ISSN 8756-6990, Optoelectronics, Instrumentation and Data Processing, 2020, Vol. 56, No. 1, pp. 1–11. © Allerton Press, Inc., 2020.
Original Russian Text © A.V. Timofeev, D.I. Groznov, 2020, published in Avtometriya, 2020, Vol. 56, No. 1, pp. 59–73.

AUTOMATION SYSTEMS IN SCIENTIFIC RESEARCH AND INDUSTRY

Classification of Seismic-Acoustic Emission Sources in Fiber-Optic Systems for Monitoring Extended Objects

A. V. Timofeev* and D. I. Groznov

LLC "Flagman-Geo",
Naberezhnaya Reki Karpovki 5G, 14, Saint-Petersburg, 197376 Russia
*E-mail: *timofeev.andrey@gmail.com*

Received May 22, 2019; revision received July 22, 2019; accepted for publication November 12, 2019

Abstract—A new method is proposed for automatic classification of seismic-acoustic emission sources in fiber-optic monitoring systems based on the principles of optical reflectometry in a time domain using the time reconstruction of the interference signal phase. The novelty of this approach lies in the original principles of forming a space of classification features and the use of ensemble classifiers. This method works provided that a relatively small training database is used. High practical effectiveness of the proposed approach is shown when working on real data. A classification reliability value previously inaccessible to alternative classification methods implemented in serial fiber-optic monitoring systems is evenly achieved for different classes of goals.

Keywords: classification, XGBoost, SVM, DAS, ϕ-OTDR.

DOI: ?

INTRODUCTION

Recently, the practice of monitoring the status of extended objects with a controlled perimeter length of 20 km or more has been actively supplemented with fiber-optic sensor systems based on the use of optical quantum effects, which make it possible to measure the dynamics of a seismic-acoustic field in a region where the system sensor is placed. Modern international scientific engineering literature refers to this class of systems as distributed acoustic sensors (DASs) [1, 2] and has accumulated a significant experience of practical use of DASs in different applications, including those for controlling a trunk pipeline protected zone and monitoring a railway ballast section, trunk channels, etc. From the standpoint of details of the physical principle of their operation, monitoring DAS systems have become widely used and can be divided into several classes that differ both by the type of back scattered radiation (Rayleigh, Raman, and Brillouin) and by the type of an information component of the back scattered radiation (amplitude, phase, polarization, and frequency). A detailed description of these classes lies beyond this study, but it is noteworthy that these systems widely used in practice belong to one of the four types: C-OTDR (Coherent Optical Time Domain Reflectometer) using amplitude [9, 10], P-OTDR using polarization [11], ϕ-OTDR using phase [12], and BOTDA using the frequency [13]. Currently, fiber-optic monitoring systems are intensely developed in a direction of phase-sensitive measurements (ϕ-OTDR systems) [14], which allows compensating for a decrease in the system sensitivity for the case of large sensor lengths (more than 20 km) and improving a series of technical parameters as compared with traditional C-OTDR systems.

The main advantages of fiber-optic monitoring systems as compared with analogs constructed, for example, on geophone networks should include: their low cost in solving monitoring problems on very extended objects (more than 20 km), resistance of the sensitive element to electromagnetic impact and interference, stealthiness of system operation (no visually detectable sensors, the system cannot be seen by metal detectors, and no radiation), and stable operation regardless of weather conditions (fog, rain, snow, and wind). An additional advantage is the possibility of combining a sensor with a communication line, temperature

monitoring subsystems, or deformation monitoring. Fiber-optic monitoring systems due to the principle of their operation have a comparatively narrow bandwidth (usually in a range from 0 to 300 Hz), which, however, is sufficient for solving the main seismic-acoustic monitoring problems. In very general terms, the operational principle of these sensors is quite simple [15–17]. A light pulse with a duration of 50 to 150 nanoseconds is sent to an optic fiber that serves as a vibration-sensitive sensor and laid along the monitoring line (along the perimeter of a protected object or along an extended object), and a back scattered signal is obtained. As noted above, the type of a DAS system can be amplitude, phase, frequency, or polarization. Depending on the system type, the corresponding characteristic parameters that describe the reflectivity of a fiber at its different sections whose length ranges from one to ten meters are determined for the back scattered signal. In accordance with the system type, the role of these parameters can be played by an amplitude, frequency, phase, or polarization indicator of a back scattered interference signal. Due to the impact of an external mechanical action (for example, seismic-acoustic one) applied at a particular segment of a sensor, there is local deviation of the scattering coefficient. In different types of DAS systems, the generalized parameter that objectively characterizes this local deviation is considered to be some analytical function of the difference of parameter values corresponding to one and the same sensor segment in adjacent probing cycles. The type of this analytical function depends on the type of a DAS system and is the practical knowledge of DAS system manufacturers. Let this generalized parameter be referred to as an information parameter (IP) or a signal. Thus, an IP has different physical nature for different types of DAS systems. The reflectivity of individual optic fiber sections is related to the value of the local scattering coefficient of light at these sections (segments). The local scattering coefficient, in its turn, depends on the frequency and intensity of vibrations of the external mechanical action on the sensor surface. This action can be, for example, the seismic-acoustic field of soil in which an optic fiber sensor is submerged. The fiber-optic sensor is subjected to pulsed probing at a certain frequency (in practice, $\sim$2000 Hz) as the values of the corresponding IP are measured, thereby recovering the amplitude-frequency response (AFR) of the seismic-acoustic field affecting the sensor. The physics of the operational mechanism of DAS systems with a monomode sensor is quite complicated. The process of forming a back-scattered signal is affected by many factors that arise due to an external mechanical action (for example, the seismic-acoustic field) on the sensor. In addition, this process is affected by the compression and tension of the optical fiber, leading to microfluctuations in the length and cross section of the sensor, as well as variations in the local refractive index of light in the fiber. The effects of changing the local placement Rayleigh scattering centers (microscopic inclusions) affect the deviation of the local scattering coefficient [16]. From a practical point of view, DAS systems of seismic-acoustic monitoring differ from each other by type, price, consumer availability, dynamic range, spatial resolution, and sensor length which can be served by one interrogator (analytical unit) of the system. The following engineering solutions are presented on the market [5, 6]: “Volk”, “Volkodav”, “Volna Al’fa”,“Voron”, “Dunai”,“Omega”,“Optoleks”,“Sokol”,“FiberPatrol”,“Optosens”, “Silixa iDAS”, and “OXY” [5].

One of the most important problems drastically affecting the practical efficiency of a seismic-acoustic monitoring system is undoubtedly the problem of highly reliable classification of the type of a seismic-acoustic emission (SAE) source, which is preliminary detected in the coverage area of the system. The factors complicating the solution of the classification problem of dynamic SAEs (pedestrian, car, tracked vehicles, etc.) include a relatively narrow frequency range of measurements (0–300 Hz), which is inherent in fiber-optic monitoring systems, the fundamental impossibility of prolonged (more than 5 to 30 seconds) data accumulation for making a classification decision, and incomplete reliability of the reconstruction of the AFR of the target SAE. All these factors significantly complicate the use of classical methods for recognizing (classifying) target objects in fiber-optic monitoring systems based on the use of spectral characteristics of their seismic-acoustic images reconstructed from the IP dynamics.

In [18–29], the problem of SAE classification is solved on the basis of using standard methods for reconstructing the AFRs of SAEs (some of them are discussed below).

For the ϕ-OTDR seismic-acoustic monitoring system proposed in this work, this approach does not provide a high classification quality uniformly across all SAE classes, which is due to nonlinear distortions of the AFRs of SAEs, introduced by the receiving path of the monitoring system. This study describes a practically effective approach that, based on the data of the ϕ-OTDR monitoring system, ensures high reliability indicators for solving the classification problem uniformly across all SAE classes under the conditions of a volume-limited training base. The proposed approach is based on an original set of classification features and on the use of modern machine learning methods, which together ensure its high practical efficiency.

1. STATEMENT OF THE PROBLEM

Let $O = \{o_1, o_2, ...,\}$ denote a set of source objects of seismic-acoustic emission, each belonging to one class, and $\Theta = \{\theta_1, \theta_2, ..., \theta_m\}$ be the finite set of indices of the classes to which the objects of the set O belong. There is a certain indicator function χ that is preliminarily unknown to the observer and that maps the set O onto the set Θ, matching any element $o \in O$ with exactly one element $\theta \in \Theta$. Otherwise, $\forall o \in O(\exists!\theta \in \Theta) \wedge (\chi(o) = \theta)$. The object $o \in O$ of a preliminarily unknown class generates the seismic-acoustic emission detected by an extended fiber-optic sensor of the ϕ-OTDR monitoring system based on the principle of coherent reflectometry with the help of the phase reconstruction of an interference signal in time [1].

The monitoring system, based on the observations accumulated during a preliminarily specified time interval and related to a classification object in fully automatic mode, should determine the class to which the observed object o belongs to the set Θ. In other words, the monitoring system should approximate the preliminarily unknown indicator function χ. It is assumed that the SAEs $o \in O$, which are subjected to automatic classification and together form the set O, are located at preliminarily unknown distances from the sensor of the monitoring system. In addition, there is the preliminarily unknown value of the absorption coefficient of the medium (soil) into which the fiber-optic sensor of the monitoring system is embedded. The propagation medium of seismic-acoustic emission is assumed to be anisotropic.

2. CONCEPTS, DEFINITIONS, AND METHOD

A monitoring channel of the ϕ-OTDR system is a linear, length-limited section of the sensor such that the monitoring system is able to receive and analyze data from this section separately from other similar sections of the sensor due to its operational principle and set tunings. The data received from a particular channel somehow characterize the stochastic dynamics of the seismic-acoustic field in its location region. The width of a ϕ-OTDR channel (spatial resolution) depends on the duration of a probe pulse of a coherent laser and on a number of other physical parameters of the system. As a rule, the channel width is smaller than or equal to the physical spatial resolution of the system. The measurement information received by the system from the ϕ-OTDR channel is digitized and used as a characteristic of the seismic-acoustic field parameters in the spatial location of the channel. The channel indices of the system together form set N. A feature of the proposed approach to the SAE classification is that it does not imply a direct estimation of the AFR parameters of the SAEs via individual channels of the monitoring system. Due to the physical principles comprising the basis of ϕ-OTDR monitoring systems, the AFR of a seismic wave coming from the SAE $o \in O$ to a group of adjacent channels of the monitoring system cannot be measured with acceptable accuracy using the data of one channel of the system (although a number of manufacturers claim the opposite). Currently, the relatively low accuracy of measurements is explained by a whole range of problems, including the very principle of coherent reflectometry in a time domain using the phase reconstruction of the interference signal in time. As the probe pulse of a coherent laser passes through the sensor body and back-scatters on the groups of microscopic inclusions (interference centers), which are always present in the optical fiber of a sensor due to the specifics of its manufacture, there is multiple light interference at the laser frequency. This interference results in is so-called interference speckles, some of whose energy is recorded by the system receiver. The speckle sequence measured at successive time points is the basis for the formation of an IP, regardless of how it is determined. It is known that a speckle dependent on the mutual arrangement of interference centers inside the body of a particular channel changes its tuning with respect to its previous state in a substantially nonlinear (almost random) way as the sensor surface has an external impact. Therefore, even with small mechanical impacts on the fiber, the speckle difference corresponding to adjacent probing cycles can be disproportionately large (or, conversely, disproportionately small) with respect to the energy of the impact applied. This disproportionality, being a substantially nonlinear distortion of the initial impact, inevitably has a nonlinear effect on the value of a certain IP, regardless of how it is calculated and whatever mathematical methods are used to smooth out the inevitable bursts of its values corresponding to adjacent measurement time intervals. In this regard, the AFRs of seismic oscillation, which has a vibrating effect on the sensor, are reconstructed quite approximately even for high probing frequencies of the optical fiber by a pulsed laser of the system. In addition, due to the relatively long widths of the channel (5–10 m), an acoustic wave arrives at such a channel with a different phase, which also significantly distortions the estimation of the frequency of seismic-acoustic waves. Thus, using the ϕ-OTDR monitoring system to measure the AFR of the seismic acoustic field results in a nonlinearly distorted estimate of the AFR of the SAE. Our practical

experience with *phi*-OTDR data shows that, even using measurements of this quality, it is possible to classify the types of SAE with high reliability.

Within the framework of the proposed concept, each channel of the ϕ-OTDR monitoring system at each time point t carries the following information: 1 means that a signal is detected in the channel, and 0 denotes that no signal is detected. For convenience, this fact is denoted by a simple indicator function:

$$Ind_j(t) = \begin{cases} 1, & \text{at the time } t \text{ a signal is detected ;} \\ 0, & \text{at the time } t \text{ no signal is detected.} \end{cases}$$

Here j is the channel number. As practice shows, the information that the indicator function $Ind_j(t)$ carries is quite sufficient to reliably classify the types of SAEs. The algorithms for preliminary channel-by-channel detection of SAEs and the principles for choosing the threshold separating a signal-noise mixture from pure noise are detailed in [3]. It is noteworthy that two approaches are practically used: the classical one with adaptive setting of a threshold within the framework of a sequential analysis and the machine learning (ML) approach, in which a binary classification problem is solved during the target detection process: the "signal + noise" class against the "noise" class. This problem is solved using a binary SVM classifier. The classical approach with an adaptive threshold is much less demanding on computing resources, and the ML-based detection method provides low errors of the first kind (< 0.01).

Let there be object o in the sensitive region of the system. Let a sequence of time points at which the monitoring system reads signals from the object o from the fiber-optic sensor be denoted as $T = \{0, 1, 2, ...\}$. At the time $t \in T$, the seismic-acoustic emission from the object o is simultaneously reflected in one or several channels of the monitoring system, whose indices together form the set $N_t = \{n | Ind_n(t) = 1\} \subseteq N, |N_t| \geqslant 1$. Moreover, due to the impacts of the channel noise and other SAEs, some of the channels with indices from the set $N_t \subseteq N$ may not be related to emission from the object o. In other words, at the time $t \in T$, this channel group provides false information and, in fact, serves as an interference generator. This situation is common for the type of monitoring systems under consideration, so the classification system should be resistant to such inaccuracies in the observed data.

Let the coordinates of the jth channel be represented by a sequence $\mathbf{c}_j = \langle c_{jl}, c_{jr} \rangle$, where c_{jl} and c_{jr} denotes the coordinates of its extreme points; $c_{jl}, c_{jr} \in R^m$; R^m is the coordinate space; $c_{jc} = \left(c_{jl} + c_{jr}\right)/2$ denotes the coordinates of the arithmetic center of the jth channel. The coordinates of the arithmetic center of the group of adjacent channels with indices from the set $N' \subseteq N$: $c_c(N) = \sum_{j \in N_t} c_{jc} \left|N'\right|^{-1}$, δ is the width of the ϕ-OTDR channel, and $|A|$ is the cardinal number of the set A.

Seismic-acoustic emission from the object o arrives at the sensor surface nonuniformly both in time and space. This is the disproportionality that carries classification features of the type of the object o. Let there be the term of a mark that is central for the proposed concept. A mark with a subscript i is the disturbance of the seismic-acoustic field, detected by the monitoring system, which lasts for a time interval $[t, t + \Delta_i]$, is reflected in $|N_i|$ adjacent channels of the receiving system, and has a width $\theta_i = \delta |N_i|$. The duration of the mark Δ_i is an extremely important parameter for solving the classification problem. Therefore, the time resolution of the receiving system should be such as to ensure guaranteed reception of signals from the SAEs having a pulse-shock nature (a crowbar impact, a shovel impact, or a human step), whose duration does not exceed 0.1 seconds. Formally, the mark of the object o is referred to as a parameter set $\lambda_i(o) = (t_i, \theta_i, \Delta_i, c_c(N_i))$, where i is the subscript of the mark, o is the SAE (unobserved parameter), Δ_i is the duration of the mark, θ_i is the width of the mark, t_i is the detection time of the ith mark, $c_c(N_i)$ is the arithmetic center of the channel group N_i. Next, we consider the set $T_o = \{t_1^{(o)}, ..., t_n^{(o)}\}$ that is a sequence of time points during which the system observes a sequence of marks $\{\lambda_i(o)\}$, corresponding to the object o, and $I_o = \{1, ..., n\}$ is the set of subscripts T_o. Within the framework of the proposed approach for solving the classification problem, a threshold of marks corresponding to the object o is used, which is presented in the form $\mathrm{M}(o|T_o) = \{\lambda_i(o) | i \in I_o\}$ from the object $o \in \mathrm{O}$. The distribution of time points on the set T_o is also a helpful feature for the classification because it implicitly characterizes the frequency properties of the object o.

In order for the marks $\mathrm{M}(o|T_o)$ to be used in classification algorithms, it is necessary to perform a centering operation on them: to transit from absolute indicators for the quantities $c_c(N_i)$ and the elements of the set T_o to relative ones. We introduce the centering operation with respect to the mark $lambda_i(o)$ in the following form:

$$\mathbf{S}(\lambda_i(o)|d, z) = (t_i - d, \theta_i, \Delta_i, c_c(N_i) - z), d \in R^1, z \in R^m.$$

The results of applying the centering operation to the sequence of marks $\mathrm{M}(o|T_o)$:

$$\mathrm{M}^c(o|T_o, d, z) = \mathbf{S}(\mathrm{M}(o|T_o)|d, z) = \{\mathbf{S}(\lambda_i(o)|d, z)|i \in Ind(T_o)\}, d \in R^1, z \in R^m.$$

Here $\mathrm{M}^c(o|T_o) = \mathrm{M}^c(o|T_o, t_1^{(o)}, \mathbf{c}_c(T_o))$ denotes the centered sequence of observations. Here $t_1^{(o)}$ is the first time point at which $o \in O$ is detected, and $\mathbf{c}_c(T_o)$ is the arithmetic center of the quantities from the set $\{c_c(N_i)|i \in Ind(T_o)\}:\quad \mathbf{c}_c(T_o) = \sum_{i \in Ind(T_o)} c_c(N_i)\,|T_o|^{-1}$. Thus, $\mathrm{M}^c(o|T_o)$ is the centralized version of the flow $\mathrm{M}(o|T_o)$. The flow of the marks $\mathrm{M}^c(o|T_o)$ is used as initial data for solving the classification problem.

When solving the SAE classification problem [18–29], the formation of classification features is based on various data processing methods: the fast Fourier transform (FFT) and similar methods of spectral analysis, the wavelet transform, the principal component method, the Gaussian mixture model (GMM), etc. For convenience, let these methods be referred to as synthetic processing methods. For example, the FFT is used for primary data processing in [18], the wavelet transform in [19], and a combination of the wavelet transform and the principal component method in [20]. Often, a processing scheme is formed with attempts to mechanically implement the schemes that provide acceptable results in solving problems from other application domains. For example, the mel frequency cepstral coefficient (MFCC) transform, which is originally developed for speaker identification, traditionally applied as part of an MFCC–GMM–SVM combination, and reduced to projecting a signal spectrum onto a so-called mel-scale, is used in [21] at the stage of feature extraction. This technique allows one to determine the frequencies most significant for human perception. This conversion has proven itself in speaker identification tasks, but its application in working with seismic-acoustic signals looks strange: a regular linear cepstrum is sufficient in such an application.

In [22], the problem of classifying seismic-acoustic events is solved using a data processing scheme, obviously traced from the well-known FFT–CNN combination (here CNN is a convolutional neural network), which shows good results in problems of classifying optical images. In order to ensure the operability of this scheme as applied to the analysis of DAS data, the observations of the ϕ-OTDR system in [22] are artificially transformed into a certain "image". In general, this scheme can be quite functional if one ignores the fact that the ϕ-OTDR system nonlinearly distorts the FDR of original seismic-acoustic signals. In view of the specifics of operation of the ϕ-OTDR system, one and the same SAE in different parts of the fiber-optic sensor inevitably forms images with incompletely identical AFR, the stochastic dynamics of whose IPs (in the case of the ϕ-OTDR system based on reconstructing the interference signal phase) is reproduced unsteadily, with significant nonlinear distortions. What happens if the synthetic data processing procedures are applied to the sequence of measured IPs? As a result of the processing, the vectors of the output parameters of these procedures to one extent or another inevitably reflect the distortions introduced into the measurement sequence of the IPs by the receiver ϕ-OTDR system. Moreover, for one and the same SAE source, different channels of the system give a significantly different set of values of the vector components of output parameters. Further, this vector can either be directly used as a classification feature, or classification features of a higher level of aggregation can be constructed on its basis. In any case, the features obtained this way are significantly distorted. Therefore, a classifier, such as CNN, should be trained on distorted and unstable images of target SAEs. This factor significantly reduces the effectiveness of any classification procedure trained using the results of applying synthetic processing procedures for IP measurements calculated separately for each channel. In this regard, the time-frequency analysis implemented by standard methods in separate channels of the system, such as the FFT, the wavelet transform, or other methods based on the use of single-channel estimates of IPs, is not very informative for the problems of classifying the type of a SAE source in ϕ-OTDR systems . This is partly the reason for the uneven quality of operation of classifiers trained using transformed channel-by-channel estimates of IPs for various SAE classes in DAS monitoring systems. Thus, in [22], a comparatively low classification accuracy (below 90%) is indicated for two of the six classes. In [23], for two of seven classes, the classification accuracy is also determined to be below 90%. As classifiers, SVM [24], the nearest neighbor method [25], GMM [18], and neural networks of various topologies [22, 23] are traditionally widely used. As part of the development of the deep learning concept, neural networks become an increasingly popular means of solving classification problems in various application domains. Particularly impressive results are achieved using neural networks to recognize optical images. In particular, in recent years, the winners of the large scale visual recognition challenge are all exclusively neural network algorithms. It should be noted that the complete training of algorithms of this class requires sufficiently powerful training corpora. In [23], a convolutional neural network is trained using corpora with a capacity of 487 000 samples. However, in most practical applications of DAS systems, training corpora of such a significant volume are practically inaccessible due to the high cost of their manufacture.

In practice, the capacity of training corpora for each class of the target usually does not exceed 200-300 samples. In this case, the practical use of neural network algorithms is impossible as the considered type of algorithms cannot be effectively trained with low-capacity training corpora. At the same time, a number of modern ensemble algorithms, such as XGBoost [30], remain operable with relatively low-capacity training corpora. It is the algorithms of this type that have shown the greatest efficiency in a number of practical applications of DAS systems [3–6].

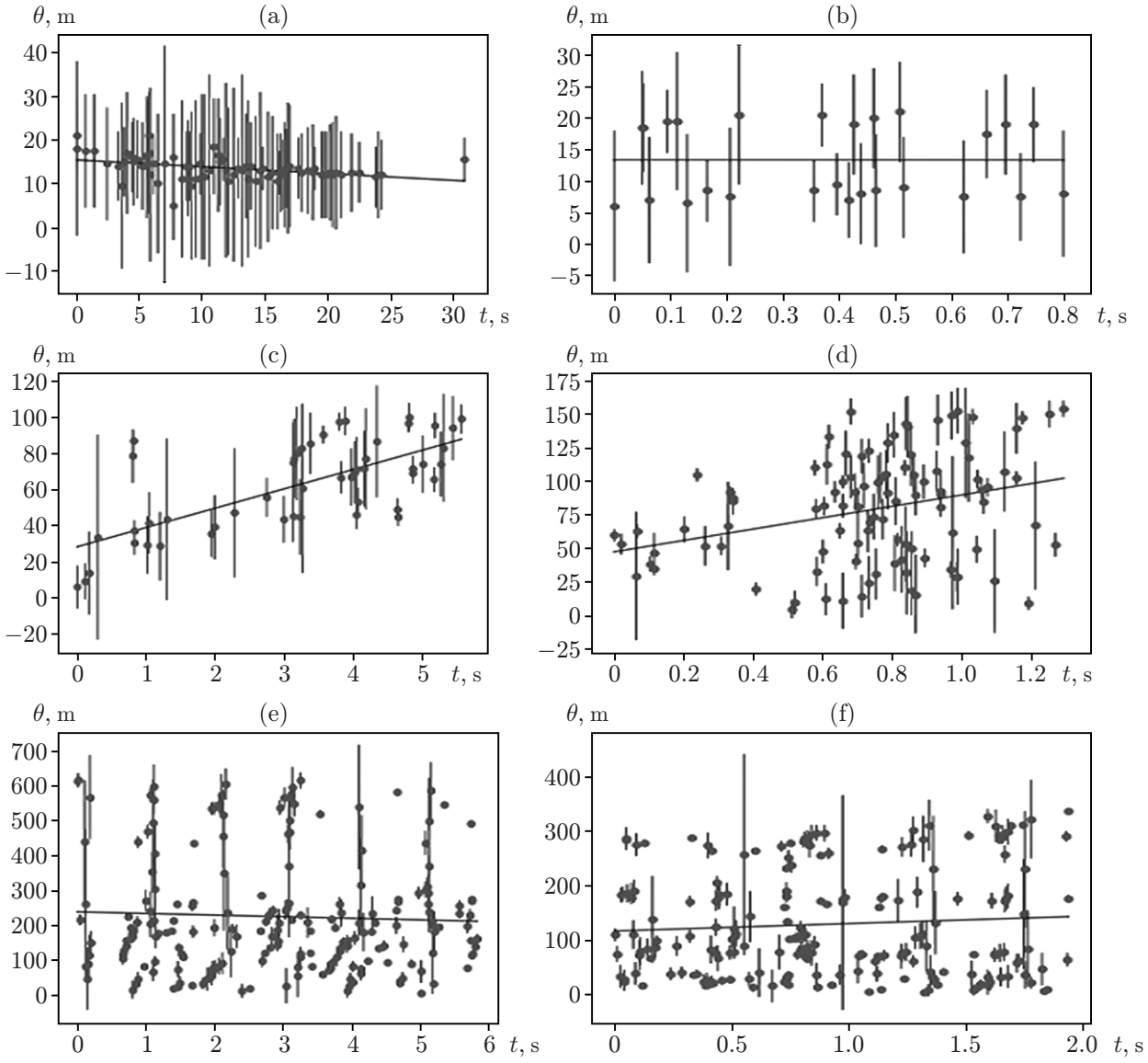
The originality of the approach proposed in this paper lies in the fact that a comprehensive, time-frequency analysis of SAEs is carried out using a special object of a high level of aggregation: the flow of marks $\mathrm{M}^c(o|T_o)$. In relation to the data of single-channel measurements of IPs, the "mark" object is more aggregated as it it is always constructed using several adjacent channels. This approach allows one to partially compensate for incorrectly measuring the AFR of the seismic-acoustic field of the SAE using the ϕ-OTDR system because the width of the mark (measured with an accuracy of up to the width of the ϕ-OTDR channel) may approximately, but objectively characterize the emission source power. As follows from this research, the indicators of spatiotemporal distribution of the marks in the flow $\mathrm{M}^c(o|T_o)$ are very powerful and original classification features. As a result, the "flow of marks" object represents an informative basis for generating original classification features that allow one to solve a classification problem with high efficiency without requiring significant computational resources, and the use of ensemble classifiers makes highly effective training possible in the case of using relatively low-capacity training corpora.

3. CLASSIFICATION OF THE TYPES OF SEISMIC-ACOUSTIC EMISSION SOURCES

The main steps of constructing a classification system are as follows:
— gathering of a marked database of "raw" information, intended for training the classifier using the ML methods;
— generation of a set of classification features;
— reduction (optimization) of this set by special methods;
— selection of the classification algorithm and its training using the ML methods, as a result of which the optimal values of the classifier parameters are determined, thereby providing acceptable values of the classification quality indicators (metrics).

3.1 Gathering of a Marked Database

Raw data samples recorded by the interrogator from the sensor of the ϕ-OTDR system and corresponding to certain classes of SAE sources are obtained, collected, and archived. In addition, each record is marked with a source type marker. Gathering the base is an extremely important step, which is impossible without the use of an expensive polygon. The database used to train the classification system is collected from two polygons. The sensor of the first polygon is placed at a distance of 3 to 5 meters from the edge of the railway ballast section at a depth of 30 to 50 cm, and the soil comprises crushed stone, sand, and clay. The sensor length is 1500 m. The polygon is deployed at "PK-2, peregon st. Sorokovaya—Astana, 497th kilometer" to tune the ϕ-OTDR system intended for railway applications as part of a pilot project with "Kazakhstan Temir Zholy" company. The sensor of the second polygon is deployed near the analog of a boundary dividing strip at a depth of 30 to 50 cm in order to tune the ϕ-OTDR system for perimeter applications. Its length is 5000 m, and the soil comprises dense sand, crushed stone, and loam. In addition to pedestrians, cars, and light wheeled tractors move along this lane. According to the terms of reference, the duration of each record corresponding to a certain type of event (class) is 5 s. As the tuned classifier observes the detected object during the same time, it should determine the type of an object and categorize it as one of five classes. Each type of class corresponds to an implemented flow of marks $\mathrm{M}^c(o|T_o)$ (record), observed for five seconds. The tuning of the ϕ-OTDR system allows one to record seismoacoustic events in a frequency range from 0 to 500 Hz. Thus, a database is gathered, which contains the records of seven types of events: "pedestrian" (101 records, with a distance to the sensor from 0 to 7 m), "manual digging of soil" (50 records, with distance to the sensor from 1 to 10 m), "passenger car" (124 records, with a distance to the sensor from 10 to 50 m),"wheeled tractor" (145 records, with a distance to the sensor from 10 to 50 m), "train" (150 records, with a distance to the sensor from 3 to 5 m), "freight train" (150 records, with a distance to the sensor from 3 to 5 m), "background class" (200 records). The latter includes the seismoacoustic events that are not target events, which is why they cannot fall into any of the first six classes. In general, the



Examples of typical flows of marks for the six classes of SAE sources: pedestrian (a), manual digging (b), passenger car (c), wheeled tractor (d), electric train (e), and freight train (f).

events of this class can be considered as background or interference events. "Background class" is comprised of records of operating excavator and caterpillar equipment, production operation, sounds of rock drills, powerful channel noise that occasionally spontaneously appear in the channels of the monitoring system, and all kinds of combinations of these events. Visual representations of typical flows of marks for the six classes are shown in the figure.

3.2 Classification Features

In choosing a set of features used for automatic classification of objects, two successive tasks are solved. First, based on the parameters of the flow $\mathrm{M}^c(o|T_o)$, a set of classification features is generated. The success of generating features is largely due to the how deeply the developers understand the physical processes underlying the generation of raw data. Let us describe the features used in more detail as selecting them is one of the central issues in solving the problem of automatic classification of SAE sources.

Let $n = |T_o|$ denote the number of marks in the flow $\mathrm{M}^c(o|T_o)$ and $\Delta T_o = t_n^{(o)} - t_1^{(o)}$ stand for the duration of the time interval of observation of the flow $\mathrm{M}^c(o|T_o)$.

For an arbitrary nonempty random set $X \subseteq R^1$, we determine a vector function $\mathbf{V}(X) = (Me(X), Mn(X), Mh(X), \sigma(X), Range(X)) \in R^5$, whose components are as follows: $Me(X)$ is the median of the set X, $Mn(X)$ is the average value of the set X, $Mh(X)$ is the average harmonic of the set X, and $\sigma(X)$ is the estimate of the standard deviation on the set $X, Rang(X) = Max(X) - Min(X)$. The following sets are considered next: $T_o^c = \{t_i - t_1^{(o)} | i \in Ind(T_o)\}$, $\boldsymbol{\theta}_o^c = \{\theta(t_i) - \mathbf{c}_c(T_o) | i \in Ind(T_o)\}$, $\boldsymbol{\Delta}_o^c = \{\Delta_i | i \in Ind(T_o)\}$, and $\mathbf{C}_o^c = \{c_c(N_i) - \mathbf{c}_c(T_o) | i \in Ind(T_o)\}$.

The elements of the set $T_o^c \otimes \boldsymbol{\theta}_o^c$ can be interpreted as time points t_i at which noisy measurements of the arithmetic coordinate center $c_c(N_i)$ of the dynamic object $o \in \mathrm{O}$ are made. In this case, $\nu(T_o^c \otimes \boldsymbol{\theta}_o^c)$ denotes the local estimate of the projection of the velocity of this object onto the sensor, obtained by the linear interpolation method. Next, the following heuristic indicators that characterize the flow $\mathrm{M}^c(o|T_o)$ are considered: $D_\Delta = nMe\left(\boldsymbol{\Delta}_o^c\right)\left(\Delta T_o \sigma\left(\boldsymbol{\Delta}_o^c\right)\right)^{-1}$ is the characteristic of the distribution density of the duration of marks over time, and $D_\theta = n\left(\Delta T_o Me\left(\boldsymbol{\theta}_o^c\right)\right)^{-1}$ and $D_{\theta L} = n\left(\Delta T_o Rang\left(\boldsymbol{\theta}_o^c\right)\right)^{-1}$ are the parameters of the spatiotemporal distribution of marks. Other heuristic indicators characterizing the flow $\mathrm{M}^c(o|T_o)$ are also considered. As soon as the set of features for the classification is formed, it is reasonably narrowed by selecting the most informative signs. In fact, several so-called "filtering" methods are used at this stage: chi-square, the Pearson correlation, and analysis of variance (ANOVA). Another method used is backward elimination. The resulting set of features that are eventually used to classify the type of a SAE source has the following form: $\{D_\Delta, D_\theta, D_{\theta L}, \nu(T_o^c \otimes \boldsymbol{\theta}_o^c), \mathbf{V}\left(\boldsymbol{\Delta}_o^c\right), \mathbf{V}\left(\mathbf{C}_o^c\right), \mathbf{V}\left(\boldsymbol{\theta}_o^c\right)\}$.

3.3 Classification Algorithms

Numerous experiments have shown the high practical effectiveness of well-known methods: a multiclass support vector machine (SVM) method and two ensemble methods which are XGBoost [30] and Gradient-Boosting (GB) [31, 32]. For various reasons, the rest of the methods provided significantly lower quality indicators for solving the classification problem on the used database. In particular, the classifier based on a multilayer neural network (ANN) in the case under study is inferior to the three above-mentioned methods by more than 5% according the "probability of correct classification" indicator due to an insufficient size of a training sample. In the case of ANN, the size of a training sample is a factor that radically determines the effectiveness of the classification. All these methods, especially the multiclass SVM method, have clear mathematical justification, which is not given here due to its large volume.

The generalizing ability of classifiers is estimated using a standard rolling control scheme also known as cross validation (CV) in a leave-one-out (LOO) version. For the SVM classifier, with account for the multiclass statement of the problem, the "one-vs-rest" strategy is used. The main parameters of the algorithms, otherwise called hyperparameters, are optimized using the Hyperopt module of the Sklearn library, Python 3.6. For example, in the XGBoost classifier, the hyperparameters are ¡eta¿ (analogous to "learning rate"), "max_depth", "min_child_weight", and "gamma". The quality of the classification algorithms is estimated using the universal indicator P_{CC} (probability of correct classification), estimated numerically for each class at the step of testing the classification algorithms. The base of precedents (signals) used for training is relatively well balanced: there is no significant difference in the power of the classes presented therein. For this reason, the exponent P_{CC} as the main classification metric in this case is quite correct. Nevertheless, other traditional classification quality metrics are actively used during the investigation, including the recall estimates and parametric F measures, which, due to the good balance of the training set, expectedly bring no additional information to estimating the quality of the classification algorithm. It is the metric P_{CC} in the context of the goal of this work that is the most indicative, well understood, and interpreted by both developers and operators of monitoring systems, and it is also an objective integral indicator of the quality of the classification system.

The classification results are summarized in the table. As follows from the data presented therein, the selected three classifiers trained on a marked database containing records of signals from six types of sources, confidently can classify these types of events. It is very surprising that the classifiers quite reliably distinguish between an electric train and a freight train by a five-second long sound and that the pedestrian class is 100% reliably classified: there is no confusion with the amplitude-frequency class of "manual digging of soil", which is close to it. The first section of the table shows classification results for seven classes. The second section contains classification results for six classes: the "electric train" and "freight train" classes are combined into a single class of "train". The third section of the table shows classification results for five classes: the "light wheeled tractor" and "passenger car" classes are combined into a single class of "wheeled vehicles".

Table

Classifier	Section 1		Section 2		Section 3	
	Object type (7 classes)	P_{CC}	Object type (6 classes)	P_{CC}	Object type (5 classes)	P_{CC}
XGBoost	Pedestrian	1.0	Pedestrian	1.0	Pedestrian	1.0
	Manual digging of soil	1.0	Manual digging of soil	1.0	Manual digging of soil	1.0
	Passenger car	0.99	Passenger car	0.99	Wheeled vehicle	0.99
	Light wheeled tractor	0.99	Light wheeled tractor	0.99		
	Electric train	0.99	Train	1.0	Train	1.0
	Freight train	0.98				
	Background class	0.99	Background class	1.0	Background class	0.99
	Average P_{CC}	0.99	—	0.99	—	0.99
Gradient-Boosting	Pedestrian	1.0	Pedestrian	1.0	Pedestrian	1.0
	Manual digging of soil	1.0	Manual digging of soil	1.0	Manual digging of soil	0.99
	Passenger car	0.97	Passenger car	0.98	Wheeled vehicle	0.99
	Light wheeled tractor	0.97	Light wheeled tractor	0.98		
	Electric train	0.98	Train	1.0	Train	1.0
	Freight train	0.98				
	Background class	1.0	Background class	1.0	Background class	0.99
	Average P_{CC}	0.98	—	0.99	—	0.99
Multi-class SVM	Pedestrian	0.99	Pedestrian	1.0	Pedestrian	1.0
	Manual digging of soil	0.99	Manual digging of soil	1.0	Manual digging of soil	0.99
	Passenger car	0.98	Passenger car	1.0	Wheeled vehicle	0.99
	Light wheeled tractor	0.98	Light wheeled tractor	1.0		
	Electric train	0.97	Train	1.0	Train	1.0
	Freight train	0.97				
	Background class	1.0	Background class	1.0	Background class	1.0
	Average P_{CC}	0.98	—	1.0	—	0.99

In general, the results indicate that the three selected classification methods reliably determine the type of class from its five-second sample. For example, in the case of six classes, the multiclass SVM method shows a result absolute in accuracy. Using the CV methodology guarantees that the generalizing ability of trained classifiers is sufficiently high, i.e., no overfitting is observed. The interference SAEs, which belong to the "background class", are also fairly reliably classified, which is extremely important for the practical use of the monitoring system for ensuring a minimum level of errors of the first kind. In particular, when testing the ϕ-OTDR monitoring system tuned for railway applications, the detection and classification subsystem equipped with the XGBoost classifier for continuous monitoring for six months obtains an average value of at least 920 hours for the "false alarm hours" parameter. In this application, the target SAEs Are: five types of trains, pedestrians, manual production operation on the railway track, and automated production operation on the railway track. The interference SAEs are the channel interference, the operation of equipment near a ballast prism, road transport (a dirt road passes near the railway track), pets, and other SAEs. The length of the fiber-optic sensor is 1 500 meters, and the correctness of recognition of target events by the system is controlled by the operator using a video camera, as well as the preliminarily known train schedule. From the standpoint of convenience and effectiveness of practical application, the classification methods under consideration on the selected feature space are approximately equivalent, while the multiclass SVM method is trained faster and the ensemble methods work more accurately with a larger number of classes.

Comparative experiments are carried out on a regular PC (quadcore Core i7-2630QM with a standard frequency of 2.0 GHz and 16 GB of RAM). With the use of the corpus of data described in sec. 3.1 and the feature space from Sec. 3.2, the average training times are 0.5 s for the multiclass SVM, 6 s for GB, and 16 s for XGBoost.

CONCLUSION

Practical studies showed that the proposed method for automatic classification of seismic-acoustic emission sources according with the use of the data from the monitoring DAS system was highly reliable and easily implemented at the modern level of computing facilities. The selected original set of classification features was sufficiently informative to ensure the classification process itself. Classifiers of the following types showed the best results: multiclass SVM, XGBoost, and Gradient-Boosting. All of the three uniformly provided a value $P_{\mathrm{CC}} \geqslant 0.98$ for the cases of seven, six, and five classes of targets. The classification reliability level reached evenly across all classes of targets is currently unavailable for alternative classification methods used in ϕ-OTDR systems [22, 23], where this indicator for some classes does not exceed 0.9. Based on the results of this study, a subsystem for the classification of seismic-acoustic events was developed, which was part of the *phi*-OTDR monitoring system "OXY". This monitoring system was included in a number of projects to ensure the control of extended facilities in the territory of the Republic of Kazakhstan.

REFERENCES

1. J. P. Dakin and C. Lamb, *GB Pat. 2222247 A. Distributed Fibre Optic Sensor System* Publ. 28.02.1990.
2. K. N. Choi, J. C. Juarez, and H. F. Taylor, "Distributed Fiber Optic Pressure/Seismic Sensor for Low-Cost Monitoring of Long Perimeters," Proc. SPIE **5090**, 134–141 (2003).
3. A. V. Timofeev and V. M. Denisov, "Multimodal Heterogeneous Monitoring of Super-Extended Objects: Modern View," in *Recent Advances in Systems Safety and Security* Vol. 62 (Springer–Verlag, Switzerland, 2016); DOI: 10.1007/978-3-319-32525-5_6.
4. A. V. Timofeev and D. V. Egorov, "Bimodal System for Surveillance of Super-Extended Objects," in *Proc. of the 6th Intern. Conf. on Electronics, Computers and Artificial Intelligence (ECAI). Bucharest, Romania, 23–25 Oct., 2014*; DOI: 10.1109/ECAI.2014.7090211.
5. A. V. Timofeev, D. V. Egorov, and V. M. Denisov, "The Rail Traffic Management with Usage of C-OTDR Monitoring Systems," in World Acad. Sci. Eng. Technol. Intern. Journ. Comput. Electr. Autom. Control Inf. Eng. **9** (7), 1492–1495 (2015).
6. A. V. Timofeev, "Monitoring the Railways by Means of C-OTDR Technology," Intern. Journ. Mech. Aerosp. Ind. Mechatron. Eng. **9** (5), 701–704 (2015).
7. A. Papp, C. Wiesmeyr, M. Litzenberger, et al., " Train Detection and Tracking in Optical Time Domain Reflectometry (OTDR) Signals," in *Pattern Recognition. GCPR 2016* Ed by B. Rosenhahn and B. Andres (Springer, Hannover, 2016). (Lecture Notes in Computer Science, Vol. 9796.)
8. V. V. Korotaev, V. M. Denisov, J. J. P. C. Rodrigues, et al., "Monitoring Industrial Facilities Using Principles of Integration of Fiber Classifier and Local Sensor Networks," Proc. SPIE **9525**, 95254J (2015); DOI: 10.1117/12.2184727.
9. A. Barrias, J. R. Casas, and S. Villalba, "A Review of Distributed Optical Fiber Sensors for Civil Engineering Application," Sensors **16** (5), 56–78 (2016).
10. K. Takada, A. Himeno, and K. Yukimatsu, "Phase-Noise and Shot-Noise Limited Operations of Low Coherence Optical Time Domain Reflectometry," Appl. Phys. Lett. **59** (20), 2483–2485 (1991).
11. A. J. Rogers, "Polarisation Optical Time Domain Reflectometry," Electron. Lett. **16** (13), 489–490 (1980).
12. J. C. Juarez and H. F. Taylor, "Polarization Discrimination in a Phase-Sensitive Optical Time-Domain Reflectometer Intrusion-Sensor System," Opt. Lett. **30** (24), 3284–3286 (2005).
13. Q. Cui, S. Pamukcu, W. Xiao, and M. Pervizpour, "Truly Distributed Fiber Vibration Sensor Using Pulse Base BOTDA with Wide Dynamic Range," IEEE Photon. Technol. Lett. **23** (24), 1887–1889 (2011).
14. Y. Wang, B. Jin, Y. Wang, et al., "Real-Time Distributed Vibration Monitoring System Using Φ-OTDR," IEEE Sensors Journ. **17** (5), 1333–1341 (2017).
15. X. Liu, B. Jin, Q. Bai, et al., "Distributed Fiber-Optic Sensors for Vibration Detection," Sensors **16** (8), 1164–1195 (2016); DOI: 10.3390/ s16081164.
16. K. V. Marchenko, O. E. Nanii, E. T. Nesterov, et al., "Protection of a FOCL by a Distributed Acoustic Sensor Based on a Coherent Reflectometer," Vestn. Svyazi, No. 9, 17–19 (2011).

17. Yu. Rusanov, "Prospects of Using Fiber-Optic Tools for Perimeter Securty," Sistemy Bezopasnosti, No. 4, 29–32 (2018).
18. R. R. Brooks, P. Ramanathan, and A. M. Sayeed, "Distributed Target Classification and Tracking in Sensor Networks," Proc. of the IEEE **91** (8), 1163–1171 (2003); DOI: 10.1109/JPROC.2003.814923.
19. X. Shen, Sh. C. Wan, H. Huo, and T. Fang, "An Improvement on Discrete Wavelet Transform-Based Algorithm for Vehicle Classification in Wireless Sensor Networks," in *Proc. of the 1st IEEE Conf. on Industrial Electronics and Applications. Singapore, Singapore, 24–26 May, 2006*; DOI: 10.1109/ICIEA.2006.257218.
20. X. Wang, H. Qi, and S. S. Iyengar, "Collaborative Multi-Modality Target Classification in Distributed Sensor Networks," in *Proc. of the Intern. Conf. on Information Fusion 2002. Annapolis, USA, 8–11 July, 2002*.
21. Y. Kim, S. Jeong, and D. Kim, "A GMM-Based Target Classification Scheme for a Node in Wireless Sensor Networks," IEICE Trans. Commun. E91B, 3544–3551 (2008).
22. M. Aktas, T. Akgun, M. U. Demircin, and D. Buyukaydin, "Deep Learning Based Multi-Threat Classification for Phase-OTDR Fiber Optic Distributed Acoustic Sensing Applications," Proc. SPIE. **10208**, 56–70 (2017).
23. A. V. Makarenko, "Deep Learning Algorithms for Signal Recognition in Long Perimeter Monitoring Distributed Fiber Optic Sensors," in *Proc. of the IEEE 26th Intern. Workshop on Machine Learning for Signal Processing (MLSP). Vietri sul Mare, Italy, 13–16 Sept., 2016*; DOI: 10.1109/MLSP.2016.7738863.
24. J. Li, C. Zhang, and Zh. Li, "Battlefield Target Identification Based on Improved Grid-Search SVM Classifier," in *Proc. of the Intern. Conf. on Computational Intelligence and Software Engineering. Wuhan, China, 11–13 Dec., 2009*.
25. M. Kandpal, V. K. Kakar, and G. Verma, "Classification of Ground Vehicles Using Acoustic Signal Processing and Neural Network Classifier," in *Proc. of the Intern. Conf. on Signal Processing and Communication (ICSC). Noida, India, 12–14 Dec., 2013*; DOI: 10.1109/ICSPCom.2013.6719846.
26. S. A. Alyamkin and S. I. Eremenko, "Pedestrian Detection Algorithms Based on an Analysis of the Autocorrelation Function of a Seismic Signal," Avtometriya **47** (2), 26–32 (2011) [Optoelectron., Instrum., Data Process. **47** (2), 124–129 (2011)].
27. D. O. Sokolova and A. A. Spektor, "Classification of Moving Objects Based on Spectral Features of Seismic Signals," Avtometriya **48** (5), 112–119 (2012) [Optoelectron., Instrum., Data Process. **48** (5), 522–528 (2012)].
28. Yu. V. Morozov and A. A. Spektor, "Object Classification Based on Analysis of Spectral Characteristics of Seismic Signal Envelopes," Avtometriya **53** (6), 49–56 (2017) [Optoelectron., Instrum., Data Process. **53** (6), 576–582 (2017)].
29. Yu. V. Morozov, M. A. Raifel'd, and A. A. Spektor, "Seismic Signal Processing for Estimating the Path of a Moving Vehicle," Avtometriya **54** (3), 32–38 (2018) [Optoelectron., Instrum., Data Process. **54** (3), 237–242 (2018)].
30. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. of the 22nd ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining. San Francisco, USA, 13–17 Aug., 2016*; DOI: 10.1145/2939672.2939785.
31. L. Mason, J. Baxter, P. Bartlett, and M. Frean, "Boosting Algorithms as Gradient Descent," in *Proc. of the 12th Intern. Conf. on Neural Information Processing Systems. Denver, USA, 29 Nov. – 4 Dec., 1999*.
32. J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," The Ann. Statist. **29** (5), 1189–1232 (2001).