Computers & Security 137 (2024) 103594

Contents lists available at ScienceDirect

Computers & Security

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

SHRIMPS: A framework for evaluating multi-user, multi-modal implicit authentication systems

Jiayi Chen [a,1], Urs Hengartner [a,*], Hassan Khan [b]

[a] University of Waterloo, 200 University Ave W., Waterloo, N2L 3G1, ON, Canada
[b] University of Guelph, 50 Stone Road East, Guelph, N1G 2W1, ON, Canada

ARTICLE INFO

Keywords:
User authentication
Evaluation framework
Implicit authentication
Biometrics
Score fusion

ABSTRACT

Smart devices are commonly used in multi-user scenarios, such as shared household devices and shared corporate devices for front-line workers. A multi-user device requires both identification and authentication to defend against unauthorized access and distinguish between legitimate users in real-time, especially when multiple users participate in the same session. Although implicit authentication (IA) has been proposed to provide continuous and transparent authentication throughout a session, most existing IA solutions are optimized for single-user scenarios. The challenges of designing multi-user IA systems include fusing multiple modalities for good accuracy, segmenting and labeling behavioral data while authenticating, and adapting IA models to new users and new incoming data. We propose SHRIMPS, an evaluation framework to support IA researchers in the design of multi-user, multi-modal IA systems. SHRIMPS allows the evaluation of multi-user IA solutions that incorporate multiple modalities and supports adding new users and automatically labeling new incoming data for model updating. SHRIMPS supports different score fusion strategies, including a novel score fusion strategy based on Dempster-Shafer (D-S) theory to improve accuracy with considering uncertainties among different IA mechanisms. SHRIMPS enables composing tasks with public datasets to evaluate and compare different IA schemes. We present and evaluate two sample use cases to showcase how SHRIMPS helps address practical design questions of multi-user, multi-modal IA systems. The evaluation results show that D-S theory based score fusion methods can effectively reject attackers and detect user switches for the multi-user scenario in real-time.

1. Introduction

Smart devices play a significant part in people's daily life. Since people are increasingly relying on smart devices to access personal and corporate data, the demand for security and usability drives the evolution of user authentication mechanisms. Researchers have introduced more usable authentication mechanisms, such as fingerprint and face recognition, to replace passwords. However, these authentication mechanisms only authenticate a user once for unlocking and fail to provide protection afterwards. Behavioral biometrics based implicit authentication (IA) (Jakobsson et al., 2009; Frank et al., 2012; Bo et al., 2014) is a promising technology that provides continuous and transparent protection by leveraging distinct users' behaviors. Most existing IA systems (Khan et al., 2014a; Crawford et al., 2013) are designed for a single user. With the expansion of usage scenarios, multi-user shared devices have become common, including shared household devices (Matthews et al., 2016; Al-Ameen et al., 2021) and mobile devices for front-line and medical workers (Microsoft Azure, 2023; Draffin et al., 2013). It is important to design an IA framework that secures sensitive data on shared smart devices in multi-user scenarios.

A multi-user IA system needs to identify a user in addition to rejecting imposters. A single behavioral biometric is insufficient to ensure good identification accuracy due to data unavailability or poor quality (Gofman et al., 2016). Existing studies (Vhaduri and Poellabauer, 2019; Hintze et al., 2019; Abuhamad et al., 2020) have shown that multi-modal authentication systems provide more accurate and robust

* Corresponding author.
E-mail address: urs.hengartner@uwaterloo.ca (U. Hengartner).
[1] Present address: Huawei Technologies Canada Co., Ltd., 300 Hagey Blvd, Waterloo, N2L 0A4, ON, Canada.

https://doi.org/10.1016/j.cose.2023.103594
Received 18 April 2023; Received in revised form 5 October 2023; Accepted 8 November 2023
Available online 10 November 2023

performance in comparison to single-modal systems. However, little effort has been made to design a multi-modal multi-user IA system.

Multi-user, multi-modal IA requires careful consideration as it is not a simple extension of a binary classification problem into a multi-class one. The challenges include: 1) *Heterogeneous authenticators.* Authenticators based on different behavioral biometrics provide different coverage (e.g., gait data is only available when a user is walking) and accuracy. A critical problem is how to organize different authenticators and aggregate their results to provide accurate identification and authentication for multiple users. 2) *Real-time detection of user switches.* It is common that the current user of an unlocked smart device changes to another valid user without an explicit account switch (Matthews et al., 2016; Al-Ameen et al., 2021). The system should recognize the valid user after a user switch in addition to rejecting attackers. 3) *New users and data.* IA mechanisms may experience accuracy degradation over time (Frank et al., 2012; Zheng et al., 2014; Chauhan et al., 2020). Existing single-user IA systems (Bo et al., 2014; Khan et al., 2014a) only need to update the device owner's IA models. However, for a multi-user system, we need to consider both new users and new data. Adding a new user requires updating the existing models in the system to distinguish the new user from existing users. When the system processes new incoming data, it should label it with the correct corresponding user. 4) *User data imbalance.* Some users (e.g., device owners) are more likely to have more training data compared to other valid users. When new users are added to the system, their training data is much less than existing users'. As a result, the system may have low accuracy for users with less training data. This user data imbalance also exacerbates the performance differences among various authenticators.

We propose SHRIMPS,[2] a novel IA evaluation framework that can model configurations where multiple modalities are used to provide transparent and continuous identification and authentication in multi-user systems. Our focus is on designing a general framework that helps security developers and researchers combine existing and new IA mechanisms to evaluate the accuracy of multi-user identification and authentication.

SHRIMPS is targeted at evaluating multi-user, multi-modal IA systems. Supporting multiple users and multiple modalities has already been studied in the context of multi-user, multi-modal *biometric* authentication systems (Ross and Jain, 2004; Oloyede and Hancke, 2016; Jing et al., 2018; Toli and Preneel, 2015). However, these systems are usually based on explicit authentication, where users are asked to take an explicit authentication action (e.g., putting their finger on a fingerprint reader, or placing their face in front of a camera). Therefore, some of the challenges faced by multi-user, multi-modal IA systems, like real-time detection of user switches, usually do not occur in multi-user, multi-modal biometric systems. However, other challenges, like heterogeneous authenticators, are similar, and solutions proposed for multi-user, multi-modal biometric systems may also apply to multi-user, multi-modal IA systems. Evaluating this applicability is not the focus of this paper. In our evaluation (see § 6), we configure different IA environments and use different state-of-the-art IA, modality fusion, and model updating algorithms to demonstrate the versatility of SHRIMPS. We acknowledge that our chosen algorithms may not necessarily be the best ones. We support researchers interested in evaluating additional algorithms by making the SHRIMPS framework available open source.

SHRIMPS is a simulation-based IA evaluation framework for evaluating multi-user, multi-modal IA systems. Such systems should not only be evaluated with SHRIMPS, but also with other tools, like user studies. However, user studies can be expensive, both in terms of time and money. SHRIMPS can be used by security developers and researchers for weeding out candidate configurations of multi-user, multi-modal IA systems that are unlikely to result in good performance in practice and for determining more promising configurations. These configurations can then be evaluated in user studies.

User studies can be performed in the lab or in the field. Field-based user studies are usually adopted for evaluating usability under practical settings. However, it is hard to capture unauthorized access and device theft in a natural setting (Hintze et al., 2019). In general, it is challenging to conduct user studies for multi-user scenarios (e.g., controlling the conditions for legitimate users and attackers). To trade off, trace-based evaluation is a good option for conducting tasks using real-world public datasets, which is the approach pursued by SHRIMPS.

By providing a framework, SHRIMPS allows IA researchers to focus on a particular research problem, like implementing a proposed IA algorithm and comparing it to existing algorithms (hopefully) already implemented in SHRIMPS, the proper hyperparameter tuning of an IA algorithm, or the proper division of data into training and test sets. SHRIMPS helps streamline this work. Other, often tedious tasks, like parsing input data or dealing with imbalanced data, are automatically taken care of by SHRIMPS.

Finally, another advantage of SHRIMPS is that it enables easier reproducibility of research results by other researchers. For example, researchers proposing a new IA algorithm can implement and configure this algorithm in SHRIMPS and release the implementation and configuration. Since SHRIMPS is open source, anyone can use the released information to reproduce the results and improve on them.

The contributions of our work include:

- SHRIMPS is the first multi-user, multi-modal IA evaluation framework for shared smart devices. The framework can detect unauthorized access from strangers and identify the current user from a group of valid users.
- The framework supports model updating with new data and users to ensure high accuracy for multiple users across sessions. It can automatically segment and label the newly collected behavioral data based on authentication results and user feedback.
- SHRIMPS supports different existing score fusion strategies. In addition, considering the performance differences among different modalities, we propose a Dempster-Shafer (D-S) theory (Sentz et al., 2002) based score fusion strategy to combine the authentication scores from multiple modalities for different users and incorporate it into SHRIMPS.
- SHRIMPS runs in a simulation environment. The environment supports easy and flexible construction of simulation tasks using public datasets, which benefits other researchers for evaluating their IA schemes in a multi-user setting.
- We conduct extensive simulation tasks to show that SHRIMPS can be used for evaluating different scenarios. For example, we show that D-S theory based score fusion achieves both low false acceptance rate and low false rejection/identification rate. Besides, we demonstrate multi-user multi-modal IA configurations that are able to detect user switches and identify the new user with low detection latency. Our comprehensive evaluation shows that with the help of SHRIMPS it is also possible to realize configurations that handle new users well and automatically label new incoming data for model updating.
- We release our implementation of SHRIMPS in open-source for other security researchers and developers.[3]

2. Related work

As our work investigates the design of *multi-user*, *multi-modal* implicit authentication schemes, we discuss related works on these two aspects. Besides, as SHRIMPS is designed for evaluating multi-user IA

[2] Short for "SHaRing-aware IMPlicit authentication System".

[3] https://github.com/cryspuwaterloo/jiayi_thesis_code/tree/main/shrimps/idauth.

systems, we also summarize the evaluation methods adopted by existing studies.

2.1. Multi-user implicit authentication

Implicit authentication (IA) transparently authenticates a user's identity to improve the security and usability of user authentication. IA leverages users' distinct device usage or behavioral patterns to distinguish a user from others in a non-intrusive way. On the one hand, IA provides an additional authentication factor to supplement explicit authentication mechanisms. Many IA mechanisms can continuously verify a user's identity in the background during device usage. For example, an attacker may launch a shoulder surfing attack to obtain the PIN code to unlock a device. Behavioral biometrics based IA mechanisms can still block the attacker from accessing the device by comparing the attacker's touch patterns or keystroke dynamics to the device owner's. On the other hand, IA helps reduce unnecessary explicit authentication requests for alleviating a user's burden faced by user authentication.

Researchers have investigated various behavioral biometrics for IA, including touch (Frank et al., 2012; Bo et al., 2014; Zheng et al., 2014), gait (Derawi et al., 2010; Zou et al., 2020), keystroke (Lamiche et al., 2019), etc. Although most of them regard the IA problem as a binary or one-class classification problem (Gupta et al., 2019), a few studies conducted preliminary explorations of multi-user scenarios recently. Ehatisham-ul Haq et al. (2018) leveraged physical activity patterns to identify the device owner and secondary users who have partial access to the device. Zou et al. (2020) used Deep Neural Networks (DNN) to conduct gait-based multi-user identification and authentication separately. However, in practice, the system is expected to detect unauthorized access and track user switches in real-time. In comparison, SHRIMPS handles both tasks simultaneously without training additional models. ContAuth (Chauhan et al., 2020) adopted iCaRL (Rebuffi et al., 2017) and EWC (Kirkpatrick et al., 2017) to address the incremental learning problem for DNN-based single-modal IA mechanisms to improve cross-session performance for multi-user scenarios. In comparison, SHRIMPS considers auto-labeling and further improves identification accuracy by incorporating multiple IA mechanisms. DriverAuth (Gupta et al., 2019) is a multi-user and multi-modal authentication solution for ride-sharing platforms. However, DriverAuth performs implicit authentication only at the beginning of a ride, while SHRIMPS targets multi-user, multi-modal IA for general purposes and supports continuous authentication.

2.2. Multi-modal authentication

Most existing work on multi-modal authentication, with the exception of DriverAuth (Gupta et al., 2019) (see above), has focused on single-user scenarios. Combining multiple behavioral biometrics enables IA to identify a user's identity with high confidence and lowers the chance of spoofing attacks. Abuhamad et al. (2020) classified the fusion methods into three levels: feature-level (Vhaduri and Poellabauer, 2019; Lamiche et al., 2019; Gupta et al., 2019), algorithm/score-level (Crawford et al., 2013; Hintze et al., 2019; Buriro et al., 2015; Saevanee et al., 2015), and decision-level (Fridman et al., 2015). Crawford et al. (2013) proposed a score-level weighted average fusion method that gives more weight to more recent detection scores. Buriro et al. (2015) calculated the weight based on the classifier performance for their weighted average fusion method. Vhaduri and Poellabauer (2019) designed a multi-modal solution for wearable devices with feature-level fusion of step counts, heart rate, calorie burn and metabolic equivalent of task. Smith-Creasey and Rajarajan (2019) adopted the Dempster-Shafer theory based score fusion for single-user scenarios. Our work extends the application of the D-S theory to cover multi-user scenarios. Shrestha et al. (2019) proposed ZEMFA to extract gait features from multiple devices to perform zero-effort authentication. CORMORANT (Hintze et al., 2019) was designed to provide risk-aware continuous authentication for single-user cross-device scenarios. It proposed two weighted score threshold fusion methods and a Kalman filter based score fusion method to fuse the authentication score from different devices. We compare the Kalman filter based fusion to our D-S theory based method. The evaluation results show that our approach outperforms the Kalman filter based method in terms of low false identification rate and false acceptance rate.

In summary, compared to existing studies, SHRIMPS provides the architecture and workflow of a multi-user IA system with considering user switches in mid-session, model updating with new data and users, and score fusion from multiple modalities. These challenges were not fully covered in existing systems and solutions.

2.3. Evaluating IA systems

Since IA is usually regarded as a classification problem, many existing IA studies (Frank et al., 2012; Bo et al., 2014; Abuhamad et al., 2020; Zou et al., 2020; Ehatisham-ul Haq et al., 2018; Khan et al., 2014b) evaluated their proposed schemes by determining common metrics, such as AUC, EER, FAR, or FRR, in an offline setting. Sugrim et al. (2019) found that such an evaluation is inadequate to show how the system performs outside ideal conditions. Eberz et al. (2017) used the Gini Coefficient to quantify the systematic errors and we adopt it to measure the error distribution change among different methods. Although researchers have evaluated their authentication systems with lab or field studies (Riva et al., 2012; Hayashi et al., 2013), it is inefficient to collect sufficient data for unauthorized access. To trade off, generating traces from public datasets for real-world scenarios has been used to evaluate a (single-user) IA system (Hintze et al., 2019). Besides, since an IA system needs to update its model from time to time, it is important to observe its performance over time (Chauhan et al., 2020). SHRIMPS fills the gap in evaluating multi-user IA systems by providing an evaluation framework that operates on real-world data. It enables IA researchers to build a multi-user IA system and compose trace-based tasks.

2.4. Multi-user, multi-modal biometric systems

Whereas research on multi-user, multi-modal IA systems is relatively recent, research on multi-user, multi-modal biometric authentication systems has been well established (Ross and Jain, 2004; Oloyede and Hancke, 2016; Jing et al., 2018; Toli and Preneel, 2015). Some of the problems studied in this existing research, like fusing modalities (Oloyede and Hancke, 2016; Jing et al., 2018; Ross and Jain, 2003; Dinca and Hancke, 2017; Ryu et al., 2021) or model (i.e., template) updating (Rattani et al., 2009; Pisani et al., 2019), also occur in multi-user, multi-modal IA systems. Therefore, solutions proposed for these systems may also be applicable to multi-user, multi-modal IA systems. Whereas studying this applicability is outside of the scope of this paper, the SHRIMPS framework can be a useful tool for undertaking such a study.

3. Problem and modeling

In this section, we formulate the multi-user, multi-modal IA problem and provide the threat model.

3.1. Authentication model

3.1.1. Definitions and assumptions

In a multi-user IA system, two or more users are allowed to access a device. We define a user who is registered and has full or partial access to the device as a *legitimate user*. We define a *session* as the period of user-device interaction that starts from when the device is unlocked with explicit authentication, such as a PIN, to when the device is locked. The IA system continuously identifies the current user and verifies their identity throughout a session. As a consequence of failed authentication,

the system locks the device and asks for explicit authentication. Inspired by recent device/account sharing studies (Matthews et al., 2016; Al-Ameen et al., 2021; Marques et al., 2019), we consider multiple users sharing the *same* smart device, and therefore, participating in the *same* session alternatingly, where there may be more than one legitimate user during a session. We assume that there is only one user interacting with the device at any moment. For example, a shared tablet is running a kiosk app for medical staff to look up and process patients' data. A session starts when a medical worker turns on the tablet, and any legitimate medical worker can access the device afterwards. A session ends when the tablet is turned off or detects unauthorized access. It requires continuously and implicitly (re-)identifying the new user from all other legitimate users in real-time during a session.

3.1.2. Problem formulation

The multi-user IA problem is a multi-class classification problem. We denote the legitimate user set as $\mathcal{U}^+ = \{u_0, u_1, \dots, u_{n-1}\}$, where n is the number of legitimate users, user u_0 is the *primary user* (i.e., *owner*) of the device, and users $u_i, i > 0$ are *secondary users*. We define a *null user* or *attacker* as a user who is not registered and has no access to the device, which is denoted as u_{-1}. The whole user space for a multi-user authentication system is defined as $\mathcal{U} = \mathcal{U}^+ \cup \{u_{-1}\}$.

For accurate identification and authentication, the system adopts multiple IA mechanisms (i.e., *authenticators*). The basic workflow of each authenticator is to extract features from sensor measurements and perform multi-class classification. An authenticator can be described as a function $\boldsymbol{s} = M(\boldsymbol{f})$, where $\boldsymbol{s} = \{s_{-1}, s_0, s_1, \dots, s_{u-1}\}$ represents the normalized scores of all instances in $\mathcal{U}$, and $\boldsymbol{f}$ is the feature vector. Then, each authenticator obtains a series of feature vectors with timestamps $\{(t_0, \boldsymbol{f}_0), (t_1, \boldsymbol{f}_1), \dots, (t_k, \boldsymbol{f}_k)\}$ and generates a series of score vectors $\{(t_0, \boldsymbol{s}_0), (t_1, \boldsymbol{s}_1), \dots, (t_k, \boldsymbol{s}_k)\}$ accordingly, where k is the number of the classification times performed within a given period. The system then identifies the user and decides whether to lock the device. Thus, the multi-user, multi-modal IA problem is about combining different authenticators to obtain who is the most likely user.

3.2. Threat model

For multi-user IA, possible attackers include strangers and legitimate users. A stranger attacker is physically close to the device and attempts to access sensitive resources, which is a lunchtime attack (Kaczmarek et al., 2018). A legitimate user attacker may intentionally or accidentally access the *previous* legitimate user's resources. For both cases, the authentication system should reject their access and de-authenticate the current user. We assume attackers do not have or know the victim's credentials (e.g., password, PIN) for explicit authentication. We also assume the device and its operating system are trusted, and attackers cannot install malicious apps or tamper with the system services (e.g., modifying sensor inputs). Since our work focuses on a general multi-user IA framework, mimicry attacks (Khan et al., 2018) that target specific behavioral biometrics are out of the scope of our paper. Nevertheless, we test the system under the scenario where the accuracy of one authenticator is significantly lower than other authenticators (see § 6.2.3).

4. Multi-user IA

SHRIMPS first addresses the multi-user IA problem in § 3.1.2 from the following three aspects: 1) a general extension strategy to extend existing binary or one-class IA algorithms into multi-user, 2) a score fusion method to combine multiple modalities, and 3) new incoming data and user enrollment for model updating.

4.1. Multi-user identification

A multi-user IA model is an $n+1$-class classifier for a system with n legitimate users u_i with $0 \le i < n$, where negative instances (i.e., imposters) are denoted u_{-1} (see § 3.1.2). Thus, we need an imposter training set to provide negative training data. In SHRIMPS, the imposter training set is sampled from multiple randomusers, which are different from the legitimate users and the users pretending to be imposters later during model testing, to represent a "general" user's behavioral biometrics.

Besides, the machine learning technique adopted by a multi-user IA system should support multi-class classification. If a multi-class IA classification algorithm is available, SHRIMPS can directly take advantage of it. For binary classification algorithms, SHRIMPS adopts the generic "one-vs-the-rest" strategy to extend their models into multi-class classifiers: 1) For each class u_i, we construct a training set with labeling u_i as positive class and all other classes as negative class. 2) We train n sub-classifiers for all n classes using the training sets constructed in step 1. 3) For authentication, the authenticator calculates the normalized scores of the positive classes from all sub-classifiers and constructs a score vector as the output. SHRIMPS also supports one-class classification algorithms, which it extends for multi-class classification in the same way as binary classification algorithms. Recent IA research comparing one-class to binary classifiers (Giovanini et al., 2022; Özlem Incel et al., 2021; Wang et al., 2023; Ray-Dowling et al., 2022; Georgiev et al., 2022b; Vhaduri et al., 2021; Cheung and Vhaduri, 2020) has consistently shown that binary classifiers perform better. Binary classifiers have the disadvantage that they require negative training data, which one-class classifiers do not.

Multi-user scenarios also result in the user data imbalance problem, where we have different amounts of training data for different users. For example, a multi-user system may collect more training data for the owner compared to the other users since the owner usually spends more time doing various activities with the device. Thus, we need to balance the training data by resampling techniques, including downsampling the data for the majority classes and oversampling the minority classes (e.g., SMOTE (Chawla et al., 2002)). But the resampling techniques cannot fully address the accuracy degradation problem (Fernández et al., 2018). We still need to consider the accuracy imbalance among different users for decision making. We elaborate on this challenge as a part of the score fusion strategy in § 4.2.

To achieve multi-user identification, SHRIMPS handles the generation of balanced training data from a user's historical data and the imposter training set, and provides a generic wrapper to extend existing IA mechanisms into multi-class classification (see § 5.1.1).

4.2. Multi-modal score fusion

We fuse the results of multiple authenticators at score-level to provide accurate identification for multiple users since it allows each modality to work separately. SHRIMPS is designed to support various score fusion methods to aggregate the results from multiple modalities to make decisions. However, the scores produced by different modalities may have different implications such as the likelihood of each user, the similarity to a user's behavioral profile, etc. Also, it is necessary to take the uncertainty of each modality into account. Thus, calculating the average score is not sufficient. In SHRIMPS, we also adopt the Dempster-Shafer theory (Sentz et al., 2002) for score fusion since it is proposed to combine evidence (i.e., scores) from different sources (i.e., modalities) with uncertainty, which is usually applied for sensor fusion problems (Wu et al., 2002). Smith-Creasey and Rajarajan (2019) adopted a D-S theory based score fusion method for multi-modal IA schemes in the single-user scenario. In our study, we explore the multi-user D-S theory based score fusion method by decomposing the problem into $n+1$ binary cases.

For $u_i \in \mathcal{U}$, there are two mutually exclusive states: positive S_i and negative $\bar{S}_i$. The frame of discernment Ω_i is defined as $\Omega_i = \{S_i, \bar{S}_i\}$. All subsets in the power set $2^{\Omega_i} = \{\emptyset, \{S_i\}, \{\bar{S}_i\}, \Omega_i\}$ are assigned a *basic belief mass* within $[0, 1]$, denoted by m, where $m(\emptyset) = 0$, $\sum_{A \in 2^{\Omega_i}} m(A) = 1$. For an authenticator M that outputs a score vec-

tor $\boldsymbol{s} = \{s_{-1}, s_0, s_1, \ldots, s_{n-1}\}$, we define its uncertainty on each class (i.e., user) as $\boldsymbol{v} = \{v_{-1}, v_0, v_1, \ldots, v_{n-1}\}$. For each class, we construct the masses attributed for all hypotheses in 2^{Ω_i} as:

$$m(\emptyset) = 0, m(\{S_i\}) = (1 - v_i)s_i,$$

$$m(\{\bar{S}_i\}) = (1 - v_i)(1 - s_i), m(\Omega_i) = v_i.$$

To combine the masses of hypothesis $A = \{P_i\}$ from two authenticators M_p and M_q (the belief functions are denoted by m_p and m_q, respectively), we use Dempster's rule of combination to calculate its joint mass as

$$m(A) = m_p(A) \oplus m_q(A) = \frac{\sum_{B \cap C = A \neq \emptyset} m_p(B) m_q(C)}{1 - \sum_{B \cap C = \emptyset} m_p(B) m_q(C)}.$$

The combined belief $\text{Bel}(\{S_i\}) = \sum_{A|A \subseteq \{S_i\}} m(A) = m(\{S_i\})$ is the fused score for u_i from multiple authenticators.

We determine the uncertainty $\boldsymbol{v}$ of each authenticator by their model accuracy based on the following observations: 1) An authenticator may have a better accuracy detecting certain classes compared to others. 2) Different authenticators may have different accuracy for the same class. Intuitively, a higher accuracy on a certain user u_i should contribute to a lower uncertainty v_i. In our work, the system leaves 10% of the collected data out of the training data for each authenticator to construct their validation sets. Then, it evaluates all IA models with their corresponding validation sets at each model training or updating. The accuracy metrics include the per-user area under the receiver operating characteristic curve (AUROC) and equal error rate (EER), the threshold for the equal false acceptance rate and false rejection rate of each user. We adopt two uncertainty functions based on either AUROC or EER. Given the authenticator M and the target user u_i, the uncertainty is:

$$v^{\text{AUC}}_{M,i} = \min(0, 1 - \text{AUROC}_{M,i}), \tag{1}$$

$$v^{\text{EER}}_{M,i} = \max(1, 2 * \text{EER}_{M,i}). \tag{2}$$

Assume there are k authenticators $\mathcal{M} = \{M_0, M_1, \ldots, M_{k-1}\}$, and the average score vector of all authenticators is denoted as $\{\bar{s}_0, \bar{s}_1, \ldots, \bar{s}_{k-1}\}$. We use the D-S theory to merge the average score vectors of all authenticators. For each class, we obtain the fused score for each user $\hat{s}_i = m_0(\{S_i\}) \oplus m_1(\{S_i\}) \oplus \ldots \oplus m_{k-1}(\{S_i\}), i \in \{-1, 0, 1, \ldots, u-1\}$. Finally, we choose the most likely user by $\text{res} = \underset{i \in \{-1,0,1,\ldots,u-1\}}{\arg\max} \hat{s}_i$ as the current user.

In addition to score fusion based on Dempster-Shafer theory, SHRIMPS also supports other fusion methods, such as average and weighted average. This demonstrates that SHRIMPS supports both simple and complex fusion strategies. Due to its open-source nature, SHRIMPS can also be used for evaluating other score fusion methods proposed in earlier work, including fusion methods proposed for multi-modal biometric authentication (Oloyede and Hancke, 2016; Jing et al., 2018; Ross and Jain, 2003; Dinca and Hancke, 2017; Ryu et al., 2021). Such an analysis is outside of the scope of this paper.

4.3. New incoming data and users

In practice, IA models are not constant: 1) When a new user is added to the system (i.e., *user enrollment*), IA models need to be updated to identify the new user as a new class. The new user needs to complete tasks or use the device for a period of time so that the system can collect and label behavioral data for initial model training. In a deployed system, user enrollment would be initiated by an administrator. In SHRIMPS, user enrollment is indicated in the storyboard underlying the evaluated IA scheme (see § 5.2.2). 2) IA mechanisms require model updating with new incoming data to mitigate accuracy degradation over time. During normal device usage, the system is also collecting biometric data while authenticating and identifying the user. Unlike user enrollment, the system does not always know the ground truth of the current user's identity. Thus, we need to address the following problems:

4.3.1. Auto labeling

The common labeling strategy of single-user IA systems (Khan et al., 2014a; Crawford et al., 2013) is to label all incoming data as the owner's if no attack is detected. However, for multi-user systems, a piece of behavioral data may involve several users given possible user switches. Thus, the system needs to split the data into segments, where each segment contains *only one* user's usage data. Then, it finds out the corresponding user for each segment. Although external signals (e.g., screen-on/off) may imply user switches and indicate the start and end moments of a segment, they are insufficient to cover all user switches in a shared session. A multi-user IA system can continuously identify the current user and provide coarse-grained segmentation—knowing who is using the device during which segment. However, the time taken to collect sufficient data for decision-making is not negligible (evaluated in § 6). If a user switch is detected based on identification results without the help of external signals, the system discards the data collected during a time period (e.g., maximum detection latency) before the detected user switch since its ownership is uncertain. For the remaining data, the system labels the pre-switch part as the former user and the post-switch part as the latter user.

4.3.2. Model updating

User enrollment and new incoming data correspond to class incremental learning and data incremental learning, respectively. There are three types of model updating strategies. 1) *Full retraining* is applicable for all IA mechanisms. Models are retrained with all new and historical data. However, it occupies the most space; 2) *Partial fitting* is applicable for implementing data incremental learning to specific machine learning techniques, such as SGD-based techniques (Moctezuma et al., 2019) and Naive Bayes classifiers. They can update a trained model with new data without keeping the historical data. 3) *Incremental learning techniques* are applicable for DNN-based IA mechanisms (Zou et al., 2020; Shin et al., 2017). ContAuth uses EWC (Kirkpatrick et al., 2017) and iCaRL (Rebuffi et al., 2017) to update a model without storing all historical data. In SHRIMPS, we determine the suitable model updating strategy based on the IA mechanisms: we apply iCaRL for DNN-based IA mechanisms (since Chauhan et al. (2020) show that it is superior to EWS); for other IA mechanisms, we adopt full retraining for class/data incremental learning or partial fitting for data incremental learning.

SHRIMPS simplifies the comparison of existing model updating strategies for adaptive IA systems (Chauhan et al., 2020; Shen et al., 2023; Giovanini et al., 2022). Moreover, we observe that various template update methods have been developed for adaptive biometric systems (Rattani et al., 2009; Pisani et al., 2019). As in the case of existing fusion algorithms for biometric systems, SRIMPS makes it possible to study whether existing template update algorithms can also be used for adaptive IA systems. Such an analysis is outside of the scope of this paper.

SHRIMPS handles user enrollment and new incoming user in two steps: automatically segmenting and labeling the collected data, and updating IA models for all authenticators with appropriate strategies. Besides, it listens to the user's feedback to correct falsely labeled data. We describe the detailed workflow in § 5.1.2.

5. The SHRIMPS evaluation framework

In this section, we propose a multi-user, multi-modal IA evaluation framework, SHRIMPS, which consists of a multi-user IA system and an evaluation environment.

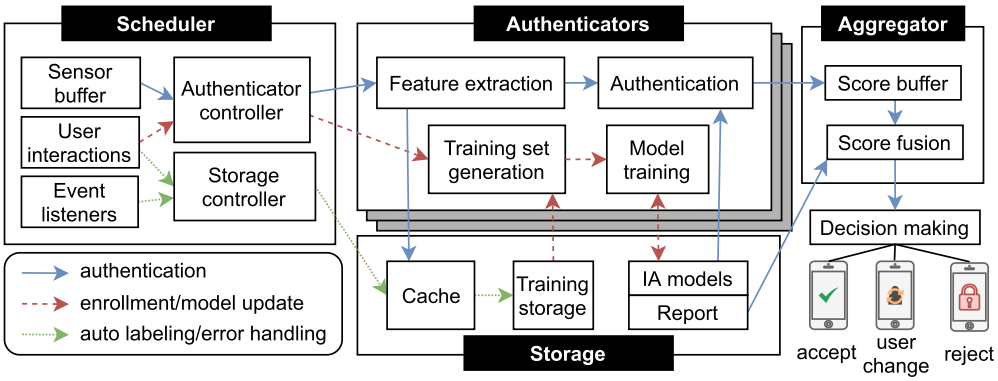


Fig. 1. Architecture of the SHRIMPS framework.

5.1. Multi-user IA system

We abstract the main components and behaviors of a multi-user IA system for shared smart devices, including user management, model training and updating, sensor data processing, and authentication.

5.1.1. Architecture

Fig. 1 shows the architecture of a multi-user IA system, which comprises of four modules: the scheduler, the authenticators, the aggregator, and the storage.

Scheduler receives sensor events and external signals, and coordinates authenticators and the storage module. The scheduler receives and caches the incoming sensor events in the sensor buffer. The authenticator controller is responsible for activating authenticators and invoking authentication or model training. Whenever there is sufficient sensor data, the authenticator controller activates that authenticator and dispatches the required sensor data. The scheduler also maintains a set of event listeners to receive and process external signals for auto labeling, model updating, and error handling (see § 5.1.2). External signals, such as screen-off, imply the end of a session or a possible user switch, resulting in clearing cached data, data segmentation and resetting the authentication status of the system. Besides, user feedback that occurs after an erroneous rejection or user switch decision is an important signal for error handling. In response, the scheduler fixes wrong labels of the cached data and sets the authentication status as authenticated.

Authenticators are responsible for providing the essential functions, including feature extraction, model training, and classification. Researchers can provide their own IA mechanisms by specifying these essential functions. If a provided IA mechanism is based on binary or one-class classification, SHRIMPS applies the multi-user extension introduced in § 4.1. For each authenticator, the feature extraction function takes raw sensor data as input and produces feature vectors as output. The authentication function feeds the feature vectors to the trained IA models to calculate the authentication scores. The model training function takes two sets of labeled feature vectors as input for training and testing, respectively. Internally, the model training function can further sample a subset of the training dataset for validation, which is usually used for tuning the hyperparameters of IA models. The testing dataset is used to pre-evaluate the accuracy of an authenticator. An authenticator needs to store the pre-evaluation results for the certainty calculation of multi-modal fusion.

The training set generation function is responsible for generating training and testing data for the authenticator. The function loads the history feature data of each user in the training storage and samples negative training data from the imposter training set. All the fetched data is used to construct a labeled dataset. It is optional to apply resampling techniques to produce a balanced dataset (i.e., all classes have the same data size). The processed data is divided into two parts in a configurable ratio for training and testing, respectively, which is provided for the model training function.

Aggregator collects and fuses authentication scores. Since scores from various authenticators arrive at the aggregator asynchronously, our strategy is to let the aggregator cache the recent score vectors within a specified time interval and fuse the scores based on the steps in § 4.2 (note: the multi-user also supports other score fusion methods such as average and weighted average). The cache is cleared at the session end or a user switch through final decisions or external signals. In addition, we adopt a (m, n)-sliding window: If at least m out of n results are the same, the aggregator adopts that result as the final decision; otherwise, it waits for more scores to make decisions. There are three types of decisions: accepting the user as the identified one, rejecting the user, and detecting a user change from one to another. Accordingly, there are three types of false decisions: 1) false acceptance (FA): the system falsely accepts an attacker, 2) false rejection (FR): the system falsely rejects a valid user, and 3) false identification (FI): the system identifies a valid user as another. We explore error handling in the next subsection.

5.1.2. Multi-user IA system workflow

We present the workflow of a multi-user IA system performing the following operations:

User Enrollment & Removal. SHRIMPS support user enrollment and removal events as external signals. For user enrollment, the system does not conduct authentication and only collects behavioral data for the new user. A piece of labeled behavioral data is directly added into the training storage. Model training is triggered as follows: authenticators fetch the training data from the storage, generate training datasets, and train their models. The models and their pre-evaluation reports are stored in the storage. User removal requires indicating the target user. SHRIMPS supports the following two options for removing a user: 1) If the system stores users' historic behavioral data, authenticators fully retrain IA models with all data except for the removed user. 2) If an IA model consists of several per-user classifiers, the system can remove that user's classifier. Their behavioral data is also removed from the training storage and excluded from any future model updating.

Authentication. The authentication system continuously collects sensor data in the scheduler. Once the authenticator controller detects sufficient data for a certain modality, it calls the corresponding authenticator with the cached sensor data. The authenticators extract features, load the saved model from the storage, and then conduct classification to obtain score vectors. Score vectors from different authenticators are sent to the aggregator for score fusing. Finally, the system determines whether to accept or reject the current user based on the fused score.

Model updating. SHRIMPS takes both external signals and identification results to segment and label the data automatically and dy-

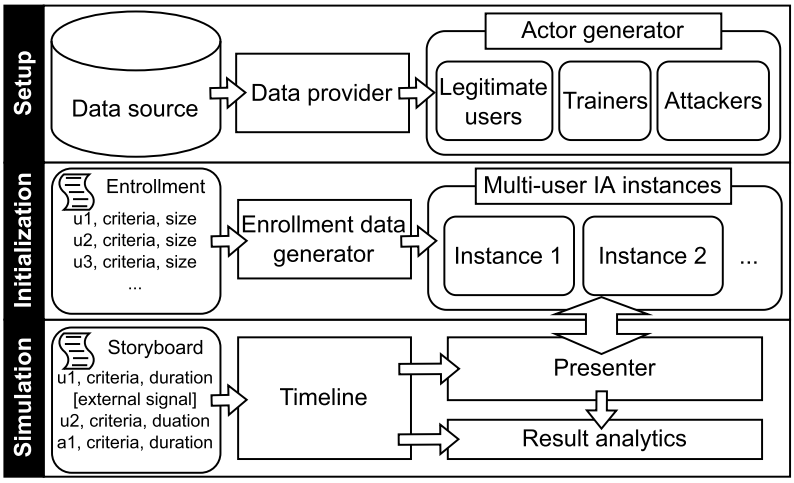


Fig. 2. Evaluation framework.

namically. Whenever the scheduler receives an external signal or the aggregator detects a user switch, SHRIMPS labels the collected data as a segment with the previously identified user and stores them in the training storage. It can automatically correct the misclassified data points of individual authenticators based on the overall decisions. For example, if a single data point is u_{-1} (i.e., attacker) and the overall decision is acceptance, the system will fix the label of this single data point. If the device mistakenly locks the user out, it can correct the detection as well as the labels for cached features based on the user's feedback (see error handling below). Once sufficient new data is collected, the authenticators update the models for the existing users based on their data incremental learning strategy (see § 4.3).

Error Processing. Error processing of an IA system takes the following measures based on the error types: 1) False acceptance may temporarily expose the device to an attacker. Since the system is continuously authenticating the user, it will stop the attacker whenever a rejection decision is made. 2) False rejection leads to explicit authentication. If SHRIMPS receives a legitimate user's feedback (i.e., the user has passed EA), it can correct the labels of the collected features and update the IA models. 3) False identification is not as obvious as the other two errors. Immediate user feedback is not guaranteed if there is no mandatory EA to verify a user's identity. Nevertheless, we can handle false identification using the following strategy: if the system detects frequent user switches within a short time (e.g., the user has changed more than two times in five consecutive decisions), it will issue a rejection decision and a request for identity confirmation. Once the system receives the user's feedback, it will correct the labels accordingly.

5.2. Evaluation framework

5.2.1. Motivation

Evaluating a multi-user IA system requires testing under various conditions. It involves measuring accuracy with different user numbers and training data sizes, and detection latency for identifying a user after a user switch. As an IA system updates its models with new incoming data and users, it is also necessary to track the overtime accuracy change considering the impact of auto labeling. Moreover, a false decision may have different implications for continuous authentication: For example, a user is more sensitive to false rejections since they interrupt device usage, while an individual false acceptance is tolerable as long as the system rejects an attacker within a reasonable time. A real-world user study is usually adopted for evaluating usability under practical settings. However, it is hard to capture unauthorized access and device theft (Hintze et al., 2019). Specifically, it is challenging to conduct user studies for multi-user scenarios (e.g., controlling the conditions for legitimate users and attackers). To trade off, trace-based evaluation is a good option for conducting tasks using real-world public datasets.

SHRIMPS enables researchers to stitch together data from public datasets and easily compose evaluation tasks based on specific requirements without falling into two common evaluation pitfalls (Georgiev et al., 2022a): 1) Non-contiguous training data selection, and 2) attacker data in training. It supports external signals, enrollment and user feedback (i.e., reactions to decisions). Besides, we introduce both decision-level and session-level metrics to compare different strategies and understand the practical performance of the system.

5.2.2. Evaluation process

As shown in Fig. 2, the evaluation process is divided into three stages. We introduce the components of the evaluation framework and their functions at each stage:

Setup. Researchers determine the data source. A data provider manages the connection to a public dataset, parses raw sensor data, and provides an interface for data retrieval. Internally, a user's data is stored in *blocks*, where each block contains sensor data of a user collected over a continuous period of time. It ensures contiguous data selection in chronological order, and no data in the training data blocks will appear in the evaluation tasks. The actor generator fetches a complete list of users via a data provider and randomly selects a specified number of actors from the list. There are three actor types: legitimate users, trainers, and attackers. As defined in § 3.1, legitimate users should be enrolled in and identified by the IA system. Trainers provide negative training data to construct an imposter training set. Attackers attempt to access the device and should be blocked by the system. SHRIMPS ensures the attackers' data will not be used in the model training of legitimate users.

Initialization. The initialization stage determines the initial system status. An enrollment script is required to determine which legitimate users have enrolled and how much training data has been collected for each user. The enrollment data generator parses the enrollment script and fetches training data via the data provider. Then, SHRIMPS instantiates the multi-user IA system, and adds the specified legitimated users and their training data for the initial model training. Multiple instances that adopt different schemes can co-exist in the same environment so that we can compare different schemes with the same conditions and inputs.

Evaluation. We introduce a *storyboard* to help researchers quickly design evaluation tasks. A storyboard lists one or a series of data blocks

with specifying the actor, the selection criteria (e.g., activity, location), and the duration. It provides the ground truth of data segmentation. To describe a session with the participation of multiple users, one can concatenate multiple data blocks of different actors. External signals, such as "screen off" and "screen on", can be added in between two data blocks to mark the start and end of a session. Besides, SHRIMPS supports adding user enrollment events during an evaluation task. We show simplified storyboards in § 6.

According to the storyboard, SHRIMPS can fetch the matched sensor data and automatically generate a *timeline* comprised of a series of events in chronological order. The timeline automatically adjusts the sensor event timestamps of each block to ensure that the new timestamps of every two consecutive blocks are coherent. Assume that a new block with m events, $\text{sess} = \{(t_0, data_0), (t_1, data_1), \ldots, (t_{m-1}, data_{m-1})\}$, is appended to a timeline, where the last event timestamp of the timeline is T. The new timestamps are adjusted as follows: $t'_i = t_i - t_0 + T + \Delta t, i = 0, 1, \ldots, m-1$, where Δt is the customized interval between two segments. The presenter is responsible for processing the timeline and communicating with the instances: While passing each event to the instances, it also receives and answers their decisions. If a false decision is made, the presenter records it and produces a user's feedback for correction. After traversing the entire timeline, SHRIMPS saves all scores and decisions. The result analytics module generates the metrics by comparing each decision with the ground truth provided by the timeline.

Measures & Metrics. Multi-user IA systems are evaluated at two levels: decision-level and session-level. At decision-level, we use three basic metrics: false acceptance rate (FAR) is the proportion of the acceptance decisions among all decisions made on an attacker, and false rejection rate (FRR) and false identification rate (FIR) are calculated as the proportion of FRs or FIs among all decisions made on a legitimate user, respectively. Eberz et al. (2017) propose the Gini coefficient (GC) to analyze the error rate distribution among users and quantify systematic errors. A high Gini coefficient means that errors are concentrated in a small group of users. SHRIMPS uses GC to supplement decision-level FAR, FRR, and FIR for analyzing error distribution. Session-level metrics aim to help understand the practical impact of false decisions on the whole session. We define session-level errors based on the following criteria: 1) False acceptance: the system fails to reject an attacker within a specific time period (i.e., valid attack window). 2) False rejection: the system makes *at least one* decision to reject a valid user during the whole session. 3) False identification: the system makes *at least one* false identification during the whole session. For user switches where the user changes from one to another without any external signals, we allow the system to take a specific delay (i.e., uninformed switch window) before making the correct decision. During this period, any false identification is ignored since it does not block the user.

Accordingly, we define session-level FAR, FRR, and FIR by dividing the corresponding error number by the total session number. In addition, we record the moment t_d of the first correct decision to measure the detection latency, which is calculated by subtracting the starting timestamp of the session t_0 from t_d.

5.3. Evaluation workflow

IA researchers can use SHRIMPS to design and evaluate multi-user IA schemes according to the following steps: The first step is to build the multi-user IA system, including adding authenticators, specifying the score fusion strategy, and adjusting the auto labeling and model updating behaviors. Researchers can choose to add their own IA mechanisms/score fusion strategies or use the built-in ones provided by SHRIMPS. The second step is to connect to a data source and generate actors. Researchers need to provide the source dataset and its data provider. SHRIMPS includes example data providers for the HMOG dataset (Sitová et al., 2015), the BB-MAS dataset (Belman et al., 2019), the IDNet dataset (Gadaleta and Rossi, 2018), and the Touchalytics dataset (Frank et al., 2012). Actor generation requires a random seed and the numbers of each actor type. The third step is to design an enrollment script and a storyboard. Then, SHRIMPS can run the evaluation task and output the raw results accordingly. Based on the external signals in the storyboard, the result analyzer can segment the evaluation results into sessions and produce the per-session results automatically.

6. Sample use cases

In this section, we present two sample use cases that use SHRIMPS to design trace-based tasks and evaluate multi-user IA schemes.

6.1. Common setup

IA mechanisms. Both use cases use the same set of behavioral biometrics for their multi-user IA schemes. For demonstration, we choose touch-based and gait-based IA mechanisms and use SHRIMPS to adapt state-of-the-art algorithms to multi-user identification: 1) touch-based IA uses 28 touch-related features based on the feature extraction algorithm of Touchalytics (Frank et al., 2012), and SHRIMPS enables the multi-class classification following the extension strategy in § 4.1. 2) gait-based IA adopts a CNN+LSTM-based gait identification algorithm (Zou et al., 2020), which already supports multi-user identification. For the gait authenticator, the sampling rate of motion sensors is set to 50 Hz. The authenticator extracts gaits from a 1024-sample segment and is set to perform authentication every 512 samples (=10.24 s). Thus, every two consecutive segments have 50% overlap.

SHRIMPS handles training data generation for each authenticator. We adopt the same data balancing settings: using SMOTE to oversample minority classes and ensuring that all classes (including the negative class) have the same training size. Note that SHRIMPS also supports researchers to compare different balancing methods and parameters to find the best settings.

Data source. In the evaluation, we use the following public datasets:

1. **HMOG (Sitová et al., 2015):** accelerometer data, gyroscope data, and touch events from 100 users performing reading, writing, and map navigation tasks. Each task lasts about 5–15 minutes. Our use cases focus on the reading and walking tasks since gait and touch data are available simultaneously for them.
2. **BB-MAS (Belman et al., 2019):** accelerometer data, gyroscope data, and touch events from 117 users. Each user completed a 25-minute typing task and a 10-minute walking task. Note: users did not perform any touch events while walking.
3. **IDNet (Gadaleta and Rossi, 2018):** accelerometer data and gyroscope data from 50 users performing 5-minute gait tasks.
4. **Touchalytics (Frank et al., 2012):** touch events from 41 users. Each user completed 3–4 web browsing tasks and 2–3 game tasks. Different from BB-MAS, the touch events are mainly vertical and horizontal swipes.

For all datasets, we select ten users as trainers to provide behavioral data for the negative class. SHRIMPS excludes these users from the legitimate user and attacker selections, ensuring no overlap between trainers and attackers to avoid the attacker-data-in-training pitfall. The evaluation datasets should include multiple users, sufficient cross-session sensor data for each user, and multiple modalities. Since only the HMOG dataset meets all requirements, our sample use cases use only it for most of the evaluations.

Compared to HMOG, BB-MAS does not provide cross-session gait data (i.e., only one 10-minute task for a user), while Touchalytics only provides touch data and IDNet only provides gait data. To address the lack of multi-modal public datasets, a compromise solution adopted by existing studies (Hintze et al., 2019; Gupta et al., 2019, 2022; Lopes

Silva et al., 2019) is to fuse multiple datasets for different modalities that are independent of each other and rely on different sensors (e.g., touch and gait). Therefore, we fuse the IDNet dataset and the BB-MAS (or Touchalytics) datasets as follows: 1) We map the 14 users who provided three or more tasks from IDNet to 14 users randomly selected from BB-MAS (or Touchalytics). 2) We randomly select ten other users from both datasets to provide data for the negative class. 3) For data fusion, we use each IDNet motion data block as a basis and extract the BB-MAS (or Touchalytics) touch events in the same duration. 4) We adjust the timestamps of the touch events to align them to those of the motion data. We acknowledge the limitation that merged user behavioral data may not be realistic. However, the fused dataset is only used to test the accuracy gain of different fusing methods for the multi-modal scenario where an authenticator is failing. It also demonstrates that SHRIMPS supports various public datasets.

6.2. Use case 1: fusion method comparison

A multi-user IA system is expected to identify each legitimate user and reject imposters under different settings. SHRIMPS enables IA researchers to compare different IA schemes and choose the best one in terms of accuracy and detection latency. Specifically, a multi-user system should detect user switches, which are common in household sharing (Matthews et al., 2016; Al-Ameen et al., 2021). Also, an attacker may grab the device from the owner, causing a sudden user change.

In this use case, we address the following questions:

1. How does adopting multiple modalities benefit multi-user IA compared to single modality solutions?
2. What score fusion method provides the highest overall accuracy considering false acceptance rate and false rejection rate?
3. Is it necessary to set the maximum user size for a multi-user IA system?
4. How fast and accurately can a multi-user IA system capture an uninformed user switch during a shared session?

We first explain what fusion methods we add to SHRIMPS. Then we describe the evaluation tasks that we execute in the framework.

6.2.1. Fusion methods

We tested different score fusion methods and compared them to single modalities to examine how they balance FAR, FRR, and FIR. The two baseline methods include single-modal gait-based IA and single-modal touch-based IA. The most widely used score fusion method is average-based fusion. Moreover, weighted average methods also take the authenticator's performance into consideration. To compare D-S theory based methods to the average-based methods, we apply the per-user AUCs and EERs as the weights for the average-based methods.

CORMORANT (Hintze et al., 2019) proposed a Kalman filter based score fusion method that is resistant to the noise of detection. We extend it into a multi-user fusion method by applying Kalman filter to multi-user scores for each user with the following settings: 1) Measurement uncertainty R is determined by the per-user EER, 2) Process uncertainty $Q = 0.25$: a large Q makes the estimated score emphasize on new scores (Hintze et al., 2019) (the selection of Q is explained in Appendix A). In summary, we compared the following methods:

- **Touch.** Applying the touch authenticator only.
- **Gait.** Applying the gait authenticator only.
- **Mean.** Calculating the average score of all authenticators.
- **Mean-AUC.** Calculating the weighted average score using AUC as the factor.
- **Mean-EER.** Calculating the weighted average score using 1-EER as the factor.
- **Kalman.** Applying Kalman Filter based score fusion.
- **DS-AUC.** Applying multi-user D-S theory based score fusion with the AUC-based uncertainty function (Eq. (1))
- **DS-EER.** Applying multi-user D-S theory based score fusion with the EER-based uncertainty function (Eq. (2))

6.2.2. Evaluation tasks

We design two groups of evaluation tasks to address the above questions. Given the limited data amount we use two fused datasets, IDNet+BB-MAS and IDNet+Touchalytics, only for the first group and adopt different settings for this dataset than for the HMOG dataset.

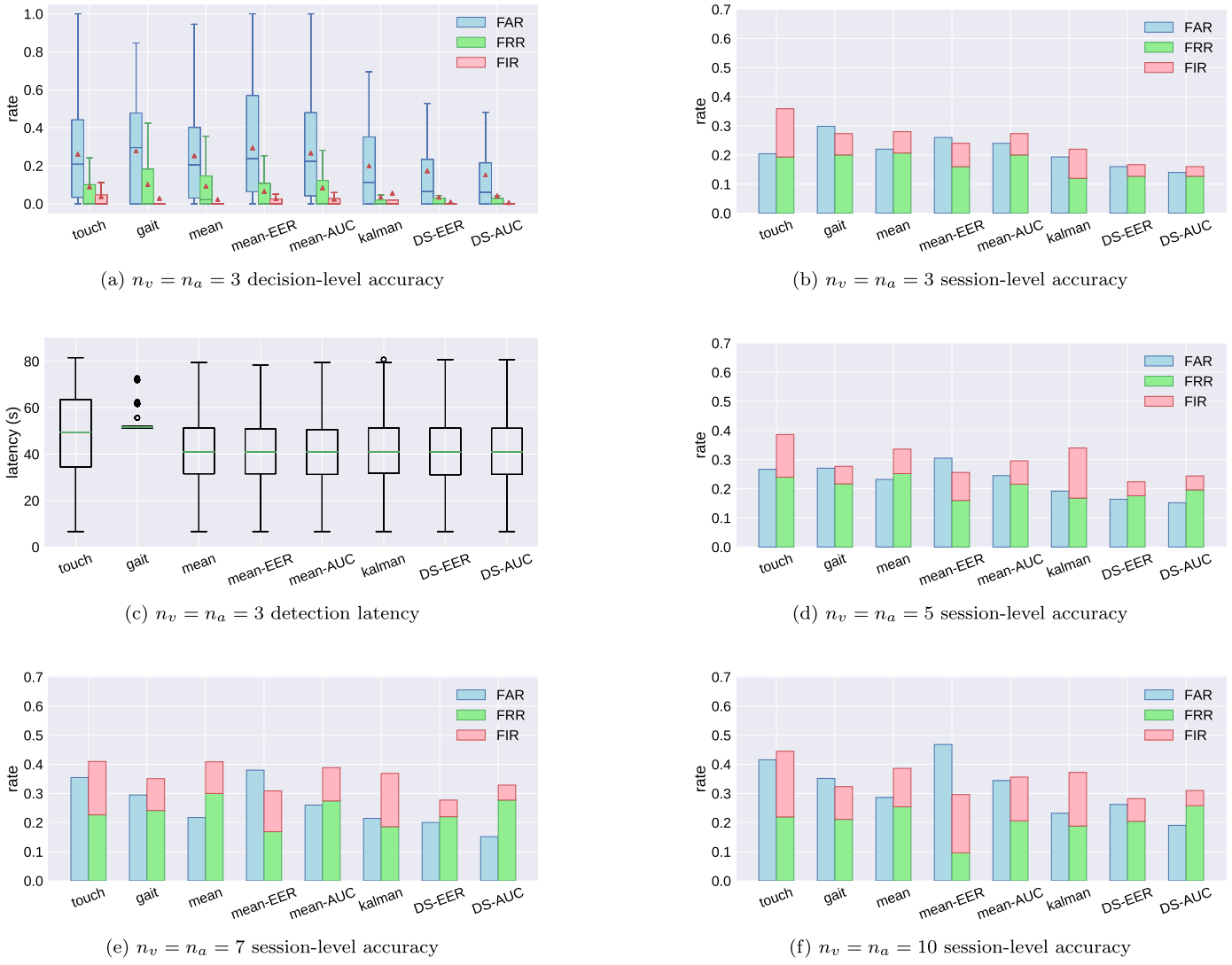
Group 1 (Accuracy evaluation) tests if the system can reject an attacker in a lunchtime attack and verify the identity of a legitimate user. The accuracy evaluation adopts a balanced static setting: there are equal numbers of legitimate users n_v and attackers n_a in each setup; each legitimate user has a fixed number of data blocks for initial enrollment and contributes to one fixed-length block for testing; therefore, there are $n_v + n_a$ blocks for each setup; we set an external signal between blocks to reset the authentication status. For HMOG, we used six enrollment data blocks for each user and set the testing block length as three minutes; we tried four different actor sizes, $n_v = n_a = 3, 5, 7, 10$ and tested 50 different actor combinations for each actor size. For IDNet+BB-MAS and IDNet+Touchalytics, we used two enrollment data blocks for each user and set the testing block length as two minutes; we tested $n_v = n_a = 3$ for 25 different actor combinations. In addition, given the length of the IDNet motion data blocks is much shorter than HMOG, we also reduce the segment size and the detection interval of the gait authenticator to 512 and 256 samples, respectively.

Group 2 (User switch evaluation) tests how each method detects user switches from a legitimate user to another legitimate user or an attacker in real-time. There is no external signal that informs the system of user switches. We assume that there are three legitimate users and three attackers in the task, where each legitimate user has six data blocks for initial enrollment. We composed three device sharing events and three attack events in the following storyboard: 1) u_0's block, u_1's block, [external signal]; 2) u_0's block, u_2's block, [external signal]; 3) u_1's block, u_2's block, [external signal]; 4) u_0's block, a_0's block, [external signal]; 5) u_1's block, a_1's block, [external signal]; 6) u_2's block, a_2's block, [external signal]. $a_{0,1,2}$ are three different attackers (i.e., u_{-1}). Each event (i.e., session) consists of two blocks from two different actors without any external signal in between to describe an uninformed user switch. The external signals in the storyboard only mark the end of each session.

6.2.3. Result analysis

We provide the result analysis as follows:

Group 1. Fig. 3a shows the results of the first group of evaluation tasks on HMOG, which includes the decision-level accuracy distributions of all eight methods at $n_v = n_a = 3$. The D-S theory based methods have the lowest FAR, FRR, and FIR, which means they can effectively reject attackers with less chance to falsely reject a legitimate user. Table 1 shows the GC of each error type (the error distribution curves are presented in Appendix C). High GCs on D-S theory based solutions imply that most errors were contributed by fewer users after applying the D-S theory based solution. The performance of score fusion methods is bounded by the fused modalities — The error rate concentrated on the users for whom both modalities have low accuracy. Session-level comparisons are in Fig. 3b. We find that the impact of FRs and FIs is magnified at session-level. Specifically, the touch-based method has a significantly high FIR. Among all methods, D-S theory based methods achieve the lowest overall false detection rate: FRR (0.13) and FIR (0.03), which



(a) $n_v = n_a = 3$ decision-level accuracy

(b) $n_v = n_a = 3$ session-level accuracy

(c) $n_v = n_a = 3$ detection latency

(d) $n_v = n_a = 5$ session-level accuracy

(e) $n_v = n_a = 7$ session-level accuracy

(f) $n_v = n_a = 10$ session-level accuracy

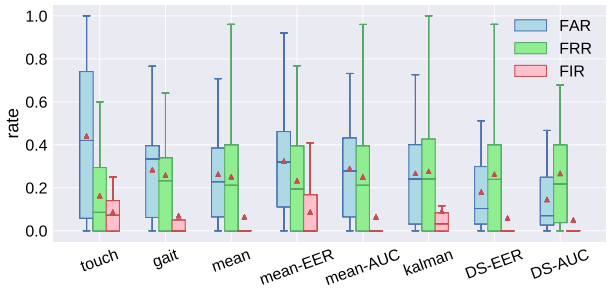
Fig. 3. Accuracy evaluation on HMOG. For each setting, the number of legitimate users (n_v) and attackers (n_a) are equal. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Table 1
Gini Coefficient of FAR, FRR, and FRR at $n_v = n_a = 3$.

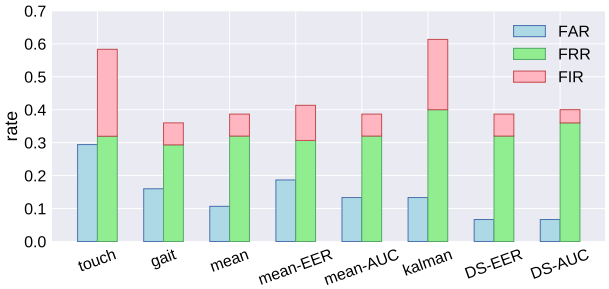
	touch	gait	mean	mean-EER	mean-AUC	kalman	DS-EER	DS-AUC
GC-FAR	0.70	0.70	0.70	0.66	0.69	0.74	0.78	0.81
GC-FRR	0.89	0.86	0.86	0.89	0.87	0.93	0.93	0.92
GC-FIR	0.91	0.94	0.95	0.95	0.95	0.94	0.98	0.98

means about 84% of the legitimate users' blocks are error-free. Although Kalman filter based fusion also achieves a low FRR (0.12), its FIR is significantly higher (0.10). We measure the latency as shown in Fig. 3c. Detection latency is determined by the adopted IA mechanisms: the touch authenticator relies on a user's interaction with the screen, and the gait authenticator using the default settings (Zou et al., 2020) performs authentication at a low frequency. Both take much time to collect sufficient data for making decisions. Since all multi-modal methods are implemented in SHRIMPS with the same configuration, there is no significant latency difference. Compared to single modalities, they improve the latency because they receive results from both modalities to make decisions earlier.

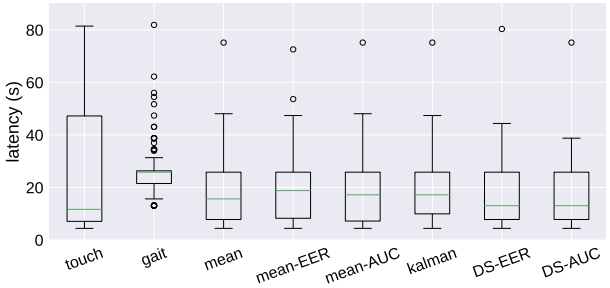
Fig. 4 shows the results on IDNet+BBMAS. This evaluation task compares the accuracy gain of different fusion methods when one modality performs significantly worse than the other. Due to the task setup, different swipe types (i.e., vertical swipes on the left and right parts of a screen and horizontal swipes on the bottom) are not evenly distributed in the time series. Consequently, patterns for some swipe types are not well learned by the touch based IA, which leads to poor accuracy. From Fig. 4a, we can see that the FAR of the touch authenticator was very high. However, the D-S theory based solutions still significantly reduced the FAR (the session-level FARs of DS-AUC and DS-EER are 0.07) compared to the other approaches. Besides, they can also improve the FIR. We can draw the same conclusion from the session-level results in Fig. 4b. For detection latency, we see the same trend for IDNet+BB-



(a) decision-level accuracy

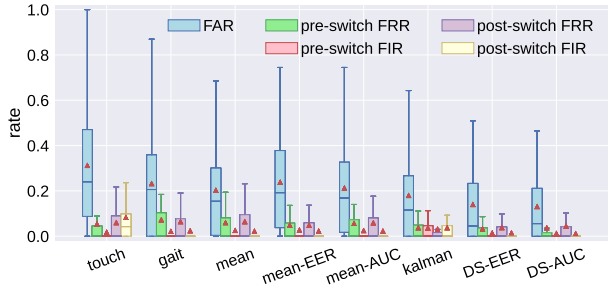


(b) session-level accuracy

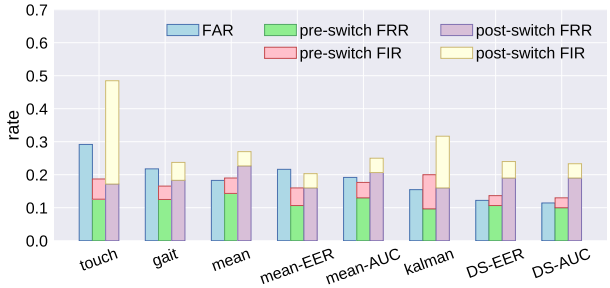


(c) detection latency

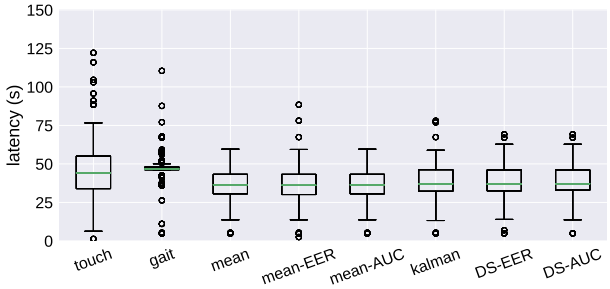
Fig. 4. Accuracy evaluation on IDNet+BB-MAS.



(a) decision-level accuracy



(b) session-level accuracy



(c) detection latency

Fig. 5. User switch evaluation results.

MAS as for HMOG. However, due to the shorter detection interval and earlier touch events, the overall latency for IDNet+BB-MAS is much shorter than for HMOG. In IDNet+Touchalytics, the touch-based IA had a higher overall accuracy. Except for lower FAR and FRR of the touch-based IA and all fusion methods, we observed a similar trend to IDNet+BBMAS and HMOG, and therefore, we put the results in Appendix B.

Based on the above results, we can answer questions 1 and 2: 1) Multi-modal methods achieve significantly better accuracy and balancing FAR, FRR, and FIR than single modalities, and 2) Among the tested fusion methods, D-S theory based methods have the lowest false detection rate.

When the legitimate user size is increased to 5, 7, and 10, we observe an increase in false decisions for all methods in Figs. 3d, 3e, and 3f. In

particular, the FAR rises significantly, which implies that the ability to detect attackers is weakened when classifying more classes. Nevertheless, DS-AUC can still well balance FAR and FRR/FIR, (FAR: 0.19, FRR: 0.26, FIR: 0.05, when u_v is 10). From the result, we can answer the third question: it is necessary to control the user size of a system to ensure high overall accuracy. IA researchers need to specify a threshold for accuracy and test different user sizes to determine the system capacity.

Group 2. Fig. 5a shows the results of the second group of tasks: the decision-level results for both pre-switch blocks and post-switch blocks are similar to the accuracy evaluation results. For attack events, DS-EER has the lowest FAR. However, for sharing activities, we can observe an increase in FIR and FRR for the post-switch blocks for all methods at session-level because of the detection latency — although the current user has changed to a different legitimate user, the authentication system still has no sufficient confidence in identifying this user. In Fig. 5b, DS-AUC and mean-EER still have better FRRs (0.19, 0.16) and FIRs (0.04, 0.04) compared to the other methods. However, the FAR of mean-EER (0.23) is much higher than that of DS-AUC (0.12). The high FIR (0.16) of Kalman filter based fusion shows that it is not a good option for handling user switches because its noise resistance makes it slow in response to sudden score changes. The detection latency results in Fig. 5c are similar to the first group of tasks. DS-AUC can provide low and stable detection latency (mean = 37.7 s, std = 10.8).

The results have shown that D-S theory based fusion methods can capture user switches during shared sessions with balancing FAR, FRR, and FIR compared to the other methods, which answers the fourth question. In addition, the results also imply the importance of external signals. If a signal, such as Android's Screen Pinning signal (Google Inc., 2023), may imply a user switch event, the system can then determine the end of a user's device use and reset the authentication status. Then, a user-switch task can be simplified as an accuracy evaluation task, where external signals assist in data segmentation to improve accuracy (see the second use case).

In summary, with the help of SHRIMPS, we were able to answer our questions for the first use case. We summarize the answers below.

1. Multi-modal methods achieve significantly better accuracy and balancing FAR, FRR and FIR than single modalities.
2. Among the tested fusion methods, D-S theory based methods have the lowest false detection rate.
3. It is necessary to control the user size of a system to ensure high overall accuracy.
4. D-S theory based fusion methods can provide low and stable detection latency for user switches.

6.3. Use case 2: multi-user model updating

Compared to the balanced and static settings adopted in the first use case, the second use case considers more factors: First, given that the owner usually spends more time with the device and contributes more training data than a secondary user, we test how the system handles imbalanced user data. Second, as new incoming data is used for model updating and new users are added into the system, the detection accuracy of the system may change over time. Third, user feedback towards false decisions may influence the identification and auto labeling processes. In addition, a user's lifting and putting down the device and other events may indicate the starting and the end of device usage and can be used to segment the data, which are considered as external signals to the multi-user IA system. The evaluation tasks should address the following questions:

1. What is the accuracy difference of the system identifying the owner and the secondary users?
2. How does new incoming data affect the accuracy of the system identifying different users?
3. How does external signals and user feedback benefit data segmenting and labeling in term of the overall accuracy?

6.3.1. Comparison strategies

From the first use case, we conclude that the D-S theory based methods can accurately detect attackers and identify legitimate users. Thus, we adopt DS-EER for score fusion. To address the above questions, we compare three strategies: 1) *baseline*: the system only supports user enrollment and does not learn from historical data (i.e., no model retraining after each part); 2) *uninformed*: the system makes decisions and performs auto labeling based on identification results and user feedback, and ignores external signals; 3) *informed*: the system additionally uses external signals for detecting a user switch and auto labeling.

6.3.2. Evaluation tasks

For setup, there are three legitimate users: the owner u_0 and the secondary user u_1 have already enrolled, and a new user u_2 will enroll in the system during the task. For initial enrollment, u_0 has three blocks, and u_1 only has one. The block length is randomly sampled, ranging from two to five minutes based on the high variance reported by Harbach et al. (2014). To show the accuracy change over time, we split the task into three parts at each model re-training and design the following storyboard:

1. (2 blocks): u_0's block, [external signal], u_1's block, [external signal], model retraining;
2. (3 blocks): u_2's enrollment, [external signal], u_0's block, [external signal], u_1's block, [external signal], u_2's block, [external signal], model retraining;
3. (3 blocks): u_0's block, [external signal], u_1's block, [external signal], u_2's block, [external signal].

External signals are only used in the informed strategy, while user feedback is used in both the informed and the uninformed strategies: External signals indicate device handoff where a legitimate user passes the device to another. User feedback indicates the current user has successfully passed the explicit authentication, which means the IA system made a false rejection. As a result, the evaluation framework will notify the system of the false decision. Model retraining is not applied to the baseline strategy. For each part, we measure the accuracy over all the blocks within. We repeat the task with 150 different actor combinations.

6.3.3. Result analysis

Table 2 shows the per-user results for each part. In Part 1, we can observe that the system had a lower false detection rate at identifying the owner than a secondary user when there is not much training data. This difference becomes smaller when more training data is available for secondary users due to the data balancing strategy, which answers the first question.

After u_2's enrollment at the start of Part 2, the system updated all IA models. By the end of Part 2, the system experienced significant accuracy degradation in identifying u_0 compared to Part 1. However, due to score fusion and data balancing, the accuracy of identifying the new user is close to identifying u_1. At the end of Part 2, the system retained all models with new data collected in Part 2. In Part 3, the false decision rate dropped. Compared to the baseline, the FRR and FIR of the informed strategy were lower for all users. The results have

Table 2
Per-part results for use case 2. Three legitimate users: u_0: primary user; u_1: secondary user; u_2: new legitimate user. False decision rate (FR) is the sum of FRR and FIR.

	user	Part 1			Part 2			Part 3		
		FRR	FIR	FR	FRR	FIR	FR	FRR	FIR	FR
Baseline	u_0	0.25	0	0.25	0.43	0.05	0.48	0.48	0.03	0.51
	u_1	0.47	0	0.47	0.5	0.05	0.55	0.53	0.02	0.55
	u_2	-	-	-	0.43	0.03	0.46	0.53	0.02	0.55
Uninformed	u_0	0.21	0.01	0.22	0.38	0.11	0.49	0.37	0.10	0.47
	u_1	0.43	0	0.43	0.37	0.03	0.40	0.43	0.07	0.50
	u_2	-	-	-	0.38	0.03	0.41	0.29	0.04	0.33
Informed	u_0	0.23	0.01	0.24	0.43	0.04	0.47	0.33	0.03	0.36
	u_1	0.38	0.01	0.39	0.27	0.02	0.29	0.26	0	0.26
	u_2	-	-	-	0.27	0.02	0.29	0.25	0.01	0.26

addressed the second question: model updating can help improve cross-session accuracy significantly.

To answer the third question, we compare the uninformed and informed strategies across all parts. The results have shown that external signals can further improve accuracy because 1) they enabled the system to reset the authentication status at a user switch to avoid false identification, and 2) they provided precise data segmentation, which makes the system correctly label more behavioral features. Despite the benefits of improving accuracy, IA researchers also need to consider the usability of the system. For example, frequently asking for a user's feedback makes the system hard to use. With SHRIMPS, IA researchers can observe the frequency of the external signals and optimize the workflow by modifying the auto labeling and model updating mechanisms.

In summary, with the help of SHRIMPS, we were able to answer our questions for the second use case. We summarize the answers below.

1. The system had a lower false detection rate at identifying the owner than a secondary user when there is not much training data. This difference becomes smaller when more training data is available for secondary users due to the data balancing strategy.
2. Model updating can help improve the cross-session accuracy significantly.
3. External signals can further improve accuracy because 1) they enabled the system to reset the authentication status at a user switch to avoid false identification, and 2) they provided precise data segmentation, which makes the system correctly label more behavioral features.

7. Discussion

Limitations. We list the following limitations of SHRIMPS or sample use cases: 1) The design of simulation tasks is restricted by the dataset. For example, for HMOG, we limit the length of use case 2 to three parts to satisfy the cross-session requirement, which leads to high error rate for all strategies. 2) Although SHRIMPS supports simulated user feedback, there is still a gap between simulation and user studies in usability evaluation. Nevertheless, SHRIMPS can be used for tuning and evaluating a multi-user IA system before user studies. 3) Since SHRIMPS is a simulation framework, it is not for implementing and developing a deployable multi-user IA system on smart devices. However, it can be easily connected with the real systems for parameter tuning.

Applications. We present two use cases to exemplify how SHRIMPS helps IA researchers and developers design and evaluate multi-user IA schemes. We note that SHRIMPS can be applied in diverse scenarios. For example, it is feasible to use the simulation environment to generate a long simulation task consisting of random sensor data blocks and random external signals to test the robustness of a multi-user IA scheme. Besides, IA researchers can explore how much training data is required for different user types (e.g., owner and secondary users) to help balance the per-user accuracy by modifying the training set generation module.

Multi-user concurrent usage. In our paper, we assume only one user operating the device at the same time. Matthews et al. (2016) listed broadcasting as a type of device sharing, where multiple people are co-using a single device simultaneously. Recognizing all present legitimate users can be regarded as a multi-label classification problem. However, if the system is always assuming the device is under concurrent usage and performing multi-label classification, its accuracy is very likely to suffer given the problem complexity. Thus, we need a certain external signal indicating the concurrent usage context and then trigger multi-label classification.

Contextual information. According to the user switch simulation tasks, an IA system may falsely reject a legitimate user if it is uninformed. However, if it knows the context of user switch through external signals, the problem can be simplified as a general identification task. Existing studies (Hintze et al., 2019; Miettinen et al., 2014) also use contextual information to adapt IA systems for better accuracy or less battery consumption. A future avenue is to embed contextual information into SHRIMPS and establish connections between context and authenticators.

8. Conclusion

We proposed SHRIMPS, the first simulation framework for designing and evaluating multi-user, multi-modal IA schemes. SHRIMPS is targeted at IA researchers and developers and allows them to easily compose and evaluate different continuous identification and authentication strategies. We also proposed a Dempster-Shafer based score fusion strategy to combine multiple modalities. Finally, we presented and evaluated two use cases that use SHRIMPS to design a multi-user IA scheme with touch-based and gait-based IA mechanisms and addressed practical design questions. The evaluation results of the use cases showed the (in)effectiveness of different IA strategies and configuration settings.

CRediT authorship contribution statement

Jiayi Chen: Conceptualization, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. **Urs Hen-**

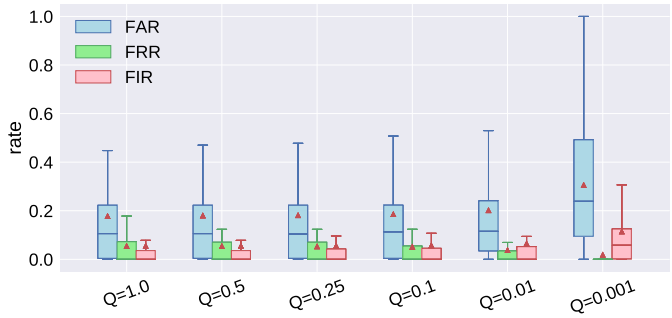
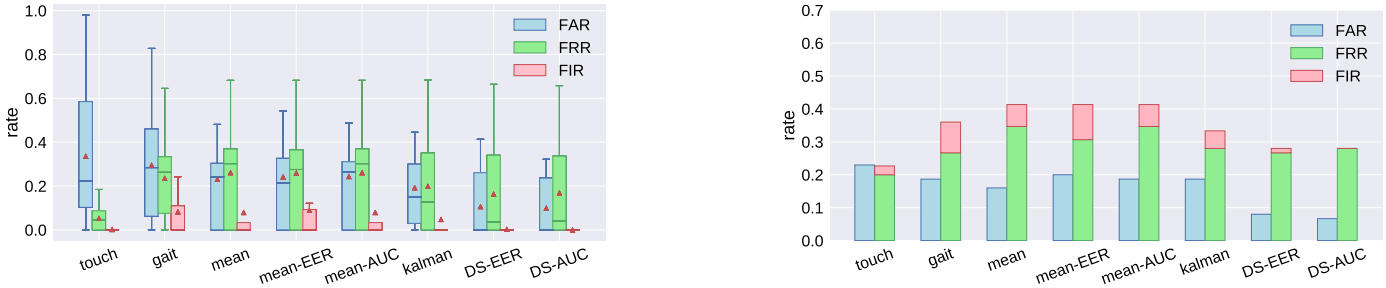


Fig. A.6. Decision-level FAR, FRR, and FIR of Kalman filter based fusion methods at different Q's ($n_v = n_a = 3$).



(a) decision-level accuracy (b) session-level accuracy

Fig. B.7. Accuracy evaluation on IDNet+Touchalytics.

gartner: Conceptualization, Funding acquisition, Supervision, Writing – review & editing. **Hassan Khan:** Conceptualization, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Urs Hengartner reports financial support was provided by Waterloo-Huawei Joint Innovation Laboratory. Jiayi Chen reports a relationship with Huawei Technologies Canada Co., Ltd. that includes: employment.

Data availability

The submission uses public data. The submission contains a link to the code.

Acknowledgements

This research has been supported by the Waterloo-Huawei Joint Innovation Laboratory.

Appendix A. Q selection for Kalman filter

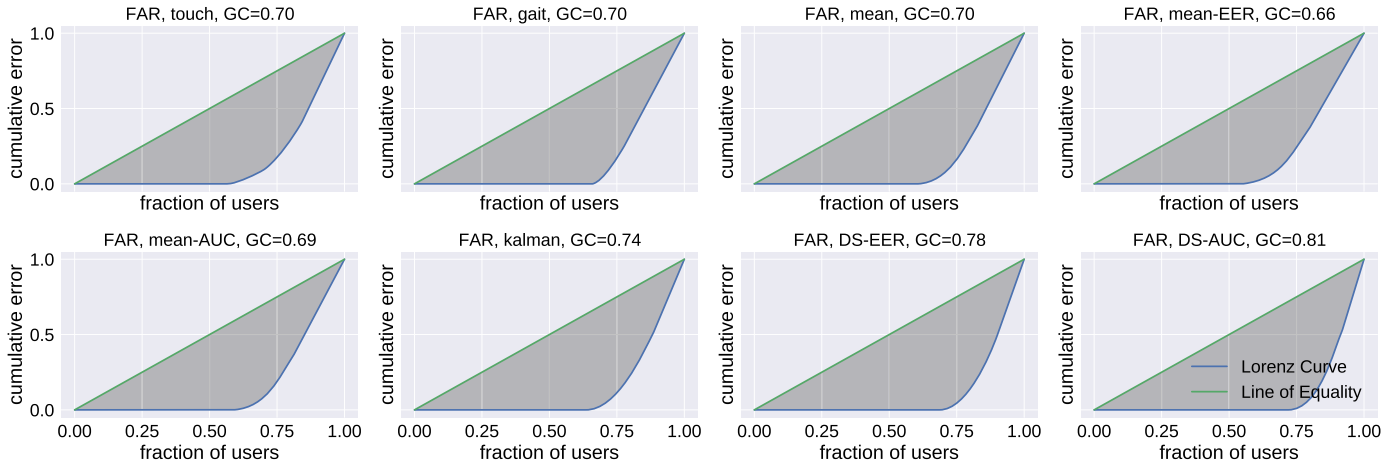
In the first use, we adopt the Kalman Filter settings from CORMORANT (Hintze et al., 2019). The Kalman filter assumes there is Gaussian noise in the state transition, which is modeled with a noise covariance matrix **Q**. CORMORANT highlighted the significant impact of the value selection of **Q**: A large **Q** has a smaller confidence in the system model and a larger confidence in the observations, which is desired by the score fusion purpose. To ensure the Kalman filter based score fusion method is optimized, we tested several Q values and compared their accuracy values using the same settings of the first use case, where $n_v = n_a = 3$. Fig. A.6 shows the decision-level metrics of different Q's. In general, a large Q results in a better FAR. An extremely small $Q = 0.001$ leads to high FAR and FIR. However, FRR and FIR do not decrease with larger Q. To balance the three metrics, we choose $Q = 0.25$ in our experiments, which has the lowest average FRR (0.05) plus FIR (0.06) with a relatively low FAR (0.18).

Appendix B. Results for IDNet+Touchalytics

In IDNet+Touchalytics, the decision-level FAR and FRR of the touch-based IA are significantly lower than IDNet+BB-MAS (see Fig. B.7). At the session level, the FAR and FRR+FIR of the touch-based IA are more balanced. After applying the score-level fusion, we can see lower FAR, while D-S theory based methods can achieve the lowest session-level FAR (DS-AUC: 0.07). However, due to the gait-based IA, all fusion methods have a higher FRR than the touch-based IA. However, DS-AUC can still achieve the lowest FRR+FIR among all methods.

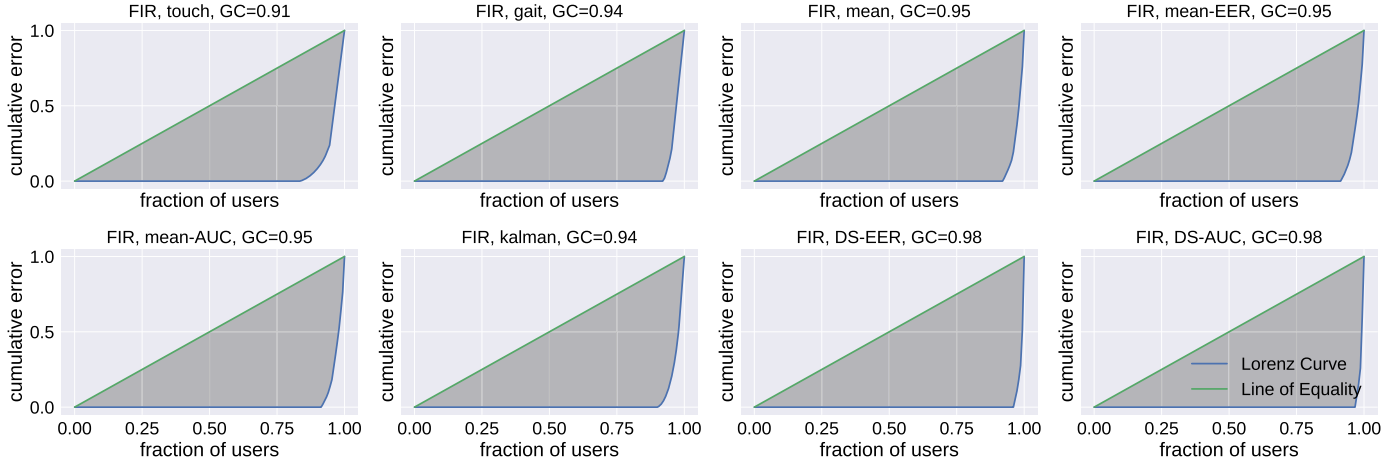
Appendix C. Gini coefficient for use case 1

The Gini Coefficient (GC) is calculated between the area between the Lorenz Curve and the Line of Equality. For evaluating IA systems, Lorenz Curve plots percentiles of the users on the x-axis according to error rate and plots cumulative error rate on the y-axis (Eberz et al., 2017). A point (x, y) on the curve indicates the normalized total error rate y contributed by the bottom x users. The Line of Equality is a straight diagonal line with a slope of 1, which represents that all users contribute to the same error rates. In Fig. C.8, we present the Lorenz Curve and the GC of all eight methods in the first group of evaluation tasks for the first use case. If we compare the D-S theory based fusion methods to other methods, we can observe fewer users contribute to more errors from the figure. For FAR, it means that the system may mis-classify a few "very successful" attackers as legitimate users. However, it is also possible that the system rejects an attacker at a certain time during continuous authentication. Thus, we also measure session-level metrics and detection latency to capture such situations. For FRR and FIR, the Lorenz Curve does not change as significantly as FAR, which means the error distribution remain similar among different methods. We can infer that score fusion strategies are effective in reducing random errors. To further improve the accuracy and reduce systematic error, a possible avenue is to incorporate new modalities.



(a) FAR

(b) FRR



(c) FIR

Fig. C.8. Lorenz Curve and Gini Coefficient of use case 1, $n_v = n_a = 3$.

References

Abuhamad, M., Abusnaina, A., Nyang, D., Mohaisen, D., 2020. Sensor-based continuous authentication of smartphones' users using behavioral biometrics: a contemporary survey. IEEE Int. Things J. 8 (1), 65–84.

Al-Ameen, M.N., Kocabas, H., Nandy, S., Tamanna, T., 2021. "We, three brothers have always known everything of each other": a cross-cultural study of sharing digital devices and online accounts. Proc. Priv. Enh. Technol. 4, 203–224.

Belman, A., Wang, L., Iyengar, S., Sniatala, P., Wright, R., Dora, R., Baldwin, J., Jin, Z., Phoha, V., 2019. SU_AIS BB-MAS (Syracuse University and assured information security-behavioral biometrics multi-device and multi-activity data from same users) dataset. IEEE DataPort.

Bo, C., Zhang, L., Jung, T., Han, J., Li, X.-Y., Wang, Y., 2014. Continuous user identification via touch and movement behavioral biometrics. In: 33rd International Performance Computing and Communications Conference (IPCCC). IEEE.

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.

Chauhan, J., Kwon, Y.D., Hui, P., Mascolo, C., 2020. ContAuth: continual learning framework for behavioral-based user authentication. In: ACM IMWUT '20 4 (4), 1–23.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.

Cheung, W., Vhaduri, S., 2020. Context-dependent implicit authentication for wearable device users. In: 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 1–7.

Crawford, H., Renaud, K., Storer, T., 2013. A framework for continuous, transparent mobile device authentication. Comput. Secur. 39, 127–136.

Derawi, M.O., Nickel, C., Bours, P., Busch, C., 2010. Unobtrusive user-authentication on mobile phones using biometric gait recognition. In: 2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing. IEEE, pp. 306–311.

Dinca, L.M., Hancke, G.P., 2017. The fall of one, the rise of many: a survey on multi-biometric fusion methods. IEEE Access 5, 6247–6289.

Draffin, B., Zhu, J., Zhang, J., 2013. Keysens: passive user authentication through micro-behavior modeling of soft keyboard interaction. In: International Conference on Mobile Computing, Applications, and Services. Springer, pp. 184–201.

Eberz, S., Rasmussen, K.B., Lenders, V., Martinovic, I., 2017. Evaluating behavioral biometrics for continuous authentication: challenges and metrics. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, pp. 386–399.

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. J. Netw. Comput. Appl. 109, 24–35.

Fernández, A., Garcia, S., Herrera, F., Chawla, N.V., 2018. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J. Artif. Intell. Res. 61, 863–905.

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.

Fridman, L., Stolerman, A., Acharya, S., Brennan, P., Juola, P., Greenstadt, R., Kam, M., 2015. Multi-modal decision fusion for continuous authentication. Comput. Electr. Eng. 41.

Gadaleta, M., Rossi, M., 2018. IDNet: smartphone-based gait recognition with convolutional neural networks. Pattern Recognit. 74, 25–37.

Georgiev, M., Eberz, S., Turner, H., Lovisotto, G., Martinovic, I., 2022a. Common evaluation pitfalls in touch-based authentication systems. arXiv:2201.10606.

Georgiev, M., Eberz, S., Martinovic, I., 2022b. Techniques for continuous touch-based authentication modeling. http://arxiv.org/abs/2207.12140.

Giovanini, L., Ceschin, F., Silva, M., Chen, A., Kulkarni, R., Banda, S., Lysaght, M., Qiao, H., Sapountzis, N., Sun, R., Matthews, B., Wu, D.O., Grégio, A., Oliveira, D., 2022. Online binary models are promising for distinguishing temporally consistent computer usage profiles. IEEE Trans. Biom. Behav. Identity Sci. 4 (3), 412–423.

Gofman, M.I., Mitra, S., Cheng, T.-H.K., Smith, N.T., 2016. Multimodal biometrics for enhanced mobile device security. Commun. ACM 59 (4), 58–65.

Google Inc., 2023. Android pin & unpin screens. https://support.google.com/android/answer/9455138?hl=en. (Accessed March 2023).

Gupta, S., Buriro, A., Crispo, B., 2019. DriverAuth: a risk-based multi-modal biometric-based driver authentication scheme for ride-sharing platforms. Comput. Secur. 83, 122–139.

Gupta, S., Kacimi, M., Crispo, B., 2022. Step & turn-a novel bimodal behavioral biometric-based user verification scheme for physical access control. Comput. Secur., 102722.

Harbach, M., Von Zezschwitz, E., Fichtner, A., De Luca, A., Smith, M., 2014. It's a hard lock life: a field study of smartphone (un) locking behavior and risk perception. In: Symposium on Usable Privacy and Security.

Hayashi, E., Das, S., Amini, S., Hong, J., Oakley, I., 2013. CASA: context-aware scalable authentication. In: Proceedings of the Ninth Symposium on Usable Privacy and Security, pp. 1–10.

Hintze, D., Füller, M., Scholz, S., Findling, R.D., Muaaz, M., Kapfer, P., Koch, E., Mayrhofer, R., 2019. Cormorant: ubiquitous risk-aware multi-modal biometric authentication across mobile devices. In: ACM IMWUT '19 3 (3), 1–23.

Jakobsson, M., Shi, E., Golle, P., Chow, R., 2009. Implicit authentication for mobile devices. In: USENIX Conference on Hot Topics in Security. USENIX Association.

Jing, K., Zhang, X., Xu, X., 2018. An overview of multimode biometric recognition technology. In: Proceedings of the 6th International Conference on Information Technology: IoT and Smart City, pp. 168–172.

Kaczmarek, T., Ozturk, E., Tsudik, G., 2018. Assentication: user de-authentication and lunchtime attack mitigation with seated posture biometric. In: International Conference on Applied Cryptography and Network Security. Springer, pp. 616–633.

Khan, H., Atwater, A., Hengartner, U., 2014a. Itus: an implicit authentication framework for Android. In: Proceedings of the 20th Annual International Conference on Mobile Computing and Networking, pp. 507–518.

Khan, H., Atwater, A., Hengartner, U., 2014b. A comparative evaluation of implicit authentication schemes. In: International Workshop on Recent Advances in Intrusion Detection. Springer, pp. 255–275.

Khan, H., Hengartner, U., Vogel, D., 2018. Augmented reality-based mimicry attacks on behaviour-based smartphone authentication. In: Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services, pp. 41–53.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al., 2017. Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. 114 (13).

Lamiche, I., Bin, G., Jing, Y., Yu, Z., Hadid, A., 2019. A continuous smartphone authentication method based on gait patterns and keystroke dynamics. J. Ambient Intell. Humaniz. Comput. 10 (11), 4417–4430.

Lopes Silva, P., Luz, E., Moreira, G., Moraes, L., Menotti, D., 2019. Chimerical dataset creation protocol based on Doddington Zoo: a biometric application with face, eye, and ECG. Sensors 19 (13), 2968.

Marques, D., Guerreiro, T., Carriço, L., Beschastnikh, I., Beznosov, K., 2019. Vulnerability & blame: making sense of unauthorized access to smartphones. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–13.

Matthews, T., Liao, K., Turner, A., Berkovich, M., Reeder, R., Consolvo, S., 2016. "She'll just grab any device that's closer" a study of everyday device & account sharing in households. In: ACM CHI '16.

Microsoft Azure, 2023. Overview of shared device mode. https://docs.microsoft.com/en-us/azure/active-directory/develop/msal-shared-devices. (Accessed March 2023).

Miettinen, M., Heuser, S., Kronz, W., Sadeghi, A.-R., Asokan, N., 2014. ConXsense: automated context classification for context-aware access control. In: Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security, pp. 293–304.

Moctezuma, D., Tellez, E.S., Miranda-Jiménez, S., Graff, M., 2019. Appearance model update based on online learning and soft-biometrics traits for people re-identification in multi-camera environments. IET Image Process. 13 (12), 2162–2168.

Oloyede, M.O., Hancke, G.P., 2016. Unimodal and multimodal biometric sensing systems: a review. IEEE Access 4, 7532–7555. https://doi.org/10.1109/ACCESS.2016.2614720.

Özlem Incel, 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.

Pisani, P.H., Mhenni, A., Giot, R., Cherrier, E., Poh, N., Ferreira de Carvalho, A.C.P.d.L., Rosenberger, C., Amara, N.E.B., 2019. Adaptive biometric systems: review and perspectives. ACM Comput. Surv. 52 (5).

Rattani, A., Freni, B., Marcialis, G.L., Roli, F., 2009. Template update methods in adaptive biometric systems: a critical review. In: Advances in Biometrics, pp. 847–856.

Ray-Dowling, A., Hou, D., Schuckers, S., Barbir, A., 2022. Evaluating multi-modal mobile behavioral biometrics using public datasets. Comput. Secur. 121, 102868.

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., Lampert, C.H., 2017. iCaRL: incremental classifier and representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Riva, O., Qin, C., Strauss, K., Lymberopoulos, D., 2012. Progressive authentication: deciding when to authenticate on mobile phones. In: 21st USENIX Security Symposium (USENIX Security 12), pp. 301–316.

Ross, A., Jain, A., 2003. Information fusion in biometrics. Pattern Recognit. Lett. 24 (13), 2115–2125.

Ross, A., Jain, A.K., 2004. Multimodal biometrics: an overview. In: 2004 12th European Signal Processing Conference, pp. 1221–1224.

Ryu, R., Yeom, S., Kim, S.-H., Herbert, D., 2021. Continuous multimodal biometric authentication schemes: a systematic review. IEEE Access 9, 34541–34557.

Saevanee, H., Clarke, N., Furnell, S., Biscione, V., 2015. Continuous user authentication using multi-modal biometrics. Comput. Secur. 53, 234–246.

Sentz, K., Ferson, S., et al., 2002. Combination of Evidence in Dempster-Shafer Theory, vol. 4015. Sandia National Laboratories Albuquerque.

Shen, Z., Li, S., Zhao, X., Zou, J., 2023. IncreAuth: incremental learning based behavioral biometric authentication on smartphones. IEEE Int. Things J. https://doi.org/10.1109/JIOT.2023.3289935.

Shin, S., Jung, J., Kim, Y.T., 2017. A study of an EMG-based authentication algorithm using an artificial neural network. In: SENSORS. IEEE.

Shrestha, B., Mohamed, M., Saxena, N., 2019. ZEMFA: zero-effort multi-factor authentication based on multi-modal gait biometrics. In: 2019 17th International Conference on Privacy, Security and Trust (PST). IEEE, pp. 1–10.

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.

Smith-Creasey, M., Rajarajan, M., 2019. A novel scheme to address the fusion uncertainty in multi-modal continuous authentication schemes on mobile devices. In: International Conference on Biometrics. IEEE.

Sugrim, S., Liu, C., McLean, M., Lindqvist, J., 2019. Robust performance metrics for authentication systems. In: Network and Distributed Systems Security (NDSS) Symposium 2019.

Toli, C.-A., Preneel, B., 2015. A survey on multimodal biometrics and the protection of their templates. In: Privacy and Identity Management for the Future Internet in the Age of Globalisation, pp. 169–184.

Vhaduri, S., Poellabauer, C., 2019. Multi-modal biometric-based implicit authentication of wearable device users. IEEE Trans. Inf. Forensics Secur. 14 (12), 3116–3125.

Vhaduri, S., Dibbo, S.V., Cheung, W., 2021. HIAuth: a hierarchical implicit authentication system for IoT wearables using multiple biometrics. IEEE Access 9, 116395–116406.

Wang, C., Xiao, Y., Gao, X., Li, L., Wang, J., 2023. A framework for behavioral biometric authentication using deep metric learning on mobile devices. IEEE Trans. Mob. Comput. 22 (1), 19–36.

Wu, H., Siegel, M., Stiefelhagen, R., Yang, J., 2002. Sensor fusion using Dempster-Shafer theory [for context-aware hci]. In: IEEE Instrumentation and Measurement Technology Conference, vol. 1. IEEE.

Zheng, N., Bai, K., Huang, H., Wang, H., 2014. You are how you touch: user verification on smartphones via tapping behaviors. In: 22nd International Conference on Network Protocols. IEEE.

Zou, Q., Wang, Y., Wang, Q., Zhao, Y., Li, Q., 2020. Deep learning-based gait recognition using smartphones in the wild. IEEE Trans. Inf. Forensics Secur. 15, 3197–3212.