Adaptive Fuzzy Chaotic Genetic Clustering Based Continuous Keystroke Authentication

M. Rathi, A. V. Senthil Kumar, Ismail Musirin

Abstract: Exponential growth in technology, has increased the security breaches day today. This is because the system which is standalone or connected with networks, are handled by different users for various purposes. The security of individuals is preserved using identification and authentication to prevent from intruders. The authentication of an individual is done either by using behavioral or physiological characteristics of the concern user. This paper aims to develop an optimized approach based on ballistic nature of typing behavior based authentication system. This system is used to identify an individual by their typing rhythm to confirm their genuineness. Determining authenticated persons by keystroke dynamics is very difficult in presence of uncertainty in their typing rhythm. This proposed model devised a fuzzy inference model with gene clustering to discover the pattern of the user as genuine or an intruder. There is no proper proof of handling indeterminacy in impersonate users as authenticated or not by discovering the keystroke pattern. Hence, this proposed work handles the indeterminacy of user keystroke recondition by applying membership degree with the obtained features involved in behavioral keystroke typing rhythm based authentication model

Index Terms: continuous keystroke dynamic, behavioral, typing rhythm, indeterminacy, Fuzzy inference model, Gene clustering, Buffalo Dataset

1. INTRODUCTION

DUE to rapid increase in demand of strong security mechanism, conventional methods fail to face the challenges due to tokens and passwords are too many to evoke. One of the important issues in a restricted access of using computer system in remote is user authentication. Continuous spread of internet usage makes effectual remote authentication a key issue providing additional difficulties. Many biometric authentication models need dedicated hardware which was unhandy for remote applications [1]. Related with other biometrics, one of the emerging and attractive user-friendly biometric mechanisms is keystroke authentication. The dynamic data of keystroke can be gathered without disturbing the activities of the corresponding user. The keystroke dynamics is mainly used for recognition of an authenticated user by their typing rhythm. While a keystroke pattern sequence successfully matches a user input, then it spots the user as an authenticated person or if it is mismatched the user is treated as an intruder. An approach which uses rhythm of typing as a pattern of biometric authentication is referred as continuous keystroke authentication. This kind of authentication not only checks the value of the password but also the typing rhythm. In addition, after the successful initial log, the system does not assume that the user changes during a session, when a user fails to log out after completing his work, or leave away for short or long period of time. This situation easily allows the impostor to access the documents, delete the content or send mail as a genuine user. To overcome this kind of problem the necessity of continuous authentication is considered as a primary authentication tool in the field of security mechanism. The major difference between the static keystroke dynamics and the continuous keystroke dynamics is that in the former static method, the typed information used for authentication is fixed, while in latter the information is never fixed [2]. The main requirement of the continuous authentication is as follows:

- During continuous authentication the user is not interrupted in their daily activities
- The system utilizes each single keystroke to discover the genuineness of the user

This paper introduced the concept of uncertainty in determining the continuous keystroke pattern when there is a high degree of similarity in typing rhythm among the normal user and the impostor. This paper introduces genetic clustering based continuous keystroke pattern recognition in an optimized way.

- *Department of Computer Technology, Dr. NGP Arts and Science College, Coimbatore, India. e-mail: rathi.vidu@gmail.com*
- *Department of MCA, Hindusthan College of Arts and Science, Coimbatore, India. e-mail: avsenthilkumar@yahoo.com,*
- *Faculty of Electrical Engineering, University Teknologi MARA, Selangor, Malaysia. e-mail: ismailbm@uitm.edu.my*

2 RELATED WORK

This section discusses about some of the existing works related to keystroke dynamics and the authentication process. Dowland et al. [3] developed a digraph, word latency and tri graph as features and for classification they used distance-based classifier for dynamic keystroke authentication over 35 users. Gunetti et al. [4] in their work, to perform keystroke dynamic authentication they used the digraph latency for extraction of features and they also used distance-based classifier to classify the users as legitimate users or imposter among 205 users. Stewart et al. [5] devised a burst authentication. Their main motive is to use the technique of burst authentication to decrease the frequency of sovereign checks of authentication. This model owns the merit of decreasing false alarm rate, evades capturing of huge volume of irrelevant data and unnecessary usage of resources to process the selected input, whilst it offers sufficient for continual biometric authentication training. The feature extraction is done on stylometry and keystroke time information along with KNN classifier and the nearest neighbours are discovered using Euclidean distance.

Messerman et al. [6] developed a non-intrusive authentication

scheme which performs continuous verification method and repetitively confirms the user individuality with free text dependent keystroke dynamics. This method is an extension of the previous work of [4], where web based applications is the target which needed the process of decision making system. They highlighted the main issues like scalability, human behavior and time of response. They tested the methodology with 55 users and achieved satisfactory outcome. Ordal et al. [7] utilized the duration of keypress and digraph latency as a feature similarity model. In this work, they attained 150 keystroke segments and used for analyzing the continuous authentication process. Ahmed et al. [8] in their work, used digraph latency and duration of key press as features for analyzing the continuous keystroke authentication using Artificial Neural Network with 53 users. Unlike other biometric authentication model, the keystrokes timing information are gathered using software alone without any special hardware or embedded techniques. Keystroke dynamics is more efficient in terms of cost effectiveness, user friendly and modality of continuous user authentication. Alshehri et al. [9] designed an approach of iterative real time keystroke continuous authentication analyzing the typing behavior as a time series based which avoided the demerit of feature vector extraction. Rathi and Senthil Kumar [10] in their work used Buffalo dataset for User Profiling Similarity Measurement and Euler Movement Firefly Algorithm to recognize the time slice of the users. They [11] also worked on similar problem with different application for Keystroke Dynamics and discovering Fixed-Text and Free-Text based validation system [12]. Lu et al. [13] developed a model which split the data of user keystroke with fixed length sequence and converting the sequence of keystroke into the vector sequence conferring with keystroke time feature. The sequence of the keystroke is learnt by developing a recursive neural network in convolutional neural network for discovering uniqueness of authentication. Yan Sun et al. [14] developed a Gaussian mixture model based clustering, which uses buffalo dataset for determining the authentication using continuous keystroke. Vural et al. [15] used a new data set with the algorithm developed by Gunetti et al. [16] to discover whether the two instances of sample belong to same person or not. Murphy et al. [17] in their work gathered a large voluminous dataset of free text keystroke, which recorded the key operation, mouse operation and software activities for two and half years. They used Gunetti and Picardi algorithm for classifying the keystroke information as authenticated user or imposter. Jiaju Huang et al. [18] introduced a kernel density estimation method which computed the distance among the training and testing samples by the means of probability density. This is to discover the authenticity of individuals based on their typing rhythm. The developed model is tested with various datasets. Shimshon et al. [19] stated that the window length effect on continuous verification and the model achieved promising result.

3 BACKGROUND STUDY

3.1 Genetic Algorithms (GA)

Genetic Algorithms (GAs) belong to the category of optimization algorithm which involves in adaptive search to discover solutions to large scale optimization issues with many local optima [23]. Standard GA uses its chromosomes to define candidate solution. Every chromosome denotes a portion or the complete solution. Through evolution process, Genetic algorithm determines the best solution by managing the set of chromosomes during searching process. During each generation, with roulette wheel selection or tournament methods parents are selected and they involve in creating new offspring with operation of crossover and mutation. The worst candidates in the present generation are replaced by the newly constructed ones. The crossover and mutation support to overcome local optimum. But Genetic Algorithm often suffers from early convergence and the selection of initial chromosomes are done at a random manner and during each generation only the offspring of the initially selected parents are used for searching process. The remaining chromosomes of the original population are entirely avoided which may gives us more useful information. To overwhelm these two issues this proposed model introduce a chaotic genetic algorithm for providing improved results in continuous keystroke authentication system and the working model is shown in Fig. 1.

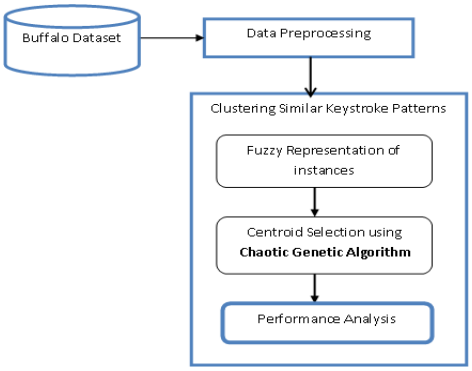

***Fig. 1**: Overall Framework of Adaptive Fuzzy Chaotic Genetic Clustering (AFCGC) for Continuous Keystroke Authentication*

This work devised an adaptive fuzzy clustering algorithm which combines the GA and fuzzy algorithm with chaotic theory-based cluster evaluation method to overcome the drawbacks of existing methods that use the GA and FCM algorithms. The clustering process is carried out by performing fuzzy representation of datasets and the chaotic genetic algorithm is used for clustering the pattern of keystrokes.

3.2 Dataset Description

The performance of the developed Adaptive Fuzzy Chaotic Genetic Clustering is done on the Buffalo dataset. This dataset is gathered by researchers at SUNY Buffalo [21], [14]. This dataset is comprised by both fixed and free keystrokes. This dataset consists of 148 users with a total of 2.14m keystrokes in three different sessions. The authors have used four different kinds of keyboards. These samples are useful for performing continuous keystroke authentication.

3.3 Data pre-processing of Keystroke Dataset

During data pre-processing, only the events which related to the keyboard usages are considered. The keyboard has been grouped based on the standard QWERTY scheme. Each event of the keyboard is stored in the ensuing ithrow of the selected input file with the vector wti as follows:

$$wti = \{prfx,tpi,id\}$$

Where prfx is an event type $prfx \in \{Ku, Kd\}$ Ku- key up and Kd- key down, tpi- event timestamp, id refers to key identifier

The text data file comprised of set of vectors wti is transformed into a vector set VCid which signifies the dependency of time among keyboard events. The file is searched for identical identifiers of two successive pair of rows and such pair is rehabilitated into vector VCid relevant to the following equation

$$\begin{matrix} wt_i = [K^u, t_i, id] \\ wt_j = [K^d, t_j, id] \end{matrix} \quad \rightarrow Vc_{id} = [t_i, t_j] i < j$$

The vectors of same type must be present in the file an even number of times. Else, the vector for which the pair was not found is measured as artefact and it will be eliminated

4 CLUSTERING SIMILAR KEYSTROKES USING FUZZY ADAPTIVE CHAOTIC GENETIC ALGORITHM

4.1 Fuzzy C-Means Clustering

Fuzzy C-Means (FCM) is an extension of standard K-Means clustering [20]. In K-Means each data belongs to one of the clusters but in FCM clustering model it allows each data to belong to more than one cluster and they are denoted using degree of membership. Discovering the pattern structure and effort to detect and quantify the non-random inaccuracy are well-treated using this clustering model. The objective function used for FCM algorithm is as follows:

$$J(K, m) = \rangle \ \rangle \ (\mathfrak{U}_{k_i})^{m} dst^2(x_i cn_k) \qquad (1)$$

where K is the number of Cluster, m is the parameter, $\mathfrak{U}_{k_i}$ is the membership degree of x_i in Cluster K, $dst^2(x_i cn_k)$ is the distance from x_i controls cn_k. The parameters in this equation are the centroid vector cn_k and the components of the membership vector $\mathfrak{U}_{k_i}$. By evaluating $\mathfrak{U}_{k_i}$, it shows the belonging ratio to a cluster k and centroid cn_k belonging expression cluster k.

To handle ambiguity in detection of similar entries in continuous keystroke analysis, fuzzy based representation is used in this work. In this methodology, each entry, say an instance or record is considered to belong to one or many clusters with varying membership degrees. This overcomes the problem, when a single instance has similarity on both clusters.

It is essential to adapt the concept of fuzziness when an instance lies in border of both the clusters. In real life applications precise representation of cluster belongingness is done not feasible, for instances lying in boundaries, these instances are denoted by terms of threshold with binary value to indicate whether an instance certainly belongs to a specific cluster or not, a membership degree is computed by fuzzy for each cluster. The output of the fuzzy membership value lies between 0 and 1. Fig. 2 shows the representation of fuzzy membership value for a solitary character, Standard Deviation values (σ) computed throughout training phase, and are attuned by fine-tuning variable γ

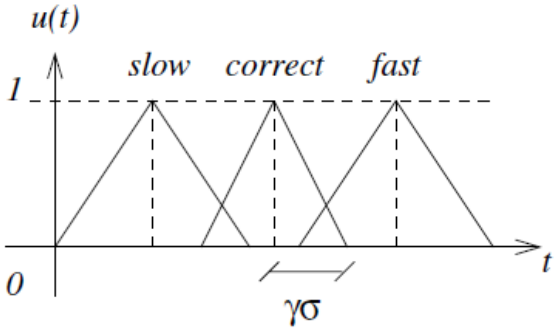

***Fig. 2**: Representation of Fuzzy functions, Standard deviation and tuning variable*

4.2 Algorithm for Fuzzy C-Means based Continuous Keystroke Clustering

Steps:

1: With selected keystroke input dataset Ks= (ks1,ks2,ks3,...,ksn), initialize the number of clusters Cs Є (2,3,..n-1),

2: During each iteration L, compute the Cs means vectors Vi, with $\mathfrak{U}_{ks_q}$, the proximity function of ksi with respect to the kth cluster

$$V_i = \frac{\sum_{q=1}^{n} u_{iq}^{m} ks_q}{\sum_{q=-1}^{n} u_{qj}^{m}}$$

3. Update the cluster centroids

$$U_{iq} = \frac{1}{\sum_{l-1}^{n} \left(\frac{d(ks_q, c_i)}{d(ks_q, c_l)} \right)^{2/(m-1)}}$$

4. If |U(L+1)– UL| < e, where e is the error, stop; else L= L+1 and go to Step 2.

4.3 Procedure for AFCGC Continuous Keystroke Authentication

One of the significant studies in mathematic field is chaos theory, which has more impact during initial condition and the action is referred as butterfly effect. It means if there is any minute modification at an initial stage then it faces a huge variation in their later stages. In Genetic Algorithm, the initial population of genes are chosen in a random manner [22]. This may lead to failure of choosing optimal population which influences the best solution accuracy. The arbitrary process of selecting populations may not guarantee the uniform dispersal in keystroke authentication. It also has a consequence of early convergences because of local optima. To overwhelm this controversy, this proposed work used chaos mapping which handles the unpredictable behavior of keystroke authentication and mapping strategy in an optimized way. It avoids the earlier convergence by selecting the population with potential searching capability. Each entry in the file is represented as a matrix of cluster centers in Fuzzy C-Means clustering. The initial population is chosen by applying chaotic mapping strategy. During each iteration of clustering, the chaotic genetic clustering selects some of the members of the population for reproduction and it is accomplished by using cross over and mutation.

Selection

Roulette wheel selection is used for choosing the population to

reproduce. Based on their fitness value, each member in the population is selected, with high probability of maximum function value. Two population members are selected for reproduction by using roulette wheel twice. Then crossover and mutation are performed on these selected members.

Crossover

Each entry feature V_{ij} of the cluster centre is used to perform crossover, as the value is represented in fuzzy terms, the bits of each feature of a cluster center separated and a cross point and block of bits are arbitrarily chosen such the number of bits of crossover point toward left and right must be same. The block is swapped with the other parent. This crossover operation is done on each cluster c for mating with parents and produces two offspring for their next generation.

Mutation

After completing the process of crossover each bit of the offspring is considered for mutation with a probability of mutation pb_m. In this process the value of chosen bit will be flipped from 0 to 1 or vice versa. The rate of mutation is set to transform one bit for each child in a cluster.

5 RESULTS AND DISCUSSIONS

This section discusses about the simulation result of proposed model AFCGC for continuous keystroke authentication using MATLAB software on the buffalo dataset which consist of 148 users with a total of 2.14m keystrokes in three different sessions. Different performance metrics are used for evaluating the performance of the AFCGC with other two clustering models namely K-Means and Fuzzy C-Means clustering. Table 1 shows the performance comparison of three different Clustering models against continuous keystroke dynamics.

Performance Metrics

$$\text{Mean Absolute Error} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \overline{y}_i\right|$$

$$\text{Root Mean Square Error} = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}$$

Table 1: *Comparison of Clustering Models*

Measures	K-Means Clustering	Fuzzy C-Means Approach	Fuzzy Gene Clustering Approach
Correctly Clustered	74.81%	82.5249 %	92.35%
Incorrectly Clustered	25.19%	17.4751 %	7.5%
Mean Absolute Error	0.3081	0.1811	0.0132
Root Mean Squared Error	.2082	0.0876	0.0041

The performance analysis of the three different clustering models based on correctly clustered instances as either normal or as intruders is shown in Fig. 3. From the output, it is observed that the use of Adaptive Fuzzy Genetic Algorithm based clustering produces better performance because of its optimality in handling the uncertain condition and the ambiguous pattern of continuous typing rhythm is discovered by the degree of membership which is clustered in an effective way. The other two algorithms K-Means and Fuzzy C-Means algorithm fails to choose the optimal cluster centroids to define an accurate degree of similarity than the proposed model.

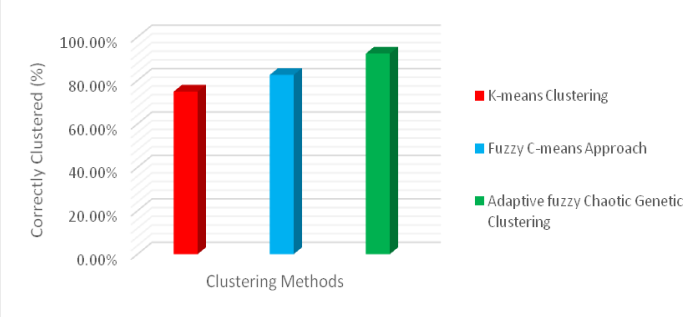

Fig. 3*: Clustering models based on correctly clustered instances*

The performance analysis of the three different clustering models based on incorrectly clustered instances as either normal or as intruders is shown in Fig.4. The issue occurs when there is ambiguity or vagueness in distinguishing the rhythm pattern as normal or impostor. This is overwhelmed by introducing each instance in the fuzzy domain representation and the genetic clustering is used for grouping the instances with similar pattern.

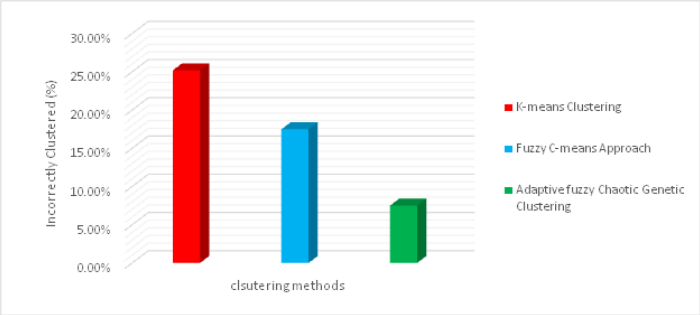

Fig. 4*: Clustering models based on incorrectly clustered instances*

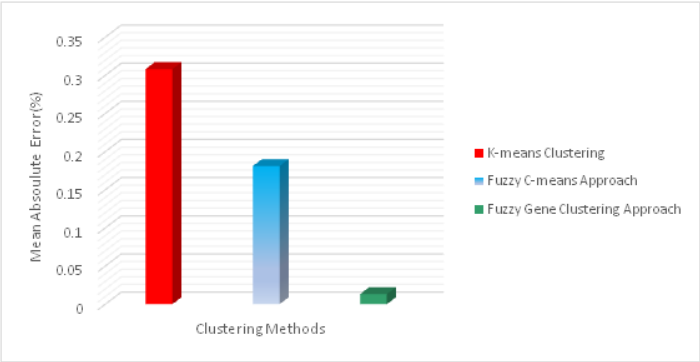

Fig. 5 *Performance comparison based on Mean Absolute Error*

The performance analysis of the three different clustering models based on Mean Absolute error and Root Mean Square Error rate are shown in Fig. 5 and Fig. 6 respectively.

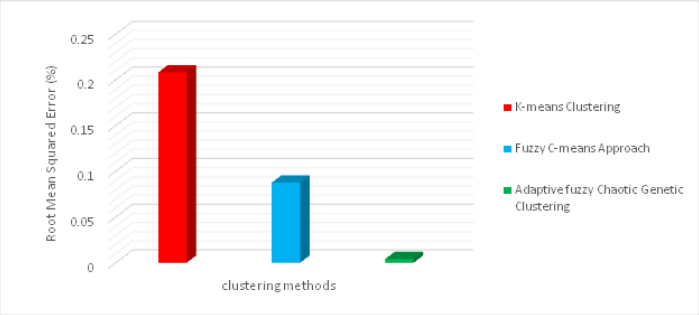

***Fig. 6**: Comparison of the clustering models based on Root Mean Square Error*

The error rate of the proposed model is reduced in a significant factor by utilizing the knowledge of chaotic theory based genetic clustering in which earlier convergence and the local optima are greatly handled with the optimized selection of population where the existing models fails to handle the ambiguity in distinguishing similar pattern of continuous keystroke authentication.

5 CONCLUSION

This paper concentrates on introducing the chaotic mechanism in genetic algorithm to enhance the process of fuzzy clustering for continuous keystroke authentication. Based on the keystroke patterns the similarity among them are identified and clustered using the proposed adaptive fuzzy chaotic genetic clustering. The main objective of this research paper is to handle the ambiguity in discovering the authenticated users and imposters in terms of continuous keystroke analyzing. The common issue of suffering from earlier convergence due to local optima is overwhelmed by using the chaotic mapping in selection of centroids as well as optimal selection of population which are highly responsible for the better accuracy. The proposed model AFCGC produces better clustering accuracy by increasing the detection rate of users as authenticated or impostors compared to the existing models considered in this work. In future, feature extraction and the deep learning models can be used for acquiring more depth knowledge about continuous keystroke authentication.

ACKNOWLEDGMENT

The authors wish to express their profound gratitude to the Managements and the Principals for their kind support and inspiration towards carrying out this research work. (DrNGPASC 2019-20 CS018)

REFERENCES

[1] R.H.C. Yap, T. Sim, G.X.Y. Kwang, and R. Ramnath. “Physical access protection using continuous authentication”. In IEEE Conference on Technologies for Homeland Security, pages 510–512. IEEE, 2008.

[2] P. Bours. “Continuous keystroke dynamics: A different perspective towards biometric evaluation”. Information Security Technical Report, 17:36–43, 2012.

[3] P.S. Dowland and S.M. Furnell. “A long-term trial of keystroke profiling using digraph, trigraph and keyword latencies”. In Security and Protection in Information Processing Systems, volume 147 of (IFIP) The International Federation for Information Processing, pages 275–289. Springer US, 2004.

[4] D. Gunetti and C. Picardi. “Keystroke analysis of free text”. ACM Transactions on Information and System Security, 8(3):312–347, 2005.

[5] J.C. Stewart, J.V. Monaco, S.H. Cha, and C.C. Tappert. “An investigation of keystroke and stylometry traits for authenticating online test takers”. In International Joint Conference on Biometrics (IJCB), pages 1–7, 2011.

[6] Messerman, T. Mustafic, S.A. Camtepe, and S. Albayrak. “Continuous and non-intrusive identity verification in real-time environments based on free-text keystroke dynamics”. In International Joint Conference on Biometrics (IJCB), pages 1–8. IEEE, 2011.

[7] P. Ordal, D. Ganzhorn, D.V. Lu, W. Fong, and J.B. Norwood. “Continuous identity verification through keyboard biometrics”. Journal of Undergraduate Research, 4(1):20–24, 2005.

[8] A.A. Ahmed and I. Traore. “Biometric recognition based on free-text keystroke dynamics”. IEEE Transactions on Cybernetics, 44(4):458–472, 2014.

[9] Abdullah Alshehri,Frans Coenen, Danushka Bollegala, “Iterative Keystroke Continuous Authentication: A Time Series Based Approach”, KI –Künstliche Intelligenz(2018) 32:231–243

[10] M. Rathi, A. V. Senthil Kumar, “Euler Movement Firefly Algorithm and Fuzzy Kernel Support Vector Machine Classifier for Keystroke Authentication”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-11, September 2019

[11] A.V. Senthil Kumar and M. Rathi, “Keystroke Dynamics – A Behavioral Biometric Model for User Authentication in Online Exams,” in Biometric Authentication in Online Learning Environments, IGI Global, 2019, pp. 183–207, DOI: 10.4018/978-1-5225-7724-9.ch008.

[12] M. Rathi, Dr. A. V. Senthil Kumar, “Exploration of Keystroke Dynamics Based Authentication on Fixed-Text and on Free-Text”, International Journal of Computer Sciences and Engineering, Vol.7 , Issue.1 , pp.807-812, 2019

[13] Lu Xiaofeng, Zhang Shengfei, Yi Shengweib, “Continuous authentication by free-text keystroke based on CNN plus RNN”, 2018 International Conference on Identification, Information and Knowledge in the Internet of Things, IIKI 2018, Procedia Computer Science 147 (2019), 314–318

[14] Yan Sun, Hayreddin Ceker, Shambhu Upadhyaya, “Shared keystroke dataset for continuous authentication”, Conference: 2016 IEEE International Workshop on Information Forensics and Security (WIFS)

[15] E. Vural, J. Huang, D. Hou, and S. Schuckers, “Shared research dataset to support development of keystroke authentication”. In IEEE International Joint Conference on Biometrics, 2014

[16] F. Bergadano, D. Gunetti, and C. Picardi. “User authentication through keystroke dynamics”. ACM Transactions on Information and System Security, 5(4):367–397, 2002.

[17] C. Murphy, J. Huang, D. Hou, and S. Schuckers, “Shared dataset on natural human-computer interaction to support continuous authentication research,” in IEEE International Joint Conference on Biometrics (IJCB), 2017, pp. 1–6.

[18] J. Huang, D. Hou, S. Schuckers, and S. J. Upadhyaya, “Effects of text filtering on authentication performance of

keystroke biometrics," in 2016 IEEE International Workshop on Information Forensics and Security (WIFS), Dec 2016, pp. 1–6.

[19] T. Shimshon, R. Moskovitch, L. Rokach, and Y. Elovici, "Continuous verification using keystroke dynamics," in 2010 International conference on computational intelligence and security,IEEE. 2010, pp. 411–415.

[20] Soumik Mondal, Patrick Bours, "Continuous Authentication using Fuzzy Logic", Proceedings of the 7th International Conference on Security of Information and Networks,pp 1-9, 2014

[21] http://cubs.buffalo.edu/research/datasets

[22] Zhaohui Jiang, Tingting Li, Wenfang Min, Zhao Qi, Yuan Rao¸ "Fuzzy c-means clustering based on weights and gene expression programming", Pattern Recognition Letters, Volume 90, 15, Pages 1-7, 2017

[23] S. Salcedo-Sanz, J. Del Ser, Z. W. Geem, "An Island Grouping Genetic Algorithm for Fuzzy Partitioning Problems", The Scientific World Journal, Article ID 916371, pp 1-15 pages volume 2015