

Authenticating Device Users via Keyboard Strokes

Olusola Gbenga Olufemi

Hood College
401 Rosemont Ave
Frederick, MD 21701

Rukayat Damilola Alimi

Hood College
401 Rosemont Ave
Frederick, MD 21701

ABSTRACT

The present-day world is filled with numerous valuable personal digital devices that need to be well safe-guarded. To see ample of U.S populace still suffering in the hands of impostors and fraudsters of different kinds, despite technological advancement in such great nation is worrisome. Other nations are not spared of this security menace and perpetrators. Cases of break-in of keyboard embedded or plugged-in devices become alarming day-in-day-out [8]. Many authentication mechanisms' failures in vital devices like smartphones, tablets and laptops have been reported recently. PIN, which stands for Personal Identification Numbers and pattern drawing have been used frequently as authentication methods, especially on devices that use keyboards [6]. However, these devices are still susceptible to shoulder surfing occurrence [6]. Alternatively, keystroke dynamics which authenticates genuine owners of these devices, based on their typing manner has been studied for many years, but less utilized in these very important gadgets. Keystroke dynamics can help more in reducing these authentication problems being faced by these owners, by examining the password typed or patterns made, more essentially with how a user types (the time speed and all others) [6].

General Terms

Keyboard, Strokes, Devices, Tablet, Smartphone, Computer, Users, Impostors, Identification

Keywords

Keystroke dynamics, Authentication, Security, Privacy, Password, Biometric, R, RStudio

1. INTRODUCTION

Keystroke dynamics gives better insights on typing rhythms, which distinguish among keyboard embedded or keyboard plugged-in device users. This has been recommended for detecting impostors, which may be internal or external in any environment. This is another biometric identifier, yet to be fully harnessed in most critical devices making use of keyboards for access. With keystroke dynamics, impostors attempting to authenticate using compromised passwords could be detected and rejected, since typing rhythms differ significantly from that of the genuine users [1].

Many different techniques and uses for keystroke dynamics have been proposed in recent researches. The main objective of this study is centered on collecting keystroke dynamics dataset already available in the public domain, evaluating and developing detection procedures from this dataset. Results from these measures and performance are then compared to other similar researches. To be specific, this research effort has considered using the work done by [1] as case study, by improving on their dataset. Essentially, this work adopted [1] dataset in realizing another refined detection effect.

2. RELATED WORK

From work that was done by these duo – [1], same password was typed in 400 times via a precise keyboard connected to a computer system by each of the 51 subjects or users employed in [1] study. A 10-character password, which was manually generated by these researchers, was typed in by [1] subjects. The 10-character password included a number, an upper-case letter, lower-case letters and a dot or punctuation character, to conform with the general perception of a strong password. Eventually, password as shown below was generated:

.tie5Roanl

Timing features' extraction: In [1] research, all the timing features derived from users' typing were used. However, some of these timing features were discovered to be correlated, and some were linearly dependent (for instance, each keydown-keydown time can be decomposed into the sum of a hold-time and a keyup-keydown time). The collected typing data from the 51 subjects, each typing 400 repetitions of the password generated, were classified as keydown-keydown times, keyup-keydown times, and hold times for all keys in the password. For this password, 31 timing features were extracted and organized into vectors. In fact, [1] laudable work did provide useful dataset and evaluation methods that are now made publicly available and shared all over the research community. It will be reiterated that the dataset fully exploited in this study is gotten from this great research effort.

However, [1] were able to evaluate each of 14 detectors (algorithms) proven to work well with authentication, using the password-timing data they successfully came up with. Each detector was trained and tested with same procedures and dataset in turn. The anomaly scores resulting from each detector were then converted into standard measures of errors, to finally come up with how each detector fared better than the other.

3. METHODOLOGY

3.1 Defining Dataset

The [1] typing samples of multiple users gotten from public domain were used to find decision frontiers that can be used to distinguish each user from the other in this study. All typing feature sets studied and extracted from this dataset from [1], i.e. features considered useful in this study are as listed:

- Enter: The Enter key is part of the password;
- Keydown-Keydown: The time between the key presses of consecutive keys;
- Keyup-Keydown: The time between the release of one key and the press of the next;
- Hold: Time between the press and release of each key.

3.2 Data Collection

The first step in this study data evaluation was to collect

sample of keystroke-timing data from work that was done on similar research on keystroke dynamics as earlier mentioned. A csv file of the timing data collected, which was thereafter loaded into R programming environment for onward processing is a file shown in parts as shown in Figure 1 & Figure 2.

3.3 R for the Analysis & Experimentation

Any authentication method should be able to determine any genuine user, if it must be fully accepted as being real. Moreover, verification of user's identity is considered as the most important means of validating any authentication method. In realizing these facts, R language was adopted in refining and analyzing dataset in this research implementations. The main objective for R use in this study is to provide statistical support in the analysis of data, assisting in biometric experiments, illustrating statistical inferences, and for the presentation of visual results [7]. R can also assist in automating scientific computing to minimize cost [7].

3.4 Title and Authors

Olusola holds BSc Computer Science, MTech Computer Science (Information Networks), & MSc Information Technology degrees. He's been admitted for a PhD study in Information Security, University of the Cumberland (UC). Olusola is a husband, a father and a life-long learner.

Damilola holds BSc, MSc Information Technology degrees, Hood College - Maryland. She's a wife and a caring mother. Damilola is also an ardent researcher with special interest in experimentation and innovations in digital device usage and safety.

3.5 Timing Features Used

Table 1. Showing study timing features from raw data

Keys pressed	Down-down (DD) representations on CSV file	Keys pressed	Down-down (DD) representation on csv file
(1) .	<i>DD.period</i>	(6) Shift+r+o	<i>DD.Shift.r.o</i>
(2) t+i	<i>DD.t.i</i>	(7) o+a	<i>DD.o.a</i>
(3) I+e	<i>DD.i.e</i>	(8) a+n	<i>DD.a.n</i>
(4) e+5	<i>DD.e.five</i>	(9) n+l	<i>DD.n.l</i>
(5) 5+Shift+r	<i>DD.5.Shift.r</i>	(10) n+Return	<i>DD.l.Return</i>

3.6 Experiment

The collected raw dataset used by these researchers [1], was pre-processed in order to have more efficient dataset to work with. However, this research study total pre-processed dataset was still 20,400 as [1], (i.e. 51 subjects multiplied by 400 times, for each subject's typing as the original dataset). In contrast, the dataset timing features was narrowed down in this study, as seen in Table 1, using only the *keydown-keydown times*, since *keyup-keydown times* and *hold times* have made up this. The pre-processed data was then loaded in R environment (See Figure 3), which then became the new test dataset for this work – named *DSLxx.csv* (See Figure 1 & Figure 2), to compute the means and standard deviations for each user's typing of password done 400 times (See Figure 4 & Figure 5). Finally, the overall mean of the means and the overall mean of the standard deviations were also computed. These overall means and overall standard deviations were eventually graphed using R as shown in Figure 6, Figure 7, Figure 8, Figure 9. The R codes' depiction for the whole computation and experimentation as shown in Figure 10.

3.7 Results

As can be seen from the below plots (See Figure 7 & Figure 9), one can deduce that each of the 51 users still has different typing speeds. These different typing speeds can be used to identify each one differently on same computer system. These dissimilar typing speeds realized can further be developed and programed into any keyboard embedded or keyboard keyed-in device, to help identify each user the next time they want to use same computer system. To comprehend these results better, these should be studied alongside the R codes that generated the results (see Figure 10). However, all results including training and classification times are directly associated with any typical keyboard enabled device.

3.8 Advantages of the Technology

The main goal is to be able to continually ascertain the identity of anyone, based on typing speed on any device that uses keyboard. Physiological biometrics are known to be more efficient and safer; yet they command high cost because expert device is needed to detect features [3]. However, of all other biometrics systems, keystroke dynamics has been perceived to be very economical, since computer keyboard and few other analytic tools are required [2]. Training time of users is minimal, and ease of use is very high. Public acceptability for keystroke dynamics is very high, since no prejudices against it yet [2]. With Keystroke dynamics, the fear of users forgetting, or misplacing access credentials will be reduced, so also the fear of having security credentials in wrong hands [4]. However, security and privacy parameters that must still be considered are user authentication, access control, data integrity, non-repudiation, content protection, and others [8]. It is also studied that more robust authentication systems can be designed and developed with keystroke dynamics than the traditional password [9].

	A	B	C	D	E	F	G	H	I	J	K	L
1	subject	DD.period	DD.t.i	DD.i.e	DD.e.five	DD.five.Sf	DD.Shift.r	DD.o.a	DD.a.n	DD.n.l	DD.l.Return	
2	s002	0.3979	0.1674	0.2212	1.1885	1.6055	0.759	0.2136	0.1484	0.3515	0.3509	
3	s002	0.3451	0.1283	0.1357	1.197	0.7822	0.7877	0.1684	0.2558	0.2642	0.2756	
4	s002	0.2072	0.1291	0.1542	1.0408	0.6203	0.7195	0.2931	0.2332	0.2705	0.2847	
5	s002	0.2515	0.2495	0.2038	1.0556	1.2564	0.755	0.153	0.1629	0.2341	0.3232	
6	s002	0.2317	0.1676	0.1589	0.8629	0.8955	0.7632	0.1975	0.1582	0.2517	0.2517	
7	s002	0.2343	0.1299	0.1412	0.9373	1.0896	0.3716	0.1287	0.1534	0.2528	0.2971	
8	s002	0.2069	0.1368	0.1407	0.7967	1.2005	0.3083	0.14	0.1204	0.1999	0.2907	
9	s002	0.181	0.1378	0.1367	0.6447	1.1876	0.3139	0.1152	0.104	0.2127	0.2776	
10	s002	0.1797	0.1296	0.1425	0.7357	0.9406	0.2257	0.126	0.1403	0.2138	0.2868	
11	s002	0.1807	0.1457	0.1241	0.755	0.8065	0.3117	0.1785	0.1162	0.2281	0.3187	
12	s002	0.166	0.156	0.1386	0.6927	0.8135	0.3157	0.1746	0.0502	0.4062	0.2897	
13	s002	0.1525	0.1516	0.1391	0.9155	0.7485	0.4426	0.2065	0.1492	0.2201	0.2599	
14	s002	0.162	0.1547	0.1349	0.7028	1.0995	0.3474	0.1967	0.1581	0.3101	0.3008	
15	s002	0.1871	0.1919	0.16	0.9165	0.764	0.3954	0.4337	0.1885	0.2827	0.3889	
16	s002	0.2562	0.1549	0.1462	1.3501	1.0669	0.6546	0.2112	0.1083	0.2072	1.1307	
17	s002	0.1839	0.1381	0.1774	0.6069	0.8047	0.202	0.1746	0.1521	0.1954	0.2643	
18	s002	0.1799	0.1434	0.1412	0.8381	0.8525	0.3701	0.1531	0.1186	0.1954	0.2385	
19	s002	0.1755	0.1391	0.1613	0.77	0.6947	0.486	0.1609	0.0697	0.1944	0.2976	
20	s002	0.2237	0.188	0.1803	0.7784	0.5635	0.2954	0.216	0.0135	0.2526	0.6565	
21	s002	0.1781	0.1418	0.1544	0.614	0.7332	0.2529	0.2297	0.3128	0.7067	0.3063	
22	s002	0.1374	0.1629	0.1521	0.7165	0.5739	0.2503	0.3026	0.1624	0.2048	0.2485	
23	s002	0.2217	0.1349	0.1716	0.7674	0.5554	0.2609	0.14	0.1373	0.2009	0.2654	

Figure 1: CSV file containing pre-processed data of 20400 entries partly shown

	A	B	C	D	E	F	G	H	I	J	K	L
1	subject	DD.period	DD.t.i	DD.i.e	DD.e.five	DD.five.Sf	DD.Shift.r	DD.o.a	DD.a.n	DD.n.l	DD.l.Return	
20380	s057	0.3275	0.1063	0.1047	0.1255	0.1882	0.1233	0.0985	0.0792	0.0359	0.1985	
20381	s057	0.1186	0.195	0.1461	0.2195	0.213	0.1537	0.0971	0.0987	0.0488	0.1695	
20382	s057	0.1089	0.0889	0.1148	0.1627	0.2193	0.1469	0.093	0.1074	0.0691	0.2154	
20383	s057	0.1004	0.1076	0.0683	0.1211	0.2411	0.1404	0.1457	0.0845	0.0987	0.1996	
20384	s057	0.1133	0.0957	0.0866	0.0638	0.223	0.141	0.094	0.099	0.0538	0.2278	
20385	s057	0.1606	0.0705	0.1348	0.2331	0.3277	0.2063	0.1299	0.076	0.0781	0.1956	
20386	s057	0.1179	0.1823	0.1216	0.2092	0.2349	0.1232	0.1154	0.0718	0.081	0.2022	
20387	s057	0.3681	0.1138	0.0823	0.1912	0.2293	0.1275	0.1122	0.0605	0.0565	0.1904	
20388	s057	0.1065	0.1105	0.0726	0.0797	0.2582	0.154	0.1074	0.0866	0.0362	0.2004	
20389	s057	0.1004	0.1103	0.0828	0.0794	0.31	0.2011	0.1159	0.0924	0.062	0.1935	
20390	s057	0.1126	0.0944	0.082	0.541	0.3408	0.1752	0.065	0.1919	0.1965	0.1822	
20391	s057	0.0946	0.1064	0.0488	0.0649	0.2333	0.1567	0.0935	0.1005	0.0501	0.207	
20392	s057	0.0754	0.1451	0.062	0.1756	0.2154	0.157	0.1246	0.1198	0.0246	0.3425	
20393	s057	0.0865	0.1274	0.0486	0.0467	0.2465	0.1497	0.1061	0.0897	0.038	0.2267	
20394	s057	0.0678	0.1601	0.0391	0.1923	0.2604	0.1219	0.1141	0.0765	0.0622	0.2047	
20395	s057	0.1147	0.1055	0.086	0.097	0.3599	0.1717	0.0962	0.0929	0.0406	0.2246	
20396	s057	0.1018	0.1158	0.0697	0.68	0.6376	0.2815	0.1296	0.0974	0.0448	0.2436	
20397	s057	0.0685	0.129	0.0757	0.0826	0.2398	0.2148	0.2066	0.1383	0.1329	0.2054	
20398	s057	0.063	0.1148	0.0636	0.0852	0.2441	0.1209	0.0977	0.0512	0.0868	0.2206	
20399	s057	0.1189	0.1122	0.0462	0.2045	0.219	0.17	0.1104	0.1169	0.1311	0.2017	
20400	s057	0.1294	0.099	0.0897	0.057	0.2881	0.1602	0.1111	0.0821	0.0697	0.1917	
20401	s057	0.131	0.1103	0.0813	0.1237	0.2831	0.2	0.1172	0.0784	0.1133	0.1993	

Figure 2: CSV file containing pre-processed data of 20400 entries partly shown

subject	DD.period.t	DD.ti	DD.i.e	DD.e.five	DD.five.Shift.r	DD.Shift.ro	DD.o.a	DD.a.n	DD.n.l	DD.I.Return
1 s002	0.3979	0.1674	0.2212	1.1885	1.6055	0.7590	0.2136	0.1484	0.3515	0.3509
2 s002	0.3451	0.1283	0.1357	1.1970	0.7822	0.7877	0.1684	0.2558	0.2642	0.2756
3 s002	0.2072	0.1291	0.1542	1.0408	0.6203	0.7195	0.2931	0.2332	0.2705	0.2847
4 s002	0.2515	0.2495	0.2038	1.0556	1.2564	0.7550	0.1530	0.1629	0.2341	0.3232
5 s002	0.2317	0.1676	0.1589	0.8629	0.8955	0.7632	0.1975	0.1582	0.2517	0.2517
6 s002	0.2343	0.1299	0.1412	0.9373	1.0896	0.3716	0.1287	0.1534	0.2528	0.2971
7 s002	0.2069	0.1368	0.1407	0.7967	1.2005	0.3083	0.1400	0.1204	0.1999	0.2907
8 s002	0.1810	0.1378	0.1367	0.6447	1.1876	0.3139	0.1152	0.1040	0.2127	0.2776
9 s002	0.1797	0.1296	0.1425	0.7357	0.9406	0.2257	0.1260	0.1403	0.2138	0.2868
10 s002	0.1807	0.1457	0.1241	0.7550	0.8065	0.3117	0.1785	0.1162	0.2281	0.3187
11 s002	0.1660	0.1560	0.1386	0.6927	0.8135	0.3157	0.1746	0.0502	0.4062	0.2897
12 s002	0.1525	0.1516	0.1391	0.9155	0.7485	0.4426	0.2065	0.1492	0.2201	0.2599
13 s002	0.1620	0.1547	0.1349	0.7028	1.0995	0.3474	0.1967	0.1581	0.3101	0.3008
14 s002	0.1871	0.1919	0.1600	0.9165	0.7640	0.3954	0.4337	0.1885	0.2827	0.3889
15 s002	0.2562	0.1549	0.1462	1.3501	1.0669	0.6546	0.2112	0.1083	0.2072	1.1307
16 s002	0.1839	0.1381	0.1774	0.6069	0.8047	0.2020	0.1746	0.1521	0.1954	0.2643
17 s002	0.1799	0.1434	0.1412	0.8381	0.8525	0.3701	0.1531	0.1186	0.1954	0.2385
18 s002	0.1755	0.1391	0.1613	0.7700	0.6947	0.4860	0.1609	0.0697	0.1944	0.2976
19 s002	0.2237	0.1880	0.1803	0.7784	0.5635	0.2954	0.2160	0.0135	0.2526	0.6565
20 s002	0.1781	0.1418	0.1544	0.6140	0.7332	0.2529	0.2297	0.3128	0.7067	0.3063
21 s002	0.1374	0.1629	0.1521	0.7165	0.5739	0.2503	0.3026	0.1624	0.2048	0.2485
22 s002	0.2217	0.1349	0.1716	0.7674	0.5554	0.2609	0.1400	0.1373	0.2009	0.2654
23 s002	0.1841	0.1568	0.1539	0.8558	0.5318	0.2989	0.1566	0.1362	0.2666	0.4134

Showing 1 to 23 of 20,400 entries

Figure 3: CSV data loaded into R environment for each user’s typing of password which shows 20,400 total entries

Group.1	DD.period.t	DD.ti	DD.i.e	DD.e.five	DD.five.Shift.r	DD.Shift.ro	DD.o.a	DD.a.r
1 s002	0.1695560	0.15019350	0.12641800	0.4943708	0.4692455	0.2050107	0.15991275	0.1
2 s003	0.1691607	0.15555200	0.14329400	0.2565380	0.3813785	0.2046078	0.13603675	0.1
3 s004	0.2010645	0.14814250	0.12631625	0.3696997	0.4363167	0.2259325	0.12140525	0.1
4 s005	0.2782405	0.22728575	0.19429225	0.3340075	0.5180872	0.3391278	0.18995675	0.1
5 s007	0.1764632	0.12066875	0.10482850	0.2493100	0.2984668	0.1633378	0.12848700	0.1
6 s008	0.2074323	0.11494950	0.11311650	0.3091873	0.3364485	0.1364415	0.09779850	0.1
7 s010	0.1490798	0.14726150	0.11904100	0.2321005	0.2403050	0.1374405	0.12156775	0.1
8 s011	0.1656643	0.14047550	0.14364350	0.2266142	0.3004932	0.2091620	0.12749525	0.1
9 s012	0.1856252	0.14918475	0.14473925	0.3422292	0.3493205	0.1967710	0.14916950	0.1
10 s013	0.1289028	0.11641050	0.10022300	0.2452533	0.3022655	0.1443542	0.09679800	0.1
11 s015	0.1680275	0.10430800	0.07761300	0.3936880	0.3054655	0.1875380	0.13159875	0.0

Showing 1 to 12 of 51 entries

Figure 4: Computed mean data for each user’s typing of password done 400 times

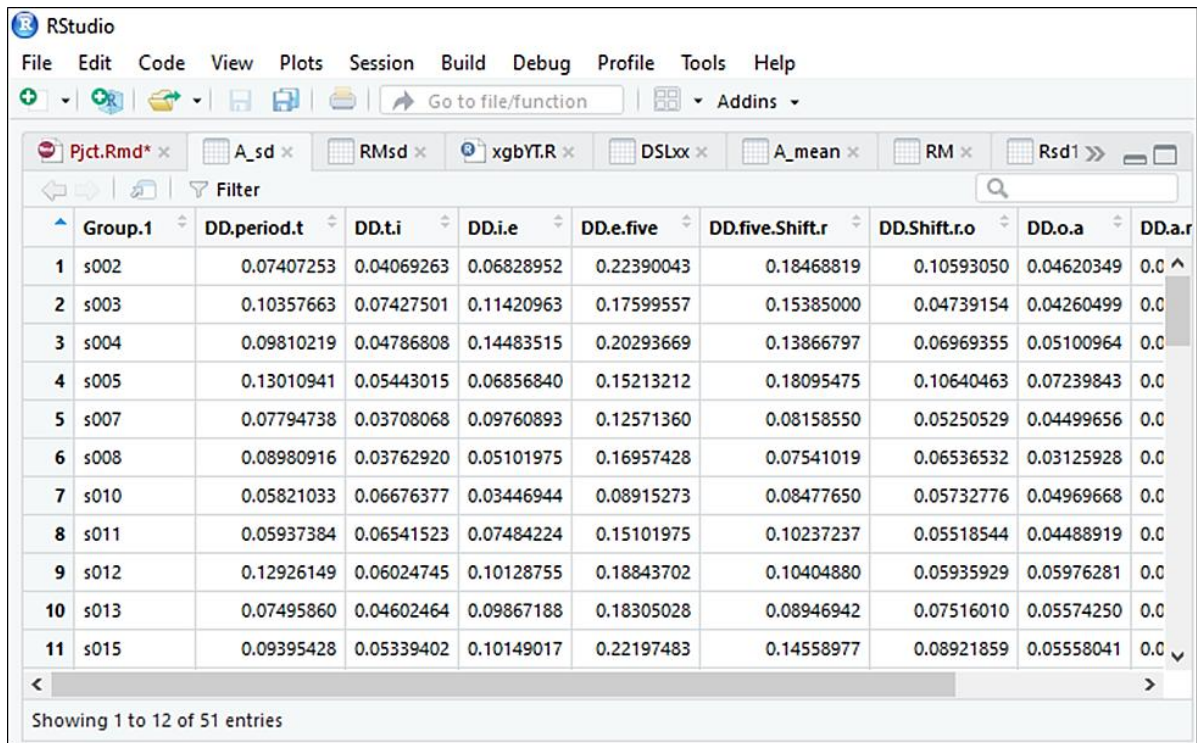


Figure 5: Computed standard deviation for each user’s typing of password done 400 times.

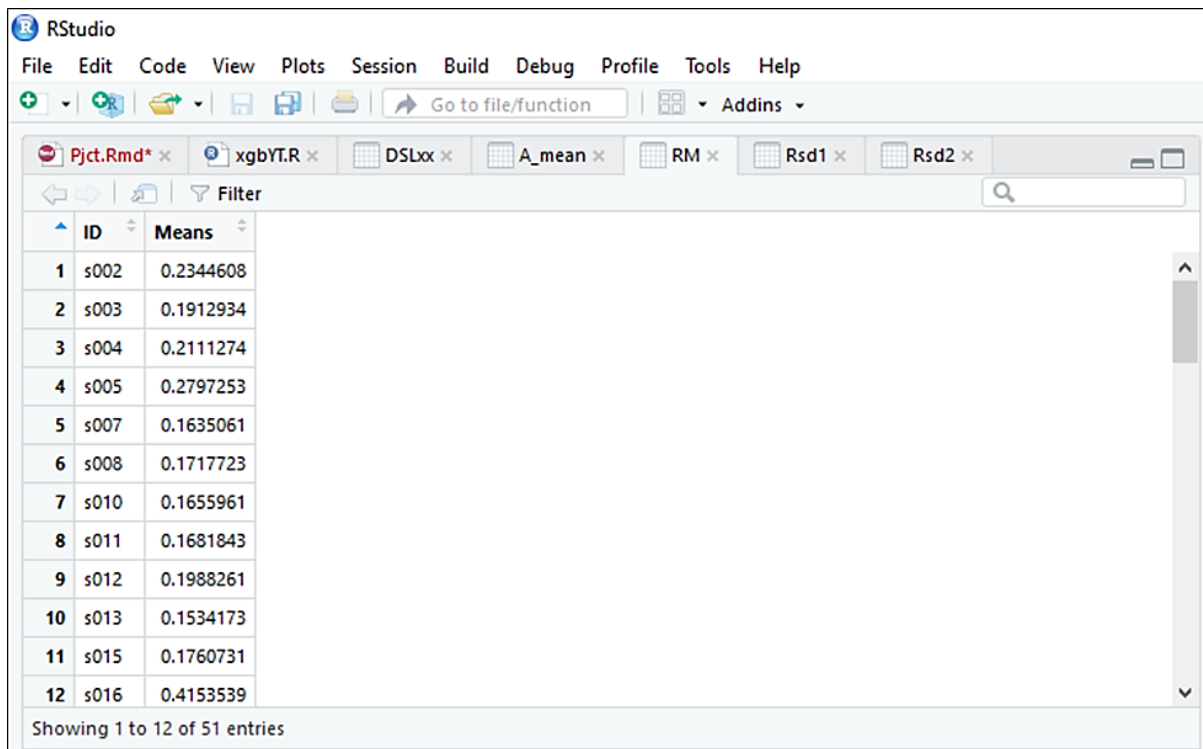


Figure 6: The calculated overall mean of means of each user’s typing of password done 400 times

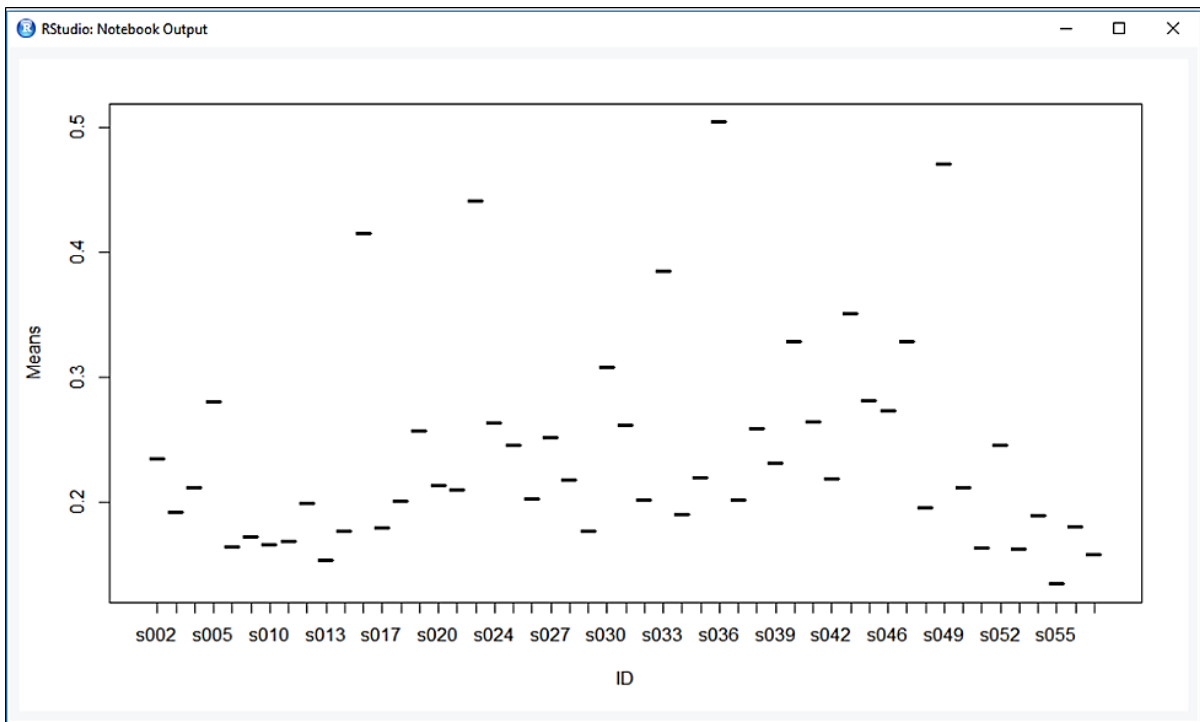


Figure 7: Plot of overall mean of means columns against Users' ID rows

ID	Means
1 s002	0.09790430
2 s003	0.08986158
3 s004	0.09438789
4 s005	0.10324220
5 s007	0.06687651
6 s008	0.07225096
7 s010	0.06130288
8 s011	0.07142409
9 s012	0.09212601
10 s013	0.07784508
11 s015	0.09976687
12 s016	0.17864864

Figure 8: The calculated overall mean of standard deviations for each user's typing of password done 400 times

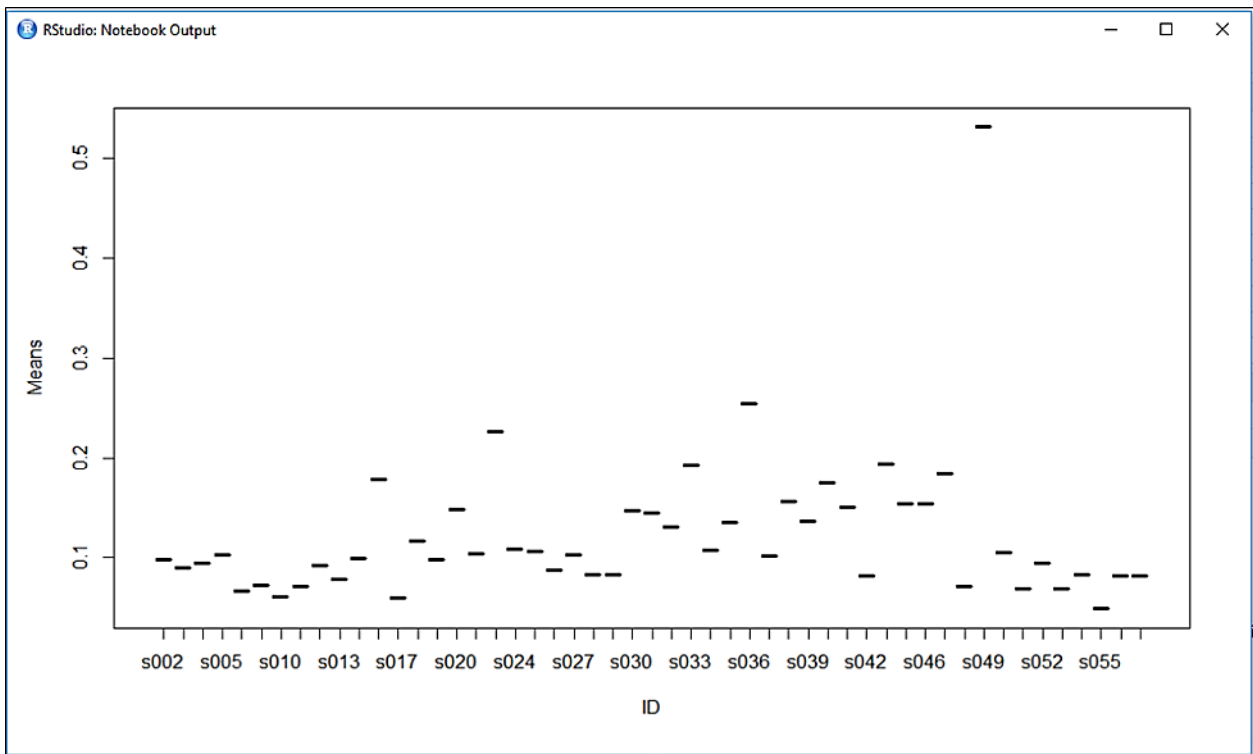


Figure 9: Plot of overall means of standard deviations column against Users' ID rows

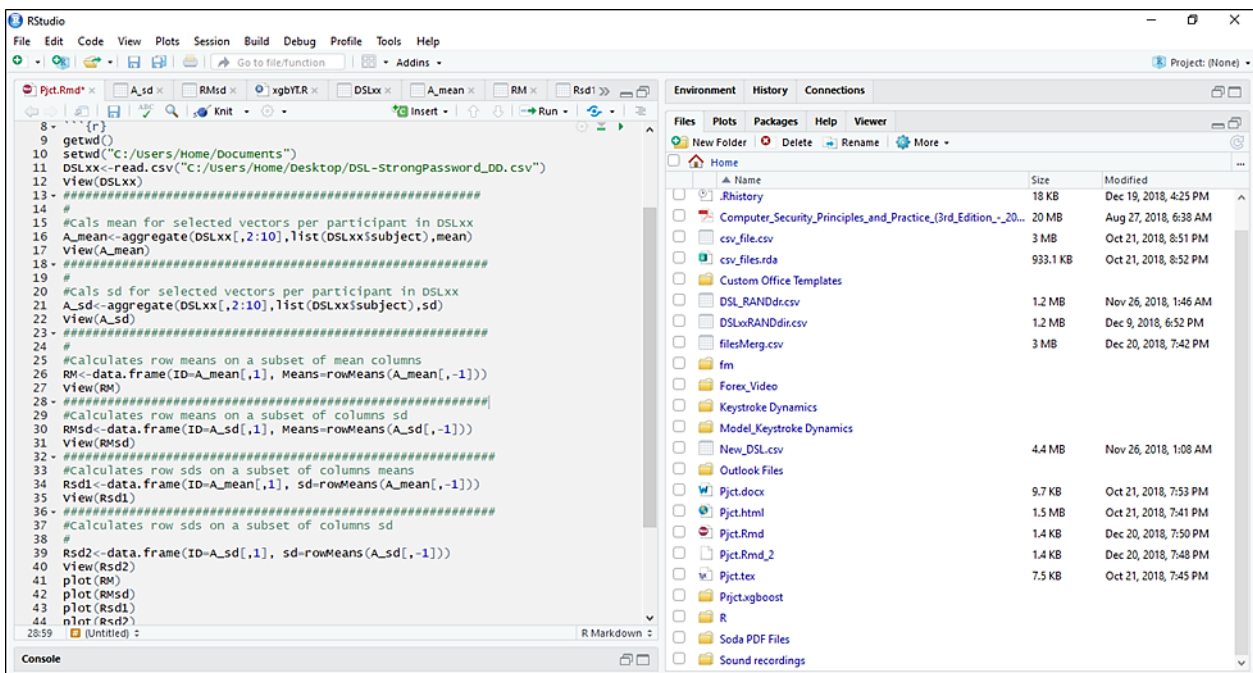


Figure 10: R Codes that generated results

4. CONCLUSION

To gain wider acceptance, a strong method of authentication should be able to provide one or more key factors of identification to improve security [4]. These factors of identification are: *something we know*, *something we have*, and lastly *something we are* [4]. However, *something we have*, based on our behavior can play far better role in proving that we are who we proclaim to be. Therefore, keystroke dynamics study, which capability is based on *something we have* should be accorded a much wider implementation in our present day vulnerable digital society,

to ensure greater safety of critical data on keyboard embedded or plugged-in devices. Rigorous continuous sourcing for rawer dataset is been made, in preparation for further work on this study. Several researches are presently on-going for keystroke dynamics to further strengthen PIN-based authentication [6]. Hence, more will be added to the body of knowledge in this key area of study until keystroke dynamics takes a complete firm root in this present time and beyond. This study pre-processed dataset can be found in <https://github.com/Olusola-cloud/Projects/blob/master/DSLxx.csv>. This is made available

to anyone that may like to explore more on keystroke dynamics, using some of the publicly available dataset.

5. ACKNOWLEDGMENTS

Thanking God Almighty for making this a possibility. Also acknowledge is our alma-mater, Hood College in Maryland, for offering sound research foundation and instructions. Appreciating our amiable faculties in this citadel of knowledge for their immense wealth of knowledge shared freely. Family members have been indirectly part of this success story too. Hence, thank you Mr. Alimi, and thank you Mrs. Irene Olufemi for your great support and understanding.

6. REFERENCES

- [1] Kevin S. Killourhy, Roy A. Maxion. 2009. Comparing Anomaly-Detection Algorithms for Keystroke Dynamics. In Proceedings of the 39th Annual International Conference on Dependable Systems and Networks (DSN-2009). IEEE Computer Society Press
- [2] Petr Svenda. 2001. Keystroke Dynamics Masaryk University, P018 - term project, 2001
- [3] Rohit A. Patil, Amar L. Renke. 2016. Keystroke Dynamics for User Authentication and Identification by using Typing Rhythm International Journal of Computer Applications (0975 – 8887), Volume 144 – No.9, June 2016
- [4] Mahnoush B, Majid B, Mohd A. 2014. Authentication Method through Keystrokes Measurement of Mobile users in Cloud Environment Int. J. Advance Soft Compu. Appl, Vol. 6, No. 3, November 2014
- [5] Rohit A., Amar L. 2016. Keystroke Dynamics for User Authentication and Identification by using Typing Rhythm, International Journal of Computer Applications (0975 – 8887) Volume 144 –No.9, June 2016
- [6] Hyungu L, Jung Y, Dong I, Shincheol L, Sung-Hoon L, Ji S. 2018. Understanding Keystroke Dynamics for Smartphone Users Authentication and Keystroke Dynamics on Smartphones Built-In Motion Sensors, Hindawi Security and Communication Networks, Volume 2018, Article ID 2567463, 10 pages <https://doi.org/10.1155/2018/2567463>
- [7] CGIAR. 2015. Biometrics and Statistical Computing - MEL – CGIAR, ICARDA’s Strategy for Biometrics and Statistics Support to its Research: 1Draft vs1, 9 September 2015
- [8] Georgios K, Dimitrios D, Dimitrios P, Emmanouil P. 2014. Introducing touchstroke: keystroke-based authentication system for smartphones. SECURITY AND COMMUNICATION NETWORKS Security Comm. Networks2016; 9:542–554Published online 1 July 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/sec.1061
- [9] Pavithra M., Sri Sathya K.B. 2015. Continuous User Authentication Using Keystroke Dynamics -.M.Pavithra et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1922-1925