Machine intelligence-CO472 R UMESH SAI

Applying LSTM for fraud detection technique 16CO240

called Keystroke Dynamics 8th Semester 9986002696

[saiumesh333@gmail.com](mailto:saiumesh333@gmail.com)

**Abstract**

Keystroke dynamics, keystroke biometrics, typing dynamics and lately typing biometrics, is the detailed timing information which describes exactly when each key was pressed and when it was released as a person is typing at a computer keyboard.This report considers an issue of protecting data from unauthorized access by users’ authentication through keystroke dynamics.The dataset consists of the pressure and time characteristics typing a common password (“shiva”) of a genuine user and many fraud users and Long Short Term Memory is used to train the dataset where we convert the train data into matrix using horizontal and vertical stack and input that to LSTM model with lstm layers and dense layers using optimizers and loss function including “softmax” activation in LSTM layer and “relu” in Dense layer. Finally we get train accuracy for each epoch and our validation accuracy is around 83% considering all epochs

**Keywords**

Keystroke Dynamics, Biometrics, Timing, Pressure characteristics.

**Introduction**

Nowadays public IT access is in the works process. An increasing number of web services are emerging. Many countries tend to build an electronic government that will provide and services to their citizens. The level of privacy in such cases services must be of the highest quality. But most of the secret leak information, and cyber attacks occur due to the web

and services. The amount of land leaking every year is growing. As of October 2016 biometric authentication is in use 57% of businesses. Biometric images of the figure (finger or iris) is not a secret, so it can be copied by making a physical or digital model (remote validation). Private biometric images contain a secret (password) about it that can provide the highest level of protection. See enter the exposed keystroke power while typing a password phrase. The weak point of the validation method by using the keystroke power for very low reliability decisions made as opportunities for false rejection error (FRR) and false access error is very important to use this method in practice.

The keystroke features of the users are calculated to obtain a unique biometric pattern of that user for authentication in the future.The data needed to analyze the keystroke's power is obtained by keystroke entry.There exists different accents of english among different people similarly there exist unique pattern among different users.Though the keystroke depends on physical and mental state it is observed that there exists a consistency while typing words.

**Literature review**

Keystroke dynamics has become an active research area due to the increasing importance of cyber security and computer or network access control. There are generally two types of authentications using these keystroke dynamics.Many of the research studies focus on the static verification of the password (i.e., a pre-enrolled string of characters chosen by the user) along with these keystroke dynamics[1].Here these keystroke biometrics recorded are used for the authentication purposes.This paper describes the biometrics can be physical or behavioural out of those keystroke biometrics are one of the features of behavioural.

In paper[2] the feature subset selection for the parameters are discussed avoiding manual data preprocessing and reduce typing inconsistencies

In paper[3] the only pressure characters are considered for determining the fraud detection and 3 pressure characters and used as parameters

In many works, keystroke biometrics research has utilized many existing machine learning and classification techniques.Both classical and advanced classifiers have been used like neural networks [4]

In their papers, Bergadano et al.[5] also studied these n-graphs.They have extracted the features using the relative order of times of duration for different n-graphs.

|  |  |  |
| --- | --- | --- |
| **Authors** | **Title** | **Contribution** |
| 1)[David Umphress](https://www.sciencedirect.com/science/article/abs/pii/S0020737385800365#!)  Department of Computer Science, Texas A&M University, College Station,  2)[Glen](https://www.sciencedirect.com/science/article/abs/pii/S0020737385800365#!)  [Williams](https://www.sciencedirect.com/science/article/abs/pii/S0020737385800365#!)  Department of Computer Science, Texas A&M University, College Station, | ->Identity Verification through Keyboard Characteristics | This paper proposes that the keystrokes can be static or continuous however continuous verification observes the patter of the user in the entire session of logging which give more security.The keystrokes are calculated as mean latency or average time between two press and score is calculated and from next time the score is calculated and evaluated with the average score. |
| 1)[Enzhe Yu](https://ieeexplore.ieee.org/author/38227583400)  Dept. of Ind. Eng., Seoul Nat. Univ., South Korea  2)[Sungzoon Cho](https://ieeexplore.ieee.org/author/37441817500)  Dept. of Ind. Eng., Seoul Nat. Univ., South Korea | -> GA-SVM wrapper approach for feature subset selection in keystroke dynamics identity verification  . | Keystroke dynamics vector is created based on typing patterns and trained.Manual cleaning of data results in errors and low accuracy and inconsistencies.Feature selection process which selects the required features and remaining features are ignored and genetic algorithm and SVM is used for results. |
| 1)[George R. Widmeyer](https://ieeexplore.ieee.org/author/37390572800)  Department of Information Systems, New Jersey Institute of Technology,  2) [Michael L. Recce](https://ieeexplore.ieee.org/author/37352076700)  Department of Information Systems, New Jersey Institute of Technology | -> An Investigation into the Efficacy of Keystroke Analysis for Perimeter Defense and Facility Access | Pressure related features are brought into consideration.Feature points are obtained using pressure sensors and a biometric keypad was made to do the classification with the keystroke parameters.3 related parameters are considered namely amplitude or peak,peak area and peak sharpness. |
| 1)[Purvashi Baynath](https://ieeexplore.ieee.org/author/37086195512)  Electrical and Electronics Engineering, University of Mauritius  2)[K.M. Sunjiv Soyjaudah](https://ieeexplore.ieee.org/author/38546500400)  Electrical and Electronics Engineering, University of  Mauritius | ->Keystroke recognition using neural network | Biometric is more complex than password and is unique for each individual. Keystroke which tells about the unique pattern of a particular user.Here dwell time and hold time is used as characteristics and multilayer perceptron is used to train the data.A classifier of neural network to detect the user and evaluate him |

**Motivation**

The main aim of this lab practical is to learn and apply new model on our dataset and understand the model well enough.LSTM is a common deep learning technique used now and then and plays a vital role in many predictions.

Hence this report proposes using a deep learning model called Long Short Term Memory(LSTM) to determine fraud detection.For our approach, the general idea is to extract the desired features of pressure keystrokes by building a pressure sensitive keyboard and time parameters using pyhook library in python and apply different algorithms on the extracted features in order to obtain better fraud detection as nowadays fraud detection is the major concern in many industries

**Concept**

LSTM is the concept being contributed through this work of fraud detection technique called Keystroke Dynamics. LSTM have been developed as the short-term memory solution, with internal mechanisms called gates/neurons that can control the flow of information. By doing so, relevant information can be passed down the long chain of sequences to make predictions. With this network, almost all state-of - the-art tests are obtained based on repeated neural networks.

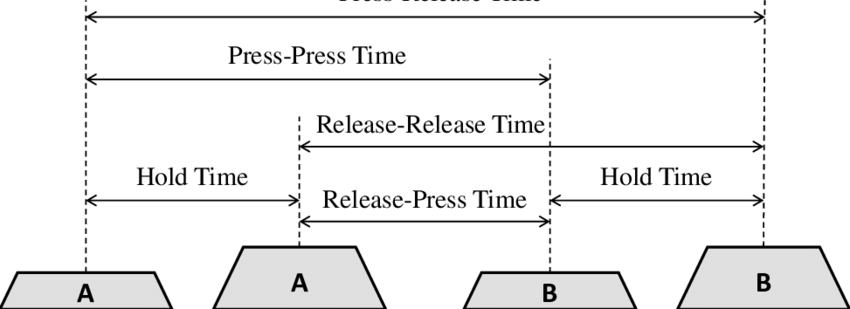
In this work we have taken the keystrokes of different users both genuine and fraud.To know the working of Long Short Term Memory and how it is applied on the labelled data to determine this fraud detection called Keystroke Dynamics.

**Methodology**

This project contains the following processes. 1)Extracting the time features using a python tool.

2)Extracting pressure features 3)Creating the dataset.and training data with LSTM model and evaluating it

**1. Extraction of time parameters**: In this step, The capture time or hold time of individual keys and the delay between two keys (the time between the release of one key and the pressing of the other key). The time parameters of the keystrokes are:



Different Timing Parameters between consecutive keys A,B

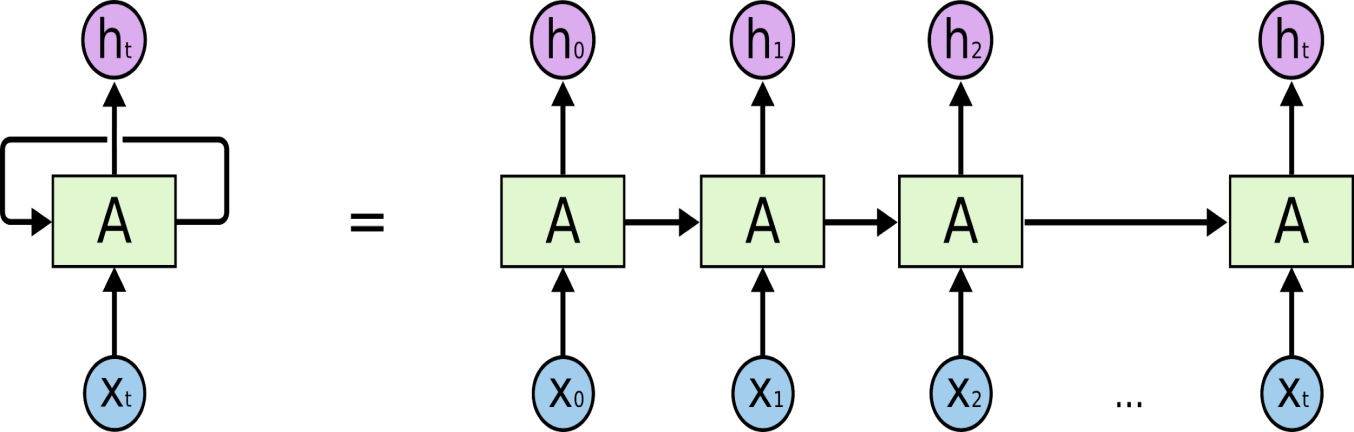
* Hold Time : key delay between pressed and key released.
* Press-Press Time: time in between two consecutive presses.
* Release-Release Time : time between two successive releases.
* Release-Press Time : time in between the current key release and the next key press.
* Press-Release Time : time between the current key press and the next key release.

There will be many more timing parameters depending on the number of characters in the password. These Key Latencies are calculated using the pyxhook package in python.For simplicity, we have considered the password to be a common name.

**2. Extraction of Pressure Parameters:** This step involves building the pressure sensitive keyboard. This involves working with the hardware components.e we are to calculate the pressure applied under each key when pressed, We need a long sensor strip which is to be placed under the keys of the same row inside the keyboard. Such sensor is FSR Interlink-408 as shown in figure. Using this sensor we can calibrate pressure along the length of the strip. It can detect pressure throughout its length.

**3. Dataset and applying LSTM :** Now we created a dataset of genuine user in which there are a total of 1000 entries labelled as 1. We also created a dataset of fraud users which consists of 500 entries and label as 0. We shuffle both datasets and the new dataset is our final one on which we apply LSTM.

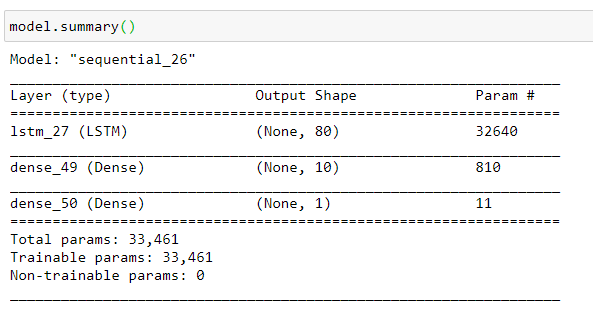
.Firstly, we take the pressure and time parameters for the name “shiva” which consists of 23 columns i.e., 23 features of the given password.Then after data preprocessing (shuffling and labelling the data) we split the data to 90% train and 10% test and train the data in batch of 3 and add layers of LSTM and dense layers and Softmax function is used as activation function along with adam optimizer and mean squared error as our loss parameter as it is accurateHere there are 23 features and LSTM models take those features into its network and as per the given data, we can set the number of time steps in the hidden layer to 1.

****

**Block diagram of basic LSTM model**

Train data is reshaped into 3D inorder to input into the Long Short Term Memory model.The model consists of LSTM layers and Dense layers, LSTM layer consists of 80 neurons whereas a dense layer has input data shape.Model trains the data in batch size of 3 of 20 epochs where trainloss,train error,validation loss,validationerror are the parameters obtained at each epoch.

**Model Summary**



**Experimental work details**

Programming language :python

Libraries used: numpy, pandas, sklearn.

Platform : Jupyter notebook

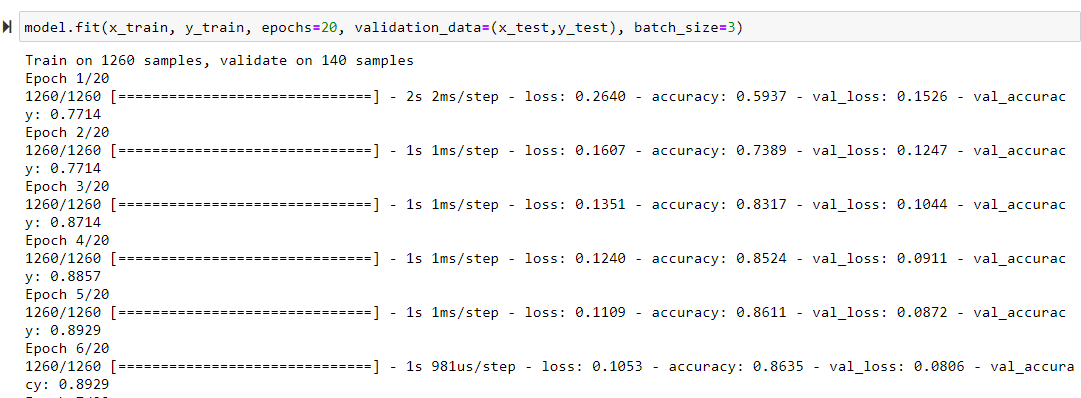
**Dataset**

Now we created a dataset of one user(genuine) in which he types the name/password(shiva) in 100 different sessions of a day in which each session contains 10 repetitions making a total of 1000 entries labelled as 1. Time parameters and pressured parameters of the user are named as a column in the dataset and used as an individual feature for that user. The data was not being collected at the same part of the day.We also created a dataset of fraud users which consists of 500 entries and label as 0. We shuffle both datasets and the new dataset is our final one on which we apply LSTM.

**Contributions**

The major concept used in this lab work is Long Short Term Memory. I have learned how to apply LSTM on the fraud detection technique called keystroke Dynamics and I have studied a few topics of deep learning.

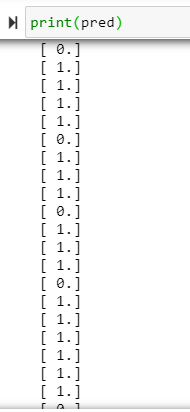
**Results**

****

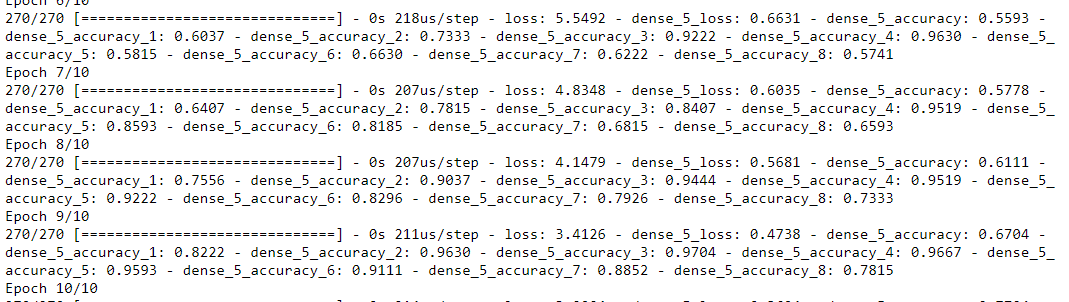
Above figure shows loss and accuracy of both train and test dataset for every epoch

And we can see that our test accuracy is around 85% after considering all epochs

.



Above figure shows the output of test data.



Above is the accuracy of proposed method using only time parameters ie, around 75%

**Conclusion and Future work**

In this report, we propose Keystroke Dynamics using Long Short Term Memory. This method gives better results when compared to applying LSTM/neural networks to the dataset that contain only time parameters.To the validation accuracy of the proposed method(85%) is better in the existing method as the third picture shows the accuracy of existing one(75%).The proposed method works better because the more features are considered which classifies the typing pattern more uniquely . This method can also be applied on large datasets and adding more LSTM and Dense layers according to our requirement.

**References**

[1] D.Umphress and G. Williams, “Identity Verification through Keyboard Characteristics”, Int’l J. Man-Machine Studies, Vol. 23, No. 3, pp. 263–273, 1985.

[2] E.Yu and S.Cho.“GA-SVM wrapper approach for feature subset selection in keystroke dynamics identity verification”, In Proc. Int’l Joint Conf. on Neural Networks (IJCNN), pp. 2253–2257, 2003.

[3] C. S. Leberknight, G. R. Widmeyer and M. L. Recce, "An Investigation into the Efficacy of Keystroke Analysis for Perimeter Defense and Facility Access," *2008 IEEE Conference on Technologies for Homeland Security*, Waltham, MA, 2008, pp. 345-350.

[4] P. Baynath, K. M. S. Soyjaudah and M. H. Khan, "Keystroke recognition using neural network," *2017 5th International Symposium on Computational and Business Intelligence (ISCBI)*, Dubai, 2017, pp. 86-90.

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