A Machine Learning Approach for User Identification using Keystroke Dynamics

S Rao Chintalapudi, Sri Sai Vinay Garapati, A Daniel Praveen Kumar, K Sri Sai Ram
Department of Computer Science and Engineering,

Pragati Engineering College (Autonomous), Surampalem, AP, India.

Abstract : The majority of computer systems employ a login ID and password as the principal method for access security. In stand-alone situations, this level of security may be adequate, but when computers are connected to the internet, the vulnerability to a security breach is increased. In order to reduce vulnerability to attack, biometric solutions have been employed. In this paper, we investigate the use of a behavioural biometric based on keystroke dynamics. Although there are several implementations of keystroke dynamics available, their effectiveness is variable and dependent on the data sample and its acquisition methodology. The results from this study indicate that the accuracy is significantly influenced by the attribute selection process and to a lesser extent on the authentication algorithm employed. Our results also provide evidence that Multi Layer Perceptron(MLP) Classifier is more accurate compared to K- Nearest Neighbor (KNN) and Support Vector Machine (SVM).

Keywords: K-Nearest Neighbor, Support Vector Machine, Multi Layer Perceptron, Keystroke Dynamics.

I. INTRODUCTION

Cyber security is the protection of information systems from theft or damage or unauthorized access to the hardware, the software, and to the information as well as from misdirection of the services they provide. Governments, military, corporations, financial institutions, hospitals and other businesses collect, process and store a great deal of confidential information on computers and transmit that data across networks to other computers. With the growing volume and sophistication of cyber attacks, ongoing attention is required to protect sensitive business and personal information, as well as safeguard national security. Cyber security involves protecting information and systems from major cyber threats, such as cyber terrorism, cyber warfare, and cyber espionage, Denial-of Service & Password Attacks. To avoid our account/machine from password attack, the login-password authentication is the most common mechanism used to grant access to user because it is low-cost, but password can be broken if it is weak or user is careless. The purpose of this paper is to improve login password authentication using Biometric characteristics. "Biometrics" means with the use of unique characteristics to identify an individual. Biometrics[8] is the development of statistical and mathematical methods applicable to data analysis problems in the biological sciences. Biometrics is used in computer science as a form of identification and access control. A number of biometric traits have been developed and are used to authenticate the person's identity. A biometric system[10] is essentially a pattern recognition system which makes a personal identification by determining the physiological or behavioral characteristic possessed by the user. Physiological characteristics include, fingerprint, face recognition, DNA, Palm print, hand geometry, iris recognition and retina. Behavioral characteristics are related to the pattern of behavior of a person, including typing rhythm and voice.

The security field uses three different types of authentication:

- Something you know password, PIN, or piece of information.
- Something you have a card key, smart card, or token
- Something you are a biometric.

More traditional means of access control include token-based identification systems, such as a driver's license or passport, and knowledge-based identification systems, such as a password or PIN. Since biometric identifiers are unique to individuals, they are more reliable in verifying identity than token and knowledge-based methods. The advantages mainly are the high levels of security it provides when compared to conventional methods, the uniqueness of biometric attributes makes them an ideal candidate authenticating users, problems associated with passwords can be avoided and that a Biometric characteristic can't be stolen as opposed to passwords etc. The various disadvantages are the low acceptance rate, high costs associated with Biometric

authentication[11] due to the integration into the current network and the acquisition of the hardware and the software, the danger of an individual's biometric data can be exploited and there are instances especially in voice recognition where the individual is restricted access due to a change in Biometric characteristics due to a cold etc. So, all these disadvantages have to be worked upon to ensure that this brilliant technology be incorporated into all the security systems to ensure safer transactions and restricted access.There are many thoughts behind the use of Keystroke Dynamic. When a person types, the latencies between successive keystrokes, keystroke durations, finger placement and applied pressure on the keys can be used to construct a unique signature[1] (i.e., profile) for that individual. Keystroke dynamics is the process of analyzing the way a user types at a terminal by monitoring the keyboard inputs thousands of times per second, and Attempts to identify them based on habitual rhythm patterns[9] in the way they type. In this paper, we experimented with different machine learning algorithms on our keystroke dynamics dataset.

II. LITERATURE REVIEW

Most of the papers used statistical methods and neural networks for keystroke based authentication[3] proposed a Monte Carlo approach for data collection and parallel decision tree (DT) for identifying the genuine user. Data collection included six basic parameters by comparing key press and key release of successive keys. A vector formed on the basis of raw data. Wavelet analysis was performed on four 16-element sub vectors by splitting the keystroke feature vectors and eight DT classifiers were trained for every user.

Shanmugapriya et. al discussed about the study in Keystroke Dynamics is one the biometric methods, authors tried and implemented to identify the genuineness of a user while the user is key in a keyboard. The verification process is done by observing the change in the typing pattern of the user. A complete survey of the various approaches of existing keystroke dynamics techniques its metrics. This research also discusses about the various security issues and challenges faced by keystroke dynamics.

Abdullah Osamah AL-Rahmani et.al enhanced the Classifier of Authentication in Keystroke Dynamics Using Experimental Data in the problem of cyber-attacks on information systems and networks, for various illegal purposes, has become a major threat to society and individuals. Computer hackers are using all possible means to get access to private data, or to destroy such data. It has become necessary to improve computer security through more advanced access control mechanisms. Recently the use of biometrics has been employed to strengthen access control through user authentication[2] that is based on users‟ measureable features.

Baljit Singh Saini et. al explained a survey of Keystroke Dynamics for Mobile Phones. The Biometric is the science of authenticating a user based on his physical or behavioral attributes. Keystroke dynamics is behavioral study which analyses the typing rhythm of the user. We adopted a systematic procedure for studying the state of the art in keystroke dynamics in mobile phones[7]. We analyzed the features extracted, the classification techniques, the input text, length of the input text, number of users, hardware used and the results.

III. USER IDENTIFICATION USING MACHINE LEARNING

3.1 K-Nearest Neighbor

It is a lazy learning algorithm that stores all instances corresponding to training data in n-dimensional space. It is a lazy learning algorithm as it does not focus on constructing a general internal model, instead, it works on storing instances of training data. The K-Nearest Neighbor[4] algorithm with k=3 is represented in Fig.1.

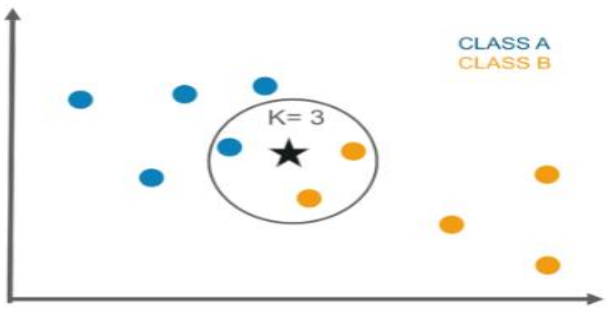

Fig. 1 K Nearest Neighbor with K=3

Classification is computed from a simple majority vote of the k nearest neighbors of each point. It is supervised and takes a bunch of labeled points and uses them to label other points. To label a new point, it looks at the labeled points closest to that new point also known as its nearest neighbors. It has those neighbors vote, so whichever label the most of the neighbors have is the label for the new point. The "k" is the number of neighbors it checks.

This algorithm is quite simple in its implementation and is robust to noisy training data. Even if the training data is large, it is quite efficient. The only disadvantage with the KNN algorithm is that there is no need to determine the value of K and computation cost is pretty high compared to other algorithms.

3.2 Support Vector Machine

The support vector machine[5] is a classifier that represents the **training data as points in space** separated into categories by a gap as wide as possible. New points are then added to space by predicting which category they fall into and which space they will belong to.

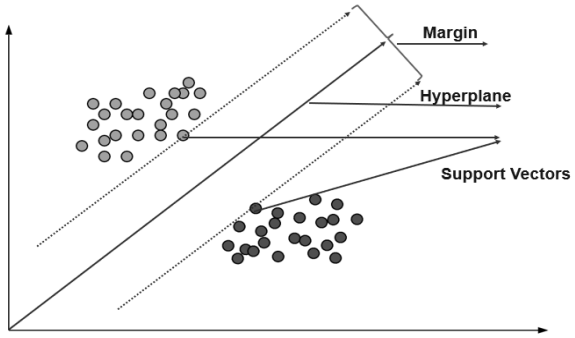

Fig .2 Support Vector Machine

It uses a subset of training points in the decision function which makes it memory efficient and is highly effective in high dimensional spaces. The only disadvantage with the support vector machine is that the algorithm does not directly provide probability estimates.

3.3 Multi Layer Perceptrons

Multilayer Perceptrons[6], or MLPs for short, are the classical type of neural network. They are comprised of one or more layers of neurons. Data is fed to the input layer, there may be one or more hidden layers providing levels of abstraction, and predictions are made on the output layer. A sample neural network is depicted in Fig.3.

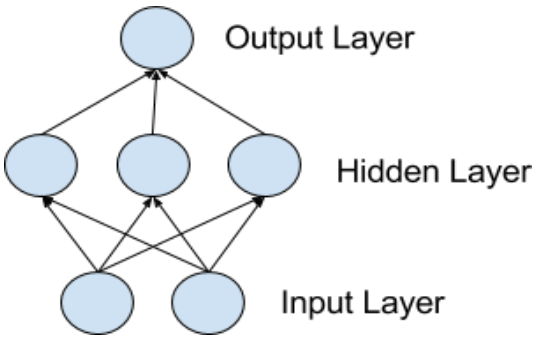

Fig. 3 A Sample Neural Network

MLPs are suitable for classification prediction problems where inputs are assigned a class or label. They are also suitable for regression prediction problems where a real-valued quantity is predicted given a set of inputs. Data is often provided in a tabular format, such as you would see in a CSV file or a spreadsheet. They are very flexible and can be used generally to learn a mapping

from inputs to outputs. This flexibility allows them to be applied to other types of data. For example, the pixels of an image can be reduced down to one long row of data and fed into a MLP. The words of a document can also be reduced to one long row of data and fed to a MLP. Even the lag observations for a time series prediction problem can be reduced to a long row of data and fed to a MLP. As such, if your data is in a form other than a tabular dataset, such as an image, document, or time series, I would recommend at least testing an MLP on your problem. The results can be used as a baseline point of comparison to confirm that other models that may appear better suited add value.

IV. RESULTS AND DISCUSSIONS

4.1 Keystroke Dynamics Dataset

It is a benchmark dataset downloaded from https://www.cs.cmu.edu/~keystroke/ .It contains keystroke information of 51 users. Each user types the strong password ".tie5Roanl".

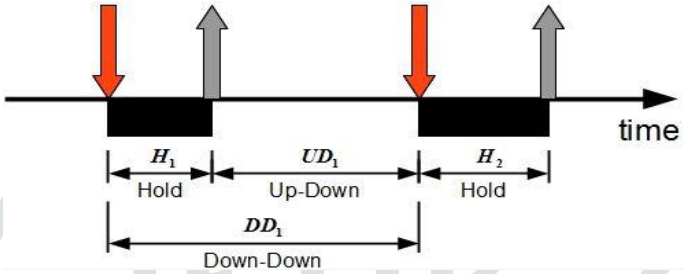

Fig.4 Features of keystroke dynamics

Data is collected in several sessions where there are 8 sessions per one user and for each session there are 50 repititions so per one user there are 400 keystroke vectors sessions are conducted for atleast one day gap

Table.1 Keystroke Dynamics Dataset

subject	sessionIndex	rep	H.period	DD.period.t	UD.period.t	…….
s002	1	1	0.1491	0.3971	0.2488	……
s002	1	2	0.1111	0.3451	0.234	……

Lets see how to interpret the values given in the dataset. Given below is a row from the .csv file in the dataset.

s002 1 1 0.1491 0.3979 0.2488 …….

These feature vector contains the data for subject s002 (user) from his first session and first repetition in that session. The feature H.key tells the hold time for the key; eg, H.period is the hold time for the dot key. DD.key1.key2 is the keydown-keydown time for the key pair key1 and key2; eg, DD.period.t is the keydown-keydown time for pressing the dot key and the t key. Lastly, UD.key1.key2 is the keyup-keydown time for the keys key1 and key2; eg, UD.period.t is the keyup-keydown time for the dot key and t key.

The columns 4-34 thus have these time information for the various keys and key pairs in **.tie5Roanl**

4.2 Performance of machine learning algorithms

The performance of the machine learning algorithms is measured using performance metric accuracy as follows.

4.2.1 Accuracy

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right It is the ratio of number of correct predictions to the total number of input samples. Accuracy can be computed using Eq. (1)

$$Accuracy = \frac{no\_of\_correct\_predictions}{Total\_no\_of\_predictions} * 100 \quad (1)$$

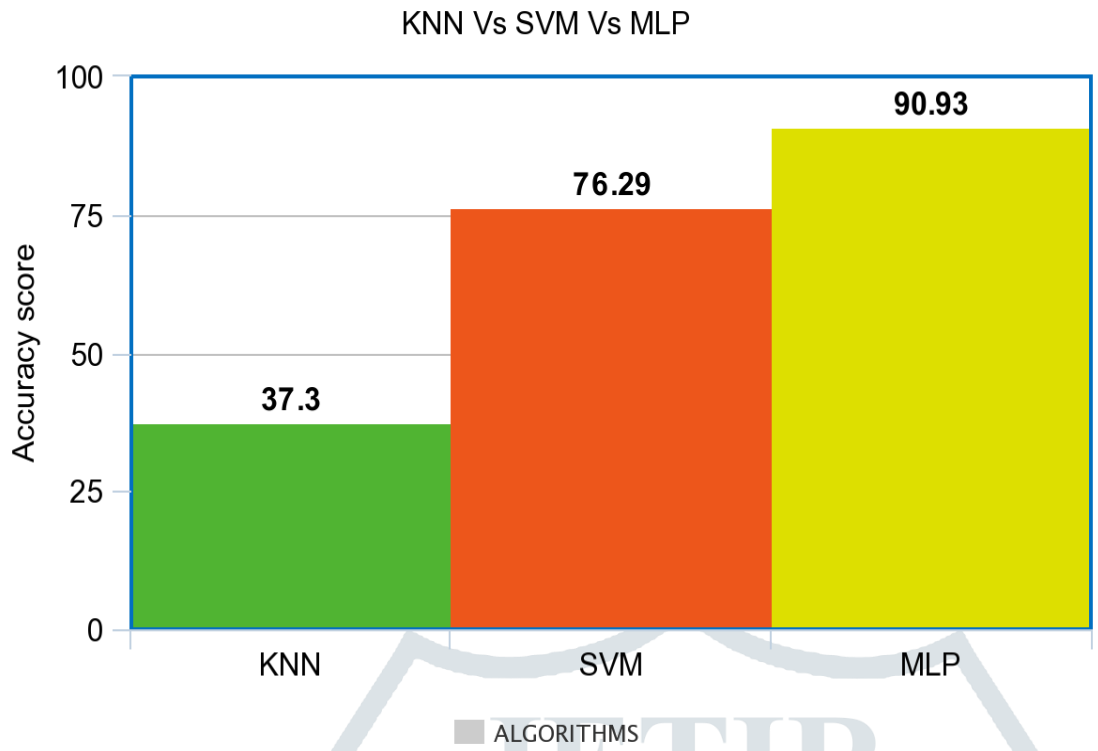

Fig.5 Performance of KNN, SVM, MLP Algorithms

In Fig.5, we have compared the performance of three machine learning algorithms namely H-Nearest Neighbour (KNN), Support Vector machine (SVM), Multi Layer Perceptron (MLP). Among these algorithms Multi Layer Perceptron Algorithm producing better model for User identification using Key Stroke Dynamics. The accuracy of these three models is listed in Table.2.

Table.2 Accuracy of KNN, SVM, MLP

Algorithm	Accuracy
KNN Classifier(KNN)	37%
Support Vector Machine(SVM)	76%
MLP Classifier(MLP)	90%

V.CONCLUSION & FUTURE SCOPE

In this paper, we experimented on benchmark dataset i.e Keystroke Dynamics dataset. It contains keystroke information of 51 users. Each user types the strong password ".tie5Roanl" and the data is collected in several sessions where there are 8 sessions per one user and for each session there are 50 repetitions so per one user there are 400 keystroke vectors ,sessions are conducted for at least one day gap so we have a dataset of 20400 keystroke vectors and we experimented our dataset with three machine learning algorithms KNN , SVM and MLP and out of three MLP Classifier is turned out to be the better model with 90% accuracy where as for KNN and SVM it is 37% and 76% accuracy.

Keystroke dynamics are very cheap in implementation and very accurate in authorization of user. Keystroke dynamics makes a username/password-based authentication procedure significantly more secure because the valid user can have the knowledge of the typing speed when the password is given. The speed of the typing may vary during the day of the user. Also the typing speed is affected if the user is in hurry. One more factor that plays a important role in the acceptance through Keystroke Dynamics is the position of the user (i.e. Typing while Sitting, Standing, and Lying)because our typing speed is differentin various positions. There are lot more studies to be done on this topic. I think the Keystroke Dynamics should be set free with their limitations i.e. Positions, because the user will like to access their account as they want. There should be no limitations on the position of user.

REFERENCES

[1] Ahmed, Ahmed Awad E., Traore, Issa, and Ahmed, Almulhem. "Digital Fingerprinting Based on Keystroke Dynamics". In: Second International Symposium on Human Aspects of Information Security & Assurance. 2008, pp. 94–104 (cited on p. 36).

[2] Al Solami, Eesa. "An examination of keystroke dynamics for continuous user authentication". PhD thesis. Queensland University of Technology, 2012 (cited on pp. 19, 43).

[3] Al Solami, Eesa, Boyd, Colin, Clark, Andrew, and Islam, Asadul K. "Continuous Biometric Authentication: Can It Be More Practical?" In: High Performance Computing and Communications (HPCC), 2010 12th IEEE International Conference on. IEEE. 2010, pp. 647–652 (cited on p. 25).

[4] Tong Xiao, Feifei Cao, Tianning Li ,Guolong Song ,Ke Zhou,Jingbo Zhu,Huizhen Wang. KNN and Re-ranking Models for English Patent Mining Task at NTCIR-7. In Proceedings of the 7th NTCIR Workshop Meeting. 2008.12. P333-338

[5] J.C. Platt, Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Microsoft Research Tech. Report MSR-TR-98-14, Microsoft, Redmond, Wash., 1998

[6] Harun, N.; Woo, W.L. ; Dlay, S.S. ;" Performance of keystroke biometrics Authentication system using Artificial Neural Network (ANN) and Distance Classifier Method " International Conference on Computer and Communication Engineering (ICCCE), 2010

[7] Antal, Margit, Szabó, László Zsolt, and László, Izabella. "Keystroke Dynamics on Android Platform". In: 8th InternationalConference Interdisciplinarity in Engineering (2014) (cited on pp. 26, 30, 42, 48)

[8] Araújo, Lívia C. F. et al. "User authentication through typing biometrics features". In: IEEE Transactions on Signal Processing 53.2 (2005), pp. 851–855 (cited on pp. 14, 25).

[9] A. Peacock, X. Ke, and M. Wilkerson, "Typing patterns: a key to user identification," *IEEE Security and Privacy*, vol. 2, no. 5, pp. 40–47, 2004.

[10] M. Karnan, M. Akila, and N. Krishnaraj, "Biometric personal authentication using keystroke dynamics: a review," *Applied Soft Computing Journal*, vol. 11, no. 2, pp. 1565–1573, 2011.

[11] Alsultan, Arwa, Warwick, Kevin, and Wei, Hong. "Free-text keystroke dynamics authentication for Arabic language". In: IET Biometrics (2016) (cited on pp. 15, 39, 47).