УДК 681.3
DOI https://doi.org/10.32838/2663-5941/2022.3/08

Касіяненко Д.В.
Київський національний університет імені Тараса Шевченка

АУТЕНТИФІКАЦІЯ КОРИСТУВАЧІВ НА ОСНОВІ АНАЛІЗУ КЛАВІАТУРНОГО ПОЧЕРКУ

Розглянуто сучасні методи розпізнавання користувачів мережевого сервісу на основі аналізу біометричних характеристик з метою аутентифікації. Зазначено, що проблема впровадження автоматизованих процедур аутентифікації у багатьох випадках показує низьку ефективність у зв'язку з недотриманням користувачами зазначеного регламенту стратегії захисту мережевих ресурсів. Вказано, що підходи, які базуються на аналізі біометричних характеристик користувачів, ефективно обмежують несанкціонований доступ до конфіденційних даних та послуг мережевого сервісу. Зазначено, що клавіатурний почерк є найбільш простим способом визначення біометричних характеристик, що не вимагає застосування додаткових апаратних засобів у рамках робочої платформи користувача. Представлена комплексна методика, що базується на основі аналізу таких параметрів як кількість помилок при наборі тексту, інтервали між натисканням окремих клавіш, час утримання окремих клавіш, наявність перекриття між клавішами, рівень аритмічності набору та швидкість набору тексту. Проведена формалізація надала можливість побудувати математичну модель, що базується на визначенні цільових показників точності аутентифікації користувачів мережевого сервісу та навантаження на обчислювальний ресурс апартно-програмного комплексу мережевого сервісу. Показано, що зазначений підхід може бути використано з одного боку як основна система захисту, а з іншого – на рівні комплексного підходу, у рамках якого підсистема аналізу клавіатурного почерку ефективно доповнює систему аутентифікації на основі введення логіну та паролю. Представлено розширену методику, що надає можливість проведення адаптації системи аутентифікації на основі біометричних параметрів, яка враховує поступову зміну клавіатурного почерку користувача та автоматично збільшує ефективність машинного аналізу на основі великих об'ємів статистичних даних.

Ключові слова: *мережевий сервіс, аутентифікація користувачів, біометричні характеристики, клавіатурний почерк, ключ-шаблон, навчальна вибірка, екстремум цільової функції.*

Вступ. Активний розвиток і поширення мережевих сервісів та, зокрема, впровадження хмарних центрів обробки даних (ЦОД), що спостерігається протягом останніх двох десятиріч призвели до необхідності розробки та оптимізації систем захисту мережевих ресурсів (Data Leakage Prevention Strategy, DLP-стратегія). Вирішення поставленої задачі включає у себе побудову високоефективних систем аутентифікації користувачів, що працюють у режимі реального часу з мінімальним навантаженням на обчислювальний ресурс апаратно-програмної платформи мережевого сервісу. У рамках сучасних підходів впровадження DLP-стратегії зумовлює повну автоматизацію процесу аутентифікації, що дозволяє зменшити об'єм роботи персоналу (особливо при масштабуванні інфраструктури мережевого сервісу) та уникнути помилок, пов'язаних з так званим «людським фактором». Водночас, має бути зазначено, що більшість користувачів мережевих сервісів досить безвідповідально відносяться до захисту персональних даних і у своїй більшості не мають розуміння про ступінь небезпеки та базові процедури по аутентифікації, застосування яких дозволить знизити рівень кіберзагрози. Тому на сьогоднішній день перевага надається біометричним алгоритмам аутентифікації, що надають можливість виключити з процесу аутентифікації активні і усвідомлені дії користувачів та водночас ефективно здійснити процедуру машинного аналізу [1-5].

Аналіз сучасних досліджень і публікацій присвячених проблемам організації автоматизованих систем аутентифікації користувачів мережевого сервісу вказав, що сучасні методи базуються на виділенні таких біометричних характеристик як риси обличчя користувача, міміка, жести та голос [6-10]. Автори досліджень зазначають, що вказана задача є нетривіальною, і тому для її вирішення використовуються як складні програмні так і нейромережеві алгоритми, які дозволяють ефективно виділити типові аудіо- та відео-зразки, але при цьому призводять до значного навантаження на обчислювальний ресурс апаратно-

програмної платформи мережевого сервісу [6, 8-10]. Крім того, виділення відповідних біометричних характеристик зумовлює наявність та налаштування додаткового апаратного обладнання, як то камера і мікрофон достатньо високої якості, взаємне розташування користувача і камери, а також користувача і мікрофона, система освітлення, тощо. Слід зазначити, що передача аудіо- та відео-даних з метою аутентифікації користувача додатково призведе до навантаження на мережевий ресурс сервісу, що у окремих випадках може бути критичним за умов необхідності у режимі реального часу. Зазначені особливості побудови автоматичної системи аутентифікації на основі біометричних параметрів вказують на переваги застосування алгоритмів машинного аналізу клавіатурного почерку (Keystroke Dynamics, KsD) користувача мережевого сервісу [11-14]. Клавіатурний почерк користувача може бути повністю описано через мінімальний набір числових параметрів, що не потребує для автоматичної обробки складних алгоритмів. При цьому підсистема аналізу клавіатурного почерку може доповнювати базову систему аутентифікації, що базується на введенні логіну та паролю. Наведені дослідження включають визначення кількості помилок при наборі тексту, інтервалів між натисканням окремих клавіш, час утримання окремих клавіш, наявність перекриття між клавішами, рівень аритмічності набору та швидкість набору тексту [12-14]. Оцінка ефективності окремих підходів, тим не менш, не дає можливість визначити цілісну методологію побудови алгоритмів машинного аналізу клавіатурного почерку, що розглядається як ***невирішена частина загального дослідження.***

Метою дослідження, таким чином, стала розробка комплексної методології побудови алгоритмів машинного аналізу клавіатурного почерку у рамках аутентифікації користувачів мережевого сервісу. Зазначений підхід включає у себе методику адаптації системи аутентифікації на основі біометричних параметрів, що враховує поступову зміну клавіатурного почерку користувача та автоматично збільшує ефективність машинного аналізу відповідно до статистичних даних, отриманих у процесі користування окремими послугами мережевого сервісу.

Виклад основного матеріалу дослідження. Формалізуємо процес машинного аналізу текстового набору системою аутентифікації мережевого сервісу. Нехай процес набору включає у себе $n \in [1; N]$ символів (з пропусками, включно). Позначимо час утримання окремої клавіші як t_n, а проміжок між натисканням клавіш n та $(n+1)$, як t^n_{n+1}, причому, якщо загальний час набору блоку N символів складає T, то зазначений показник може бути розраховано як $T = \sum_{n=1}^{N} t_n + \sum_{n=1}^{N-1} t^n_{n+1}$. Це дозволяє навести наступні біометричні характеристики, на основі яких може бути визначено клавіатурний почерк користувача (рис. 1):

- середній час утримання окремої клавіші $\underline{t_n} = \frac{\sum_{n=1}^{N} t_n}{N}$;
- аритмія утримання клавіші, як максимальна різниця у часі утримання окремої клавіші $\Delta t_n = t_n - t_n$;
- середній проміжок між натисканням клавіш $\underline{t^n_{n+1}} = \frac{\sum_{n=1}^{N-1} t^n_{n+1}}{(N-1)}$;
- аритмія набору символів, як максимальна різниця у проміжках між натисканням клавіш $\Delta t^n_{n+1} = t^n_{n+1} - t^n_{n+1}$;
- середня швидкість набору $V = \frac{N}{T}$;
- загальна кількість помилок у наборі N_E ;
- загальна кількість перекриттів між клавішами (Overlap Between Keys, OBK) як N_{OBK} .

Розширена модель у рамках аналізу відповідних параметрів набору $\{t_n, t^n_{n+1}\}$ може включати визначення залежностей для окремих клавіш або, радше, груп клавіш, що разом зі збільшенням точності процедури аутентифікації надає можливість зменшити навантаження на обчислювальний ресурс. У даному дослідженні пропонується виділити такі групи клавіш:

- група клавіш, що відповідають літерам, яка може бути розбита на дві рівні підгрупи: підгрупу клавіш у центрі клавіатури $\{S^0_i\}$ та підгрупу клавіш у периферійній області – $\{S^+_i\}$, де $i \in [1; I]$;
- група клавіш, що відповідають цифрам $\{S^N_j\}$, де $j \in [0; 9]$;
- група клавіш, що відповідають знакам пунктуації $\{S^P_m\}$, де $m \in [1; M]$;
- група клавіш $\{S^A_k\}$, де $k \in [1; K]$, що відповідають додатковим елементам, як то «Shift», «Ctrl», «Alt», для яких особливо важливо враховувати перекриття клавіш.

Нарешті, необхідно розглянути задачу адаптації розробленої моделі для вирішення задачі побудови комплексного алгоритму аутентифікації, що базується на введенні пари логін-пароль та доповнюється аналізом біометричних характеристик користувача мережевого сервісу. Застосування моделі аутентифікації користувача на основі клавіатурного почерку у рамках даної задачі має ряд переваг:

1. Дві послідовності логін і пароль є достатньо короткими і визначеними заздалегідь. Це надає можливість не поділяти символьний ряд на окремі групи

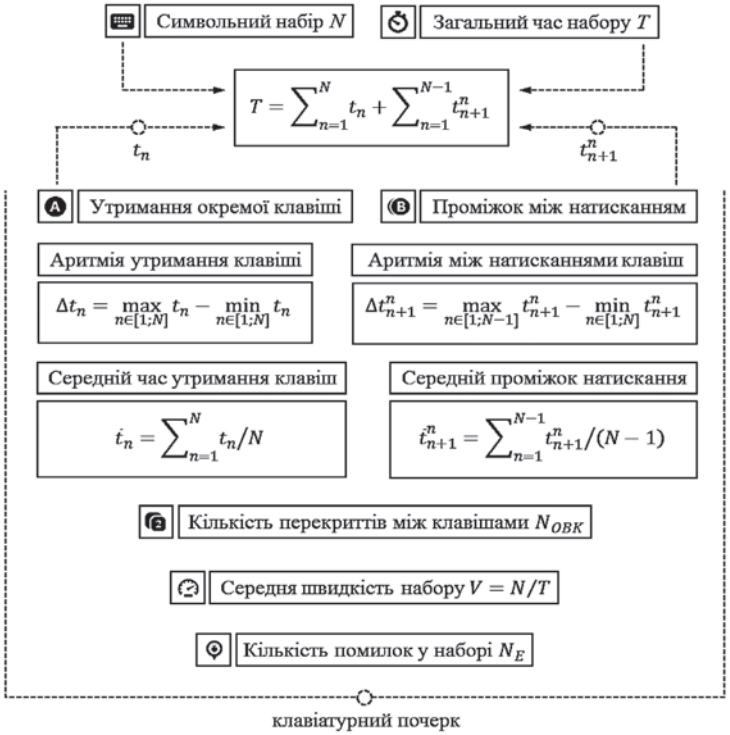

Рис. 1. Базова модель виділення набору біометричних параметрів, що відповідають клавіатурному почерку користувача мережевого сервісу

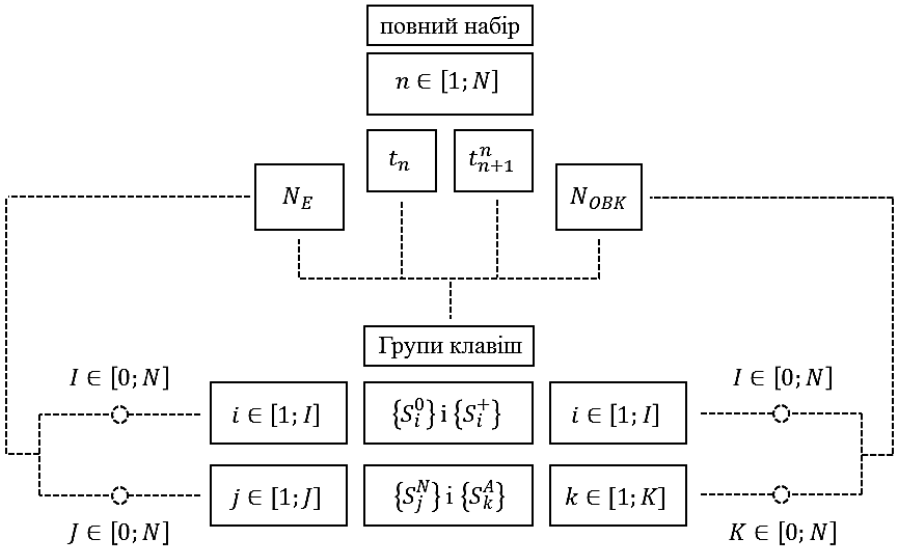

Рис. 2. Розширена модель машинного аналізу клавіатурного почерку користувача відповідно груп клавіш

відповідно їх розташування на клавіатурі, а аналізувати кожен символ окремо відповідно його положенню у текстових послідовностях пари логін-пароль.

2. Набір коротких послідовностей логін-пароль у рамках навчання користувача сервісу через повторення, а також у процесі безпосереднього користування послугами сервісу відбувається несвідомо, що зумовлює стабільність відповідних біометричних параметрів, що відслідковуються системою аутентифікації.

3. Поєднання двох максимально простих систем аутентифікації надає можливість суттєво

збільшити рівень захисту без застосування додаткових апаратних засобів та при мінімальному навантаженні на обчислювальний ресурс апаратно-програмної платформи. Слід також зауважити, що з точки зору користувача процес навчання включає лише серію повторень при введенні логіна і пароля, що дозволить надійно запам'ятати дані аутентифікації облікового запису.

Формалізуємо на математичному рівні процес аутентифікації користувача мережевого сервісу на основі пари логін-пароль, введення якої підлягає машинному аналізу відповідно клавіатурному почерку. Нехай логін складається за $n \in [1; N_{LOG}]$ символів, а пароль – з $n \in [1; N_{PAS}]$, причому після створення зазначеної пари користувач мережевого сервісу має ввести його $r \in [1; R]$ разів. Повторення введення пари логін-пароль надає можливість отримати R груп значень, що відображають кількість похибок при наборі $\{N^E_{LOG}(r)\}$ і $\{N^E_{PAS}(r)\}$, відповідно, затримок при натисканні окремої клавіші $\{t^n_{LOG}(r)\}$ і $\{t^n_{PAS}(r)\}$, відповідно, а також проміжки часу між натисканням клавіш $\{t^{n|n+1}_{LOG}(r)\}$ і $\{t^{n|n+1}_{PAS}(r)\}$.

Статистичний аналіз для отримання на основі відповідного масиву зразку клавіатурного почерку включає у себе визначення наступних показників:

- середні значення відповідно рівня помилок при введенні логіну та паролю, що розраховуються як $\underline{N}^E_{LOG} = \sum_{r=1}^{R}\left(\frac{N^E_{LOG}(r)}{R}\right)$ і $\underline{N}^E_{PAS} = \sum_{r=1}^{R}\left(\frac{N^E_{PAS}(r)}{R}\right)$, відповідно;
- середні значення часу утримання окремої клавіші при введенні логіну та паролю як $\underline{N}^E_{LOG} = \sum_{r=1}^{R}\left(\frac{N^E_{LOG}(r)}{R}\right)$ і $\underline{N}^E_{PAS} = \sum_{r=1}^{R}\left(\frac{N^E_{PAS}(r)}{R}\right)$, відповідно;
- середні значення проміжку між натисканням клавіш символьних послідовностей логіну та паролю як $\underline{t}^{n|n+1}_{LOG} = \sum_{r=1}^{R}\left(\frac{t^{n|n+1}_{LOG}(r)}{R}\right)$ і $\underline{t}^{n|n+1}_{PAS} = \sum_{r=1}^{R}\left(\frac{t^{n|n+1}_{PAS}(r)}{R}\right)$, відповідно;
- дисперсія рівня помилок, квадрат якої розраховується як $(\sigma^E_{LOG})^2 = \sum_{r=1}^{R}\left(\frac{\left|\underline{N}^E_{LOG} - N^E_{LOG}(r)\right|}{R}\right)$ і $(\sigma^E_{PAS})^2 = \sum_{r=1}^{R}\left(\frac{\left|\underline{N}^E_{PAS} - N^E_{PAS}(r)\right|}{R}\right)$, для логіну та паролю, відповідно;
- дисперсія часу утримання окремої клавіші, квадрат яких розраховується як $(\sigma^T_{LOG})^2 = \sum_{r=1}^{R}\left(\frac{\left|\underline{t}^n_{LOG} - t^n_{LOG}(r)\right|}{R}\right)$ і $(\sigma^T_{PAS})^2 = \sum_{r=1}^{R}\left(\frac{\left|\underline{t}^n_{PAS} - t^n_{PAS}(r)\right|}{R}\right)$ для символьних послідовностей логіну та паролю, відповідно;
- дисперсія у значенні проміжку між натисканням клавіш символьних послідовностей, квадрат яких $(\sigma^{TT}_{LOG})^2 = \sum_{r=1}^{R}\left(\frac{\left|\underline{t}^{n|n+1}_{LOG} - t^{n|n+1}_{LOG}(r)\right|}{R}\right)$ і $(\sigma^{TT}_{PAS})^2 = \sum_{r=1}^{R}\left(\frac{\left|\underline{t}^{n|n+1}_{PAS} - t^{n|n+1}_{PAS}(r)\right|}{R}\right)$ для логіну та паролю, відповідно.

На основі наборів середніх значень та значень дисперсії через введення математичного очікування E, t-критерію Стьюдента, та значення вірогідності наявності помилок першого роду P_{EI}, а також визначення загальної кількості зразків навчальної вибірки $l \in [1; L]$ у рамках розробленої математичної моделі процесу аутентифікації можуть бути розраховані інтервали допустимих значень цільових показників, що відповідають клавіатурному почерку, як це показано на рис. 3:

- інтервали для допустимої кількості помилок $N^E_{LOG}(r) \in \left[N^{E\downarrow}_{LOG}; N^{E\uparrow}_{LOG}\right]$ і $N^E_{PAS}(r) \in \left[N^{E\downarrow}_{PAS}; N^{E\uparrow}_{PAS}\right]$ для символьних послідовностей логіну та паролю, відповідно;
- інтервали для допустимого інтервалу часу утримання окремої клавіші $t^n_{LOG}(r) \in \left[t^{T\downarrow}_{LOG}; t^{T\uparrow}_{LOG}\right]$ і $t^n_{PAS}(r) \in \left[t^{T\downarrow}_{PAS}; t^{T\uparrow}_{PAS}\right]$ для символьних послідовностей логіну та паролю, відповідно;
- інтервали для допустимого проміжку між натисканням клавіш $t^{n|n+1}_{LOG}(r) \in \left[t^{TT\downarrow}_{LOG}; t^{TT\uparrow}_{LOG}\right]$ і $t^{n|n+1}_{PAS}(r) \in \left[t^{TT\downarrow}_{PAS}; t^{TT\uparrow}_{PAS}\right]$ для символьних послідовностей логіну та паролю, відповідно.

Зазначені інтервали дозволяють на кількісному рівні через статистичний аналіз однозначно визначити відповідність клавіатурного почерку окремому користувачу. Для підтвердження відповідності біометричних характеристик має бути виконано повний набір умов, вказаних вище.

Оптимізація алгоритму відбувається, з одного боку через варіювання показників E, t-критерію Стьюдента, L та P_{EI}, а з іншого шляхом розширення роботи з користувачем мережевого сервісу. У першому випадку оптимізація відбувається через пошук глобального екстремуму цільової функції точності машинного аналізу на основі методу градієнтного спуску або аналогічних методів. Другий підхід зумовлює необхідність повторення проходження користувачем процедури повторного введення пари логін-пароль M разів поспіль. Це допомагає як додатково уточнити допустимі інтервали, що визначають клавіатурний почерк, так і відслідкувати зміни у клавіатурному почерку, що може відбуватись протягом тривалого часу.

Висновки. В результаті проведеного дослідження було розглянуто методи аутентифікації користувачів мережевого сервісу на основі клавіатурного почерку. Була представлена методика, що базується на статистичному аналізі таких параметрів як кількість помилок при наборі тексту, інтервали між натисканням окремих клавіш, час утримання окремих клавіш, наявність перекриття між клавішами, рівень аритмічності набору та швидкість набору тексту.

Таким чином, було побудовано:

- базову модель виділення набору біометричних параметрів, що відповідають клавіатурному почерку користувача мережевого сервісу;

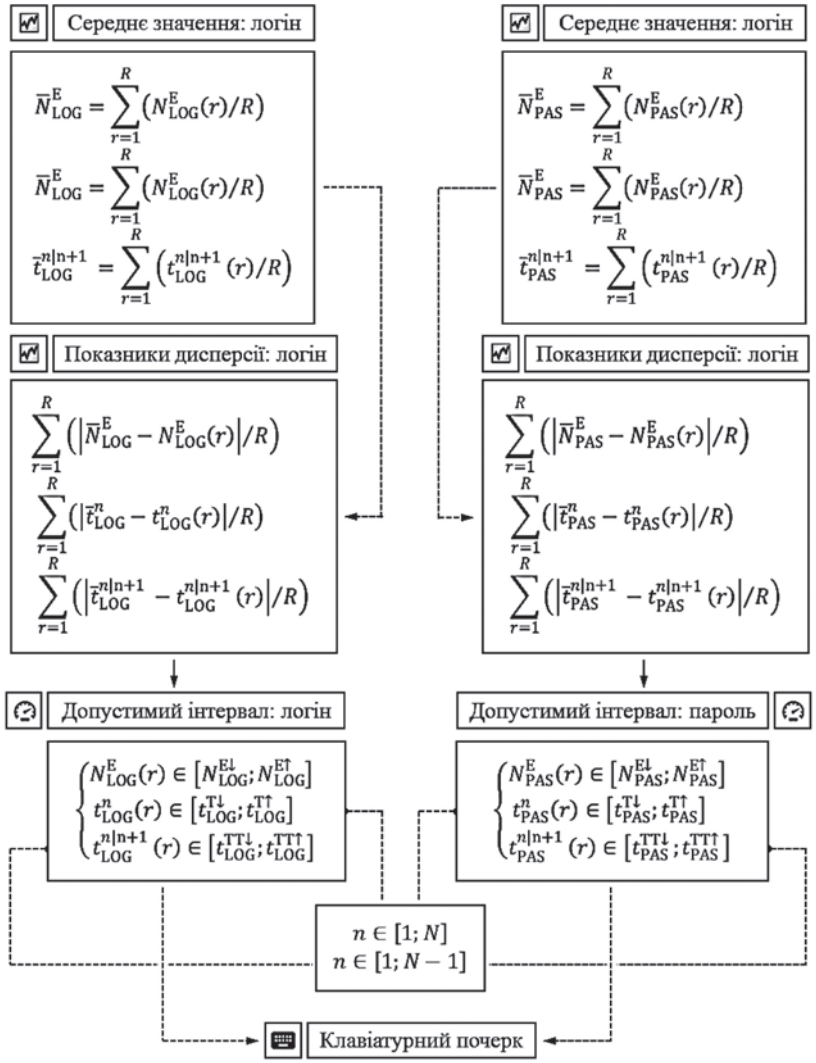

Рис. 3. Методика аутентифікації, що базується на введенні пари логін і пароль та машинному аналізі клавіатурного почерку користувача

- розширену модель машинного аналізу клавіатурного почерку користувача мережевого сервісу відповідно окремих груп клавіш;
- методику аутентифікації, що базується на введенні пари логін і пароль та додатковому машинному аналізі клавіатурного почерку користувача мережевого сервісу.

Представлена методика надає можливість суттєво підвищити рівень ефективності аутентифікації користувачів мережевого сервісу через аналіз зразків клавіатурного почерку. На основі побудованої моделі проводиться аналіз зміни клавіатурного почерку користувача та автоматично збільшується ефективність машинного аналізу завдяки роботи з великими об'ємами статистичних даних.

Список літератури:

1. Zhu, H.H. et al., Voiceprint-biometric template design and authentication based on Cloud Computing Security. *2011 International Conference on Cloud and Service Computing*. 2021. p. 34-49.

2. Gawade S. et al., Biometric authentication using software as a service in cloud computing. *International Journal Of Engineering And Computer Science*. 2017. p. 18-22.

3. Bharti A., Raj A. Sensory launches cloud-based voice and facial ID service. *Biometric Technology Today*. № 2022 (1) p. 2205-2211.

4. Lu Y., Zhao D. Providing impersonation resistance for biometric-based authentication scheme in Mobile Cloud Computing Service. *Computer Communications*. 2022. № 182, p. 22-30.

5. Hu H. et al., Toward a biometric-aware cloud service engine for multi-screen video applications. *Proceedings of the 2014 ACM Conference on SIGCOMM*. 2014. p. 11-17.

6. Zhang Z., Gong C., Liu R. Face Detection Based on Method Combined RVM and SVM. *Computer Science and Artificial Intelligence*. 2017. p. C20-C24.

7. Tsai C., Lee Y. The parameters effect on performance in ANN for hand gesture recognition system. *Expert Systems with Applications*. 2011, №38 (7). p. 7980-7983.

8. Ganakwar D.G., Kadam, V.K. Face Detection Using Boosted Cascade of Simple Feature. *2019 International Conference on Recent Advances in Energy-efficient Computing and Communication (ICRAECC)*. 2017. p. A35-A39.

9. Viola P., Jones M.J. Robust real-time face detection. *International Journal of Computer Vision*, № 57 (2). 2014. p.137-154.

10. Alorf A.A. Performance evaluation of the PCA versus improved PCA (IPCA) in image compression, and in face detection and recognition. *2016 Future Technologies Conference (FTC)*. 2016. p. 29-37.

11. Cascone L., et al. Touch keystroke dynamics for demographic classification. *Pattern Recognition Letters.* 2022. p. 19-26.

12. Alsultan A., Warwick K., Wei H. Non-conventional Keystroke Dynamics for user authentication. *Pattern Recognition Letters*, №89. 2017. p. 53–59.

13. Tsai C.J., Huang P.H. Keyword-based approach for recognizing fraudulent messages by keystroke dynamics. *Pattern Recognition*. 2020, № 98 (10). p. 67-70.

14. Kim J., Kang P. Freely typed keystroke dynamics-based user authentication for mobile devices based on heterogeneous features. Pattern Recognition. 2020, № 108. p. 107556-107564.

Kasiianenko D.V. USER AUTHENTICATION BASED ON KEYSTROKE DYNAMICS ANALYSIS

Modern methods of recognizing network service users based on the analysis of biometric characteristics for authentication are considered. It is noted that the problem of implementing automated authentication procedures in many cases shows low efficiency due to non-compliance with the regulations of the network protection strategy. It is stated that approaches based on the analysis of user biometric characteristics effectively limit unauthorized access to confidential data and network service. It is noted that keystroke dynamics analysis is the easiest way to determine biometric characteristics, which does not require the use of additional hardware within the user's work platform. A comprehensive technique based on the analysis of such parameters as the number of typing errors, intervals between pressing individual keys, time to hold individual keys, the presence of overlap between keys, the level of typing arrhythmia and typing speed. The formalization provided an opportunity to build a mathematical model based on the definition of targets for the accuracy of authentication of network service users and the load on the computing resource of the hardware and software complex of the network service. It is shown that this approach can be used on the one hand as the main security system, and on the other – at the level of integrated approach, in which the keyboard handwriting analysis subsystem effectively complements the authentication system based on login and password. An advanced technique is presented, which allows the adaptation of the authentication system based on biometric parameters, which takes into account the gradual change of the user's keyboard handwriting and automatically increases the efficiency of machine analysis based on large amounts of statistics.

Key words: *network service, user authentication, biometric characteristics, keystroke dynamics, key-template, training set, extremum of the target function.*