Intrusion detection system on the basis of data mining algorithms in the industrial network of automated process control system

M.A. Gurin[1], A.M. Vulfin[1], V.I. Vasilyev[1], A.V. Nikonov[1]

[1]Ufa State Aviation Technical University, K. Marks St. 12, Ufa, Russia, 450008

Abstract. The purpose of the work is to increase the security of the industrial network of an automated process control system based on intelligent network traffic analysis algorithms. The analysis of the problem of detecting and recording actions of violators on the implementation of a network attack on an automated process control system in the industrial network of an enterprise has been performed. A structural and functional model of the monitoring system of the industrial network of industrial control systems is proposed. An algorithm is developed for the intellectual analysis of network traffic of industrial protocols and a software package that implements the proposed algorithms as part of a monitoring system to evaluate the effectiveness of the proposed solution on field data.

1. Introduction

Security of the critical infrastructure of automated process control system (APCS) under the conditions of the automation level of modern production in the Russian Federation and around the world is becoming an increasingly priority task. The imperfection of the protection and vulnerability of modern SCADA-systems (Supervisory Control and Data Acquisition systems) is due to a number of features of the organization of such systems. Special viruses and target attacks, sponsored by terrorist groups or governments of competing countries, increasingly began to target at the industrial production facilities [1, 2, 3]. The Internet of things gradually comes to the enterprises networks, expanding the already extensive list of industrial protocols and forming the concept of an industrial Internet of things (IIoT) [2, 3]. The means to ensure the information security of process control systems at this stage of their development are not able to withstand such threats [4, 5].

Network security is becoming one of the main directions in the development of information security through the use of a set of technical means [1]. Since any computer process control system can be attacked, which usually results in serious technical, reputation and economic losses, it is necessary to timely detect both known and previously unknown attacks in industrial networks. Attacks of malicious persons are constantly improving, becoming combined and spread almost instantly. Intrusion detection systems (IDS) implement monitoring functions and detect attacks that have bypassed the firewall. IDS informs the administrator, who, in turn, takes a further decision on the response to the attack.

Thus, it can be concluded that the network attacks detection systems based on the use of artificial intelligence methods as a key element of ensuring cybersecurity of the critical infrastructure of the APCS in the concept of the development of the digital economy are of relevance and need to be improved.

The research goal is to increase the effectiveness of network attack detection system by using a neural network analysis module as part of the IDS. To achieve this goal, it is necessary to solve the following tasks:

- Analysis of the problem of detecting network attacks in industrial networks APCS.
- Development of the structure of the system for monitoring the industrial network of APCS;
- Development of algorithms for intellectual analysis of network traffic of industrial networks;
- Development of a software package that implements the proposed algorithms as part of a monitoring system, and an assessment of the effectiveness of the proposed solution on full-scale data.

2. Analysis of the problem of detecting network attacks in industrial networks

Please The process of automation of industrial production continues to evolve: the number of "intelligent" terminal devices is increasing, the number of microcontroller-based computing systems involved in the process control and process control is growing. Under these conditions, the role of data collected at all levels of the process control system significantly increases. Requirements imposed by consumers of this information are increasingly being tightened in terms of the volume, speed and reliability of data acquisition, as well as information security of the entire system [3]. In turn, increasing degree of automation of the enterprise functioning promoted the mutual integration of information (IT) and so-called operational (OT) technologies [5].

An industrial network is a data transmission environment that must meet a variety of diverse, often contradictory requirements; a set of standard data exchange protocols that allow to link equipment together (often from different manufacturers), and also to ensure interaction between the lower and upper levels of the enterprise management system.

In IIoT, the main types of "things" that need to be connected to the network are various types of sensors and actuators. These devices, on the one hand, have an interface with a communication network, and on the other hand, an interface that provides physical interaction with the process to be monitored (Ethernet, Wi-Fi, cellular networks, Sigfox, LoRa, ZigBee, etc.).

Not so long ago, the hierarchy of the APCS had a clear boundary between the levels. The trends of recent years have made this structure much more complex and diffuse. The automated process control system is more and more integrated with the automated control system, and through it inevitably enters the sphere of Internet technologies. Unification of the corporate and industrial network of an enterprise inevitably poses a serious problem of information security of the industrial network of industrial control systems.

The traditional process control system is a real-time system. To ensure error-free process control, continuous process operation monitoring is necessary. If IT security methods are applied in the process control system, in the event of possible data compromетation, the security system may limit access to this data. This, in turn, can lead to loss of control over the TP and man-made or environmental catastrophe (in critical infrastructure, petrochemical industry and other industries). Therefore, in relation to industrial control systems, the inverse distribution of the significance of safety aspects is widely used [6,7]:

- availability;
- integrity;
- confidentiality.

The following main threats to the security of an industrial network can be identified [6, 7]:

- Traditional virus software (malware);
- Targeted attacks;
- Unintentional staff errors;
- Suppliers of equipment and software, partners, contractors;
- extortion programs;
- Internal and external sabotage;
- Errors of specialized industrial control systems;
- Failure of hardware.

Summary information of the information security systems of automated process control systems shown in Table 1.

Table 1. Information security support systems in APCS.

Product name	**Kaspersky Industrial CyberSecurity** [7,8]	**Security Matters SilentDefense** [10,11]	**Positive Technologies Industrial Security Incident Manager (PT ISIM)** [9]	**Honeywell Risk Manager**
Meeting the requirements of regulators (FSTEC №31)	+	-	+	-
Security audit	+	-	+	+
Creating rules for the operation of technological processes	+	+	+	-
Intgration with Human-Machine Interface (HMI)	+	-	-	-
System distribution	KICS for Nodes, KICS for Networks, Security Center	Sensors + Command Center	Full distribution	A single control center that collects information from external monitoring and security systems
Recommendations for elimination	-	-	+	+
Intervention in technological process	Uses a copy of network traffic (SPAN / TAP), but contains an intrusion prevention system	Uses copy of network traffic (SPAN-ports)	Uses copy of network traffic (unidirectional gateway)	Data collection without intervention, integration with intrusion prevention system is possible
Software developer certification for APCS	Siemens (WinCC, WinCC OA), Emerson	-	-	Honeywell Experion

3. Development of the structure of the system for monitoring the industrial network of APCS

Figure 1 shows network structure of an enterprise with tools for collecting and analyzing network traffic of a network intrusion detection system (IDS).

The structure of the network attack detection system based on data mining is shown in the Figure 2.

At the first stage, network traffic is captured. In Figure 1, the numbers indicate the following components: 1 – router as a means of collecting incoming / outgoing network traffic, 2 – router as a means of collecting traffic within the enterprise network. The collection of necessary data is performed using the package sniffer.

The second stage identifies the most significant parameters that characterize network activity.

At the third stage, detection and classification of attacks is carried out. The results of this recognition are transmitted to related systems for reporting and visualization, depending on the capabilities and specifics of adjacent systems. In addition, information about the attack on the APCS is added to a special archive designed to investigate cybersecurity incidents by authorized specialists and managers.

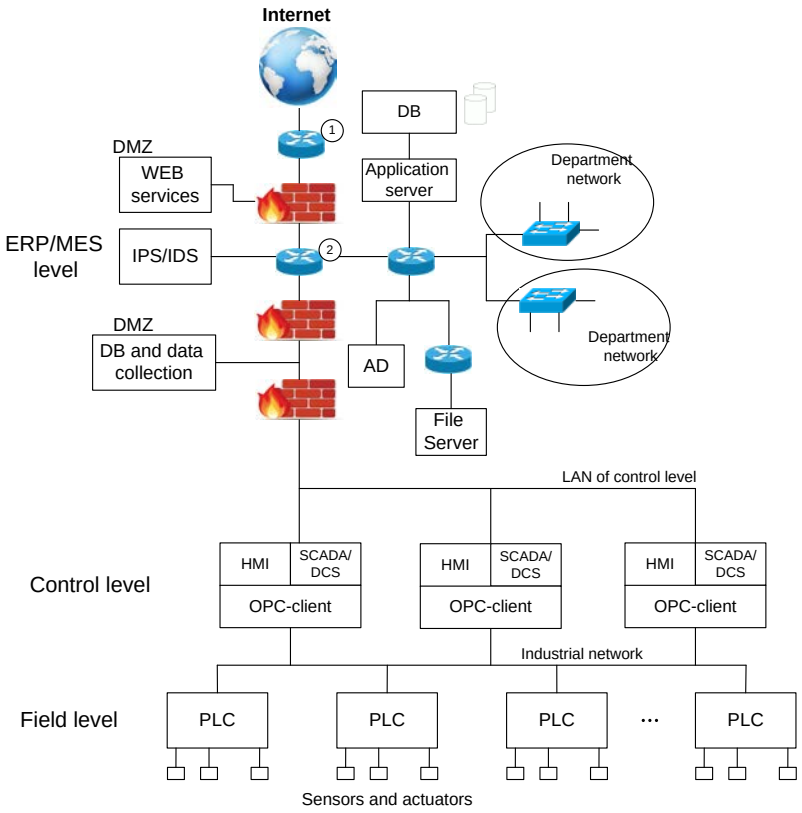

Figure 1. The structure of the enterprise network in which information is collected.

4. Development of algorithms for intellectual analysis of network traffic of industrial networks

An effective network attack detection system based on artificial intelligence methods can be built only with a high-quality dataset of training and test samples that simulates various intrusions.

KDDCUP99 – intrusion detection dataset based on the data set DARPA 98, is one of the only publicly available labeled data set [12]. Dataset NSL-KDD proposed to improve KDD dataset. This dataset has the following advantages over the KDD dataset:

- it does not include redundant entries in the training set, therefore classifiers will not be retrained due to the frequency of such entries;
- there are no duplicate entries in the proposed test suites;
- number of records in the training and test sets is optimal, which makes it possible to conduct experiments on the full set.

Each entry has 41 attributes describing the various functions of the connection, and the label assigned to each of them: attack or normal connection.

Dataset UNSW-NB15 [13] contains data of normal traffic in modern networks and network traffic of synthesized networks.

Each entry in this set contains attributes that describe the various functions of the connection, and the label assigned to each of them: attack or normal connection [13].

The comparative table (Table 2) of the NSL-KDD and UNSW-NB15 methods is shown below.

Dataset UNSW-NB15 is selected for use in the system:

- number of classes of attacks is more than 2 times;
- test stand contained 33 subnets (NSL-KDD – 2 subnets);
- when collecting traffic on the network, 45 IP addresses participated in the exchange of information against 11 in NSL-KDD;
- traffic was collected by several means (in NSL-KDD - Bro-IDS);
- UNSW-NB15 set contains more attributes for the record (49 vs. 42 in NSL-KDD).

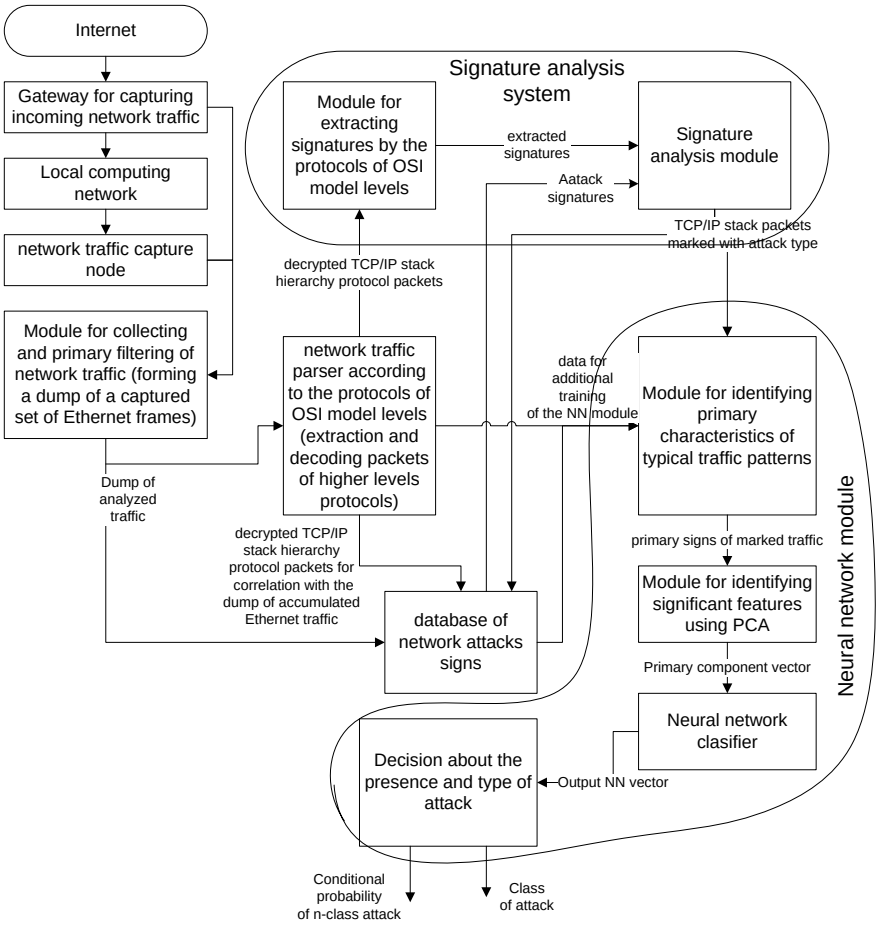

Figure 2. Structural scheme of the network attack detection system.

Table 2. Comparative table of the NSL-KDD and UNSW-NB15 methods.

№	Parameters	NSL-KDD	UNSW-NB15
1	Number of networks	2	33
2	Number of different IP addresses	11	45
3	Traffic simulation	Yes	Yes
4	Duration of data collection	5 weeks	16 days 15 hours
5	Format of data collected	3 types (tcpdump, BSM and dump-files)	PCAP-files
6	Attack classes	4	9
7	Feature Extraction Tools	Bro-IDS	Argus, Bro-IDS and etc.
8	Number of attributes in the record	42	49

At the moment, in relation to industrial networks the following types of network attacks can be distinguished (Table 3).

Of all types of attacks implemented in the industrial network, network attack detection systems are able to most effectively cope with network intelligence, DoS attacks, as well as various types of injections and buffer overflow attacks. IDS is a practically universal tool capable of detecting most types of attacks implemented on an industrial network.

Main steps of the network traffic analysis algorithm in the industrial network are presented in the table 4.

Table 3. Network attack methods comparison.

Attack Type	Description	Implementation Features	method of combating
Buffer overflows	Search for vulnerabilities that can violate the memory boundaries, execute an arbitrary binary code on behalf of an authorized user	1. Preparation of code to be executed in the context of a privileged program. 2. Changing the sequence of program execution with transfer of control to the prepared code.	• Adjustment of the source code of the program. • The use of non-executable buffers. • The use of checks overstep the border. • Conduct integrity checks.
Specialized programs	Viruses, Trojan horse, sniffer, rootkit	The hidden nature of the functioning in the system, data collection, avalanche dissemination	• Anti-virus tools and regular updating of their signatures; • Encryption; • Antisniffera; • Firewalls; • Anti-rootkits [4].
Network intelligence	Collect network information using publicly available data and attack planning applications.	Network intelligence is conducted in the form of DNS queries, ping sweep, and port scanning	• Disable ICMP echo and echo reply on peripheral routers. • The use of intrusion detection systems (IDS).
IP-spoofing	The attacker impersonating an authorized user of the system	Insert false information or malicious commands into the normal data stream	• Access control • The use of cryptographic authentication.
Injections	SQL injection, crossite scripting (XSS attack), XPath injection.	Changing the query parameters to the database, embedding arbitrary code in the web page.	• Rules for building SQL queries; • Encoding data and control characters; • Regular update.
Denial of Service (DoS)	Creating conditions under which legitimate users cannot access the system.	Keeping all connections in busy state. During DoS attacks, normal Internet protocols (TCP and ICMP) can be used.	• Anti-spoofing functions. • Anti-DoS features. • The use of network attack detection systems.
Phishing-attacks	Cheating or social development of enterprise employees to steal their identity and transfer them for criminal use.	Using spam-mailing via e-mail or instant messengers, the use of computer-bots, methods of social engineering.	• The use of proven resources; • Antivirus tools and signature database updates; • Education and training of staff.

5. Development of a software package that implements the proposed algorithms as part of a monitoring system

When pre-processing the parameters of the selected data set UNSW-NB15, the attack classes containing less than 5000 examples are excluded from the training set (Table 5).

Categorical variables are coded into numeric ones. The entire data set is divided into a training and test sample in the ratio of 75% to 25%.

In order to compare the effectiveness of the use the classifier for a specific task, it is necessary to compare the learning results of these classifiers on real data sets. To quantify the classifiers, the following coefficients are applied [15]:

1) False Positive Rate – FPR;
2) True Positive Rate – TPR;
3) Sensitivity;
4) Specificity;
5) Proportion of correctly recognized examples – Correct Rate.

Table 4. Characteristics and tools for analysis.

Analysis stage	Characteristics and tools used
Extract traffic	To solve the problem of capturing traffic, it is proposed to use switches with port mirroring and connecting devices with the sniffer and packet analyzer installed.
Feature selection	When analyzing the main parameters of network traffic, one has to deal with an interconnected system of input parameters (factors). Not all of the factors studied are essentially interconnected, but separate groups of input parameters. A transition is needed to a set of independent parameters containing the necessary information about the variation or dispersion of the initial set of factors of the process under study [14]. It is proposed to use: • Principal Component Analysis, PCA; • Neual network autocoder; • Neual network autocoder on the basis of convolutional neural network.
Classification	In relation to the problem of classification of network traffic and network discovery it is proposed to use: • Artificial neural networks (multilayer perceptrons); • Decision Tree Ensemble; • **Classifier k nearest neighbors (KNN).**

Table 5. The attack classes.

id	Class name	Number of records
1	DoS	16353
2	Exploits	44525
3	Fuzzers	24246
4	Generic	58871
5	Normal	93000
6	Reconnaissance	13987

The sensitivity of the algorithm is equal to the proportion of false positive classifications FPR (a, X).

$$Sen = FPR(a, X)$$

A sensitive diagnostic test is called overdiagnosis – the maximum prevention of missing malicious code.

The specificity of the algorithm is calculated as follows:

$$Spe = 1 - TPR(a, X)$$

A specific diagnostic test only diagnoses for certain traffic related to network attacks.

In the course of the research, a series of experiments were carried out, the essence of which consists in determining the presence of an attack and attributing it to a specific class (Table 6).

Dependence of the Correct Rate indicator on the number of main components is presented in Figure 4.

A comparison of all methods is presented in table 9. The results given in the table are indicated with an accuracy of 0.01%.

Table 6. Classifier Parameters.

Classifier	Basic Parameters
Decision Trees Committee (RFT)	The maximum number of nodes in the decision tree is assumed to be 250.
Multilayer perceptron (MLP)	The number of neurons in the hidden layer was selected during training to achieve the minimum error on the test sample, the activation function of the hidden layer neurons is the hyperbolic tangent; The number of 5000 epochs of learning, the learning algorithm is conjugate gradients.
Decision Trees Committee (RFT) + main component method for feature selection	Before the classification, features are selected by the method of principal components. The maximum number of nodes of the decision tree is assumed to be 100. The results of the work of the "decision trees" method using feature selection by the principal component method on the test sample are presented in Table 8. The maximum number of nodes in the decision tree is assumed to be 250. The results of the "decision trees" method are presented in table 7.
Classifier based on k-nearest neighbors	Parameter k was hit to achieve optimal error on the test sample. $k \in [5; 100]$
Multilayer perceptron + Autocoder	Before making a classification, features are selected using a two-layer neural network autocoder

Table 7. Inaccuracy matrix for decision trees.

	C_1	C_2	C_3	C_4	C_5	C_6
C_1	**23317** **100%**	0 0%	0 0%	0 0%	0 0%	0 0%
C_2	0 0%	**2677** **76.55%**	172 1.52%	5 0.08%	28 0.68%	1 0.01%
C_3	0 0%	711 20.33%	**10156** **90.31%**	689 11.43%	3410 82.25%	239 1.65%
C_4	0 0%	15 0.43%	252 2.24%	**5256** **87.21%**	68 1.64%	23 0.16%
C_5	0 0%	94 2.69%	649 5.77%	73 1.21%	**632** **15.24%**	24 0.17%
C_6	0 0%	0 0%	17 0.15%	4 0.07%	8 0.19%	**14226** **98.02%**
Total	**23317** 100%	**3497** 100%	**11246** 100%	**6027** 100%	**4146** 100%	**14513** 100%

Table 8. Indicators of detection efficiency of the "decision trees" method depending on the number of components on the test sample.

Number of components	**2**	**4**	**6**	**8**	**10**	**12**	**14**	**16**
Average proportion of correctly recognized examples (correct rate)	0,695	0,725	0,755	0,764	0,771	0,776	0,805	0,811
Scatter	0,001	0,001	0,001	0,001	0,002	0	0,001	0

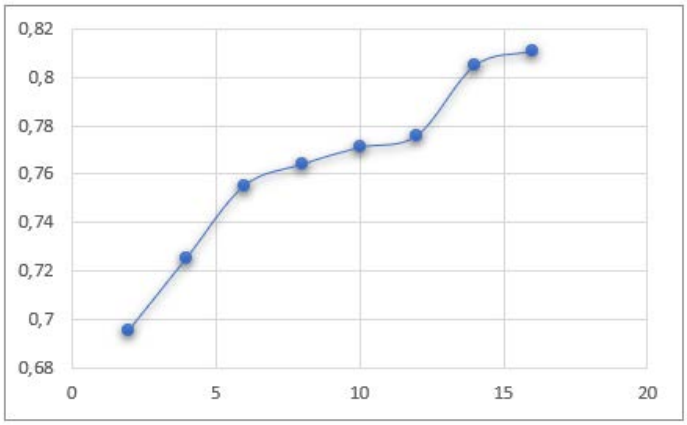

Figure 4. Dependence of the Correct Rate indicator on the number of main components.

Table 9. Comparative experimental results.

Name of the indicator	FitKNN	RFT	MLP	Autocoder	RFT+PCA
Sensitivity	0.1746	0.9877	0.1356	0.0842	0.9168
Specificity	0.9776	0.9897	0.9888	0.9917	0.9335
Correct rate,%	86.24%	89.67%	89.06%	88.58%	81.10%

As can be seen from the summary table, in the course of the experiments, the best indicators of correctly recognized images were shown by the algorithms “decision trees” (89.67%) and the multilayer perceptron (89.06%). Sensitivity indicators for FitKNN, MLP and autocoder methods do not exceed 20%.

When using the “decision trees” method together with the principal component method for decreasing the dimension, the indicators decrease (sensitivity - by 8%, specificity - by 6.6%, the proportion of correctly recognized examples - by 8.5%), and require more time and computational resources.

6. Conclusion

During the research the following tasks were solved:

1) The main security threats and the types of intruders in the industrial network of the enterprise are considered. A comparative analysis of software systems to ensure the safety of automated process control systems was conducted: Kaspersky Industrial CyberSecurity, Silent Defense, PT Industrial Security Incidents Manager, Honeywell Risk Manager.
2) A structural scheme of a network attack detection system based on data mining techniques has been developed.
3) Analyzed the data sets of network traffic, suitable for modeling the traffic of the industrial network of enterprises: KDD99 CUP, NSL-KDD, UNSW-NB15 for the task of detecting network attacks. The UNSW-NB15 set is selected for use in the system, since the number of attack classes is twice as large; test stand contained 33 subnets (NSL-KDD – 2 subnets); in collecting traffic on the network, 45 IP addresses participated in the exchange of information against 11 in NSL-KDD; traffic collection was carried out by several means (in NSL-KDD – Bro-IDS); the UNSW-NB15 set contains more attributes in the record (49 vs. 42 in NSL-KDD).
4) A software package has been developed that implements a comparative analysis of network attack detection algorithms. The most effective is the “decision trees” method with sensitivity indicators Sen = 1, specificity Spe = 0.9877, and the mean correct rate MCR = 89.67%.

7. References

[1] Montgomery, G. SCADA: Threat landscape [Electronic resource]. – Access mode: https://energy.gov/sites/prod/files/cioprod/documents/Cracking_Down_SCADA_Security_Garrett_Montgomery.pdf (20.09.2018).

[2] Langner, R. To kill a centrifuge – a technical analysis of what Stuxnet's creators tried to achieve [Electronic resource]. – Access mode: http://www.langner.com/en/wp-content/uploads/2013/11/To-kill-a-centrifuge.pdf (17.09.2018).
[3] Alert IR-ALERT-H-16-056-01 Cyber-Attack Against Ukrainian Critical Infrastructure [Electronic resource]. – Access mode: https://ics-cert.us-cert.gov/alerts/IR-ALERT-H-16-056-01 (23.09.2018).
[4] Ginter, A. SCADA Security. What's broken and how to fix it // Abterra Technologies, 2016. – 165 p.
[5] Steenstrup, K. IT and Operational Technology Alignment Innovation Key Initiative Overview [Electronic resource]. – Access mode: https://www.gartner.com/doc/2691517/it-operational-technology-alignment-innovation#a-98481934 (23.09.2018).
[6] Meltzer, D. Industrial Cyber Security for dummies / D. Meltzer, J. Lund [Electronic resource]. – Access mode: http://www.vectorinfotech.com/assets/files/Industrial-Cyber-Security-for-dummies.pdf (21.09.2018).
[7] Kaspersky Industrial CyberSecurity [Electronic resource]. – Access mode: https://ics.kaspersky.ru/ (23.09.2018).
[8] Kaspersky Industrial Cybersecurity. [Electronic resource]. – Access mode: https://softprom.com/sites/default/files/materials/KICS_rus_0816.pdf (06.10.2018).
[9] Positive Technologies Industrial Security Incident Manager [Electronic resource]. – Access mode: https://www.ptsecurity.com/ru-ru/products/isim/ (06.10.2018).
[10] Product | Security Matters SilentDefense [Electronic resource]. – Access mode: https://www.secmatters.com/product (07.10.2018).
[11] SilentDefense datasheet [Electronic resource]. – Access mode: https://www.secmatters.com/hubfs/Security_Matters-March2017/PDF/SilentDefense-Datasheet.pdf (07.10.2018).
[12] Kashyap, S. Soft Computing Based Classification Technique Using KDD 99 Data Set for Intrusion Detection System / S. Kashyap, P. Agrawal, V.C. Pandey, S.P. Keshri // International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering. – 2013. – Vol. 2(2). – P. 1398-1405.
[13] Nour, M. UNSW-NB15: a comprehensive data set for network intrusion detection system (UNSW-NB15 network data set) / M. Nour, J. Slay // Military Communications and Information Systems Conference (MilCIS). – Canberra, Australia, 2015.
[14] Perrin, Ch. The CIA Triad [Electronic resource]. – Access mode: https://www.techrepublic.com/blog/it-security/the-cia-triad/ (20.09.2018).
[15] Easton, V.J. Hypothesis testing / V.J. Easton, J.H. McColl [Electronic resource]. – Access mode: http://www.stats.gla.ac.uk/steps/glossary/hypothesis_testing.html (17.10.2018).

Acknowledgments
This work was supported by the Russian Foundation for Basic Research, research №17-48-020095.