User Authentication in Critical Infrastructure Information Systems Using a Keyboard Handwriting Biometric Model

Vitalii Fesokha[1], Nadiia Fesokha[1], Ihor Subach[1,2], Artem Mykytiuk[2] and Ivan Horniichuk[2]

[1] *Kruty Heroes Military Institute of Telecommunications and Information Technologies, st. Knyaziv Ostrozkyh, 45/1, Kyiv, 01011, Ukraine*
[2] *National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", st. Verkhnoklyuchova, 4, Kyiv, 03056, Ukraine*

Abstract

In the context of enhancing the protection of critical infrastructure information systems against unauthorized access, this work considers the user authentication procedure based on an improved keyboard handwriting biometric model grounded in fuzzy logic. A primary prerequisite for the development of the proposed biometric model is the need to expand the formalization of user uniqueness within information systems during the registration stage. The limited feature space of existing biometric models, which arises from the constraints of ordinary keyboard properties, negatively impacts the reliability of the authentication procedure.

The construction of the biometric model relies on engineering behavioral patterns within a statistical dataset of keyboard handwriting, followed by the generation of new features and their description using fuzzy linguistic terms. During the configuration of the access control and user differentiation system in the information system, users are given the option to select the type of feature space for the keyboard handwriting biometric model: either a shared space for all users or a personalized one.

Furthermore, it is planned to detect any drift in the values of the user's keyboard handwriting features based on Kullback-Leibler divergence to ensure timely adaptation of the biometric model to the dynamics of the user's behavior. A comparative analysis of the results from user authentication experiments based on the proposed approach and existing authentication methods is also presented.

Keywords

Cyber security, unauthorized access, authentication, biometric model, keyboard handwriting, information systems, fuzzy logic, principal component method, data drift.

CEUR-WS.org/Vol-3887/paper7.pdf

1. Introduction

As of today, ensuring cybersecurity for critical infrastructure facilities, whose functions are directly tied to technological processes and/or services essential for national security, is a strategic priority for any nation.

Given that many cyber threat methods, including various types of cyberattacks—such as phishing, viruses, spyware, "man-in-the-middle" attacks, software vulnerabilities, and social engineering—share the objective of gaining unauthorized access to critical infrastructure information systems (IS), the task of ensuring data confidentiality, availability, and integrity is particularly crucial.

Access control and user differentiation systems are typically responsible for countering unauthorized access, especially during authentication, when the claimed identity of a user is verified for further authorization [1-3]. However, current authentication methods often fall short in effectively safeguarding IS from cyber threats, as evidenced by numerous recent incidents of security breaches [4-6]. This is primarily due to attackers' evolving strategies, which necessitate new authentication methods and solutions for IS users.

ITS-2023: Information Technologies and Security, November 30, 2023, Kyiv, Ukraine

vitaliifesokha@gmail.com (V. Fesokha); nadya19020202@gmail.com (N. Fesokha); igor_subach@ukr.net (I. Subach); mukuta8888@gmail.com (A. Mykytiuk); horniychuk.ivan@gmail.com (I. Horniichuk)
0000-0001-6612-1970 (V. Fesokha); 0000-0002-9797-5589 (N. Fesokha); 0000-0002-9344-713X (I. Subach); 0000-0002-8307-9978 (A. Mykytiuk); 0000-0001-6754-4764 (I. Horniichuk)

© 2023 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

An analysis of relevant literature [2,3,7-9] reveals that one of the most effective ways to prevent unauthorized access to IS resources is through access control and user differentiation systems based on analyzing users' behavioral biometric characteristics at the authentication stage, as they are practically impossible to fake. This approach involves identifying users based on their subconscious sensory and motor skills throughout their interaction with the IS, allowing the detection of the substitution of an already authorized user.

In IS, the most common practice for analyzing behavioral biometric characteristics at the authentication stage is through keyboard handwriting (KH), assessing typing indicators such as speed, rhythm, pressure, press duration, and time between key presses during password entry or typing of arbitrary text [2,3,7-9].

This highlights the need for further research to improve user authentication effectiveness in critical infrastructure IS based on KH.

2. Biometric model of keyboard handwriting of users

In most scientific works [8,10-13] focused on developing keyboard handwriting biometric models, the task of formalizing the uniqueness (individual subconscious characteristics) of users is constrained by a limited feature space in keyboard handwriting (typing speed, rhythm, key pressure, key press duration, and time between key presses). This limitation prevents these models from achieving adequate representation and, consequently, high accuracy in the authentication process. Therefore, this article examines a keyboard handwriting biometric model for users in critical infrastructure information systems, as proposed in [3,7], which allows for expanding the keyboard handwriting feature space through feature engineering, using fuzzy logic to generate additional features.

The construction of this keyboard handwriting biometric model involves the following stages.

2.1. The synthesis of the initial keyboard handwriting feature space

The initial keyboard handwriting feature space, denoted as S_{start}, whose features include (1):
d – typing dynamics – the time between key presses and the duration of key presses;
t_s – typing speed – the number of keystrokes divided by the typing duration.

$$S_{start} = \{d, t_s\}. \tag{1}$$

2.2. Formation of new features

It is evident that the features selected in the previous stage do not sufficiently capture the uniqueness of information system (IS) users. Thus, following the approach proposed in [3,7], new features are generated by defining behavioral patterns (templates) from a statistical dataset of keyboard handwriting during control text (password) input. Specifically, the duration of key presses by the user, denoted as Δt_i^r, (where i is the key identifier), is considered. This indicator is chosen due to its minimal variability for each user compared to other indicators. For example, the duration between key presses reflects the time required for the user's hand to move across the keyboard, which inherently displays excessive variability.

Behavioral patterns are established through statistical analysis of the Δt_i^r indicator by repeatedly entering the control text. This is then represented as a variation curve on a graph. Figure 1 presents the behavioral pattern function of the first author's keyboard handwriting, plotted at 12 points while entering their own 12-character password.

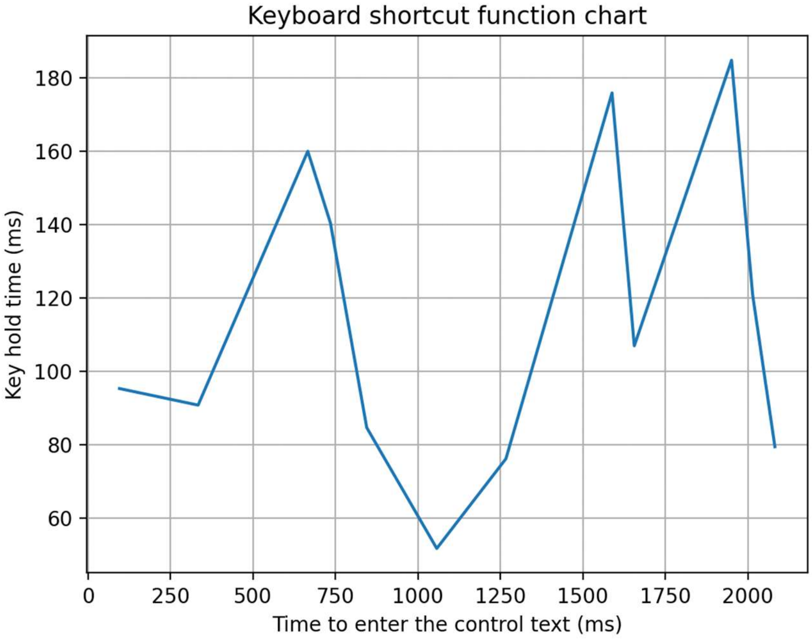


Figure 1: The graph of the keyboard handwriting behavioral pattern function of the first author while entering their own password

This curve describes the rate of change (the geometric interpretation of differentiation), which can be approximated by trigonometric functions to engineer new features. In line with [3,7], a subset of new keyboard handwriting features for the IS user is defined to create the final feature space for the keyboard handwriting biometric model M_{u_i}. Consequently, the presented curve is segmented into equal-length time windows t_i^w, which form a new subset of keyboard handwriting features S_{new} (Fig. 2).

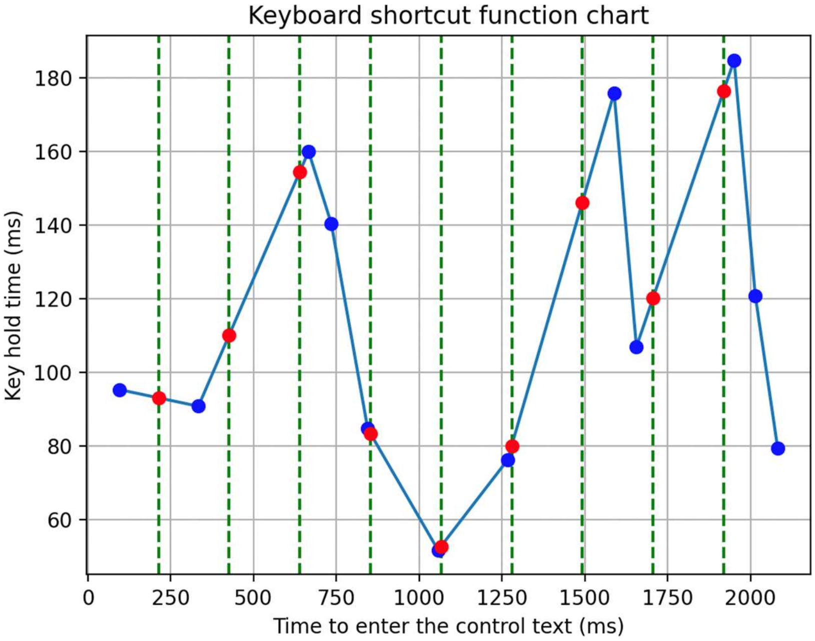


Figure 2: The graphical representation of the decomposition of the investigated curve $\Delta t_i^r$by the user u_i into time windows t_{1-10}^w.

As shown in Figure 2, the horizontal time axis is divided into 10 windows t_{1-10}^{w} by dashed lines. Blue dots represent key presses, while red dots mark the intersections of the Δt_i^r curve with the time window boundaries. This yields a new subset of features, S_{new} (2) where each feature corresponds to a specific time window t_i^w. The number of windows is chosen based on the average time required to enter passwords of 8 to 15 characters.

$$S_{new} = \{t_1^w, t_2^w, \dots, t_{10}^w\}. \tag{2}$$

Figure 3 displays curves from five attempts by the first author at entering the control text. These curves show a high degree of similarity, enabling the identification of a unique keyboard handwriting pattern for the IS user.

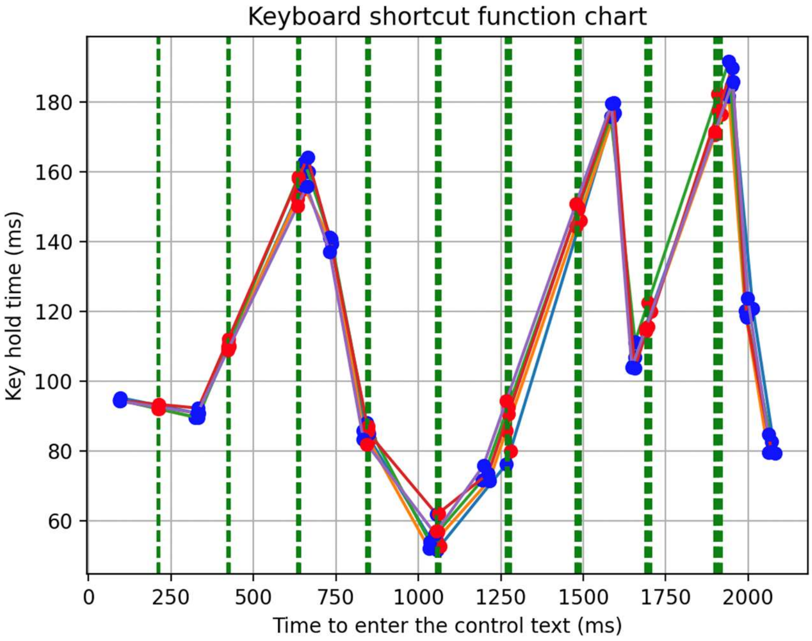


Figure 3: The graphical representation of the result of entering control text, with a total of 5 attempts, is presented in the form of curves

To ensure minimal variability in the keyboard handwriting pattern values, the control text should be practiced thoroughly until it becomes automatic. Without this level of familiarity, the analysis of arbitrary text input loses its effectiveness for identifying a behavioral pattern in keyboard handwriting, which is difficult to falsify.

2.3. Description of features using fuzzy linguistic terms

Since the values of a user's keyboard handwriting features exhibit some variability, the task of authenticating an IS user is essentially a process of iteratively assessing the degree of correspondence between their keyboard handwriting and the biometric model M_{u_i} using fuzzy logic methods [14, 15]. Here, the input linguistic variables are elements from the subsets. S_{start} and S_{new}. However, each time window t_i^w contains a different number of segments (piecewise-linear functions) sub_i of the curve that describes the keyboard handwriting behavioral pattern. Consequently, the expanded feature space can be represented analytically (3).

$$M_{u_i} = (S_{start} \cup S_{new}) \rightarrow \{d, t_s, t_1^w = \{sub_1, \ldots, sub_n\}, \dots, t_{10}^w = \{sub_1, \ldots, sub_n\}\}. \tag{3}$$

To describe the features S_{start} using fuzzy linguistic terms, we propose an approach that automatically determines the number of linguistic terms without requiring expert input, based on the statistical Silhouette method (4). This method calculates the optimal number of terms $mf_i, \dots, mf_n, i = \overline{1,n}$, where mf_i represents a triangular-shaped fuzzy linguistic term [16], and n is the number of terms.

The optimal number of terms mf_i is selected to maximize the silhouette indicator (4):

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, \tag{4}$$

where $s(i)$ – silhouette value of i for term T;

$a(i)$ – the average value of intra-term distance;

$b(i)$ – distance between terms in features d, t_s.

To describe the features S_{new} using fuzzy linguistic terms, we calculate the angle for each segment of the curve obtained by determining the value in degrees between the horizontal axis and the segment, using the cosine theorem (5) [7].

$$\cos\alpha = \frac{b^2 + c^2 - a^2}{2bc}. \tag{5}$$

After calculating the angle value, it is matched to the corresponding fuzzy term on a scale for fuzzy term determination (Fig. 4), with increments of 15 degrees.

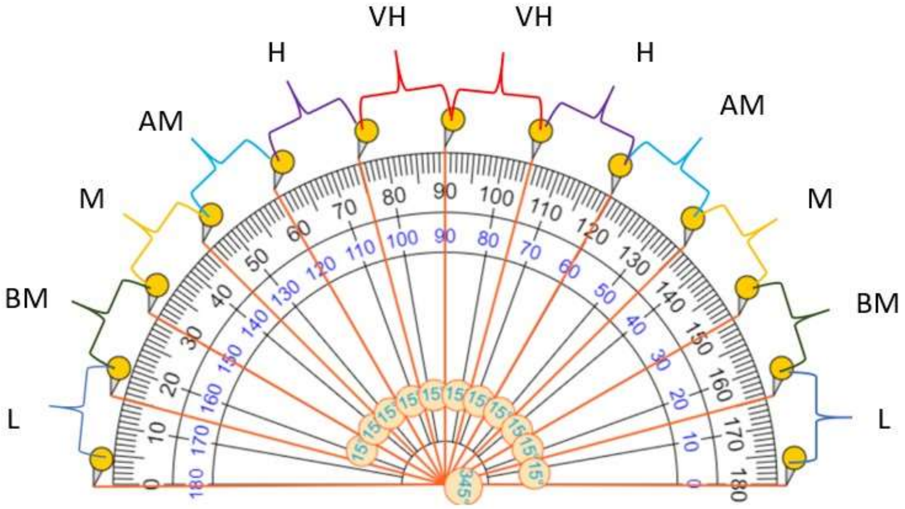
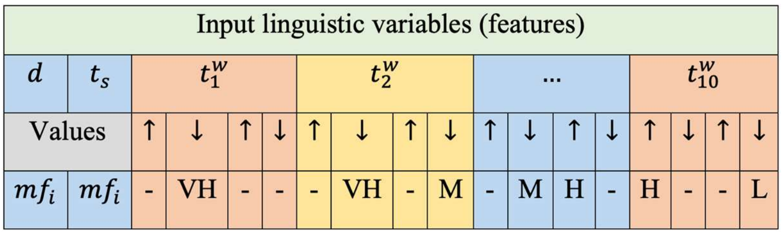


Figure 4: Scale for determining the fuzzy term of the linguistic variable

Using this scale (Fig. 4) enables the description of:

- increasing curve ↑ (very high, high, above average, medium, below average, low) – ranging from 0° to 90°;
- decreasing curve ↓ (very high, high, above average, medium, below average, low) – ranging from 90° to 180°.

Figure 5 shows the final feature space of the proposed keyboard handwriting biometric model for information system users.



Input linguistic variables (features)																	
d	t_s	t_1^w				t_2^w				...				t_{10}^w			
Values		↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓
mf_i	mf_i	-	VH	-	-	-	VH	-	M	-	M	H	-	H	-	-	L

Figure 4: The final feature space of the keyboard handwriting biometric model for information system users

To ensure effective model training, it is recommended that during the user registration phase, the user repeatedly enters the control text at least twice as many times as the number of features.

2.4. Selecting the type of feature space for the keyboard handwriting biometric model

In contrast to the approach proposed in [3, 7], which constructs a keyboard handwriting biometric model based on a common feature space for all users in the system, this work proposes allowing the cybersecurity administrator to select the type of feature space for the keyboard handwriting biometric model during the configuration of the user access control and segregation system in the information system (IS). The options include a common feature space for all users or a personalized feature space for individual users. Furthermore, the use of a common feature space introduces additional computational overhead. This is because, during the description of the time windows t_i^w derived from the decomposition of the user's keyboard handwriting curve, it is necessary to formalize not only the varying number of upward and downward trends within each t_i^w but also to maintain their precise sequence.

Additionally, when each system user is represented as a point in an nnn-dimensional common feature space, this space is often not linearly separable, as keyboard handwriting patterns of different individuals may intersect at certain points. This, in turn, can negatively impact model accuracy [7].

Therefore, it is recommended to construct the keyboard handwriting biometric model for critical infrastructure system users using a personalized feature space.

2.5. Detection of keyboard handwriting features drift

Over time, the effectiveness of the keyboard handwriting biometric model for IS users may decline, as typing speed tends to change with age or experience [7,17]. Thus, the task of detecting data drift—the change in statistical properties of data over time [18,19] – becomes essential to allow timely retraining of models with updated datasets. This involves applying methods to identify when model updates are necessary.

In this stage, the difference or similarity between the distributions of the model's training dataset $P(x)$ (historical data) and the new (accumulated) dataset $Q(x)$ is calculated using the Kullback-Leibler divergence [17]. The Kullback-Leibler divergence (or relative entropy) measures the difference between two probability distributions, indicating how much the information entropy of one distribution differs from another. This asymmetric measure ranges from 0 to infinity, where 0 indicates identical distributions. The Kullback-Leibler divergence is calculated for distribution Q relative to P using the following analytical expression (6):

$$KL(P \parallel Q) = P(X) log \frac{P(x)}{Q(x)}. \tag{6}$$

It is advisable to periodically analyze the divergence values obtained between the training dataset and the accumulated statistics at least once every six months. This ensures the adaptation of the proposed keystroke biometric model to changes in the characteristics of information system (IS) users. If constant values of $KL(P||Q) > 0$, are obtained, it is necessary to initiate the retraining of the keystroke biometric models for IS users.

2.6. Authentication of critical infrastructure information system users based on the keyboard handwriting biometric model

During attempts to access information system (IS) resources, users present a personalized identifier, which is authenticated using a password.

The next step involves periodic additional authentication of the IS user $u_i \in U$ throughout the entire session, based on the proposed approach to keyboard usage. The user recognition procedure for u_i among all system users U consists of evaluating the expression (7):

$$u_i = \{d, t_s, t_1^w = \{sub_1, .., sub_n\}, \dots, t_{10}^w = \{sub_1, .., sub_n\}\}. \tag{7}$$

With a specific periodicity or an event-based scheme configured by the administrator, the identified user $u_i \in U$ may be blocked from accessing the IS, prompting them to enter a control text to verify the authenticity of their claimed personalized identifier. The event-based scheme responds to any keyboard or mouse activity if the user has been inactive for more than 5 minutes. If no actions are performed with the respective devices during this time, the user is not prompted for control text. However, if an event occurs after the specified interval, the user will be asked to enter the control text. The allowable variability in the specific keyboard characteristics of the user is monitored through ranges defined by linguistic terms within the keyboard handwriting biometric model.

If there is a mismatch between the characteristics of the claimed personalized identifier and the keyboard handwriting biometric model of user $u_i \in U$, he current session will be terminated, and an appropriate message will be sent to the IS security system.

3. Evaluation of effectiveness

To assess the effectiveness of authentication systems, metrics for first and second kind errors are employed: the False Rejection Rate (FRR)—the probability of incorrectly rejecting a registered user—and the False Acceptance Rate (FAR)—the probability of granting access to an unregistered user [20-22]. These metrics are calculated as follows:

$$FRR = \frac{FN}{FN + TP}, \tag{8}$$

$$FAR = \frac{FP}{FP + TN}, \tag{9}$$

where *FN (False Negative)* – the number of times a registered user has been denied access;
TP (True Positive) – the number of times a registered user has been granted access;
FP (False Positive) – the number of times an unregistered user was granted access;
TN (True Negative) – the number of times an unregistered user was denied access.

The effectiveness of access control and segregation systems is greater when the values of FRR and FAR are minimized. Typically, one of these metrics is prioritized; specifically, prohibiting access to illegitimate users is considered more critical. To achieve this, it is essential to minimize the FAR. By reducing the False Acceptance Rate (*FAR*), the system can effectively thwart unauthorized access attempts, prioritizing security measures that maintain the integrity and confidentiality of the information system. This focus on minimizing *FAR* highlights the necessity of stringent authentication protocols to reduce the risk of unauthorized access and potential security breaches.

In commercial biometric authentication systems, the maximum acceptable value of *FAR* typically ranges from 10^{-3} to 10^{-6}. In systems with a large user base and a high level of security, this value can drop to as low as 10^{-9}. Meanwhile, the *FRR* may vary between 0.025 and 0.01; for systems with many users, this rate should not exceed 0.001 to 0.0001. These thresholds provide benchmarks for evaluating the performance and reliability of biometric authentication systems, ensuring they meet stringent security requirements while balancing user convenience and system efficiency [23, 24].

The analysis of the results demonstrates the practicality of the proposed solutions for user authentication in information systems. Specifically, the developed methodology enhances the reliability of information system authentication by reducing the *FRR* (type II error) by 2-3% compared to research results where the *FAR* (type I error) equals zero, and by 10-15% compared to research results where the *FAR* is greater than zero. This achievement aligns with the objectives of this work. A comparative analysis of the calculations based on the results of user authentication in information systems using the proposed approach versus existing methods [24] is presented in Table 1, focusing on the *FAR* and *FRR* metrics.

Table 1
Results of IS user authentication based on the proposed approach and existing methods in terms of FAR and FRR indicators

Researchers	Methods	FAR	FRR
European Access Control Standard	-	0,001%	1%
Bleha and Obaidat	perceptron	8%	9%
Nguyen, Le	neural network	4.12%	5.55%
Draffin, Zhu	neural network	14%	2.2%
Ahmed and Traore	neural network	0%	5.01%
Modi and Elliott	standard deviation	0.33%	94.87%
Trojahn, Arndt	mean square deviation	4.19%	4.59 %
Corpus, Gonzales	standard deviation	7%	40%
De Ru and Eloff	fuzzy logic	0-15%	-
Proposed solution	fuzzy logic, principal component analysis	0%	3.2%

4. Conclusion

In summary, the findings underscore the effectiveness of utilizing users' dynamic biometric traits for authentication, providing a robust safeguard for information security. Authentication decisions in these systems hinge on comparing the user's biometric model with data collected during the authentication process. The user's biometric model is developed through an analysis of specific individual characteristics, making systems that leverage keyboard handwriting recognition particularly valuable.

A biometric model of keyboard handwriting for users in critical infrastructure information systems has been proposed. This model expands the feature space of keyboard handwriting by analyzing behavioral patterns within a statistical dataset, generating new features, and describing them using fuzzy linguistic terms.

Additionally, the proposed model includes the detection of drift in users' keyboard handwriting characteristics using Kullback-Leibler divergence. This ensures timely adaptation of the biometric model to the dynamics of user behavior.

The practical application of the improved biometric model of keyboard handwriting patterns has demonstrated its effectiveness in recognizing users within access control and segregation systems in critical infrastructure facilities. This model enhances the reliability of user authentication in information systems by reducing the false rejection rate (*FRR*) by 2-3% compared to previous research results where the false acceptance rate (*FAR*) is zero, and by 10-15% compared to studies where the *FAR* exceeds zero.

References

[1] Chunariova A.V., Chunariov A.V. Analysis of existing authentication systems of information and communication systems and networks. Ukrainian Scientific Journal of Information Security, 2012 №2 (18). P. 65-70.

[2] Fesokha V. V. Analysis of existing solutions for authentication of users of information systems and special purpose networks/ V. V. Fesokha, N. O. Fesokha, O. D. Dobroshtan // Collection of scientific papers MITIT. 2020. № 3. P. 129–136.

[3] Fesokha V. V., Fesokha N. O. Model of fuzzy authentication of users of information systems of military administration based on behavioral biometrics // Protection of information. 2021. V. 23, № 2. P. 116–123.

[4] Hacks that became top news in 2023. https://10guards.com/ua/articles/data-breaches-that-hit-the-headlines-in-2023.

[5] Cyber security in the information society: Informational and analytical digest / resp. ed. O. Dovgan; according to O. Dovgan, L. Lytvynova, S. Dorogykh; State scientific institution «Institute of information, security and law of the National academy of legal sciences of Ukraine»; National Library of Ukraine named after V. I. Vernadskyi. – К., 2023.– No9 (september). – 351 p.

[6] Top 10 Cloud Security Incidents in 2022. https://www.immuniweb.com/blog/top-10-cloud-security-incidents-in-2022.html.

[7] Fesokha V. V., Fesokha N. O. The method of regularization of the feature space of the biometric model of keyboard handwriting of users of military information systems based on factor analysis. Collection of scientific papers MITIT. Kyiv. 2023. № 3. P. 152–162.

[8] Liashenko H. Y., Astrakhantsev A. A. Study of the effectiveness of biometric authentication methods. Information processing systems. 2017. V. 2 (148). P. 111–114. DOI: https://doi.org/10.30748/soi.2017.148.20.

[9] Yevetskyi V., Horniichuk I. Analysis of stability of the user’s keyboard handwriting characteristics in the biometric authentication systems // Collection “Information technology and security”. 2018. Vol. 6, no. 2. P. 19–28. URL: https://doi.org/10.20535/2411-1031.2018.6.2.153487.

[10] Kim J., Kang P. Recurrent neural network-based user authentication for freely typed keystroke data. 16 Jun 2018. URL: https://arxiv.org/abs/1806.06190.

[11] Continuous authentication by free-text keystroke based on CNN and RNN / X. Lu et al. Computers & Security. 2020. Vol.96. P. 101861. DOI: https://doi.org/10.1016/j.cose.2020.101861.

[12] Krutohvostov D., Khitsenko V. Password Authentication and Continuous Authentication by Keystroke Dynamics Using Mathematical Statistics. Voprosy kiberbezopasnosti. 2017.

[13] Chalaia L. User identification model based on keyboard handwriting. Artificial intelligence. 2004. V.4. P. 811–817.

[14] Analysis of research and publications [User identification model based on keyboard handwriting: website.URL: http://www.iai.dn.ua/public/JournalAI_2004_4/Razdel8/06_Chala ya.pdf .

[15] Chalaya L. E. System for recognizing keyboard handwriting of users based on a poly-Gaussian algorithm: article. Artificial Intelligence: scientific-theoretical magazine. 2004. № 4.

[16] Fesokha V. V., Subach I. Y., Kubrak V. O., Mykytiuk A. V., Korotaiev S. O. Zero-day polymorphic cyberattacks detection using fuzzy inference system. Austrian Journal of Technical and Natural Sciences. 2020. № 5–6. P. 8–13.

[17] Fesokha N. O. Determining the necessity of data drift of the biometric model of keyboard handwriting of users of military information systems based on the Kullback-Leibler distance. *InterConf*: Proceedings of the 2nd International Scientific and Practical Conference «Science and Education in Progress», Dublin, 16–18 June 2023. 2023. P. 353–354.

[18] MLOps с Python-библиотекой Evidently: обнаружение дрейфа данных в ML-моделях. URL: https://medium.com (дата звернення: 22.08.2022).

[19] Zheng, Shihao, et al. "Labelless concept drift detection and explanation." NeurIPS 2019 Workshop on Robust AI in Financial Services: Data, Fairness, Explainability, Trustworthiness, and Privacy. 2019.

[20] Increasing the reliability of user authentication based on protected electronic key and behavioral biometrics / O. V. Saliieva et al. Visnyk of vinnytsia politechnical institute. 2023. Vol. 167, no. 2. P. 102–111.

[21] M. Sivaram, M. Ahamed, D. Yuvaraj, G. Megala, V. Porkodi, M. Kandasamy, Biometric Security and Performance Metrics: FAR, FER, CER, FRR, in: Proceedings of 2019 International Conference on Computational Intelligence and Knowledge Economy, ICCIKE, IEEE, Dubai, United Arab Emirates, 2019. doi: 10.1109/iccike47802.2019.9004275.

[22] D. Thakkar, False acceptance rate (FAR) and false recognition rate (FRR), 2020. URL: https://www.bayometric.com/false-acceptance-rate-far-false-recognition-rate-frr/

[23] Shih N., Shakleina I. Analysis of biometric user identification methods in controlled access systems. Bulletin of the National University of Water Management and Nature Management", series "Technical Sciences".2016. № 3(75). P. 120–130.

[24] Alshehri A., Coenen F., Bollegala D. Keyboard usage authentication using time series analysis. Big data analytics and knowledge discovery. Cham, 2016. P. 239–252.