Classification and Feature Extraction for User Identification for Smart Home Networks Based on Apps Access History

Yosef Ashibani
Department of Electrical, Computer and Software Engineering
Ontario Tech University
Oshawa, Ontario, L1G 0C5 Canada
yosef.ashibani@ontariotechu.net

Qusay H. Mahmoud
Department of Electrical, Computer and Software Engineering
Ontario Tech University
Oshawa, Ontario, L1G 0C5 Canada
qusay.mahmoud@ontariotechu.net

Abstract—**Advancements in smartphones has led to increasing dependency on them as an essential part of IoT smart home networks. Although this increase provides convenience to users, many challenges, including user identification and authentication, need to be considered. Many related works consider smartphone user authentication that assumes that the same user will be using the device throughout the access session. However, after the login stage, the device could be used by others, either in the home environment or outside, leading to undesired access to home appliances. In this paper, we present a continuous user identification approach that uses implicit features that can be integrated as a second layer of identification beyond the login step, based on user behavior on smartphones. The proposed method is mainly based on user interaction on the smartphone, which thus reflects the user's pattern that in turn provides the impetus to employ user identification as a complementary approach beyond the user login stage.**

Keywords—smart home networks, user identification, user authentication, multi-class classification

I. Introduction

One of the important IoT applications is the smart home network, also known as home automation. The popularity and adoption of smart home networks has increased in the last few years. In addition to accessing, operating, and controlling home appliances, advanced smart home networks provide many other services to home residents, such as entertainment storage information, and personal files. All these services are, in most cases, accessible to users by their smartphones/tablets, which have become global end-user devices. In addition to the primary use of performing calls, smartphones/tablets have become essential tools for performing different functions, such as text messaging, Internet browsing, social media application hosting, email checking, and game playing. Additionally, smartphones/tablets have become a main part of the smart home; they are the means from which home appliances are remotely controlled and operated, either locally through WiFi communication or broadly through the Internet.

As an example, Samsung's SmartThings, Wink and HomeKit are smart home platforms, as presented by the industry, which are built based on the cloud-back end service where control management and authentication are performed mostly through an installed application on end-user devices. Despite the increased adoption of smart home frameworks by consumers, many challenges still need to be considered, especially user identification. One of these challenges is that smartphone devices are susceptible to unauthorized access by other users when, for example, either lost or, in the home environment, where other household members or visitors are in a shared place. In consequence, after the login step, the home services (appliances) will remain accessible during the access session, hence could be used by unauthorized users. Consequently, this problem introduces security and privacy issues that need to be considered. Thus, there is a need to guarantee that the legitimate user is still using the end-device.

A solution can be achieved by utilizing continuous authentication and identification of the current user of the smartphone beyond the login stage. The authentication process verifies the user's identity, namely whether legitimate or unknown, while the user identification process verifies the current user's identity among other enrolled users. Hence, continuous authentication is the process of continuously checking the user' identity beyond the login stage, whereas continuous identification is the process of continuously checking user identity among other enrolled users. This brings the need to include an implicit authentication and identification mechanisms that derives features from the user while interacting with the mobile device, and utilizes these features for continuous authentication and identification. In turn, user interaction behavior can be implicitly modeled using machine learning classification strategies for users, which can be continuously utilized in the background as a second layer of authentication and identification, in addition to the main entry authentication method. This method will guarantee that the device is being accessed by the approved registered user.

The aim of this work is to identify the current user who requests access to the home network in order to test if this request is coming from the right user on a registered end-device. In this paper, we propose a remote implicit continuous user identification for accessing smart home networks. The proposed is based on the behavior pattern on the smartphone by applying classification-based machine learning. This approach tracks the users' access patterns (history) on their mobile devices and utilizes this behavior to identify the current user at access request and during access to the home networks. To this end, the contribution of this paper is a multi-user continuous identification approach for smart home networks. This approach:

- Is able to identify registered users utilizing app interactions on smartphones with a considerable high F-measure;
- Uses implicit features that can be collected (or generated) in the background without requiring user intervention in the identification process;
- Ensures that the utilized features are generalized and so can be extracted from most mobile devices regardless of the device operating systems on these devices or type of hardware;

The rest of this paper is organized as follows. Section II provides a related work summary. The proposed approach is presented in Section III, while Section IV presents the performance evaluation and results. A discussion is presented in Section V and Section VI concludes the paper and suggests future research directions.

II. RELATED WORK

User authentication and identification have been considered in both academia and the industry. As an example, the work in [1] introduces a user identification method utilizing users' routines, including location and text messages. The study in [2] presents a user identification approach using text messages, voice and apps and obtains an equal error rate (EER) of 7.03%. When the authors in [3] studied the inclusion of user access behavior on web browsers in order to identify users, the results achieved a false acceptance rate (FAR) and false rejection rate (FRR) of 24%. Other studies focus their research on modeling user behavior patterns by utilizing sensors while users are interacting with their mobile devices. As an example, the research in [4], which uses gait recognition and keystroke dynamics for continuous user authentication. Identifying users based on typing pattern, pressure, rotation and vibration sensors is proposed in [5].

An additional study [6] presents a user identification method, based on interaction with mobile devices using information from the sensors, magnetometer and gyroscope. The presented approach refers to the mobile's position, such as in the pocket or in the user's hands. Additionally, as presented in [7], low swipe gestures are encountered when users are interacting with their mobile phone apps. Device accelerometers have been recently used for profiling and identifying user behavior. The authors in [8] devised a method for utilizing accelerometers in television remote controls in order to identify individuals. In [9], an approach that identifies and authenticates users based on accelerometer data is proposed. The achieved accuracy in this work is 72.2%, and the used dataset was generated by repeating pre-defined activities. In [10], the authors described a method, also based on gait recognition, for determining whether the owner is using the device.

Different features, including battery level, transmitted information, ambient light and noise, are considered in [11] in order to obtain a user classification-based identification. A continuous authentication system presented in [12] combines the features movement data, location, touch screen and voice. SensifyID, a task-based authentication technique that combines location, the user's writing techniques, and sensors is presented in [13]. The user in the proposed approach is required to take specific action to prove user legitimacy. The main focus of the literature is on single modality in which the built models target single users for the used device. In other words, the focus of most of the listed studies is on the client side; from activities on the mobile device, the built profile detects illegal usage of the device from the modeled user profile. Despite the listed works in the literature, we believe that insider user identification has not received enough attention when considering the attack scenario presented later in Section III of this paper.

III. THE PROPOSED APPROACH

This section presents an identification approach that builds a user profile, based on previous access history, in order to make the right decision for subsequent access requests regarding legitimate user identification. The goal of the proposed approach is to check access patterns to apps on the smartphone and determines if the current user is the one registered with this mobile device. The approach presented in this paper, as shown in Fig. 1, is independent of the main authentication method utilized at the entry-point (e.g. a PIN or password) and will be used at and beyond the login stage on the mobile device.

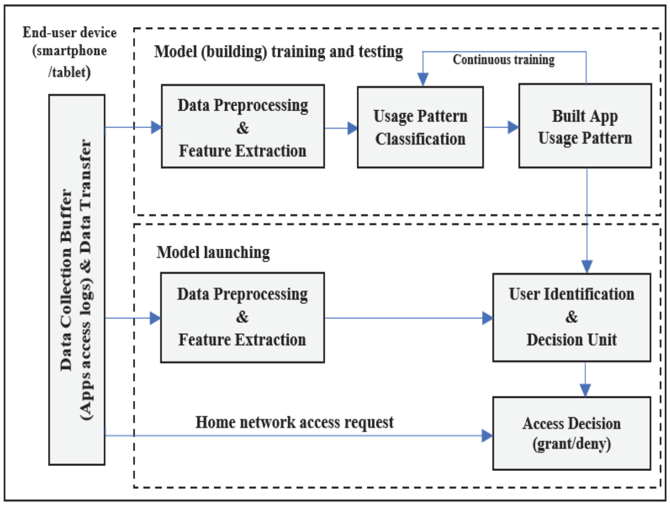
End-user device (smartphone /tablet)
Model (building) training and testing
Continuous training
Data Preprocessing & Feature Extraction
Usage Pattern Classification
Built App Usage Pattern
Data Collection Buffer (Apps access logs) & Data Transfer
Model launching
Data Preprocessing & Feature Extraction
User Identification & Decision Unit
Home network access request
Access Decision (grant/deny)

Fig. 1. Architecture of the proposed user identification method

In addition, this approach can be integrated as a second layer of user identification in an implemented framework in [14]. In this framework, a central controller hub, namely a smart home hub that functions as the network controller, is responsible for user registration, authentication, and identification as well as features collection and generation. The homeowner can add, remove and update access to users, other home members, and visitors through an application installed on the used smartphone. In addition, pairing and adding new devices to the network can be locally configured through, for example, a wired channel or via WiFi.

A. Threat Model

Accessing the smart home network and controlling appliances is mainly achieved through registered smartphones by known users. However, access permissions can be given to other users, such as visitors, friends, and relatives who will be able to access the home network using their smartphones.

Consequently, there are two main security points where unauthorized access to the home network could occur:

- Access through a registered authorized user's device by another authorized user (insider) with unauthorized access permission for the latter to a part of the home services.
- Access through a registered authorized user's device by unauthorized users (outsiders) who have no access permission to the home network. Achieving access will cause a dramatic security and privacy concern to the home network.

Unauthorized access to smart home devices could result in: accessing private information; turning on the heating, causing power consumption; or shutting off home security cameras. Each of these scenarios could cause physical, privacy-related or financial damage.

B. Workflow of the Proposed Approach

The framework works by, after the registration step, first collecting user access logs and training user behavior during access sessions to apps on mobile devices, then identifying users based on the built behavior. The following steps show the workflow of the proposed method:

- *Data Collection Buffer and Transfer*

This unit collects app access events whenever the user interacts with foreground apps on the smartphone. The access logs will then be continuously sent to the smart home hub which, in turn, pre-processes the received information and stores it in an anonymous form for training. After building the model, the access logs will not be sent directly but when reaching a specific number of logs in a group, or when there is an access request to the home network.

- *Data Preprocessing and Feature Extraction*

The features included in the proposed approach comprise a usage session and interaction time. Even though there are different definitions in the literature regarding these two terms, including work [15], we define the usage session as that in which the user is accessing the app without interruption, and the interaction time as the total time spent in accessing the session. The study in [16] shows that the authentication accuracy is subject to the day of the week. In addition, authors conclude that the categorization of the weekends should be given more weight in feature selection as some apps are mostly accessed during weekends. Hence, the feature extraction step is based on generating implicit features from the received logs including, for example, access time, day of the week and the year, app name and type, type of access event, and size of the generated data during access sessions. The generated features will then be stored in a raw form in the database for training and testing processes. The number of required usage sessions mainly depends on the user's interaction, which can be determined in a continuous manner during model training and testing.

- *Classification Strategy*

An appropriate classifier will be applied to app access events, with the prepared features from the previous step included in the model training stage. To make sure that the presented model is not classification algorithm specific, three classification algorithms are used in the training. The selected classifiers in this research, which are mostly used in the literature, such as in [17], include three different classification methodologies. The first classifier is the random forest (RF) classifier, which fits a number of decision tree classifiers on various subsamples of instances and utilizes the average in order to improve accuracy and eliminate over-fitting. The second classifier is the gradient boosting classifier (GBC) that provides several hyperparameter tuning options that provide the function with a very flexible fit. However, its computation cost is high, especially when the data size is large, in addition to the number of generated trees which may exceed 1000.

The third classifier used in our evaluation is the k-neighbors classifier (KNN), which applies the k-nearest neighbors' vote. The KNN classifier identifies new data entry based on historical data and labels them with the majority class according to the nearest neighbor. Although the fact that the KNN classifier is easy to implement, the training data has to be saved at the classification time. For the usage scenario, as we are targeting multi-user identification, the multi-class classification will be selected, with the result that each access event will be classified as related to one of the enrolled known users. For training the model, increasing training and testing methodology on an incremental usage basis will be adopted, in which training the model will be applied within a specific time interval and testing the model will be applied on the unseen data.

- *Decision Unit*

In the proposed approach, access logs, which are translated events with extracted appropriate features, are classified as either for an enrolled user or not. This paper focuses solely on user identification, assuming that all users are known and previously enrolled in the smart home network. In this unit, the decision (d_i) will be made based on the last two accessed events, $(a_{l-1} and\ a_l)$. Consequently, when the last two classified events are identified for a specific user (u_i), the next access request will be accepted. In other words, if the last accessed events are identified to the current user, the next request will be accepted, otherwise it will be denied, and the user will be requested to undergo a second-factor authentication in order to prove identity.

$$\mathrm{d_i} = \begin{cases} \text{if } (\mathrm{a}_{\mathrm{l}-1} \text{and } \mathrm{a_l}) \in \mathrm{u_i}, \text{ Permit access} \\ \text{if } (\mathrm{a}_{\mathrm{l}-1} \in \mathrm{u_i}) \text{ and } (\mathrm{a_l} \notin \mathrm{u_i}), \text{ Deny access} \\ \text{if } (\mathrm{a}_{\mathrm{l}-1} \notin \mathrm{u_i}) \text{ and } (\mathrm{a_l} \in \mathrm{u_i}), \text{ Deny access} \end{cases} \quad (1)$$

IV. EVALUATION AND RESULTS

To evaluate the performance of the presented approach, the dataset UbiqLog4UCI [18] collected from real users is utilized, and the identification performance is considered as the accuracy metric when classifying an access session to one of the enrolled users. Information of 20 users was used because these users access most apps in relatively the same period for three months. The first experiment is performed on the access session lengths based on different ranges of access duration that users spend interacting with apps. As an example, selected session access times, ranging from 5 seconds (sec) up to 60 minutes (min), are shown in Fig. 2 in the logarithmic scale. We can see that, from the same figure, most of the access sessions fall in the range of

5 sec to 10 min. In the second evaluation, we examine the model based on different access session lengths.

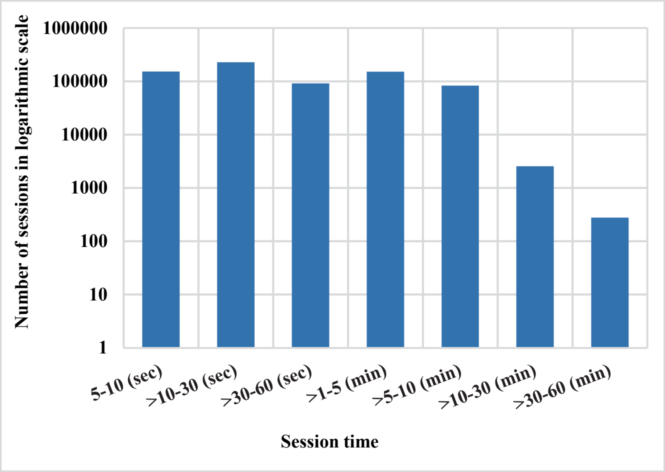

Fig. 2. Number of accessed sessions per time interval

As shown in Fig. 3, the presented results indicate that the accuracy remains low when access sessions ranged between 5 sec and 5 min. However, as the access session time increases, accuracy increases, and the best accuracy is achieved when the access sessions exceed 5 min. The alternative evaluation experiment is user identification in a short time period. Thus, we examine the model performance based on 24 hours, to test the capability of the presented method in identifying legitimate users. Therefore, we consider the hourly access time.

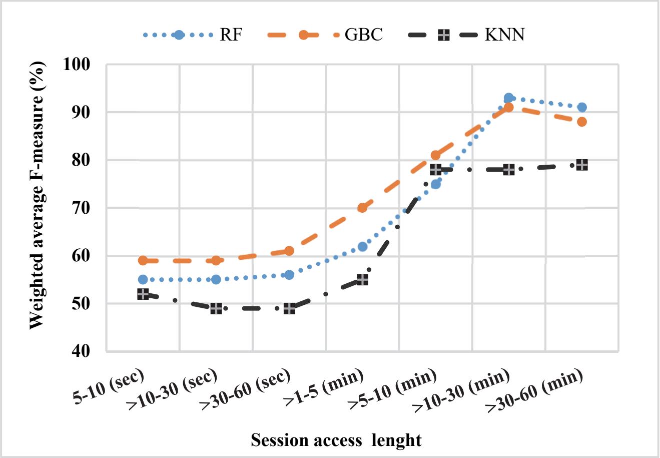

Fig. 3. Performance evaluation based on access session length

From the results, which are shown in Fig. 4, there is a variance in classification performance due to the change variation in app starting time. Hence, to overcome this issue, the considered time range should be increased. The next evaluation is based on the number of users involved and the weighted average of the F-measure, as shown in Fig. 5. The evaluation is achieved on unseen sessions. It can be observed from the results that there is little change in the weighted average F-measure when increasing the number of users, and that it ranges from 90% to 100% when considering the GBC classifier. Furthermore, the GBC and RF present the highest accuracy compared with the KNN.

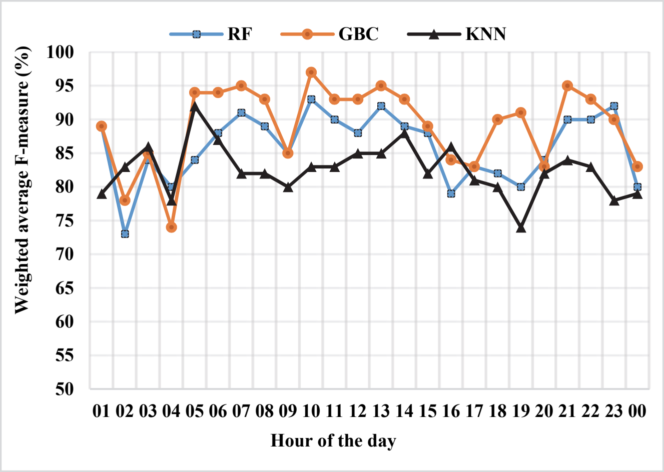

Fig. 4. Performance evaluation based on the hour of day

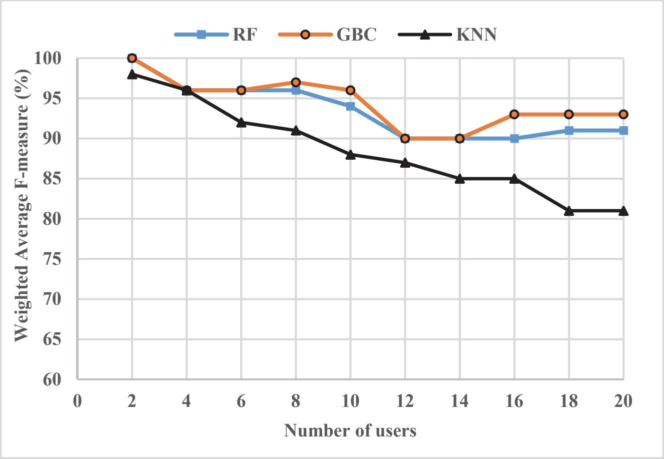

Fig. 5. Performance evaluation based on the number of enrolled users

Accordingly, for authorizing new access based on the classification of the previous (last) access events, the proposed approach performs a minimum of 90% of access decisions when considering the access sessions that exceed 5 min.

A. Limitations

Although the presented results are promising, there are some limitations in this work. As an example, neither newly launched apps nor the sequence order of app access are considered in this study. Additionally, we discard apps used for a limited number of times to avoid their impact on accuracy. However, user access patterns may change over time and this change, such as using new apps or stopping the use of others, should be considered. Furthermore, we consider only the identification aspect as well as assume that all users are registered and known. However, authentication is also an important issue that has to be considered in order to detect unauthorized access from unknown users (outsiders). Although it is difficult to build a model for unknown user behavior, this issue will be the focus of future work. Additionally, a study in [19] considers traffic generated app access sessions and concludes that users can be authenticated based on this traffic. However, traffic generated information is not considered in this work due to the nature of the utilized dataset, and we believe the availability of this feature will considerably improve the accuracy of the identification.

V. DISCUSSION

From the results, it can be seen that it is possible to obtain user identification based on access sessions of apps on the same device and utilizing them to identify the current user of the device. In this work, we consider all used apps that have been used for at least one week in a period of three months and ignore the remainder, the newly launched apps, during the model training and testing. By this step, we discard apps that are used a limited number of times to avoid their effect on accuracy. To avoid any impact on accuracy, app usage change can be solved by including all launched apps in addition to discarding those apps not in use. However, this solution needs a dynamic adaptation of app usage change leading to continuous model training and testing. We partially avoid this issue, after building the model, by including only the last two access sessions for decision making. Moreover, as seen in Fig. 5, the performance of the presented approach consistently remains at 90 % when considering GBC and RF classifiers. However, as the number of users increases, the performance slightly decreases. This decrease in performance is mainly because there is a close similarity in the users' access sessions. As usage sessions may increase or decrease, and similarity among users may be present, other features, such as access order, need to be extracted, in addition to ignoring apps that are not being used. The new extracted features should be implicitly generated from user usage patterns and should not require specific action from the user.

VI. CONCLUSION AND FUTURE WORK

This paper presents a multi-user identification approach that continuously identifies users based on their interaction with apps on smartphones. The presented method employs implicit features that do not require extra user action to be generated and applied in order to offer multi-modal user identification for smart home networks. This approach is able to provide continuous user identification by tracking user access sessions without interpreting user current actions. The achieved results from the evaluation are promising; however, some limitations to this work should be considered. Authentication, which is an important aspect, especially in detecting unauthorized access from unknown users, is not considered in this research. In addition, in order to improve the scalability of the presented approach, the number of users, as well as the length of the interaction session, should be larger. These aspects, will be considered in future work.

REFERENCES

[1] E. Shi, Y. Niu, M. Jakobsson, and R. Chow, "Implicit Authentication Through Learning User Behavior," Springer, Berlin, Heidelberg., pp. 99–113, 2011.

[2] F. Li, N. Clarke, M. Papadaki, and P. Dowland, "Misuse Detection for Mobile Devices using Behaviour Profiling," in International Journal of Cyber Warfare and Terrorism (IJCWT), 2011, pp. 41–53.

[3] M. Abramson and D. W. Aha, "User Authentication from Web Browsing Behavior," in The Twenty-Sixth International FLAIRS Conference, 2013, pp. 268–273.

[4] H. Saevanee, N. Clarke, S. Furnell, and V. Biscione, "Text-Based Active Authentication for Mobile Devices," in International Federation for Information Processing IFIP, Springer, Berlin, Heidelberg, 2014, pp. 99–112.

[5] C. Bo, L. Zhang, and X.-Y. Li, "SilentSense: Silent User Identification via Dynamics of Touch and Movement Behavioral Biometrics," in 19th Annual International Conference on Mobile Computing and Networking (MobiCom), 2013, pp. 187–190.

[6] M. Ehatisham-ul-Haq et al., "Authentication of Smartphone Users Based on Activity Recognition and Mobile Sensing," Sensors (Switzerland), vol. 17, no. 9: 2043, 2017.

[7] V. Sharma and R. Enbody, "User Authentication and Identification From user Interface Interactions on Touch-Enabled Devices," IEEE transactions on information forensics and security, vol. 8, no. 1, pp. 136–148, 2013.

[8] K. Chang, J. Hightower, and B. Kveton, "Inferring Identity Using Accelerometers in Television Remote Controls," In International Conference on Pervasive Computing, pp. 151–167, 2009.

[9] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Cell Phone-Based Biometric Identification," 4th International Conference on Biometrics: Theory, Applications and Systems, BTAS , IEEE, pp. 1–7, 2010.

[10] D. Gafurov, K. Helkala, and T. Søndrol, "Biometric Gait Authentication Using Accelerometer Sensor," Journal of Computers, vol. 1, no. 7, pp. 51–59, 2006.

[11] D. Fuentes, J. Maria, L. Gonzalez-Manzano, and A. Ribagorda, "Secure and Usable User-in-a-Context Continuous Authentication in Smartphones Leveraging Non-Assisted Sensors," Sensors (Switzerland), vol. 18, no. 4, p. 1219, 2018.

[12] W. Shi, J. Yang, Y. Jiang, F. Yang, and Y. Xiong, "SenGuard: Passive User Identification on Smartphones Using Multiple Sensors," Proceedings of the 7th International Conference on Wireless and Mobile Computing, Networking and Communications, pp. 141–148, 2011.

[13] V. Maojo, F. Martín, and J. M. Vázquez-Naya, Inteligencia Artifical y ComputacióN Avanzada. 2015.

[14] Y. Ashibani, D. Kauling, and Q. H. Mahmoud, "Design and Implementation of a Contextual-Based Continuous Authentication Framework for Smart Homes," Applied System Innovation, vol. 2, no. 1, pp. 1–20, 2019.

[15] D. Hintze, S. Scholz, R. D. Findling, R. Mayrhofer, and M. Muaaz, "Diversity in Locked and Unlocked Mobile Device Usage," in ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, 2014, pp. 379–384.

[16] Y. Ashibani and Q. H. Mahmoud, "A Machine Learning-Based User Authentication Model Using Mobile App Data," in International Conference on Intelligent and Fuzzy Systems (INFUS), 2019, pp. 408–415.

[17] H. Cao and M. Lin, "Mining Smartphone Data for app Usage Prediction and Recommendations: A Survey," Pervasive and Mobile Computing, vol. 37, pp. 1–22, 2017.

[18] R. Rawassizadeh, E. Momeni, C. Dobbins, and P. Mirza-babaei, "Lesson Learned from Collecting Quantified Self Information via Mobile and Wearable Devices," Journal of Sensor and Actuator Networks, vol. 4, no. 4, pp. 315–335, 2015.

[19] Y. Ashibani and Q. H. Mahmoud, "A User Authentication Model for IoT Networks Based on App Traffic Patterns," in 9th Annual IEEE Information Technology; Electronics and Mobile Communication Conference (IEEE IEMCON), 2018, pp. 632–638.