Extraction of Features Characterizing Keystroke Dynamics in the User Identification problem

Makhamadaziz Rasulmukhamedov[1[0000-0001-8404-3013]], Gulmira Mirzaeva[2[0000-0002-0412-3951]], Nomaz Mirzaev[2[0000-0002-0657-2654]], Nuraddin Gafforov[1,3[0000-0003-4905-3675]]

[1]Tashkent State Transport university, Tashkent, Republic of Uzbekistan
[2]Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Republic of Uzbekistan
[3]Digital technologies and artificial intelligence research institute, Tashkent, Republic of Uzbekistan
mrgaffarov28@gmail.com

Abstract. This article examines the separation of characteristics describing the user's behavior in the matter of identifying the user based on keyboard dynamics. The study analyzed time sequences recorded by users while working with the keyboard. Haar wavelet transformation was used to distinguish characteristics that characterize user properties. The obtained wavelet coefficients were used as characters characterizing users. The advantages of the highlighted features were checked using SVM, MLP, Random Forest, and KNN algorithms. The results obtained can be used in the development of biometric authentication systems.

Keywords: Keyboard dynamics, user identification, Haar wavelet transform, SVM, MLP, KNN, Random Forest, time series, hold time, flight time.

1 Introduction

In the modern digital technology environment, ensuring information security remains one of the pressing problems. In particular, methods that ensure reliable user authentication are an integral part of any information system. Traditional authentication systems based on passwords, PIN codes, and physical tokens, despite their ease of use, have serious shortcomings in terms of information security. In addition, the high probability of these tools being forgotten under the influence of subjective factors, appropriated by third parties, as well as detected or compromised as a whole as a result of various cyberattacks reduces their reliability [1]. Biometric authentication technologies, including fingerprint or facial recognition, voice identification methods, despite their high accuracy, are not widely used in all fields due to the need for special hardware and software tools for their implementation, as well as the existence of possible risks to the confidentiality of personal biometric data [2].

Therefore, in recent years, passive biometric systems based on tracking user-device interaction have attracted special attention. In particular, dynamic authentication methods based on the user's keyboard typing actions are considered one of the promising directions in this direction. Each user's typing style when working with the keyboard, in particular the interval between pressing and releasing keys, the time intervals between keystrokes, typing speed, and rhythmic characteristics, are individual, and it is possible to identify a person based on these characteristics [3, 4].

This study proposes the use of wavelet transformation to implement authentication based on user keyboard typing actions. Wavelet transform allows for the identification of complex signal features through time-frequency analysis, and in the process serves to extract unique user-specific parameters [5]. This approach is especially effective in determining the temporal and local characteristics of user activity.

2 Methodology

2.1 General Description of the Study

This research aims to solve the problem of biometric authentication based on keyboard dynamics and aims to deeply analyze typing time signals in the user identification process. The main focus of the research is on identifying the time-varying and uneven properties of data entered through the keyboard, as well as identifying biometric features specific to the user from these characteristics [7].

The research process includes the following sequential steps: using a ready-made keyboard dynamics dataset, pre-processing the raw data, feature extraction based on wavelet, user identification using machine learning algorithms, and evaluating the final results based on statistical criteria. These stages are interconnected and serve to ensure the reliability and practical effectiveness of the proposed methodology.

The research results create a scientific basis for further improvement of authentication systems based on keyboard dynamics and their implementation in real information security systems [8].

2.2 Dataset

To study the problem of identifying users based on keyboard dynamics, an open and widely used dataset called “DSL-StrongPasswordData.csv” was used [3, 6]. This dataset was presented by K.S. Killourhy and R.A. Maxion at the DSN-2009 conference. A total of 51 users participated in the dataset, with each user entering the same predetermined text sequence (similar to a static password) several hundred times [4]. As a result, there are an average of 400 writing samples for each user, which allows for the study of stable statistical characteristics of individual writing style [9,16].

Each record in the dataset is based on key press and key release timestamps. The total data volume is over 20,400 records. Each record contains over 30 time-related features, mainly of the Hold Time (Dwell Time) and Flight Time types. Hold Time represents the difference between the time a button is pressed and released, while Flight Time describes the time interval between two consecutive button presses. These time parameters incorporate the speed, rhythm, and motor characteristics specific to individual users [3,4].

One of the important advantages of the dataset is that it is open source and allows for direct comparison of research results with other scientific works. Therefore, using this dataset allows for an objective assessment of the reliability and effectiveness of the proposed wavelet-based approach.

2.3 Data Preprocessing

In the study, data describing keyboard typing behavior was expressed as a time-dependent feature vector for each user [10]. Each writing sample is represented by a set of properties of the following type:

$$X = \{x_1, ..., x_i, ..., x_n\} \tag{1}$$

here x_i– is one of the time parameters recorded during keyboard operation, n– is the number of common features allocated for one recording sample.

Determining incorrect and extreme values. Extreme values arising as a result of random pauses or technical delays that may occur in the set of properties can negatively affect the model's efficiency. Therefore, each element in vector X was passed through a statistical filter:

$$x_i \in \left[\mu_x - 3\sigma_x,\ \mu_x + 3\sigma_x\right]$$

Values outside this interval were assessed as noise, removed, or replaced with threshold values [11].

Normalize features. In order to reduce the overall typing speed differences between users and bring all features to a single scale, Z-score normalization was applied. As a result, the normalized feature vector is expressed as:

$$x_i^{norm} = \frac{x_i - \mu_x}{\sigma_x}$$

As a result of this process, all features were centered around zero and had a variance of one, which ensures stable and accurate operation of machine learning algorithms.

Standardization of time series. To apply the wavelet transform, all X vectors were scaled to the same size. To do this, short sequences were zero-padded, and long sequences were truncated to a specified maximum length:

$$x_i \in \mathrm{R}$$

As a result, the pre-processed feature set was made ready for wavelet-based multi-stage analysis [14,15].

2.4 Wavelet-Based Feature Extraction

Since time series related to keyboard dynamics are highly variable and have local characteristics, it is advisable to use a multi-resolution approach in their analysis [17]. Therefore, the discrete Haar wavelet transform was used in this study. The Haar wavelet has a simple structure, is effective in detecting sharp signal changes, and allows identifying time symbols specific to the keyboard recording process [18].

Let's isolate N sequentially from the previously given data (1), which can be represented as:

$$\mathfrak{X} = (x_1, \ldots, x_i, \ldots, x_N), \tag{2}$$

Here x_i– elements consisting of the values of Hold Time and Flight Time recorded during keyboard recording ($i = \overline{1, N}$; $N = 2^n$).

It should be noted that the sequence of elements of the vector $\mathfrak{X}$ consists of discrete values, measured in milliseconds. In order to apply the Haar wavelet transform, it was ensured that N has a degree of 2.

The Haar wavelet transform processes the signal based on pairwise elements. Each pair is decomposed using the following two orthogonal basis functions

$$\varphi = \left(\frac{1}{\sqrt{2}}; \frac{1}{\sqrt{2}}\right), \; \psi = \left(\frac{1}{\sqrt{2}}; -\frac{1}{\sqrt{2}}\right)$$

We will group the selected (2) initial signal sequence in pairs:

$$(x_1, x_2), (x_3, x_4), \ldots, (x_{N-1}, x_N) \tag{3}$$

The approximation and detail coefficients are calculated for each of the separated (3) pairs:

$$C_{1,k} = \frac{x_{2k-1} + x_{2k}}{\sqrt{2}}, \tag{4}$$

$$d_{1,k} = \frac{x_{2k-1} - x_{2k}}{\sqrt{2}}, \tag{5}$$

$$k = 1,2 \ldots, [N/2].$$

Here:

$C_{1,k}$— represents the low-frequency (general) component of the signal,

$d_{1,k}$— represents the high-frequency (rapidly changing) component.

The obtained approximation coefficients are decomposed again using the Haar wavelet in the next step. Thus, using formulas (4) and (5) sequentially, we generate the following vector:

$$\left(C_J, d_J, d_{J-1}, \ldots, d_1\right),$$

Here $J = \log_2 N$ – maximum decomposition rate. As a result, all the coefficients obtained fully represent the original signal and incorporate features at different time scales.

In this study, the approximation and detail coefficients obtained as a result of the Haar wavelet transform were used directly as a set of features, without performing additional statistical aggregation [19,20]. That is, the feature vector for each writing sample was formed as follows:

$$\mathrm{F} = [C_J, d_J, d_{J-1}, \ldots, d_1]$$

This approach allows for maximum preservation of the time-scale structure of keyboard dynamics and serves to transfer user-specific typing patterns to classification algorithms without losing them [19,21].

2.5 User identification algorithms

Several machine learning algorithms were tested to identify users based on the extracted features. Including:

- Support Vector Machine (SVM);
- K-Nearest Neighbors, (KNN);
- Random Forest (RF);
- Multilayer Perceptron (MLP);

Each model was trained using 70% of the dataset and tested on the remaining 30%. This distribution was applied to evaluate the models' generalization capability and to ensure the objectivity of the results.

3 Results

3.1 Analysis of extracted features using Haar wavelet transform

The features obtained based on the Haar wavelet transform made it possible to reflect the time structure of the keyboard typing process at different scales. The zoom and detail coefficients obtained as a result of multi-step steps clearly demonstrated individual differences in the typing rhythm and speed of users. In particular, significant differences were observed in the distribution of wavelet coefficients between typing samples belonging to different users [22].

The results of the analysis showed that low-level approximation coefficients reflected the general direction of the writing process, while high-level detail coefficients represented rapid changes over a short period of time. It was these detailing coefficients that became important in distinguishing users and made it possible to identify subtle and stable styles inherent in the writing process. As a result, it was found that features extracted based on the Xaar wavelet have high discrimination ability in user identification [19,20].

3.2 Test results of user identification algorithms

The feature set generated based on the Haar wavelet transform was tested using several machine learning algorithms. In particular, the K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest (RF) models were used. During model evaluation, the data was divided into training and test sets, and each algorithm was tested under the same conditions [11,12,13].

The results of various classification algorithms were evaluated based on the criteria of Accuracy, Precision, Recall, and F1-score. The obtained results are presented in Table 1, which enables the comparison of model effectiveness and the identification of the most optimal algorithm.

Table 1. User identification results

Algorithm	Accuracy	Precision	Recall	F1-score
KNN	86.7	0.88	0.87	0.87
Random Forest	90.1	0.90	0.90	0.90
MLP	91.5	0.92	0.92	0.92
SVM	**92.6**	**0.93**	**0.93**	**0.93**

As can be seen from the table, all algorithms showed high accuracy with features formed on the basis of the Haar wavelet transform. This case confirms that preserving the time-scale structure of keyboard typing dynamics significantly improves user recognition efficiency. In particular, a comparative analysis was performed in Table 2 based on the accuracy and balanced evaluation indicators F1-macro, F1-weighted, and F1-score of the KNN, Random Forest, MLP, and SVM models.

Table 2. Classification algorithm results (based on F1-score)

Algorithm	Accuracy	F1-macro	F1-weighted	F1-score
KNN	86.7	0.8680	0.8681	0.87
Random Forest	90.1	0.9015	0.9015	0.90
MLP	91.5	0.9158	0.9158	0.92
SVM	**92.6**	**0.9272**	**0.9272**	**0.93**

According to the results, the SVM algorithm achieved the highest accuracy rate (92.6%). This is explained by the fact that the SVM model can effectively form optimal splitting limits in the space of high-dimensional wavelet features [13]. The MLP (Neural Network) model also demonstrated high accuracy and proved effective in studying nonlinear connections [12,23]. While the Random Forest algorithm provided stable and balanced results, the KNN algorithm had a relatively lower accuracy due to the complexity of the feature space [11].

4 Discussion

The obtained experimental results showed that the effectiveness of classical machine learning algorithms in biometric identification systems is different. Among the four models studied during the experiment - SVM, MLP, Random Forest, and KNN - the

SVM (Support Vector Machine) algorithm showed the highest indicators in all metrics.

Table 3. Overall performance indicators of models

Algorithm	Accuracy	F1-macro	F1-weighted	FAR	FRR	EER
KNN	0.865236	0.868122	0.868122	0.013768	0.053137	0.033453
Random Forest	0.901167	0.900075	0.900075	0.023462	0.023529	0.023496
MLP	0.916085	0.915871	0.915872	0.015226	0.015686	0.015457
SVM	**0.927647**	**0.927261**	**0.927262**	**0.012541**	**0.012353**	**0.012447**

According to the data in Table 3, the SVM model significantly surpasses other models with an accuracy of 92.76%. The closeness of the F1-macro and F1-weighted (0.927) indicators confirms the model's robustness to class imbalance in the dataset and its high generalization ability. This once again proves the effectiveness of SVM in the isolation of biometric features in multidimensional space.

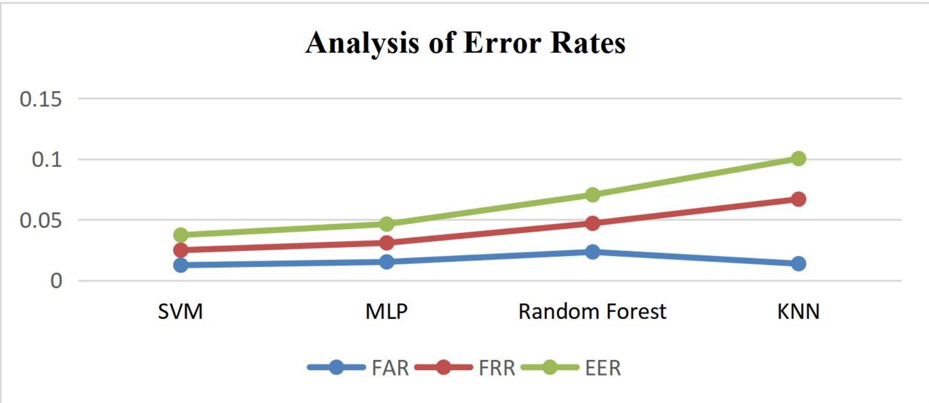

Fig. 1. Comparative diagram of the selected models by error levels (FAR, FRR, EER)

The effectiveness of classification models is determined not only by their overall accuracy, but also by the features of the errors made. Especially in biometric identification and security systems, the balance between FAR (False Acceptance Rate) and FRR (False Rejection Rate) is important (Fig. 1).

According to the results, the EER (Equal Error Rate) indicator in the SVM model was the lowest and amounted to 0.0124. The low EER value indicates that the model was able to find the optimal balance point between FAR and FRR. This effectively prevents unauthorized access to the system while avoiding unnecessary obstacles for legitimate users [24]. The MLP model (EER=0.0154) also showed high stability, confirming the ability of neural networks to reduce errors by learning nonlinear relationships.

Analysis of errors showed that the KNN model is the weakest link in this group. Its FRR (0.0531) is significantly higher compared to other models. A high FRR indicator causes inconvenience to system users, meaning the system often rejects genuine users.

This can be explained by the fact that the KNN algorithm is sensitive to data noise and makes more errors in boundary cases when making decisions based on the "nearest neighbor" principle.

Although the Random Forest model performed moderately well, its error rate was not as low as that of SVM. This means that the tree ensemble cannot accurately adapt to specific changes (variations) in the dataset, such as SVM [25].

5 Conclusion

In this study, a feature extraction approach based on Haar wavelet transformation was proposed and experimentally evaluated for user identification based on keyboard record dynamics. In the research process, the time sequences recorded by users during keyboard typing were subjected to Haar multilevel wavelet decomposition, and the obtained approximation and detail coefficients were directly used as symbols in the classification process.

Experimental results showed that features extracted based on the Xaar wavelet effectively reflect the time-scale structure inherent in users' writing process and significantly improve identification accuracy. As a result of tests using various machine learning algorithms, it was found that the SVM (RBF) and MLP (neural network) models achieved the highest accuracy and F1 scores, which is explained by the ability to deeply study a complex feature space consisting of wavelet coefficients.

Within the framework of further research, it is planned to introduce features based on Haar wavelet into deep learning models, increase the number of users, and assess the stability of the system under various input conditions (free text, different types of keyboards). The results of this work serve as a reliable scientific basis for the development of biometric authentication systems based on the dynamics of keyboard typing.

References

1. Smith J. (2019). Modern Authentication Mechanisms: Security and Limitations. Journal of Cybersecurity Research, 12(3), 145–160.
2. Kumar, R., & Lee, J. (2020). Biometric Systems and Privacy Concerns. International Journal of Information Security, 8(1), 22–34.
3. Kevin S. Killourhy and Roy A. Maxion. "Comparing Anomaly Detectors for Keystroke Dynamics," in Proceedings of the 39th Annual International Conference on Dependable Systems and Networks (DSN-2009), pages 125-134, Estoril, Lisbon, Portugal, June 29-July 2, 2009. IEEE Computer Society Press, Los Alamitos, California, 2009.
4. T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer. "ROCR: visualizing classifier performance in R," *Bioinformatics* 21(20):3940-3941 (2005).
5. Mallat Stéphane, A Wavelet Tour of Signal Processing (Third Edition), Academic Press, 2009, Pages 377-434, ISBN 9780123743701, https://doi.org/10.1016/B978-0-12-374370-1.00012-4.
6. Kevin Killourhy and Roy Maxion. Keystroke dynamics-benchmark data set. Carnegie-MellonUniversity, http://www. cs. cmu. edu/˜ keystroke/# sec2, 2009.

7. S. Cho, C. Han, D. H. Han, and H. Kim. Web-based keystroke dynamics identity verification using neural network. Journal of Organizational Computing and Electronic Commerce, 10(4):295–307, 2000.
8. Yu Zhong, Yunbin Deng, and Anil K Jain. Keystroke dynamics for user authentication. In 2012 IEEE computer society conference on computer vision and pattern recognition workshops, pages 117–123. IEEE, 2012.
9. Yan Sun, Hayreddin Ceker, and Shambhu Upadhyaya. Shared keystroke dataset for continuous authentication. In *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2016.
10. Pranit Shinde, Saideep Shetty, and Mahendra Mehra. Survey of keystroke dynamics as a biometric for static authentication. International Journal of Computer Science and Information Security, 14(4):203, 2016.
11. Jiankun Hu, Don Gingrich, and Andy Sentosa. A k-nearest neighbor approach for user authentication through biometric keystroke dynamics. In 2008 IEEE International Conference on Communications,pages 1556–1560. IEEE, 2008.
12. Norhaslinda Harun, Satnam Singh Dlay, and Wai Lok Woo. Performance of keystroke biometrics authentication system using multilayer perceptron neural network (mlp nn). In *2010 7th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP 2010)*, pages 711–714. IEEE, 2010.
13. Romain Giot, Mohamad El-Abed, and Christophe Rosenberger. Keystroke dynamics with low constraints svm based passphrase enrollment. In 2009 IEEE 3rd International Conference on Biometrics:Theory, Applications, and Systems, pages 1–6. IEEE, 2009.
14. Alaa Darabseh and Akbar Siami Namin. Effective user authentications using keystroke dynamics based on feature selections. In 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), pages 307–312. IEEE, 2015.
15. Ned Bakelman, John V Monaco, Sung-Hyuk Cha, and Charles C Tappert. Keystroke biometric studies on password and numeric keypad input. In 2013 European Intelligence and Security Informatics Conference, pages 204–207. IEEE, 2013.
16. Alsultan, A., & Warwick, K. (2013). Keystroke dynamics authentication: a survey of free-text methods. International Journal of Computer Science Issues (IJCSI), 10(4), 1.
17. Radomir S. Stanković, Bogdan J. Falkowski, The Haar wavelet transform: its status and achievements, Computers & Electrical Engineering, Volume 29, Issue 1, 2003,Pages 25-44, ISSN 0045-7906, https://doi.org/10.1016/S0045-7906(01)00011-8.
18. D. Sundararajan, "The Haar Discrete Wavelet Transform," in *Discrete Wavelet Transform: A Signal Processing Approach* , Wiley, 2015, pp.97-130, doi: 10.1002/9781119113119.ch8.
19. Mirzaev. N., Urinboev. J., Gafforov. N., Nugmanova. M. Analysis of Speech Signal Processing Methods in Speech Recognition Systems. 2025 6th International Conference on Problems of Cybernetics and Informatics (PCI), Baku, Azerbaijan, 26-28 August 2025, pp.
20. Opanasenko, V. M., Fazilov, S. K., Mirzaev, N. M., & Kakharov, S. S. (2025). A Model of Recognition Algorithms Based on Threshold Functions and Object Proximity Assessment. Cybernetics and Systems Analysis, 61(1), 148-160.
21. E. Yu and S. Cho. GA-SVM wrapper approach for feature subset selection in keystroke dynamics identity verification. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), pages 2253–2257. IEEE Press, 2003.
22. B. Hwang and S. Cho. Characteristics of auto-associative MLP as a novelty detector. In Proceedings of the IEEE International Joint Conference on Neural Networks, volume 5, pages 3086–3091, 10–16 July, 1999, Washington, DC, 1999.

23. S. Cho, C. Han, D. H. Han, and H. Kim. Web-based keystroke dynamics identity verification using neural network. Journal of Organizational Computing and Electronic Commerce, 10(4):295–307, 2000.
24. R. S. Gaines, W. Lisowski, S. J. Press, and N. Shapiro. Authentication by keystroke timing: Some preliminary results. Technical Report R-2526-NSF, RAND Corporation, May 1980.
25. Ramzi Saifan, Asma Salem, Dema Zaidan, Andraws Swidan, et al. A survey of behavioral authentication using keystroke dynamics: Touch screens and mobile devices. Journal of Social Sciences (COES&RJ-JSS), 5(1):29–41, 2016.