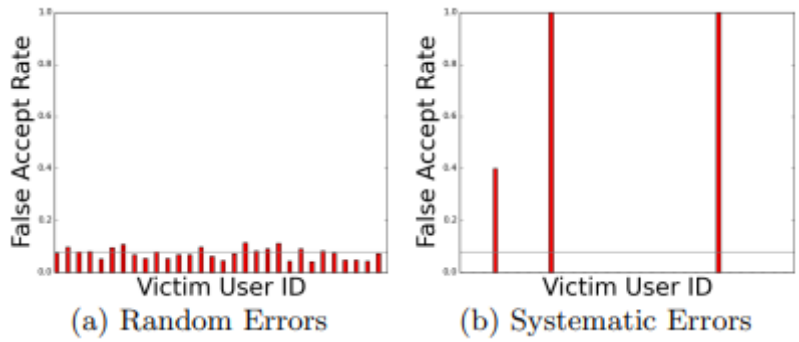
Behaviorální autentizace

Analýze dat za účelem nalezení vhodného způsobu pro behaviorální autentizaci se věnuje několik veřejně dostupných článků s různými způsoby měření a vyhodnocování. Hlavní příčinou nekonzistencí mezi studiemi je způsob sběru dat a využití vytěžených znalostí. Zatímco některé týmy sbírají data ze senzorů a dotykového displeje nepřetržitě na pozadí, ostatní se omezují na použití specifické aplikace. V prvním případě je kladen důraz na denní rutinu uživatele a dlouhodobou práci se zařízením. Typicky sledují způsob chůze, způsob jakým je telefon odkládán a zvedán, kde se uživatel pohybuje nebo jaké aplikace používá. Druhý případ pak předpokládá krátkodobé použití specifické aplikace pro specifický úkon, jako je psaní, scrollování na webu nebo swipování v galerii.

Jen málo studií zveřejní získaný dataset nebo zdrojový kód použitý k analýze dat. Často tak není možné jejich vyhodnocení ověřit, ani získat podrobnější postup zkoumání. To vede k tomu, že týmy provádějí vlastní sběr dat jen nad pár desítkami uživatelů a výkonnostní měření provádějí nad neporovnatelnými metrikami.

Začněme tedy nastavením rámce pro postup analýzy dat pro behaviorální autentizaci. Velké množství studií při tvorbě trénovacího a testovacího vzorku nerespektují dělení datové sady na jednotlivá sezení a jejich časovou posloupnost. Přitom způsob, kdy jsou vzorky voleny náhodně napříč sezeními, vede k výraznému nadhodnocení schopnosti modelu správně klasifikovat. Realistickým scénářem je použít několik, dle času vzniku, prvních celých sezení jako trénovací a zbytek jako testovací. Při vyhodnocování schopnosti klasifikátoru odhalit negativní třídu vzorků se vždy předpokládá takzvaný zero-effort útok, tedy útok při kterém se útočník nesnaží napodobit chování oběti. Důvodem je, že tento styl testu se dá simulovat pomocí dat ostatních uživatelů. Častým prohřeškem studií je použití dat sezení útočníka pro trénování modelu. Tím totiž vznikne předpoklad, že možný útočník využívá stejný systém pro behaviorální autentizaci. Obecně správným postupem při trénování binárního klasifikátoru je zvolit několik okolních uživatelů za útočníky a vůbec je nezohledňovat při testování, tak nevznikne příliš optimistické měření. Posledním významným prohřeškem, který souvisí s výše zmíněným využitím autentizačního systému, je volba velikosti okna, ve kterém je vstup či výstup modelu agregován. Některé studie vůbec nezohledňují potřebnou dobu pro detekci útočníka. [1]

S nároky z předchozího odstavce je vhodné navíc pamatovat na metriku vyhodnocování pro usnadnění porovnávání různých modelů. Většina studií se opírá o průměrné EER, FAR nebo FRR, a zatají, jak původní rozdělení metrik před zprůměrováním vypadalo – tedy zda je model schopný odhalit každého útočníka s podobnou pravděpodobností, nebo existují útočníci, kteří nebudou vůbec detekováni (jak naznačují sloupcové grafy níže). Pro kvantifikaci systematické chyby je doporučen [2] gini koeficient.



(a) Random Errors (b) Systematic Errors

Veřejně dostupné datasety

Vliv prohřešků studií popsané na předchozí straně byly zkoumány oxfordským výzkumným týmem [1] nad vlastním datasetem, tvořen scroll a swipe gesty uživatelů iPhone. Kromě reportu zveřejnili i používanou smartphone aplikaci i zdrojový kód použitý k analýze dat. Mimo již zmíněného poukázali také na vliv velikosti displeje telefonů vstupující do evaluace na chybovost modelu – gesta z velikostně podobného telefonu si jsou více podobná, než z telefonu rozdílné velikosti.

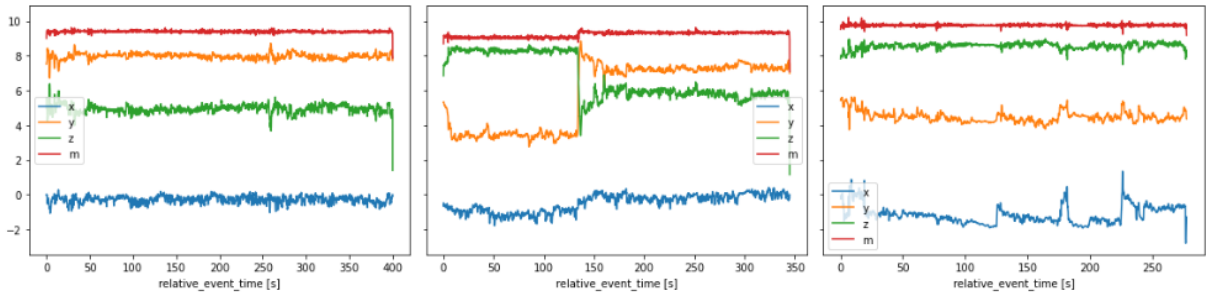
Data dalšího veřejně dostupného datasetu [3] byly shromážděny prostřednictvím hry Brain Run. Informace k datasetu uvádějí, že aplikaci využilo více než 2000 lidí s širokým spektrem zařízení. Aplikace se skládá z několika krátkých miniher, k jejichž vyřešení je potřeba swipovat a klikat na dotykovou plochu. Bližší průzkum ale ukázal, že data ze senzorů (akcelerometru, gyroskopu a magnetometru) pro většinu uživatelů chybí. Navíc data ze senzorů postrádají časové razítko a evidentně data před uložením prošla neznámým zpracováním.

Pro analýzu veřejně dostupných dat tedy zbyl dataset pojící se k jednomu z nejvíce citovanému článku v oboru [4]. Sběr dat probíhal kontrolovaně s použitím deseti Samsungů Galaxy S4, celkem se zúčastnilo 100 lidí. Každý uživatel psal odpovědi na otázky v sedě a za chůze, prohlížel mapu v sedě a za chůze a četl článek v sedě a za chůze. Každá taková aktivita byla provedena čtyřikrát. Dataset se skládá z raw dat z akcelerometru, gyroskopu, magnetometru a interakcí s displejem. Navíc byl zapojen i android GestureDetector, třída, která na základě přijatých MotionEventů rozhodne, zda se jednalo o gesto (swipe, fling, long tap) či o jediný klik. Data z GestureDetectoru byla porovnána s raw daty a bylo nalezeno několik chybějících motion eventů.

Dataset - pozorování

Akcelerometr

Analýza byla omezena pouze na psaní v sedě a držení telefonu v režimu portrét. Kromě tří os $[x, y, z]$ ze senzoru byla přidána norma $m = \sqrt{x^2 + y^2 + z^2}$. Časová řada těchto dat pak může vypadat například následovně.

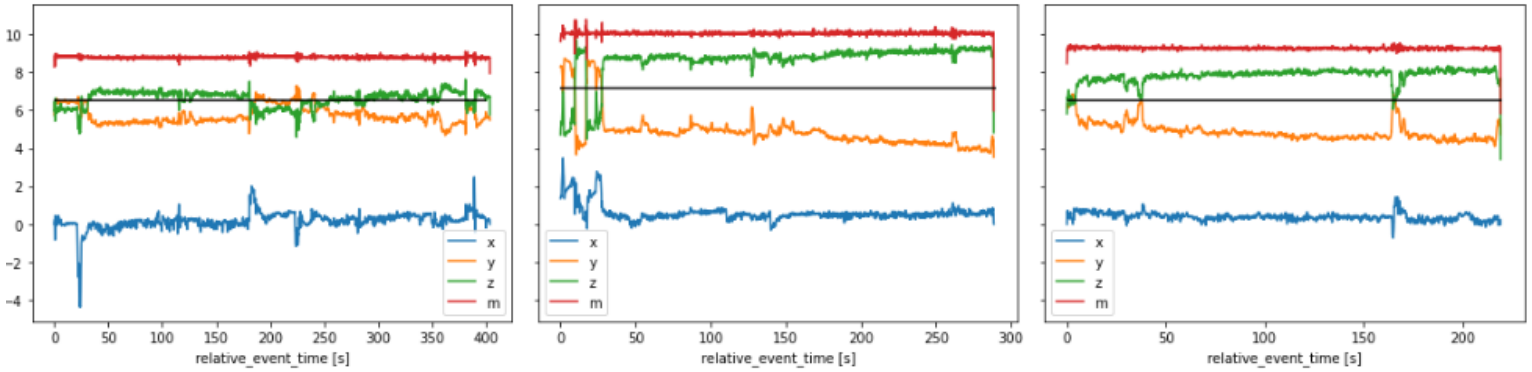


Všechny tři průběhy zobrazují držení zařízení v režimu portrét (osa *x* blízká nule). První a poslední průběh zobrazuje velice stabilní držení telefonu. Osa *z* v posledním případě je blízká normě, což odpovídá držení telefonu spíše rovnoběžně se zemí, naopak první případ dokládá držení více kolmé k zemi. Na prostředním grafu je vidět záporná korelace mezi osami *z* a *y*. Pokud by byl telefon otočen do režimu landscape, vyměnil by se význam hodnot v ose *x* a *y* (průběh *y* by byl blízko nule, *x* v záporný nebo kladný v závislosti na otočení proti nebo po směru hodinových ručiček). Norma m byla přidána

z důvodu, že je invariantní vzhledem k orientaci [1]. Správně kalibrovaný senzor by měl v klidovém (nebo rovnoměrném přímočarém pohybu) dosahovat normy *9.81*.

Průběhy níže jsou tři různé sezení téhož uživatele, vždy v režimu portrét. Černé vodorovné osy naznačují hodnotu střetu osy *y* a *z*, který je rovněž výrazně posunutý.

$mean_m = 8.80$ $mean_m = 10.06$ $mean_m = 9.27$



Podobný efekt byl pozorován u více uživatelů, někdy byl posun průměru pozorovatelný jen u jednoho ze čtyř sezení. Přestože byly pro sběr dat použity smartphony stejného modelu, patrně dochází i tak k získání rozdílných hodnot z různě kalibrovaných akcelerometrů. Tento jev nebyl řešen v žádném článku zkoumající tento dataset. Pro další analýzu byl posun odstíněn normalizací jednotlivých os normou *m*, pro osu *x* například $x = \frac{x}{m}$, a řada *m* byla diferencována.

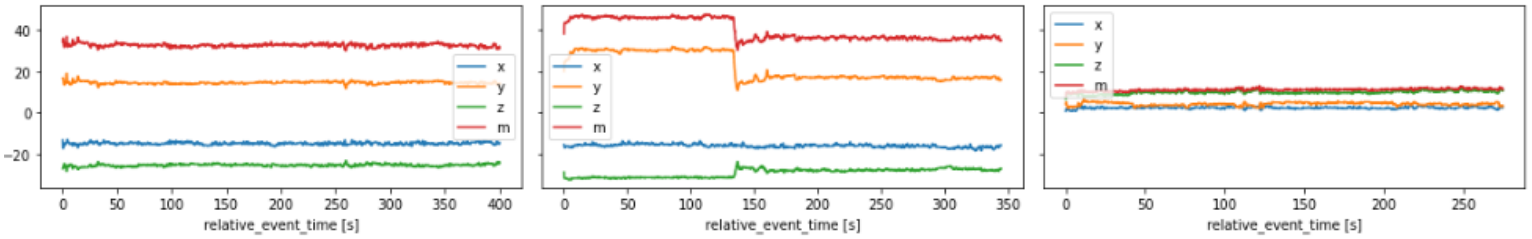
Dále byly jednotlivé časové řady rozděleny do sekundových oken, z kterých byly vytěženy příznaky, jako je maximum, minimum, rozptyl, křivost, šikmost, interquartil range, průměr (pro diferencovanou řadu průměr absolutních hodnot), energie (normovaný součet mocnin), entropie [5], hodnota dvou nevyšších amplitud a frekvence nejvyšší amplitudy [6].

Gyroskop

Gyroskop měří aktuální úhlovou rychlost, v klidovém stavu jsou všechny osy nulové. U dat gyroskopu žádný neobvyklý posun hodnot v závislosti na sezení nebyl pozorován, nebylo nutné hodnoty normovat ani diferencovat. Příznaky byly použité stejné jako pro akcelerometr.

Magnetometr

Hodnoty z magnetometru jsou extrémně citlivé na změnu držení zařízení nebo změnu polohy vůči světovým stranám. Navíc obsahují velké množství šumu v závislosti na okolí. Rozdělení dat z magnetometru nemá u mnoha uživatelů žádný překryv napříč sezeními. Data mohou být přínosem jen jako další sledování mírných otřesů zařízením. Tři průběhy znázorněny na grafech níže pochází od jednoho uživatele ve třech sezeních.



[1] Na zařízení Xiaomi Mi A1 jsem zjistil, že norma není invariantní vzhledem k orientaci, ale pohybuje se mezi *9.0* a *10.1* v závislosti na způsobu držení zařízení.

Displej

Z jednoduchých dotyků je získán čas držení displeje při dotyku, časová prodleva mezi koncem jednoho dotyku a začátkem dalšího a velikost dotykové plochy. Gesta jsou vytěžována dle [7].

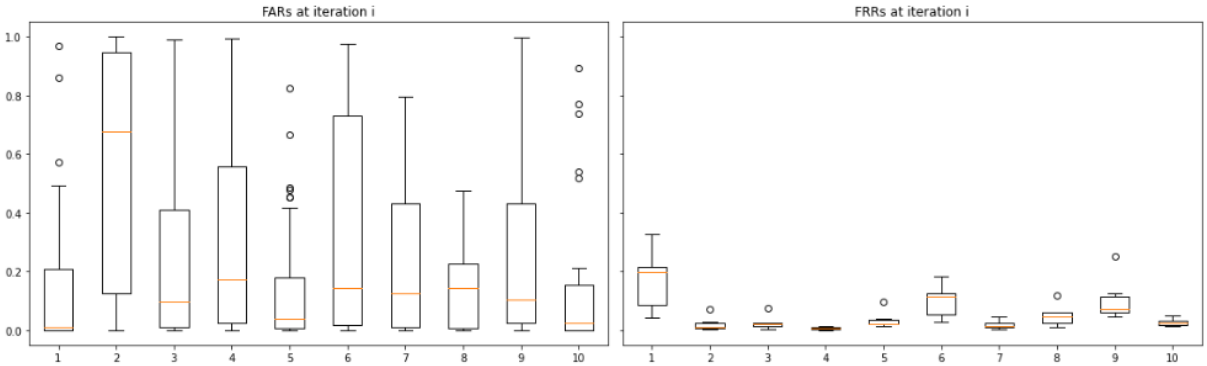
Klasifikátory

Mezi nejčastěji používané binární klasifikátory v rámci studií behaviorální autentizace patří Support Vector Machine a Random forest. Existuje však významná skupina výzkumných týmů, které považují detektory anomálií za lepší variantu. Mezi ně řadí Local outlier factor, One Class SVM, Isolation Forest a různé „template“ metody [8], čímž je myšleno, že z trénovacích dat se vytvoří jeden bod (například jako jejich průměr) a testování probíhá na základě vzdálenosti od vytvořeného předlohového bodu. Hlavní argumentací proti binárním klasifikátorům je nezbytnost mít v době trénování modelu data útočníka, což je obzvlášť problematické, pokud se uvažuje autentizační systém běžící přímo na uživatelském zařízení.

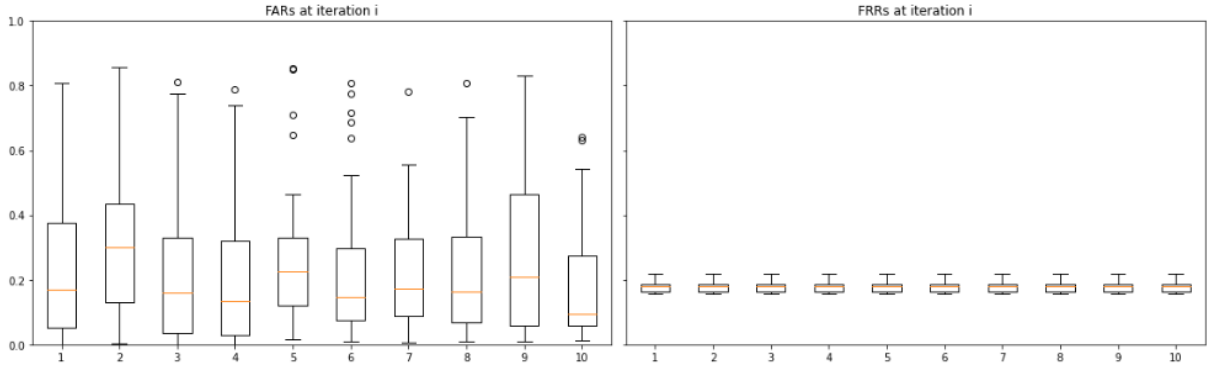
Nyní jsou uvažována pouze ta sezení, při kterých uživatelé psali v sedě odpovědi na tři otázky. Protože každý uživatel dokončil čtyři sezení, je k dispozici 12 psacích aktivit od každého uživatele. První dvě sezení, respektive 6 aktivit, byla použita pro trénování, zbylé aktivity byly určeny k testování. Po vzoru [1] byl vybrán vždy jeden uživatel jako oprávněný a zbylá skupina uživatelů simulovala útočníky. Skupina útočníků byla rozdělena na dvě skupiny. Z první skupiny uživatelů bylo náhodně zvoleno šest útočníků a z každého byla použita jedna aktivita pro trénování negativní třídy. Testování negativní třídy bylo provedeno na jedné náhodné aktivitě každého útočníka z druhé skupiny.

Rozdíl v použití binárního klasifikátoru a detektoru anomálií byl zkoumán nad daty z akcelerometru. Aby byly výsledky přesnější, bylo nejdříve zvoleno 10 nejlepších příznaků metodou MRMR [9], tato metoda dala přednost příznakům z os *[x, y, z],* pokud byl uživatel stabilní v držení zařízení v trénovacích sezeních a naopak upřednostňovala příznaky vzniklé z normy pro uživatele s nestabilním držením telefonu.

U binárních klasifikátorů se ukázal vliv náhodného výběru negativního vzorku pro trénování jako nezanedbatelný. Patrně jsou si nějací uživatelé v daných aktivitách podobnější než jiní a výběr šesti nepostihne dostatečně kvalitní vzorek dat. Při použití vyššího množství aktivit negativní třídy pro trénování zase přichází vliv nerovnováhy. Navíc často dochází k „systematické chybě“ FAR představené na první straně. Pro znázornění vlivu volby vzorků z negativní třídy bylo desetkrát spuštěno trénování a testování SVM klasifikátoru se standardizovanými daty. Nutno dodat, že model neprošel laděním hyperparametrů.



Oproti tomu jsou detektory [8] mnohem stabilnější, protože v době trénování nejsou potřebná data z negativní třídy. Navíc měření FAR neprokazuje tak extrémní systematické chyby. V průměru ale nedosahují tak dobrých výsledků jako binární klasifikátory. Níže je deset měření výkonosti Mahalanobis detekoru, jehož vzdálenostní threshold byl odhadnut jako průměr ze vzdálenosti od trénovacích dat.



Vliv chůze

Zkoumaný dataset kromě aktivity psaní v sedě obsahuje i aktivity psaní při chůzi. Na několika uživatelích bylo otestováno, že model naučený na aktivitách v sedě nerozezná aktivitu za chůze od téhož uživatele. To platí pro data z akcelerometru i gyroskopu, pro binární klasifikátor i detektor anomálií.

Soukromý dataset

Android aplikace pro sběr dat byla upravena, aby kromě běžně zkoumaných senzorů, jako je akcelerometr, gyroskop a magnetometr, shromažďovala navíc data z lineárního akcelerometru (bez vlivu gravitačního zrychlení), gravitačního senzoru a rotation vektor senzoru. Ke každému senzoru je navíc zjištěna i aktuální přesnost senzoru na škále 0-3. Obohaceny byly i data z dotykové plochy. Nově obsahují informaci elipse major a minor, což by mělo poukazovat na orientaci plochy dotyku. S touto úpravou se zatím podařilo získat data od pěti uživatelů, navíc 3 z nich byli dále řízeni k použití jednoho sdíleného telefonu pro lepší simulaci krádeže zařízení a odhalení nekonzistencí způsobené různou kalibrací senzorů.

Efekt posunutých hodnot akcelerometru byl pozorován i u soukromého datasetu. Modely telefonů skutečně mají různě kalibrovaný senzor, což vlivem gravitačního zrychlení způsobuje posun hodnot.

U lineárního akcelerometru nebyl podobný vliv pozorován. Naopak uživatel dosahoval podobných průměrných hodnot napříč různými zařízeními.

Rotation vektor senzor je fúzí akcelerometru a gyroskopu, jehož hlavním využitím jsou hry a rozšířená realita. Jeho hodnoty se mění v závislosti na změně orientace v prostoru. Pro behaviorální autentizaci nejspíš nepřinese žádnou další informaci.

V datasetu bylo nalezeno zařízení, které nemá dostupný gyroskop, ani lineární akcelerometr.

U dat z dotykové plochy dochází k asi největším rozdílům napříč zařízeními. Například Google Pixel má velmi citlivý senzor dotyku. Předává přesná data o velikosti plochy dotyku, dokonce i její orientaci. Navíc frekvence vzorkování je významně vyšší než u ostatních zařízení. Levná Xiaomi zařízení jsou tak mnohem méně přesné v měření doby, po kterou byl vykonáván dotyk a velikost dotykové plochy není předávána vůbec nebo jen jako násobek nějaké hodnoty.

Závěr

Veřejně dostupný dataset byl použit jako jakási zkratka k získání lepší představy o tom, jak by mohl systém pro behaviorální autentizaci fungovat. Zkoumaný dataset byl shromážděn pod dohledem výzkumného týmu za použití jedné sady smartphonů. Přesto existuje nezanedbatelná závislost výkonosti binárního klasifikátoru na vybraném vzorku zero-effort útočníků. Problém výběru vhodných útočníků k trénování modelu bude mnohem komplikovanější ve škálovatelném systému bez omezení na použitém zařízení. Vhodný trénovací útočník by měl používat zařízení, které systému předává podobně kvalitní data, má podobně stejnou velikost, tvar a navíc je samotný útočník jako osoba podobně uživatelsky schopný. Z toho důvodu považuji za rozumnější nalézt vhodný detektor anomálií, který je navíc výhodný z hlediska výpočetních zdrojů, paměťové náročnosti a snadného doučení modelu v případě driftu dat.

Dalším bodem je klasifikace aktivity uživatele, tedy minimálně toho zda je uživatel v pohybu nebo ne, nebo zda používá telefon položený na stole a data ze senzorů tak není žádoucí vyhodnocovat. Systém android poskytuje detekci aktivity pomocí Activity Recognition API. To však od androidu 10 potřebuje explicitní povolení při startu aplikace. Článek [6] to řeší pomocí vlastního klasifikátoru, rozeznávání mezi pohybem a setrváním považuje za dostatečné.

V průběhu několika měření byla vždy data senzorů agregována do sekundových oken. Rozhodně je nutné nad soukromým datasetem vyzkoušet jiné přístupy, jako například zohledňování dat jen v době dotyku. Nakonec bude potřeba nalézt způsob spojení jednotlivých modalit. Dosavadní analýzy nasvědčují, že bude vhodnější zvolit score-level fusion, tedy neslučovat příznaky do jednoho modelu, ale tvořit rozhodnutí na základě výstupu několika modelů.

Reference

[1] M. Georgiev, S. Eberz, T. H. G. Lovisotto a I. Martinovic, „Common Evaluation Pitfalls in Touch-Based Authentication,“ University of Oxford, 2022. [Online]. Available: https://www.semanticscholar.org/reader/cbf6956df6ce82d5c3cdd66c143565378d9d62d8.

[2] S. Eberz, K. Rasmussen, V. Lenders a I. Martinovic, „Evaluating Behavioral Biometrics for Continuous,“ University of Oxford, 2017. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3052973.3053032.

[3] M. Papamichail, K. Chatzidimitriou, T. Karanikiotis, N.-C. Oikonomou, A. Symeonidis a S. Saripalle, „BrainRun: A Behavioral Biometrics Dataset towards Continuous Implicit Authentication,“ Aristotle University of Thessaloniki, 2019. [Online]. Available: https://doi.org/10.3390/data4020060.

[4] Z. Sitová, J. Šeděnka, Q. Yang, G. Peng, G. Zhou, P. Gasti a K. S. Balagani, „HMOG: New Behavioral Biometric Features for Continuous Authentication of Smartphone Users,“ 2014. [Online]. Available: https://hmog-dataset.github.io/hmog/.

[5] C. Shen, „Performance Analysis of Multi-Motion Sensor Behavior for Active Smartphone Authentication,“ 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8006292.

[6] W.-H. Lee a R. B. Lee, „Implicit Smartphone User Authentication with Sensors and Contextual Machine Learning,“ 2017. [Online]. Available: https://arxiv.org/abs/1708.09754.

[7] M. Frank, „Touchalytics: On the Applicability of Touchscreen Input as a Behavioral Biometric for Continuous Authentication,“ 2012. [Online]. Available: https://ieeexplore.ieee.org/document/6331527.

[8] K. S. Killourhy a R. A. Maxion, „Comparing anomaly-detection algorithms for keystroke dynamics,“ 2009. [Online]. Available: https://ieeexplore.ieee.org/document/5270346.

[9] smazzanti, „mRMR (minimum-Redundancy-Maximum-Relevance) for automatic feature selection at scale.,“ 2021. [Online]. Available: https://github.com/smazzanti/mrmr.