Method of User Authentication by Keyboard Handwriting based on Neural Networks and Genetic Algorithm

Andrii Pryimak[a], Yurii Yaremchuk[a], Olha Salieva[a], Vasyl Karpinets[a] and Nataliia Kunanets[b]

[a] *Vinnytsia National Technical University, Khmelnytsky highway 95, Vinnytsia, 21000, Ukraine*
[b] *Lviv Polytechnic National University, 12 Bandera street, Lviv, 79013, Ukraine*

Abstract

A method of user authentication based on keyboard handwriting with error injection was proposed. It is based on a two-level neural network architecture using five-time functions and built-in sigmoid activation function to increase the efficiency of the neural network. An error code injection was also introduced, which allowed to collect more accurate data on human handwriting and increase the accuracy of correct recognition of the user and his successful authentication by 3-11% compared to existing methods. The use of a hash function based on a genetic algorithm is proposed, which provides the security of storing a code word in the database.

Keywords 1

Information security, user authentication, neural network, keyboard handwriting, genetic algorithm.

1. Introduction

Given the rapid pace of development of information technology, increasing the number of information threats, the degree of uncertainty of their origin and implementation, as well as the complexity of information security systems and their specialized focus, the task of building an information security system becomes relevant. One of the methods of information protection is user authentication. User authentication is the verification that the user being authentication is who he or she claims to be [1]. For correct user authentication, it is necessary for the user to present some unique information, which should be owned only by him and no one else.

Traditional methods of identification and authentication, based on the use of smart cards, USB keys, electronic keys or other portable identifiers, as well as passwords and access codes, have significant disadvantages, such as: the possibility of stealing the item from the user; the need for special equipment for working with magnetic cards, smart cards and others; the ability to make a copy of a unique item. In general, the main disadvantage of such methods is not always reliable authentication [2]. This shortcoming can be eliminated by using biometric authentication methods, such as the dynamics of keystrokes by the user. Biometric characteristics are an integral part of human beings and therefore cannot be falsified, lost or forgotten.

Keystroke dynamics, which represent the typing rhythms that the user performs while typing on the keyboard, provide a high level of security and also have advantages in practical application, as inexpensive implementation of this method is an important indicator compared to scanning fingerprints or irises eyelids that require additional equipment to achieve authentication [3].

Another modern approach to solving the problem of authentication is the use of neural networks. There are a number of architectures that have already become classic - maximum search network, input

International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2021), September 20–21, 2021, Kharkiv, Ukraine
EMAIL: andrii.pryimak@live.com (A. Pryimak); yurevyar@gmail.com (Y. Yaremchuk); salieva8257@gmail.com (O. Salieva); karpinets@gmail.com (V. Karpinets); nek.lviv@gmail.com (N. Kunanets)
ORCID: 0000-0001-9695-0462 (A. Pryimak); 0000-0002-6303-7703 (Y. Yaremchuk); 0000-0003-2388-7321 (O. Salieva); 0000-0001-8148-2002 (V. Karpinets); 0000-0003-3007-2462 (N. Kunanets)

and output star, single-layer perspective, BSP, network with radial basis function (RBF), Hopfield network, Hemming, Cosco, McCulloch-Pitts, Kohonen, Grosnign, and Ymovir APT network. Thus for each class of applied problems the architecture of a neural network is used. Applying a neural network approach to the authentication problem can improve the accuracy of user authentication, because this approach has the property of filtering random interference present in the input data, which allows to abandon the algorithms for smoothing experimental dependencies required for statistical data processing [4].

2. Related works

Methods of user authentication by keyboard handwriting have been practiced relatively recently. In their work Leggett, Umphress and Williams [5] conducted an experiment with 17 programmers. They used measured keystrokes, known as digraphs. The first set consisted of 1400 characters and was used at the learning stage, the second set consisted of 300 characters and was used for verification. In his report, Leggett specified an authentication probability of 89.5%. In their experiments, the authors suggested a possible deviation of the mean retention time of bigram equal to 0.5, the user was considered recognizable if 60% or more of the time delay coincided with the allowable deviation of the sample.

Other work in this area was carried out by Garris, Young and Hammon [6]. In their proposed approach, a matrix of changes in the associated delay vectors was used as a quantity (parameter) that contains data on individual handwriting. The Mahalanobis distance function was then used to determine the similarity between the handwriting identified and the user profile. Unlike others, Young and Hammon used the Euclidean distance between two vectors to compare the number of attributes.

Another well-known work was presented by Rastorguev [7]. In his monograph, the author divided the authentication procedure into two types. The first type is password authentication, where the user goes through the authentication procedure by password, the second type is the authentication of users by a set of random phrases. He also singled out two modes of the authentication procedure: the system setup procedure and the authentication procedure.

When identifying users by typing free text in the mode of setting up the authentication system, the keyboard was divided into four parts. When the user worked on the keyboard, the time intervals between these four parts were calculated, regardless of which key was pressed in these parts. In the authentication mode, the current values were compared with the reference and the system made decisions. The main assumption was that the distribution of temporal characteristics of users seemed to be a normal Gaussian law. Algorithms for excluding gross errors, system settings and authentication were also presented in the paper.

Another paper was presented by Saket Maheshwari and Vikram Pudi [8], who propose a method for identifying a user using keyboard handwriting based on a five-level neural network. The authors investigated in detail the possibility of using three different methods of building a neural network architecture (exclusion method, rectification method, packet normalization method). Studies have shown that the maximum accuracy of user identification is 85.22-93.59%. However, it should be noted that the best accuracy result was achieved by using a five-level neural network, which significantly increases the learning time of this network (up to 9 minutes) per user.

Nura Hanura's [9] work focuses on using the time interval between keystrokes as a feature of a set of individual characters to identify authentic users. A four-level neural network with a multilinear perceptron (MLP) with a built-in error propagation method (BP) is used to train and test functions. The results of this study showed that the accuracy of user identification is in the range of 90-92%.

In the previous work of the authors "Method of user identification by keyboard handwriting based on neural networks" [10] a study of the possibility of using a neural network, which is a two-layer system of direct access to a network with 70 sigmoid hidden neurons and 10 sigmoid source neurons. The obtained indicators of user identification accuracy using the proposed method are 88-93%.

Most of the considered works are based on geometrical methods of recognition, using various degree of closeness between the handwriting sample and its standard (Euclidean, Mahalanobis, etc.). The maximum found probability of authentication of such systems is 90%. Methods based on the use of

neural networks have higher accuracy (85.22-93.59%), but it should be noted that the research was conducted using multilevel neural networks, which significantly affected the speed of their learning.

The results of comparing the accuracy of user authentication by existing methods are presented in the Table 1 below.

Table 1
Comparative characteristics of existing methods

Method name	Accuracy, %
The method of Leggett, Umphress and Williams	89,5
The method of Rastorguev	90
The method of Maheshwari and Pudi	85,22–93,59
The method of Hanura	90–92
The method proposed in the previous work of the authors	88–93

Based on the analysis of existing methods of user identification by keyboard handwriting, it is seen that the accuracy of identification is in the range from 85.22% to 93.59%, so it remains important to increase the accuracy of identification and development of the appropriate method.

Therefore, it makes sense to pay attention to the injection of an error in the word to verify the user, which can allow to collect more accurate data on human handwriting and reduce the risk of forgery. It can also be borne in mind that a person's keyboard handwriting, like normal handwriting, can change, leading to incorrect validation methods, and the user's reaction to the error remains relatively constant, regardless of changes in speed or correct character set by the user. To increase the security of codeword storage, it can be considered using a genetic approach to generate a hash function of the word.

3. Problem Statement

To study the possibility of using a neural network and a hash function based on a genetic algorithm to improve the accuracy of user authentication based on keyboard handwriting with error injection, as well as to propose a method based on this mathematical apparatus. Also compare the proposed method of authentication with existing ones.

4. Proposed method

It is proposed to use the injection in the form of error generation to collect the necessary information on the user's keyboard handwriting to verify the correctness of the entered word by the user, as well as to use the hash function based on a genetic algorithm to increase the security of code word storage. The neural network architecture, which is a two-tier system of direct access to a network with 70 sigmoid hidden neurons and 10 sigmoid output neurons, using the sigmoid activation function, was chosen to study keyboard writing directly.

The proposed method of user authentication (Figure 1) consists of the following steps:

1. Entering a code word by the user (code word pre-stored in the database as a hash with a use of genetic algorithm - H_1).
2. Generate a random word character number to be changed.
3. Generate a random number of the word symbol to which the symbol from the previous step will be changed.
4. Replacing two characters with places - creating an error in the code word.
5. The user makes corrections in the word. Based on its correction, a hash is generated with a use of genetic algorithm - H_2.
6. Comparison of two hashes with each other. If they are the same, then the user has made a fix and the method continues to work. If the hashes do not match, the user returns to step 5.
7. User identification based on a trained neural network.
8. If all data converges, the user successfully passed authentication process, if not, he returns to step 1.

This method uses a hash function, the feature of which is that it is based on a genetic algorithm, the main operators of which are crossover on the worst gene, mutation of the two worst genes and fitness function, which uses analysis based on five statistical tests (monobit test, poker test, start test, longruns test and autocorrelation test). Using this approach to generating a hash function allows you to increase the security of code word storage.

The stage of user authentication based on a trained neural network includes nine main stages of collecting information using time functions and its further processing. The main stages are: collection of all necessary data; data preparation and normalization; operation of synchronization functions; basic component analysis; automatic selection of learning parameters; network training; checking the correctness of training; adjustment of parameters; readiness for further use.

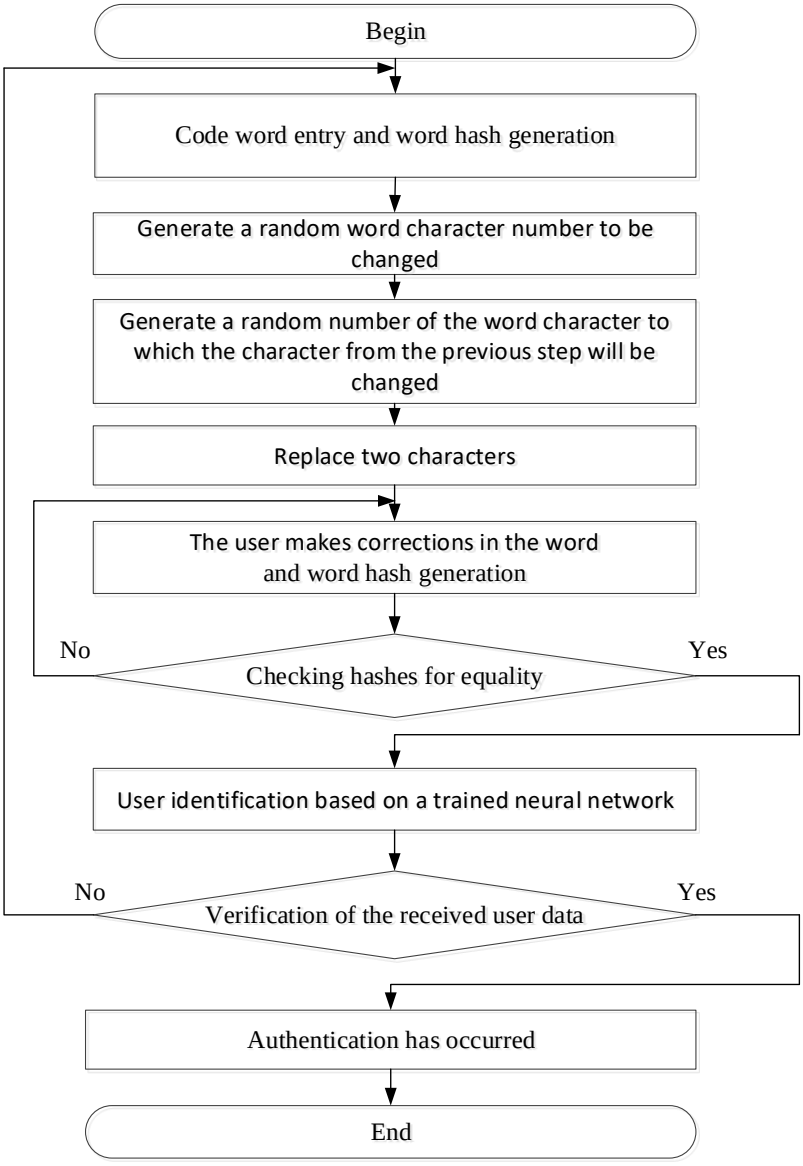

Figure 1: Flowchart of the proposed method

To collect information, the proposed method contains five-time functions (delay time, up-down delay, down-down delay, up-up delay, total time) that collect the information needed for comparison and identification of the user by the neural network. These are the five-time functions:

1. Delay time - the time of keystroke, which is determined by the following equation

$$KeyDuration = R_i - P_i, \quad (1)$$

where R_i- release time of the i-th key; P_i is the time of pressing the i-th key.

2. «Up-down» delay - the time difference between releasing a key and pressing the next key

$$UpDownLate = P_{i+1} - R_i, \quad (2)$$

where R_i- release time of the i-th key; P_i is the time of pressing the i-th key.

3. «Down-down» delay - the time difference between releasing the same key twice

$$DownDownaLatency = P_{i+1} - P_i, \quad (3)$$

where P_i is the time of pressing the i-th key.

4. «Up-up» delay is the time difference between pressing the same key twice

$$UpUpLatency = R_{i+1} - R_i, \quad (4)$$

where R_i- release time of the i-th key.

5. Total time - the time required to enter all the text

$$TotalTyping = R_{i=N} - P_{i=1}, \quad (5)$$

where R_i- release time of the i-th key; P_i is the time of pressing the i-th key; N is the number of characters in the text.

It is proposed to use a neural network for further processing of the collected information.

To solve the problem of authentication, it is necessary to calculate the expressions for the given parameters. Mathematically, a neuron is a weighted adder, the only output of which is determined through its inputs and the weight matrix so that

$$y = f(u),\ u = \sum_{i=1}^{2} w_i x_i + w_0 x_0, \quad (6)$$

where $f(u)$ - activation function; u - induced local field; w_i - entrance weight; x_i - signal at the input of the neuron; w_0 - additional entrance; x_0 - the weight corresponding to it.

Let the number of input parameters be two. To begin with, it is necessary to investigate one hidden layer of neurons. The number of elements on it should be determined using the Arnold-Kolmogorov-Hecht Nelson formula

$$\frac{N_y Q}{1+\log_2(Q)} \le N_w \le N_y\left(\frac{Q}{N_x}+1\right)\left(N_x+N_y+1\right)+N_y \quad (7)$$

where N_y is the dimension of the output signal; Q is the number of elements of the set of educational examples; N_w - the required number of synaptic connections; N_x - the dimension of the input signal.

After performing the calculations, we can conclude that the required number of synaptic connections is in the range of $7 < N_w < 20$. To find out the number of required neurons in the hidden layer, you must use the formula

$$N = \frac{N_w}{N_x + N_y}. \quad (8)$$

Thus, the number of neurons in the hidden layer will be in the range of $1 < N < 70$. To study the entire range, you need to select the number of neurons in the hidden layer at which the learning error will be less. In this case, it would be advisable to select 70 neurons on the hidden layer.

You also need to select activation functions for each layer. Neurons of the input and output layers are responsible only for data input and output, their functions can be left linear. The main calculated load falls on the neurons of the hidden layer, so its activation function should be made sigmoid.

In direct propagation neural networks, synaptic connections are organized in such a way that each neuron in a given level of the hierarchy receives information only from some non-empty set of neurons that are located at a lower level. The name of the networks indicates that they have a dedicated direction of propagation of signals that move from the input through one or more hidden layers to the output layer. It is easy to see that a multilayer neural network can be obtained by cascading single-layer networks with weights matrices

$$W^1, W^2, \dots, W^p, \quad (9)$$

where p is the number of layers of the neural network.

In the case of linearity of activation functions, a multilayer neural network can be reduced to an equivalent single layer with a matrix of weights $W = W^1 * W^2 * \dots * W^p$, so the formation of such structures makes sense only if nonlinear activation functions are used in neurons.

A neural network is proposed and presented in Figure 2, which is a two-layer system of direct access to a network with 70 sigmoid hidden neurons and 10 sigmoid output neurons.

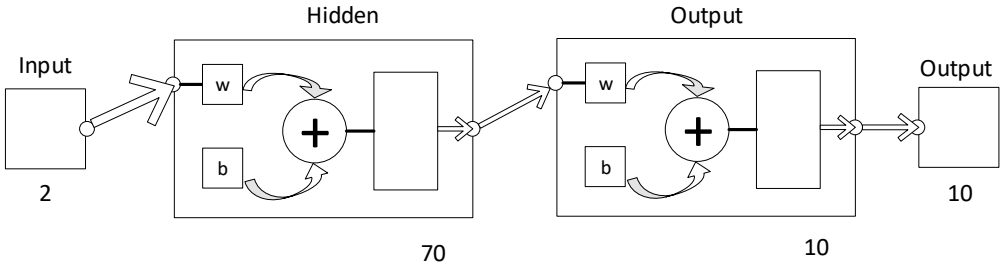

Figure 2: The offered neural network architecture

A detailed diagram of the proposed neural network is presented in Figure 3.
The input for the network will be the key hold time and the time intervals between keystrokes.
Training is carried out as follows:

1. All weights of the network are randomized to small values.
2. The input training vector X is fed to the network input and the NET signal from each neuron is calculated using the standard expression
$$NET_j = \sum_i xw. \quad (10)$$
3. The value of the activation threshold function for the NET signal from each neuron is calculated.
4. The error for each neuron is calculated by subtracting the output from the desired output
$$error_j = target_j - OUT_j. \quad (11)$$
5. Each weight is modified as follows
$$W_{ij}(t+1) = w_{ij}(t) + a_x error_j. \quad (12)$$
6. Repeat steps two through five until the error is small enough.

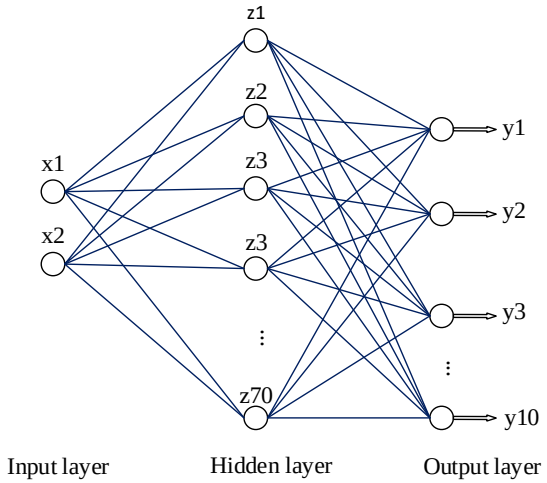

Figure 3: Scheme of the proposed neural network architecture

The results of the accuracy of the user recognition as well as comparison with existing methods are presented in the next chapter of this work.

5. Results and discussion

Let us estimate the probability p of correctly recognizing the user by his frequency in n independent experiments. With the help of the developed software, we will conduct an experiment. To do this, 5 users consistently entered their code word 100 times. The number of access denials is indicated and shown in Table 2.

Table 2
Data on the number of denials of access for a real user

User №	Number of false authentications	Frequency of denied access	Frequency of confirmed access
1	6	0.06	0.94
2	3	0.03	0.97
3	3	0.03	0.97
4	7	0.07	0.93
5	5	0.05	0.95

The average value of the correct frequency of user recognition in a series of 100 experiments is 0.96. To check the applicability of the normal distribution law, the values of np and nq are estimated. Assuming that $p \approx p^*$ we obtain:

$$\begin{aligned} np &\approx np^* = 96, \\ nq &\approx n(1-p^*) = 4, \end{aligned} \tag{13}$$

where p^* - average access frequency.

The obtained values give grounds to believe that in this case the normal distribution law can be applied. According to the tables, we find $t_\beta = 1.652$ for $\beta = 0{,}9$. Next, calculate p_1 and p_2 by the following formulas:

$$p_1 = \frac{p^* + \frac{1}{2}\frac{t_\beta^2}{n} - t_\beta\sqrt{\frac{p^*(1-p^*)}{n} + \frac{1}{4}\frac{t_\beta^2}{n^2}}}{1 + \frac{t_\beta^2}{n}}, \tag{14}$$

$$p_1 = \frac{0{,}96 + 0{,}0135 - 0{,}0366}{1{,}027} = 0{,}92,$$

$$p_2 = \frac{p^* + \frac{1}{2}\frac{t_\beta^2}{n} + t_\beta\sqrt{\frac{p^*(1-p^*)}{n} + \frac{1}{4}\frac{t_\beta^2}{n^2}}}{1 + \frac{t_\beta^2}{n}}, \tag{15}$$

$$p_2 = \frac{0{,}96 + 0{,}0135 + 0{,}0366}{1{,}027} = 0{,}98.$$

Thus, the probability of correct user recognition is in the range from 92% to 98%.

A comparison of the proposed method with the existing methods (Figure 4) showed that the proposed method has better accuracy by 2.5-8.5% than the accuracy of the method of Leggett, Umphress and Williams, by 2-8% better accuracy than the method of Rastorguev, by 4.41-6.78% better than the method of Maheshwari and Pudi and 2-6% better than the method of Hanura. The performance of the recognition accuracy of the method proposed in the previous work was also improved by 4-5%.

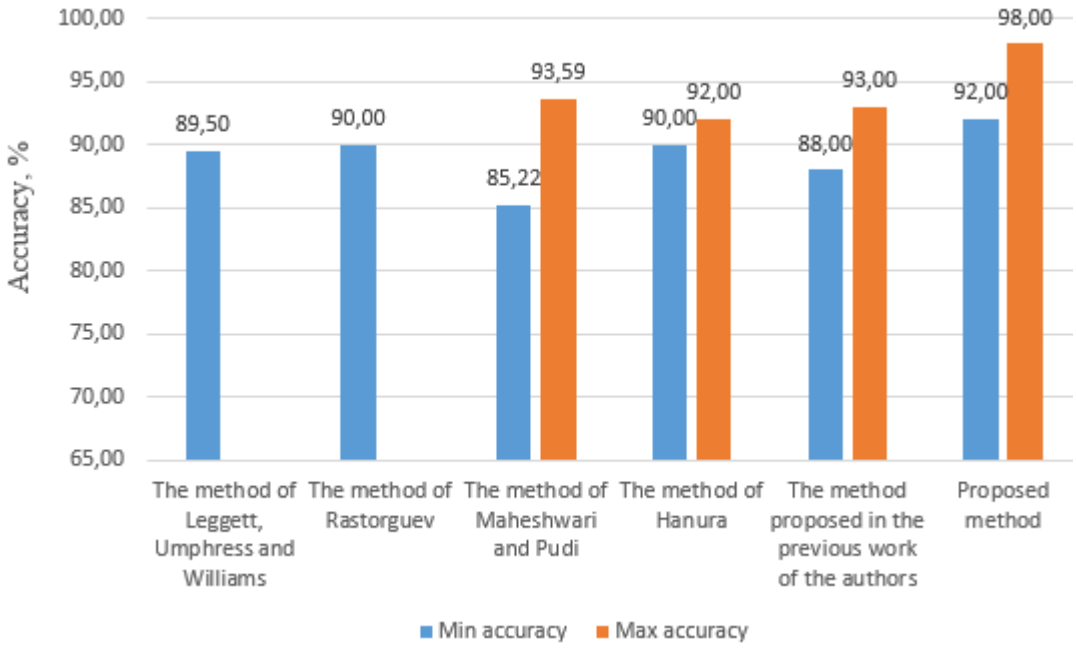

Figure 4: Recognition accuracy comparison of proposed and existing methods

So the scientific novelty of the work is the proposed method of user authentication by keyboard handwriting based on neural network and genetic algorithm, the feature of which is the use of a neural network in the form of a two-level system of direct access to a network with 70 sigmoid hidden neurons, 10 sigmoid source neurons and sigmoid activation function, the use of error injection into the code word, and the use of a hash function based on a genetic algorithm with crossover on the worst gene, mutation of the two worst genes and fitness function, which uses analysis based on five statistical tests. This allowed to increase the recognition accuracy of the user's keyboard handwriting to 92-98%, which is better by 3-11% compared to existing methods.

6. Conclusions

An experimental study of the possibility of using a neural network and a genetic algorithm to improve the accuracy of user identification based on keyboard handwriting with error injection was made and as a result a method of user authentication was proposed, which is based on a two-level neural network architecture using five-time functions and built-in sigmoid activation function to increase the efficiency of the neural network. An error injection was also introduced, which allowed to collect more accurate data on human handwriting and increase the accuracy of correct recognition of the user and his successful authentication by 3-11% compared to existing methods.

The use of a hash function based on a genetic algorithm is proposed, which is aimed at increasing the security of storing a code word in the database and the impossibility of making any changes to it, as during the authentication process the code word is not compared by itself, but the hash values.

7. References

[1] Gavan Leonard Tredoux, Steven J. Harrington. Method and system for providing authentication through aggregate analysis of behavioral and time patterns. Xerox Corporation, Norwalk, CT, 2016.

[2] El-Hajj, M., Chamoun, M., Fadlallah, A., & Serhrouchni, A. “Analysis of authentication techniques in Internet of Things (IoT).” In: 2017 1st Cyber Security in Networking Conference (CSNet). IEEE, 2017, pp. 1-3. doi: 10.1109/CSNET.2017.8242006.

[3] Salminen, J., Jung, S. G., Chowdhury, S., Sengün, S., & Jansen, B. J. “Personas and analytics: A comparative user study of efficiency and effectiveness for a user identification task.” In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020, pp. 1-13. doi: 10.1145/3313831.3376770.
[4] Luiz G. Hafemann, Robert Sabourin, Luiz S. Oliveira. Learning features for offline handwritten signature verification using deep convolutional neural networks. Computer Vision and Pattern Recognition, 2017, pp 163–176.
[5] Umphress David & Williams Glen. Identity verification through keyboard characteristics. International Journal of Man-Machine Studies, 1985, pp. - 263-273. doi: 10.1016/S0020-7373(85)80036-5.
[6] Young J.R. and Hammon R.W. Method and Apparatus for Verifying an Individual's Identity. Patent Number 4,805,222, U.S. Patent and Trademark Office, Washington, D.C., Feb., 1989.
[7] Rastorguev S.P. Software methods for protecting information in computers and networks. Moscow: «Yakhtsmen» Agency Publishing House, 1993, 188 p.
[8] Saket Maheshwary, Soumyajit Ganguly, Vikram Pudi. Deep Secure: A Fast and Simple Neural Network based approach for UserAuthentication and Identification via Keystroke Dynamics. Conference: 2017 International Joint Conference on Artificial Intelligence (IJCAI), At Melbourne, Australia, 2017.
[9] Harun N., Woo W.L. and Dlay S.S. Performance of KeystrokeBiometrics AuthenticationSystem Using Artificial Neural Network (ANN) and Distance Classifier Method. International Conference on Computer and Communication Engineering (ICCCE 2010). 11–13 May 2010, Kuala Lumpur, Malaysia, 2010.
[10] Danilyuk I.I., Karpinets V. V., Pryimak A.V., Yaremchuk, Y. Y., Kostyuchenko O.I. Neural network based method of a user identification by keyboard handwriting. Data recording, storage & processing, 20(2), 2018, pp. 68-76. doi: 10.35681/1560-9189.2018.20.2.142913.