Performance of Keystroke Biometrics Authentication System Using Artificial Neural Network (ANN) and Distance Classifier Method

N. Harun, W. L. Woo and S.S. Dlay
Electrical, Electronic and Computer Engineering
Newcastle University
United Kingdom

Abstract—**Having a secure information system depends on successful authentication of legitimate users so as to prevent attacks from fraudulent persons. Traditional information security systems use a password or personal identification number (PIN). This means they can be easily accessed by unauthorized persons without access being noticed. This paper addresses the issue of enhancing such systems using keystroke biometrics as a translucent level of user authentication. The paper focuses on using the time interval (key down-down) between keystrokes as a feature of individuals' typing patterns to recognize authentic users and reject imposters. A Multilayer Perceptron (MLP) neural network with a Back Propagation (BP) learning algorithm is used to train and validate the features. The results are compared with a Radial Basis Function (RBF) neural network and several distance classifier method used in literature based on Equal Error Rate (EER).**

Keywords- Keystroke, Biometrics, Multilayer Perceptron (MLP) Neural Network, Back Propagation (BP), Verification, Security

I. INTRODUCTION

The number of computer use's has increased rapidly and so too has the use of internet applications such as e-commerce, online banking services, webmail, and blogs. All internet applications require the user to use a password authentication scheme to make sure only the genuine individual can login to the application. Passwords and personal identification numbers (PIN) have traditionally been used to access such applications [1, 2, 5]. However, it is easy for unauthorized persons to access these systems without detection. In order to enhance such password authentication systems, typing biometrics, known as keystroke, can be used as a transparent layer of user authentication.

The idea of keystroke dynamics for recognition has been since World War II [7]. Operators were able to easily identify the sender from their key rhythms. Since then, many adaptations of this phenomenon have been studied. Keystroke dynamics is a process of analyzing keyboard typing characteristics or keyboard typing rhythms by monitoring keyboard inputs [3, 4, 5]. In other words, the system verifies how a person types. Keystroke verification techniques can be categorized as either static or continuous. Static verification system approaches study keystroke characteristics at a specific time. Although they are more robust they cannot detect a substitution of the user after initial verification. Continuous verification, on the other hand, examines the user's typing behaviour throughout the interaction time. Time-features can be extracted from keystroke data in many ways, such as studying keystroke latency, duration of key hold, pressure of keystroke, frequency of word errors, and typing rate [9, 10]. However, not all of these methods are widely used. Keystroke solutions are usually measured in three ways: dwell time – how long a key is pressed, flight time – how long it takes to move from one key to another, and key code [7, 9, 11]. Keystroke dynamics is one of the novel and creative biometric techniques. It is not only nonintrusive, but also transparent and inexpensive [5].

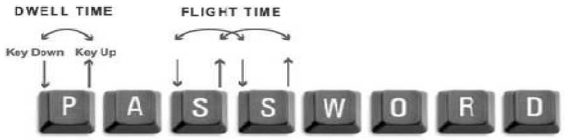

Figure 1. Time measurement for a keystroke [11].

The way a person types can verify their identity with a False Acceptance Rate (FAR) of approximately 0.01% and a False Reject Rate (FRR) of approximately 3.0% [7]. Figure 2 shows a comparison between FAR and FRR of a keystroke system and other biometrics.

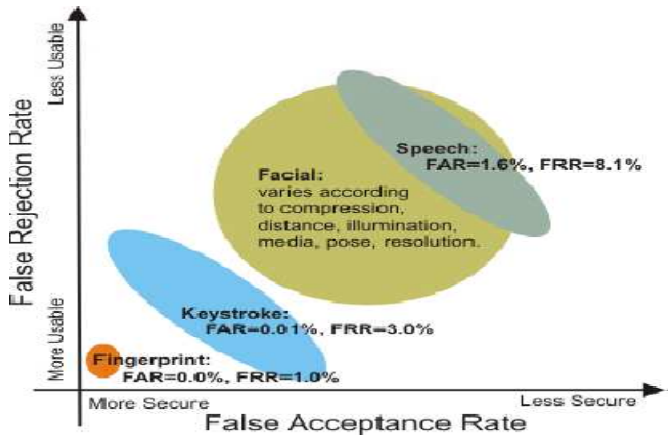

Figure 2. Error rate comparison [7].

978-1-4244-6235-3/10/$26.00 ©2010 IEEE

II. RELATED WORK

Some aspects can be utilized to create a keystroke verification system, for example by using a target string that will be typed by the user and monitored by the system, or a number of samples collected during the enrolment process to compound the training set. Two widely used features are duration of the key time interval that key remains pressed, and keystroke latency time interval between successive keystrokes. A more robust way is to use a combination of these features to analyze a keystroke system. Time accuracy, trials of authentication and classifiers are other examples of ways to analyze a keystroke authentication system [5, 6, 8, 9].

Statistical models and diagraph latencies were found to be the first techniques used to analyze keystroke biometrics. Then the neural network (NN) approach was developed by Brown and Rogers [12], they used a simple MLP with BP. Their work was extended by D.T. Lin [13], who considered the deviation on the architecture and parameters of the neural network with customized keystroke latency and gave a 1.1% FAR and 0% impostor pass rate (IPR).

N. Capuano [14] used the MLP with RBF as a transfer function, rather than a sigmoid one used previously by others. It resulted in 97% correct authentication with 0% intrusions. M.S. Obaidat and D.T Maccahairolo [15] achieved 97.5% correct classification by using a combination of multilayer feedforward with BP algorithm (MFN/BP) and sum of product (SOP) network with keystroke time interval. M.S Obaidat continued this work with S Sadoun and used key hold times for classification, comparing the performance with the former interkey time based technique and then combining interkey and hold times for the identification process. An identification accuracy of 100% was achieved when hold and interkey times were combined and trained using fuzzy ARTMAP, RBF and learning vector quantization (LVQ) neural networks [16].

In the work of L. K. Maisuria, C. S. Ong and W. K. Lai [17] keystroke was classified based on an MLP approach and K-means cluster algorithm. Both the MLP and K-means gave an 84% and 85% acceptance rate and a 69% and 85% impostor rejection rate. Alternatively, Hasimah Ali, Wahyudi and Momoh J.E Salami [18] designed and developed a system that combined the maximum pressure applied on the keyboard and time latency between the keystrokes as features to create typing patterns for each user. They combined Artificial Neural Network (ANN), with Multilayer Feedforward Network (MFN), and Adaptive Neuro-Fuzzy Interference System (ANFIS) as classifiers to authenticate individual users. The classification rate gave 100% with 0.9094 sec in average training time. But in A. Sulong, Wahyudi and M.U Siddiqi work, they only analysed the combination of the maximum pressure that apply on the keyboard and time latency between the keystrokes as features to create typing patterns for each users, using Radial Basis Function Network (RBFN) [19]. 100% classification rate with 22.4 sec in average training time shows that the RBFN-based authentication system suitable for keystroke analysis functions. In this study we investigate a MLP neural network with BP algorithm used as classifier and compare the result with RBF neural network and some of the distance classifier used in literature [9, 20, 23, 24].

III. DATABASE DESCRIPTION

A. Databases

In this paper, all databases used are based on the work of [20], which can be accessed from the Biochaves site [21]. It consists of four databases (A, B, C and D) with down-down (DD) time intervals only. A total of 47 people took part in the experiment.

In Database A, 10 people were asked to type a set of four words, *chocolate, zebra, banana, taxi*, 10 times; 5 times (five samples) during the first session and then another 5 times (five samples) a month later. Database B was built up in a similar way to Database A, but only 8 people took part. Also, the duration between the first and second sessions was shorter, being only a week [20].

In Database C, the DD time intervals were recorded by typing two fixed words in Portuguese, *computador calcula*. There were 14 people involved in this database. They were given copies of the sampling program and were free to type the words when and where they liked. In Database D, 15 people were given copies of the program and were asked to type a set of 10 freely typed rows of text with about 110 strokes per row in two sessions; 5 rows (five samples) during the first session and then the remaining 5 rows a week later. They were free to type the words when and where they liked [20].

In databases A, B and C, if the subjects pressed 'Delete' or 'Backspace' then they had to retype the string from the beginning in order to reduce the chance of recording poor samples. On the other hand, in Database D the program collected everything including the pressing of 'Delete' or 'Backspace' keys. Also text typed in Databases A, B and C categorized as static text while database D have free text typed [20].

B. Pre-processing the databases

All the databases were normalized using an equalization histogram which is a nonlinear transformation. A straightforward equalization transform obtain using following equation:

$$q(x) = \frac{1}{1 + \exp\left(-\dfrac{K(\log_e(x) - \mu_y)}{\sigma_y}\right)} \tag{1}$$

where K = 1.7, μ_y = -1.56, σ_y = 0.65 (estimated from Databases A, B and C) and $y = \log_e(x)$ with x is given in seconds. The results in this paper using this equalized data, *q(x)*, and the result was compared with the non equalized data [20].

IV. NEURAL NETWORK

MLP neural network and RBF networks have become the most widely used network architectures in pattern classification problems. The general difference between the two neural

networks is that MLP is a more distributed approach compared to RBF, which only responds to a limited section input space.

A. Multilayer Perceptron (MLP) Network

The MLP is a feed forward neural network pattern that maps groups of input data onto a set of target outputs. Figure 3 shows the structure of the MLP network used in this paper. It consists of three main parts: an input layer, one or more hidden layers, and an output layer.

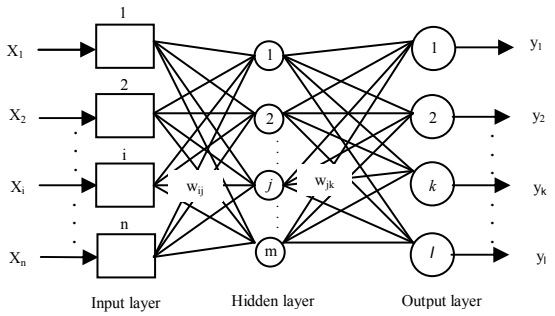

Figure 3. Architecture and signal flow of an MLP neural network.

The input layer distributes the input data to the processing elements in the next layer. The second stage is the hidden layer which incorporates the nonlinearity behaviour and the last stage shows the output layer. Input and output are directly accessible, while the hidden layers are not. Each layer consists of several neurons. The architecture in this paper uses only one hidden layer and the structure has an input x_1, x_2, …, x_n and output y. Neurons are connected between different layers using weight and bias.

The output of neuron j in the hidden layer is given by:

$$H_j = f\left(\sum_{i=1}^{n} w_{ji} x_i + b_i\right) \tag{2}$$

where w_{ij} and b_i are the hidden layer neurons weight and bias, and $f(.)$ is the non-linear activation function (sigmoid function). Then the output of the network is:

$$y = f\left(\sum_{j=1}^{m} w_{kj} H_j + b_o\right) \tag{3}$$

where f(.), w_{kj} and b_o are the output layer neuron activation function (again tansig function was used), weights and bias respectively. The BP algorithm was chosen to minimize the mean square error (MSE) based on the set of N given a training data pattern as following equation [18]:

$$E = \frac{1}{2}\sum_{n=1}^{N} (d_i - y_i)^2 \tag{4}$$

where d is referred as the target or desired output and y_i is neural network output. The training set is repeatedly presented to the network until the output of the neural network, y_i, is steady and close to the target, d. Gradient descent with momentum and adaptive learning rate backpropagation is a network function used in this work that updates weights and bias. The weights are updated to get a minimum E. During the simulations, the number of input nodes, learning rate value, the number of hidden nodes, momentum value and performance goal value were changed so as to find the most suitable parameter values.

B. Radial Basis Function (RBF) Network

The popular alternative neural network architecture is RBFN. RBFN normally configured with three different layers; input layer, single hidden layer of units and output layer as shown in Figure 4.

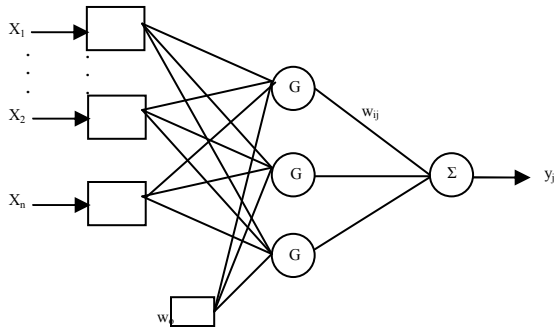

Figure 4. Architecture of RBF neural network [19].

The structure of this network is similar to that of MLP, but it only produces one output. An activation function is used in the single hidden layer from a class of basis functions such as Gaussian and multiquadric. The commonly used basis function is the Gaussian basis function which has the parameter centre and width or spread. A Gaussian basis function monotonically decreases with distance from the centre and are local (give a significant response only in a neighborhood near the centre) and are more commonly used than multiquadric which have a global response. The centre of the Gaussian basis function for a node i at the hidden layer is a vector c_i. The Euclidean distance d_i is computed between the input signal x and the centre c_i as follows [19]:

$$d_i = \| x - c_i \| = \sqrt{\sum_{i=1}^{n} (x - c_i)^2} \tag{5}$$

Then the output for each hidden layer unit is calculated by using Gaussian basis function G to this Euclidean distance d_i [19]:

$$h_i = G(d_i, \sigma) = \exp\left(-\frac{d_i^2}{2\sigma^2}\right) \tag{6}$$

The spread **σ** corresponding to the variance has a peak at zero distance and it decreases as the distance from the centre increases. The change from the input to the hidden layer is

nonlinear, whereas linear transformation from the hidden to the j^{th} output. So, output can be written as:

$$Y_j = f(u) = w_o + \sum_{i=1}^{N} w_{ij} G(\| x - c_i \|) \qquad (7)$$
$$j = 1,2,..., m$$

where w_o is the bias, w_{ij} is hidden-to-output weight, and N is the number of hidden units. During the simulations, the spread constant and number of hidden units was changed in order to find the most suitable parameter values.

V. DISTANCE CLASSIFIER

Several distance classifier used in this work to authenticate keystroke biometrics. All these classifiers have been used in previous researchers to classify groups of typing biometrics.

A. Euclidean distance (normed)

This classifier was illustrated by Bleha et al. [23] and described it the 'normalized minimum distance classifier'. The squared Euclidean distance between the test data, y, and the mean vector from training data, μ_i, is calculated and normalized by the product of norms of the two vectors (i.e if d is squared Euclidean distance, then the normalized distance is $d/(|| \mu_i|| ||y||)$).

B. Mahalanobis (normed)

This classifier also was illustrated by Bleha et al. [23] and described it the 'normalized Bayes classifier'. The Mahalanobis distance between the test data, y, and the mean vector from training data, μ_i, also covariance matrix of timing vectors, C, is calculated and normalized by the product of norms of the mean and test vectors (i.e if d is squared Euclidean distance times covariance matrix, C, then the normalized Mahalanobis distance is $d/(|| \mu_i|| ||y||)$).

C. Manhattan distance

This distance classifier was described in [24]. The distance is calculated by summation of absolute differences between the mean vector from training data, μ_i, and test data, y.

D. Manhattan distance (scaled)

This distance classifier used by Arau´jo [9]. The mean vectors, μ_i, and the mean absolute deviation, m_i, for each time interval is calculated. The Manhattan distances is compute similar as above and divide it with m_i. The distance is similar to the normal Manhattan distance except it is scaled by mean absolute deviation.

E. Euclidean distance (normed)

This classifier is used by J. R. Montalvao [20] and similar to the one that described by Bleha but instead of $d/(|| \mu_i|| ||y||)$), they calculated by squared of differences between test data and mean vectors (i.e $||y- \mu_i||$).

VI. RESULTS AND DISCUSSION

Table I and Table II show the simulation results in EER for both MLP and RBF neural network. As shown in both tables, all the data gave good EER with the equalization histogram method. It gave large improvement in the test data in range of 2 to 6 times greater EER compared without equalization data. Therefore it can be concluded that the simple equalization databases can enhance the classification or authentication system.

TABLE I. THE SIMULATION RESULT OF MLP NEURAL NETWORK

DATABASE	MLP	
	WITHOUT EQUALIZATION	WITH EQUALIZATION
A	EER=12%, MSE=0.0006	EER=2%, MSE=0.0045
B	EER=10%, MSE=0.0656	EER=2.5%, MSE=0.0003
C	EER=10%, MSE=0.0186	EER=2.9%, MSE=0.0049
D	EER=64.3%, MSE=0.0671	EER=22.9%, MSE=0.0123

TABLE II. THE SIMULATION RESULT OF RBF NEURAL NETWORK

DATABASE	RBF	
	WITHOUT EQUALIZATION	WITH EQUALIZATION
A	EER=28%, MSE=0.0078, σ=2.5	EER=16%, MSE=0.0052, σ=3.0
B	EER=30%, MSE=0.0058, σ=2.5	EER=12.5%, MSE=0.0225, σ=2.5
C	EER=28.5%, MSE=0.0090, σ=1.0	EER=14.3%, MSE=0.0045, σ=1.5
D	EER=85.7%, MSE=0.0070, σ=2.4	EER=75.7%, MSE=0.0040, σ=2.5

Table I also demonstrated that by using MLP with the BP algorithm EER gives better performances compared to the RBF network in Table II. As mentioned before, MLP neural network as well as RBF neural network is layered feedforward networks that produce nonlinear function mappings but the only different is the hidden and output layers of MLP are both nonlinear, while only the hidden layer of RBF is nonlinear (the output layer is linear).This makes MLP more suitable and highly accurate to classifier non linear keystroke data.

Also the nodes in the hidden and output layers of MLP use the same or monotonic activation functions where it computes inner products from the input and the incoming weights; thus the activation of a hidden unit in a MLP is constant on hyperplanes surfaces. In contrast, the hidden unit in a RBF network uses Euclidean distance between the input and the centre and makes this activation constant on concentric hyperspheres dimension.

A MLP also forms a distribution representation where many hidden units will typically contribute to the determination of the output values make it more accurate than RBF network that normally only a few hidden units are active for a given input.

All these advantages by of the MLP show in improvement of 3 to 8 times better EER in range compared to RBF in equalized data while 2 to 3 times better performance for data without equalization. Also simulation result proves that MLP improves in EER between 3 and 5 times better compared to [20]. The exception is for Database D that slight high EER compared to [20].

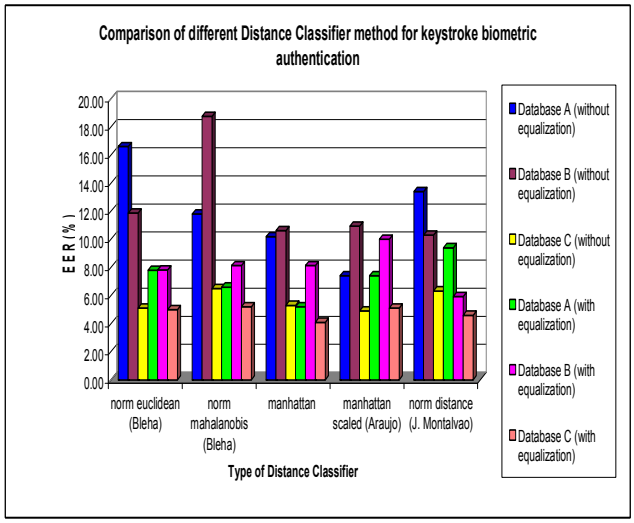

Figure 5. Comparison in EER of different distance classifier method used in [9, 20, 23, 24]by using the described database

Figure 5 shows result in EER for differences distance classifier as explained before. Form the graph we can conclude that these classifiers is not good enough to classify groups of keystroke biometrics. They give 5% to 19% EER values for non equalization data but with equalized data EER much better 4% to 10% EER error rate.

VII. CONCLUSION

This paper shows that keystroke is a special behavioural biometric that can be used as features for an additional and transparent layer of user authentication. This is demonstrable by using a NN platform like MLP and RBF. The simulation results revealed that MLP with BP network is more suitable to discriminate and classify a nonlinear keystroke database as low as 2% in EER. It also shows that MLP with the BP algorithm shows greater promising result for improving EER in order to verify the authorized user as compared to RBF network. Even though the result for the free text (Database D) is high (23%) compared with others databases, however, it is still proved that MLP/BP outperforms the RBF. Moreover this work proves that MLP gives better accuracy and improvement in classifying nonlinear data compared to linear classifier. However, further study has to be done to improve the EER for the free text data as well as enhance the level of security of the system.

This works also demonstrate result in EER for differences distance classifier as explained before. It can be concluded that distance classifiers is not good enough to classify groups of keystroke biometrics. They give 5% to 19% EER values for non equalization data but with equalized data EER much better 4% to 10%. However these classifiers not suitable to classify or authenticate time intervals of typing biometrics because the distance of every feature is very close each other.

ACKNOWLEDGMENT

N. Harun would like to thank Universiti Teknologi Malaysia (UTM) for scholarship of doctoral study and also for their supporting.

REFERENCES

[1] Jain, A. R. a. A. "Information fusion in biometrics." Pattern Recognition Letters, 24(13), pp. 2115-2125, (2003).

[2] U.Dieckmann, R. W. F. a. "Bioid: A multimodal biometric identification systems." IEEE Computer, 33(2), pp. 64-68, (2000).

[3] Davide Maltoni, D. M., Anil K.Jain, Salil Prabhakar Handbook of fingerprint recognition. New York: Springer, (2003).

[4] Anil K. Jain, Arun Ross, and Salil Prabhakar, "An Introduction to Biometric Recognition." IEEE Transactions On Circuits And Systems For Video Technology, 14(1), pp. 4-20, (2004).

[5] F. Monrose, a. D. R. "Keystroke Dynamics as a Biometric for Authentication." Future Generation Computing Systems (FGCS), 12(12), pp. 351-359, (2000).

[6] Modi*, S. K., & Elliott, S. J. "Keystroke Dynamics Verification Using Spontaneously Generated Password", 40th IEEE International Carnahan Conferences Security Technology. Lexington, Kentucky, (2006).

[7] Checco, J. C. "Keystroke dynamics & corporate security". WSTA, 241 Maple Avenue, Red Bank, NJ 07701, 2006. http://www.wsta.org/publications/articles/1003_article06.html

[8] R. Stockton Gaines, William Lisowski, S. James Press, and Norman Shapiro "Authentication by keystroke timing: Some preliminary results". Rand Report R-256- NSF, (1980).

[9] Arau´ jo, L.C.F., Sucupira Jr., L.H.R., Liza´rraga, M.G., Ling, L.L., Yabu-Uti, J.B.T., "User authentication through typing biometrics features". IEEE Trans. on Signal Processing, 53 (2), 851–855, (2005).

[10] Pin Shen Teh, Andrew Beng Jin Teoh, Thian Song Ong, Han Foon Neo ,"Statistical Fusion Approach on Keystroke Dynamics". Third International IEEE Conference on Signal-Image Technologies and Internet-Based System (SITIS '07), pp. 918-923, (2007).

[11] BioPassword: History and Science of keystroke dynamics. http://www.biopassword.com/resources/

[12] M. Brown and S.J Rogers, "User identification via keystroke characteristics of type names using neural networks." International journal of Man Machine studies, vol 39, pp 999-1014 ,(1993).

[13] D.T.lin, "Computer Access authentication with neural network based keystroke indentity verification", Proc IEEE Intl Conf Neural Networks ,pg 174-178, (1997).

[14] N. Capuano, M.Marsella, S.Miranda and S. Salerno, "User Authentication with Neural networks", Univerity of Salerno Italy. http://www.capuano.biz/Papers/EANN_99.pdf

[15] M.S. Obaidat and D.T Macchairolo, "A multilayer neural network system for computer access security", IEEE transactions on Systems, Machine and Cybernetics, vol 24(5), (1994).

[16] M.Obaidat and S Sadoun, “Verification of computers users using keystroke dynamics”, IEEE Transactions on systems, Man ad cybernetics Part B Cybernetic, Vol 27, pp 261-269, (1997).

[17] L. K. Maisuria , C. S. Ong and W. K. Lai, “ A comparison of artificial neural network and cluster analysis for typing biometrics authentication”, International Joint Conference on Neural Network, IJCNN’99, vol.5, pp 3295-3299, (1999).

[18] Hasimah Ali, Wahyudi and Momoh J.E Salami, “Intelligent Keystroke Pressure-Based Typing Biometrics Authentication System by Combining ANN and ANFIS-Based Classifiers”, International Colloquium on Signal Processing & Its Applications (CSPA), pp198, (2009).

[19] A. Sulong, Wahyudi and M.U Siddiqi, “Intelligent Keystroke Pressure-Based Typing Biometrics Authentication System by Using Radial Basis Function Network”, International Colloquium on Signal Processing & Its Applications (CSPA), pp151, (2009).

[20] J. R. Montalvao and E. O. Freire, "On the equalization of keystroke timing histograms," Pattern Recognition Letters, vol. 27, pp. 1440-1446, (2006).

[21] http://itabi.infonet.com.br/biochaves/en/download.htm

[22] M. Negnevitsky, Artificial Iintelligence, 1st edition, Pearson Education Limited, 2002.

[23] S. Bleha, C. Slivinsky, and B. Hussien. "Computer access security systems using keystroke dynamics." IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(12):1217–1222, (1990).

[24] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern Classification,2nd edition, John Wiley & Sons, Inc .,2001.