International Journal of Research in Advent Technology, Vol.6, No.5, May 2018
E-ISSN: 2321-9637
Available online at www.ijrat.org

Machine Learning approaches to Keystroke dynamics-based authentication system with Enhanced Features

Gunjan Goswami
Assistant Professor, JSS Academy of Technical Education, Noida, U.P.-201301, India
E-mail:gunjanpahuja.04@gmail.com

Abstract- Most of us use the combination of username and password to authenticate ourselves to the system. Maintaining a consistent password for multiple systems makes it easy for hackers to crack the systems and steal the information. Thus cyber security has become a critical issue these days. A better way to strengthen the password is to combine the passwords with biometrics *(behavioral characteristics)*. Keystroke dynamics-is one of the behavioral biometric that is based on typing rhythm of the individuals. Many features like dwell time, flight time, latency etc. can be used for the analysis of typing rhythm to distinguish the users. Various classifiers exist in the literature that can be used to authenticate the users but it is difficult to compare them because of the inconsistent evaluation conditions. Thus the objective of this paper is to develop a evaluation procedure that can be used to measure and compare the performance of a range of classifiers. For this purpose we have collected the keystroke-dynamics data from 100 users typing 20 passwords each. Three types of passwords i.e. weak, medium and strong varying in strength are used and the strength of passwords is checked by Microsoft Security essentials.

Index Terms- Biometrics Systems; behavioral biometric; keystroke dynamics; pattern recognition; features.

1. INTRODUCTION

Since 1960's Pattern recognition has become more and more popular and is being used in wider areas. Pattern recognition is a collection of mathematical, statistical, heuristic and inductive techniques of fundamental role in executing the tasks like human being on computers. Duda and Hart (2012) defined the pattern recognition as a field concerned with machine recognition focuses on recognition of patterns and regularities in noisy or complex environments. The applications of Pattern Recognition can be found everywhere. Examples include biometrics, computer vision (Guzmán, A., 1968), bioinformatics (Liew, Yan, & Yang, 2005), agriculture geography, engineering and military affairs. In this paper we would delve into the details of biometric technology in the area of cyber security. Biometric technologies are defined as "automated methods of verifying or recognizing the identity of a living person based on physiological characteristics or behavioral characteristics (Miller, 1994). Physiological or static characteristics include fingerprint, palm print, shape of face, retina, iris, pattern of blood veins and are related with the shape of body. On the other hand, behavioral or non static characteristics are related to the pattern of behavior of a person, like signature, voice, typing rhythm, gait etc. Physiological characteristics are considered as constant physical features that owned by a person while behavioral characteristics are the characteristics that are learned or acquired over time.

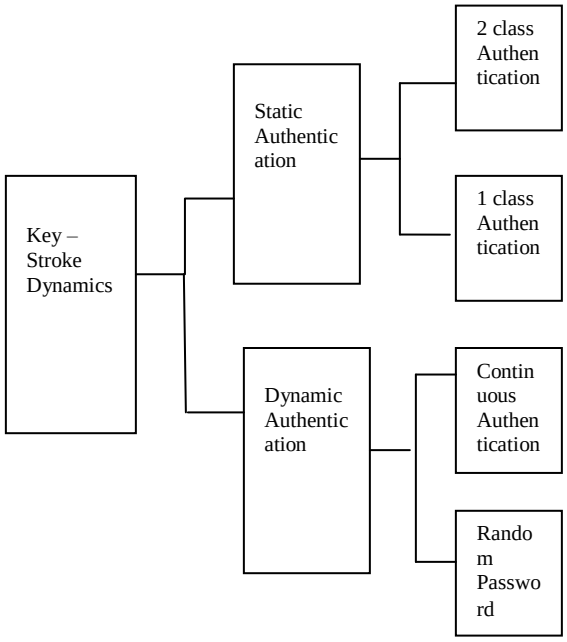

Fig 1: Topology of KD Family

Because biometrics provides highest level of security, thus these technologies are gaining popularity. Also at the time of authentication users don't have to remember or carry anything is the main advantage of biometrics.

Authentication based on personal information like passwords or PINs is the most widely used security mechanism. They are easy to obtain and can be easily stored and have low maintenance.

Keystroke Dynamics is one of the behavioral biometrics authentication methods, based on typing rhythm of a person (Yu & Cho, 2004). The advantage of using KDA is that no additional device is required and thus cheaper than physiological based biometrics systems. Also KD is user friendly, non-invasive and the typing rhythm of the person can't be lost or forgotten. If stolen or lost, the new one can be easily generated (Hwang, Lee, & Cho, 2009). Fig 1 describes the topology of keystroke dynamics family.

2. KEYSTROKE DYNAMICS AS BIOMETRICS

A noticeable amount of work has already done in the field of Biometric authentication. Keystroke dynamics is not concerned with what the user's type but deals with how they type. Keystroke dynamics has been extensively used in authentication systems previously. Collecting data, feature extraction, normalization, feature subset selection and classification are the steps involved in pattern recognition (Duda & Hart, 2012). In keystroke dynamics authentication, the raw data that can be obtained from the user while typing via the keyboard are the key press time and release time. After getting the raw data, various features such as duration, latency or flight time can be extracted.

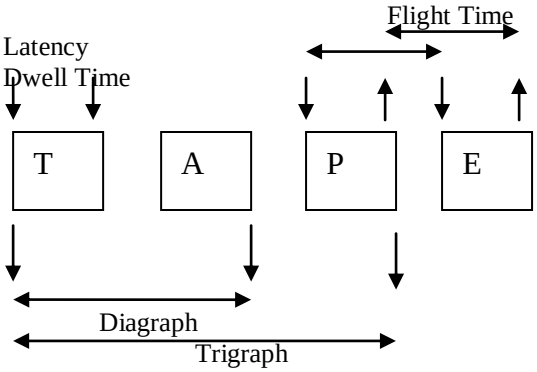

Fig 2: Dwell time, Flight time, Diagraph & Trigraphs

Duration is the amount of time the key is pressed and flight time or Latency is the difference of time between two key actions.

Thus from the raw data, three timing features can be extracted are press-to-press, release-to-release and release-to-press. Other timing information like time it takes to write a word, digraph (two letters) or trigraph (three letters) can also be extracted. After the collection of keystroke timing data, processing is done to get the simple patterns derived from statistics of the features such as mean and standard deviations.

Teh, Yue, & Teoh (2012) incorporated the time interval between a key press and the next key press (D2) and the time interval between a key release and the next key release (D4) along with time interval between a key press and its release (D1) and the time interval between a key release and the next key press (D3). Gaussian Probability Density Function and Direction Similarity Measure Techniques were used to transform the keystroke latency into similarity scores.

Killourhy, K., & Maxion, R. (2008) used dwell time, hold time and flight time and measured the performance of many anomaly detection algorithms on an equal basis. The data was collected from 51 users, 400 repetitions each.

Ahmed & Traore(2014) used the features monographs and digraphs to train ANN, then to predict missing digraphs based on the relation between the monitored keystrokes.

Purgason & Hibler (2012) validated URIEL method for analyzing behavioral biometric data. The method used the timing information (time to transition from one finger to another) while typing and feed-forward neural network was used for the analysis purpose.]

Bours (2012) measured mean and standard deviation of feature values such as key up, key down and latency to evaluate the performance of biometric authentication system. Deutschmann et. al (2013) proved that keystroke dynamics is most appropriate for continuous behavioral biometric. The authors used the fuzzy sets to build the users profiles using the features like keyboard, mouse interactions and Bayesian network were used to compare the profiles with the data. The required training times for users shows that a profile can be trained with only (median) 103 keyboard and 6.606 mouse interactions.

Loy, Lai & Lim(2005) extracted the features such as the fundamental frequency, root mean square, arithmetic mean, energy; peak, noisy and distorted signals, total harmonic distortion and skewness from keystroke pressure. By combining both latency and pressure patterns, it was found EER of the system has been improved.

Robinson et al. (1998) found that the performance of inductive learning approach is the best out of three approaches of classification of typed login signatures viz. minimum intra-class distance, nonlinear and inductive learning approaches. The hold and interkey times were used as features and the authors found that

hold times alone was better than using interkey times alone, and for the MICD (minimum interclass distance) and nonlinear classifiers, hold times alone performed better than both hold and interkey.

Araujo et al.(2005) measured four features like key code, key down and up times and key duration for static user authentication. The authors when evaluated the results found that the best results (FRR=1.45%, FAR= 1.89%) were obtained when all the features were utilized. Schclar et. al (2012) incorporated keystroke dwell time and latency for user authentication based on the keystroke dynamics of the password entry. Rather than using the complete dataset for training, a small subset of users, referred to as representatives, was used along with the password entry keystroke dynamics of the examined user. By doing this the possibility of overfitting gets reduced, while allowing scalability to a high volume of users.

Table1: A Survey of Features used for Biometric Authentication

Authors	**Features selection**	**Features Extraction methods**	**Observations**
Hosseinzadeh et al. (2008)	Up-Up & down-down Keystroke latency	Gaussian Mixture Model + leave-one-out method	EER=4.4%
Wangsuk, & Anusas-amornkul, (2013)	Hold time, interkey time ,latency	C# language (KD on username)	Accuracy=96 %
Hwang, Lee, & Cho(2009)	Artificial rhythms and cues	Hypothesis test	The authentication accuracies improved a lot if very small number of patterns of Artificial rhythms with cues are used for training.
Akila & Kumar (2011)	Latency, duration & diagraph	Mean, median, standard deviation	When compared with other features, the combination of diagraph with median gives the good results.
Monrose, Reiter & Wetzel (2002)	Latency & Duration	Standard Deviation & Mean	For both online and offline text, the system is more efficient than conventional method.

3. PROBLEM STATEMENT

As clear from table 1, that it is unsound to compare the results obtained by using different detectors. The relative performance of the detectors is different because too many factors are involved and they vary from one evaluation to another. Thus lack of generic evaluation method is one of the major drawbacks of such systems. Therefore aim of this research paper is to develop an evaluation procedure by collecting our own keystroke rhythm data and then to measure the performance of a range of classifiers so that the results can be compared on an equal basis. Also the classifier accuracy depends upon number of samples used in a particular study. From the literature it is clear that Performance, satisfaction & security are the three aspects for KD evaluation.

3.1. *Methodology*

In this study, three types of passwords (weak, medium and strong) varying in strength has been collected and analyzed. The strength of the passwords is checked by Microsoft Security essentials.

It is an easy task to type the weak passwords but typing medium and strong passwords is not, because different cognitive activities are involved while typing medium and strong passwords. Strong passwords are the non-routine strings of characteristics that include special characters, lower and upper case letters, either user have used caps lock key or they had just press shift+ etc. It is observed while collecting the data from the users that cognitive load increased while users jump from weak to medium and medium to strong passwords. Thus there is a need to include special characters; upper and lower case letters so that the people became conscious of what is being typed.

When users type the weak passwords that is not a non-routine string of characters, the typing behavior is efficient, fluent and consistent than their behavior while typing strong passwords.

3.2. *Interface for capturing the keystroke data*

Since data collection is the first step of pattern recognition system, therefore we developed the software to collect data and to extract attributes from collected data. The capturing process takes place at the enrollment time and verification time. During the enrollment, the users are required to enter the data several times in order to build the model. On the other hand a single sample is collected during the verification time. From this single sample various features are extracted which can then be compared to biometric model of the pretender. In this study, the users were asked to enter weak, medium, strong strength passwords. Fixed-text passwords were collected using Java-based software. This application was created to record the respective keys and time in milliseconds at which each key was pressed and released. Ms-Excel was used as a database to store the statistical data collected by the application. 100 subjects were identified within the educational institute. 2 data-collection sessions were completed by the subjects (of 10 passwords each), for a total of 20 password-typing samples for weak, 20 for medium and 20 for strong. Thus for one user 60 samples were taken. The users waited for one day between sessions, to capture some of the day to day variations. The subjects selected were from the age group 18-22 years. Since the data was collected from the engineering institute, so the users were having educational background and computer experience. Also some kind of prior knowledge was given to the users so that they become aware of this process. Each piece of password is roughly of same length, the length of weak password is 6 characters, medium password is of 9 characters (consists of both lower and upper case letters) and strong password is of 14 characters (consists of special case characters, lower and upper case letters). A snapshot for data collection software is shown in fig.3.

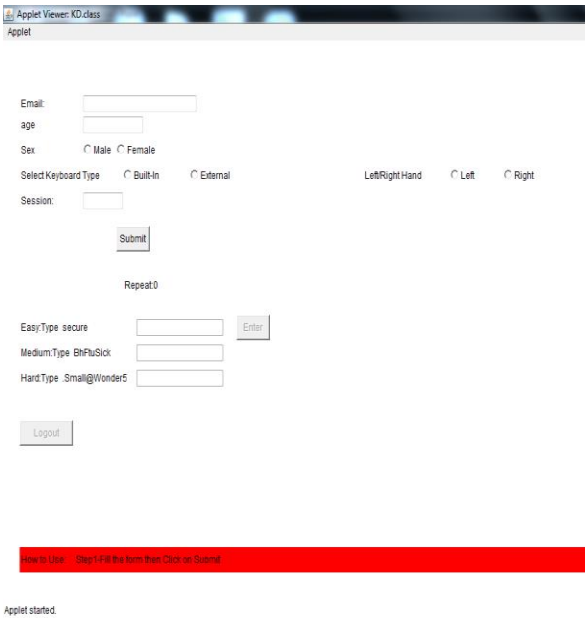

Fig.3: Interface for collecting the keystroke data

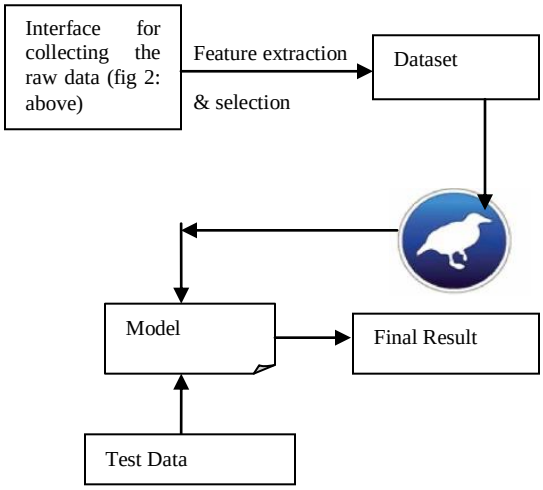

Fig 4: Flow Diagram for training & testing the authenticity.

This software captured the vital data of the users in a log file while they key in the password (i.e. key-down or key-up time). The application displays the password with a text-entry field. Each user was requested to enter data 10 times in a day. Thus 2 days were required by each user to complete the entry process. This ensured data collection under different states of the same user and hence a change of their typing speeds. After entering weak type of password 10 times, user was allowed to enter medium password

and so on. If any errors are detected while entering the password, then the subject is required to retype the password. Thus during the typing of weak, medium, strong type of passwords precise timing data for keystrokes were captured to build a model predicting the authentic passwords.

3.3. *Extracting the attributes of keystroke rhythm*

As it is already said that data collected by the application was stored in MS-Excel, later on these files were used to build the models using Weka. Different researchers extract different combinations of features as clear from the feature selection column of Table 1. Dwell time, flight time, hold time etc. are the features that are commonly used for authentication process. In this research paper, the new features being used are: name, age, gender, LR hand along with dwell time, flight time, hold time, and key that was used by the user to type capital letters or special characters e.g. it may be shift + (left or right) or caps. Thus 26 features from weak password, 38 features from medium strength password and 53 features from strong password are extracted from the data which is entered by various users. As clear from fig 3 that after the data collection, Weka tool is used to suggest most weight carrying attributes out of a collection of attributes.

Attribute selection is done on the basis of searching. Searching is done through all possible combinations of attributes in the data to find out which attributes or collection of attributes will work best for prediction. Thus attribute selection is a two step procedure. First the attribute evaluator i.e. a machine learning algorithm evaluates the attributes and assigns a weight or value to each subset of attributes and second the search method which provides the option of choosing top-down or bottom-up style of searching. We have implemented and evaluated kNN, J48, Naïve Bayes , List may be presented with each item marked by bullets and numbers. Multilayer Perceptron, LWL (Euclidean Distance) & SMO (polykernal) classifiers on the same data set.

3.4. *Performance Metrics for Keystroke Dynamics*

Till Date various classifiers are available to evaluate the behavioral biometrics like Keystroke dynamics, so these models are validated based on security metrics like False Acceptance rate (FAR), False Rejection rate (FRR) and Equal error rate (EER). FAR gives the number of frauds or imposters who are inaccurately allowed as genuine users. On the other hand FRR gives the number of genuine users who are rejected from using the system. Higher FRR is preferred in high security systems. Third security metric which is used as performance parameter is EER which is the ratio of FAR divided by FRR. Lower value of EER signifies a better system.

FAR can be defined as ratio of number of false matches divided by total number of fraud match attempts. FRR is the ratio of number of false rejections divided by total number of genuine match attempts.

4. RESULTS & DISCUSSIONS

After the collection of keystroke data from 104 users, the next step is to find the best features which improve the efficiency of keystroke dynamics system. Feature selection is done by Correlation Attribute, Gain Ratio Attribute, Info Gain Attribute, Symmetrical Uncertainty Attribute evaluation methods. Features are ranked according to the weights of different features. After the feature selection, we tried to implement various classifiers from the pattern recognition literature and results are shown in Table 2 below. The captured dataset has been divided in training and test datasets. The training dataset is used to train and build the models while test datasets are evaluated against these models.

Table 3 (a): Classifiers performance on Weak Password

	Weak Password					
Classifier	KNN	J48	Naïve Bayes	SMO (Polykernal)	Multilayer Perceptron	LWL(Euclidean distance)
Evaluation on test set ===						
Time taken to test model on supplied test set	0.3 sec	0.04 sec	1 sec	8.32 sec	0.47 sec	20.26 sec
Correctly Classified Instances	95.356 %	96.8%	62.3%	90.2 %	95.85%	75.2%
Incorrectly Classified Instances	4.643%	3.15%	37.6%	9.74 %	4.145%	24.78 %
Mean absolute error	0.0021	0.0006	0.0067	0.0172	0.9581	0.0168
Root mean squared error	0.0284	0.0231	0.0756	0.0926	0.002	0.0909
Relative absolute error	12.3 %	3.49%	38.5%	99.1 %	11.7 %	96.88%

Root relative squared error	30.4 %	24. 7%	81. 0%	99.23 %	26.2 3%	97.50%
Coverage of cases	96.6 8%	98. 01 %	69. 32 %	100%	99.8 3%	93.37%

The algorithms that are used in this research paper are: kNN, J48, Naïve Bayes , Multilayer Perceptron, LWL (Euclidean Distance) & SMO (polykernal). The results of different classifiers showing different levels of accuracies are shown in table 3[(a), (b), (c)].

Table 3 (b): Classifiers performance on Medium Password

	Medium Password					
Classifier	KNN	J48	Naïve Bayes	SMO (Polykernal)	Multilayer Perceptron	LWL(Euclidean distance)
Evaluation on test set ===						
Time taken to test model on supplied test set	0.39 sec	0.03 sec	1 seconds	9.24 sec	0.44 sec	44.71 sec
Correctly Classified Instances	100%	100 %	85.65 %	100%	100 %	79.91%
Incorrectly Classified Instances	0%	0 %	14.34 %	0%	0%	20.08%
Mean absolute error	0.001	0	0.0026	0.0175	0.0007	0.0171
Root mean squared error	0.0052	0	0.0445	0.0934	0.0041	0.0916
Relative absolute error	5.57%	0 %	14.60 %	99.15 %	3.96 %	96.70%
Root relative squared error	5.57%	0 %	47.26 %	99.26 %	4.35 %	97.36%
Coverage of cases	100%	100 %	91.80 %	100%	100 %	100%

From the above tables it is clear that the rate for correctly classified instances for medium passwords is more than the weak passwords. The values for other parameters can also be compared from the above mentioned tables. Thus it is clear from table [3a, 3b, 3c] that accuracy for medium and strong passwords is more than weak passwords. Although the use of keystroke dynamics is cheap and does not required any additional hardware, but still the evaluation of KD modality is less in number as compared to other types of modalities such as fingerprint, palmprint modalities. Also very less public databases exists that could be used by the researchers to evaluate the keystroke dynamics authentication systems. From the results presented in table3, it is clear that existing keystroke dynamics methods provide promising results in terms of correctly classifying the claimer and these kinds of systems are well perceived and accepted the users. Trustable Keystroke-Based Authentication known as TOKEN (Nauman & Ali 2010) , Psylock (German company that develops the security solutions based on keystroke dynamics for implementations on different platforms from MS Windows login, to web login, to Citrix and VPN integration), BehavioSec (Swedish company that develops IT security systems based on the integration of keystroke dynamics and mouse dynamics) etc. are some of the examples which are using Keystroke Dynamics for security of the systems. As we are moving towards the digitization, ii is believed that Keystroke Dynamics may be implemented in ATM (Automated Teller Machines) and for e-commerce applications where the users have threat that their passwords may be stolen.

Table 3 (c): Classifiers performance on Strong Password

	Strong Password					
Classifier	KNN	J48	Naïve Bayes	SMO (Polykernal)	Multilayer Perceptron	LWL(Euclidean distance)
=== Evaluation on test set ===						
Time taken to test model on supplied test set	0.53 sec	0.05 sec	1.54 sec	10.59 sec	0.54 sec	169.49 seconds
Correctly Classified Instances	100%	100 %	90.63 %	100%	100 %	81.87%
Incorrectly Classified Instances	0%	0 %	9.37%	0%	0%	18.13%
Mean absolute error	0.0009	0	0.017	0.0167	0.0006	0.0163
Root mean squared error	0.0051	0	0.0382	0.091	0.0037	0.0897
Relative absolute error	5.55%	0 %	9.92%	99.19 %	3.72 %	97.31%
Root relative squared error	5.55%	0 %	41.70 %	99.30 %	4.09 %	97.88%
Coverage of cases	100%	100 %	92.87 %	100%	100 %	100%

5. CONCLUSION

In this study, overview of Keystroke Dynamics has been presented. Also the results have been proved on Desktop application. The future of the same is not only limited to desktop applications but it can be extended to Mobile and Internet applications, because mobile phones and Internet are getting more popularity than the desktop systems. When various applications being run on the Mobile like Internet Banking, E-mail Verification, payment of Bills etc. all require the authentication of username and passwords to verify the identity of a user. Integration of Keystroke Dynamics verification with the existing authentication systems would harden the authentication process. The field of keystroke dynamics is still an emerging field, where most of the challenges need to be overcome in order for it to become an effective biometric.

REFERENCES

[1] Miller, B. (1994): Vital signs of identity [biometrics]. IEEE spectrum, ***31***(2), 22-30.

[2] Guzmán, A. (1968, December) : Decomposition of a visual scene into three-dimensional bodies. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I* (pp. 291-304). ACM.

[3] Monrose, F., Reiter, M. K., & Wetzel, S. (2002): Password hardening based on keystroke dynamics. *International Journal of Information Security*, **1**(2), 69-83.

[4] Adams, A., & Sasse, M. A. (1999): Users are not the enemy. *Communications of the ACM*, **42**(12), 40-46.

[5] Yu, E., & Cho, S. (2004):Keystroke dynamics identity verification—its problems and practical solutions. *Computers & Security*, **23**(5), 428-440.

[6] Liew, A. W. C., Yan, H., & Yang, M. (2005): Pattern recognition techniques for the emerging field of bioinformatics: A review. *Pattern Recognition*, **38**(11), 2055-2073.

[7] Loy, C. C., Lai, W., & Lim, C. (2005): Development of a pressure-based typing biometrics user authentication system. *ASEAN Virtual Instrumentation Applications Contest Submission*.

[8] Araújo, L. C., Sucupira, L. H., Lizarraga, M. G., Ling, L. L., & Yabu-Uti, J. B. T. (2005): User authentication through typing biometrics features. *IEEE transactions on signal processing*, **53**(2), 851-855.

[9] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016): *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[10] Liu Jie, Jigui Sun , and Shengsheng Wang (2006): Pattern recognition: An overview." IJCSNS International Journal of Computer Science and Network Security, **6.6**, 57-61.

[11] Killourhy, K., & Maxion, R. (2008, September): The effect of clock resolution on keystroke dynamics. In *International Workshop on Recent Advances in Intrusion Detection* , **5230**,331-350.

[12] Hosseinzadeh Danoush and Krishnan Sridhar (2008): Gaussian Mixture Modeling of Keystroke Patterns for Biometric Applications. *IEEE transactions on systems, man, and cybernetics—part c: applications and reviews*, **38**(6), 816-826.

[13] Hwang, S. S., Cho, S., & Park, S. (2009): Keystroke dynamics-based authentication for mobile devices. *Computers & Security*, **28**(1-2), 85-93.

[14] Hwang, S. S., Lee, H. J., & Cho, S. (2009): Improving authentication accuracy using artificial rhythms and cues for keystroke dynamics-based authentication. *Expert Systems with Applications*, **36**(7), 10649-10656.

[15] Akila, M., & Kumar, S. S. (2011): Improving feature extraction in keystroke dynamics using optimization techniques and neural network, *Proceedings of the International Conference on Sustainable Energy and Intelligent Systems(SEISCON 2011)*,891-898.

[16] Purgason, B., & Hibler, D. (2012): Security through behavioral biometrics and artificial intelligence. *Procedia Computer Science*, **12**, 398-403.

[17] Bours, P. (2012):Continuous keystroke dynamics: A different perspective towards biometric evaluation. *Information Security Technical Report*, **17**(1-2), 36-43.

[18] Schclar, A., Rokach, L., Abramson, A., & Elovici, Y. (2012):User authentication based on representative users. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **42**(6), 1669-1678.

[19] Duda, R. O., Hart, P. E., & Stork, D. G. (2012):*Pattern classification*. John Wiley & Sons.

[20] Teh, P. S., Yue, S., & Teoh, A. B. (2012): Feature fusion approach on keystroke dynamics efficiency enhancement. *International Journal of Cyber-Security and Digital Forensics (IJCSDF)*, **1**(1), 20-31.

[21] Ahmed, A. A., & Traore, I. (2014): Biometric recognition based on free-text keystroke

dynamics. *IEEE transactions on cybernetics*, **44**(4), 458-472.

[22] Wangsuk, K., & Anusas-amornkul, T. (2013): Trajectory Mining for Keystroke Dynamics Authentication. *Procedia Computer Science*, **24,** 175-183.

[23] Deutschmann, I., Nordström, P., & Nilsson, L. (2013):Continuous authentication using behavioral biometrics. *IT Professional*, **15**(4), 12-15.

[24] Robinson, J. A., Liang, V. W., Chambers, J. M., & MacKenzie, C. L. (1998): Computer user verification using login string keystroke dynamics. *IEEE transactions on systems, man, and cybernetics-part a: systems and humans*, **28**(2), 236-241.

[26] Nauman, M., & Ali, T. (2010, June): Token: Trustable keystroke-based authentication for web-based applications on smartphones. In *International Conference on Information Security and Assurance* , 286-297. Springer, Berlin, Heidelberg.

[27] Wangsuk, K., & Anusas-amornkul, T. (2013): Trajectory Mining for Keystroke Dynamics Authentication. *Procedia Computer Science*, **24**,175-183.

(A.1)