Performance Perspective of Different Classifiers on Different Keystroke Datasets

Soumen Roy[#1], Utpal Roy[*2], D. D. Sinha[#3]

[#1,3]*Department of Computer Science and Engineering*
University of Calcutta, 92 APC Road, Calcutta -700 009, INDIA.
[*2]*Department of Computer & System Sciences,*
Visva-Bharati, santiniketan -731235, INDIA
[1]soumen.roy_2007@yahoo.co.in
[2]roy.utpal@gmail.com
[3]devadatta.sinha@gmail.com

Abstract- According to SplashData (who gathered data from millions of stolen passwords posted online), the top three passwords in the year 2013 are "123456," "password" and "12345678". So we can say most of the people are uninspired while choosing a healthy password because we, as people are still very lazy. It increases the probability of guessing attacks. To minimize these attacks, we pick up some words for password from relatively small dictionary and decorate it by adding extra texts or combine capital, small letter with some symbols. It increases the complexity of password which is very difficult to remember and we forget to distinguish this type of healthy passwords for different access control systems. To solve this problem, here, in this paper, we investigated fixed-text user authentication through keystroke dynamics. Here our typing style is also accounted with the pressed password and user ID. It is established that our typing style is a behavioral biometric characteristic relates the issue of human identification or authentication. But the accuracy level of this technique is not much accepted in practice. In order to realize this technique demands higher level of security with accepted level of error 0.000000…1. Hence, it is highly needed to identify the keystroke dynamics features or combination of features and analyses the accuracy with different classification techniques.

Keywords: Keystroke Dynamics, EER, Behavioral Biometric, Canberra , Chebyshev, Czekanowski, Gower, Intersection, Kulczynski, Lorentzian, Minkowski, Motyka, Ruzicka, Soergel, Sorensen, Wavehedges, Manhattan Distance, Euclidean Distance, Mahanobolis Distance, Z Score, KMean, SVM, NaiveBaysian, ROC Curve.

I. INTRODUCTION

Passwords or PINs are used to recognition the user in knowledge-based user authentication technique. But it is unsafe while all around the areas are covered by video cameras or spy cameras. Basically in Bank is a public place, where pressing password slowly can be traceable by bank customers. If we pick up some words from our small dictionary for password then an attacker may collect our personal information and can check one by one until the actual result is obtained. So there is a probability of Brute force attack or shoulder surfing attack or dictionary attacks. Token-based user authentication scheme uses some physical items called tokens such as smart cards, user ID card, driving licence and some PIN. Here, something we have that is token and something we know that is PIN which can create

problems like, tokens may be stolen or we can forget to carry all the time and to access these token one extra security apparatus is needed, to carry these device to work at anywhere at any time is quite difficult. Among these three popular user authentication techniques the strongest one is biometric user authentication scheme, here biometric human characteristics are recognized such as fingerprint, face, hand geometry, voice print, retina etc. But it also has some shortcomings such as aging problem in face recognition, different picture qualities of different cameras that may affect the way, voice print in crowd. Here one extra device is required to recognize human characteristics such as camera in face recognition, microphone in voice recognition etc. In our system we used keystroke dynamics.

System takes comprising of characters as well as the typing style of each subsequent character entered. It facilitates that no one can track the time or presses the character of password in same rhythm. It will prevent our system from off-line guessing attracts and also prevent to track by un-authorised people. Our objective is to minimize the probability of off-line guessing attacks, hide the password from public, minimize the hardware cost and minimize the software cost by making faster pattern recognition.

Keystroke Dynamics as biometrics characteristics is not a new one. First time, in the year 1897, Bryan and Harter investigated keystroke dynamics. In 1975, Spillane described the concept of keystroke dynamics and suggested in an IBM technical bulletin that typing rhythms might be used for identifying the user at a computer keyboard. Forsen et al. in 1977 conducted preliminary tests of whether keystroke dynamics could be used to distinguish typists. Gaines et al. in 1980 produced an extensive report of their investigation with seven typists into keystroke dynamics. After then S. Bleha submitted his PhD thesis on Recognition system based on keystroke dynamics in 1988 [1]. R. Joyce and G. Gupta proposed an identity authentication based on keystroke latencies in 1990 [2]. F. Monrose et al. [3] proposed keystroke dynamic as a biometric for authentication in 2000. Different online and offline applications already have been done by fixed text and free text keystroke dynamics. Keystroke dynamics research has been going on for the more than thirty three years. Many methods have been proposed during that time. Methods based on traditional statistics-such as mean times and their standard deviations are common. Over the years, different pattern recognition methods have come into vogue and been applied to keystroke dynamics; neural networks, Fuzzy logic and support vector machines among others. Many classification algorithms have been proposed and many databases are available in the Internet. In evaluation process of different classifiers on different databases, we have obtained different average Equal Error Rates (EERs) because selecting the string for the database and considering the features for classification affect the error rate. It has been established that our typing styles are similar for the common daily used words (name, address, e-mail ID etc.). In this connection we have chosen the daily used words to train the system.

We have collected press and release time of 12096 keystrokes of 1440 samples of patterns from 12 different individuals in 4 different sessions with minimum of one month interval for five different common words (“kolkata123”, ”facebook”, ”gmail.com”, ”yahoo.com”, ”123456”) in our experiment. Then we have considered all 8 different features and combination of features then we have executed 22 different classification models on that collected data. In our observation we got 1.9% of EER for the classifier Lorentzian by taking all the features in our consideration. In second position OutlierCount classifier given 2.3% of EER when we have taken in our consideration all the features and all 4 strings

("kolkata123","facebook", "gmail.com", "yahoo.com"). So the adaptation of keystroke dynamics technique in any existing system increases the security level upto 97.6% to 98.7%.

II. KEYSTROKE DYNAMICS

A. *Basic Idea*

Keystroke dynamics is a behavioral biometrics which is the method of analysing the way a user types on a keyboard and classify him based on his regular typing rhythm. It is the study of whether people can be well-known by their typing rhythms, much like handwriting is used to recognize the author of a written text. A user's typing pattern may be unique because similar neuro-physiological factors that make written signatures unique.

B. *Science and Features Selection*

Our typing style can be easily calculated by simple program which can calculate key pressing and releasing time of each key and then generates key-hold time and key latency times. Let K_i represent entered character set and P_i and R_i represent the corresponding key press and key release time where $1 \leq i \leq$ length of the entered word. The features of the keystroke dynamics as follows:

Key Duration (T1)=R_i-P_i (1)
Up Up Key Latency (T2)=R_{i+1}-R_i (2)
Down Down Key Latency (T3)=P_{i+1}- P_i (3)
Up Down Key Latency (T4)=P_{i+1}-Ri (4)
Down Up Key Latency (T5)=R_{i+1}- P_i (5)
Total Time Key Latency (T6)=R_n-P_1 (6)
Tri-graph Latency (T7)=R_{i+2}- P_i (7)
Four-graph Latency (T8)=R_{i+3}- P_i (8)

Some new features also can be calculated like key pressure (Pressure sensitive keyboard is require), finger tips size (Touch screen keypad is needed), finger placement on keyboard (Camera is needed), keystroke sound (microphone is needed), error correcting mechanism, sequence of left-right control keys.

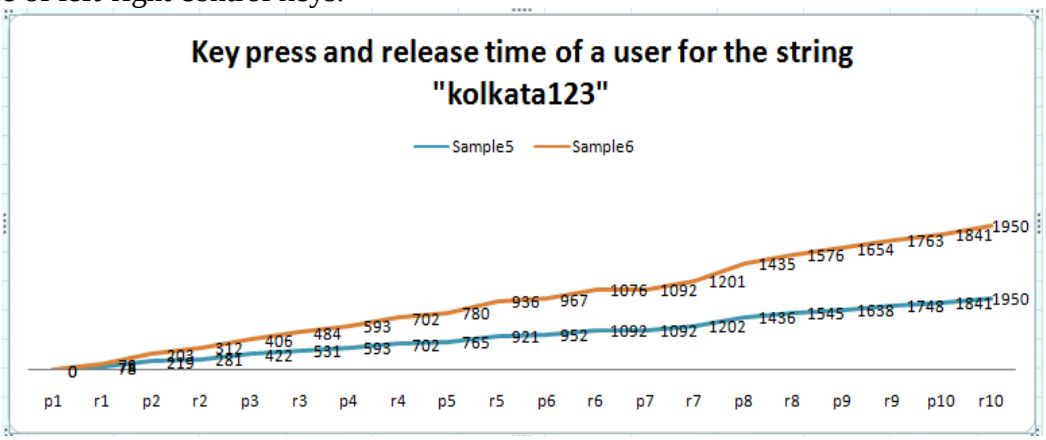

Fig. 1. Two different samples from a user in line chart

C. *Keystroke Dynamics as User Authentication*

There are different ways in which a user can be authenticated. However all of these ways can be categorized into one of three classes: "Something we know" e.g. password, "Something we have" e.g. token, "Something we are" e.g. biometric property.

The keystroke Dynamics characteristics is the behavioural biometric characteristics what we have learned in our life not what the properties we are born with which is the good human characteristics that can be used to distinguish people.

D. *Security Issues*

Among various user authentication techniques knowledge-based, token-based and biometric-based authentication techniques, biometric authentication is most popular for their uniqueness characteristics and cannot be stolen or there is no chance to loss. Keystroke Dynamics is a behavioral characteristic which is unique and can be effectively implemented with the existing system with minimal alternation. It can be used as a safe guard of our password from different type of attacks.

E. *Factors Affecting Performance*

Some of the factors which affect the way of keystroke Dynamics as follows: Text length, sequences of character types, word choice, and number of training sample, statistical method to create template, mental state of the user, tiredness or level of comfort, keyboard type, keyboard position and height of the keyboard, hand injury, weakness of hand mussel, shoulder pain, education level, computer knowledge, and category of users.

F. *Algorithms*

In this section, we have defined all the models, where P refers to the training set and Q refers to the test set. Mean and standard deviation is represented by μ and α respectively.

Canberra:

$$D_{car}=\sum_i^n \frac{|P_i-Q_i|}{P_i+Q_i} \quad (9)$$

Chebyshev:

$$D_{cheb}=\sum_i^n \max\ |P_i-Q_i| \quad (10)$$

Czekanowski:

$$D_{cze}=\frac{\sum_i^n\ |P_i-Q_i|}{\sum_i^n\ (P_i+Q_i)} \quad (11)$$

Gower:

$$D_{gow}=\frac{1}{n}\sum_i^n |P_i-Q_i| \quad (12)$$

Intersection:

$$D_{ins}=\frac{1}{2}\sum_i^n |P_i-Q_i| \quad (13)$$

Kulczynski:

$$D_{kuld}=\frac{\sum_i^n\ |P_i-Q_i|}{\sum_i^n\ min(P_i,Q_i)} \quad (14)$$

Lorentzian:

$$D_{lor}=\sum_i^n \ln\ (1+|P_i-Q_i|) \quad (15)$$

Minkowski:

$$D_{mink}=\sqrt[p]{\sum_i^n |P_i-Q_i|^p} \quad (16)$$

Motyka:

$$D_{mot}=\frac{\sum_i^n\ \max\ (P_i,)|}{\sum_i^n\ (P_i+Q_i)} \quad (17)$$

Ruzicka:

$$D_{ruz}=1-\frac{\sum_i^n\ min(P_i,Q_i)}{\sum_i^n\ max(P_i,Q_i)} \quad (18)$$

Soergel:

$$D_{soe}=\frac{\sum_i^n |P_i-Q_i|}{\sum_i^n max(P_i,Q_i)} \quad (18)$$

Sorensen:

$$D_{sor}=\frac{\sum_i^n |P_i-Q_i|}{\sum_i^n (P_i+Q_i)} \quad (19)$$

Wavehedges:

$$D_{wv}=\frac{\sum_i^n |P_i-Q_i|}{\sum_i^n max(P_i,Q_i)} \quad (20)$$

Manhattan Distance:

$$M=\sum_{i=1}^{n}(|P_i - Q_i|) \quad (21)$$

Euclidean Distance:

$$E = \sqrt[2]{\sum_i^n(|P_i - Q_i|)^2} \quad (22)$$

Mahanobolis Distance:

$$Eh = \sqrt[2]{\sum_i^n((|P_i - Q_i|)/\alpha_i)^2} \quad (23)$$

Z Score:

$$Z=\sum_{i=1}^{n}(|P_i| - \mu(|Q_i|))/\alpha_i \quad (24)$$

KMean:

It uses the k-means clustering algorithm to identify clusters in the training vectors, and then calculates whether the test vector is close to any of the clusters. In the training phase, the detector simply runs the k-means algorithm on the training data (with k = 3). The algorithm produces three centroids such that each training vector should be close to at least one of the three centroids. In the test phase, the anomaly score is calculated as the Euclidean distance between the test vector and the nearest of these centroids.

SVM:

It incorporates an algorithm called a support-vector machine (SVM) that projects two classes of data into a high dimensional space and finds a linear separator between the two classes. A "one-class" SVM variant was developed for anomaly detection. It projects the data from single class and finds a separator between the projection and the origin. In the training phase, the detector builds a one-class SVM using the training vectors. In the test phase, the test vector is projected into the same high-dimensional Space and the (signed) distance from the linear separator are calculated. The anomaly score is calculated as this distance, with the sign inverted, so that positive scores are separated from the data. The SVM parameter ν was set to 0.5; the source study set ν with a "heuristic search" but did not elaborate.

NaiveBaysian:

The normal naive Bayes assumes independence of the predictor variables, and Gaussian distribution of metric forecasters.

G. *Types*

Generally the version of keystroke dynamics are free text keystroke dynamics (like paragraph of written text) and fixed text keystroke dynamics (like signature of written text).

H. *Advantages*

It is a low implementation cost, very simple, strong as biometric characteristics, continuous monitoring method.

I. *Disadvantages*

External factors such as injury, fatigue, type of keyboard, position of keyboard are affected the system. Typing pattern of a human may vary during a day or between two days which is depends on mental state of the user. More data set is required to train the system. It takes huge time.

J. *Application Areas*

This technique can be effectively applied in application areas such as student or employee attendance system, distance based examination, password recovery mechanism, emotion recognition, private data encryption, continuous user verification, criminal investigation, identifying backdoor accounts, free-text user authentication etc.

III. EXPERIMENTAL SETUP

We have implemented a program in JAVA Applet to collect the raw data, which has the capability of capturing all key pressing and releasing events, which are used to create the database of different sample of passwords and timing templates. It can also calculate different score or distance between vectors.

We have collected press and release time of 12096 keystrokes of 1440 samples of patterns from 12 different individuals in 4 different sessions with minimum of one month interval for five different common words (“kolkata123”, ”facebook”, ”gmail.com”, ”yahoo.com”, ”123456”) in our experiment. Then we have considered all 8 different features and combination of features then we have executed 22 different classification models on that collected data.

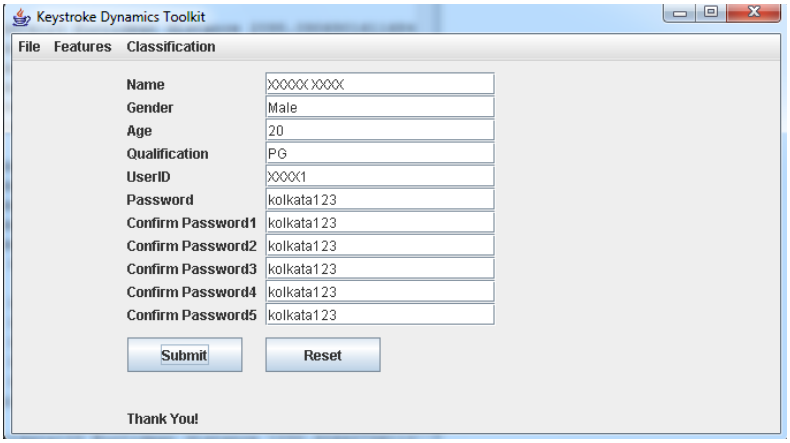

Fig. 2. Simulator to collect the raw data

IV. EVALUATION AND ANALYSIS

We have implemented 22 different classification models in R statistical programming language and evaluated on different keystroke databases and we got the average equal error rate represented in the following table no. 1.

TABLE I
(AVERAGE EQUAL ERROR RATE FOR ALL CLASSIFICATION MODELS)

Models	All strings	kolkata123	facebook	gmail.com	yahoo.com	123456
Canberra	0.037	0.089	0.108	0.125	0.129	0.152
Chebyshev	0.125	0.135	0.159	0.162	0.200	0.192
Czekanowski	0.114	0.144	0.146	0.159	0.193	0.185
Gower	0.514	0.526	0.533	0.538	0.533	0.499
Intersection	0.602	0.637	0.540	0.547	0.555	0.591
Kulczynski	0.114	0.144	0.146	0.159	0.193	0.185
Kulczynskis	0.114	0.144	0.146	0.159	0.193	0.185
Lorentzian	**0.019**	**0.091**	**0.099**	**0.106**	**0.131**	**0.164**
Minkowski	0.240	0.188	0.191	0.229	0.214	0.205
Motyka	0.114	0.144	0.146	0.159	0.193	0.185
Ruzicka	0.114	0.144	0.146	0.159	0.193	0.185
Soergel	0.114	0.144	0.146	0.159	0.193	0.185
Sorensen	0.113	0.144	0.146	0.159	0.193	0.185
Wavehedges	0.114	0.144	0.146	0.159	0.193	0.185
Euclidean	0.228	0.160	0.189	0.191	0.224	0.202
Manhattan	0.125	0.135	0.159	0.162	0.200	0.192
ScaledManhattan	0.089	0.118	0.098	0.119	0.147	0.176
OutlierCount	0.023	0.098	0.100	0.134	0.121	0.163
Mahalanobis	0.262	0.154	0.163	0.163	0.268	0.317
KMeans	0.184	0.152	0.132	0.139	0.170	0.181
SVM	0.182	0.150	0.114	0.145	0.160	0.159
naiveBayes	0.160	0.146	0.215	0.243	0.250	0.340

Here we got 1.9% of EER for the model Lorentzian, which is the minimum EER among all the models.

V. EVALUATION AND ANALYSIS

After getting all the EER for different classification models, we analysed the results by Weka Software and we observed that all the models are suitable to recognise the keystroke pattern except 2 to 4 classification models for different pattern of string, but Lorentzian model is the better where we got 98.1% of accuracy.

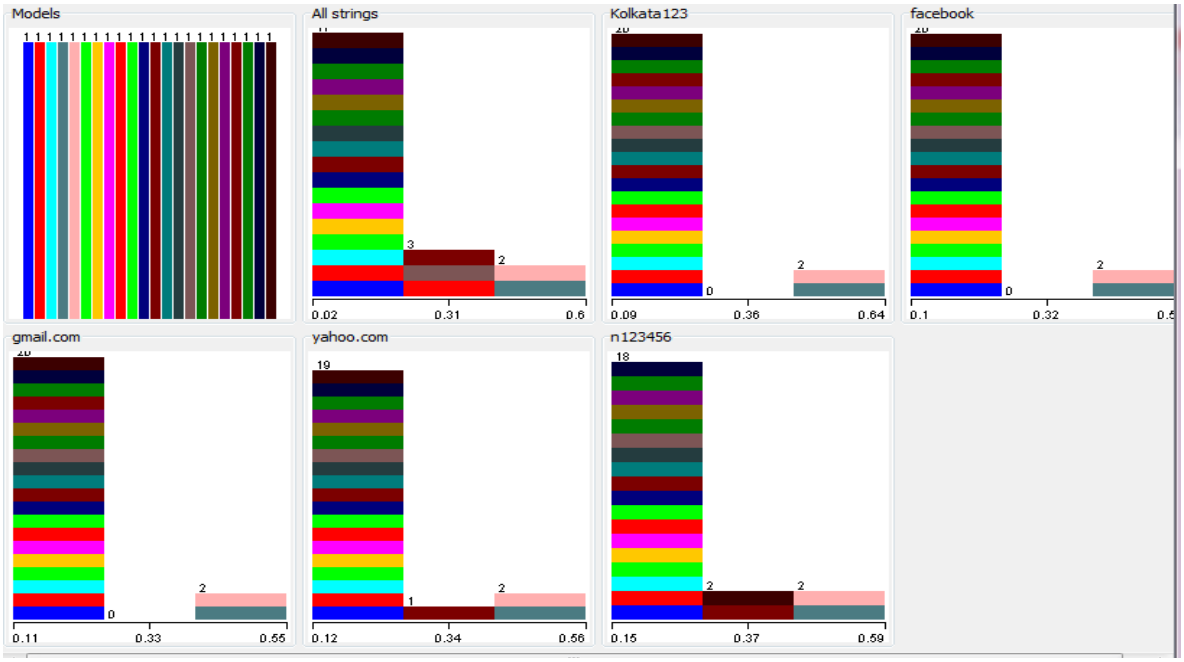

Fig. 3. Histogram of average EER of different pattern of strings

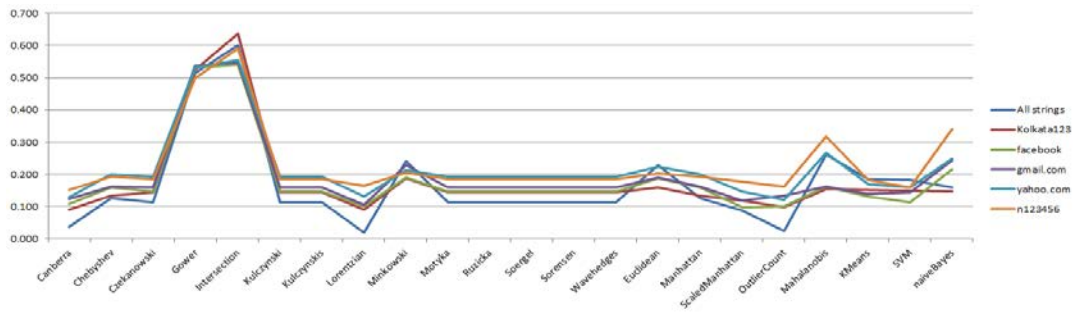

Fig. 4. Line chart of different pattern of strings

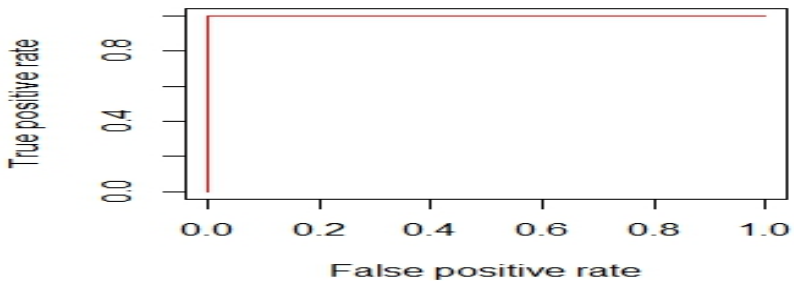

Fig. 5. ROC curve of Lorentzian model for all strings

Keyboard is essential for a computer device, which can recognize our typing style and very much unique as per our experiment and cannot be copied or stolen. It can be used as safe guard of our password in any access control system. This technique can be used in online criminal investigation, back door account identification, online typing examination, emotion recognition, lab attendant system many more.

Different type of keyboard such as keypad, desktop keyboard or standard keyboard, and screen touch keypad may affect the way of keystroke dynamics. Basically keypad is not changing frequently in mobile phone. So this technique can be effective for mobile security as per Trojahn, M. and Ortmeier, F. otherwise artificial keystroke dynamics or keystroke sound implemented by Roth, J. et al. and Metaxas D. would be introduced.

Characteristics of human may change over time. So update mechanism is needed to update template after acceptable verification or identification. This technique can be effectively applied in application areas such as student or employee attendance system, distance based examination, password recovery mechanism, emotion recognition by Kolakowska, A., private data encryption, continuous user verification, gender identification, criminal investigation, identifying backdoor accounts, free-text user authentication etc.

Sometimes, score of different algorithms varies. It would be better if we combined all scores in a single equation like mean value calculation with given weights of all scores.

VI. CONCLUSION

We have implemented and evaluated 22 different classification models on 5 similar keystroke database taking in our consideration all 8 features and subset of features and also we have taken all five string as a keystroke database. In our evaluation process, we have identified the best model is Lorentzian. It achieved 98.1% of accuracy for all 5 strings. Z-score classification model given the second highest accuracy upto 97.6% for all the string patterns. We also have tested this algorithm on the entire strings database separately and we obtained impressive results. So it has been established that Lorentzian is the best model on keystroke dynamics database, so keystroke dynamics can be effectively implemented in any existing knowledge-based user authentication technique.

ACKNOWLEDGEMENT

Authors acknowledge Mr. Champak Chakraborty, Department of Physics, Bagnan College for reading the manuscript carefully.

REFERENCES

[1] Bleha, S. et al. (1990). Computer-access security systems using keystroke dynamics. IEEE Transactions on Pattern Analysis and Machine Intelligence 12, 1217–1222.
[2] Joyce, R. & Gupta, G. (1990). Identity authorization based on keystroke latencies. Communication of ACM 33 (2) 168–176.
[3] Monrose, F. & Rubin, A. D. (2000). Keystroke dynamics as a biometric for authentication. Future Generation Computer Systems, Vol. 16, No. 4, pp. 351–359.
[4] Killourhy, K. S. (2012). A Scientific Understanding of Keystroke Dynamics. PhD thesis, Computer Science Department, Carnegie Mellon University, Pittsburgh, US.
[5] E. Yu and S. Cho, “Novelty detection approach for keystroke dynamics identity verification,” in Intelligent Data Engineering and Automated Learning, vol. 2690, pp. 1016–1023, Springer, Berlin, Germany, 2003.

[6] P. Kang, S. S. Hwang, and S. Cho, “Continual retraining of keystroke dynamics based authenticator,” in Advances in Biometrics, Proceedings, vol. 4642, pp. 1203–1211, Springer, Berlin, Germany, 2007.
[7] S. Haider, A. Abbas, and A. K. Zaidi, “A multi-technique approach for user identification through keystroke dynamics,” in Proceedings of the 2000 IEEE Interantional Conference on Systems, Man and Cybernetics, vol. 2, pp. 1336–1341, October 2000.
[8] R.Giot, M. El-Abed, andC. Rosenberger, “GREYCkeystroke: a benchmark for keystroke dynamics biometric systems,” in Proceedings of the IEEE 3rd International Conference on Biometrics: Theory, Applications and Systems (BTAS ’09), pp. 1–6, September 2009.
[9] Pin Shen Teh, Andrew Beng Jin Teoh, and Shigang Yue, A Survey of Keystroke Dynamics Biometrics, The ScientificWorld Journal Volume 2013, Article ID 408280, 24 pages
[10] S. Roy, U. Roy, D.D. Sinha, “Enhanced Knowledge-Based User Authentication Technique Via Keystroke Dynamics”, International Journal of Engineering and Science Invention (IJESI), Vol 3, Issue 9, Sep, 2013, 41-48.