One of the biggest reasons that lead to violations of the security of companies' services is obtaining access by the intruder to the legitimate accounts of users in the system. It is almost impossible to fight this since the intruder is authorized as a legitimate user, which makes intrusion detection systems ineffective. Thus, the task to devise methods and means of protection (intrusion detection) that would make it possible to identify system users by their behavior becomes relevant. This will in no way protect against the theft of the data of the accounts of users of the system but will make it possible to counteract the intruders in cases where they use this account for further hacking of the system. The object of this study is the process of protecting system users in the case of theft of their authentication data. The subject is the process of identifying users of the system by their behavior in the system. This paper reports a functional model of the process of ensuring the identification of users by their behavior in the system, which makes it possible to build additional means of protecting system users in the case of theft of their authentication data. The identification model takes into consideration the statistical parameters of user behavior that were obtained during the session. In contrast to the existing approaches, the proposed model makes it possible to provide a comprehensive approach to the analysis of the behavior of users both during their work (in a real-time mode) and after the session is over (in a delayed mode). An experimental study on the proposed approach of identifying users by their behavior in the system showed that the built patterns of user behavior using machine learning methods demonstrated an assessment of the quality of identification exceeding 0.95

Keywords: information protection, user identification, behavior model, machine learning methods

UDC 004.492.3
DOI: 10.15587/1729-4061.2022.259099

DEVISING AN APPROACH TO THE IDENTIFICATION OF SYSTEM USERS BY THEIR BEHAVIOR USING MACHINE LEARNING METHODS

Vitalii Martovytskyi
Corresponding author
PhD, Associate Professor
Department of Electronic Computers*
E-mail: vitalii.martovytskyi@nure.ua
Oleksandr Sievierinov
PhD, Associate Professor
Department of Information Technology Security*
Oleksii Liashenko
PhD, Associate Professor
Department of Electronic Computers*
Yuri Koltun
PhD, Associate Professor
Department of Information and Network Engineering*
Serhii Liashenko
Doctor of Technical Sciences, Professor
Department of Life Safety**
Viktor Kis
PhD, Associate Professor
Department of Mechatronics and Mashine Elements**
Vladyslav Sukhoteplyi
Senior Instructor
Department of Radioelectronic Systems of Control Points of Air Forces***
Andrii Nosyk
PhD, Senior Researcher
Department of Multimedia Information Technologies and Systems
National Technical University «Kharkiv Polytechnic Institute»
Kyrpychova str., 2, Kharkiv, Ukraine, 61002
Dmytro Konov
Researcher
Research Laboratory***
Dmytro Yevstrat
PhD, Associate Professor
Department of Information Systems
Simon Kuznets Kharkiv National University of Economics
Nauky ave., 9-A, Kharkiv, Ukraine, 61166
*Kharkiv National University of Radio Electronics
Nauky ave., 14, Kharkiv, Ukraine, 61166
**State Biotechnological University
Alchevskikh str., 44, Kharkiv, Ukraine, 61002
***Ivan Kozhedub Kharkiv National Air Force University
Sumska str., 77/79, Kharkiv, Ukraine, 61023

Received date 13.04.2022
Accepted date 15.06.2022
Published date 30.06.2022

How to Cite: *Martovytskyi, V., Sievierinov, O., Liashenko, O., Koltun, Y., Liashenko, S., Kis, V., Sukhoteplyi, V., Nosyk, A., Konov, D., Yevstrat, D. (2022). Devising an approach to the identification of system users by their behavior using machine learning methods. Eastern-European Journal of Enterprise Technologies, 3 (3 (117)), 23–34. doi: https://doi.org/10.15587/1729-4061.2022.259099*

1. Introduction

With the development of computing capabilities, the increasing popularity of social networks, and the use of various Web services, modern business is digitalizing its assets. Additionally, the global coronavirus pandemic prompted the transition of companies to the Internet space. In order to remain competitive and evolve, companies need to change

the approach to organizing their work and communication among all participants [1]. One of these approaches is the use of cloud technologies.

Cloud technologies are one of the main tools of digitalization, without which it is difficult to imagine not only the development of business of any scale but also the life of modern person. Currently, cloud technologies are partially or fully used in various business sectors. Moreover, these approaches are much more important for companies than it may seem at first glance.

For technologically advanced companies that employ distributed teams, cloud solutions make it possible to organize modern mechanisms of remote work. And for medium and small companies that do not have enough resources to build their own high-quality infrastructure and server capacities, they make it possible to delegate part of their work to the cloud.

To ensure the safety of the use of cloud technologies, the following rules should be followed:

– protection of confidential information is the area of responsibility of the company itself at all levels, from the manager to ordinary employees;

– for traffic encryption, it is usually enough to use SSL/TSL but be sure to take into consideration the relevance of certificates. One can use a VPN, thereby almost completely guaranteeing the security of traffic when moving an unprotected channel;

– clear prescribing of SLA infrastructure settings and their verification. The SLA describes the conditions for the provision of services and establishes a list of such services, as well as the rules by which the customer will use these services. At the same time, SLA is one of the main mechanisms that make it possible to manage the quality of IT services.

However, no matter how companies follow these recommendations and no matter how protected the vendors who provide them with services are protected, there is still a risk of breaking the system. And the main weak point in the system has always been and remains the users themselves [2].

In 2021, cybercriminals are still taking advantage of the situation around the COVID-19 pandemic. Remote work mechanisms, isolation of employees, and the current situation with vaccination increase the interest of cybercriminals in social engineering methods. And thus, intruders receive legitimate data from the official accounts of users with the help of which they then steal data or bypass the security systems of the enterprise.

The ITRC's 2021 Business Impacts Report shows that more than half of small businesses have been affected by data breaches or security breaches, and a third of companies have been hacked at least three times.

ITRC requested information about the impact of cyberattacks on the company directly from small business owners and executives affected by security and data breaches. In the 2021 Business Aftermath Report, 417 small business executives and 1,050 ordinary consumers answered questions in two separate surveys on the impact of cybercrime on small businesses. The conclusions include the following facts:

– fifty-eight (58) percent of small businesses are faced with data breaches, security breaches, or both. Three-quarters of these businesses faced at least two hacks, and one-third faced at least three hacks;

– forty-four (44) percent of small businesses spent USD 250,000 to USD 500,000 to cover the costs associated with the data breach. Sixteen (16) percent of small businesses spent USD 500,000 to USD 1 million;

– thirty-six (36) percent of small businesses took out loans to cover the cost of security breaches, and 34 percent used cash reserves;

– fifteen (15) percent reduced the number of staff to reduce costs;

– external threats are responsible for 40 percent of attacks. Intruders and contractors are responsible for 35 percent of the attacks.

Another important finding presented in the 2021 business impact report is that 42 percent of small businesses take one to two years to return to normal operation. Twenty-eight (28) percent of respondents say their business takes three to five years to fully recover.

One of the biggest reasons that led to security breaches of companies' services is to gain access by an intruder to legitimate user accounts of the system. It is almost impossible to fight this since the intruder is authorized as a legitimate user, which makes intrusion detection systems ineffective.

Thus, the task of devising methods and means of protection (intrusion detection) that would make it possible to identify system users by their behavior becomes relevant. This will in no way protect against the theft of the data of the accounts of users of the system but will make it possible to counteract the intruders in the cases where they use this account for further hacking of the system.

2. Literature review and problem statement

Cybersecurity has been and remains one of the priority areas of development in many countries. The growth of research related to cybersecurity is clearly seen today, due to the growing number of cybercrimes and cases of cyberterrorism. Hacker attacks are recorded in all corners of the world. Among the most resonant is the spread of WannaCry viruses [4], Petya/NotPetya [5], which caused significant damage to banking systems and large companies in different countries.

The authors of [6] conducted a study into the processes of ensuring the protection of the Web application from attacks aimed at obtaining unauthorized access to the functions of the administrator of the content management system. As a result, a method for selecting protection measures for a Web application was presented, which is based on a method for assessing the success rate of an attack. Since all protection measures differ in cost, efficiency, and impact on different vectors of attacks, the choice determines a set of countermeasures, which provides the maximum decrease in the success rate of the attack. Therefore, the change in the set of countermeasures leads not only to a change in their parameters but also to a change in the parameters of the attack tree. The task of choosing protection measures is a nonlinear task of integer programming with Boolean variables [6].

In response to the growing number of vulnerabilities of organizations to data breaches, the authors of [7] presented an integrated model of data leak management risk, based on a systematic review of the literature. Theoretical research expands the totality of knowledge about data leakage management by identifying and updating conceptual ideas about the risks of data leakage and permits (actions), as well as by providing a basis for organizations to respond to data leakage incidents (heuristics). In practice, the study provides key information that practitioners can use to organize effective data leakage management based on comprehensive risk element profiles and elimination methods.

Cloud computing technology provides access to a pool of configured resources, including storage space, applications, services, and an on-demand network. The use of cloud technologies in the organization minimizes the efforts of the organization to meet the needs of its customers. One of the main advantages of cloud computing is the Single Sign-On Method (SSO), which allows the user to access multiple application services using single user credentials. There are many issues and problems in cloud computing that need to be discussed. However, preventing attacks on the security system is much more difficult while maintaining the confidentiality of agent users. In [8], the authors propose an SSO-based biometric authentication architecture for cloud computing services to overcome security and privacy attacks. Biometric authentication is effective for resources controlled by end devices when accessing cloud services as these devices are computationally inefficient for processing user information during authentication. Accordingly, with the help of the proposed architecture, an attack on security in cloud computing is minimized. The proposed architecture also includes a new approach in which there is a relationship between the user agent and the service provider. In this, users' agents can use their fingerprint when requesting registration and access to various cloud application services in the cloud. Based on a comparative study with several existing architectures, the main points of the proposed architecture were presented. However, this approach requires the use of additional equipment, which is not always advisable in the development of certain products of the company that rely on cloud computing technologies.

In the modern era of corporate computing, enterprise application integration (EAI) is a well-known and industry-recognized architectural principle based on a loosely connected application architecture where service-oriented architecture (SOA) is an architectural template for implementation. Although SOA can be implemented in a wide range of technologies, the implementation of SOA through web services is becoming a popular choice because of its simplicity, based on basic Internet protocols. Web services technology defines several supporting protocols and specifications, such as SOAP and WSDL, to communicate with the client and server for data exchange. In the 2000s, SOA had a new architectural paradigm called REpresentational State Transfer (REST), which is also used by system integration consortiums to integrate loosely connected service components called RESTful web services. This SOA implementation does not contain adequate security solutions, and its security depends entirely on network/transport security, which is outdated due to the latest web technologies such as Web 2.0 and its updated version of Web 3.0. Supplier security products have severe implementation limitations, such as the need for a secure organizational environment and violations of SOA specifications, which leads to new vulnerabilities.

Therefore, paper [9] proposes an adaptive security solution for REST, which uses public key infrastructure methods to improve the security architecture. A new security component called «intelligent security mechanism» is presented, which studies possible cases of security threats in SOA using algorithms for training artificial neural networks. This component predicts potential attacks on SOA based on the results obtained using the developed theoretical security model, and written algorithms as part of a security solution to prevent SOA attacks. Such solutions are quite promising in terms of ensuring the desired level of security in modern software products that are focused on cloud technologies.

In the current era of the cloud paradigm, the flow of services, applications, and data access is growing over the Internet. Typically, users need to authenticate multiple times to gain credentials and access the services or programs they want. In [10], the authors proposed a completely secure scheme to mitigate multiple authentications required of a particular user. In the proposed model, federal trust is created between two different domains: the consumer and supplier. All traffic entering the service provider is further divided into three stages, depending on the risks associated with the data of the relevant user. Single sign-on (SSO) and multi-factor authentication (MFA) are deployed to provide authentication, authorization, accounting, and availability (AAAA) to ensure the security and confidentiality of end-user credentials. The proposed solution uses the conclusion that the MFA achieves a better AAAA vs. SSO and Availability (AAAA) template to ensure the security and confidentiality of end-user credentials. This approach complicates the process of using the data of accounts of legitimate users of the system but if the intruder managed to penetrate the system with the help of these credentials, then further hacking of the system cannot be avoided.

Paper [11] reports the study of methods for identifying malicious software in computer systems. One of the most significant threats to the security of computer systems and information, in general, is malware, or computer viruses. It should be noted that this problem is exacerbated by the dynamic growth of the number of mobile devices, the general transition to cloud technologies, and the spread of Internet technologies, which leads to an increase in the number of malicious software. Study [11] considers software that generates the functions of the transitions of the store machine in accordance with the specified rules of grammar. Next, the incoming file is analyzed for the presence of specified features characteristic of malicious software and simulates the operation of a deterministic downstream store machine. As a result of the work of the store machine, a conclusion is formed about the possibility of infection of the computer system.

It is advisable to use such systems along with other methods and means of ensuring security. If an intruder manages to penetrate the system as a legitimate user, then such a subsystem will not allow him to use malicious software and remain invisible. However, such subsystems will in no way save the system data to which the legitimate user has access.

The authors of [12] analyzed the personalized preferences of users to their naming in terms of displayed names and then used various methods for calculating similarity to determine the similarity of functions contained in the displayed names. In addition, the authors also measured and analyzed the user interest schedule to further improve user identification efficiency. The authors combined one-to-one constraints with the Gale-Shapley algorithm to eliminate the one-to-many and «many to many» account ratio problems that often arise in the process of compiling results. Experimental results have shown that the proposed method makes it possible to identify the user using only a small amount of online data. However, due to the use of the Gale-Shapley algorithm, this method has all its shortcomings, and therefore, with a large number of users, the user-profile ratio will be a rather long operation. However, the results of the cited study prove the operability of methods for identifying users by their behavior within a certain system.

Additionally, in [13], researchers propose a neural network designed for a large number of users to identify the user in systems with a common account. That is, the work presents a solution to the problem of user identification, taking

into consideration sessions. MISS consists of two main components: one of them is the Dwell Graph neural network (DGNN), which includes the time an element stays in a neural network with a closed graph to record a change in user interest between sessions. The other is a user identification module (IM) employed to distinguish the behavior of different users under the same account. This method is used to give users recommendations for viewing content based on their preferences, that is, based on their behavior during the session. The cited study proves the possibility and effectiveness of using neural networks to build a function that describes the user's behavior during a session in the system. This will make it possible to build additional user session protection.

Paper [14] offers a simple but powerful approach to building a profile of user behavior when browsing in order to identify the user. The authors create user profiles that capture the power of user behavioral patterns that can be used to identify users. Their experiments show that these profiles may be more accurate at identifying users than decision trees when there is enough web activity and can achieve greater efficiency than support vector machines.

Based on current publications, it can be concluded that the development of modern identification algorithms makes it almost impossible to hack the methods of identification of users of the system. However, these methods are in no way protected from cases when the user voluntarily transfers his credentials to an intruder. And after analyzing modern approaches to identifying users by their behavior, one may say that these approaches are quite effective and are used in predicting user preferences.

Based on the above review, we can conclude that the task of identifying users by their behavior is similar to the task of detecting abnormal behavior and classification since the main goal of the methods used is recognition. Modern systems for analyzing user behavior are divided into two groups:

– systems that detect users with respect to certain templates (signatures) [6, 8, 11, 14];

– systems that detect abnormal behavior of a particular user [12, 13].

In the first case, the system has some database consisting of behaviors that make it possible to identify a third-party user if one appears in the system. In the second case, the system abstracts from the constituent processes of the components and tries to recognize the presence of behavior that is not characteristic of the user in general.

The signature analysis takes precedence over the analysis of abnormal behavior because the signature analysis is simpler. Additionally, this approach provides lower error rates of the first and second kinds. The system of recognition of abnormal behavior is able to catch qualitatively new objects (users) even if it has not met them before. However, with this approach, there is a high value of the frequency of errors of the second kind when the normal process is taken by an abnormal process. Therefore, it is necessary to develop approaches that combine the advantages of both signature analysis and methods for detecting anomalies.

3. The aim and objectives of the study

The purpose of this work is to devise an approach to identifying users of the system by their behavior during the session to detect intrusions into the system by intruders using legitimate user data. This will make it possible to build additional subsystems of protection that could be invisible to the intruder and which would be difficult to get around. This, in turn, will increase the overall protection of the system from reckless actions of users who accidentally or intentionally transfer their accounts to an intruder.

To accomplish the aim, the following tasks have been set:

– to build a functional model of the process of ensuring the identification of users by their behavior in the system;

– to conduct an experimental study on the proposed approach.

4. The study materials and methods

The task of identification is to establish mathematical relationships between the measured inputs and outputs at given measurements in time [15].

Identification is carried out using a model that can be configured, a particular structure whose parameters can be changed. The functional identification scheme can be represented in the following form shown in Fig. 1.

At each point in time, $t=1, 2, ..., n$, an external signal $u(t)$ is applied to the inputs of the configurable object and model. The object is also disturbed by some random variable $\xi(t)$. The output value of the object $y(t)$ depends on both the external influence and interference (noise) and the unknown vector of parameters w'. The output value of the model configured $y'(t)$ depends on the vector of the parameters that are configured. They are recalculated according to the algorithm that processes the vector of all observations $z(t)$. The set of these observations depends on certain identification tasks.

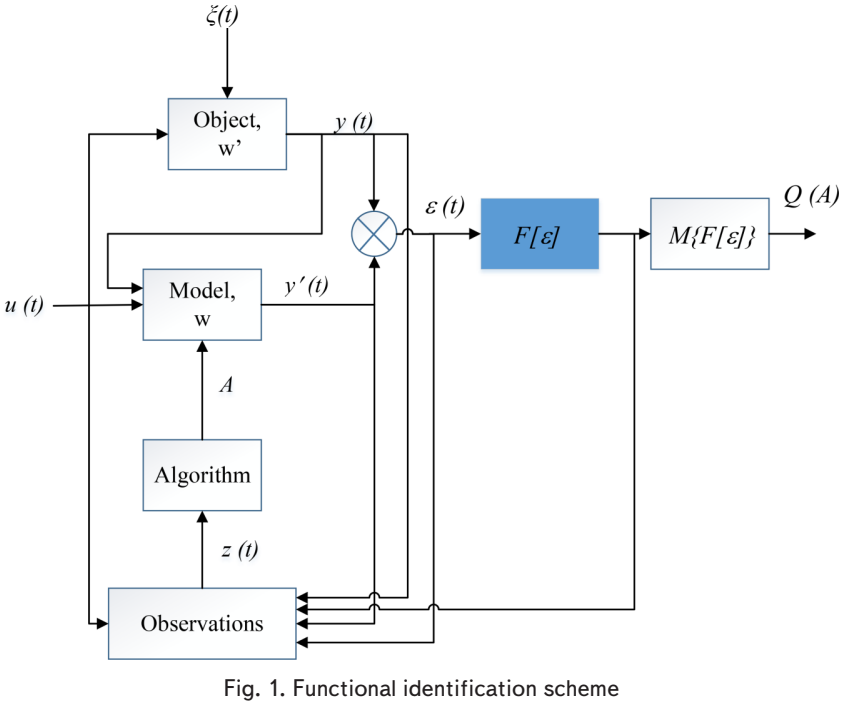

Fig. 1. Functional identification scheme

The difference in the initial values of the customized object and model forms some error:

$$\varepsilon\big(z(t), A\big) = y(t) - y'(t), \tag{1}$$

which is fed to the input of the functional converter shown in Fig. 2 in the blue rectangle. Further, it is always implied that the object operates under a stationary mode, that is, the probable characteristics of sequences $y(t)$, $y'(t)$, and, therefore, $z(t)$ do not depend on the time t. This regime is called normal operation mode.

Compliance with the customizable model and the object, that is, the quality of identification, is assessed by the criterion of the quality of identification:

$$Q(A) = M\Big\{F\Big[\varepsilon\big(z(t), w\big)\Big]\Big\}, \tag{2}$$

where F is the loss function and M is the symbol of mathematical expectation.

The criterion of quality of identification (2) is the average loss (error). The lower the average loss, the better the quality of identification. Improving the quality of identification is carried out by properly selecting the structure of the customizable model and selecting its parameters. The configuration of these attributes is carried out by an identification algorithm in the process of training the model.

The identification algorithm is determined by the loss function and the structure of the customizable model. According to observations of the input influence of the initial values of the object and the model, the identification algorithm changes the parameters of the latter so that the average losses reach a minimum. These conditions correspond to identification under the normal operation mode of the object.

To solve the task of identification, it is necessary, as follows from the functional scheme (Fig. 1):

– to determine the classes of objects;

– to select a model for this class of objects for configuration, that is, a model whose parameters can be changed;

– to select the criterion of quality of identification, which would characterize the difference between the initial values of the object and the results of the model;

– to form an identification algorithm that, using the values of input and output values available for observation, would change the parameters of the customized model so that the average losses with the growth of t reach a minimum.

In general, the task of identification can be reduced to a simple classification of objects, then machine learning methods can be used to build an identification model, such as:

– Linear Regression [16];

– Logistic Regression [17];

– K-Nearest Neighbors [18];

– Decision Trees and Random Forests [19];

– Support Vector Machines [20];

– Artificial Neural Networks [21].

Since the choice of machine learning method depends entirely on the complexity of the classification model itself, it is proposed to use logistic regression to demonstrate the presented identification approach. In the experimental part of the work, to demonstrate performance, a relatively non-complex model of user behavior is proposed, which can easily be described by logistic regression. And as the experiments have shown, other training methods produce very similar results.

Logistic regression is a way of constructing a linear classifier, which makes it possible to evaluate the posterior possibilities of ownership of objects to classes and is a separate case of generalized linear regression. It is assumed that the dependent variable acquires two values and follows a binomial distribution [16].

The main area of use of the proposed approach is the use of intrusion detection systems. Therefore, it is necessary to implement one of the simplest binary classifiers, identifying a legitimate user or violator, using logistics regression and its training with the help of conventional (full) and stochastic gradient descents.

In logistic regression, a linear algorithm for classifying a: $X \to Y$ of the following form is built:

$$a(x, w) = \operatorname{sign}\left(\sum_{j=1}^{n} w_j f_i(x) - w_0\right) = \operatorname{sign}\langle x, w\rangle, \tag{6}$$

where w_j is the weight of the j-th attribute, w_0 is the decision threshold, $w = (w_0, w_1, ..., w_n)$ is the weight vector, $\langle w, x\rangle$ is the scalar product of the object's features per weight vector. It is assumed that somebody artificially introduced «constant» zero feature: $f_0(x) = -1$.

The task of training logistics regression with L_2 regularization can be represented as follows:

$$Q(w, X) = \\ = \frac{1}{m}\sum_{i=1}^{m} \log\Big(1 + \exp\big(-y_i(x_i \cdot w)\big)\Big) + \frac{\lambda_2}{2}\|w\|^2 \to \min_w. \tag{7}$$

We all believe that $y_i \in \{-1, +1\}$, and the zero attribute is made single (that is, w_0 corresponds to the free term). We shall search with a gradient descent:

$$w^{(k+1)} = w^{(k)} - \alpha \nabla_w Q(w, X). \tag{8}$$

In the case of full gradient descent, $\nabla_w Q(w, X)$ is counted as is, that is, using all sampling objects. In the case of stochastic gradient descent $\nabla_w Q(w, X) \approx \nabla_w q_{i_k}(w)$, where i_k is the randomly selected number of the term from the functionality (the regularizer can be added to the amount by multiplying and dividing by m). The length of the step $\alpha > 0$ within this task is proposed to be taken equal to some relatively small constant.

The gradient by object x_i is calculated from the following formula:

$$\nabla_w Q(w, x_i) = -\frac{y_i x_i}{1 + \exp\big(y_i(x_i \cdot w)\big)} + \lambda_2 w. \tag{9}$$

As a criterion for stopping, it is necessary to use (simultaneously):

– verification of Euclidean rate of weight difference on two adjacent iterations (for example, less than some small number of about 10^{-6});

– reaching the maximum number of iterations (for example, 1000).

One can initialize the weight randomly or with a zero vector.

The probability of belonging of the object x to the class +1 is calculated as follows:

$$P(y = +1 \mid x) = \frac{1}{1 + \exp\big(-\langle w, x\rangle\big)}. \tag{10}$$

The matrix of objects-attributes X must be pre-normalized.

In logistics regression, L_1 regularization can also be used. Then the term $\lambda_1\|w\|_1$ is added to the loss function. In the formula for calculating the loss gradient by the coefficient vector, this term will correspond to $\lambda_1 sgn(w)$, where sgn is the calculation of the number sign applied to the vector of the coefficients by element-to-element.

5. Results of studying the approach to identifying users by their behavior in the system

5. 1. Functional model of the process of building a profile of user behavior and ensuring identification based on it

Identifying the user using behavior patterns is a relatively new and interesting task in terms of providing additional methods for protecting information and computer systems. Our paper builds on the study reported in [22–25] and proposes a simple but effective (as shown by experiments) method of identifying the user according to the profile of his behavior in the system.

To this end, describe the information (computer) system with the following tuple:

$$Sys=\{U,O\}, \quad (11)$$

where U is a set of users of the system, and O is a set of operations observed on certain programs of the information (computer) system for the corresponding users of the set U and is described by the following tuple:

$$O_i=\{A_i,Op_i,F_i\}, \quad (12)$$

where A_i is a program used by a system user, Op_i is an operation that has been used in the A_i program, and $F_i=[f_{i,1}, ..., f_{i,j}]$, $j\geq 1$ Op_i is a set of features of the observed operation.

Next, for a specific user U_k, we shall determine the user's behavior over the period $T=[t_0, t_1]$, as a set of sessions in the system $\{S_{k,1}(T), ..., S_{k,q}(T)\}$, where $S_{k,q}(T)$ is a subset of observable operations on certain O_i programs, for the k-th user of the system over the period T.

Thus, according to formula (2), for the user U_k whose behavior is $S_{k,q}(T')$, one needs to find the best and most reliable way to identify it. This technique determines whether the U_k profile at time T' is similar enough to the built U_k profile under a normal operation mode, using the methods presented in chapter 4 based on $S_{k,q}(T)$ sessions.

Additionally, in the process of identifying a user in the system by his behavior, one can face such a problem as a large number of errors of the second kind. This is due to the fact that in the process of long work in the system, the user gradually adopts and learns from the experience of other users. Thus, his profile of behavior gradually begins to differ from his own model of behavior built at the stage of the normal operation of the system. To solve this problem, it is proposed to estimate the probability P that the $S_{k,q}(T)$ session belongs to the U_k user. And if $1-P\geq\alpha$, where α is the permissible level of error, then, on the basis of valid data of $S_{k,q}(T)$ sessions for the k-th user, the user behavior model is rebuilt.

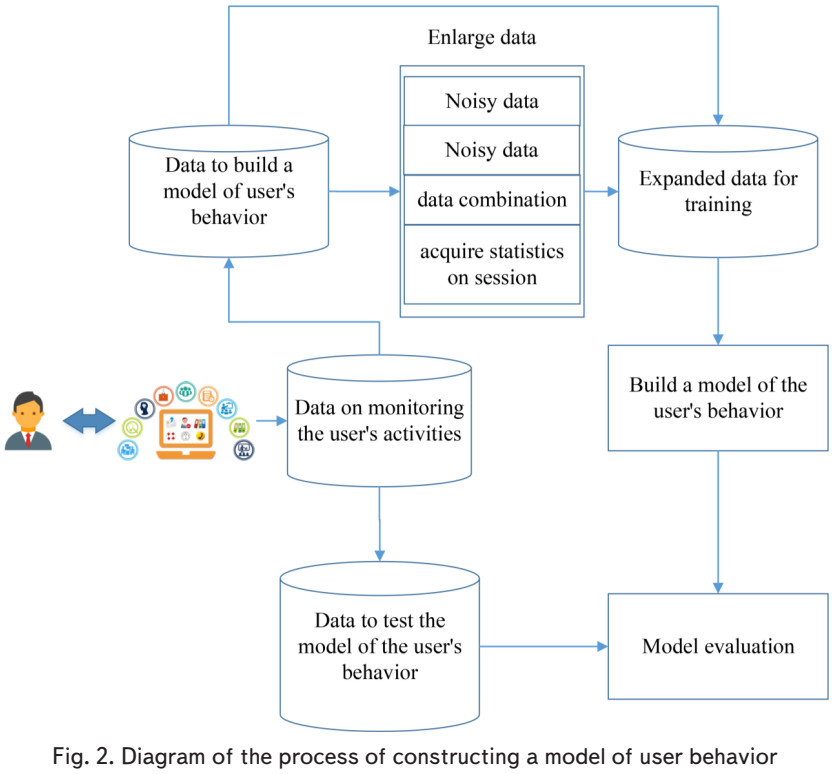

Fig. 2. Diagram of the process of constructing a model of user behavior

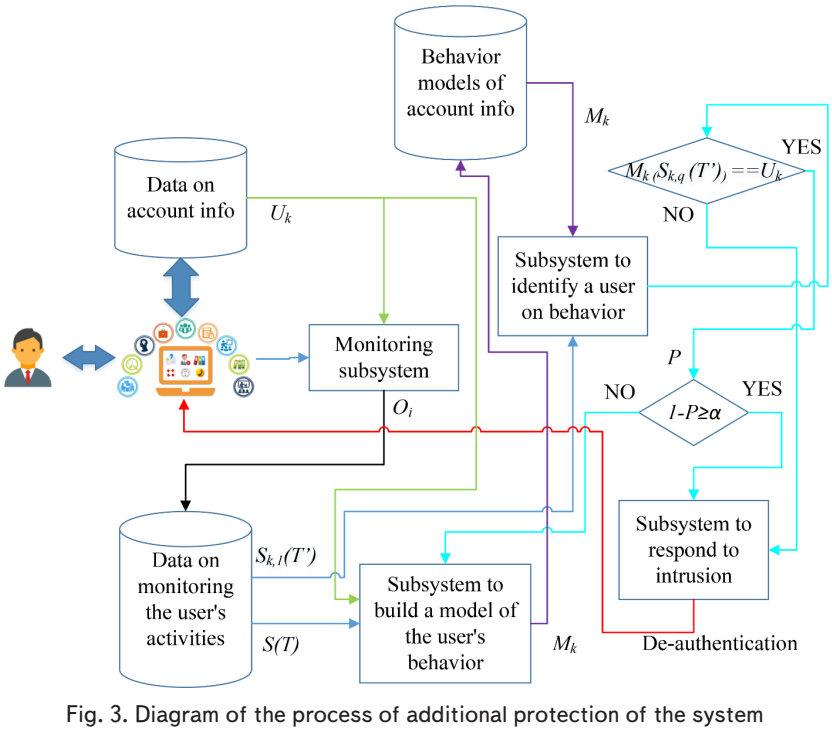

Fig. 3. Diagram of the process of additional protection of the system from intrusion by identifying user behavior

Fig. 2 shows in general form the diagram of the process of constructing a model of user behavior.

Fig. 3 shows a diagram of the process of additional protection of the system from intrusion by identifying user behavior.

In the process of working in the information system, the monitoring subsystem collects data on the user's activity in the O_i system, thereby forming a set of $S_{k,q}(T)$ sessions that will be used to identify and build a behavior model. Further, based on these data, according to the scheme shown in Fig. 2, the user behavior model is built. This process occurs when a new user is added to the system and under a normal operation mode. Next, the model is stored in the database for further use. After that, the identification subsystem, by using data about the current $S_{k,q}(T')$ session and the M_k user behavior model, assesses the compliance of the current user behavior with his model. Further, if the user's behavior does not correspond to his model, then the user is de-authentified. If the behavior coincides, then the value of the probability of compliance with the correct behavior $1-P\geq\alpha$ is checked and, in this case, the model of the behavior of this user is rebuilt.

5. 2. Experimental study of the approach to identifying users by their behavior in the system

For the experimental study, data were used from [26] presented on Kaggle, a platform for analytics competitions and predictive modeling.

These data are collected from the proxy servers of Blaise Pascal University. It consists of $17\cdot10^6$ strings of logs connected from more than 3,000 users and contains a user ID, time stamp, and domain names for each string. When forming a sample, two types of filters were applied to domain names: blacklist filters and filters based on HTTP requests. The authors used multiple lists of domain names to remove all domains that are treated as advertising. They also filtered the data by the status code obtained after a simple HTTP request for a domain name. After these steps, the authors received $4\cdot10^6$ strings. The authors split the file among 3,000 users to retrieve class files. More information about this dataset can be found in [27]. The study was conducted for 150 users with the highest number of requests.

To assess the performance of the proposed approach, an experiment was conducted to identify the user with the following restrictions:

1. As a set A, which describes the set of programs used, we shall consider the set of available domain names.

2. As operations in the program are monitored, the Op set will contain only one operation (visiting the site).

3. As attributes of each event (visiting a certain site), the F_i set will be the index of the visited site in the session and the timestamp.

4. The user sessions $S_{k,q}(T)$ are separated in such a way that they cannot be longer than 10 sites or at T equal to 30 minutes.

Thus, the tuple $S_{k,q}(T)$ will take the form:

$$S_{k,q}(T)=\left\{site^k_{1,q},time_{1,q},...,site^k_{10,q},time_{10,q}\right\}, \quad (12)$$

where *site* is the site indexes, and *time* is the timestamp when the site was visited.

Missed values may occur in sessions, which means that the session is less than 10 sites. These values will be replaced by 0.

Thus, a model of user behavior was built by analyzing sequences from several websites that were visited in a row by the same user, and determine whether it was Alice (legitimate user) or intruder (another person).

According to the scheme of the process of constructing a model of user behavior (Fig. 2), having data for building a model, one needs to increase the number of attributes. Since data about the site and timestamps are not enough to build a high-quality behavior model, one needs to create a more informative set of features using all the methods illustrated in Fig. 2.

The first thing one needs is to analyze the data for the presence of missed values and the distribution of sessions by Alice and the intruder.

After analyzing the distribution of sessions, the following values were obtained: Alice's sessions – 2297, and the sessions of intruders – 251264. It was then analyzed for missed values. The result is given in Table 1.

Thus, we can see that we have an uneven distribution of data and many different types given with missed values. In accordance with the scheme of the process of constructing a model of user behavior, shown in Fig. 2, we shall use methods to increase the number of attributes.

Table 1

Analysis of missing values in data

Attribute	Number of non-zero values	Attribute	Number of non-zero values	Attribute	Number of non-zero values
session_id	253,561	site4	244,321	time7	237,297
site1	253,561	time4	244,321	site8	235,224
time1	253,561	site5	241,829	time8	235,224
site2	250,098	time5	241,829	site9	233,084
time2	250,098	site6	239,495	time9	233,084
site3	246,919	time6	239,495	site10	231,052
time3	246,919	site7	237,297	time10	231,052

Since it is impossible to build a qualitative model of user behavior with such data, a list of attributes was created on their basis. As attributes, the number of sites in the session, the duration of the session, the day of the week, the hour of the beginning of the session, the hour of the end of the session, the beginning and end of the session, the minute, the day of the month, the index of time were chosen. At the same time, we immediately carry out the normalization of data.

The next stage was the analysis of some attributes to demonstrate their informativeness. To do this, distribution plots were built separately for Alice and intruders for some attributes. Fig. 4, 5 show the distribution of sessions by day of the week.

From Fig. 4, 5, one can see that the distribution of sessions by day for Alice differs from all others with less information about Wednesday, Saturday, and Sunday. However, there is a lot of input on Monday. Therefore, there is reason to believe that it is better to leave a weekday as a fictitious attribute and not to group them on this basis.

Then another important feature was analyzed – the hours when the sessions took place. The distribution of data by hours is presented separately for Alice and other users (intruders) in Fig. 6, 7, respectively.

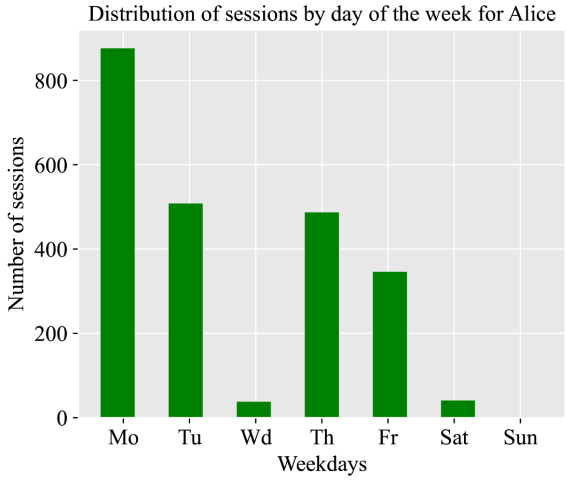

Fig. 4. Distribution of sessions by day of the week for user Alice

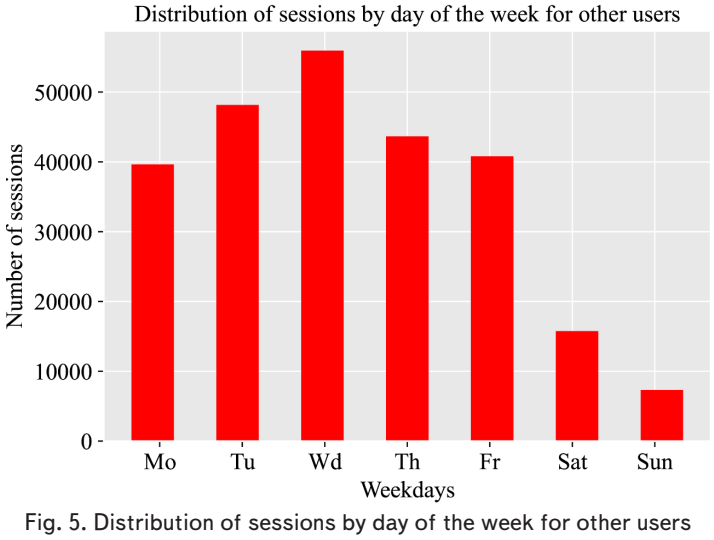

Fig. 5. Distribution of sessions by day of the week for other users

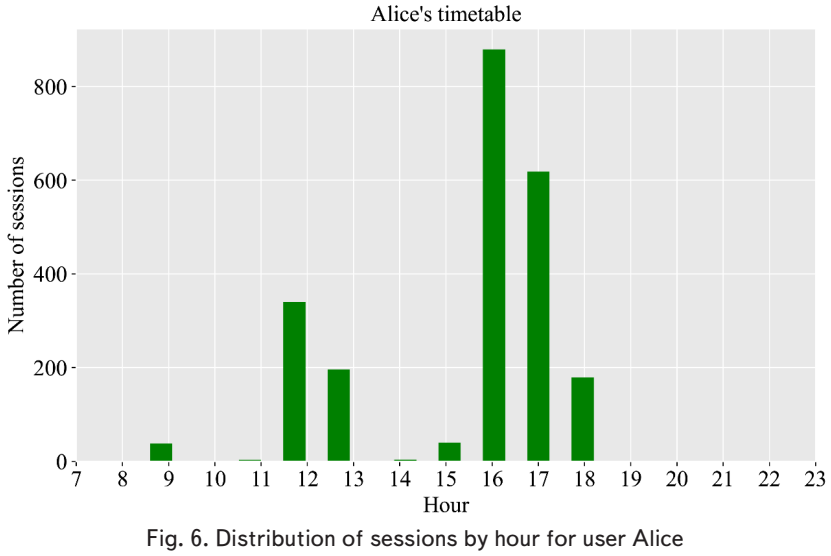

Fig. 6. Distribution of sessions by hour for user Alice

To better demonstrate the distribution of session duration, a logarithm of the session duration was used.

As one can see from Fig. 8, 9, there is a slight difference between the histograms that is not critical. However, this attribute can be left because in a combination with others, it gives a small increase in the accuracy of the model.

Further, based on these features, using machine learning methods, four models of user behavior were built, namely:

– logistic regression;
– algorithm of *k*-nearest neighbors;
– random forest;
– method of support vectors.

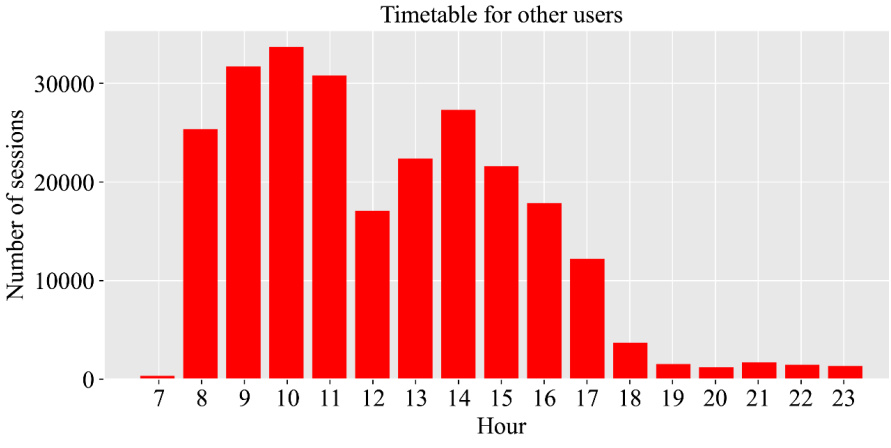

Fig. 7. Distribution of sessions by hour for other users

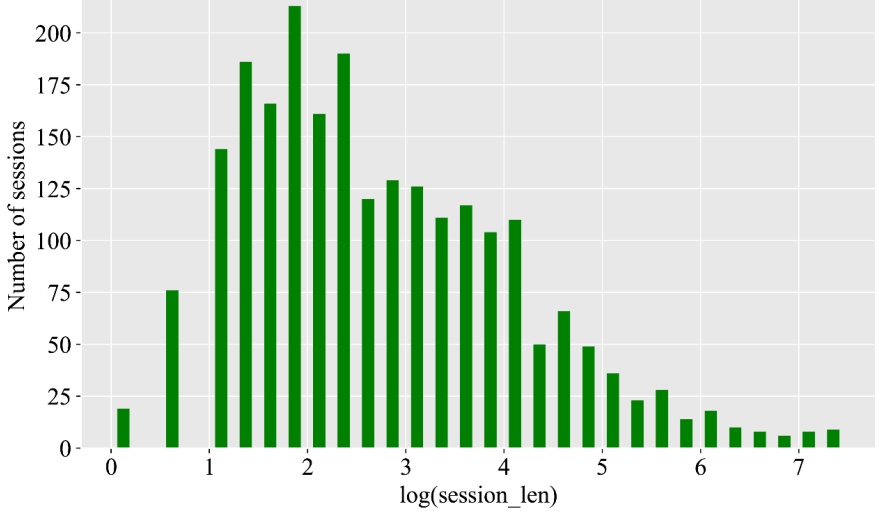

Fig. 8. Distribution of logarithm of the duration of sessions for Alice

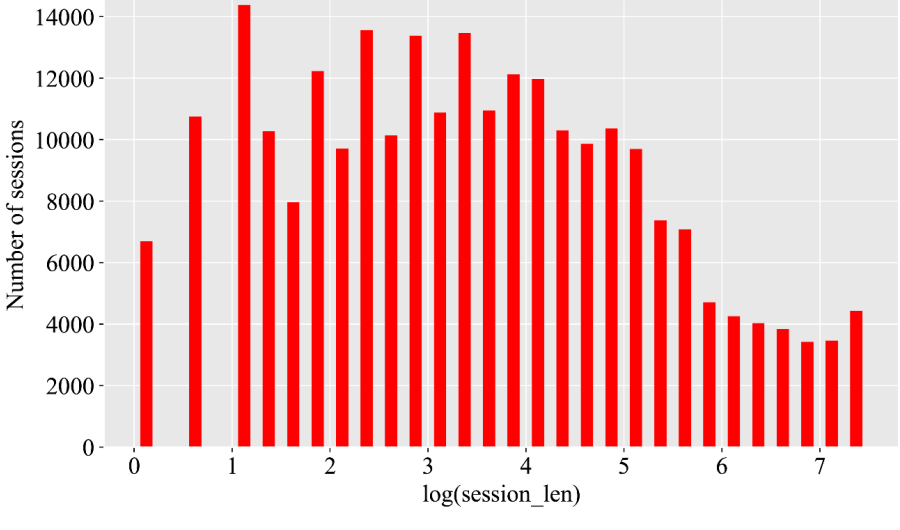

Fig. 9. Distribution of logarithm of the session duration for other users

Fig. 10 shows a ROC curve for the logistic regression.

For three other models, the ROC curve has a similar shape and, therefore, those plots are not given in the original work.

Table 2 gives an assessment of the quality of all four models.

As one can see from Table 2, all behaviors that are constructed using machine learning methods have a high assessment of the quality of models.

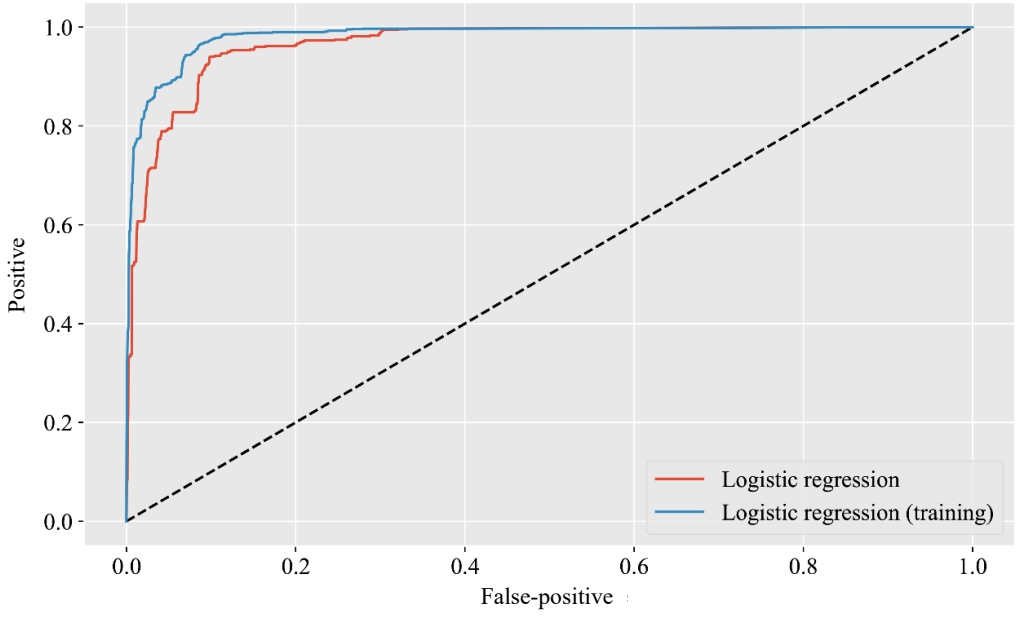

Fig. 10. ROC curve for logistic regression

Table 2

Assessment of the quality of models

Methods of constructing identification models	Quality assessment on a validation sample	Quality assessment on a test sample
Logistic regression	0.9830909934669896	0.9661144047123622
Algorithm of *k*-nearest neighbors	0.9831242141667668	0.9560378703913886
Random forest	0.9831829172812581	0.9671711140956807
Method of support vectors	0.9835188819490981	0.9656421000597966
Method from [13]	0.949346789111634	0.8765977451873890

6. Discussion of results of studying the approach to identifying users by their behavior in the system

The experiment consisted of three main stages:

1. Analysis of data for monitoring user actions and conducting filtering operations and expanding the set of features, by using statistical methods and a combination of available ones.

2. Constructing user behavior patterns using several machine learning methods.

3. Assessing the quality of built user behavior patterns.

Looking at the histograms in Fig. 4–9, one may notice that there are two main factors in Alice's dataset. The first is uneven distribution, which complicates the process of building a qualitative model of user behavior. The second is the repeatability of Alice's actions, which will enable machine learning methods in specific cases to quickly find patterns of user behavior, thereby improving the quality and stability of the model in general.

Fig. 4, 5 show the distribution of sessions by day of the week for user Alice and other users. From these distributions, we can see that Alice's day is different from all the others. This indicates that this is a good attribute for identification but it is impossible to completely rely on it. Additionally, this distribution shows that in real systems there may be a lack of some information and its noise and high-quality models should fight this.

In Fig. 6, 7, the distribution of hours of active work of users shows that Alice works on a fairly clear schedule. The distribution of hours is really different. One can observe periods when Alice does not use the Internet at all. This means that this feature can also be used quite effectively by machine learning methods to build a high-quality user model.

Additionally, one of the attributes that have been analyzed is the duration of the sessions. To better demonstrate the distribution of session duration, a logarithm of the session duration was used where one can observe the difference between Alice and other users' sessions. Although this difference is negligible but, as one can see in Fig. 10, its use also allowed us to build a quality model.

The ROC-curve plot for the logistic regression shows, however, that the built identification model based on the logistic regression very well revealed patterns of Alice's user behavior. The ROC curve plots for other models that were built using other machine learning techniques were not presented in the original work because they look quite similar. This indicates that a good set of features for building a model has been chosen. In contrast to the method of identification of user behavior, which is reported in [11], machine learning methods do a pretty good job of building a model of user behavior.

In contrast to the approach presented in [12], the proposed model makes it possible to provide an integrated approach to the analysis of user behavior, both during his operation in real-time) and after the end of the session (de-

ferred mode). This is achieved due to the fact that the computational complexity of the presented models is small, except for a model based on the random forest method.

To confirm this, one can analyze the data from Table 2 and see that all models are of high quality. Among them, a model is noticeable that is built using the random forest method. However, given that this is an ensemble method of machine learning, it uses more computational resources than other methods to build a model of behavior.

Our results in Table 2 show that, by using behavior models that are constructed using machine learning techniques with a probability of more than 0.95, it is possible to correctly determine whether the user's behavior profile corresponds to his authentication data or not. Since the presented approach is used as an additional identification factor for multifactor authentication, therefore, according to the NIST testing standard, the presented quality of models above 0.95 is acceptable.

Additionally, when comparing the developed approach with the approach to identifying users by web sessions reported in [13], we observe that the quality of the model of the proposed approach is 0.09 better. Such results can be explained by two factors:

– in addition to the original dataset from [23], which is the same in both cases, additional statistical analysis was used and the initial dataset was expanded by this data;

– since the main purpose of the developed approach to identification is used as an additional factor in multifactor authentication to identify non-legitimate users, the identification model is based on binary classification. This, in turn, allows machine learning models to more qualitatively approximate data to build a behavior model. However, unlike the methods in [13], the presented approach will not be effective in other areas of application of identification methods, such as identification for recommendation systems or identification for directional contextual advertising. This is due to the fact that in such cases a multiclass classification is used.

These results confirm that the model presented in the current work for building a profile of behavior and identifying the user behind it can be used as an additional means of ensuring the security of information systems.

The disadvantage of this approach to identifying users is that it takes some time to collect information about user behavior. This, in turn, allows the intruder to carry out a certain number of malicious operations.

For further development of this approach, attention should be paid to the devising an interactive model of behavior that accounts for the dynamics of user behavior and the trend analysis module, designed to identify possible changes in user behavior. The use of these models will predict the actions of the user, which could reduce the total time of identification of the user by his behavior.

7. Conclusions

1. A functional model of the process of ensuring the identification of users by their behavior in the system has been developed, which makes it possible to create additional means of protecting system users in the case of theft of their authentication data. The identification model takes into consideration the statistical parameters of user behavior (user signature) that were obtained during the session. This approach has allowed us to improve the quality of the identification of user behavior by 0.09 compared with the classical method reported in [13]. The following advantages of the proposed model should be highlighted:

– independence from the number of users in the system since this approach uses a model for assessing the behavior of an authorized user and current user characteristics;

– the ability to identify hidden patterns in user behavior; this is achieved through the use of machine learning methods;

– adaptation to changing user behavior. Since the user is gradually learning during work in any system, the model of behavior that was built in normal operation will gradually differ from the real behavior of the user. That is why an additional criterion for assessing the permissible deviation of current behavior from the constructed one was introduced. In the case of exceeding this criterion, the process of restructuring the user's behavior model on current data takes place.

2. An experimental study was conducted on the proposed approach of identifying the user by his behavior in the system. The constructed models of user behavior using machine learning methods showed an assessment of the quality of identification exceeding 0.95. The results of identification using the logistic regression and the «random forest» method showed almost the same result of identification. Therefore, it is not necessary to use complex models when describing user behavior, the main thing is to form informative attributes and correctly form a validation and test sample.

References

1. Lutsenko, I. (2016). Principles of cybernetic systems interaction, their definition and classification. Eastern-European Journal of Enterprise Technologies, 5 (2 (83)), 37–44. doi: https://doi.org/10.15587/1729-4061.2016.79356
2. The cyber-threat landscape: The digital rush left many exposed. Available at: https://www.pwc.com/us/en/services/consulting/cybersecurity-risk-regulatory/library/2021-digital-trust-insights/cyber-threat-landscape.html
3. The Identity Theft Resource Center's Inaugural 2021 Business Aftermath Report Shows the Impacts Identity Crimes Have on Small Businesses. Available at: https://www.idtheftcenter.org/post/the-identity-theft-resource-centers-inaugural-2021-business-aftermath-report-shows-the-impacts-identity-crimes-have-on-small-businesses/
4. Ghafur, S., Kristensen, S., Honeyford, K., Martin, G., Darzi, A., Aylin, P. (2019). A retrospective impact analysis of the WannaCry cyberattack on the NHS. Npj Digital Medicine, 2 (1). doi: https://doi.org/10.1038/s41746-019-0161-6
5. Gohwong, S. G. (2019). The State of the Art of Cryptography-Based Cyber-Attacks. International Journal of Crime, Law and Social Issues, 6 (2). doi: https://doi.org/10.2139/ssrn.3546334
6. Tetskyi, A. (2018). The method of selecting measures to protect the web application against attacks. Advanced Information Systems, 2 (4), 114–118. doi: https://doi.org/10.20998/2522-9052.2018.4.19

7. Khan, F., Kim, J. H., Mathiassen, L., Moore, R. (2021). Data breach management: an integrated risk model. Information & Management, 58 (1), 103392. doi: https://doi.org/10.1016/j.im.2020.103392
8. Alemu, B., Kumar, R., Sinwar, D., Raghuwanshi, G. (2021). Fingerprint Based Authentication Architecture for Accessing Multiple Cloud Computing Services using Single User Credential in IOT Environments. Journal of Physics: Conference Series, 1714 (1), 012016. doi: https://doi.org/10.1088/1742-6596/1714/1/012016
9. Beer, M. I., Hassan, M. F. (2017). Adaptive security architecture for protecting RESTful web services in enterprise computing environment. Service Oriented Computing and Applications, 12 (2), 111–121. doi: https://doi.org/10.1007/s11761-017-0221-1
10. Hussain, M. I., He, J., Zhu, N., Sabah, F., Zardari, Z. A., Hussain, S., Razque, F. (2021). AAAA: SSO and MFA Implementation in Multi-Cloud to Mitigate Rising Threats and Concerns Related to User Metadata. Applied Sciences, 11 (7), 3012. doi: https://doi.org/10.3390/app11073012
11. Gavrylenko, S., Chelak, V., Vassilev, V. (2018). Malicious software identification system provision on the basis of context-free grammars. Advanced Information Systems, 2 (2), 101–105. doi: https://doi.org/10.20998/2522-9052.2018.2.17
12. Xing, L., Deng, K., Wu, H., Xie, P., Gao, J. (2019). Behavioral Habits-Based User Identification Across Social Networks. Symmetry, 11 (9), 1134. doi: https://doi.org/10.3390/sym11091134
13. Wen, X., Peng, Z., Huang, S., Wang, S., Yu, P. S. (2021). MISS: A Multi-user Identification Network for Shared-Account Session-Aware Recommendation. Lecture Notes in Computer Science, 228–243. doi: https://doi.org/10.1007/978-3-030-73200-4_15
14. Yang, Y. (Catherine). (2010). Web user behavioral profiling for user identification. Decision Support Systems, 49 (3), 261–271. doi: https://doi.org/10.1016/j.dss.2010.03.001
15. Billings, S. A. (1980). Identification of nonlinear systems – a survey. IEE Proceedings D Control Theory and Applications, 127 (6), 272. doi: https://doi.org/10.1049/ip-d.1980.0047
16. Su, X., Yan, X., Tsai, C.-L. (2012). Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics, 4 (3), 275–294. doi: https://doi.org/10.1002/wics.1198
17. LaValley, M. P. (2008). Logistic Regression. Circulation, 117 (18), 2395–2399. doi: https://doi.org/10.1161/circulationaha.106.682658
18. Kramer, O. (2013). K-Nearest Neighbors. Intelligent Systems Reference Library, 13–23. doi: https://doi.org/10.1007/978-3-642-38652-7_2
19. Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1 (1), 81–106. doi: https://doi.org/10.1007/bf00116251
20. SVMLight. Support Vector Machine. Available at: https://www.cs.cornell.edu/people/tj/svm_light/
21. Zell, A. (1994). Simulation Neuronaler Netze. Chap. 5.2. Addison-Wesley.
22. Martovytskyi, V., Ruban, I., Sievierinov, O., Nosyk, A., Lebediev, V. (2020). Mathematical Model of User Behavior in Computer Systems. 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T). doi: https://doi.org/10.1109/picst51311.2020.9467944
23. Ruban, I. V., Martovytskyi, V. O., Kovalenko, A. A., Lukova-Chuiko, N. V. (2019). Identification in Informative Systems on the Basis of Users' Behaviour. 2019 IEEE 8th International Conference on Advanced Optoelectronics and Lasers (CAOL). doi: https://doi.org/10.1109/caol46282.2019.9019446
24. Ruban, I., Martovytskyi, V., Lukova-Chuiko, N. (2018). Approach to Classifying the State of a Network Based on Statistical Parameters for Detecting Anomalies in the Information Structure of a Computing System. Cybernetics and Systems Analysis, 54 (2), 302–309. doi: https://doi.org/10.1007/s10559-018-0032-1
25. Ruban, I., Martovytskyi, V., Lukova-Chuiko, N. (2016). Designing a monitoring model for cluster super–computers. Eastern-European Journal of Enterprise Technologies, 6 (2 (84)), 32–37. doi: https://doi.org/10.15587/1729-4061.2016.85433
26. Kahn, G., Loiseau, Y., Raynaud, O. (2016). A tool for classification of sequential data. ECAI 2016 (Workshop FCA4AI). Available at: https://hal.archives-ouvertes.fr/hal-02024913/document
27. Dia, D., Kahn, G., Labernia, F., Loiseau, Y., Raynaud, O. (2020). A closed sets based learning classifier for implicit authentication in web browsing. Discrete Applied Mathematics, 273, 65–80. doi: https://doi.org/10.1016/j.dam.2018.11.016