Computers & Security 121 (2022) 102868

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

Evaluating multi-modal mobile behavioral biometrics using public datasets

Aratrika Ray-Dowling[a], Daqing Hou[a,*], Stephanie Schuckers[a], Abbie Barbir[b]

[a] Department of Electrical and Computer Engineering, Clarkson University, 8 Clarkson Avenue, Potsdam 13699, New York, USA
[b] Mobile Security Group, CVS Health, USA

ARTICLE INFO

Article history:
Received 21 February 2022
Revised 5 June 2022
Accepted 3 August 2022
Available online 5 August 2022

Keywords:
Performance evaluation
Behavioral biometric
Continuous authentication
Multi-Modality
Likelihood ratio-based score fusion
Support vector machine

ABSTRACT

Behavioral biometric-based continuous user authentication is promising for securing mobile phones while complementing traditional security mechanisms. However, the existing state of art perform continuous authentication to evaluate deep learning models, but lacks examining different feature sets over the data. Therefore, we evaluate the performance of user authentication based on acceleration, gyroscope (angular velocity), and swipe data from two public mobile datasets, HMOG (Hand-Movement, Orientation, and Grasp) (Sitová et al., (2015) dataset et al. (2015)) and BB-MAS (Behavioral Biometrics Multi-device and multi-Activity data from Same users) (Belman et al., (2019) dataset et al. (2019)) extracted with different feature sets to observe the variation in authentication performance. We evaluate the performances of both individual modalities and their fusion. Since the swipe data is intermittent but the motion event data continuous, we evaluate fusion of swipes with motion events that occur within the swipes versus fusion of motion events outside of swipes. Moreover, we extract Frank et al.'s (2012) Touchalytics features Frank et al. (2012) on the swipe data but three different feature sets (median, HMOG (Sitová et al. (2015)), and Shen's (Shen et al. (2017))) on the motion event data, among which the Shen's features were shown to perform the best. More specifically, we perform score-level fusion for a single modality utilizing binary SVMs (Support Vector Machine). Furthermore, we evaluate the fusion of multiple modalities using Nandakumar's likelihood ratio-based score fusion (Nandakumar et al. (2007)) by utilizing both one-class and binary SVMs. The best EERs (Equal Error Rates) of fusing all three modalities when using the one-class SVMs are 8.8% and 0.9% for HMOG and BB-MAS respectively. On the other hand, the best EERs in the case of binary SVMs are 1.5% and 0.2% respectively. Observing the better performances of BB-MAS compared to HMOG in swipe-based experiments, we examine the difference of swipe trajectory between the two datasets and find that BB-MAS has longer swipes than HMOG which would explain the performance difference in the experiments.

1. Introduction

Mobile phones serve manifold purposes in our everyday life. In addition to making calls, people use them to socialize with each other, store and process multimedia, shop online, make payments, and transfer funds. Such sensitive information and transaction require protection from non-legitimate users (impostors). Currently mobile phones are typically protected with one-time entry point authentication mechanisms like PINs, passwords, or biometrics that can be compromised easily. In addition, MFA (multi-factor authentication) is often used to further strengthen the entry-point authentication. Unfortunately the use of MFA often introduces friction, such as by requiring additional effort from the user to obtain and enter a security token (one time password).

On the other hand, behavioral biometrics authenticate users based on an analysis of data passively logged from the phone's embedded sensors. In contrast to conventional MFA, one advantage of behavioral biometrics is that they can be frictionless as the user is not required to perform any additional activities for securing the mobile phone. Motion events such as acceleration and angular velocity logged from phone's accelerometer and gyroscope sensors respectively can be used for securing the device against unauthorized users. The logged motion events capture a user's hand micromovements and are believed to capture an individual's unique behavioral traits. According to Sitová et al. (2015), when interacting with mobile phones in hand, users strive to achieve stability and precision. As a result, each user develops their own postural preference. Furthermore, factors such as hand size, grip strength,

* Corresponding author.
E-mail address: dhou@clarkson.edu (D. Hou).

https://doi.org/10.1016/j.cose.2022.102868
0167-4048/

age, and gender constitute a user's physiological traits. Both postural preference and physiological traits are believed to contribute to the uniqueness of user behaviors (Sitová et al., 2015). Due to their high availability, behavioral biometrics can also be used to continuously authenticate a user beyond the initial log-in.

The state of the art on behavioral biometrics-based authentication have been evaluated on private datasets with generally a small number of users (Amini et al., 2018; Deb et al., 2019; Ehatisham-ul Haq et al., 2018; Incel et al., 2021; Kumar et al., 2015; 2016a; Roy et al., 2015). However, there are inadequate evaluations of algorithms on public datasets (Centeno et al., 2017; Karakaya et al., 2019; Neverova et al., 2016; Volaka et al., 2019). To establish the generalizability of the performance of behavioral biometrics, more benchmarking of existing algorithms on large public datasets is necessary. Therefore, we have authenticated users by fusion of multi-modalities (acceleration, gyroscope, and swipe) from two large public datasets, namely, Sitová et al.'s (2015) HMOG (Hand Movement, Orientation, and Grasp) (dataset et al., 2015) and Belman et al.'s (2019) BB-MAS (Behavioral Biometrics Multi-device and multi-Activity data from Same users) (dataset et al., 2019), both of which have at least 100 users. We choose the HMOG dataset as it logs the users hand micro-movements through sensor events. On the other hand, the BB-MAS dataset is a new dataset with modalities and user behavior of interest similar to HMOG. Given that BB-MAS is relatively new, evaluations of the motion event modalities and its fusion with touch event data from this dataset have not been done so far. Moreover, we extract three different kinds of feature sets from each of the motion event modalities, which are, low-level median feature, Sitová et al.s (2015) (Sitová et al., 2015), Shen et al.s (2017) (Shen et al., 2017) statistical features. No prior work has evaluated two public datasets over the three different feature sets. On the other hand, there are existing state of art (Abuhamad et al., 2020; Buriro et al., 2021; Centeno et al., 2017; Neverova et al., 2016; Volaka et al., 2019) that utilize deep learning architecture to authenticate users but without explicitly computing features on the data.

In our feature-based evaluation of multiple modalities from the two datasets we choose HMOGs grasp and stability based features (Sitová et al., 2015) because it captures the hand-micromovements of users while they are sitting and typing. In the case of stationary user behavior, the micro-movements of hand gestures are believed to be unique across users. We select Shen et al.s (2017) statistical features (Shen et al., 2017) because it will capture the unique distribution characteristics of motion event data across users. Lastly, we also extract low-level median feature to observe how far behind it falls from the two high-level feature sets. From the swipe data we extract Frank et al.s (2012) Touchalytics features (Frank et al., 2012) which capture the unique trajectory and speed of swipes across all users. Table 3 lists all the features under each feature set which we extract from motion event and swipe data. The data extracted with the above features is classified using both one-class classifier (OCC) and binary classifier (BC).

Our work is the first to apply Nandakumar et al.'s (2007) likelihood ratio-based score fusion (LR) to multi-modal behavioral biometrics, whereas Nandakumar et al. (2007) applies LR to only physiological biometrics (iris, fingerprint, and speech) datasets (Nandakumar et al., 2007). In their work, they categorize score level fusion techniques into transformation-based, classification-based, and density-based. The authentication performance of likelihood ratio-based score fusion is better than any other score fusion techniques. The Gaussian Mixture Model (GMM) used proves to be effective in modeling the genuine and impostor score densities. The combined match scores vector constitutes scores of all three physiological biometrics. Using the match scores the likelihood ratio is calculated as ratio of the genuine to impostor distribution (Nandakumar et al., 2007). In this work we utilize the generated match scores of two (acceleration and gyroscope) and three (acceleration, gyroscope, and swipe) behavioral biometric modalities to perform likelihood ratio-based score fusion.

We evaluate the performance of user authentication through both single modality and the fusion of two/three modalities. Furthermore, we estimate the overall authentication performance based on the availability of modalities by a weighted sum of the swipe-driven authentication and continuous motion events through a hybrid (combined) experiment. Between the two datasets, BB-MAS (dataset et al., 2019) consistently performs well than HMOG (dataset et al., 2015) in both one-class and binary classification experiments. Therefore, we investigate the cause of the difference in performances between the two datasets across most experiments. We notice the effect of concept drift due to HMOG's data collection method which adds the effect of behavioral adaptation in users visiting for days. Whereas, this factor is absent in the case of BB-MAS. Additionally, we observe the nature of swipes between the two datasets. Examining the nature of swipes, we find significant difference in the statistics of swipe trajectories throughout the swipe lengths between the two datasets. HMOGs (dataset et al., 2015) swipes being shorter than BB-MAS (dataset et al., 2019) produce less significant feature magnitudes throughout its trajectories which impact authentication performance in the swipe-based experiments.

Therefore, the contributions of this work can be summarized as follows:

1. Evaluating the authentication performance of fusing three modalities (acceleration, gyroscope, and swipe) using two public datasets (HMOG (dataset et al., 2015) and BB-MAS (dataset et al., 2019)). No prior work has done this with both datasets.
2. Evaluating three different feature sets (median, HMOG (dataset et al., 2015), and Shen (Shen et al., 2017)) on acceleration and gyroscope where each set highlights different characteristics of the two motion modalities.
3. Evaluating both binary and one-class classifiers using both datasets. Although it is common to use binary classifiers in authentication, a one-class classifier becomes necessary when impostor data is not available.
4. Applying Nandakumar et al.s (2007) likelihood ratio-based score level fusion (Nandakumar et al., 2007) to multi-modal behavioral biometric data.
5. Exploring possible causes for BB-MAS outperforming HMOG (Section 5.3): a) the average swipe trajectory of BB-MAS is longer than HMOG, b) the presence of concept drift in HMOGs acceleration and gyroscope.

The rest of the paper is organized as follows: Section 2 presents related work. Section 3 describes the two public datasets. In Section 4 we discuss the experimental procedures. Section 5 reports all the experimental results. Lastly, Section 6 concludes our study.

2. Related work

The criteria of selecting existing state of art closely related to our study is to choose works that have motion events (acceleration and/or gyroscope) as modalities logged during different user behaviors (typing, swiping, picking up phone, sitting, standing, walking, and others) for authentication. Table 1 describes selected state of art on behavioral biometric-based authentication that has both acceleration and gyroscope as modalities or both in combination with other modalities. To put this work in the context of the related work, the table highlights, for each related work, the datasets, number of data providers, user behavior, duration of device usage,

Table 1
Comparative literature review. AB: AdaBoost, ACC: Accuracy, AUC: Area Under the Curve, BC: Binary Classifier, DT: Decision Trees, EE: Elliptic Envelope, EL: Ensemble Learning, EER: Equal Error Rate, FAR: False Acceptance Rate, GBC:Gradient Boosting Classifier, HMM: Hidden Markov Model, IF: Isolation Forest, kNN: k Nearest Neighbor, LDA: Linear Discriminant Analysis, LOF: Local Outliers Factor, LReg: Logistic Regression, LSTM: Long Short-Term Memory, MLP: Multilayer Perceptron, NB: Naive Bayes, OCC: One Class Classifier, PABG: Phone Acceleration-based Gait Biometric, RF: Random Forest, RNN: Recurrent Neural Network, SE: Scaled Euclidean, SM: Scaled Manhattan, SVM: Support Vector Machine, SPMP: Swiping and Phone Movement Patterns, Sess.: Session, TAR: True Acceptance Rate. WABG: (Smart)watch Acceleration-based Gait, WRBG: (Smart)watch Rotation-based Gait.

Study & Dataset	#User	Behavior	Duration	Modality	Sampling (Hz)	Algorithm	Fusion	Best performance
Roy et al. (2015), own	42	handwriting	1 Sess.	Accel, Gyro, Touch	-	OCC-HMM	feature	0%, EER
Deb et al. (2019), own	37	routine usage	15 days	Accel, Gyro, Magneto, Rotate, Key, GPS, Gravity	1	OCC-LSTM	score	99.92%, TAR at 0.1% FAR
Abuhamad et al. (2020), own	84	routine usage	5 days	Accel, Gyro, Touch, Magneto, Elevation	64	BC-RNN	sensor (data)	0.09%,EER
Li et al. (2018), own	100	routine usage	6 hr	Accel,Gyro	100	OCC-SVM	feature	4.66%,EER
Gascon et al. (2014), own	315	typing	1 Sess.	Accel, Gyro, Rotate, Key	-	BC-SVM	feature	80%,AUC
Kumar et al. (2016a), WABG, WRBG (own)	40	wear smartwatch and walk	2 Phase (3 months apart) each with 2 Sess. (10 min apart)	Accel, Gyro	25	BC-kNN, LReg, RF, MLP,	feature, score	95%,ACC
Papamichail et al. (2019), BrainRun (own)	2218	game playing	-	Accel, Gyro, Swipe, Magneto, Tap	10	-	-	-
Li et al. (2020), own, BrainRun (Papamichail et al., 2019)	100, 82	routine usage	24 Sess. (∼60 hr)	Accel, Gyro	100, 10	OCC-SVM	sensor (data)	5.14%, EER
Kumar et al. (2018), PABG (Kumar et al., 2015), WABG (Kumar et al., 2016a), WRBG (Kumar et al., 2016a), SPMP (Kumar et al., 2016b)	18, 40, 28	gait, typing in sitting	-	Accel, Gyro	46, 25	OCC-SVM, LOF, IF, EE; BC-AB, NB, kNN, LDA, LReg, MLP, RF, SVM	score, decision	94.22%, ACC
Amini et al. (2018), own	47	browse shopping app	10-13 min	Accel, Gyro	100	LSTM, SVM, RF, LReg, GBC	sensor	96.7%, ACC
Incel et al. (2021), own	45	phone usage in sitting and standing	15 Sess. (22.5 min)	Accel, Gyro, Magneto, Scroll	100	BC-SVM, kNN, MLP, DT, RF, NB, EL; OCC-SVM	Feature	3.2%, EER
Volaka et al. (2019), HMOG dataset et al. (2015)	100	read, write, navigate (sit/walk)	24 Sess. (∼60 hr)	Accel, Gyro, Touch	100	BC-LSTM	Feature	15%, EER
Ehatisham-ul Haq et al. (2018), own	10	walking, sitting, standing, running, walking up and down stairs	90 min	Accel, Gyro, Magneto	50	BC-SVM, DT, kNN	Feature	100%, ACC (SVM)
Shen et al. (2017), own	102	routine usage	3 rounds	Accel, Gyro, Swipe, Magneto, Rotate	-	OCC-HMM	feature	4.74%, EER
Belman et al. (2019), BB-MAS (own)	117	multiple activities	1 Sess. (1.8 hr)	Accel, Gyro, Swipe, Key, Mouse	100	-	-	-
Sitová et al. (2015), HMOG (own)	100	read, write, navigate (sit/walk)	24 Sess. (∼60 hr)	Accel, Gyro, Swipe, Tap, Key, Magneto, Pinch	100	OCC-SM, SE, SVM	score	7.16%, EER
Ray et al.(2021)(Our previous work) (Ray et al., 2021), own	49	Android form filling in siting	2 Sess. (intra, inter)	Accel, Gyro	2	BC-SVM	score (weighted, LR)	2.4%, EER (intra); 6.9%, EER (inter)
This work, HMOG (dataset et al., 2015), BB-MAS (dataset et al., 2019)	100, 115	typing in sitting	4 Sess., 25 min	Accel, Gyro, Swipe	100, 100	OCC, BC-SVM	score (LR Nandakumar et al. (2007))	0.2%, EER (BC-Accel, Gyro, Swipe); 0.9%, EER (OCC-Accel, Gyro, Swipe)

modalities, sampling rate of motion events, algorithms evaluated, fusion type, and performance measurement.

Non-gait-based studies involving only acceleration and gyroscope are discussed as follows. Li et al. (2020) evaluate SCANet continuous authentication platform over 100 recruited volunteers for their own dataset and 82 selected user data from the BrainRun (Papamichail et al., 2019) dataset. For their own data the user behavior includes reading, writing, and map navigation activities over 24 sessions. Investigating the combination of acceleration and gyroscope over one-class SVM, they achieve an EER of 2.35% as the best performance on their own dataset. Kumar et al. (2018) compare the performances of several one-class classifiers (OCC) with binary classifiers (BC) utilizing four datasets. Among experiments performed on individual OCC, BC and fusion of multiple OCCs, the kNN (k-Nearest Neighbor) BC produces the best result of 94.22% accuracy. However, the user activity of the four datasets include both non-gait and gait-based behaviors.

The following state of art involve both acceleration and gyroscope in combination with other modalities to enhance the authentication performance. Abuhamad et al. (2020) evaluate their authentication platform AUToSen on their own dataset of 84 volunteers. Combining several motion events and a touch event at sensor (data) level, they achieve a best performance of 0.09% EER. Roy et al. (2015) implement an HMM (Hidden Markov Model)-based multi-sensor system, which is evaluated on their own dataset of 42 volunteers. The user activity includes reading Wikipedia articles and filling out a questionnaire through which they log modalities like swipe, tap, acceleration, and gyroscope. Utilizing a single swipe observation they achieve an EER of 13.29% which improves to 0% when 19 consecutive swipes are combined. A similar pattern is observed in the case of taps where the EER improves from 16.55% to 1% when 17 consecutive taps are consolidated. Shen et al. (2017), combine multiple sensor events like acceleration, gyroscope, magnetometer, and orientation and a subset of the four sensor events. However, the combination of all the four motion sensor modalities produce the best EER of 4.74%. Incel et al. (2021) investigate authentication performance over 15 sessions when users are interacting with smartphones in hand while sitting and standing and in sitting when the device is on the table. Applying binary classifier on the entire collected dataset they achieve a best performance of 3.5% EER. Other similar studies are by Gascon et al. (2014) and Cai and Chen (2012) where motion events are combined with keypress.

A single motion event modality (one of acceleration or gyroscope) combined with other modalities is utilized to perform authentication in the following studies. Kumar et al. (2016b) investigate the fusion of phone movement patterns (acceleration) with typing and swiping when a user uses a web browser in sitting, achieving an accuracy of 93.33% for a feature fusion of movement and swipes, and 89.31% for a score fusion of movement and typing. Kim and Kang (2020) authenticate users based on typing in English and Korean languages where keypress is fused with acceleration and touch events logged during typing. They authenticate 50 users achieving EERs as low as 0%. On a dataset of 39 users Crawford and Ahmadzadeh (2017) perform authentication based on keypress and gyroscope achieving 97.7% Area Under the Curve (AUC).

Motion events are also captured while users perform special activities like picking up a phone call. In Carlson et al. (2015), acceleration and gyroscope are utilized to capture the user behavior of taking out phone from pocket to ear, holding phone to ear, and putting back phone from ear to pocket. In total 10 users are classified for each of the three gestures using Multi-Layer Perceptron. Out of the three gestures, the best accuracy of 88% is achieved when users are holding phone to an ear. In a similar study by Buriro et al. (2015) users perform a special behavior of slide swiping while unlocking phone, then putting phone to an ear, and speak over phone while sitting and walking. Both acceleration and gyroscope are logged when users put phone to their ears after unlocking. They achieve a Half Total Error Rate (HTER) of 7.33% as the best using Bayesian Network among other classifiers.

Frank et al. (2012) authenticate 41 users over 3 sessions through only swipe modality. On combining up to 12 consecutive swipes they achieve 0%, 2–3%, and 4% EERs in the intra, inter, and inter-week sessions respectively. The work by Xu et al. (2014) also authenticate users through touch events (keypress, swipe, and pinch). They perform both discrimination and authentication experiments achieving a best EER of 2% on combining 11 consecutive swipes.

The state of art in using acceleration and/or gyroscope for authentication through gait-based user activity are discussed as follows. Thang et al. (2012) perform user authentication while users are walking a distance while the sensor devices (logging acceleration) are attached to their hips. Ehatisham-ul Haq et al. (2018) perform six user behaviors, namely walking, sitting, standing, running, walking upstairs, and walking downstairs when the sensor device is kept in 5 locations of a user's body (left, right jeans pocket, waist, upper arm, and wrist). They authenticate users through acceleration, gyroscope, and magnetometer achieving 100% accuracies from walking and running behaviors.

In our previous work (Ray et al. (2021)) utilizing acceleration and gyroscope we perform continuous authentication on mobile devices collecting our own dataset of 49 seated users. Fusing the two modalities at weighted score level and likelihood ratio-based score level, we observe best EERs of 2.4% and 6.9% for intra- and inter-session experiments respectively. Between the two score fusion techniques, the likelihood ratio-based score fusion performs the best in both intra-session and inter-session (with effect of concept drift) experiments. Therefore, in the present work we utilize two larger public datasets, HMOG (dataset et al., 2015) and BB-MAS (dataset et al., 2019), with 100 and 115 users respectively and perform fusion-based experiments using likelihood ratio-based score fusion. We extract the motion event data over three different feature sets and the swipe data over Touchalytics feature set (Frank et al., 2012). We perform multi-modalities fusion experiments taking both one-class and binary SVMs.

3. The public datasets

In this work, we use two public datasets, namely, HMOG (dataset et al., 2015) and BB-MAS dataset et al. (2019). These are large-scale datasets having multiple modalities collected from at least 100 users. The objective of HMOG is to capture data through three different activities (reading, writing, and map navigation) which are used to evaluate the new modality, HMOG, as described in Sitová et al. (2015). The HMOG dataset involves two user behaviors, namely, sitting and walking, while performing each of the above activities whereas BB-MAS captures routine usage traits of the same user across different devices (desktop, tablet, and Android mobile phones). It involves several user behaviors while logging the data like sitting, walking on the corridor, and walking up and down a staircase. We are interested in authenticating users utilizing three modalities (from each dataset) namely acceleration, gyroscope, and swipe which are logged from mobile devices while each user is sitting and typing/writing.

3.1. HMOG dataset

The HMOG (dataset et al., 2015) data has 100 recruited participants. The data is collected on Android mobile phones only. There are $8*3 = 24$ sessions in total involving several activities like reading, writing, and map navigation. Out of the 8 typing/writing sessions, we extract data from the 4 sessions (3, 9, 15, 21) that require users to sit and type. This activity takes approximately 20 to

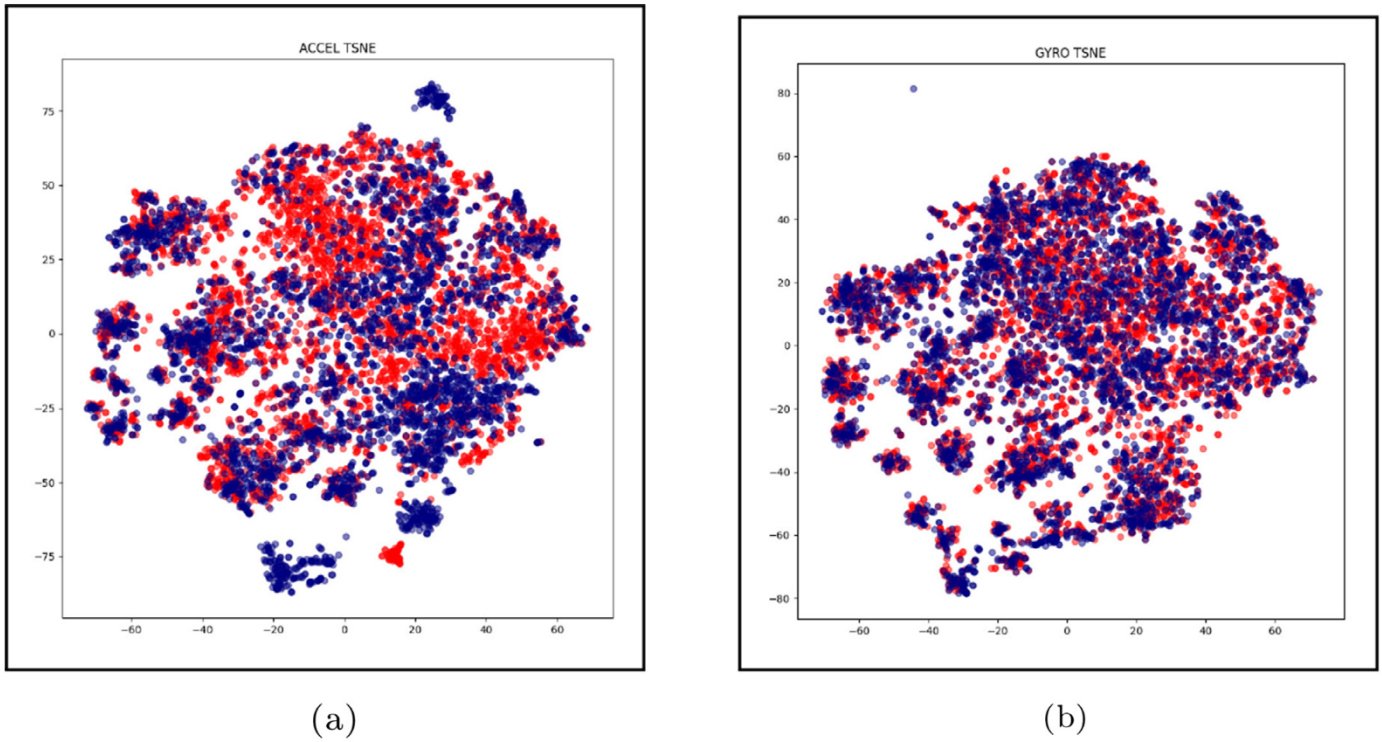

Fig. 1. t-SNE plot of BB-MAS motion data with HMOG features (poor performing genuine user) (a) acceleration, (b) gyroscope..

60 minutes to complete. Each user is asked to perform three free text typing tasks where each answer is of approximately 250 characters. Users visit for multiple days to finish the entire task.

3.2. BB-MAS dataset

The public data, BB-MAS (dataset et al., 2019), has 117 recruited volunteers who provide data on multiple interfaces, while performing several activities. We are interested in the mobile phone data which is logged when users are typing while sitting. Out of the 117 users, there are two users whose data cannot be used. The sequence of activities that the users need to perform on mobile phones are typing two pieces of static texts of approximately 112 characters each, followed by ten questions whose answers must be of at least 50 characters each. The layout of the questions makes users swipe vertically and horizontally in between. Within one visit users need to finish the entire task of logging data on multiple devices which takes around 2 hours (110 minutes) in total. The duration spent on mobile phone while users are seated is around 25 minutes.

3.3. Data statistics, feature extraction, and visualization

Each motion event data is logged along x, y, and z axes and we further compute the resultant of the motion event as the square root of the sum of the squares of the motion event along x, y, and z axes:

$$resultant = \sqrt{x^2 + y^2 + z^2}$$

The motion events are originally logged at 100 Hz sampling rate per sensor. There are two cases of feature extraction from such motion events. In the first case we extract features from a time window of 500 ms. However, in our pilot studies, we try 50 ms, 100 ms, 200 ms, and 500 ms time windows where 500 ms produces the best results. The second case is driven by the availability of swipes where features are extracted from all the motion events that fall within each swipe. In both cases, we extract three kinds of features from the motion event, namely, the medians of the motion events, Shen et al. (2017) features and HMOG features (Sitová et al., 2015). Shen's features include both descriptive statistics and intensive features (i.e., energy and entropy). HMOG's features measure stability and precision of hand micromovements of users. In case of swipes, we have extracted Frank et al.s (2012) Touchalytics features (Frank et al., 2012), which measure the distance, movement, and temporal attributes of swipes. Table 3 shows the computed features per feature set. We do not perform any feature selection method because we want to evaluate the public data on the entire feature set as proposed and experimented in the original state of art (Frank et al., 2012; Shen et al., 2017; Sitová et al., 2015). Table 2 shows the statistics of the modalities of interest, per dataset, after they are processed for our experiments.

As shown in Figs. 1, 2, 3, and 4, we apply the t-SNE (t-distributed stochastic neighbor embedding) (Van der Maaten and Hinton, 2008) dimensionality reduction algorithm to visualize the high-dimensional motion event data, where blue and red dots represent genuine and impostor samples respectively. This process reveals that in the BB-MAS motion event data, extracted with HMOG (Sitová et al., 2015) features, there is a considerable overlap between genuine and impostor user samples in case of both motion modalities (acceleration and gyroscope), when the genuine user is a bad performing user or does not perform well in the authentication (Fig. 1). On the other hand, as shown in Fig. 2, in case of a good performing genuine user, the acceleration modality is the primary reason for the user to perform well, whereas the gyroscope alone is not enough to successfully authenticate a good performing user. Similar patterns are observed in case of the HMOG data when extracted with Shen et al.'s (2017) features (Shen et al., 2017), as shown in Figs. 3 and 4. Hence, from the visualizations we understand that fusion of the modalities would enhance the authentication performance.

Fig. 5 shows the t-SNE plots of BB-MAS (dataset et al., 2019) and HMOG (dataset et al., 2015) swipe data extracted with Touch-

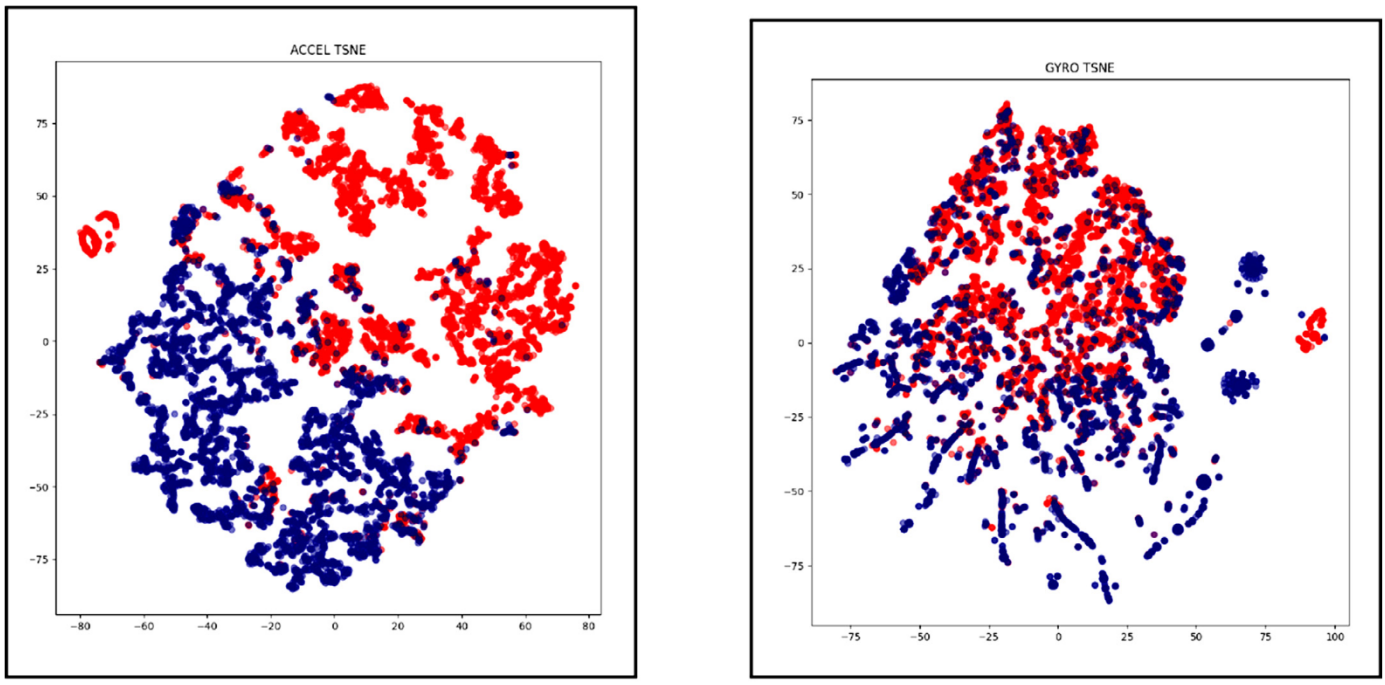

Fig. 2. t-SNE plot of BB-MAS motion data with HMOG features (good performing genuine user) (a) acceleration, (b) gyroscope.

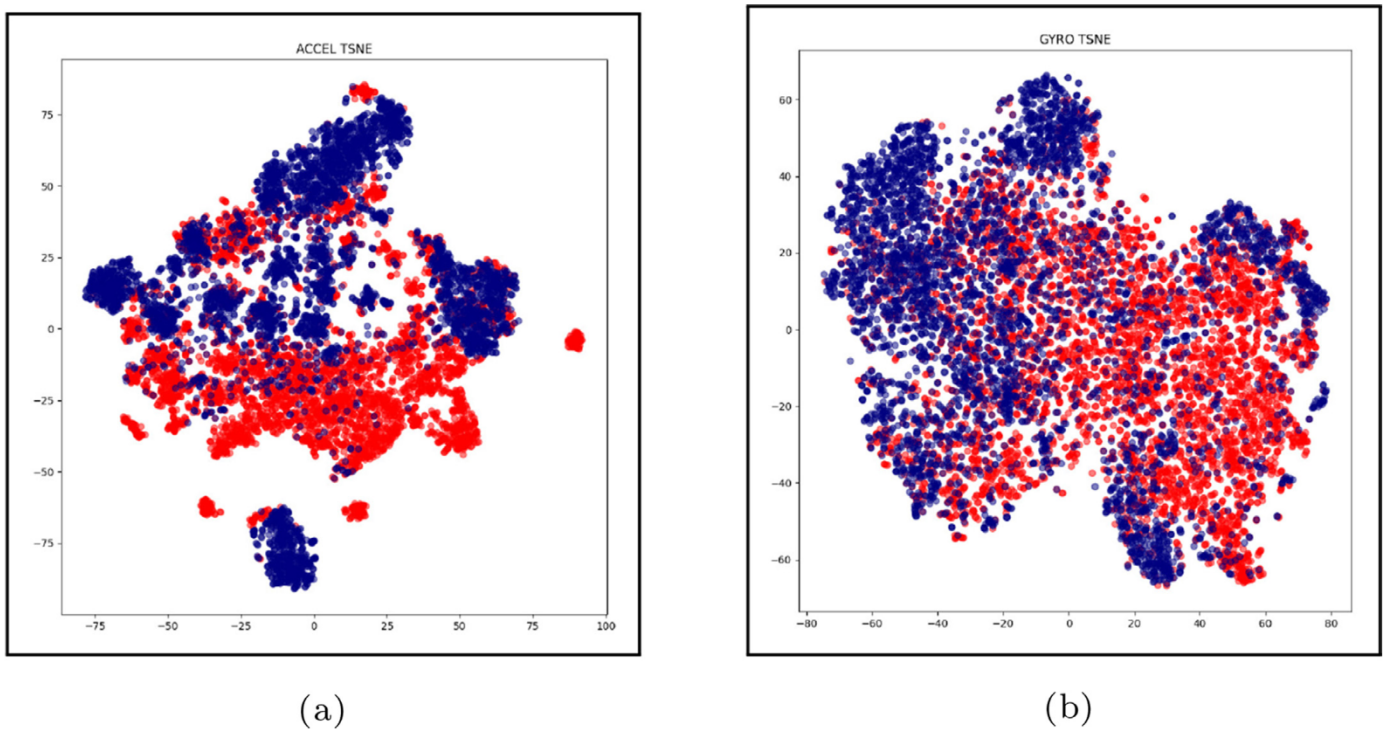

Fig. 3. t-SNE plot of HMOG motion data with Shen features (poor performing genuine user) (a) acceleration, (b) gyroscope..

Table 2
Statistics of the number of rows of data of the two public data sets: HMOG (dataset et al., 2015) and BB-MAS (dataset et al., 2019) after processing for experiments. AVG-Average, MED-Median, MIN-Minmum, MAX-Maximum, STDV-Standard Deviation.

	HMOG data statistics						BB-MAS data statistics					
Input Data	Total	AVG	MED	MIN	MAX	STDV	Total	AVG	MED	MIN	MAX	STDV
Only swipes	46,235	462.35	376.5	25	2521	395.14	22,265	193	185	85	385	51
Motion within swipes	2,716,415	27,164	25,046	353	75,174	15,326	1,483,225	12,897	11,519	4,319	44,645	6,180
Motion outside swipes	17,720,578	177,205	164,248	50,265	445,780	73,765	23,825,826	207,181	202,793	113,541	296,010	38,483

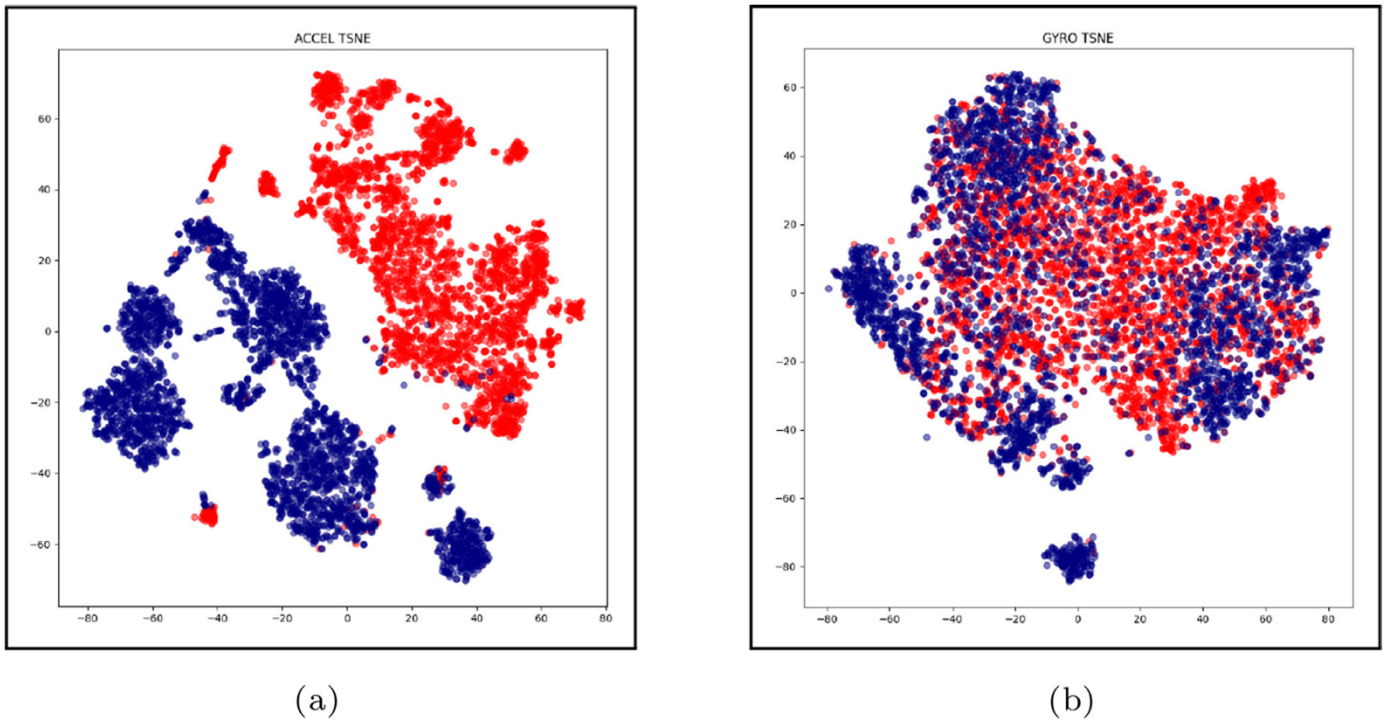

Fig. 4. t-SNE plot of HMOG motion data with Shen features (good performing genuine user) (a) acceleration, (b) gyroscope..

alytics features (Frank et al., 2012). In case of BB-MAS swipe data, Figs. 5(a) and 5(b) show less and more overlap respectively of genuine (blue) samples with impostor (red) samples. More overlap between genuine and impostor samples results in poor authentication performance for a genuine bad performing user. On the other hand less overlap shows better authentication performance for a good performing genuine user. Similar trend is observed in case of HMOG swipe data as shown in Figs. 5(c) and 5(d). In addition to this, the t-SNE plots for both HMOG and BB-MAS swipe data show a cluster formation in the genuine (blue) samples in case of the good performing user. However, the bad performing user does not show such pattern. Therefore, fusion of swipes with other modalities may enhance the authentication performance of bad performing genuine users.

Therefore, to enhance the authentication performance of the bad performing genuine users, our hypothesis is to fuse maximum number of modalities available at an instant such that the strongest modality among multiple can overcome the misclassification caused by the weaker performing modality/modalities.

4. Experimental procedures

This section describes the training and testing split for both binary classifier (BC) and one-class classifier (OCC), grid search ranges for tuning classifier hyperparameters and fusion parameters, and fusion methodologies (score fusion of k readings, likelihood ratio).

4.1. Design overview

In this work we have utilized both one-class and binary SVMs. Although it has been common to evaluate multiple classifiers on same datasets, our focus is not to evaluate performance of multiple classifiers but rather to compare the performance of the fusion methods over two datasets extracted with three different feature sets. Since SVM is considered a state of the art machine learning classifier that has the capability to classify data with large input feature dimensions, we select it as the classifier in our classification pipeline (as a controlled variable).

We train an SVM classifier for each of acceleration, gyroscope, and swipe modalities for both HMOG (dataset et al., 2015) and BB-MAS (dataset et al., 2019) datasets. Using SVM as our classifier, we perform score-level fusion for single modality-based authentication (Section 4.2) and likelihood ratio-based score fusion (Section 4.3) to combine multiple modalities.

We experiment with both binary and one-class SVMs with the Radial Basis Function (RBF) as kernel. We split each users data into training and testing sets. A 10-fold shuffling is performed where the data per user is divided into ten equal portions and from each of the ten shuffles, 9/10 of the portions is used for training and the remaining 1/10 portion is used for testing. This method will mitigate the effect of overfitting. We stop at 10-fold shuffle to have substantial data points per shuffle. The binary classifier, gets trained with both genuine and impostor data. We train the SVM using one genuine user and 50% random users from the impostor set (half from the total impostor set) following the 10-fold shuffling. The 1/10 portion of the data from the genuine users and each of the other random 50% impostors (other non-overlapping half from the impostor set) are used as the test data. Therefore, we train and test the classifier with non-overlapping sets of impostors which ensures robustness of the system.

In case of the one-class classifier, we train using the 9/10 portions of data of the genuine user. We then test the classifier with the other 1/10 portion of the genuine users data and 1/10th portion from each of the 50% random impostors data (random half from the impostor-set).

We perform grid searches to tune several parameters, namely, k (sliding window: these are the number of the consecutive scores generated by an SVM per modality which are fused by averaging the distance scores to reach a final decision); n (step size of k); binary SVM parameters C and *gamma*; one-class SVM parameters *gamma* and *nu*; the K_g (genuine Gaussian components) and K_i (im-

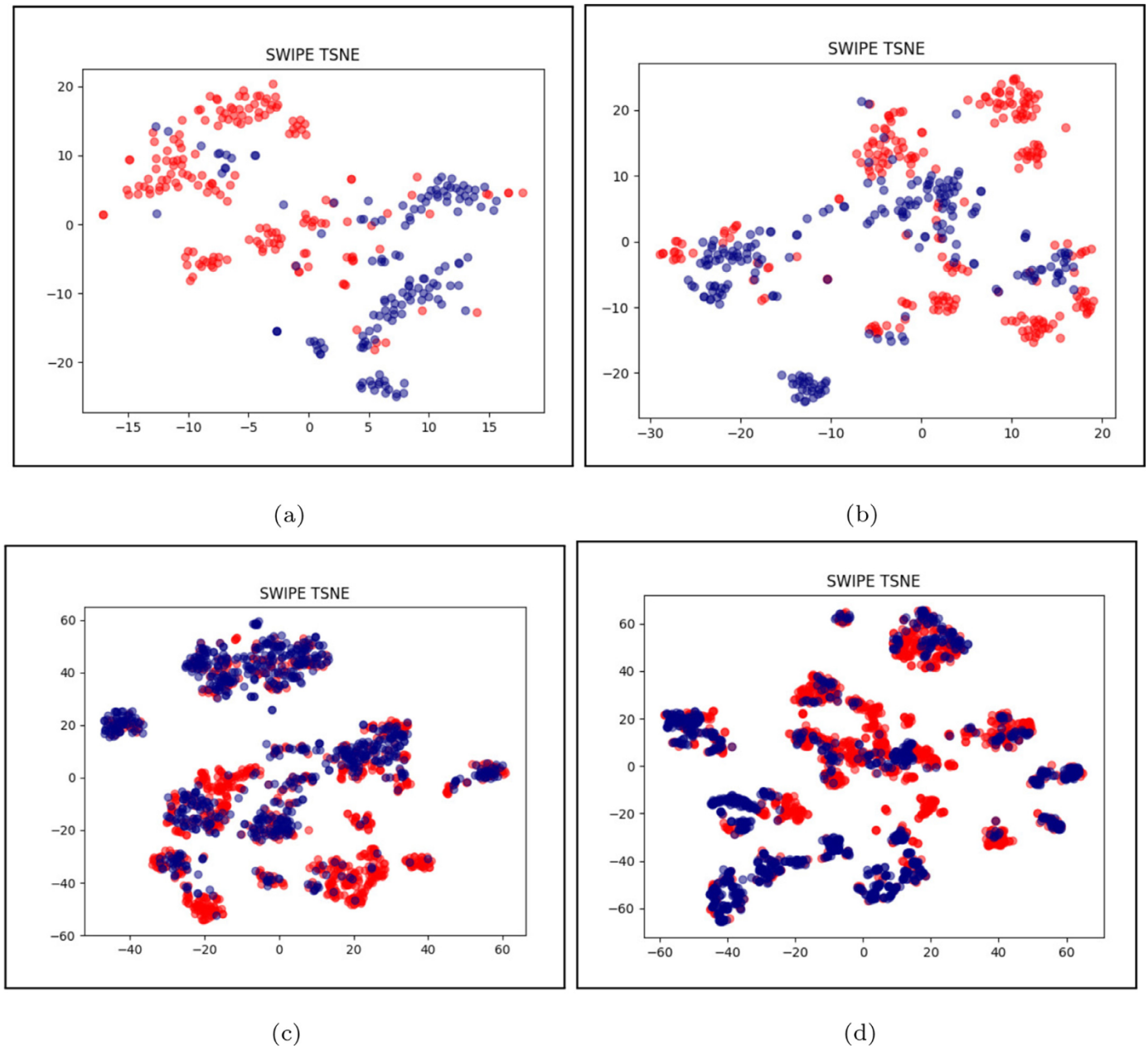

Fig. 5. t-SNE plot of swipe data extracted with Touchalytics features (a) BB-MAS swipe data good performing genuine user, (b) BB-MAS swipe data poor performing genuine user, (c) HMOG swipe data good performing genuine user, (d) HMOG swipe data poor performing genuine user..

postor Gaussian components) in the likelihood ratio-based score-level fusion.

We measure the authentication performance with Equal Error Rate (EER). EER is a performance metric utilized in measuring the biometric performance of a user authentication system. It predetermines the threshold value at which the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) are equal. We calculate EER per user in all experiments and estimate the overall performance of the system by computing average, median, minimum, maximum and standard deviation of EERs across all users.

4.2. Score-level fusion for single modality

In the single modality experiment, we utilize swipe data and train one binary SVM per user to measure the authentication performance of that genuine user against all impostors. To improve the performance, we apply a score-level fusion by averaging the distance scores of k consecutive swipe readings (scores) from the binary classifier and calculate an EER for each user.

4.3. Likelihood Ratio (LR)-based Fusion (Nandakumar et al., 2007) for multiple modalities

We apply Nandakumar et al.'s (2007) likelihood ratio-based score fusion (Nandakumar et al., 2007). First we train an SVM classifier for each of acceleration, gyroscope, and swipes. In case of the two modalities fusion, we take the two dimensional vectors of match scores of acceleration and gyroscope from their respective SVM classifiers and create genuine and impostor distributions. Similar genuine and impostor distributions are created during the fusion of all three modalities (acceleration, gyroscope, and swipes). *LR*, which is defined as the ratio of the genuine to the impostor distribution, is then used as a new match score for a test sample:

Table 3
Features extracted on motion event and swipe data.

Study	Modalities on which features are extracted	Features extracted
Shen (Shen et al., 2017)	Acceleration and Gyroscope	**Descriptive features**-Mean, minimum, maximum, range, variance, kurtosis, 30 to 80 quantiles, skewness, cross mean rate (median absolute deviation); **Intensive features**-energy, entropy. Each feature extracted over all columns (x, y, z, resultant) per motion-event.
HMOG (Sitová et al., 2015)	Acceleration and Gyroscope	**Grasp resistance features**-Mean during swipes/time-window, standard deviation during swipes/time-window, difference in reading before and after a swipe/time-window, net change in readings caused by a swipe/time-window, maximum change in readings caused by a swipe/time-window; **Grasp stability features**-time duration to achieve movement and orientation stability after a swipe/time-window, normalized time duration for mean sensor to change before and after a swipe/time-window, normalized time duration for mean sensor value to change from maximum swipe/time-window to average 100ms after swipe/time-window. Each feature extracted over all columns (x, y, z, resultant) per motion-event.
Median Touchalytics (Frank et al., 2012)	Acceleration and Gyroscope Swipe	Median during swipes/time-window. Each feature extracted over all columns (x, y, z, resultant) per motion-event. inter-stroke time, stroke duration, start x, start y, stop x, stop y, direct end-to-end distance, mean resultant length, up/down/left/right flag, direction of end-to-end line, 20% 50% 80% percentiles of pair-wise velocity, 20% 50% 80% percentiles of pair-wise acceleration, median velocity at last 3 points, largest deviation from end-to-end line, 20% 50% 80% percentiles deviation from end-to-end line, average direction, length of trajectory, ratio of end-to-end distance and length of trajectory, average velocity, median acceleration of first 5-point, mid-stroke pressure, mid-stroke area covered, mid-stroke finger orientation

$LR = \hat{f}_{gen}(x)/\hat{f}_{imp}(x)$ where $\hat{f}_{gen}(x)$ and $\hat{f}_{imp}(x)$ are the estimated genuine and impostor density functions, respectively, and x is a 2 or 3 dimensional vector of match scores of [acceleration, gyroscope] or [acceleration, gyroscope, swipe]. By taking the mean of k LRs as a final match score, we calculate an EER for each user authenticated against other impostors.

The genuine and impostor distributions are modeled as a mixture of Gaussian components. The genuine distribution is defined as:

$\hat{f}_{gen}(x) = \sum_{j=1}^{M_{gen}} P_{gen,j}\phi^K(x; \mu_{gen,j}, \Sigma_{gen,j})$ and the impostor distribution is defined as:

$\hat{f}_{imp}(x) = \sum_{j=1}^{M_{imp}} P_{imp,j}\phi^K(x; \mu_{imp,j}, \Sigma_{imp,j})$

Note that ϕ^K is a K-variate Gaussian density function with mean μ, and covariance matrix Σ:

$$\phi^K(x; \mu, \Sigma) = (2\pi)^{-K/2}|\Sigma|^{-1/2}exp(-1/2(x-\mu)^T\Sigma^{-1}(x-\mu))$$

M_{gen} (M_{imp}) is the number of mixture components used to model the density of the genuine (impostor) scores. $P_{gen,j}(P_{imp,j})$ is the weight assigned to the j^{th} mixture component in $\hat{f}_{gen}(x)$ ($\hat{f}_{imp}(x)$). The weights assigned to the j-components must sum up to one:

$\sum_{j=1}^{M_{gen}} P_{gen,j} = 1$ and $\sum_{j=1}^{M_{imp}} P_{imp,j} = 1$

$\mu_{gen,j}$ ($\mu_{imp,j}$) and $\Sigma_{gen,j}$ ($\Sigma_{imp,j}$) are the mean and covariance matrix of the jth Gaussian, respectively.

In the OCC scenario, we do not have the impostor distribution. To calculate LR, we express $\hat{f}_{imp}(x)$ as follows:

$\hat{f}_{imp}(x) = 1 - \hat{f}_{gen}(x)$

4.4. Binary Classifier (BC) experiments

We apply binary SVMs on both datasets and conduct experiments with 1, 2, and 3 modalities. Grid search is used to tune the best SVM hyperparameters, C and *gamma*, which are 100 and auto respectively.

In case of the 1 modality experiment (Section 4.2) we grid search to tune the sliding window (k) with values of 12, 13, 14, 15, 20, 40, and 50 and find that $k = 50$ produces the best result. For fusing 2 and 3 modalities we use the LR-based score level fusion (Section 4.3), where we tune the genuine (K_g) and impostor (K_i) Gaussian components using grid search. We search for (K_i, K_g) with the following tuples (1, 2); (2, 3); (2, 4); (2, 5); (2, 10); (2, 15); (3, 3); (3, 9); (6, 9); and (12, 18), where $(K_i, K_g) = (2, 5)$ produces the best results. Here, we take the mean of k LRs as a final match score and we calculate an EER per user. We grid search for k with values of 5, 10, 15, 20, 40, 50, 60 and observe that $k = 50$ produces the best result. The step size (n) of the sliding window is always set to 2.

4.5. One-Class Classifier (OCC) experiments

We train one-class SVMs only on genuine samples, to perform multi-modalities fusion experiments. We perform factorial and best guess methods of grid search Alpaydin (2010) to tune the OCC hyperparameters, *nu* and *gamma* over all values within $(0.0, 0.9]$ and 10th multiples of scale and auto respectively for 2 modalities experiments. For the 3 modalities experiments, between auto and scale the *gamma* hyperparameter is tuned to scale keeping *nu* as default. We use the same values of the fusion parameters (k and n) as pre-tuned during the grid search of BC (Section 4.4).

5. Experimental results

This section reports the results obtained from both BC and OCC experiments. We have both tabular and graphical representations of the reported results. The classifier hyperparameter and fusion

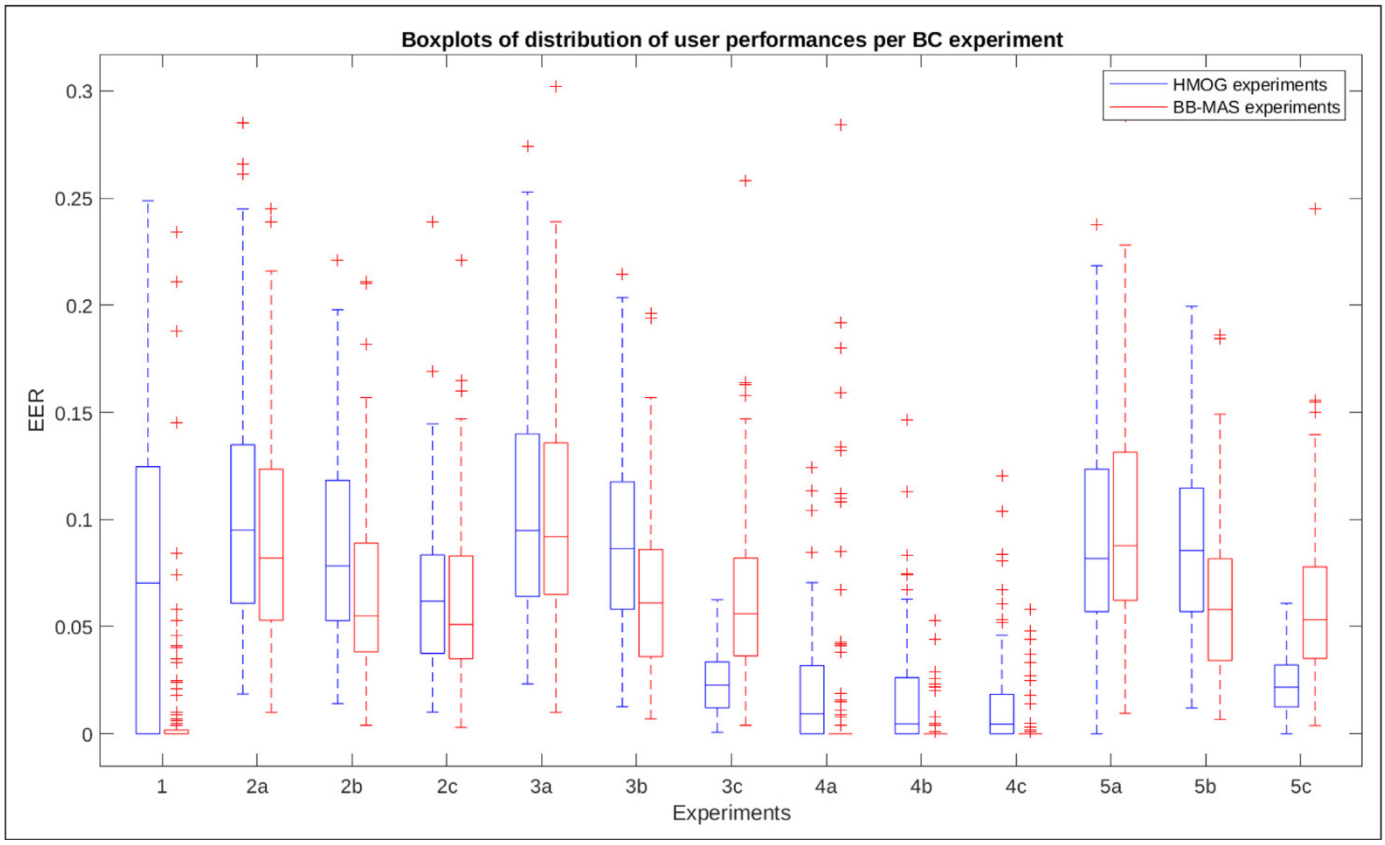

Fig. 6. Performance distribution of all users (HMOG (dataset et al., 2015) and BB-MAS (dataset et al., 2019) datasets) over binary classifier experiments. The tick labels correspond to the experiment numbers in Table 4.

parameter tuning (Section 4) through grid search sets the best possible combination of all the controlled variables (hyperparameters and parameters). However, through factorial and best guess methods of grid search (Alpaydin, 2010) we aim at obtaining the best configuration of OCC hyperparameters for the two modality fusion experiments using BB-MAS (dataset et al., 2019) data. The BC perform better than OCC in all experiments across the two datasets. Between HMOG (dataset et al., 2015) and BB-MAS (dataset et al., 2019) datasets the latter performs consistently better because of the absence of the concept drift factor and shorter swipes unlike in HMOG. Among the three feature sets, Shens statistical features (Shen et al., 2017) performs the best.

5.1. Binary classifier results

Table 4 shows results of the experiments we perform on HMOG (dataset et al., 2015) and BB-MAS (dataset et al., 2019) data. The table contains numerical values of overall performance for each experiment in terms of average, median, minimum, maximum, and standard deviation statistics across all users in a dataset. For each experiment we take the individual performance of all users in a dataset and visualize the spread and skewness trend of a distribution (formed by the average performances of each user across all the 10-fold shuffle) using box plots in Fig. 6. The design overview, score level fusion methodology, LR-based fusion method, and classifier set up are discussed in the Sections 4.1, 4.2, 4.3, and 4.4 respectively. The effect of different combinations of modalities gets evaluated through these experiments. First we evaluate the capacities of individual modalities followed by fusion of the entire motion event modalities. As swipes are sporadic, we need to investigate the authentication performance of the two motion events when they are independent of swipes. While swipes are present we evaluate fusion of all the three modalities. In the end, we need to evaluate the overall system performance in presence of continuous motion events and sporadic swipe events utilizing the hybrid experiment. Motion events produce best results when extracted with Shen's (Shen et al., 2017) statistical features.

Experiment 1: 1 modality - In our prior work (Ray et al. (2021)), we evaluate score-level fusion on a small mobile dataset where the performance of individual motion event modality was poor, with EERs of 20.5% and 18.3% for acceleration and gyroscope, respectively. The t-SNE plots (Figs. 1, 2, 3, and 4) using the public datasets in this study also show that the individual motion event modality cannot successfully authenticate users since there are a lot of overlap between the motion event samples of both genuine and impostors. We did not have swipe as a modality in our previous study. Therefore, in this study we evaluate only swipe as the single touch event-based modality to authenticate users. The experimental procedure is discussed in the Section 4.2. The swipe data statistics for both datasets are shown in Table 2. Between the two datasets, BB-MAS performs the best with an average EER of 1.3%. The standard deviation in case of the best result is 3.8% (Table 4), which shows less variation in the performance across all users. See the box plots labeled 1 in Fig. 6 where there are no outliers for the HMOG data but the spread is large. The data is skewed to the right. Whereas, for BB-MAS data under the same label 1 in Fig. 6, the spread is low, although there are outliers. Comparing the overall average EERs across all users in each dataset, BB-MAS swipes perform better than HMOG.

Experiment 2: 2 modalities in full- The entire motion event data, both outside and within swipes (Table 2) are fused. Between both datasets, the BB-MAS motion data extracted with Shen et al.s (2017) Shen et al. (2017) features performs the best, with an average EER of 6%. The standard deviation of the EERs which cor-

Table 4
Results on HMOG (dataset et al., 2015) and BB-MAS (dataset et al., 2019) datasets, using binary classifiers. Reported statistics (average, median, minimum, maximum, and standard deviation) across all users. Statistics augmentation: AVG-Average; MED-Median; MIN-Minimum; MAX-Maximum; STDV-Standard Deviation. .

Exp.	Modalities	Features	EER (%), HMOG dataset et al.(2015)	**EER (%), BB-MAS** dataset et al. (2019)
1	Swipe	Touchalytics (Frank et al., 2012)	AVG (MED, MIN, MAX, STDV)	AVG (MED, MIN, MAX, STDV)
2	Accel, Gyro, full	a) Median	7.2 (7.1, 0.0, 24.9, 6.7)	1.3 (0.0, 0.0, 23.4, 3.8)
		b) HMOG (Sitová et al., 2015)	10.4 (9.5, 1.9, 28.5, 5.8)	9.2 (8.2, 1.0, 24.5, 5.3)
		c) Shen (Shen et al., 2017)	8.7 (7.8, 1.4, 22.1, 4.5)	6.4 (5.6, 0.0, 19.8, 3.8)
3	Accel, Gyro, outside of swipe	a) Median	6.7 (6.2, 1.0, 23.9, 3.8)	6.0 (5.1, 0.0, 22.1, 3.8)
		b) HMOG (Sitová et al., 2015)	10.7 (9.5, 2.3, 27.4, 5.6)	10.1 (9.2, 1.0, 30.2, 5.2)
		c) Shen (Shen et al., 2017)	9.1 (8.6, 1.3, 21.5, 4.5)	6.5 (6.1, 0.7, 19.6, 4.0)
4	Accel, Gyro, Swipe, within swipe	a) Median, Touchalytics Frank et al. (2012)	2.4 (2.3, 0.0, 6.3, 1.4)	6.3 (5.6, 0.4, 25.8, 3.9)
		b) HMOG (Sitová et al., 2015), Touchalytics (Frank et al., 2012)	1.9 (0.0, 0.0, 12.4, 2.6)	1.5 (0.0, 0.0, 23.4, 4.4)
		c) Shen (Shen et al., 2017), Touchalytics (Frank et al., 2012)	1.7 (0.0, 0.0, 14.7, 2.6)	**0.2** (0.0, 0.0, 5.3, 1.0)
5	Accel, Gyro, Swipe, hybrid	a) Median,Touchalytics (Frank et al., 2012)	**1.5** (0.0, 0.0, 12.0, 2.3)	0.3 (0.0, 0.0, 5.8, 1.0)
		b) HMOG (Sitová et al., 2015), Touchalytics (Frank et al., 2012)	9.4 (8.6, 1.9, 23.8, 4.9)	9.8 (8.8, 1.0, 28.8, 5.0)
		c) Shen (Shen et al., 2017), Touchalytics (Frank et al., 2012)	8.7 (8.6, 1.2, 20.0, 4.3)	6.2 (5.8, 1.0, 18.6, 3.5)
			2.3 (2.2, 0.0, 6.1, 1.4)	6.0 (5.3, 0.3, 24.5, 3.7)

responds to the best performance is 3.8% which shows less performance variation across all the users. See Table 4. The BB-MAS full motion-data performs better than HMOG by only 0.7%. This is supported by the box plots labeled as 2c in Fig. 6. There is no significant difference between the two box plots where both have few outliers. However, the BB-MAS box plot (labeled 2c) in Fig. 6 shows that the data is skewed to the left which is not seen in the corresponding HMOG box plot. Therefore, this difference justifies the slightly better performance of BB-MAS full motion-data.

Experiment 3: 2 modalities outside of swipes- These experiments aim at authenticating users based on the 2 motion events outside swipes (Table 2). Between the two datasets, HMOG extracted with Shens (Shen et al., 2017) statistical features performs the best with an average EER of 2.4%. The standard deviation corresponding to the best performance as shown in Table 4 is as low as 1.4%. The HMOG box plot in Fig. 6 labeled as 3c shows the low spread of the data without outliers. On the other hand the corresponding BB-MAS box plot in Fig. 6 shows more spread of the data compared to HMOG with outliers which explains the worse performance of BB-MAS compared to HMOG.

Experiment 4: 3 modalities swipe-based- In this experiment, user authentication is performed when all the three modalities are simultaneously present. The data statistics of swipes and motion event within swipes are shown in Table 2. Between the two datasets, BB-MAS extracted with Sitová et al. (2015) HMOG features performs the best, producing an average and a standard deviation EERs of 0.2% and 1% respectively. This is the overall best performance between the two datasets when classification is performed using BC. The corresponding average EER of the HMOG data (when extracted with HMOG features) is 1.7%. The HMOG box plot in Fig. 6 labeled as 4b shows more spread than the corresponding BB-MAS. However the HMOG box plot is skewed towards left which means most users exhibit low EERs.

Experiment 5: Weighted EER - The hybrid experiment estimates the system performance on sporadic swipe events and continuous motion events. Therefore, we consider individual user's performances in experiments 3 and 4. We assign weights of 0.95 and $(1-0.95)=0.05$ to individual user's EER achieved in experiments 3 and 4 respectively. Then a weighted sum is calculated per user using the formula $0.95*EER1+0.05*EER2$. Therefore, it generates a new weighted EER per user and then the five statistics are computed across all the users to estimate the final system performance (See Table 4). As the occurrence of the 2 modalities outside swipes is more compared to the simultaneous occurrence of all the 3 modalities, a larger weight is assigned to the former. In this case, the HMOG dataset with Shens (Shen et al., 2017) features performs the best with an average EER of 2.3%. The box plots labeled as 5c in Fig. 6 show the difference in the performances between the two datasets. Given the hybrid experiment is a combination of motion events outside swipe (experiment-3) and fusion of three modalities experiments (experiment-4), the performance of HMOG data in experiment-3 enhances the performance of HMOG data in the hybrid experiments.

The above experiments are performed to authenticate users based on the availability of the modalities. We evaluate the authentication capacity of modalities from the two datasets in different combinations. Given continuous motion events and sporadic swipes the hybrid experiment represents the overall system performance where user authentication is based on the switching of available two or three modalities.

5.2. One-Class Classifier results

In this section, we discuss the experimental results of the one-class classifier. When performing the 2 modality experiments taking the two motion event data of BB-MAS dataset et al. (2019) we

observe that the performance is not comparable to BC. However, as shown in Table 5, we observe comparable results with the BC in case of three modalities.

2 modalities fusion on BB-MAS dataset et al. (2019) motion data:

For these experiments we perform LR-based fusion of the motion events occurring outside the swipe data. We fine tune the one-class SVM hyperparameters (nu and gamma) utilizing both factorial and best guess (based on t-SNE visualization) grid search methods where we alter one factor at a time and proceed further depending on the previous result. We stop at a configuration when no further improvements in the result are observed (Alpaydin, 2010).

We perform the following grid searches on the 2 modalities (outside swipes) extracted with median feature. Keeping *nu* as default (0.5), we grid search *gamma* with scale/10, scale*10, scale/100, and scale*100 values. The *nu* value of 0.5 implies at most 50% of the training samples are allowed to be misclassified or are considered as outliers by the decision boundary. At *nu*=0.5 and at *gamma* = scale*100, for both the SVMs of acceleration and gyroscope we obtain an EER of 10.3%, which is the lowest so far. As *gamma*=scale*100, produces the lowest EER, we tune *gamma* with values of scale times multiples of 100. Keeping *nu* as default in both SVMs, we grid search with *gamma* values of scale*200, scale*300, scale*400, and scale*500. None of these combinations exceed the performance of 10.3%. Thereafter, keeping *gamma*=scale*100 in both SVMs, we vary the *nu* values to 0.25 as a lower *nu* than default in one experiment and 0.75 as a higher *nu* than default in another experiment. No further improvement in the EER is observed. Considering the bad performing user in the t-SNE plots of Fig. 1, we increase the value of *nu* to 0.75 in one experiment and 0.9 in another as there are substantial overlap between genuine and impostor samples. Each of these *nu* values is combined with each of *gamma* = scale/100, scale/10, and scale*10. These runs also do not produce EER lower than 10.3%. Now in the next set of searches we change *gamma* to auto, auto/10, auto/100, auto*10, and auto*100 and combine each *gamma* with default *nu*. At *gamma*=auto*10 and default *nu* of 0.5, the lowest EER value of 10.3% returns. Therefore, *nu* values greater than 0.5 will not lead to the best possible configuration. Keeping both *gammas* at auto*10, we set nu of acceleration SVM to 0.5 and that of gyroscope SVM to 0.2 because the gyroscope t-SNE in Figs. 1 a and 1 b show higher concentration of data points towards the center than the acceleration data points. We obtain an EER of 10.4% which is close to the best performance of 10.3%. Keeping the *nu* values same as the last search we alter both *gamma* values of acceleration and gyroscope SVMs to scale*10 where we again obtain the lowest EER of 10.3%. Therefore, we stop at this configuration when *gamma* is set to scale*10 with *nu* of acceleration and gyroscope SVMs being 0.5 and 0.2 respectively producing the lowest EER of 10.3%.

Another extensive grid search based on factorial method (Alpaydin, 2010) is performed where the BB-MAS (dataset et al., 2019) motion event modalities occurring outside the swipe is extracted using HMOGs motion event feature set. We first grid search with each *gamma* values of scale, scale*10, scale/10, scale*100, and scale/100 in both SVMs where each of the *nu* of acceleration and gyroscope are tuned with all the values from the set, {0.2, 0.4, 0.6, 0.8}. From all possible combinations as above the lowest EER is obtained with *gamma*=scale*10 when *nu* of acceleration and gyroscope SVMs are 0.2 and 0.8 respectively. So we proceed to search with *gamma* values of scale*20, scale*30, scale*40. This time also we combine the *nu* of acceleration SVM with values from the set, {0.2, 0.4, 0.6, 0.8} with the *nu* of gyroscope SVM from the same set. See Fig. 7. The *gamma*=scale*40 configuration produces a better result compared to *gamma*=scale*10 so we perform a fine-grained search combining the *nu* of acceleration from the set {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9} with the *nu* of gyroscope SVM from {0.5, 0.6, 0.7,0.8, 0.9} keeping *gamma*=scale*40. See Fig. 8. We find further improvement than the previous configuration. Therefore, to our understanding, when motion event data is extracted with HMOG features, the lowest EER of 18.2% is obtained with *gamma*=scale*40, *nu* of acceleration SVM=0.5, and *nu* of gyroscope SVM=0.8.

We extract Shens feature set on BB-MAS motion event data occurring outside of swipe and perform two random runs keeping *gamma* as scale*100 and auto*10 with the *nu* values set to default where we obtain EERs of 22.8% and 17.2% respectively. As in the above best guess based grid search Alpaydin (2010) *gamma*=auto*10 produces better result, we further search keeping *gamma*=auto*10 combining the *nu* of acceleration SVM with values from the set {0.2, 0.4, 0.6, 0.8} with the *nu* of gyroscope SVM with values from the same set. At *nu* of acceleration 0.2 and *nu* of gyroscope 0.4 we obtain an EER of 12.9% which shows further improvement. As a next step we perform a fine grain search of the *nu* set where we combine *nu* of acceleration SVM from the set {0.1, 0.2, 0.3} with all the *nu* of gyroscope SVM values of {0.1, 0.2, 0.3, 0.4, 0.5}. We do not find better EER than 12.9%. We perform further runs with *gamma*=auto*40 and *nu* of acceleration SVM with values {0.1, 0.3, 0.5, 0.7} in combination with *nu* of gyroscope SVM of values {0.1, 0.3, 0.5, 0.7} which show no improvement. Thereafter, we try with *gamma* as scale and its multiples. We tune *gamma* with values of scale, scale*10, scale*20, scale*30, scale*40 and try each *gamma* values with set of *nu* of acceleration SVM of {0.1, 0.3, 0.5, 0.7, 0.9} in combination with set if *nu* if gyroscope SVM with values of {0.1, 0.3, 0.5, 0.7, 0.9}. See Fig. 9. None of these runs produce a better error rate than 12.9%. We stop at this configuration. Therefore, the best EER we obtain by extensive grid search over BB-MAS motion data extracted with Shen feature set is 12.9% when *gamma*=auto*10, and *nu* of acceleration and gyroscope SVMs are 0.2 and 0.4 respectively. See Fig. 10.

3 modalities fusion on both datasets (dataset et al., 2019; 2015):

We perform LR-based fusion of a swipe with acceleration and gyroscope events that fall within the swipe utilizing OCC. In case of the HMOG dataset (dataset et al., 2015), the best results are obtained in two experiments with the same EERs of 8.8% when the motion event data is extracted with median feature and Shen (Shen et al., 2017) feature. The standard deviation obtained are EERs of 9.7% and 9% respectively. This shows that, there has been few poor performers whose individual performances increased the overall average error rate to 8.8%. See Table 5. In case of the BB-MAS dataset (dataset et al., 2019), we obtain the best performance of 0.9% EER when the swipe is extracted with Touchalytics features and motion event is extracted with HMOG (dataset et al., 2015) features. In the best performance scenario, the standard deviation of EERs across the users is as low as 3.3% which shows that among 115 BB-MAS dataset et al. (2019) users, most of them perform well. See Table 5 for numerical values of overall performance for each experiment in terms of average, median, minimum, maximum, and standard deviation statistics across all users in a dataset. For each experiment we take the individual performance of all users in a dataset and visualize the spread and skewness trend of a distribution (formed by the average performances of each user across all the 10-fold shuffle) using box plots in Fig. 11 for both HMOG (dataset et al., 2015) and BB-MAS (dataset et al., 2019) datasets. The box plots labeled as 1a, 1b, and 1c show motion events extracted with median, HMOG Sitová et al. (2015), and Shen (Shen et al., 2017) features respectively. The larger spread in EERs per HMOG (dataset et al., 2015) experiment and minimum spread in EERs per BB-MAS (dataset et al., 2019) experiments justify BB-MAS (dataset et al., 2019) performing better than HMOG (dataset et al., 2015).

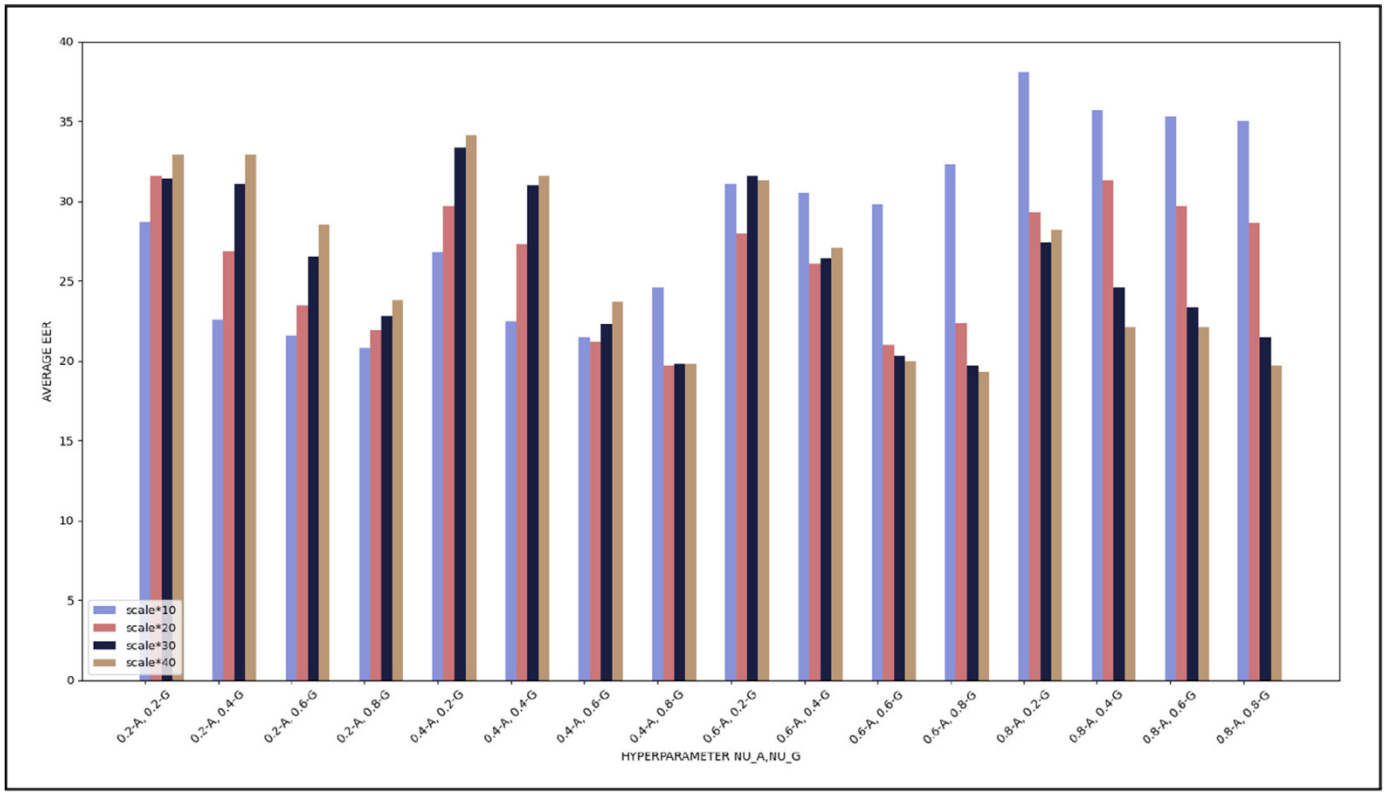

Fig. 7. Grid search on BB-MAS motion data (dataset et al., 2019), with HMOG features (Sitová et al., 2015); gamma = scale*10, scale*20, scale*30, scale*40; nu of acceleration={0.2, 0.4, 0.6, 0.8}; nu of gyroscope={0.2, 0.4, 0.6, 0.8}..

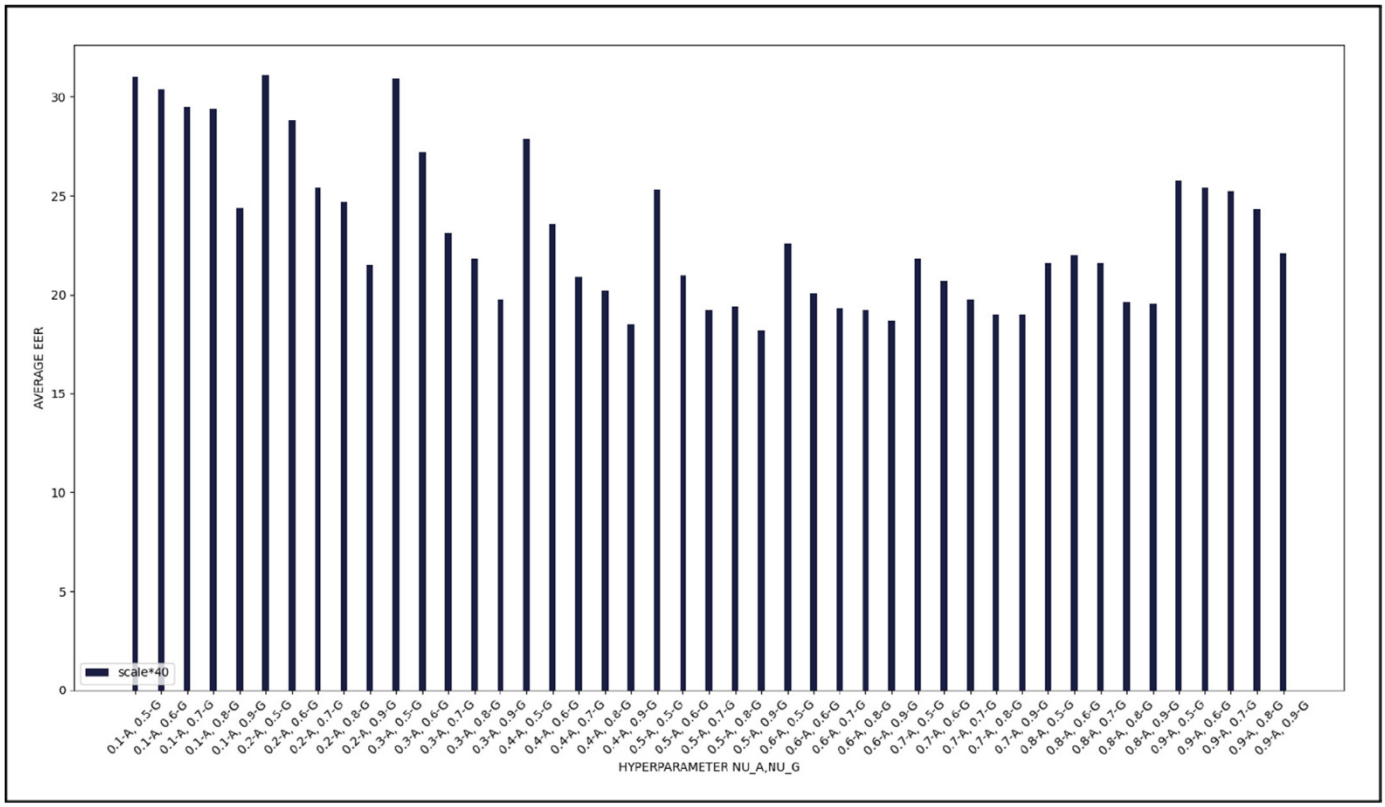

Fig. 8. Grid search on BB-MAS motion data (dataset et al., 2019), with HMOG features (Sitová et al., 2015); gamma=scale*40; nu of acceleration={0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}; nu of gyroscope={0.5, 0.6, 0.7, 0.8, 0.9}.

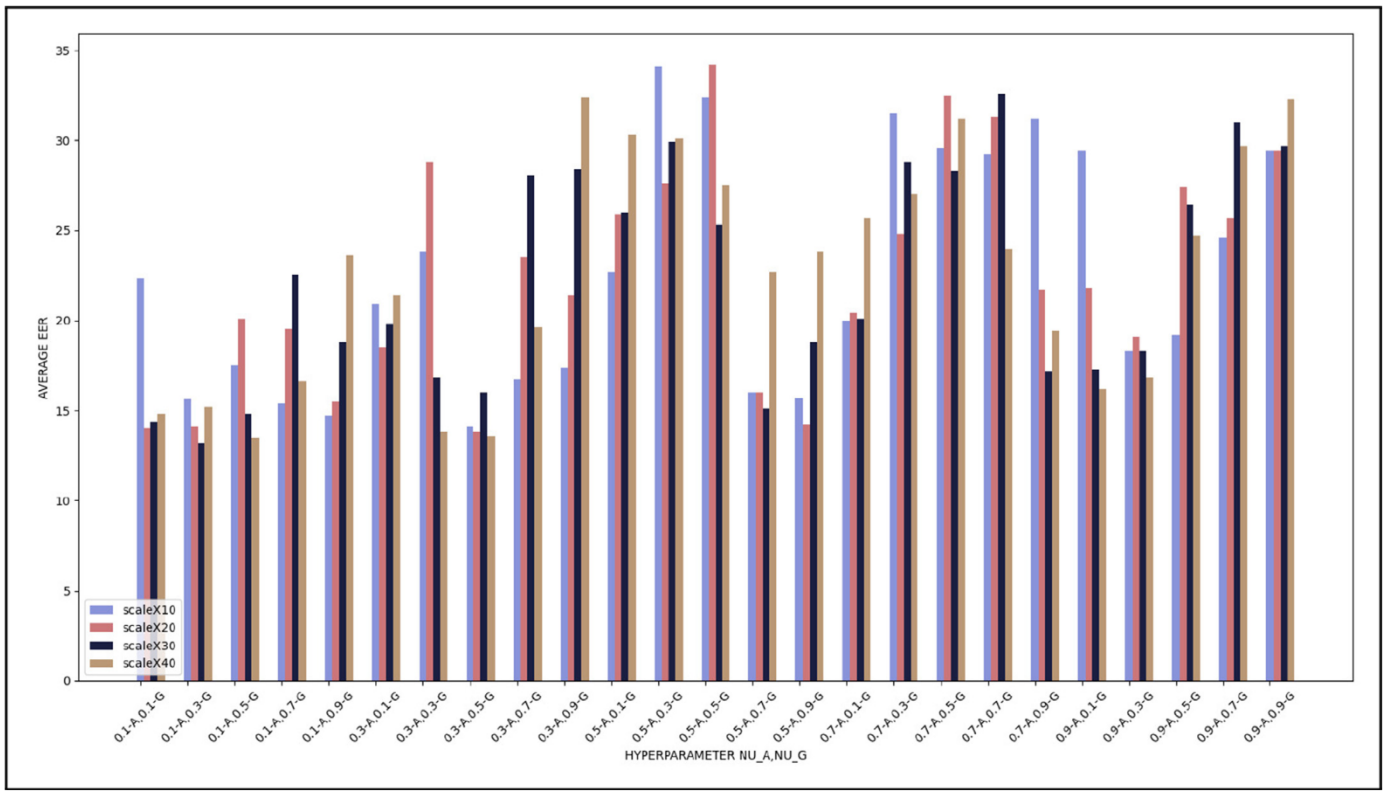

Fig. 9. Grid search on BB-MAS motion data (dataset et al., 2019), with Shen features (Shen et al., 2017); gamma = scale*10, scale*20, scale*30, scale*40; nu of acceleration={0.1, 0.3, 0.5, 0.7, 0.9}; nu of gyroscope={0.1, 0.3, 0.5, 0.7, 0.9}..

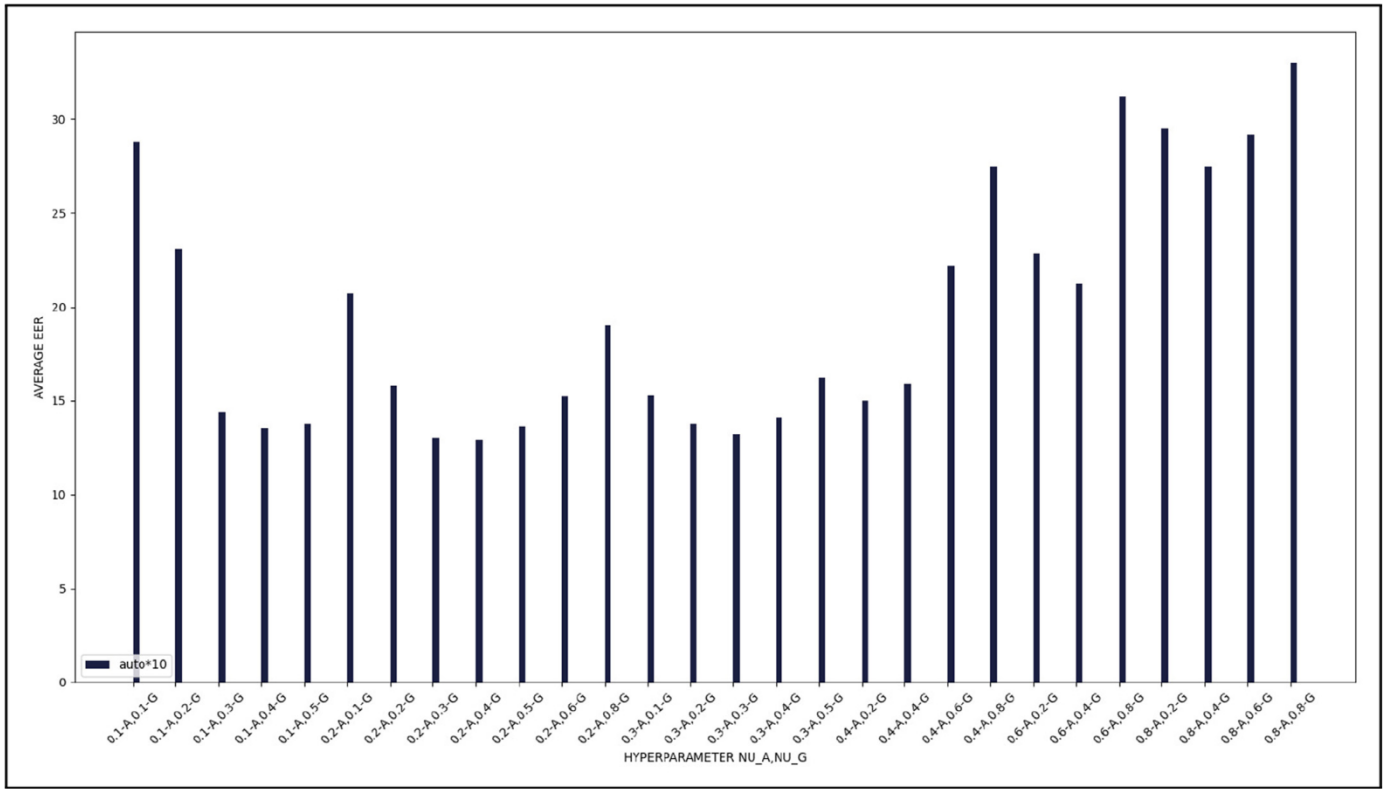

Fig. 10. Grid search on BB-MAS motion data (dataset et al., 2019), with Shen features (Shen et al., 2017); gamma = auto*10; nu of acceleration={0.1, 0.2, 0.3}; nu of gyroscope={0.1, 0.2, 0.3, 0.4, 0.5} and nu of acceleration={0.2, 0.4, 0.6, 0.8}; nu of gyroscope={0.2, 0.4, 0.6, 0.8}.

Table 5
Results on HMOG (dataset et al., 2015) and BB-MAS (dataset et al., 2019) dataset, using one-class classifiers. Reported statistics (average, median, minimum, maximum, and standard deviation) across all users. Statistics augmentation: AVG-Average; MED-Median; MIN-Minimum; MAX-Maximum; STDV-Standard Deviation. .

			EER (%), HMOG (dataset et al., 2015)	EER (%), BB-MAS (dataset et al., 2019)
Exp.	**Modalities**	**Features**	AVG (MED, MIN, MAX, STDV)	AVG (MED, MIN, MAX, STDV)
1	Accel, Gyro, Swipe (within swipe)	a) Median, Touchalytics (Frank et al., 2012)	**8.8** (6.6, 0.0, 37. 6, 9.7)	1.4 (0.0, 0.0, 22. 3, 4.0)
		b) HMOG (Sitová et al., 2015), Touchalytics Frank et al. (2012)	13.6 (13.5, 0.0, 37.6, 12.3)	**0.9** (0.0, 0.0, 20.9, 3.3)
		c) Shen (Shen et al., 2017), Touchalytics Frank et al. (2012)	**8.8** (6.1, 0.0, 32.6, 9.0)	1.0 (0.0, 0.0, 18.7, 3.0)

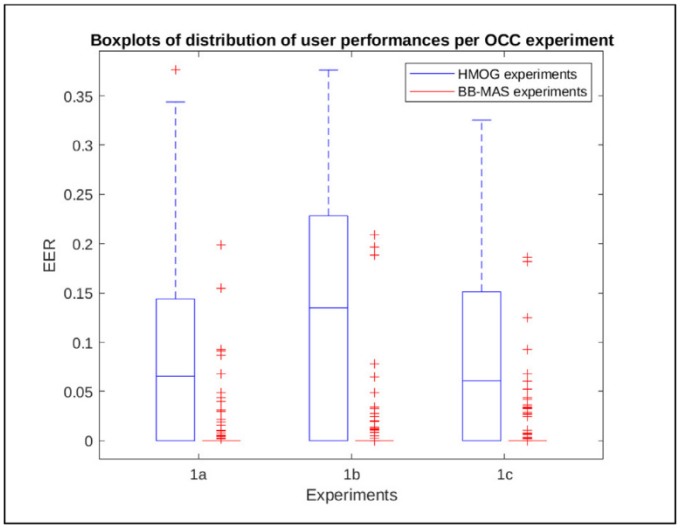

Fig. 11. Performance distribution of all users (HMOG (dataset et al., 2015) and BB-MAS (dataset et al., 2019) datasets) over OCC experiments. The tick labels correspond to the experiment numbers in Table 5.

5.3. Overall trend and justification

In most BC and OCC experiments BB-MAS (dataset et al., 2019) data performs consistently better than HMOG (dataset et al., 2015) data. The BC experiments are performed uniformly taking both datasets. However, while performing 2 modalities OCC experiments with BB-MAS dataset (dataset et al., 2019), we run extensive grid searches to fine-tune the OCC hyperparameters (*nu* and *gamma*). Given this and the t-SNE plots in Figs. 3 and 4 that show overlap between genuine and impostor samples, we did not perform 2 modalities OCC experiments with HMOG data (dataset et al., 2015). The box plots of BB-MAS in Fig. 6 (labeled 1, 4a, 4b, and 4c) and Fig. 11 (labeled 1a, 1b, and 1c) show minimal spread compared to HMOG since several BB-MAS users have achieved 0% or close EERs. Therefore, we examine the difference in the data collection methods of HMOG (dataset et al., 2015) and BB-MAS (dataset et al., 2019). In case of BB-MAS (dataset et al., 2019), the researchers made sure that users swipe in between answering questions so that swipes of considerable lengths and trajectories can be logged. Whereas in HMOG (dataset et al., 2015), when users are sitting and writing, no swipe-specific actions are performed. Users have the liberty to swipe if needed while typing. Therefore, we notice that the length of the swipes in HMOG [1] is shorter than BB-MAS (dataset et al., 2019) for which the unique swipe trajectory required to identify users is more prominent in case of BB-MAS [2] compared to HMOG (dataset et al., 2015).

Figs. 13 and 14 show 20 sample swipes of a random user from each of HMOG (dataset et al., 2015) and BB-MAS (dataset et al., 2019) datasets respectively. From the general visualizations one can see that the HMOG (dataset et al., 2015) sample swipes (Fig. 13) are shorter compared to the BB-MAS (dataset et al., 2019) sample swipes (Fig. 14). The x and y axes in both figures have the same ranges for fair comparison. To further study the nature of swipe length in both datasets we calculate all the swipe lengths per user (from the raw swipe data before processing) and plot the distributions of swipe lengths. The total swipe length is calculated by summing the Euclidean distances between every pair of coordinates of intermediate touch events which constitute the swipe. See Fig. 12 where the box plots show the skewness and spread of swipe lengths across all users in both datasets. The numerical values of the swipe length statistics (average, median, 25 percentile, minimum, maximum, and standard deviation) are shown

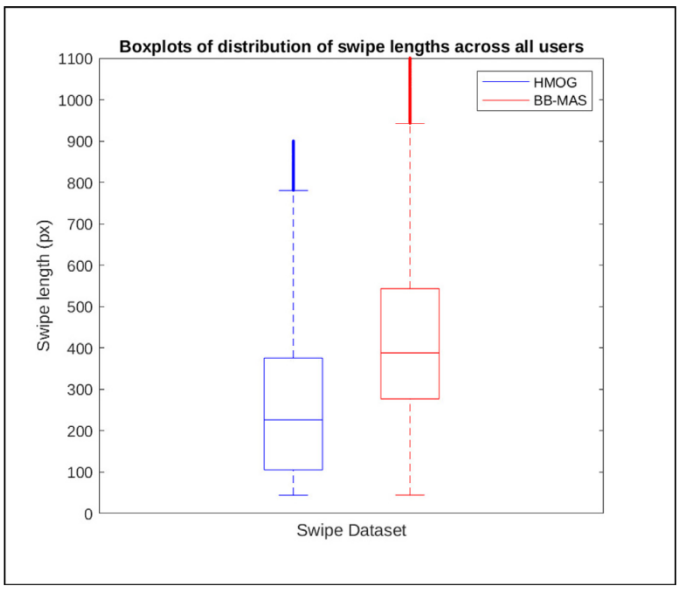

Fig. 12. Swipe length distribution of all users (HMOG (dataset et al., 2015) and BB-MAS (dataset et al., 2019) datasets).

in Table 6. The statistics in support of the box plots show that the maximum length of BB-MAS (dataset et al., 2019) swipe is larger than HMOG (dataset et al., 2015) swipe. There are short length swipes in HMOG data (dataset et al., 2015) for which the trajectories do not add sufficient uniqueness to authenticate users as it does in case of BB-MAS (dataset et al., 2019). Same trend is shown in case of other statistics like the average, median, and the 25 percentile. Therefore, such difference in the swipe length trend is a reason why in all swipe-based experiments, BB-MAS (dataset et al., 2019) outperforms HMOG (dataset et al., 2015).

In the case of HMOG, the data collection for the task of sitting and typing on mobile devices has been done across four sessions (3, 9, 15, and 21). As shown in Table 7, there appear to be variable time gaps between the sessions.

Since we have combined all four sessions in HMOG to make a user profile, the presence of concept drift would impact performance. Furthermore based on the observed time gaps between sessions, we speculate there is concept drift in HMOG. To quantify this, we perform statistical hypothesis tests between two sessions (session-3 and session-21). We first perform the Kolmogorov-Smirnov tests on acceleration along x, y, and z axes respectively, and gyroscope along x, y, and z axes respectively. Since we find that none of the data is normally distributed, we perform Mann-Whitney tests to see if a users individual axis data (x/y/z) from session-3 and their corresponding individual axis data from

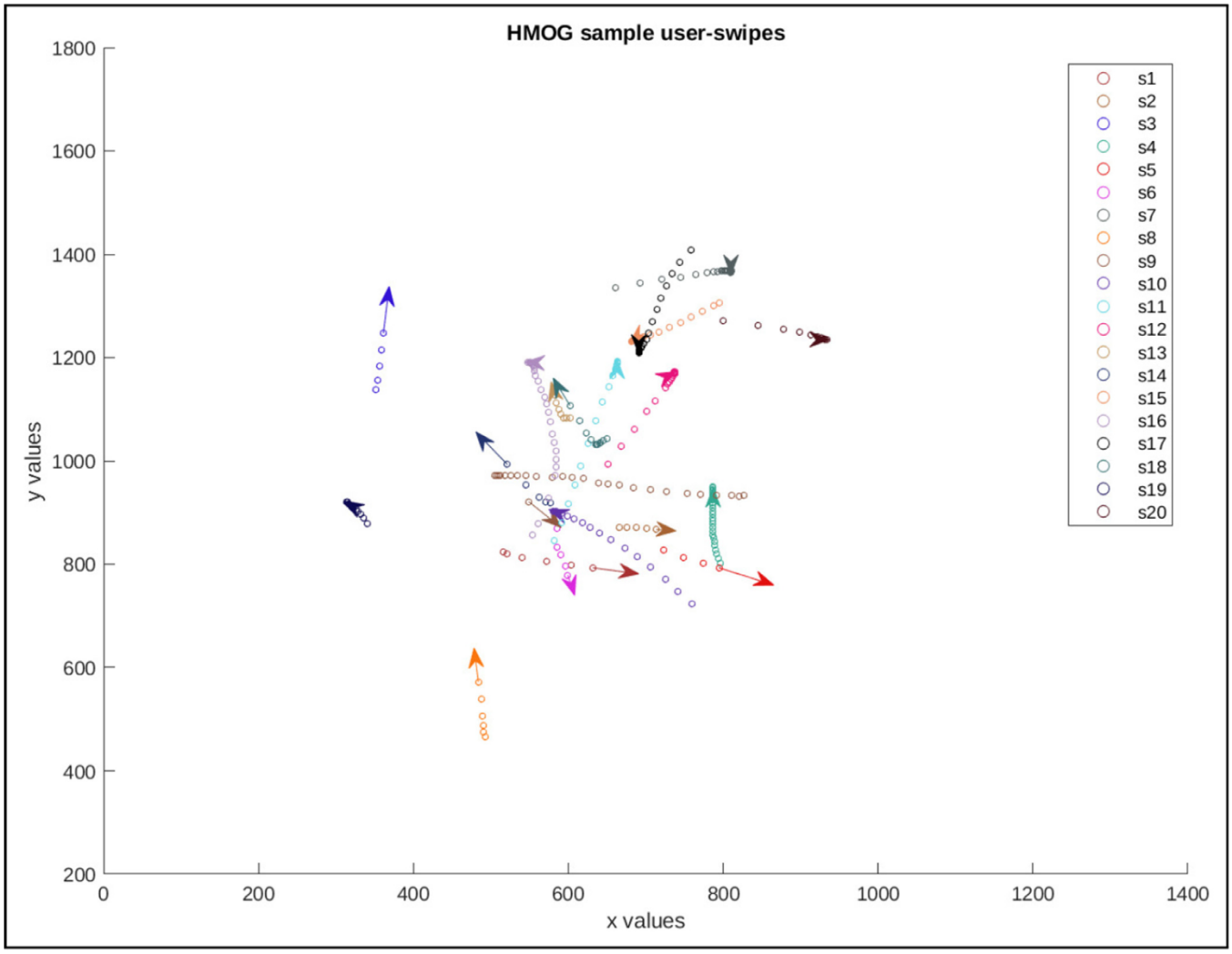

Fig. 13. 20 sample swipes of one user from HMOG dataset (dataset et al., 2015). Range of x-axis:[0,1400]. Range of y-axis:[200,1000].

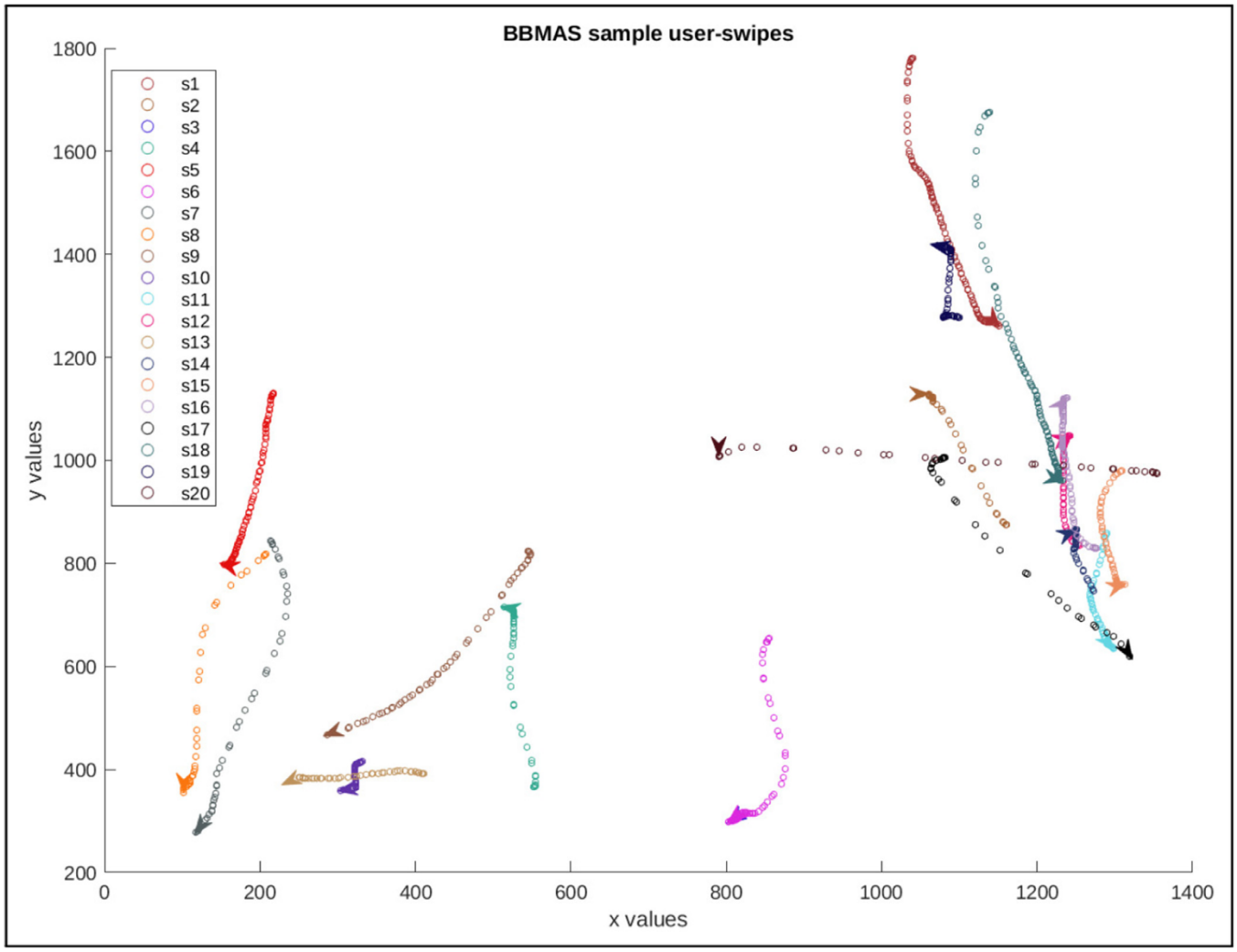

Fig. 14. 20 sample swipes of one user from BB-MAS dataset (dataset et al., 2019). Range of x-axis:[0,1400]. Range of y-axis:[200,1000].

Table 6
Statistics of swipe lengths across users in the two public data sets: HMOG dataset et al. (2015) and BB-MAS dataset et al. (2019) before processing for experiments. Statistics augmentation: AVG-Average, MED-Median, 25 PERC-25th Percentile, MIN-Minimum, MAX-Maximum, STDV-Standard Deviation. .

	Swipe-length statistics (pixel)					
Dataset	AVG	MED	25 PERC	MIN	MAX	STDV
HMOG (dataset et al., 2015)	264.53	226.43	105.56	44.11	899.71	183.19
BB-MAS (dataset et al., 2019)	436.92	388.01	276.93	44.49	3950.3	238.61

Table 7
Time-gap between sessions for three randomly picked HMOG (dataset et al., 2015) users.

Between sessions	User-100669	User-171538	user-186676
3 and 9	≈ 1 day	≈ 2 days	≈ 1 day
9 and 15	≈ 1 day	≈ 2 days	≈ 1 day
15 and 21	≈ 2 days	same day	≈2 days

session-21 belong to the same distribution. The p value is always less than 0.05 value. Therefore we conclude that no pairs belong to the same distribution, and thus these statistical tests confirm that there is the presence of concept drift in the data across sessions in HMOG. In contrast, data from BB-MAS does not have this issue as all users provide data in one attempt in a single day.

The acceleration and gyroscope data visualizations for a randomly selected user 100,669 of HMOG are shown in Figs. 15 and 16, which provide further evidence of the existence of concept drift.

We have pointed out several similarities between the two public datasets, such as the user behavior of sitting and typing is the same, the choice of the same modalities, the number of users being very close to each other, and the feature sets we applied on the datasets are the same. Above all, the user authentication algorithm that we apply has been held to be the same on both datasets. Therefore, we compare the performances of the two datasets under similar scenarios and observe BB-MAS to be consistently better than HMOG. We further explore the two possible causes for the performance difference between the two datasets. We hope others will find this useful to know.

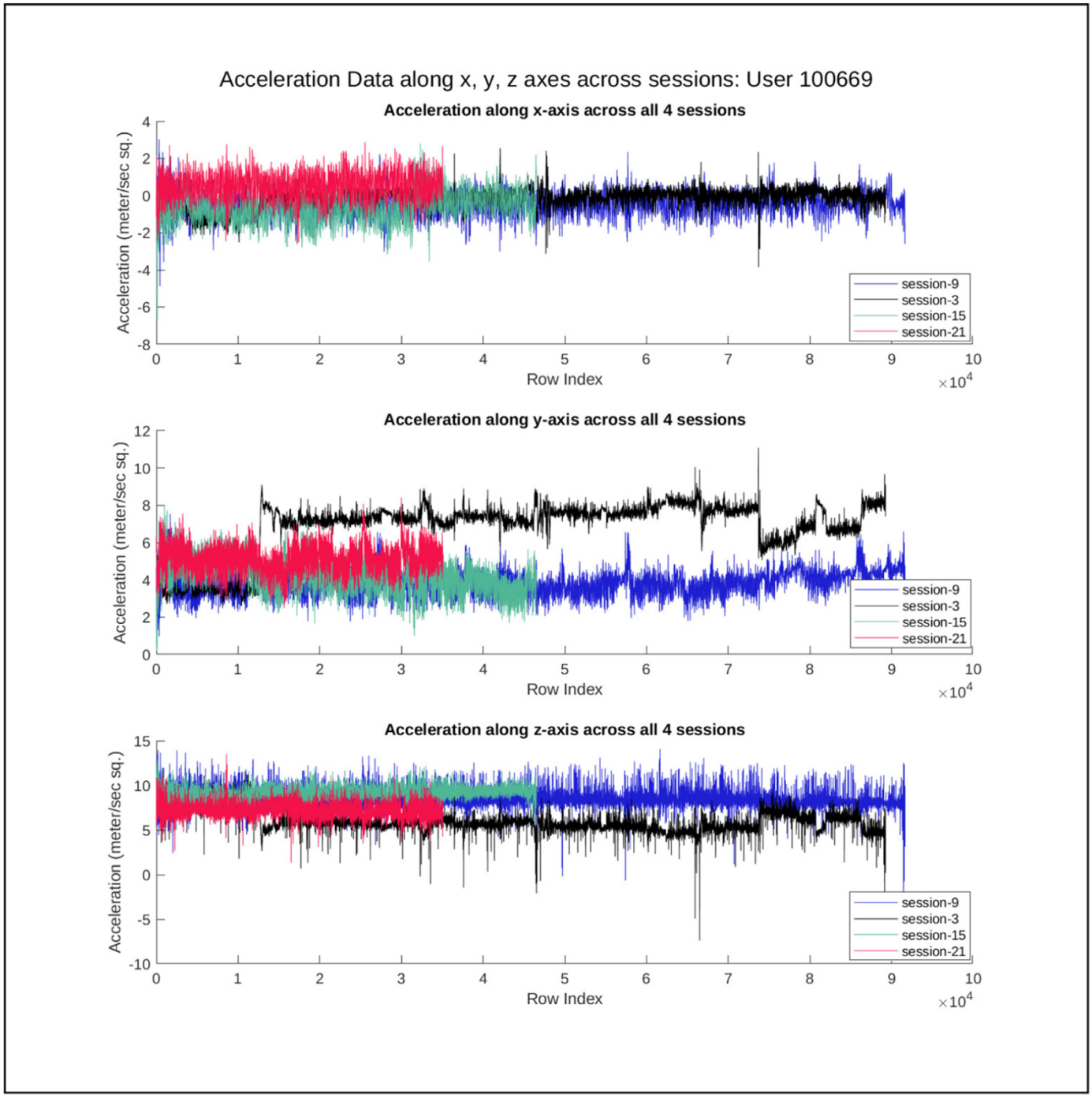

Fig. 15. Acceleration along x, y, and z axes for user 100,669 across all four sessions demonstrates the existence of concept drift in HMOG.

5.4. Time per decision during classification

We fuse 50 scores to make a decision in every experiment. In case of the swipe-based experiments with the HMOG dataset (dataset et al., 2015), it takes 2.5 minutes to gather 50 swipes to make a decision. On the other hand, the two motion modalities-based experiments take 25 seconds per decision. For any experiment with the BB-MAS (dataset et al., 2019) dataset that involve swipes, the time taken to gather 50 swipes is about 8 minutes. On the other hand, classification based on the two motion modalities take 25 seconds per decision.

5.5. Computational cost

The computer used to run all experiments is a HP Z620 workstation with 94 GB RAM and Intel Xeon E5-2670 processor with hyperthreading enabled. It runs the Kubuntu 18.04.6 LTS operating system.

The computation time of pre-processing (cleaning and feature extraction) and training (for both SVM and GMM) for binary classifiers ranges from 1.9 to 38.5 second and 0.004 to 28.13 second respectively in case of HMOG, and from 0.9 to 61 second and 0.004 to 21.39 second respectively in case of BB-MAS. For one-class classifiers, the computation time of pre-processing and training ranges

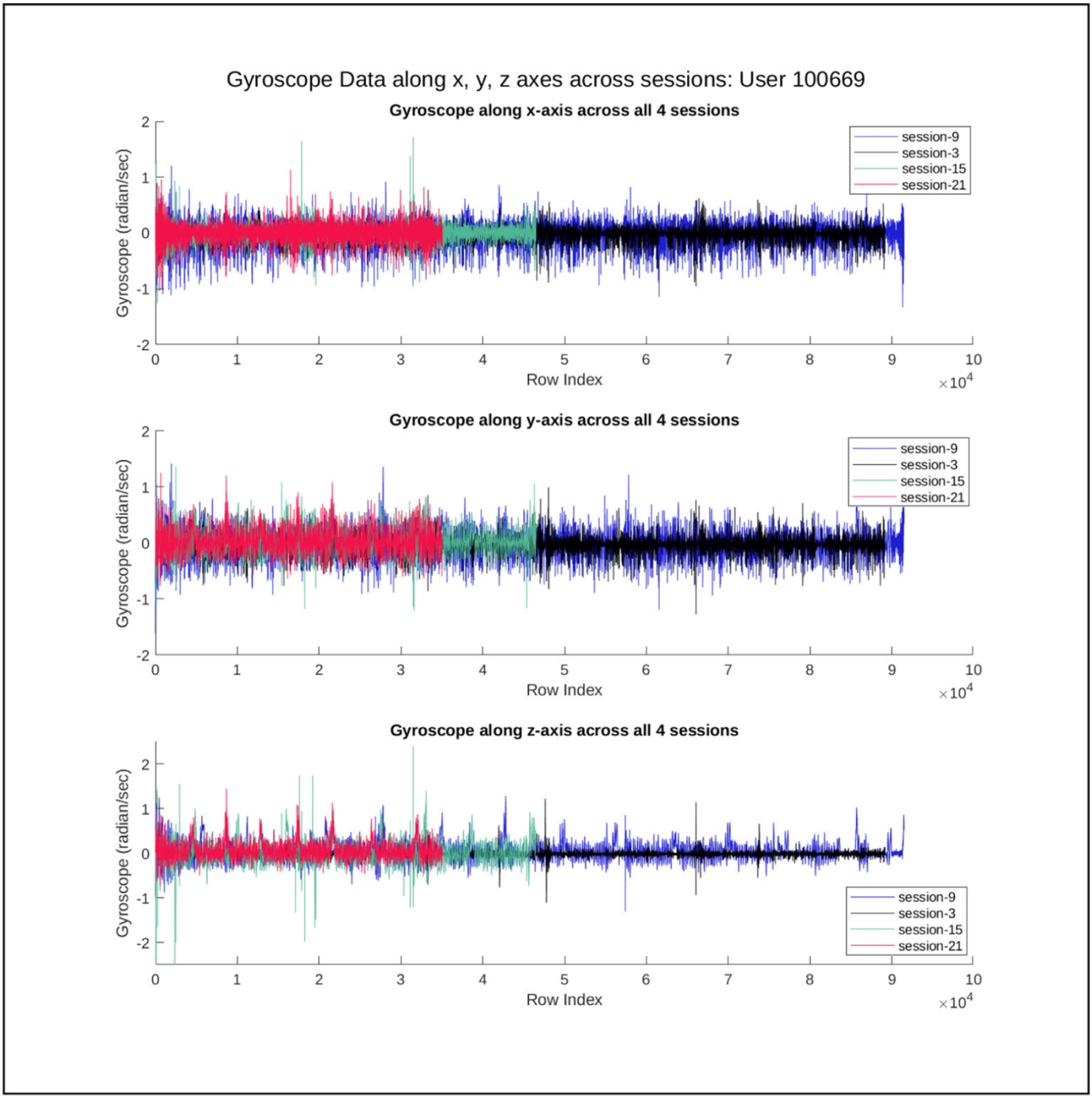

Fig. 16. Gyroscope along x, y, and z axes for user 100,669 across all four sessions demonstrates the existence of concept drift in HMOG.

from 10.7 to 38.5 second and 0.004 to 0.005 second respectively for HMOG, and from 58.8 to 61 second and 0.005 to 0.009 second respectively for BB-MAS. The three modality OCC experiment has the same processing time as that of three modality BC experiments (under group 4) since the data used are the same. However, given only genuine samples are trained on the models (SVM and GMM) and the data samples for three modality swipe-based experiments are fewer than those of the other group of experiments, the training time of OCC is shorter than BC training time.

6. Conclusions and future works

Our previous work (Ray et al. (2021)) involve acceleration and gyroscope to perform continuous authentication on our own collected dataset of 49 seated users. The likelihood ratio-based score fusion performs better than weighted score fusion in both intra-session and inter-session experiments. Therefore, the present work evaluates the performance of our authentication system on two large public datasets to establish the generalizability of behavioral

biometric-based authentication. We extract the motion event data over three different feature sets and observe the difference in performances across feature sets. We evaluate performances of multi-modalities fusion experiments using likelihood ratio taking both one-class and binary SVMs.

We evaluate our authentication platform by utilizing acceleration, gyroscope, and swipe modalities from the two large public datasets HMOG (dataset et al., 2015) and BB-MAS (dataset et al., 2019). We extract three different feature sets (median, HMOG (Sitová et al., 2015), and Shen (Shen et al., 2017)) from the motion event data as well as Touchalytics features (Frank et al., 2012) from the swipe data. In the case of the swipe-based single modality experiment we have used score level fusion that shows the authentication capacity of the individual sporadic touch event. Depending on the availability of the swipes, when there are swipes present we perform three modality fusion and when there is no swipe we fuse two motion modalities. The results of the experiments match with our hypothesis of enhanced performance when swipe is fused with the motion events. We train both binary and one-class SVM classifiers per modality. For fusing two or more modalities, we apply Nandakumars (Nandakumar et al., 2007) LR-based score fusion.

Binary SVMs achieve the best EERs of 1.5% (HMOG) and 0.2% (BB-MAS) when all three modalities are fused, whereas the one-class SVMs produce EERs of 8.8% and 0.9% respectively for the same experiment. Across most experiments, BB-MAS performs better than HMOG due to the absence of concept drift factor in the data collection process and for longer swipe lengths compared to HMOG. The binary SVMs achieve the best EERs of 2.4% (HMOG) and 6.3% (BB-MAS) when two motion modalities outside the swipe are fused, whereas in the same experiment with BB-MAS data the one-class SVMs show poor performance. Although we have shown that binary is better than one-class SVM, our goal is not to only demonstrate that. Rather, we want to evaluate both binary and one-class classifiers using both datasets. Although it is common to use binary classifiers in authentication, a one-class classifier becomes necessary when impostor data is not available.

In this work, we have evaluated multi-modal behavioral biometrics using two datasets. In the future, we plan to evaluate our user authentication model on other public datasets on mobile behavioral biometrics.

We also plan to further evaluate the authentication performance by fusing additional modalities. In HMOG (dataset et al., 2015) and BB-MAS (dataset et al., 2019) public datasets there are additional modalities, namely, tap, double-tap, pinch, long press (which are similar touch events like swipe), rotation, and magnetometer (which are similar sensor events like acceleration and gyroscope). In the future, we will extend our authentication model to fuse these modalities.

In this work, we did not evaluate the performance of deep learning algorithms. Conventional machine learning algorithms, like SVM, can be more cost-effective in terms of computation and space compared to deep learning algorithms. On the other hand, the existing state of the art (Volaka et al. (2019), Amini et al. (2018), Buriro et al. (2021), Centeno et al. (2017), Neverova et al. (2016), Deb et al. (2019), and Abuhamad et al. (2020)) evaluate deep learning-based authentication models. In fact most of these studies implement recurrent neural networks (LSTM-Long Short Term Memory) as their deep learning models. However, most of the studies utilize a private dataset. Among these studies, Volaka et al. (2019) and Centeno et al. (2017) are exceptions as they utilize HMOG (dataset et al., 2015), but their best performances using deep learning models did not outperform the results obtained in our study. In the future, we plan to extend our work to evaluate deep learning models under a common set of datasets.

Author Credits

Aratrika Ray-Dowling (co-author) Implemented the study. Daqing Hou (corresponding author) Conceptualized and designed the study. Stephanie Schuckers (co-author) Contributed to the design of the study. Abbie Barbir (co-author) Contributed to the design of the study.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Daqing Hou reports financial support was provided by National Science Foundation.

Acknowledgments

This work was supported by the Center for Identification Technology Research (CITeR) and the National Science Foundation [grant number 1650503].

References

Abuhamad, M., Abuhmed, T., Mohaisen, D., Nyang, D., 2020. Autosen: deep-learning-based implicit continuous authentication using smartphone sensors. IEEE Internet Things J. 7 (6), 5008–5020. doi:10.1109/JIOT.2020.2975779.

Alpaydin, E., 2010. Introduction to machine learning. In: Design and Analysis of Machin Learning Experiments. MIT press, pp. 475–515.

Amini, S., Noroozi, V., Pande, A., Gupte, S., Yu, P.S., Kanich, C., 2018. Deepauth: A framework for continuous user re-authentication in mobile apps. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 2027–2035. doi:10.1145/3269206.3272034.

Belman, A.K., Wang, L., Iyengar, S.S., Sniatala, P., Wright, R., Dora, R., Baldwin, J., Jin, Z., Phoha, V.V., 2019. Insights from BB-MAS–a large dataset for typing, gait and swipes of the same person on desktop, tablet and phone. arXiv:191202736.

Buriro, A., Crispo, B., Del Frari, F., Klardie, J., Wrona, K., 2015. Itsme: Multi-modal and unobtrusive behavioural user authentication for smartphones. In: International Conference on Passwords. Springer, pp. 45–61. doi:10.1007/978-3-319-29938-9_4.

Buriro, A., Gupta, S., Yautsiukhin, A., Crispo, B., 2021. Risk-driven behavioral biometric-based one-shot-cum-continuous user authentication scheme. J Signal Process Syst 93 (9), 989–1006. doi:10.1007/s11265-021-01654-2.

Cai, L., Chen, H., 2012. On the practicality of motion based keystroke inference attack. In: International Conference on Trust and Trustworthy Computing. Springer, pp. 273–290. doi:10.1007/978-3-642-30921-2_16.

Carlson, C., Chen, T., Cruz, J., Maghsoudi, J., Zhao, H., Monaco, J.V., 2015. User authentication with android accelerometer and gyroscope sensors. Proceedings of Student-Faculty Research Day, CSIS, Pace University.

Centeno, M.P., van Moorsel, A., Castruccio, S., 2017. Smartphone continuous authentication using deep learning autoencoders. In: 2017 15th Annual Conference on Privacy, Security and Trust (PST). IEEE, pp. 147–1478. doi:10.1109/PST.2017.00026.

Crawford, H., Ahmadzadeh, E., 2017. Authentication on the go: Assessing the effect of movement on mobile device keystroke dynamics. In: Thirteenth Symposium on Usable Privacy and Security ({SOUPS} 2017), pp. 163–173.

dataset, Belman, A.K., Wang, L., Iyengar, S.S., Sniatala, P., Wright, R., Dora, R., Baldwin, J., Jin, Z., Phoha, V.V., 2019. BB-MAS. IEEE DataPort. 1021227/rpaz-0h66

dataset, Sitová, Z., Šeděnka, J., Yang, Q., Peng, G., Zhou, G., Gasti, P., Balagani, K.S., 2015. HMOG. College of William and Mary. https://wwwcswmedu/~qyang/hmoghtml

Deb, D., Ross, A., Jain, A.K., Prakah-Asante, K., Prasad, K.V., 2019. Actions speak louder than (pass) words: Passive authentication of smartphone users via deep temporal features. In: International Conference on Biometrics. IEEE, pp. 1–8. doi:10.1109/ICB45273.2019.8987433.

Frank, M., Biedert, R., Ma, E., Martinovic, I., Song, D., 2012. Touchalytics: on the applicability of touchscreen input as a behavioral biometric for continuous authentication. IEEE Trans. Inf. Forensics Secur. 8 (1), 136–148. doi:10.1109/TIFS.2012.2225048.

Gascon, H., Uellenbeck, S., Wolf, C., Rieck, K., 2014. Continuous authentication on mobile devices by analysis of typing motion behavior. Sicherheit 2014–Sicherheit, Schutz und Zuverlässigkeit 1–12.

Ehatisham-ul Haq, M., Azam, M.A., Naeem, U., Amin, Y., Loo, J., 2018. Continuous authentication of smartphone users based on activity pattern recognition using passive mobile sensing. Journal of Network and Computer Applications 109, 24–35. doi:10.1016/j.jnca.2018.02.020.

Incel, O.D., Günay, S., Akan, Y., Barlas, Y., Basar, O.E., Alptekin, G.I., Isbilen, M., 2021. Dakota: sensor and touch screen-based continuous authentication on a mobile banking application. IEEE Access 9, 38943–38960. doi:10.1109/ACCESS.2021.3063424.

Karakaya, N., Alptekin, G.I., Özlem Durmaz, I., 2019. Using behavioral biometric sensors of mobile phones for user authentication. Procedia Comput Sci 159, 475–484. doi:10.1016/j.procs.2019.09.202.
Kim, J., Kang, P., 2020. Freely typed keystroke dynamics-based user authentication for mobile devices based on heterogeneous features. Pattern Recognit 108, 107556. doi:10.1016/j.patcog.2020.107556.
Kumar, R., Kundu, P.P., Phoha, V.V., 2018. Continuous authentication using one-class classifiers and their fusion. In: IEEE 4th International Conference on Identity, Security, and Behavior Analysis, pp. 1–8. doi:10.1109/ISBA.2018.8311467.
Kumar, R., Phoha, V.V., Jain, A., 2015. Treadmill attack on gait-based authentication systems. In: 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, pp. 1–7. doi:10.1109/BTAS.2015.7358801.
Kumar, R., Phoha, V.V., Raina, R., 2016a. Authenticating users through their arm movement patterns. arXiv preprint arXiv:160302211.
Kumar, R., Phoha, V.V., Serwadda, A., 2016b. Continuous authentication of smartphone users by fusing typing, swiping, and phone movement patterns. In: IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, pp. 1–8. doi:10.1109/BTAS.2016.7791164.
Li, Y., Hu, H., Zhou, G., 2018. Using data augmentation in continuous authentication on smartphones. IEEE Internet Things J. 6 (1), 628–640. doi:10.1109/JIOT.2018.2851185.
Li, Y., Hu, H., Zhu, Z., Zhou, G., 2020. Scanet: sensor-based continuous authentication with two-stream convolutional neural networks. ACM Transactions on Sensor Networks (TOSN) 16 (3), 1–27. doi:10.1145/3397179.
Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. Journal of machine learning research 9 (11).
Nandakumar, K., Chen, Y., Dass, S.C., Jain, A., 2007. Likelihood ratio-based biometric score fusion. IEEE Trans Pattern Anal Mach Intell 30 (2), 342–347. doi:10.1109/TPAMI.2007.70796.
Neverova, N., Wolf, C., Lacey, G., Fridman, L., Chandra, D., Barbello, B., Taylor, G., 2016. Learning human identity from motion patterns. IEEE Access 4, 1810–1820. doi:10.1109/ACCESS.2016.2557846.
Papamichail, M.D., Chatzidimitriou, K.C., Karanikiotis, T., Oikonomou, N.C.I., Symeonidis, A.L., Saripalle, S.K., 2019. Brainrun: a behavioral biometrics dataset towards continuous implicit authentication. Data 4 (2), 60. doi:10.3390/data4020060.
Ray, A., Hou, D., Schuckers, S., Barbir, A., 2021. Continuous authentication based on hand micro-movement during smartphone form filling by seated human subjects. In: ICISSP, pp. 424–431. doi:10.5220/0010225804240431.
Roy, A., Halevi, T., Memon, N., 2015. An HMM-based multi-sensor approach for continuous mobile authentication. In: MILCOM IEEE Military Communications Conference. IEEE, pp. 1311–1316. doi:10.1109/MILCOM.2015.7357626.
Shen, C., Li, Y., Chen, Y., Guan, X., Maxion, R.A., 2017. Performance analysis of multi-motion sensor behavior for active smartphone authentication. IEEE Trans. Inf. Forensics Secur. 13 (1), 48–62. doi:10.1109/TIFS.2017.2737969.
Sitová, Z., Šeděnka, J., Yang, Q., Peng, G., Zhou, G., Gasti, P., Balagani, K.S., 2015. Hmog: new behavioral biometric features for continuous authentication of smartphone users. IEEE Trans. Inf. Forensics Secur. 11 (5), 877–892. doi:10.1109/TIFS.2015.2506542.
Thang, H.M., Viet, V.Q., Thuc, N.D., Choi, D., 2012. Gait identification using accelerometer on mobile phone. In: 2012 International Conference on Control, Automation and Information Sciences (ICCAIS). IEEE, pp. 344–348. doi:10.1109/ICCAIS.2012.6466615.
Volaka, H.C., Alptekin, G., Basar, O.E., Isbilen, M., Incel, O.D., 2019. Towards continuous authentication on mobile phones using deep learning models. Procedia Comput Sci 155, 177–184. doi:10.1016/j.procs.2019.08.027.
Xu, H., Zhou, Y., Lyu, M.R., 2014. Towards continuous and passive authentication via touch biometrics: An experimental study on smartphones. In: 10th Symposium On Usable Privacy and Security ({SOUPS} 2014), pp. 187–198.

Aratrika Ray-Dowling (co-author) received Bachelor of Engineering Degree in Computer Science and Engineering from Visvesvaraya Technological University, Karnataka, India in 2014. She received Masters of Science Degree in Electrical Engineering from Clarkson University, USA in 2019. She is currently pursuing Ph.D. degree in the department of Electrical and Computer Engineering at Clarkson University, USA. Her research interest includes Mobile/Computer Security, Biometrics, Machine Learning, and Data Analytics.

Daqing Hou (corresponding author) received Bachelor of Science Degree in Computer Science from Peking University, Beijing, China in 1992. He received Masters of Science Degree in Computer Science from Peking University, Beijing, China in 1995. He received Ph.D. in Computer Science from University of Alberta, Edmonton, Canada in 2004. He has completed Post-Doctoral research at Avra Software Lab. Inc., Edmonton, Canada from 2004 to 2006. He is currently Professor and Director of Software Engineering at Clarkson University, USA. His research interests include software engineering, behaviroal biometrics, high-performance computing, and education.

Stephanie Schuckers (co-author) is the Paynter-Krigman Endowed Professor in Engineering Science in the Department of Electrical and Computer Engineering at Clarkson University and serves as the Director of the Center of Identification Technology Research (CITeR), a National Science Foundation Industry/University Cooperative Research Center. She received her doctoral degree in Electrical Engineering from The University of Michigan. Professor Schuckers research focuses on processing and interpreting signals which arise from the human body. Her work is funded from various sources, including National Science Foundation, Department of Homeland Security, and private industry, among others.

Abbie Barbir (co-author) serves as a Senior Security Advisor in the areas of identity management, mobile devices and authentication at Aetna Global Information Security (a CVS Health Company). He has worked with many organizations on developing next generation authentication technologies. He represents Aetna on the FIDO Board of Directors. Barbir holds a Ph.D. in computer engineering from Louisiana State University, USA. He has been Professor of Computer Science, Application Developer, Data Compression and Encryption Inventor, Systems Architect, Security Architect, Engineering Manager, Consultant, Author and Inventor of numerous security algorithms and articles.