

УДК 004.056.523:57.087.1

**ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ ІДЕНТИФІКАЦІЇ ОСОБИ  
ЗА КЛАВІАТУРНИМ ПОЧЕРКОМ ДЛЯ СЦЕНАРІЮ  
ВИКОРИСТАННЯ БІЛІНГВАЛЬНИХ ПАРОЛІВ. ЧАСТИНА 1**

Пономаренко В.О., Кривцун С.Ю.

Науковий керівник – к.т.н., доц. Горелов Д.Ю.

Харківський національний університет радіоелектроніки, каф. КРiСТЗi,  
студентський науковий гурток «Біометричні технології контролю доступу»  
м. Харків, Україна

e-mail: serhii.kryvtsun@nure.ua

Using the "Bilingual Keystroke Dynamics Dataset" database and the Orange software, a study of the influence of a set of passphrases in native and non-native languages on the accuracy of keystroke identification was carried out. It has been experimentally confirmed that the transition to two-class classification allows not only to significantly increase the accuracy of identification, but also illustrates the advantage of using passwords entered in English over passwords entered in their native language for users with unformed keystroke.

Одним з головних недоліків більшості біометричних систем захисту є необхідність додаткових пристроїв (і їх вартість), які вимірюють біометричні характеристики. Тому рядовим користувачам ПК або керівникам невеликих організацій економічно не вигідно купувати біометричні СКУД та забезпечувати їх постійну технічну підтримку.

Ця проблема не актуальна для систем, що використовують для вимірювання біометричних характеристик пристрої, що входять до базової комплектації будь-якого ПК. Саме до таких систем відносяться біометричні системи на основі клавіатурного почерку, адже для того щоб отримати зразок біометричної характеристики будь-якого користувача потрібна лише клавіатура.

Більшість авторів наукових праць присвячених клавіатурному почерку серед основних факторів, що впливають на точність ідентифікації, відзначає такі: алгоритм класифікації користувачів, кількість учасників експерименту з різним досвідом роботи з клавіатурою, спосіб та організація введення даних та апаратна платформа, на базі якої проводиться тестування системи аутентифікації.

При цьому такий найважливіший фактор, як мова введення паролю не враховується. Історично склалась ситуація, за якої більшість використовуваних паролів складаються з букв англійського алфавіту. Тобто для більшості користувачів у світі паролі складаються з букв нерідного алфавіту. Отже, актуальним є дослідження впливу на точність ідентифікації за клавіатурним почерком набору паролівних фраз на рідній та нерідній мовах.

На даний момент у вільному доступі розташовано більше 10 баз даних

параметрів клавіатурного почерку. Проте тільки один датасет містить часові характеристики вводу паролів на двох мовах – Bilingual Keystroke Dynamics Dataset [1]. Ця база даних складається з чотирьох датасетів. Для перших двох датасетів паролі англійською та арабською мовою обрано з наступних міркувань. По-перше, обидва слова є значущими: «бандит-173» англійською та «актор-173» арабською. По-друге, в обох розкладках використовуються ті ж самі клавіші (рис. 1). Отже, різниця у ритмі набору паролі фрази визначається лише мовою походження паролю.

Для 3-го та 4-го датасетів паролі фрази англійською та арабською мовою обрано з наступних міркувань. По-перше, вони обидві складаються зі значущих слів: «черга тремтіння світло помилка» англійською та «сонце мулла бадьорий слабкий струмінь» арабською. По-друге, в обох розкладках більшість використовуваних клавіш співпадає (рис. 2). По-третє, в обох паролях використовуються часто уживані двобуквені сполучення англійської та арабської мов. На рисунку зліва зірочкою позначені біграми, часові параметри вводу яких відображено у дослідних датасетах.



Рисунок 1

the	queue	shiver	light	false
نفاث	ضعيف	ساهر	مهلا	ثمن

Рисунок 2

В якості програмного засобу для проведення досліджень було використано інструмент інтелектуального аналізу даних Orange. В якості алгоритму класифікації використовувався метод випадкових лісів [2]. В якості інструменту оцінки точності класифікації користувачів використовувалась крос-валідація за 10 блоками [3].

Результати мультикласової класифікації користувачів датасетів «BKSD Password-English» та «BKSD Password-Arabic» наведено на рис. 3. Можна зробити наступні висновки:

1) для більшості користувачів (зелений колір) набір паролю рідною мовою призводить до зменшення рівня FRR за незмінно високого рівня FAR. Крім того в датасетах присутні користувачі, для яких подібний перехід значно покращує точність ідентифікації – користувачі S3, S6 та S117 – жовтий колір;

2) в датасетах присутні користувачі з добре сформованим клавіатурним почерком на англійській мові – користувачі S32, S43 та S33 – синій колір, для яких перехід до набору паролю рідною мовою призводить навпаки до зменшення точності ідентифікації;

Користувач	FRRen, %	FRRar, %	FRRen-FRRar, %	Користувач	FRRen, %	FRRar, %	FRRen-FRRar, %
S8	19.2	53.6	-34.4	S36	13.5	9.5	4
S1	27	50.1	-23.1	S10	11.8	7.8	4
S32	5	23.1	-18.1	S24	16.7	12.3	4.4
S9	19.9	41.9	-15	S21	6.3	1.8	4.5
S43	2.8	15.8	-13	S27	11.7	7.1	4.6
S33	0	10.6	-10.6	S2	11.5	6.8	4.7
S29	12.2	22.5	-10.3	S7	9.3	4.6	4.7
S121	3.5	13.5	-10	S44	27.9	23.1	4.8
S45	7.6	16	-8.4	S114	18.1	13.3	4.8
S15	11.9	18.2	-6.3	S14	23.5	18	5.5
S18	10	14.9	-4.9	S1	12	6.5	5.5
S47	15	19.6	-4.6	S5	14.3	8.4	5.9
S19	5.3	8.6	-3.3	S46	13	6	7
S48	9	11.3	-2.3	S16	19.4	11.6	7.8
S30	26.2	28	-1.8	S119	31.5	23	8.5
S31	4.3	6	-1.7	S42	14.7	6	8.7
S11	19.3	20.1	-0.8	S6	11.2	2	9.2
S25	8.3	8.3	0	S41	17.8	8.5	9.3
S118	17.6	17.2	0.4	S12	13.4	3.5	9.9
S39	22.1	21.1	1	S116	25.2	9.3	15.9
S50	28.2	27	1.2	S3	22	4.8	17.2
S125	11.5	8.5	3	S22	29.3	11.1	18.2
S17	19.8	16.8	3	S120	31.2	10.8	20.4
S115	16	12.8	3.2	S117	27.5	5.1	22.4
S49	4.7	1	3.7	S20	37.3	13.2	24.1

Рисунок 3

FRR/FAR, %	S4	S6	S8	S9	S19	S30	S31	S32	S33	S43	S117	
S3	7.0/0.7	0.2/1.0	2.0/2.5	2.7/0.5	0.2/0.0	0.7/1.2	0.2/0.3	7.6/0.5	0.0/0.3	0.0/0.2	0.0/0.0	
S4		2.2/2.0	3.7/4.2	0.5/4.5	0.5/0.0	1.5/2.0	0.5/0.0	7.6/5.9	0.0/0.0	0.0/0.0	0.0/0.0	
S6			2.0/0.7	0.8/1.0	0.3/0.0	0.5/0.0	0.0/0.0	0.3/0.5	0.0/0.0	0.0/0.0	0.0/0.0	
S8				1.2/3.5	0.5/0.5	0.7/1.5	0.0/0.0	4.0/5.2	0.0/0.0	0.3/0.0	0.0/0.0	
S9					0.3/0.3	1.0/0.3	0.5/0.0	0.7/2.8	0.5/0.0	0.5/0.5	0.0/0.0	
S19						1.8/1.5	0.8/0.0	0.5/0.0	0.0/0.0	0.0/1.0	0.0/0.0	
S30							2.5/0.8	0.7/1.7	0.0/0.0	0.2/1.0	0.0/0.0	
S31								0.7/0.8	0.0/0.0	0.0/0.3	0.0/0.0	
S32		BKSD Password-English								0.5/0.3	0.0/0.3	0.0/0.0
S33											0.0/0.0	0.0/0.0
S43												0.0/0.0
FRR/FAR, %	S4	S6	S8	S9	S19	S30	S31	S32	S33	S43	S117	
S3	3.3/1.2	0.0/0.0	1.5/2.0	0.0/0.5	0.3/0.3	0.3/0.5	0.0/0.0	2.0/2.1	0.8/1.3	0.0/0.0	0.0/0.0	
S4		0.0/0.0	0.5/0.5	0.3/0.5	1.0/0.3	0.3/0.5	0.0/0.0	3.3/5.2	1.0/2.6	0.0/0.03	0.0/0.0	
S6			1.0/0.0	0.0/0.0	0.5/0.0	0.0/0.0	0.0/0.0	0.0/0.05	0.0/0.0	0.0/0.0	0.0/0.0	
S8				0.5/0.3	1.5/1.3	0.5/1.1	0.0/0.5	4.0/0.8	1.3/1.3	0.0/0.0	0.0/0.0	
S9					1.5/1.3	0.5/1.0.5	1.8/3.0	0.8/0.5	0.5/0.0	4.0/0.3	1.3/1.3	
S19						0.1/0.5	0.3/1.0	0.3/0.5	0.3/0.8	0.7/0.3	0.0/0.0	
S30							1.3/0.8	1.8/0.8	0.5/0.3	2.2/1.1	0.8/0.5	
S31								0.5/0.3	0.0/0.0	3.2/0.4	0.0/0.0	
S32		BKSD Password-Arabic								7.2/2.0	1.0/1.0	0.0/0.0
S33											0.0/0.0	0.0/0.0
S43												0.5/0.8

Рисунок 4

3) в датасетах присутні користувачі з добре сформованим клавіатурним почерком на обох мовах – користувачі S19 та S31 – помаранчевий колір;

4) в датасетах присутні користувачі з майже не сформованим

клавіатурним почерком на обох мовах – користувачі S8, S4, та S9 та S30 – червоний колір;

5) різниця в точності ідентифікації для обох мов у випадку конкретного користувача може становити порядок, але рекомендувати абсолютно всім користувачам використовувати паролі виключно на рідній або на англійській мові неможна.

Підвищити точність ідентифікації можна за рахунок переходу до двокласової класифікації (рис. 4), тобто сценарію, на комп'ютері зареєстровано лише одного користувача, а всі інші автоматично відносяться до класу «зловмисник». Для експерименту було відібрані користувачі, що було позначено кольорами на рис. 3. Як можна бачити, кількість зелених комірок, тобто випадків, коли FRR та FAR менше 1-го відсотка (високий рівень точності) для обох датасетів однакова – 47 або 71.2 %. При цьому точність класифікації користувачів з добре сформованим клавіатурним почерком (жовтий, синій та помаранчевий кольори) для сценарію набору пароллю англійською мовою вища ніж для сценарію набору пароллю арабською мовою.

#### Список використаних джерел:

1. Bilingual Keystroke Dynamics Dataset. URL: <https://github.com/ntwaijry/BKSD> (дата звернення: 20.05.2023)
2. Hastie T., Tibshirani R., Friedman J. Random Forests. Chapter 15 // The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer-Verlag, 2009. 746 p.
3. Cross-validation: evaluating estimator performance. URL: [https://scikit-learn.org/stable/modules/cross\\_validation.html](https://scikit-learn.org/stable/modules/cross_validation.html) (дата звернення: 20.12.2023).
4. Aliksieiev Vasyl, Elena Sharapova, Olena Ivanova, Gorelov Denis, Synytsia Yuliia. Web-Based Application to Collect and Analyze Users Data for Keystroke Biometric Authentication. In Proceedings of the First IEEE Ukraine Conference on Electrical and Computer Engineering (UKRCON). Pages 917-922, 2017.
5. Vasyl Aliksieiev, Aleksey Strelnitskiy, Dmitry Gavva, Denis Gorelov, Yuliia Synytsia. Studying of keystroke dynamics statistical properties for biometric user authentication. Proceedings of 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), Pages 559-563, 2018.
6. Дослідження інформативних параметрів диграфів клавіатурного почерку для задач ідентифікації користувачів комп'ютерних мереж / Д.Ю. Горелов, О.О. Іванова, О.В. Кокорін, Д.В. Маслій, О.В. Литвиненко // Радіотехніка: Всеукр. Міжвід. Наук.-техн. Зб. – 2020. – вип. 201. – с. 194 – 200.