the status of the stress-strain state of structural elements at ultra-low temperatures (below −45 °C). This question, like many others, is the subject of future research.

References

1. Hashemian, H.M., Bean, W.C.: State-of-the-art predictive maintenance techniques. IEEE Trans. Instrum. Meas. **60**(10), 3480–3492 (2011). https://doi.org/10.1109/TIM.2009.2036347
2. Carnero, M.C.: Selection of diagnostic techniques and instrumentation in a predictive maintenance program. A case study. Decis. Support Syst. **38**(4), 539–555 (2005)
3. Swanson, D.C.: A general prognostic tracking algorithm for predictive maintenance. In: 2001 IEEE Aerospace Conference Proceedings (Cat. No.01TH8542), Big Sky, MT, USA, vol. 6, pp. 2971–2977. https://doi.org/10.1109/aero.2001.931317 (2001)
4. Zhou, X., Xi, L., Lee, J.: Reliability-centered predictive maintenance scheduling for a continuously monitored system subject to degradation. Reliab. Eng. Syst. Saf. **92**(4), 530–534 (2007)
5. Kaiser, K.A., Gebraeel, N.Z.: Predictive maintenance management using sensor-based degradation models. IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum. **39**(4), 840–849 (2009)
6. Grall, L. Dieulle, C.B., Roussignol, M.: Continuous-time predictive-maintenance scheduling for a deteriorating system. IEEE Trans. Reliab. **51**(2), 141–150 (2002). https://doi.org/10.1109/tr.2002.1011518
7. PricewaterhouseCoopers: https://cdn-sv1.deepsense.ai/wp-content/uploads/2018/11/pwc-predictive-maintenance-4-0.pdf
8. Cline, B., Niculescu, R.S., Huffman, D., Deckel, B.: Predictive maintenance applications for machine learning. In: 2017 Annual Reliability and Maintainability Symposium (RAMS), Orlando, FL, pp. 1–7 (2017). https://doi.org/10.1109/ram.2017.7889679
9. Butte, S., Prashanth, A.R., Patil, S.: Machine learning based predictive maintenance strategy: a super learning approach with deep neural networks. In: 2018 IEEE Workshop on Microelectronics and Electron Devices (WMED), Boise, ID, pp. 1–5 (2018). https://doi.org/10.1109/wmed.2018.8360836
10. Timofeev, A.V., Denisov, V.M.: Multimodal heterogeneous monitoring of super-extended objects: modern view. recent advances in systems safety and security, 06/2016: chapter. In: Volume 62 of the series Studies in Systems, Decision and Control: pp. 97–116. Springer International Publishing, Berlin. ISBN: 978-3-319-32523-1. https://doi.org/10.1007/978-3-319-32525-5_6

11. Anger, C.: Hidden semi-Markov models for predictive maintenance of rotating elements. Technische Universität, Darmstadt (Ph.D. Thesis) (2018)
12. Bredensteiner, E., Bennett, K.: Multicategory classification by support vector machines. Comput. Optim. Appl. **12**, 53–79 (1999)
13. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (eds.) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17. ACM. pp. 785–794 (2016). arXiv:1603.02754. https://doi.org/10.1145/2939672.2939785
14. American Society for Testing and Materials E 1211-97
15. European Committee for Standardization E1211-97: Standard practice for leak detection and location using surface-mounted acoustic emission sensors
16. Savina, A.V.: Analysis of the risk of accidents when justifying safe distances from the main pipelines of liquefied petroleum gas to objects with the presence of people. Ph.D. Thesis: 05.26.03. Scientific-Technical Center of Research Industrial Problems Security, Moscow, vol. 121, p. il (2013). RSL OD, 61 14-5/120
17. Safety Guide: Methodical recommendations for the quality risk analysis of accidents in hazardous production facilities of main oil pipelines and main oil products. Approved by Order of the Federal Service for Environmental, Technological and Nuclear Supervision of June 17, 2016 n. 228: http://docs.cntd.ru/document/456007201 (2016)
18. Annual Report on the Activity of the Federal Service on Environmental, Technological and Atomic Supervision in 2014: Federal Service for Ecological, Technological and Nuclear Supervision of the Russian Federation, Moscow. http://www.gosnadzor.ru/public/annual_reports/ (2015)
19. Timashev, S.A., Bushinskaya, A.V.: Probabilistic methods for predictive maintenance of pipeline systems. In: Proceedings of the Samara Scientific Center of the Russian Academy of Sciences, vol. 12, no. 1–2, pp. 548-555 (2010)
20. Powers, D.M.W.: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation). J. Mach. Learn. Technol. **2**(1), 37–63 (2011)