King's Research Portal

Document Version
Peer reviewed version

[Link to publication record in King's Research Portal](#)

Performance Analysis of Free Text Keystroke Authentication using XGBoost

Ievgeniia Kuzminykh[1,2], Saransh Mathur[1], Bogdan Ghita[3]

1. Department of Informatics, King's College London, 30 Aldwych, London WC2B 4BG, UK

2. Department of Infocommunication Engineering, Kharkov National University of Radio Electronics, Nauki av. 14, 61000, Kharkiv, Ukraine

3. School of Engineering, Computing and Mathematics, University of Plymouth, Drake Circus, Plymouth PL4 8AA, UK

ievgeniia.kuzminykh@kcl.ac.uk

Abstract: Authentication based on keystroke dynamics is a form of behavioral biometric authentication that uses the user typing patterns and keyboard interaction as a discriminatory input. This type of authentication can be coupled with a fixed text password in a traditional login system to contribute to a multifactor authentication or provide continuous user authentication in a usable security system, where the typing patterns are continuously analysed to validate the user at run time. This paper investigates the effectiveness of free text keystroke for continuous authentication in real-world systems. Evaluation is performed using XGBoost multiclass classification, applied to an unbalanced free-text keystroke dataset. The introduction of additional activity-based features and removal of inaccuracies in the timing between keys allowed a reduction of the EER for the Clarkson II dataset from 14-24%, as achieved by previous studies, to 8% when employing the proposed method.

Keywords: Usable security; Keystroke dynamics; Continuous authentication; XGBoost.

1. Instruction

Recent years witnessed an increasing interest towards unobtrusive and usable security systems as the surface area of the Internet continuously expands, particularly to accommodate transparent authentication systems. In this context, keystroke dynamics provides a solid solution and has become the de-facto solution during the past decade. Nevertheless, it is yet to see a promising mass-scale deployment for keystroke dynamics-based authentication systems due to a number of inherent challenges, including the need for an extensive analysis of the keystroke dynamics-based authentication systems to be put in use and, amongst those, to establish if the systems are accurate enough to deploy. A significant number of studies, such as [1-4], investigated the feasibility of keystroke dynamics as a unique or additional means for user continuous authentication. Amongst their conclusions, the authors agreed that

performance of the free-text dynamics for user authentication is always worse when compared to specific text, such as username and password, as the variability of the input, has a significant impact on accuracy. In this study we also investigate whether the performance of free-text keystroke authentication can be improved by implementing a novel set of features for free-text keystroke dynamics.

The aim of this study is to investigate the applicability of free-text for biometric authentication with applying additional activity-based characteristics. One of the challenges encountered was that the initial sanitising of the dataset led to inaccuracies in the timing between successive keys; additional filtering was necessary to remove these artifacts and correct timing. This allowed us to reduce the EER for Clarkson II dataset from 14-24% that was achieved by the authors in [5] to 8%.

2.   Related Works

The summary of the related studies is presented in Table 1 and shows the machine learning algorithms that researchers were used for classification, the dataset used and what type of keystroke, free text or fixed text, was under experiment.

A recent study by Daribay et al. [6] focused on the performance evaluation of a fixed-text keystroke dynamics showed promising results with XGBoost Classifier giving an accuracy of 90.91%. XGBoost is a powerful machine learning algorithm, but it requires parameter tuning to leverage the full potential, which was missing from this analysis. While the overall result is encouraging, it does not address the continuous authentication aspect due to the fixed text employed.

Baynath et al. [7] further tested the large-scale applicability of keystroke dynamics, a dataset size for this study was also way larger than the previous ones, as they worked on a combination of the Killourhy Database (CMU database) [3] and their own inbuilt database consisting of fixed text of four different strong passwords. One of the most important conclusions of the study was that the cost of implementation for such system remains low even for large datasets, both computationally and financially.

There have also been some efforts to increase the usability component of continuous authentication, such as the unsupervised approach by Ananya and Singh [8] which did not require any preregistration or the method proposed by Sim et al. [9], where the mouse could be mounted with a fingerprint sensor for initial login and then periodic re-verification in conjunction with continuous authentication via keystroke dynamics. Since eliminating the onboarding process does not seem feasible, one way to increase the usability can be by reducing the number of keystrokes or inputs required from the user itself, particularly appealing for free-text keystroke systems, but prone to error as keystroke sample size is directly correlated with accuracy. To solve this issue, one can use the strategy adopted by Ayotte et al. in [10], aiming for frequent and cumulative authentication, using flexible thresholds. The authors proposed an instance-based tail area density (ITAD) metric to help reduce the number of keystrokes required to perform authentication.

Several other works [2], [3], [11-13] studied using keystroke dynamics for user continuous authentication. In works [11, 12] the authors considered continuous

authentication for users who remotely access the desktop machine via RDP protocol using distance-based algorithms to identify differences in the keystroke patterns. The authors of study [13] collected and used a mixed dataset, consisting of fixed-text and free text keystrokes, which was used by many subsequent studies [10, 14, 15].

Table 1. Related studies and their results (in the chronological order)

Study	Dataset, subjects	Type	Classifier	Metrics	Accuracy, %
A. Lo et al 2020 [16]	133	Fixed	RF, SVM, Manhattan, Euclidian	Accuracy	74.4-95.6
S. Singh et al (CMU) 2020 [17]	51	Fixed	KNN, SVC (RBF), RF, XGBoost	Accuracy	70.4-93.6
A. Daribay et al. [6] 2019	51	Fixed	XGBoost	Accuracy	90.91
K. Elliot et al. 2019 [18]	23	Fixed	RF, NN, DT, SVM	Accuracy	71-100
C. Murphy, et al. (Clarkson II) 2017 [19]	103	Free	Degree of disorder, n-graph reject ratio	EER	88.64
A. Bansal 2016 [20]	5	Free	GMM	Accuracy	78.4
Y. Sun et al. (Buffalo) 2016 [13]	148	Free	GMM	EER	96.6
A. Darabseh et al. 2015 [21]	28	Fixed	KNN, SVM	Accuracy	81-84
E. Vural et al. (Clarkson I) 2014 [22]	39	Mixed	Degree of disorder, n-graph reject ratio	FAR, FRR, EER	96.9
J. Roth (MSU) 2014 [23]	51 30	Fixed Free	Distance, n-graph reject ratio	EER	94.5
A. Messerman et al. 2011 [24]	55	Free	Degree of disorder	FAR, FRR	N/A
K. Killourhy and Maxion (CMU) 2009 [3]	51	Fixed	Manhattan, Mahalanobis and 12 more	EER, FAR	63-90
C. Loy, et al. 2007 [25]	100	Fixed	ARTMAP- FD	EER	88
D. Gunetti and C. Picardi 2005 [26]	40	Free	Distance, n-graph reject ratio	FAR, IPR	N/A
D.T. Lin 1997 [27]	125	Fixed	BPNN	FAR, IPR	N/A
R. Joyce and G. Gupta 1990 [28]	27	Fixed	Manhattan (filtered)	IPR, FAR	86.7

To summarise, it is apparent that the area of free-text based keystroke dynamics, although extremely capable, has presented less interest, particularly due to its potentially higher computational demands and error rates. Most of the studies test the practicality of a keystroke dynamics-based authentication system using the password-based fixed-text, which cannot be used for continuous, transparent authentication. Within the free text keystroke dynamics field, studies focused on a similar feature-set, including digraphs and n-graphs. Although, these methods have generally resulted in an accuracy approaching almost 90% most of the time, the potential of other types of features remains untapped.

3.   Methodology

This study aims to investigate and improve the accuracy of free text keystroke dynamics as a method for user authentication by introducing an additional, activity-based set of features, which were not used by prior research relating to free text patterns. The proposed features are derived from the physical characteristics of the keyboard; the user profile aggregates the hold times of specific keys in combination with their location. The process of finding out how good or bad such system would perform is done through XGBoost classification algorithm. The methodology, outlined in Fig.1, consists of data pre-processing to filter inaccurate data, followed by feature extraction, classification and evaluation steps [19]. Based on the conclusions of prior research, combined with the emergency of better performing algorithms, the process is using XGBoost for classification.
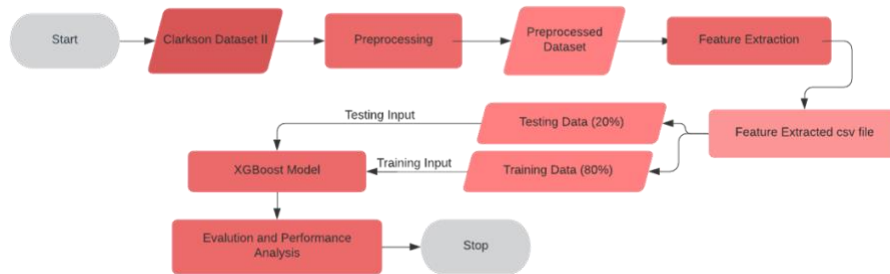
Fig. 1. Flowchart for implementation

3.1 Dataset

Given the requirements of the study and the available datasets, the Clarkson II dataset [20] was used as input. This dataset includes keystroke timing information for 103 subjects in a completely uncontrolled environment collected over a period of 2.5 years using a keylogger tool installed on each computer to record user interaction. The subjects used different hardware and OS platforms, different keyboards, different browsers, different software, and even different tasks. For each key interaction, the dataset contains the user ID, the key event (0 if pressed and 1 if released), the timestamp,

logged in .NET ticks, and the key name, as shown below in Table 2.

Table 2. Sample of dataset

User ID	Time Stamp (ticks)	Action Type	Key Name
4302075	636172286538589004 (2016-12-13 12:24:13)	'KeyDown'	'A'
4302075	636172286539669002 (2016-12-13 12:24:13)	'KeyDown'	'Space'
4302075	636172286541684820 (2016-12-13 12:24:14)	'KeyUp'	'A'

3.2 Data Sanitation and Pre-processing

As discussed in [10] and [16], the performance of algorithms on the Clarkson II dataset compared to other more controlled free text datasets is always worse. The root cause of the difference is the fact that users have a specific pattern when typing text sequences that they are familiar with, such as usernames and passwords, but use less discriminative typing for free text. The Clarkson database includes an additional element of error, as the authors sanitised the text to remove any sensitive sequences; in the process, timing of adjacent key presses was therefore also affected. To alleviate these artifacts, the dataset was filtered to remove incomplete patterns or accidentally pressed keys.

Based on the ability to filter and the removal of incomplete data (such as key events where only the key press was registered, with no key release), 24 users were removed from the dataset, leaving a total of 79 users with data suitable for analysis. The pre-processing also converted the .NET ticks to a *YYYY-MM-DD T HH:MM:SS.zzzz* format as well as replaced the key names with their respective ASCII values.

3.3 Feature Extraction

A set of 5 features was extracted from the pre-processed keylogger data, which differs conceptually from the ones used in the reviewed literature. All the studies in past have relied on digraph/n-graph based timing characteristics for the keys K_i and K_{i+1} pressed subsequently during typing. This is accurate, but it does not capture the physical characteristics of the keyboard. This study aims to provide a robust, less computationally intensive set of features, less prone to false negative errors, hence more appropriate for a transparent, additional layer of authentication. Several studies [11, 12, 21, 30, 31] have used letter position on the keyboard as one of the features, but applied it to fixed text keystroke dynamics. Besides features based on how a user's hand interacts with letters on the keyboard, the dataset also captures the use of shift and CAPSLock, as well as backspace and space keys are also assigned their separate features. All the features are based on average hold times and calculated per 10 minutes of keylogging. The labelled features are shown in the Table 3.

Table 3. Features extracted from keylogger data

Name	Label	Description
l	f2	Hold time(ms) of keys at left part without shift
r	f3	Hold time (ms) of keys at right part without shift
L	f0	Hold time(ms) of keys at left part with shift
R	f1	Hold time(ms) of keys at right part with shift
SPACE	f4	Hold time(ms) of Space key
BACKSPACE	f5	Hold time(ms) of Backspace key

The keyboard was divided into two parts- left and right - as per the traditional placement of a hand on the keyboard, as shown in Fig. 2.

Fig. 2. Finger positions on a keyboard

3.4 Machine Learning Algorithms

The dataset, after filtering, includes 79 users and the process is aiming for authentication, the classification problem is an imbalanced multiclass classification. Based on the existing studies, as summarised in section 2, XGBoost appears to be one of the most promising and successful in resolving such problems [32,33] and is a public domain classifier, hence, it will be used as part of this study. Gradient boosting (GB) methods are usually very powerful classifiers because of the ensemble training techniques that typically perform very well on unbalanced data. Amongst the available solutions, the sklearn library includes an excellent Python implementation of this algorithm, which was also used for this study.

For creating an XGBoost classifier model, the dataset must be split into input and target arrays. The input array contains all the feature rows, while the target array includes the corresponding entries for usernames. The input array is assigned to a variable 'x' and target array is assigned to a variable 'y'. We also split the data into training and testing datasets in proportion 80% and 20%, respectively.

4.   Results and Analysis

4.1.   Feature Importance

It is essential to evaluate the relative contribution and efficiency of the features as discriminants for the classes of outcomes of the model. Conducting the evaluation on the model gave the following results presented on Fig.3.
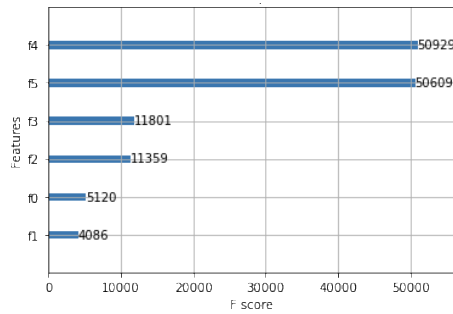
Fig. 3. Feature importance

The labels from f0 to f5 correspond to the features from Table 3. The F scores are the relevance indicators of the respective features when applied to the decision tree; the higher the F score of a feature, the higher its relative importance.

According to Fig.3, the combination of *space* and *backspace* average hold-time features have the highest importance in predicting the users. This is followed by the *l* and *r* features, with the lowest importance being for that of *L* and *R*. The low F Scores for *L* and *R* features are justifiable, as the frequency of uppercase characters is significantly lower than their lowercase counterparts. One possible explanation for the *backspace* feature is that the typing mistakes a user makes while typing are a good discriminator for the behaviour of different users. As, for *space*, having the highest importance could be a result of the frequency of use of space bar in general. Given their limited impact, the *L* and *R* features may be removed from the list of inputs with a minimal impact on the model accuracy; to be traded for a significant reduction in computational complexity.

4.2 Classification Accuracy

The accuracy of model is the primary concern in classification model that gives us a fraction for the samples that were predicted correctly. The accuracy can be calculated using Equation (1):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$ (1)

where True Positive (TP) is the rate of correct positive predictions, True Negative (TP)

is the rate of correct negative predictions, False Positive (FP) is the rate of incorrect positive prediction, and False Negative (FN) is the rate incorrect negative predictions.

In order to get an accurate indication of the ability of the model to generalise, the testing subset accuracy is also determined. After training the model with the training subset, test subset to the model, the model was tested against the testing subset. The results indicated an accuracy of is 91.72% for the testing subset. This accuracy is among highest among the related studies from Table 1 in area of free-text keystroke dynamics, and highest for such a large, unbalanced dataset. The original paper [19] that used the Clarkson II dataset reported an accuracy of 88.64% and a 10.36% EER, while our study showed only 8.28% EER. This result shows that the set of features presented in Table 3 is more effective than the equivalent set of timing features which does not take into account the key location.

4.3 Classification Report

We used three common metrics to evaluate the accuracy of the classifiers for each of the 79 users: precision, recall and F1-score. Precision is the true positive predictive value of a class, representing the ratio between the number of (TP) and the total number of predicted positive class. Recall evaluates the correctness of the class, defined as the ratio between the number of true positives and the total number of predictions of the respective class. The F1-score is the harmonic mean of recall and precision.

The results presented in Table 4 showed that the precision, recall, and F1 score for most of the users are high (0.85 - 1.0), except for those where support, which represents number of actual occurrences of the class in our dataset, has low values, as there are not enough training samples for these users to form a unique enough signature. As the dataset used is imbalanced, support in the training data will differ for each class. Support doesn't change between models but instead influences the evaluation process.

Table 4. A snippet of the results of classification for 6 users

User	Precision	Recall	F1-Score	Support
0	0.98	0.98	0.98	165
1	0.95	0.87	0.91	112
2	0.88	0.51	0.64	55
3	0.88	0.90	0.89	1110
16	0	0	0	1
17	0	0	0	5

From the report, it is apparent that classes with a lower number of samples are generally being predicted incorrectly and perform poorly. From a modelling perspective, there is not enough training data for these users to form a unique enough signature, hence they are classified as belonging to other classes. The users with few samples are the ones contributing the most to the False Positives for the other users, and False Negative for their own classes.

5. Conclusion

This study analyses the efficiency of deploying free text keystroke dynamics-based authentication as a continuous authentication method. The results showed an overall accuracy of 91.72%, and up to 98% for unique users, and the XGBoost based classifier can be implemented as a continuous authentication system, which ensures that the user does not change after an initial sign-on.

The paper proposes a user classification approach using a novel features set for free text keystroke dynamics, focused on the positioning of keys on the keyboard. The additional features, with the average hold-times of frequently pressed keys (backspace, space bar, and shift), lead to a higher accuracy in comparison with prior studies.

The results also indicated that accuracy for specific users is highly dependent on the amount of training data available, therefore users with a limited amount of data are likely to be incorrectly classified by the system.

References

1. Embroker Team, 2022 Must-Know Cyber Attack Statistics and Trends, Embroker Business and Advice Research, Jan 31, 2022. Accessed on: Mar. 15, 2022. [Online]. Available: https://www.embroker.com/blog/cyber-attack-statistics
2. Kang P, Cho S. Keystroke dynamics-based user authentication using long and free text strings from various input devices. Information Sciences, 2015, (308): 72-93.
3. Killourhy K S, Maxion R A. Comparing anomaly-detection algorithms for keystroke dynamics. Proc. of the IEEE/IFIP Int. Conf. Dependable Syst & Netw., 2009. 125-134.
4. Leggett J., Williams G, Usnick M, Longnecker M. Dynamic identity verification via keystroke characteristics. Int. J. Man-Machine Studies, vol. 1991, 35 (6): 859-870.
5. Lu X, Zhang S, Hui P, and Li P. Continuous authentication by free-text keystroke based on CNN and RNN. Comput. Security, 2020, 96: 101861.
6. Daribay A, Obaidat M S, Krishna P V. Analysis of Authentication System Based on Keystroke Dynamics. Proc of the Int. Conf. Computer, Inform. and Telecomm. Syst. (CITS), 2019. 1-6.
7. Baynath P, Soyjaudah K M S, Khan M H-M. Machine learning algorithm on keystroke dynamics pattern. Proc of the IEEE Conference on Systems, Process and Control (ICSPC), 2018. 11-16.
8. Ananya and Singh S. Keystroke Dynamics for Continuous Authentication. Proc of the 8th Int. Conf. Cloud Comp., Data Sci. Engin. (Confluence), 2018, 205-208.
9. Sim T, Zhang S, Janakiraman R, Kumar S. Continuous verification using multimodal biometrics. IEEE Trans. Pattern Anal. Mach. Intell., 2007, 29 (4): 687-700.
10. Ayotte B, Banavar M, Hou D, Schuckers S. Fast free-text authentication via instance-based keystroke dynamics. IEEE Trans. Biom. Behav. Ident. Sci., 2020, 2 (4): 377-387.
11. Kuzminykh I, Ghita B, Silonosov A. On Keystroke Pattern Variability in Virtual Desktop Infrastructure. Proc. of the Int.Worksh. Comput. Model. and Intell. Syst., 2021. 238-248.
12. Kuzminykh I, Ghita B, Silonosov A. Impact of Network and Host Characteristics on the Keystroke Pattern in Remote Desktop Sessions. arXiv:2012.03577 [cs], 2020.
13. Sun Y, H. Ceker H, Upadhyaya S. Shared keystroke dataset for continuous authentication. Proc of the IEEE Int. Workshop Inf. Forens. Secur. (WIFS), 2016. 1-6.

14. Kiyani A T, Lasebae A, Ali K, Rehman M, Haq B. Continuous User Authentication Featuring Keystroke Dynamics Based on Robust Recurrent Confidence Model and Ensemble Learning Approach", IEEE Access, 2020, (8): 156177-156189.

15. Huang J, Hou D, Schuckers S, Law T, Sherwin A. Benchmarking keystroke authentication algorithms. IEEE Workshop on Inf.Foren. and Sec.(WIFS), 2017. 1-6.

16. Lo A, Ayma V H, Gutierrez-Cardenas J. A Comparison of Authentication Methods via Keystroke Dynamics," Proc of the IEEE Engineering International Research Conference (EIRCON), 2020. 1-4.

17. Singh S, Inamdar A, Kore A, Pawar A. Analysis of Algorithms for User Authentication using Keystroke Dynamics. Proc of the International Conference on Communication and Signal Processing (ICCSP), 2020. 0337-0341.

18. Elliot K, Graham J, Yassin Y, et al. A comparison of machine learning algorithms in keystroke dynamics. Proc of the Int. Conf. Comp. Scie. and Comp. Intell., 2019. 127-132.

19. Saransh Mathur. Performance analysis of Free-Text Keystroke Dynamics based authentication using Machine Learning. MSc project, King's College London, 2021.

20. Murphy C, Huang J, Hou D, Schuckers S. Shared dataset on natural human-computer interaction to support continuous authentication research. Proc. of the IEEE Int. Joint Conf. Biometr. (IJCB), 2017. 525-530.

21. Bansal A, keystrokeDynamics/Readme, Mar. 2017 Accessed on: Mar. 15, 2022. [Online]. Available: https://github.com/ankiteciitkgp/keystrokeDynamics/blob/master/Readme.md

22. Darabseh A, Namin A S. On accuracy of classification-based keystroke dynamics for continuous user authentication. Proc of the Int. Conf. on Cyberworlds, 2015. 321-324.

23. Vural E, Huang J, Hou D, Schuckers S. Shared research dataset to support development of keystroke authentication. Proc. of the IEEE Int. Joint Conf. Biometr., 2014. 1-8.

24. Roth J, Liu X and D. Metaxas D. On Continuous User Authentication via Typing Behavior. IEEE Transactions on Image Processing, 2014, 23 (10): 4611-4624.

25. Messerman A, Mustafifa T, Camtepe S A, Albayrak S. Continuous and non-intrusive identity verification in real-time environments based on free-text keystroke dynamics. Proc. of the IEEE Int. Joint Conf. Biometr., 2011, 1-8.

26. Loy C C, Lai W K, Lim C P. Keystroke patterns classification using the ARTMAP-FD neural network. Proc of the Int. Conf. Intell. Inform. Hiding and Multimedia Sig. Proces., 2007, vol 1: 61-64.

27. Gunetti D, Picardi C. Keystroke analysis of free text. ACM Trans. Inform. and Syst. Secur. (TISSEC), 2005, 8 (3): 312-347.

28. Lin D-T. Computer-access authentication with neural network based keystroke identity verification. Proc. of the Int. Conf. Neural Networks, 1997, vol. 1: 174-178.

29. Joyce R, Gupta G. Identity authentication based on keystroke latencies. Commun. of the ACM, 1990, 33 (2): 168-176.

30. Alsultan A, Warwick K, Wei H. Improving the Performance of Free-text Keystroke Dynamics Authentication by Fusion. Applied Soft Computing, 2018, 70:1024-1033.

31. Singh P I. Robust Security System for Critical Computers. International Journal of Information Technology and Computer Science (IJITCS), 2012, 4 (6): 24-29.

32. Dhar P, Guha S. Fish Image Classification by XgBoost Based on Gist and GLCM Features. I.J. Information Technology and Computer Science (IJITCS), 2021, 4: 17-23.

33. Manju N, Harish B S and Prajwal V. Ensemble Feature Selection and Classification of Internet Traffic using XGBoost Classifier. International Journal of Information Technology and Computer Science (IJITCS), 2019, 7: 37-44.