Applied Computer Science, vol. 17, no. 4, pp. 75–83
doi: 10.23743/acs-2021-30

Submitted: 2021-10-08 | Revised: 2021-12-10 | Accepted: 2021-12-20

Keywords: keystroke dynamics analysis, machine learning, neural network, supervised learning, classification problem

Nataliya SHABLIY [0000-0002-1125-4757]*, *Serhii LUPENKO* [0000-0002-6559-0721]*, *Nadiia LUTSYK* [0000-0002-0361-6471]*, *Oleh YASNIY* [0000-0002-9820-9093]*, *Olha MALYSHEVSKA* [0000-0003-0180-2112]**

KEYSTROKE DYNAMICS ANALYSIS USING MACHINE LEARNING METHODS

Abstract

The primary objective of the paper was to determine the user based on its keystroke dynamics using the methods of machine learning. Such kind of a problem can be formulated as a classification task. To solve this task, four methods of supervised machine learning were employed, namely, logistic regression, support vector machines, random forest, and neural network. Each of three users typed the same word that had 7 symbols 600 times. The row of the dataset consists of 7 values that are the time period during which the particular key was pressed. The ground truth values are the user id. Before the application of machine learning classification methods, the features were transformed to z-score. The classification metrics were obtained for each applied method. The following parameters were determined: precision, recall, f1-score, support, prediction, and area under the receiver operating characteristic curve (AUC). The obtained AUC score was quite high. The lowest AUC score equal to 0.928 was achieved in the case of linear regression classifier. The highest AUC score was in the case of neural network classifier. The method of support vector machines and random forest showed slightly lower results as compared with neural network method. The same pattern is true for precision, recall and F1-score. Nevertheless, the obtained classification metrics are quite high in every case. Therefore, the methods of machine learning can be efficiently used to classify the user based on keystroke patterns. The most recommended method to solve such kind of a problem is neural network.

1. INTRODUCTION

It is hard to imagine modern world without different novel technologies. Therefore, the task of data protection is of high importance.

The authentication problem is as follows. Two parties are talking with each other, and one or both want to send their identity to the other (Gebrie & Abie, 2017). Authentication is

* Ternopil Ivan Puluj National Technical University, Faculty of Computer Information Systems and Software Engineering, Computer Systems and Networks Department, Ternopil, Ukraine, natalinash@gmail.com, serhii.lupenko@gmail.com, lutsyk.nadiia@gmail.com, oleh.yasniy@gmail.com
** Ivano-Frankivsk National Medical University, Department of Hygiene and Ecology, Ivano-Frankivsk, Ukraine, o16r02@gmail.com

the process of verifying the physical identity of a person and digital identity of a computer. User authentication is a cornerstone of any information system.

The principles, that are the basis of identification and authentication methods, can be divided into four groups (Gebrie & Abie, 2017):

- traditional password protection;
- verification of physical parameters of human (fingerprints, iris scanning, etc.) (Dhir et al., 2010);
- assessment of psycho-physical parameters;
- estimation of user information interests and dynamics of its change.

The password-based authentication is widely used in identity verification (Hwang, Lee & Cho, 2009). Nevertheless, it becomes unsafe when a password is obtained by third-party. Keystroke dynamics-based authentication (KDA) was invented that propose increased security (Gaines, Lisowski, Press & Shapiro, 1980). KDA is based on the fact that a user's keystroke patterns repeat themselves and are unique (Umphress & Williams, 1985). It can be employed in internet banking, ATM, and smartphones, which require high level of security. It is possible to add fingerprint, iris, and voice to the traditional password-based authentication (Jain, Bolle, & Pankanti, 2006; Dhir et al., 2010). Also, KDA needs special equipment and requires several actions of user (Ru & Eloff, 1997; Monrose, Reiter, & Wetzel, 2002).

It is clear that any biometric is not the best recognition method in all cases and its selection is specific for certain application. A comparison of features on seven factors is provided in Table 1 (Jain, Ross & Prabhakar, 2004).

Some systems require user to provide a card before it can get access the data of the system. The examples of such cards are credit cards, debit cards, cash-machine cards. Cards can have either a magnetic strip or a computer chip. Cards containing a computer chip are also known as smart cards. With this system, the user must provide such card before the machine will allow that person to access any information. With a key-lock system, a person must unlock the computer to get access to the system. Most PCs have a key-lock installed that allows the authorized user to lock out the keyboard. When the system is locked, keyboard input is not recognized. Cards and keys can be lost, stolen, or forged. Also, the key-locks on PCs can be disabled if a person can remove the case of the machine. This radical method is generally not necessary, since most PC locks use the same type of key. If a person has a computer with a key-lock, then it is possible that his or her key can open or close the lock on another unauthorized computer (Fischer, Halibozek & Walters, 2019).

PINs, passwords, and digital signatures are compatible with any computer system. PINs work in conjunction with various types of card systems. With this system, one inserts a card and then enters the PIN, a security number known only to the user. Passwords are special words required to access a computer system. Companies should require passwords to contain at least eight characters that could be any combination of special symbols, capital and lowercase letters, and numbers. Easily guessed or obvious passwords should not be employed in practice. Finally, the company may assign passwords to employees that are random combination of numbers, letters, or special symbols. If the system requires a higher degree of security, then a password should only be used once. Those are so called one time passwords (Fischer, Halibozek & Walters, 2019).

Tab. 1. Features of most common biometrics characteristics (Jain, Ross & Prabhakar, 2004)

Biometric characteristic	Universality	Distinctiveness	Permanence	Collectability	Performance	Acceptability	Circumvention
Facial thermogram	H	H	L	H	M	H	L
Hand vein	M	M	M	M	M	M	L
Gait	M	L	L	H	L	H	M
Keystroke	L	L	L	M	L	M	M
Odor	H	H	H	L	L	M	L
Ear	M	M	H	M	M	H	M
Hand geometry	M	M	M	H	M	M	M
Fingerprint	M	H	H	M	H	M	M
Face	H	L	M	H	L	H	H
Retina	H	H	M	L	H	L	L
Iris	H	H	H	M	H	L	L
Palmprint	M	H	H	M	H	M	M
Voice	M	L	L	M	L	H	H
Signature	L	L	L	H	L	H	H
DNA	H	H	H	L	H	L	L

Digital signatures system uses a public/private key system. The sender creates the signature with a public key, and the receiver reads it with a second, private key. The two largest drawbacks of the mentioned above systems are associated with passwords and PINs (Fischer, Halibozek & Walters,2019).

Passwords can be guessed. Users tend to use real words or dates (their name, birth date, friends' or children's names, user initials, ids, and so on). Some users even do not replace the default initial password. PINs and passwords are often written down by users in places that can be easily accessed by others (Fischer, Halibozek & Walters,2019).

Biometrics methods are based on measuring individual body features. Fingerprints, hand geometry, retinal characteristics, voice recognition, keystroke dynamics, signature dynamics are common ways to identify authorized users. The computer compares the item being scanned with a copy of the item stored in the computer's memory. If the compared items match, the computer allows access, or denies otherwise (Fischer, Halibozek & Walters, 2019).

The one of the most important issues with KDA is in the fact that keystroke patterns from unauthorized users are not available while training classifier (Hwang, Lee & Cho, 2009). Therefore, it is very hard to build binary classifier. This can be eliminated using novelty detection framework. The idea of novelty detection method is to identify the novel or abnormal patterns that occur in a large amount of normal data (Miljković, 2010). Novelty or outlier is a pattern in the data that signifies unexpected behavior. The aim of novelty detection is to determine abnormal system behaviors which differs from the normal state of a system (Chandola, Banerjee & Kumar 2009; Markou & Singh, 2003, Miljković, 2010).

In the study (Hwang, Lee & Cho, 2009), there was proposed to use artificial rhythms and tempo cues to provide consistency and uniqueness of typing patterns. Different novelty detectors were built based on various artificial rhythms and/or tempo cues. It was shown show that artificial rhythms and tempo cues improve authentication accuracies and can be implemented in real world authentication systems.

However, the approach with binary classifier does not take into account the patterns of another user trying to impersonate the one, its password it is typing. To overcome this limitation, instead of binary classifier, proposed in novelty detection approach, in the current study, the multiclass classifier was employed that enables authentication of specific user that enters the password. This allows adding extended security to the computer system.

There was performed the analysis of keystroke dynamics based on methods of machine learning. The time of delay while pressing the keyboard buttons was measured and was used to predict the user of the system among the known list of authorized persons. In this case, the classification task was solved.

2. METHODS

KDA was performed using the following supervised methods of machine learning: logistic regression, support vector machines, random forest, and neural networks.

Logistic regression, despite its name, is a classification model rather than regression model (Subasi, 2020). Logistic regression is method that allows determining the probability of a discrete outcome given an input variable. The most common logistic regression model deals with binary outcome; something that can take two values such as true/false, yes/no, 1/0, etc. Multinomial logistic regression is a model where there are more than two possible discrete outcomes. Logistic regression is used for classification tasks (Edgar & Manz, 2017). Python Scikit-learn module contains an optimized logistic regression implementation, which allows multiclass classification (Raschka, 2017).

Support vector machine (SVM) works as follows (Vaibhaw, Sarraf & Pattnaik, 2020). A hyperplane or a set of hyperplanes is created, that separate the feature vectors into several classes. It selects the hyperplane which is at the maximum distance from the nearest training samples. SVM determines the hyperplane with the maximal margin by mapping input data into high-dimensional space. SVM also employs regularization to prevent artifacts. Nonlinear SVM have a nonlinear decision boundary that is based on kernel function.

Random forest (RF) models are machine learning algorithms that make predictions by combining outcomes from a set of regression decision trees (Williams et al., 2020). Each tree is built independently and is based on a random vector sampled from the input data, with all the trees in the forest having the same distribution. The predictions from the forests are averaged using bootstrap aggregation and random feature selection. RF models are robust predictors for both small sample sizes and high dimensional data (Biau & Scornet, 2016).

Neural networks (NN) are computing systems inspired by the biological neural networks (Kohonen, 1982), which learn to solve tasks by considering examples without being programmed with any specific rules. The neural networks were applied to solve the variety of classification tasks in (Alyamani & Yasniy, 2020; Dewi & Utomo, 2021; Sridharan et al., 2021). The training of neural networks can be achieved in several ways, using, for instance, the approach called particle swarm optimization that was employed in (Al-Awad, Abboud & Al-Rawi, 2021).

3. RESULTS AND DISCUSSION

Each of three users typed the same word that has 7 symbols 600 times. The row of the dataset consists of 7 values that are the time period during which the particular key was pressed. The ground truth values are the user id (either 1, 2 or 3). The task was to predict the user based on the typed word.

Data normalization scales the feature values to make them belong to the same interval, and, therefore, have the same importance. Because most machine learning algorithms produce better models when the data are normalized, the numerical data should be normalized or standardized before classification. There are three most commonly employed normalization techniques: z-score normalization, min-max normalization, and normalization by decimal scaling (Javaheri, Sepehri & Teimourpour, 2013). For this study the z-score normalization was applied. The data were normalized using its mean and standard deviation. After the preprocessing, all features have a mean of zero and a standard deviation of one. For each variable, this was performed by subtracting the mean of the variable and dividing by the standard deviation.

The dataset was divided into two unequal parts, namely, the training set and the testing set. The testing set contained 33% of the dataset, while the training set consisted of remaining 67% of the entire dataset.

Four methods of supervised learning were employed: logistic regression, support vector machines, random forest, and neural networks, similarly to (Alyamani & Yasniy, 2020). For each method, the normalized confusion matrices were obtained. Fig. 1 shows the normalized confusion matrices, built by means of machine learning methods for the mentioned above dataset.

The obtained results are based on the modern methods of machine learning and main postulates of statistics and probability theory.

The confusion matrix is commonly used measure that is employed while solving classification tasks. It can be equally applied to binary classification as well as for multiclass classification task. Confusion matrices contain counts from predicted and actual values.

To obtain the normalized confusion matrix, the corresponding row of original confusion matrix was divided into number of dataset samples that were created by each user. In this study, this number was equal to 600.

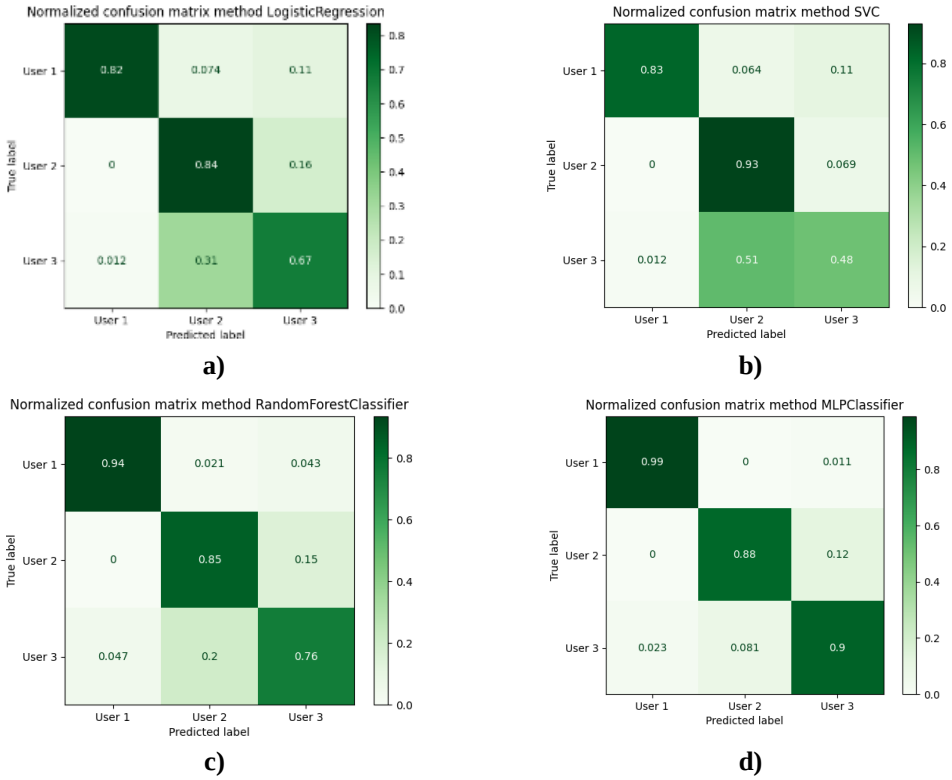

Fig. 1. Normalized confusion matrices obtained by various methods of machine learning: a) Logistic regression, b) Support vector machines, c) Random forest, d) Multilayer Perceptron (Neural network)

Using neural network approach, user 1 was detected in 99% of cases. The second place has user 3 with 90% of detection. The last place had user 2 with 88% of detection. In general, user 3 was misclassified most frequently as user 2 in the methods of support vector machines in around 51% of cases. The method of logistic regression classified 31% percent of user 3 samples as user 2. The same pattern is true for Random forest classifier with 20% of user 3 samples misclassified as user 2.

The classification metrics were obtained for each applied method. The following parameters were determined: precision, recall, f1-score, support, prediction, and area under the receiver operating characteristic curve (AUC). AUC provides a measure of performance across all possible classification thresholds. AUC takes value from [0, 1] (Bradley, 1997).

In case of neural network, its topology and hyperparameters were as follows: there were 3 hidden layers with numbers of neurons on each layer equal to ith element of the tuple (150, 10, 10), the employed algorithm was limited memory Broyden-Fletcher-Goldfarb-Shanno L-BFGS, the maximum number of iterations was equal to 1000000, the learning rate alpha was equal to 0.001.

Tab. 2. Logistic regression

Class	Precision	Recall	F1-score	Support	Predicted	AUC
1	0.987	0.819	0.895	94	78	0.984
2	0.740	0.836	0.785	116	131	0.906
3	0.667	0.674	0.671	86	87	0.863
Avg/Total	0.797	0.784	0.787	296	296	0.928

Tab. 3. Support vector classifier

Class	Precision	Recall	F1-score	Support	Predicted	AUC
1	0.987	0.829	0.901	94	79	0.989
2	0.683	0.931	0.788	116	158	0.903
3	0.694	0.476	0.565	86	59	0.841
Avg/Total	0.783	0.766	0.759	296	296	0.924

Tab. 4. Random forest

Class	Precision	Recall	F1-score	Support	Predicted	AUC
1	0.956	0.936	0.946	94	92	0.995
2	0.838	0.853	0.846	116	118	0.956
3	0.755	0.755	0.755	86	86	0.933
Avg/Total	0.852	0.851	0.851	296	296	0.968

Tab. 5. Neural network

Class	Precision	Recall	F1-score	Support	Predicted	AUC
1	0.978	0.989	0.984	94	95	0.985
2	0.935	0.879	0.906	116	109	0.971
3	0.836	0.895	0.865	86	92	0.948
Avg/Total	0.920	0.918	0.919	296	296	0.977

Tables 2–5 contain the classification metrics (precision, recall, F1-score, support, predicted, as well as AUC score for each class and for the dataset in total). The obtained AUC score is quite high. The lowest AUC score that was equal to 0.928 was achieved in the case of linear classifier. The highest AUC score was in the case of neural network classifier. The method of support vector machines and random forest showed slightly lower results as compared with neural network method. The same pattern is true for precision, recall and F1-score. Therefore, the methods of machine learning can be efficiently used to classify the user based on keystroke patterns. The best method that solved this task is neural network. Particularly, the proposed approach can be used in computer information systems to add another level of security and provide increased protection from potential intruders.

4. CONCLUSIONS

The task of users classification based on their keystrokes patterns was solved using the methods of supervised machine learning: logistic regression, support vector machines, random forest, and neural network. The multiclass classifier was built that allows determining the user based on its keystroke dynamics analysis with high accuracy. The method with highest classification score was neural network. The method with the lowest classification metrics was logistic regression. In general, the AUC score, obtained with each method, was more than 0.92. Therefore, such kind of task can be efficiently solved by means of machine learning approaches. This approach can be used in computer information systems to add another level of security and provide additional protection from potential intruders. In the future research, there can be used the extended dataset that includes data from a larger amount of users. Also, the hyperparameter optimization can be performed to increase the classification metrics.

REFERENCES

Al-Awad, N. A., Abboud, I. K., & Al-Rawi, M. F. (2021). Genetic Algorithm-PID controller for model order reduction pantographcatenary system. *Applied Computer Science, 17*(2), 28-39. https://doi.org/10.23743/acs-2021-11

Alyamani, A., & Yasniy, O. (2020). Classification of EEG signal by methods of machine learning. *Applied Computer Science*, *16*(4), 56-63. https://doi.org/10.23743/acs-2020-29

Biau, G., & Scornet, E. (2016). A Random Forest Guided Tour. *Test*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 1–58. https://doi.org/10.1145/1541880.1541882

Dewi, W., & Utomo, W. H. (2021). Plant classification based on leaf edges and leaf morphological veins using wavelet convolutional neural network. *Applied Computer Science*, *17*(1), 81–89. https://doi.org/10.23743/acs-2021-08

Dhir, Vijay, Singh, A., Kumar, R., & Singh, G. (2010). Biometric Recognition: A Modern Era For Security. *International Journal of Engineering Science and Technology*, *2*(8), 3364–80.

Edgar, T. W., & Manz, D. O. (2017). *Research Methods for Cyber Security*. Syngress.

Fischer, R. J., Halibozek, E. P., & Walters, D. C. (2019). Holistic Security Through the Application of Integrated Technology. *Introduction to Security, 2019*, 433–62. https://doi.org/10.1016/b978-0-12-805310-2.00017-2.

Gaines, R. S., Lisowski. W., Press, S. J., & Shapiro, N. (1980). *Authentication by Keystroke Timing*. The Rand Corporation.

Gebrie, M. T., & Abie, H. (2017). Risk-Based Adaptive Authentication for Internet of Things in Smart Home EHealth. *Proceedings of the 11th European Conference on Software Architecture: Companion Proceedings (ECSA'17)* (pp. 102–108). Association for Computing Machinery. https://doi.org/10.1145/3129790.3129801

Hwang, S.-S., Lee H., & Cho, S. (2009). Improving Authentication Accuracy Using Artificial Rhythms and Cues for Keystroke Dynamics-Based Authentication. *Expert Systems with Applications*, *36*(7), 10649–56. https://doi.org/10.1016/j.eswa.2009.02.075

Jain, A. K., Bolle, R. M., & Pankanti, S. (2006). *Biometrics. Personal Identification in Networked Society*. Springer.

Jain, A. K., Ross, A., & Prabhakar, S. (2004). An Introduction to Biometric Recognition. *IEEE Trans. on Circuits and Systems for Video Technology*, *14*(1), 4-19.

Javaheri, S. H., Sepehri, M. M. & Teimourpour, B. (2013). Response Modeling in Direct Marketing. A Data Mining-Based Approach for Target Selection. Data Mining Applications with R (pp. 153-180). Elsevier Inc. https://doi.org/10.1016/B978-0-12-411511-8.00006-2

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69.

Markou, M., & Singh, S. (2003). Novelty detection: a review—part 1: statistical approaches. *Signal Processing, 83*(12), 2481–2497. https://doi.org/10.1016/j.sigpro.2003.07.018

Miljković, D. (2010). Review of novelty detection methods. *The 33rd International Convention MIPRO* (pp. 593-598). IEEE.

Monrose, F., Reiter, M. K., & Wetzel, S. (2002). Password Hardening Based on Keystroke Dynamics. *International Journal of Information Security*, *1*(2), 69–83. https://doi.org/10.1007/s102070100006

Raschka, S. (2017). *Python Machine Learning. Second edition.* Packt Publishing Ltd.

Ru, W.G., & Eloff, J.H. (1997). Enhanced Password Authentication through Fuzzy Logic. *IEEE Expert, 12*, 38-45.

Sridharan, M., Rani Arulanandam, D. C., Chinnasamy, R. K., Thimmanna, S., & Dhandapani, S. (2021). Recognition of font and tamil letter in images using deep learning. *Applied Computer Science, 17*(2), 90–99. https://doi.org/10.23743/acs-2021-15

Subasi, A. (2020). *Practical Machine Learning for Data Analysis Using Python*. Academic Press.

Umphress, D., & Williams, G. (1985). Identity verification through keyboard characteristics. *International Journal of Man-Machine Studies*, *23*(3), 263–273. https://doi.org/10.1016/S0020-7373(85)80036-5

Vaibhaw, Sarraf, J., & Pattnaik, P.K. (2020). Brain–Computer Interfaces and Their Applications. *An Industrial IoT Approach for Pharmaceutical Industry Growth*, *2*, 31-54. https://doi.org/10.1016/b978-0-12-821326-1.00002-4

Williams, B., Halloin, C., Löbel, W., Finklea, F., Lipke, E., Zweigerdt, R., & Cremaschi, S. (2020). Data-Driven Model Development for Cardiomyocyte Production Experimental Failure Prediction. *Computer Aided Chemical Engineering*, *48*, 1639-1644. https://doi.org/10.1016/B978-0-12-823377-1.50274-3