IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Echo-ID: Smart User Identification Leveraging Inaudible Sound Signals

Syed W. Shah[1,2], Arash Shaghaghi[2,1], Salil S. Kanhere[1], Jin Zhang[3], Adnan Anwar[2], and Robin Doss[2]

[1]School of Computer Science and Engineering, The University of New South Wales, Sydney, Australia
[2]Deakin University, Geelong, Australia - Centre for Cyber Security Research and Innovation (CSRI)
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.

Corresponding author: Syed W. Shah (e-mail: z5038389@zmail.unsw.edu.au).

ABSTRACT In this paper, we present a novel user identification mechanism for smart spaces called Echo-ID (referred to as E-ID). Our solution relies on inaudible sound signals for capturing the user's behavioral tapping/typing characteristics while s/he types the PIN on a PIN-PAD, and uses them to identify the corresponding user from a set of *N* enrolled inhabitants. E-ID proposes an all-inclusive pipeline that generates and transmits appropriate sound signals, and extracts a user-specific imprint from the recorded signals (E-Sign). For accurate identification of the corresponding user given an E-Sign sample, E-ID makes use of deep-learning (i.e., CNN for feature extraction) and SVM classifier (for making the identification decision). We implemented a proof of the concept of E-ID by leveraging the commodity speaker and microphone. Our evaluations revealed that E-ID can identify the users with an average accuracy of 93% to 78% from an enrolled group of 2-5 subjects, respectively.

INDEX TERMS Smart-Spaces, User Identification, Sound-Signals.

I. INTRODUCTION

The recent evolution of pervasive computing technologies have brought to fruition the concept of smart spaces which aim at seamless provisioning of customized services to their inhabitants. For example, a shared smart office may identify a particular occupant and automatically turn on his computer (and other devices) and adjust the temperature and light settings of his cubicle as per his preferences. Similarly, it may also restrict the entry to a designated place (e.g., record or server rooms) only to a few individuals, and may in addition keep a record of the authorized person(s) who accessed that particular area. Likewise, a smart home may not allow the vulnerable inhabitants (e.g., children and elderly people) to operate risky appliances (e.g., oven). It may also restrict the content viewed on the TV or the Internet for some individuals (e.g., children). For all of the aforementioned operations of smart environments, it is essential to establish the identity of person(s) currently using the space. RFID swipe cards (i.e., possession-factor relying upon something user has) are widely used in smart spaces for authenticating the inhabitants. However, the requirement of carrying a physical card is onerous for the users. In contrast, PIN-PADs (i.e., knowledge-factor relying upon something user knows) - which are ubiquitously used for authentication in smart-spaces, do not require the user to carry a dedicated element (e.g., swipe card). However, it is not possible to determine the identity of the corresponding subject who has entered the correct PIN (since the same PIN is generally used by all authorized individuals). Physical biometrics (i.e., inherence-factor relying upon something user is) such as fingerprints, face-images, and iris are increasingly being adopted for human identification in smart spaces. However, they are shown to be vulnerable to subversion. For example, fingerprints may be collected from a surface which the victim may have touched and used to circumvent the fingerprint based authentication [1]. Likewise, facial-recognition may also be spoofed by using the victim's facial photograph (which is easy to find on Internet) or a 3-D printed head [2] [3]. Similarly, iris based authentication may also be breached by using the victim's photograph superimposed with the contact lens [4] [5] [1]. Camera-based gait recognition for human identification has associated privacy concerns [6]. Other mechanisms either require specialized floor-embedded sensors [7], or wearables (e.g., smartwatch) for identifying the occupants [8]. Both of these approaches are considered onerous by the users. A few works such as [9], [10] have utilized the pervasive WiFi signals to non-intrusively capture the user's gait pattern (i.e., behavioral biometrics) for human identification in smart

spaces. However, these works require the user to walk along a straight pre-defined path to perform the identification, which may not be feasible for constrained smart spaces (i.e., where such paths are not available). Likewise, authors in [11] also leveraged WiFi signals to capture the user's cardiopulmonary activity and demonstrated its appropriateness for human identification in small smart spaces. Although this approach does not demand any explicit activity from the user (unlike [9], [10] which necessitates to walk), it requires the user to stand (or sit) still in front of a commodity WiFi device for a duration of at least 20 seconds to perform the identification, which presents usability challenges.

In this paper, we present a novel non-intrusive human identification system called Echo-ID (referred to as E-ID hereon) for small smart spaces that does not demand any explicit effort from the user. E-ID endeavours to re-purpose in -situ PIN-PADs for human identification in smart spaces. PIN-PADs are pervasive in smart-environments and are generally used to restrict entry to a shared smart space (e.g., office or home), and allow access only to the authorized users who know the PIN (usually a 4 - 8 digit code). To identify an individual from an authorized list (i.e., enrolled-set), we aim to utilize the user's tapping/typing behaviour (i.e., the way user moves his fingers and hand) while entering the PIN on a PIN-PAD deployed in smart space. There is a strong evidence that different users have unique habitual tapping (or, typing) pattern while entering the PIN which has been used for user authentication on smartphones [12]. However, smartphones are integrated with a plethora of sensors (e.g., accelerometer, gyroscope, and touch sensors, etc) which makes it possible to capture the user's unique tapping/typing behaviour while entering the PIN. In contrast, the PIN-PADs are generally not equipped with the aforementioned sensors which thus renders the above methods to be not relevant. In view of this, we propose to use sound signals to capture the way user types/taps and moves his fingers and hand while entering the PIN. The PIN-PADs are generally equipped with a speaker by default to facilitate the PIN entry process (e.g., to let user know whether the entered PIN is correct or not), which can be used to transmit a sound signal while user is entering the PIN.

To make E-ID completely invisible to the user, we make use of the inaudible frequencies as an audible sound would be annoying. The commodity speakers (e.g., on PIN-PADs, mobile phones, laptops, etc) can support a sampling frequency in excess of 48KHz, and hence are capable of generating and transmitting inaudible sound frequencies (i.e., 18KHz-22KHz). When the user types the PIN, the inaudible sound frequencies transmitted in parallel through the speaker get reflected from the user's moving fingers and hand, which in turn leave a unique user-specific imprint in the sound-echos recorded by a commodity microphone (which is easy to interface with the existing PIN-PADs). We refer to such an imprint as Echo-Signature (referred to as E-Sign hereafter) and use it to identify the corresponding user from a set of N enrolled inhabitants. To the best of our knowledge, E-ID is the first work that utilizes the inaudible echos from the user's hand while entering the PIN to establish his identity.

Figure 1 depicts the typical usage scenario of the E-ID. We assume that a shared smart space has N authorized users (e.g., employees in a smart office) who know the PIN (e.g., 1234) to access the space. To identify the corresponding user from a correct PIN entry (i.e., 1234), E-ID triggers the PIN-PAD to simultaneously transmit and record the inaudible sound while user types (or taps) the PIN and extract the E-Sign from the recorded sound. E-ID utilizes the extracted E-Sign to identify the corresponding user by comparing it with the E-Sign samples of authorized users for whom the enrollment samples are collected a priori. Once the corresponding user is identified, the smart space may enable a number of customized services for the identified user (e.g., automatically turn-on the user's computer and adjust temperature and light settings as per the likings of the identified subject). Note that, the E-ID is completely transparent to the user (i.e., user only needs to type/tap the PIN as usual) and does not demand any particular action unlike the prior works of this nature [9]–[11].

The problem of identifying a subject from a large set of enrolled users using an E-Sign sample is arguably complex. To make the problem tractable, we consider a simple scenario of small smart spaces, i.e., where the number of inhabitants is limited to a maximum of 5-6 users. Although the evaluated group-sizes may apparently seem small, they are the representations of an average OECD household (average occupants of 3) and micro-enterprises (with 5-6 occupants) [13], [14]. Even with a smaller group-size, there are numerous challenges involved in realizing a system such as E-ID. The *first technical challenge* is to generate an appropriate signal that can be used to record the echos that may be reflected from the user's moving fingers and hand while entering the PIN. To this end, we generate and transmit an inaudible chirp signal with an appropriate silence period so as to record the required echos (Section II-A). The *second technical challenge* is to track the echos that may be changing due to the user's moving fingers and hand while entering the PIN. These changing echos represent the user's habitual tapping/typing pattern, which generate the E-Sign of the user. To achieve this, we devise a threshold-based mechanism to align the transmitted and recorded signals and compute the cross-correlation between them (Section II-B). We form an E-Sign by leveraging these correlation values and transforming them into a matrix by doing an appropriate computation (Section II-C). The *third technical challenge* is to identify the corresponding individual from an enrolled-set by leveraging the E-Sign samples. For this purpose, we make use of deep-learning. Specifically, we utilize the Convolutional Neural Network (CNN) for extracting the discriminating features from the E-Sign samples. We then feed these features to a multi-class SVM classifier which identifies the corresponding subject given an E-Sign sample (Section II-D). Since conventional PIN-PADs are generally not equipped with microphone(s), we implemented a proof-of-concept of E-ID

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031899, IEEE Access

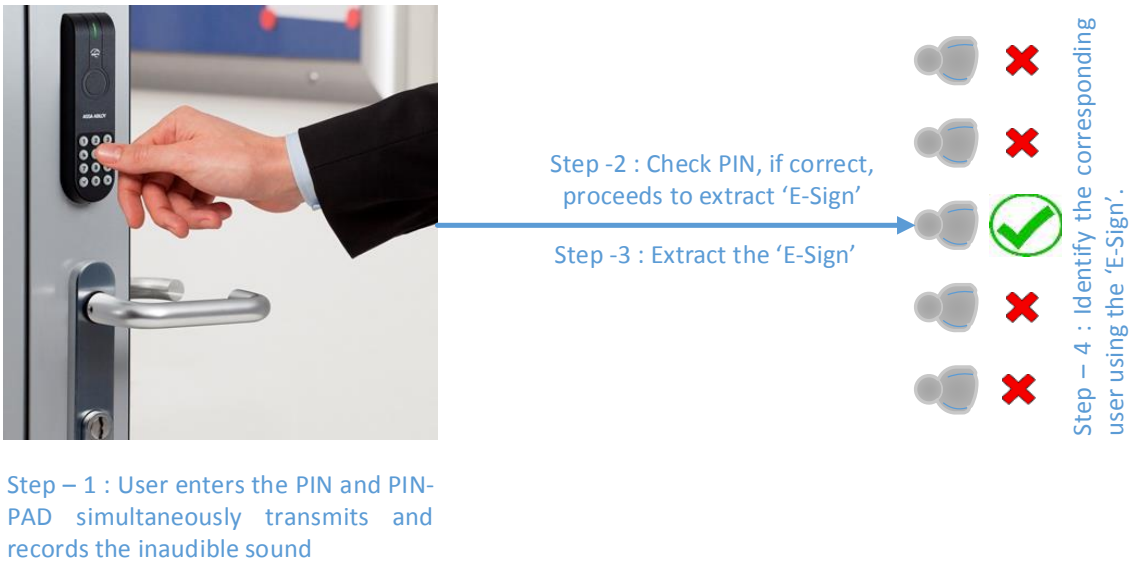

FIGURE 1: Echo-ID Usage Scenario

by leveraging the commodity speaker (i.e., built-in speaker of a laptop) and microphone (i.e., interfaced to laptop via a USB cable), while the users type the PIN on a smartphone placed in the vicinity of the microphones (See Section III-A for the details of set-up). Our extensive evaluations reveal that, E-ID can identify the users with an average accuracy ranging from 93.7% to 78.2% for a group-size of 2-5 individuals, respectively.

The main contributions of this article are as follows:

- We propose an user identification (*E-ID*) for small smart spaces that leverages the inaudible sound signals for capturing the user's behavioral tapping/typing behaviour while entering the PIN on a conventional PIN-PAD.
- We propose an all-inclusive pipeline that generates and transmits appropriate sound signals, and segregates *echo-signatures* representing the user's habitual tapping/typing behaviour.
- We present a deep-learning model that can leverage the *echo-signatures* to identify the corresponding subject from an enrolled-set of users.
- We implement a proof-of-concept of *E-ID* by leveraging the commodity speaker and off-the-shelf microphone, and demonstrate that *E-ID* is capable of identifying users with an accuracy of 93% to 78% for group-sizes of 2-5 subjects, respectively.

The rest of the paper is organized as follows. Section II presents the details of E-ID. Section III describes the evaluation methodology along with the results. Section IV details the related work, and finally the concluding remarks appear in Section V.

II. ECHO-ID IDENTIFICATION WORK FLOW

In this section, we present the details of the proposed Echo-ID system. Note that, we will only discuss the details related to the generation of E-Sign and proposed deep-learning model for identify the corresponding user, as the verification of PIN(s) is a standard. Since the same PIN is generally used for accessing a restricted space, we do not present a mechanism for detecting the unseen (i.e., not enrolled) users. This will be accomplished through the usage of PIN (i.e., only the authorized subjects will know the PIN). Figure 2 shows the different steps involved in E-ID. We will discuss each of these steps in detail in the subsequent sections.

A. SOUND SIGNAL GENERATION

Since E-ID is based upon tracking the echos that change due to the movements of user's fingers and hand while entering the PIN, the first step is the generation of an appropriate sound signal that can serve our purpose. Given that we want E-ID to be completely transparent to the user, the sound signal that we transmit from the speaker should be in inaudible frequency (i.e., beyond 18KHz). Although the audible sound signals are conveniently transmitted by the commodity speakers, they can be annoying for the users in the anticipated usage scenario. The commodity speakers and microphones support a sampling frequency of up to 48KHz. Therefore, theoretically (in accordance with Nyquist theorem [15]) they can transmit and record signals of up to 24KHz which falls within the inaudible range. With this possibility, we use a commodity speaker to generate a chirp signal of frequency $18 - 22$KHz of duration 1ms (48 samples at Fs = 48KHz, i.e., 1ms x Fs). To record the echos that reflect from the user's moving fingers and hand, we add a silence period of 4.3ms ($\approx$208 samples at Fs = 48KHz, i.e., 4.3ms x Fs). Note that, we

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031899, IEEE Access

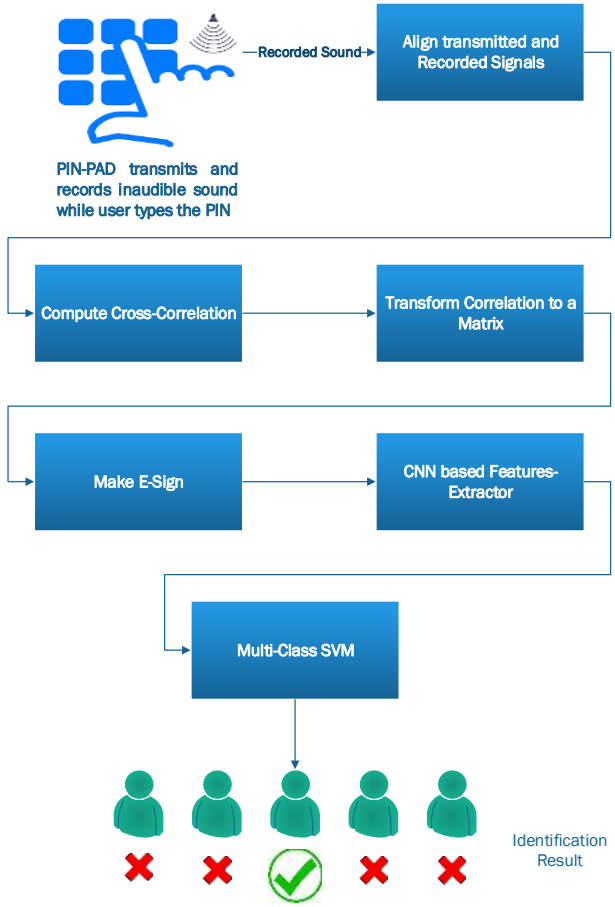

FIGURE 2: Echo-ID Work-Flow for Identification

have empirically tested different values for the duration of the chirp and silence signals, and selected the values that can best capture the required echos. Consequently E-ID transmits a signal of duration 5.3ms (i.e., 256 samples) from the speaker. Figure 3 shows the signal that we transmit to record the E-Sign of the user. This signal is transmitted for a total time duration of 3 seconds (i.e., which corresponds to the typical time required for PIN entry) at an approximate frame rate of 188 (i.e., 1/5.3ms), so as to record the echos due to the fingers and hand movement during the entire PIN entry. Figure 4 shows a chunk of the transmitted signal, while Figure 5 shows the spectrogram which depicts that the frequencies of this signal lie in the inaudible range.

B. PRE-PROCESSING OF THE RECORDED SOUND

In order to generate the E-Sign, E-ID involves a number of pre-processing steps which are discussed below.

1) Filtering the low-frequency noise

Figure 6 shows the sound signal recorded using a commodity microphone. The corresponding spectrogram of this recording is depicted in Figure 7. It is evident from the spectrogram that the recorded signal captures the signal of interest -i.e., high energy is visible in the $18 - 22$KHz band. However, the recording also demonstrates the strong presence of low-frequencies that may correspond to other audible signals in the surroundings (e.g., people talking). Such noise components are conspicuous in the spectrogram. In order to generate the E-Sign of a user from the transmitted signal, E-ID uses a band-pass filter with stop-band frequencies of 17.5KHz and 22.5KHz, so as to alleviate the impact of unwanted frequencies that fall outside the desired range. Figure 8 shows a chunk of filtered recording, while Figure 9 shows the corresponding spectrogram. It is evident from the spectrogram that the low-frequency signals are successfully eliminated, while the filtered signal prominently contains the frequencies of our interest. This also depicts that the E-ID may not be impacted by audible sounds such as music or conversations. Since the commodity speaker and microphone are not perfectly synchronized, there is always a delay before the microphone starts recording the transmitted chirp signal. This is conspicuous in a filtered recording as shown in Figure 8. In order to generate an E-Sign from the recording, it is important to eliminate this delay and align the recorded signal with the transmitted signal - i.e., each chunk of 256 samples of the recorded signal must contain the chirp signal arriving

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031899, IEEE Access

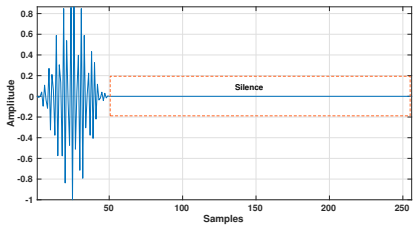

FIGURE 3: Generated Signal with Silence period

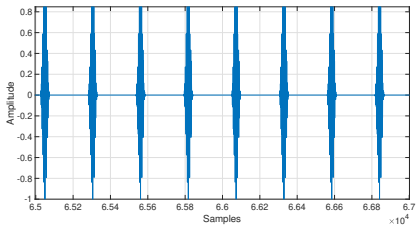

FIGURE 4: A Chunk of transmitted Signal

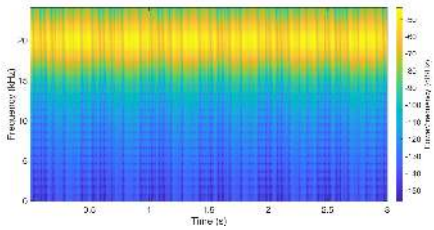

FIGURE 5: Spectrogram of the transmitted Signal

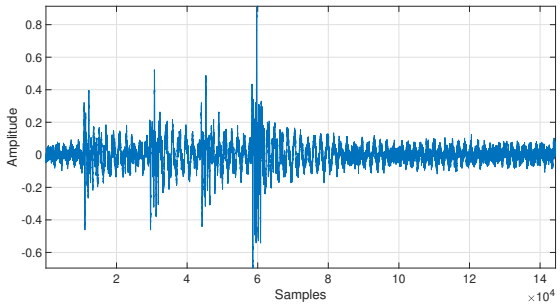

FIGURE 6: Recorded Signal

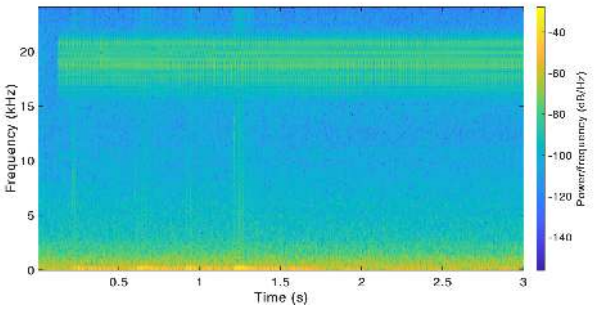

FIGURE 7: Corresponding Spectrogram

at the microphone from the direct-path at the start followed by a section that records echos (i.e., silence portion). Next, we develop a strategy to align these signals, which helps in generation of the E-Sign.

2) Alignment of Transmitted and Recorded Signals

To eliminate the aforementioned delay in the recorded signal, we identify the prominent local peaks in the signal. Figure 10 shows the peak values plotted against the number of peaks detected from the recording shown in Figure 8. For eliminating the delay in the recording, we identify the first peak that crosses a threshold value. We analytically set this threshold to half the mean peak value which helps in successful elimination of the aforementioned delay. Once the first peak that corresponds to the first prominent direct-path sound signal received by the microphone is identified, we set the start of the recording to the sample number located at length(chirp)/2 samples behind the identified peak. Our reasoning is that, the identified peak corresponds to the centre of the chirp signal that arrives at the microphone through the direct-path, and hence the actual start of this signal will be located halfway behind. Figure 11 shows the resultant signal after the elimination of delay from the recorded signal shown in Figure 8. It is evident that the delay in the recorded signal is successfully removed. In addition to removing the delay, we also normalize the recorded signal to [-1 1], since the transmitted high-frequency signals attenuate quickly, and the signal received by the microphone has much lower amplitude than the transmitted chirp as can be visualized in Figure 8. This also helps in identifying the echos that change due to the user's moving fingers and hand. The resultant recording after the elimination of delay is aligned with the transmitted signal - i.e., the first 256 samples of the resultant recording (and all other succeeding chunks of 256 samples) correspond to the transmitted chirp signal received from the direct-path along with the echos that reflect from the user's moving fingers and hand during the silence period. Figure 12 shows the first 256 samples of the recorded signal along with the transmitted signal (i.e., chirp + silence). It is evident that, these two signals are aligned approximately. In addition, the echos are also conspicuous in the recording during the silence period of the transmitted signal. We utilize these echos to generate the E-Sign to identify the corresponding individual from the enrolled-set.

3) Computing Cross-Correlation

Once the signals are aligned in accordance with the method detailed in the previous sub-section, the next step is to identify the reflections (or echos) that occur due to finger and hand motion of the user. In a scenario when there is no finger or hand movement, the static reflections (i.e., due to the static objects in the vicinity) will appear same in the consecutive recorded pulses (i.e., each of 256 samples). This is because when there is no movement all the echos in consecutive pulses will take same time to reach the microphone after the reflections. For example, Figure 13 shows two superimposed recorded pulses which are almost identical, showing that there was no movement (of fingers or hand) when these

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031899, IEEE Access

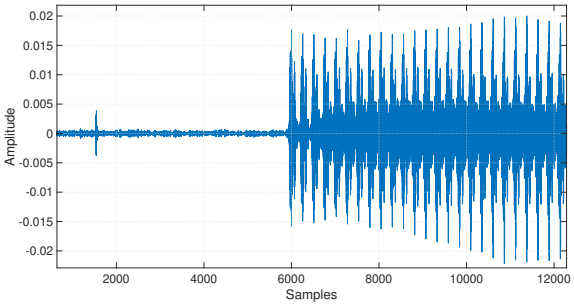

FIGURE 8: Filtered Signal

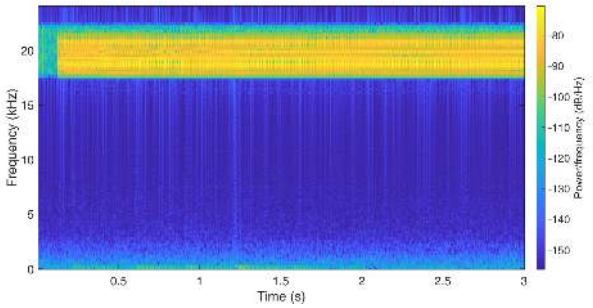

FIGURE 9: Corresponding Spectrogram

pulses were transmitted by the speaker and recorded by the microphone. In contrast, Figure 14 show two other recorded pulses where the variations that may correspond to the user's fingers and hand movement are visible (see the zoomed-in section). This is because when user moves his finger(s) and hand to enter the PIN, the echos that get reflected from the fingers and hand arrive at the microphone at a different time than the previous pulse. We are interested in extracting these variations only, i.e., changes that correspond to the user's moving fingers and hand while entering the PIN. To track these changes, we compute cross-correlation between the transmitted chirp signal and all consecutive pulses of the recorded signal (i.e., chunk of 256 samples). When there is no movement, the computed correlation values of the consecutive recorded pulses will be identical, while in case of movement these values will change. This will help in identifying the locations in every pulse that correspond to the user's fingers and hand movement. Since we have already aligned the recorded sound with the transmitted signal (see sub-section II-B), we only retain the correlation values from lag zero onward. We also normalize the computed correlation values to [-1 1] so that changes corresponding to the movements become visible as compared to other sections of the pulse.

C. GENERATION OF ECHO-SIGNATURE

We utilize the computed correlation values to generate the Echo-Signature (i.e., E-Sign) of the user, and feed this to deep-learning model (see Section II-D for details of the model) to identify the corresponding user from the enrolled-set. This section entails the steps involved in generation of E-Sign.

1) Transforming Correlation Values to Matrix

To make an E-Sign, we first transform the computed correlation values into a matrix. For this purpose, we form a matrix of order 256×564, where 256 corresponds to correlation values of each pulse of the recorded signal, while 564 represent the total frames that were transmitted (i.e., Frame Rate (188) x Time (3 sec) = 564). We append the correlation values of the consecutive recorded pulses (i.e., 256 samples) in successive columns of this matrix. Note that, since we eliminate a certain portion of the recorded signal (i.e., delay - see sub-section II-B), we may not end up having exact 564 chunks (of 256 samples) of the recorded signal. Hence, the last few values (or columns) in the matrix are set to zero to maintain a consistent matrix size. Figure 15 shows the correlation matrix of a sample recording while the user enters the PIN.

2) E-Sign Generation

To make an E-Sign of the user, we compare the columns of the correlation matrix so as to detect only the prominent changes that correspond to the user's moving fingers and hand. To achieve this, we transform the correlation matrix to a new matrix whose i_{th} column is computed as $CM(1:256, i+3) - (CM(1:256, i)$, where CM represents the correlation matrix. The column threshold (i.e., 3) is set empirically (by changing it randomly between 1 and 100) for detecting the changes occurring due to user's fingers and hand. A similar method is also used in [16] and [17] to track the finger movements and use them for detecting $2-D$ gestures and snooping the unlock patterns of mobile devices, respectively. Figure 16 shows the newly generated matrix by comparing the columns of the correlation matrix shown in Figure 15 in accordance with the aforementioned method. We refer to this new matrix as E-Sign of the user and utilize this to identify the corresponding user from an enrolled-set. However, in order to successfully use the E-Sign for user identification, it should be consistent for the same user across his multiple attempts of entering the same PIN. Likewise, it must demonstrate distinctiveness from the E-Sign samples of the other individuals for the same PIN. To demonstrate this, we show four E-Sign samples belonging to two different subjects in Figures 17 -20 (i.e., 2 of each subject). It is conspicuous that E-Sign samples of the same subject are similar in multiple instances of entering the same PIN (e.g., see Figs 17 & 18 for subject #1, and Figs 19 & 20 for subject #2), while they are different for different subjects even if they enter the same PIN (e.g., see Figs 17 & 19 to visualize the difference in E-Signs of subject # 1 & 2). This lends credence to our idea of using E-Sign for identifying

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031899, IEEE Access

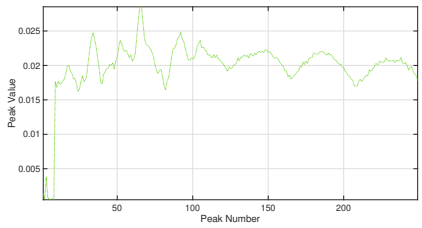

FIGURE 10: Detected Peaks

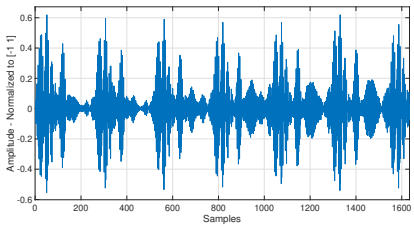

FIGURE 11: Recorded Signal with delay removed

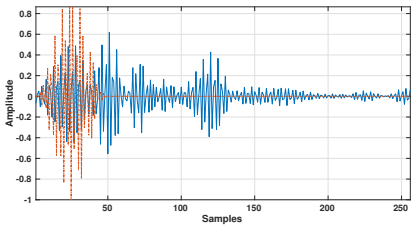

FIGURE 12: Aligned Transmitted and Recorded Signal-Superimposed

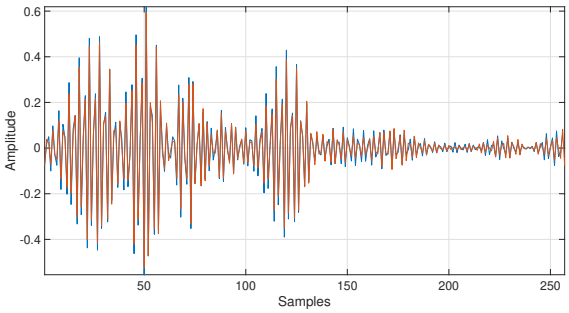

FIGURE 13: Two Consecutive Pulses of Recorded Sound - with no movement

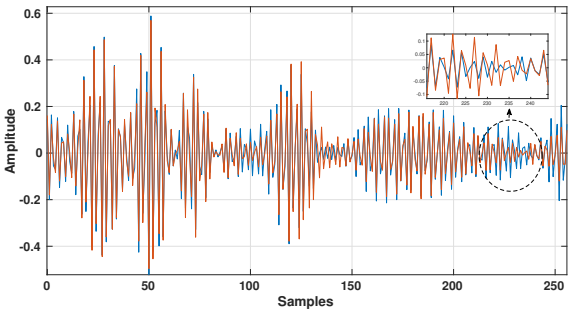

FIGURE 14: Two Pulses of Recorded Sound - with movement visible

the corresponding individual from an enrolled-set of subjects. To achieve this, we feed the E-Sign samples to our designed deep-learning model that can learn the appropriate features from the E-Sign samples and is capable of identifying the corresponding user given an E-Sign sample. In particular, we use a Convolutional Neural Network (CNN) for feature extraction. The reason behind using CNN is that, it helps in automatic extraction of the discriminating features from the E-Sign and has also shown success in prior works on human identification like [18] [19]. The deep-learning also results in superior identification performance than the conventional machine learning which often requires the manual extraction of the features [18]. Next, we describe the structure of the deep-learning model that we employed as a part of E-ID, and also present the key parameters that we used for performing the user identification.

D. DEEP-LEARNING MODEL FOR USER IDENTIFICATION

The aim of the model is to learn the representation of user's typing behaviour while entering the PIN on the PIN-PAD. Let a single E-Sign sample be represented as $E_i \in \mathbb{R}^k$, where k represents dimensions of E-Sign (i.e., 256×564). We feed the E-Sign to a CNN structure that helps in learning the spatial-features. Once the features are learnt, we feed them to a secondary SVM classifier which helps in discriminating the subjects. Figure 21 shows the employed deep-learning model. In this section, we present the details of the CNN structure.

CNN as Feature-Extractor

The CNN structure employed in our implementation for extracting the features is shown in Figure 21. The model is stacked as follows: input layer, first convolutional layer (C1), first Rectified Linear Unit (ReLU) layer, first pooling (max) layer (P1), second convolutional layer (C2), second ReLU layer, second pooling (max) layer (P2), fully-connected layer and output layer. Table 1 shows the details of the parameters used in the CNN. The dimensions of the E-Sign computed in accordance with the method described in previous subsection is [256, 564, 1]. We first pass these E-Sign samples through a set of convolution filters which scan the entire E-Sign sample and learn different local features. In C1, we empirically choose a total of 20 filters with a size of [3,3] and stride [1,1], resulting in an output of dimensions [254, 562, 20]. The output of C1 is then passed through a ReLU) which performs a thresholding operation by maintaining only the positive values. Afterwards, we feed the output of ReLU to a maximum pooling layer (i.e., P1) which performs the down-sampling and helps in prevention of over-fitting of the model [18]. We set the window-size and stride to [2, 2] in the P1, which results in an output of order [127, 281, 20]. In C2, we choose a total of 25 filters of same size and stride as in C1 (i.e., [3, 3] and stride [1,1]). We use the padding in C2 in such a way that the size of the output remains the same as that of the input (i.e, padding = 1). We then feed the output of C2 to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031899, IEEE Access

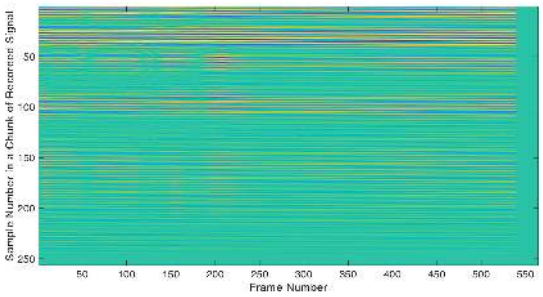

FIGURE 15: Correlation Matrix of a Sample Recording

FIGURE 16: Generated E-Sign

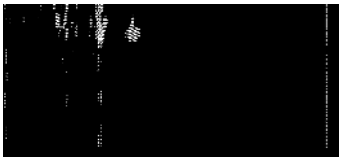

FIGURE 17: Subject # 1 - Sample 1

FIGURE 18: Subject # 1 - Sample 2

FIGURE 19: Subject # 2 - Sample 1

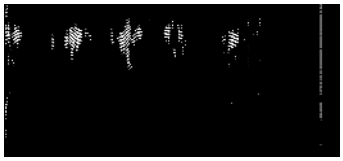

FIGURE 20: Subject # 2 - Sample 2

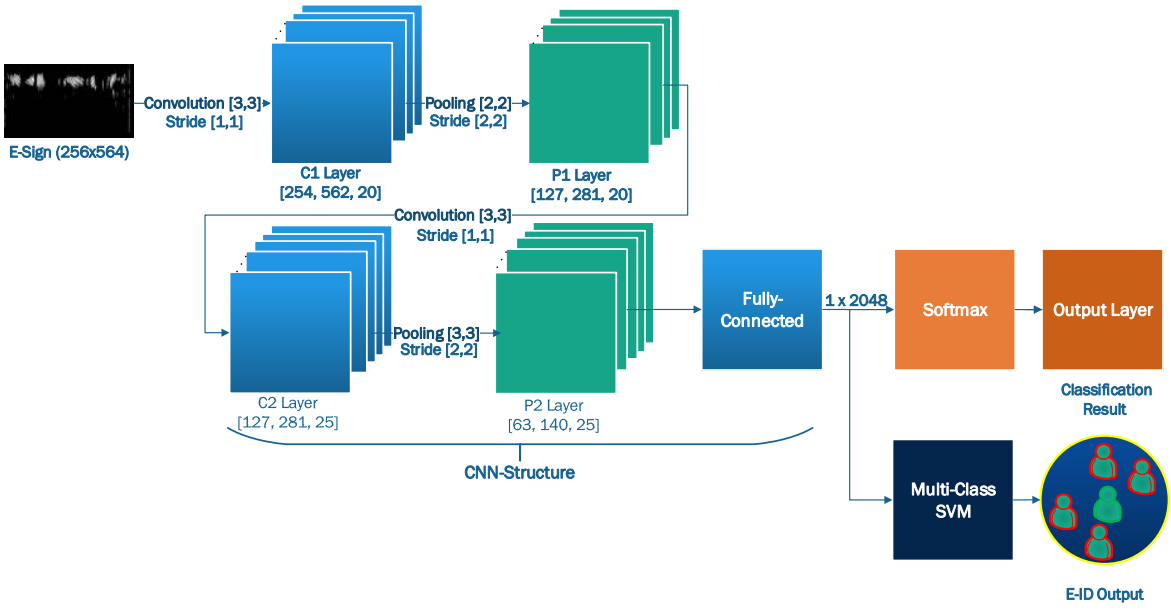

FIGURE 21: Deep-learning model employed in E-ID

P2, for which the window-size is empirically set to [3,3] with a stride of [2, 2], resulting in an output of dimensions [63, 140, 25].

The output of P2 represents the local features learned from different receptive fields of the input E-Sign sample. To make an identification decision using the learnt features, we unfold these features to flatten vectors (i.e., of order [1, 220500]). This vector is then fed to a fully-connected layer whose size is empirically set to 2048. As the name suggests, all the neurons in the fully-connected layer are connected to those in the previous layer and we treat its output as the feature-vector(s). As these features can be seen as time-series, the LSTM structure seems an appropriate choice since this helps in determining the temporal-relevance in the sequential data. LSTM networks have also shown promise in similar classification tasks [18]. However, our analysis has revealed that feeding the features learnt by the CNN to a secondary LSTM network results in a degraded performance of the E-ID. Our

analysis also revealed that by feeding the CNN features to a Support Vector Machine (SVM) results in good performance. Therefore, in our implementation, we treat CNN as a feature-extractor, and then employ a secondary SVM classifier which helps in identifying the corresponding individual given an E-Sign sample (i.e., learnt features).

III. EVALUATION SETUP AND EXPERIMENTAL METHODOLOGY

In this section, we discuss the evaluation setup, experimental methodology, and evaluation results of the E-ID.

A. EVALUATION SETUP

Figure 22 shows the evaluation setup of E-ID. Since the stand-alone PIN-PADs are not equipped with the microphones, we implemented the concept of E-ID by leveraging a commodity speaker and a microphone. As can be seen in Figure 22, we used a laptop (Hp Folio 9480m) to generate and transmit the inaudible sound frequencies through its built-in speaker (see section II-A for details of signal generation). For recording these frequencies and their echos, we interfaced an external microphone with this laptop. Although, the built-in microphone of the laptop can also be used to record the inaudible sound, our analysis revealed that these recordings were too noisy as compared with the external microphone (which may be due to the differences in build quality). In addition, the external microphone also allows to place the key-pad (i.e., upon which user enters the PIN) in its close vicinity, which would be a representation of the anticipated usage-scenario of the E-ID (i.e., on a typical PIN-PAD microphone would be close to the key-pad). For entering the PIN, we placed a mobile phone (upon which user enter the PIN on the usual key-pad) in close proximity (about 10cm) of the microphone so as to record the user's E-Sign while entering the PIN. Note that, the difference between the virtual buttons on a mobile phone and physical buttons on a PIN-PAD may impact the generated E-Sign. However, if the physical button are used instead of virtual buttons, the enrollment process will cater for this difference, i.e., the enrollment procedure will capture the E-Sign accordingly. Hence, we anticipate that E-ID is likely to work for any type of PIN-PADs (i.e., with virtual or physical buttons). The laptop that emits the inaudible sound was placed at an approximate distance of 0.4m from the microphone and key-pad (i.e., on the mobile phone). This whole setup is thus a close representation of the real-world setting in which E-ID may be used (i.e., a PIN-PAD device that can record the inaudible sound frequencies in parallel while a user enters the PIN).

B. EXPERIMENTAL METHODOLOGY

For evaluating the performance of E-ID in identifying the individual from an enrolled-set, we recruited a total of 5 subjects (4M+1F). All of these subjects were PhD students aged 25-35 years. Although the number of enrolled subjects may appear less, it is representative of a typical small smart space which on an average has less than 5 inhabitants. For example, in OECD countries a typical micro-enterprise has around 5-6 occupants, while an average home has 2.4 inhabitants on average [13] [14]. For generating the user's E-Sign, we asked each user to type the same PIN (i.e., 7913, randomly selected) on the key-pad (of the mobile phone as shown in Figure 22), and collected around 50 (± 5) data samples from each user leading to a total of 250 samples (i.e., $50 \times 5 = 250$). The length of the PIN may impact the E-Sign. However, this will also be catered by the enrollment process and will not affect the working of E-ID. While the user enters the PIN, the laptop transmits the inaudible chirp signals, which are simultaneously recorded by the microphone. The transmitted signals are reflected from the user's moving fingers and hand (due to entering the PIN), which appear as a unique pattern in the recorded signals. We process the recording in accordance with the method detailed in subsections II-A -II.C to generate the E-Signs and use these to identify the corresponding user from the enrolled set by leveraging the deep-learning model described in sub-section II-D. Out of 50 E-Sign samples/user, we used 35 samples for training the deep-learning model, while the rest are used for testing the performance of the E-ID. We also varied the number of training samples from $15 - 35$ to analyze the impact of the number of training samples upon the accuracy of the E-ID (see Section III-D for details). Note that, during the data collection process, the subjects were allowed to converse with the co-author (who was facilitating the experimentation). This depicts that our experimental scenario is close to real-world setting where the user is likely to talk with others (e.g., an accompanying friend) while entering the PIN.

C. PERFORMANCE EVALUATION

We use the following metrics for evaluating the performance of E-ID: i) Accuracy - which shows the percentage of correctly identified E-Sign samples across a set of enrolled users. ii) Confusion Matrix - which shows the performance of E-ID across all the test E-Sign samples in a matrix form. iii) Receiver Operating Characteristics (ROC) - which plot the TPR (True Positive Rate) vs FPR (False Positive Rate) at various thresholds.

D. EVALUATION RESULTS

For evaluating the performance of E-ID, we analyzed every possible combination of N subjects (i.e., $\binom{5}{N}$), where N represents the size of the evaluated group. For every combination of N subjects, we train a separate model with 35 training E-Sign samples, and test its performance on remaining 15 samples (see Section III- B for details related to collection of E-Sign samples). However, it is noteworthy that we do not do any fine tuning of parameters or layers for different group sizes – i.e., we keep our model generalized, and test it with every possible combination of N users. The evaluation of every possible combination helps us to ascertain the average accuracy of E-ID for different group sizes.

Table 2 shows the accuracy of E-ID for every possible combination of N subjects, while Figure 23 shows the per-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031899, IEEE Access

TABLE 1: Parameters of CNN structure employed in E-ID

No.	Parameter	Description	No.	Parameter	Description
1	Input Layer	256x564x1 E-Signs with zero-centre normalization	2	1st Conv Layer	20 3x3x1 convolution filters with stride [1 1] and padding [0 0 0 0]
3	Batch Normalization	with 20 Channels	4	Activation	ReLu
5	Pooling	2x2 max pooling with stride [2 2] and padding [0 0 0 0]	6	2nd Conv Layer	25 3x3x20 convolution filters with stride [1 1] and padding [1 1 1 1]
7	Batch Normalization	with 25 Channels	8	Activation	ReLU
9	2nd Pooling Layer	3x3 max pooling with stride [2 2]	10	Fully Connected	2048
11	Activation	Softmax	12	Cost Function	Cross-Entropy
13	Optimizer	SGDM	14	Learning Rate	0.0001
15	Max Epochs	15	16	Layers	10

FIGURE 22: Evaluation setup of E-ID

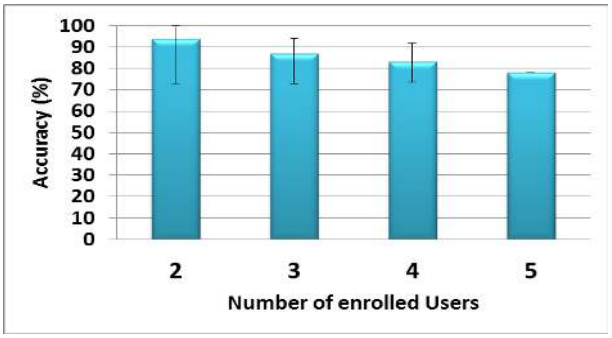

FIGURE 23: E-ID Performance

centage accuracy of the E-ID across the group-sizes of 2-5 individuals. Note that, the accuracy figures in Table 2 are arranged in descending order and have no bearing with the combinations. It is conspicuous from the Figure 23 that, the accuracy drops as the size of enrolled set increases. A similar trend is also observed in prior research [9]–[11] that utilize the WiFi signals to capture the user's gait characteristics or cardiopulmonary activity to perform the identification in small smart spaces. This may be due to the fact that the inclusion of more individuals in an enrolled set increases the chances of having similar E-Sign samples in the group, resulting in a decrease in the accuracy. Similarly, we notice that some of the combinations have relatively lower accuracy. This may be due to noisy E-Sign -i.e., due to the movement of some other body parts while entering the PIN which and impact the echo reflection. This is also observed in prior identification works such as [9]–[11] where accuracy for a particular group-size fluctuate. We believe that techniques such as ICA or PCA may help in segregating the echos of different body parts and may be an interesting direction for future research works. However, it is noteworthy that the accuracy of E-ID remains above 78% for groups of any size of up to 5 subjects. This is comparable with the prior user identification mechanisms [9], [10] that require user to walk explicitly on a long predefined path to perform the identification. In contrast, the E-ID is completely transparent to the user (i.e., user only enters the PIN as usual), and does not demand any explicit activity (e.g., walk) as in prior works [9], [10]. E-ID outperforms [11] that uses the imprints of user's cardiopulmonary activity manifested in the WiFi signals for human identification in smart spaces by approximately 5-15% for group-sizes of 2-5 individuals, respectively. In addition, unlike E-ID, [11] requires the user to sit (or stand) still in front of a commodity WiFi device for a duration of at-least 20 seconds, which may be onerous for the user. This suggests that E-ID offers significant benefits over the other state-of-the-art mechanisms.

Figures 24 -31 show the confusion matrices and ROCs of one combination (randomly selected) of group-sizes of 2-5 subjects, respectively. The confusion matrices also depict that as the number of enrolled subjects increases, some of the E-Sign samples belonging to different users are incorrectly classified (or mis-classified) by the E-ID, resulting in a drop in the overall accuracy. Note that in ROCs, the $class_i$ represents the i_{th} enrolled subject. The curves in ROCs are more spread-out as the group-size is increased. For example, it is conspicuous from Figure 31 that, for 4 subject (i.e., $class_1$-$class_4$) the corresponding ROC curves cluster around the top-left corner, which confirms that E-ID can discriminate the E-Sign samples of these subject with high accuracy. However, for the fifth user (i.e., $class_5$), the corresponding ROC curve is more spread-out at varying threshold values, which shows that E-Sign samples of this subject are confused with others. This is also evident from the corresponding confusion matrix (see Fig 30), where it can be observed that 8 evaluated E-Sign samples of other subjects are mistakenly classified as those of subject 5. This may be due to the resemblance in the E-Sign samples of this individual with others. Prior research has also shown the evidence of similarity amongst the behavioral biometrics (e.g., gait pattern) data of some individuals [20]. Nevertheless, even with all of these possibilities, the average accuracy of E-ID is above 78% for all group-sizes. E-ID is thus the first-step towards realizing an inaudible sound based human identification system for small

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031899, IEEE Access

TABLE 2: Performance of E-ID across different combination of N enrolled subjects

Group-Size	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	Avg Accuracy
2	100	100	97.30	97.14	97.06	96.88	96.15	93.75	87.10	72.4	**93.77%**
3	94.12	93.88	93.48	90.38	89.36	86.96	85.37	83.67	79.07	72.73	**87%**
4	91.80	87.88	85.7	75.86	73.44	-	-	-	-	-	**83%**
5	78.2	-	-	-	-	-	-	-	-	-	**78.2**

smart spaces (i.e., with 2-5 inhabitant) that does not demand any participation from the user (unlike prior works [9]–[11]).

Table 3 shows the evaluation results of E-ID for every possible combination of N subjects obtained by feeding the CNN features to the Softmax classifier. These results show that, for all the group-sizes, SVM outperforms the softmax classifier by approximately 2-7%. This confirms the effectiveness of our approach (i.e., feeding CNN features to SVM) in identifying the corresponding subject from an enrolled-set given an E-Sign sample. We also implemented a LSTM based model which treats the CNN features as a time-series and feed these time-series to a secondary LSTM structure to learn the temporal-relevance in the learnt features. Our analysis revealed that this approach fails to achieve an acceptable performance for any group-size.

1) Impact of Number of Training Samples

Recall from Section 2.3 that, we train our deep-learning model with 35 (out of 50) E-Sign samples/user (results presented above are with 35 training samples). In this subsection, we analyze the performance of E-ID by varying the training samples from 15 to 35 with an increment of 5 in each iteration. Figure 32 shows the performance of E-ID with different number of training samples (and with group-size of 5 subjects). It is conspicuous that the accuracy initially increases as the number of training samples are increased from 15 to 25 (e.g., 74% with 15 samples vs 78% with 25 samples). However, beyond that, the accuracy stays almost stable (i.e., around 78%). As discussed in Section I, E-ID requires one-time enrollment of the user. The accuracy of over 75% with less training samples (i.e., 20 - 25 samples) will mean a swifter enrollment process which will aid in real-world applicability of the E-ID. Note that, since the deep-learning is highly dependent on the size of training data, we anticipate that training with more E-Sign samples may result in even better performance. However, this may not be practically convenient for the users to provide a large number of training samples (e.g., 100 or more) as a part of the enrollment. In view of this, we have only tested the E-ID with less number of training samples (i.e., 15 - 35) to demonstrate its practical applicability.

2) Impact of Learning Rate

To achieve good performance, learning rate is an important hyperparameter to tune. It controls the speed at which the deep-learning model learns to approximate the corresponding subjects given the training E-Sign samples. We optimize the learning rate for CNN and select the one that achieve the best performance. Figure 33 shows the E-ID performance (with group-size of 5) at varying learning rates of CNN structure (i.e., $1e^{-1}, 1e^{-2}, 1e^{-3}$, and $1e^{-4}$). It is evident that the accuracy improves by reducing the learning rate (i.e., 78% at $1e^{-4}$ vs 0% at $1e^{-2}$). However, our analysis revealed that by decreasing the learning rate below $1e^{-3}$, the CNN fails to learn any discriminating features from the E-Sign samples that may help in identifying the corresponding individual (see Fig 33). Therefore, we select the learning rate to be $1e^{-4}$ as it results in best performance for different group-sizes of the enrolled subjects.

IV. RELATED WORK

The closest work to ours is [21] that leverages the sound (both audible and inaudible) reflected from the person's ear canal for authenticating the user. Since different individuals have a different ear canal shape, the sound reflected from such a cavity shows a discriminating frequency response that may be used for establishing the identity of the user. However, this mechanism demands an earpiece with a microphone for sending the probe sound signal in the ear canal and recording the echos reflected from the ear cavity. This requirement may be arduous for the users which may hinder the adoption of this approach in our anticipated usage scenario (i.e., smart spaces). In contrast, E-ID does not demand any special hardware (e.g., earpiece) for performing human identification in smart spaces. Likewise, authors in [22] presented a mechanism where a user's identity is confirmed by utilizing his breathing sound. However, this work requires the user to place the microphone very close to the nose and also require one to undertake a deliberate action (i.e., deep breath or sniff). Contrary, E-ID neither demands a user to undertake any particular action (e.g., breath gesture) nor requires any explicit interaction with the hardware (e.g., placing a microphone close to noise). Similarly, the authors in [23] proposed a sound-based user authentication mechanism for online services that may be extended to perform identification in smart spaces. This mechanism records the ambient sound on two co-existing devices of the user (e.g., user's laptop and mobile phone), and if both the devices record a similar sound then they are deemed to be in proximity and results in successful authentication of the user. This approach may be modified to perform human identification in smart environments (e.g., user's mobile phone and PIN-PAD records the ambient sound). However, this mechanism requires user to generate some sound (e.g., clearing throat) when no ambient

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031899, IEEE Access

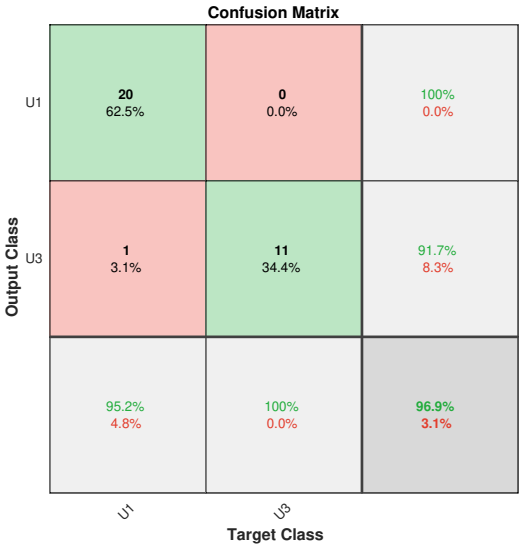

FIGURE 24: Confusion Matrix - 2 Users

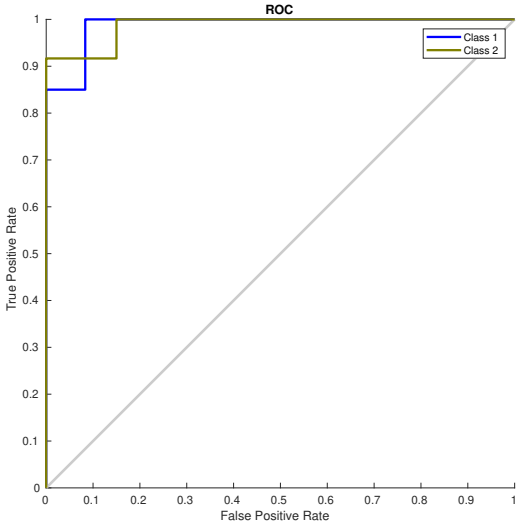

FIGURE 25: ROC - 2 Users

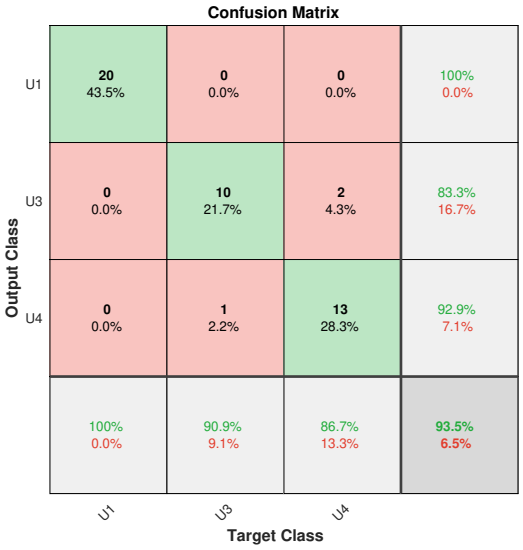

FIGURE 26: Confusion Matrix - 3 Users

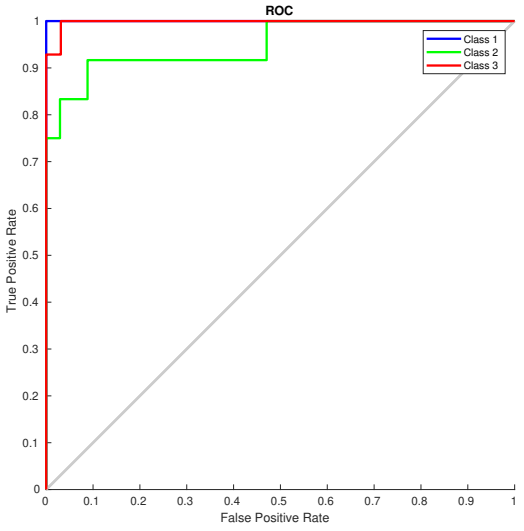

FIGURE 27: ROC - 3 Users

TABLE 3: Performance of E-ID with CNN only

Group-Size	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	Avg Accuracy
2	100	97.14	96.55	94.12	93.75	92.5	90.62	87.10	80.77	62.07	**89.4%**
3	91.84	89.11	87.76	86.54	86.27	85.37	85.11	82.6	69.77	68.18	**83.2%**
4	86.89	81.82	79.37	71.88	60.34	-	-	-	-	-	**76.06%**
5	76.92	-	-	-	-	-	-	-	-	-	**76.92%**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031899, IEEE Access

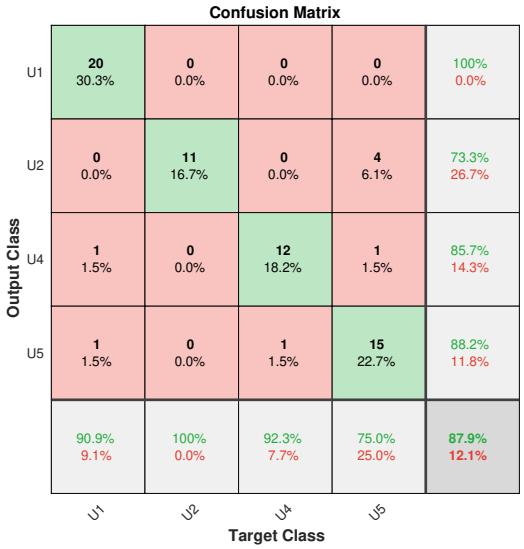

FIGURE 28: 4 Users

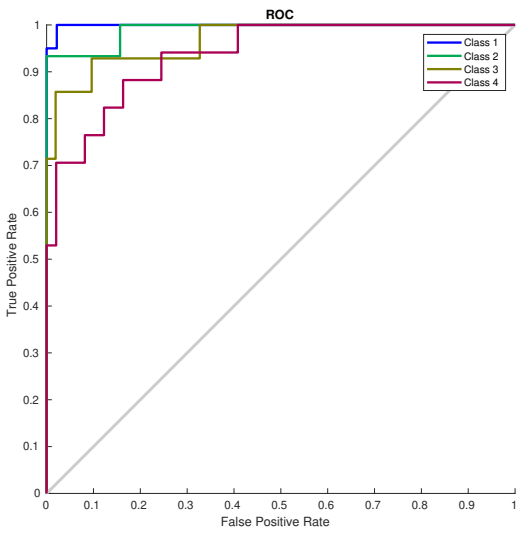

FIGURE 29: ROC-4 Users

sound is present which may be annoying for the other people in the vicinity [24]. Additionally, this approach may not work if the user's mobile phone is lost, stolen, or discharged. Furthermore, a person in unlawful possession of the victim's mobile-phone can potentially access the smart-environment under the victim's identity. In contrast, E-ID is not dependant upon any secondary device like mobile phone. Authors in [6] have utilized a camera for capturing the user's gait-pattern and used them for identification. However, unlike E-ID, this approach has the privacy issues. Similarly, a few works have used the sensors (e.g., accelerometer, gyroscope, etc) embedded in the smartphone [25] [26] or in the smartwatch [8] to capture the user-specific data (e.g., gait-pattern or arm-motion) for identifying the user. However, the requirement to carry a particular device for human identification may be deemed to be onerous by the users. E-ID, on the other hand does not require a user to carry such devices (i.e, smartwatch or phone) for its operation. [27] shows the possibility of using user's voice to establish his identity. However, this approached may be spoofed by furtively recording the victim's voice and launching the playback attack. E-ID utilizes the user's behavioral characteristics while entering the PIN, which are hard to capture (or imitate) by a potential adversary. A few works have leveraged the ubiquitous WiFi signals available in smart spaces to establish the identities of the inhabitants. For example, [9], [10] used the WiFi signals to capture the user's gait-pattern and subsequently use the corresponding WiFi perturbations to establish the identity of the user. Both of these approaches have demonstrated a similar accuracy as that of E-ID for a maximum group-size of 5-6 subjects. However, unlike E-ID, these works require a user to walk on a predefined path, which may be burdensome for the user and not always possible. In addition, the small smart spaces are unlikely to have a long straight path (e.g., 2.4m required in [9]) for the successful operation of these mechanisms. Similarly, the authors in [11] showed that it is possible to utilize the user's cardiopulmonary activity manifested as perturbations in pervasive WiFi signals to perform the human identification in small smart spaces. However, unlike E-ID, this work requires user to stand in-front of a WiFi device for a minimum duration of 20 seconds. E-ID outperforms this work by around 10% for a group-size of 5 individuals without necessitating any participation from the user (e.g., stand in-front of a commodity WiFi device). Furthermore, all of the aforementioned mechanisms that utilize WiFi (i.e., [9]–[11]) requires a controlled environment - i.e., only the authenticating person should be present in the vicinity of the WiFi transceivers, which may not always be possible in a real-world scenario. Unlike these approaches, E-ID can operate in a real-world setting.

While there are numerous other mechanisms like fingerprint, face-recognition, and iris scans that are already used for identifying the individuals, they suffer from a number of well-known vulnerabilities. For example, fingerprints can easily be collected from a surface that a victim may have touched and used to circumvent the fingerprint based authentication [28]. Likewise, face-recognition may be spoofed by using the victim's photograph (which is easy to find over social media) or 3-D printed head [2], [3]. Similarly, iris based mechanisms are also prone to subversion by using a victim's photograph (even captured from a long distance of up to 5m) superimposed with a contact lens [5]. Unlike these approaches, spoofing E-ID is difficult as it utilize the user's behavioral characteristics while entering the PIN, which in general are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031899, IEEE Access

FIGURE 30: 5 Users

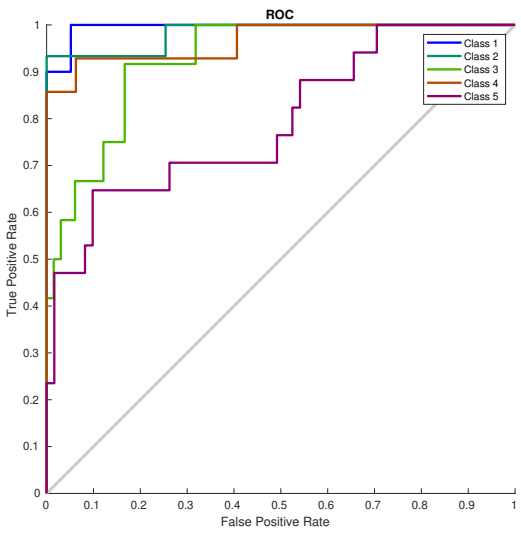

FIGURE 31: ROC-5 Users

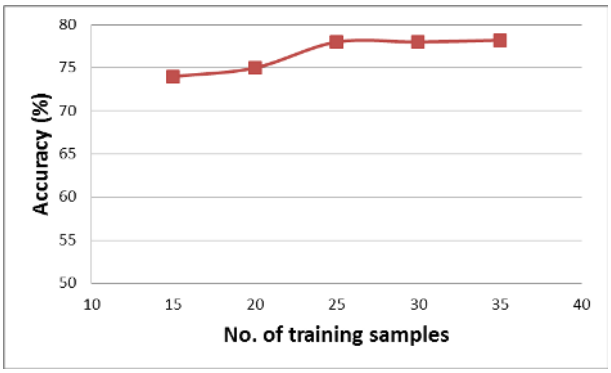

FIGURE 32: E-ID performance with varying number of training samples

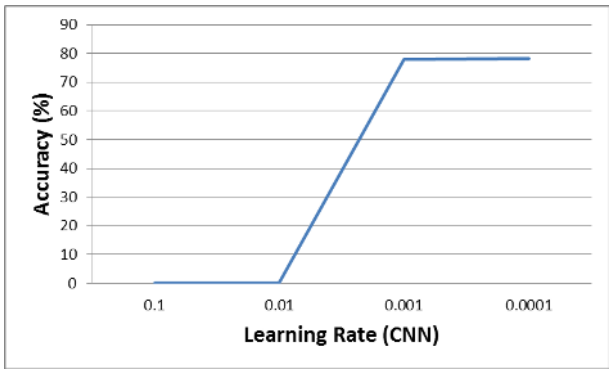

FIGURE 33: Impact of Learning Rate (CNN)

not easy to imitate. This shows that E-ID offers significant benefits over the other state-of-the-art mechanisms.

V. CONCLUSION

This paper undertook an investigation into the use of inaudible sound signals for capturing the user's behavioral biometrics (i.e., habitual characteristics) while s/he taps (or types) the PIN. We analysed the possibility of using these characteristics for identifying the corresponding user from an enrolled-set. To this end, we present E-ID, a novel non-intrusive identification system that leverages the commodity speaker and microphone for capturing the user's behavioral tapping/typing patterns in an inaudible range. We present a comprehensive processing pipeline - i.e., from transmission of signal to generation of user-specific imprints in the recorded echos. We also craft a deep-learning based identification strategy that helps in accurate identification of subjects from an enrolled-set of N subjects. Our evaluations revealed that, E-ID can identify an individual with an average accuracy 93% to 78% for a group-size of 2-5 individuals, respectively. In future, we plan to extend E-ID for bigger enrolled-sets (e.g., 10 or more subjects). We also plan to implement E-ID on small form factor devices with embedded speaker and microphone. This will represent a more realistic usage scenario and will help to analyse the impact on different practicalities such as processing time and accuracy in real-world situations. In addition, another interesting future research direction could be to use E-Signs as a second-factor of authentication in situations where every user has a separate PIN. This would need comparing the E-Sign of a user with his/her enrollment samples instead of comparing it with the samples of all enrolled-set. We also plan to conduct these investigations in future.

REFERENCES

[1] S. W. Shah and S. S. Kanhere, "Recent trends in user authentication – a survey," IEEE Access, vol. 7, pp. 112 505–112 519, 2019.
[2] L. H. Newman. (2016) Hackers Trick Facial-Recognition Logins with photos from Facebook (What else?). [Online]. Available: //www.wired.com/2016/08/
[3] T. Brewster. (2018) We Broke Into A Bunch Of Android Phones With A 3D-Printed Head. [Online]. Available: https://www.forbes.com/
[4] Samsung. (2016) Iris Recognition on Galaxy S8. [Online]. Available: https://www.samsung.com/au/iris/
[5] D. Goodin. (2016) Breaking the iris scanner locking Samsung's Galaxy S8 is laughably easy. [Online]. Available: https://arstechnica.com/
[6] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu, "Silhouette analysis-based gait recognition for human identification," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 12, pp. 1505–1518, Dec 2003.
[7] J. Cheng, M. Sundholm, B. Zhou, M. Kreil, and P. Lukowicz, "Recognizing Subtle User Activities and Person Identity with Cheap Resistive Pressure Sensing Carpet," in 2014 International Conference on Intelligent Environments, pp. 148–153.
[8] A. S. Guinea, A. Boytsov, L. Mouline, and Y. Le Traon, "Continuous Identification in Smart Environments Using Wrist-Worn Inertial Sensors," ser. MobiQuitous '18. New York, NY, USA: ACM, 2018, pp. 87–96.
[9] Y. Zeng, P. H. Pathak, and P. Mohapatra, "WiWho: WiFi-based person identifcation in Smart Spaces," in ISPN 2016.
[10] J. Zhang, B. Wei, W. Hu, and S. S. Kanhere, "WiFi-ID: Human Identification Using WiFi signals," in 2016 International Conference on Distributed Computing in Sensor Systems, pp. 75–82.
[11] S. W. Shah and S. S. Kanhere, "Smart user identification using cardiopulmonary activity," Pervasive and Mobile Computing, vol. 58, p. 101024, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1574119218305145
[12] N. Zheng, K. Bai, H. Huang, and H. Wang, "You are How You Touch: User Verification on Smartphones via Tapping Behaviors," in 2014 IEEE 22nd International Conference on Network Protocols, pp. 211–221.
[13] OECD. (2015) OECD Family Database. [Online]. Available: https://www.oecd.org/els/family
[14] OECD SME. (2015) Small and Medium-Size Enterprises (SMEs). [Online]. Available: https://stats.oecd.org/
[15] H. Nyquist, "Certain topics in telegraph transmission theory," Transactions of the American Institute of Electrical Engineers, vol. 47, no. 2, pp. 617–644, April 1928.
[16] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota, "FingerIO: Using Active Sonar for Fine-Grained Finger Tracking," in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ser. CHI '16. New York, NY, USA: ACM, 2016, pp. 1515–1525. [Online]. Available: http://doi.acm.org/10.1145/2858036.2858580
[17] P. Cheng, I.E. Bagci, U. Roedig and J. Yan, "Sonarsnoop: Active acoustic side-channel attacks," International Journal of Information Security, vol. 19, pp. 213–228, https://doi.org/10.1007/s10207-019-00449-8 2019.
[18] X. Zhang, L. Yao, S. S. Kanhere, Y. Liu, T. Gu, and K. Chen, "MindID: Person identification from brain waves through attention-based recurrent neural network," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 2, no. 3, pp. 149:1–149:23, Sep. 2018. [Online]. Available: http://doi.acm.org/10.1145/3264959
[19] Q. Zou, Y. Wang, Q. Wang, Y. Zhao, and Q. Li. (2020) Deep Learning Based Gait Recognition Using Smartphones in the Wild. [Online]. Available: https://arxiv.org/abs/1811.00338
[20] T. T. Ngo, Y. Makihara, H. Nagahara, Y. Mukaigawa, and Y. Yagi, "The largest inertial sensor-based gait database and performance evaluation of gait-based personal authentication," Pattern Recognition, vol. 47, no. 1, pp. 228 – 237, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S003132031300280X
[21] A. Takayuki, "Ear acoustic authentication technology: Using sound to identify the distinctive shape of the ear canal," NEC Technical Journal-Special Issue on Social Value Creation Using Biometrics, vol. 13, no. 2, pp. 87–90, 2019. [Online]. Available: https://www.nec.com/en/global/techrep/journal/g18/n02/pdf/180219.pdf
[22] J. Chauhan, Y. Hu, S. Seneviratne, A. Misra, A. Seneviratne, and Y. Lee, "BreathPrint: Breathing Acoustics-based User Authentication ," in MobiSys 2017.
[23] N. Karapanos, C. Marforio, C. Soriente, and S. Capkun, "Sound-Proof: Usable Two-Factor Authentication Based on Ambient Sound," in 24th Usenix Security Symposium 2015, pp. 483–498.
[24] S. W. Shah and S. S. Kanhere, "Wi-Sign: Device-Free Second Factor User Authentication," in Proceedings of the 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, ser. MobiQuitous '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 135–144. [Online]. Available: https://doi.org/10.1145/3286978.3286994
[25] R. Damaševičius, R. Maskeliūnas, A. Venčkauskas, and M. Woźniak, "Smartphone user identity verification using gait characteristics," Symmetry, vol. 8, no. 10, 2016. [Online]. Available: https://www.mdpi.com/2073-8994/8/10/100
[26] W. Shi, J. Yang, Yifei Jiang, Feng Yang, and Yingen Xiong, "Senguard: Passive user identification on smartphones using multiple sensors," in 2011 IEEE 7th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), Oct 2011, pp. 141–148.
[27] J. Angelo. (2017) Two-Factor Authentication Using Biometrics. [Online]. Available: https://auth0.com/blog/two-factor-authentication-using-biometrics/
[28] S. W. Shah and S. S. Kanhere, "Wi-Access: Second Factor User Authentication leveraging WiFi Signals," in 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), 2018, pp. 330–335.

SYED W. SHAH (Member, IEEE) received his Ph.D. degree in Computer Science and Engineering from University of New South Wales (UNSW), Sydney, Australia, and M.S. degree in Electrical and Electronics Engineering from University of Bradford, U.K. He is currently a Postdoctoral Writing Fellow at Faculty of Engineering, UNSW Sydney, Australia. His research interests include pervasive/ubiquitous computing, user authentication/identification, Internet of Things, signal processing, data analytics, privacy and security.

ARASH SHAGHAGHI (Member, IEEE) received his PhD in Computer Science and Engineering at UNSW Sydney, Australia and MSc in Information Security at University College London (UCL), UK. He is currently a Lecturer at Centre for Cyber Security Research & Innovation (CSRI) of Deakin University. He also holds affiliation with UNSW Sydney as a Visiting Fellow. Previously, he has also held positions with University of Melbourne, and the University of Texas at Dallas. His research interests include Access Control, Network Security, Internet of Things, and Blockchain. He is also passionate in developing usable and verifiable Privacy Enhancing Technology (PET) solutions. He is a board member at Australian Privacy Foundation (APF), a member of Australian Computer Society (ACS), and Australian Information Security Association (AISA).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031899, IEEE Access

SALIL S. KANHERE (Senior Member, IEEE) received the M.S. and Ph.D. degrees from Drexel University, Philadelphia. He is currently a Professor of Computer Science and Engineering with UNSW Sydney, Australia. He also holds affilations with CSIRO's Data61 and the Cyber Security Cooperative Research Centre (CSCRC). His research interests include the Internet of Things, cyberphysical systems, blockchain, pervasive computing, cybersecurity, and applied machine learning. He is a Senior Member of the ACM, a Humboldt Research Fellow, and an ACM Distinguished Speaker. He serves as the Editor in Chief of the Ad Hoc Networks journal and as an Associate Editor of the IEEE Transactions On Network and Service Management, Computer Communications, and Pervasive andMobile Computing. He has served on the organising committee of several IEEE/ACM international conferences. He has co-authored a book titled Blockchain for Cyberphysical Systems.

JIN ZHANG (Member, IEEE) received his Ph.D. degree in Computer Science and Engineering from University of New South Wales in 2017. He is a postdoc in Lab for human machine control at Shenzhen institutes of advanced technology, Chinese Academy of Sciences. His research interests include WiFi human sensing, machine learning and computer networks.

ADNAN ANWAR (Member, IEEE) is a Lecturer and deputy director for the postgraduate cybersecurity studies at the School of Information Technology, Deakin University. Previously he has worked as a Data Scientist at Flow Power, an energy management and solution company. He has over 8 years of research, and teaching experience in universities and research labs including NICTA, La Trobe University, and University of New South Wales. He is broadly interested in the security research for critical infrastructures including Smart Energy Grid, SCADA system, and application of machine learning and optimization techniques to solve cyber security issues for industrial and IoT systems. He has authored over 30+ articles including journal, conference articles and book chapters in prestigious venues.

ROBIN DOSS (Senior Member, IEEE) is a Professor and the Research Director of the Strategic Centre for Cyber Security Research & Innovation (CSRI) at Deakin University. In this role, he provides scientific leadership for this multidisciplinary research centre focused on the technical, business, human, policy and legal aspects of cybersecurity. In addition, he also leads the Next Generation Authentication Technologies theme for the Critical Infrastructure Security research program of the national Cyber Security Cooperative Research Centre (CSCRC). Prior to this role, he was the Deputy Head of School for the School of Information Technology at Deakin University. Robin has an extensive research publication portfolio and in 2019 was the recipient of the 'Cyber Security Researcher of the Year Award' from the Australian Information Security Association (AISA). His research interests include the broad areas of system security, protocol design and security analysis with a focus on smart, cyberphysical and critical infrastructures. His research program has been funded by the Australian Research Council (ARC), government agencies such as the Defence Signals Directorate (DSD), Department of Industry, Innovation and Science (DIIS) and industry partners. He has contributed to large multiyear projects under the European Union's Framework Program (FP6) and been funded by the Indian Government under the Scheme for Promotion of Academic and Research Collaboration (SPARC). He is a member of the executive council of the IoT Alliance Australia (IoTAA). He is founding chair of the Future Network Systems and Security (FNSS) conference series and is an associate editor for the Journal of Cyber Physical Systems.