



Cybersecurity of Smart Home Systems: Sensor Identity Protection

Yazan Alshboul¹ · Abdel Al Raouf Bsoul² · Mohammed AL Zamil² · Samer Samarah³

Received: 20 July 2020 / Revised: 16 January 2021 / Accepted: 3 February 2021 /
Published online: 4 March 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC part of Springer Nature 2021

Abstract

Smart home systems are designed as platforms for connecting sensors, home appliances, and devices to exchange data and, ultimately, to provide useful services to home residents. However, such systems are vulnerable to Cybersecurity attacks that can affect the reliability and integrity of the delivered services. Sensors, planted at smart homes or equipped with smart appliances, are highly exposed to identity theft. Intruders can recognize through the understanding of the exchanged data, their locations, or knowing their associated services. Such information might make the home resident vulnerable to life attacks. Therefore, protecting sensors identities in smart home systems is of high interest in this domain. This paper introduces a novel technique that protects sensors' identity from being recognized through cordless communication environments. Our proposed approach utilizes a three-phase technique that controls a synchronized queue among connected sensors and keeps their identity hidden from outsiders. The proposed approach preserves the linearity of time that is required to manage the protection of the home network. To validate the performance of our proposed approach, we conducted experiments on four different smart homes datasets. Furthermore, we performed a sensitivity analysis to measure how our proposed approach is affected by different environmental variables. The results indicated that the proposed approach provides a significant performance in protecting sensors identities in smart home area networks. Furthermore, during the sensitivity analysis, we found that our proposed technique's performance is highly affected by the threshold value that defines each sensor's time interval.

Keywords Identity theft · Intrusion detection · Sensitive environment · Cybersecurity · Cyber-attacks · Machine learning · Internet-of-things

✉ Mohammed AL Zamil
Mohammedz@yu.edu.jo

Extended author information available on the last page of the article

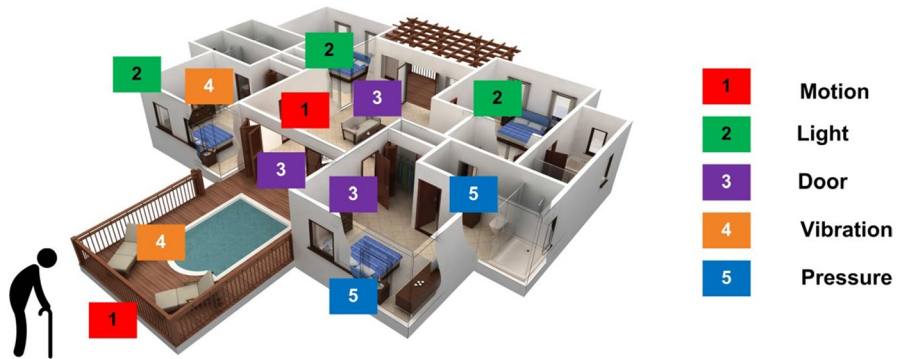


Fig. 1 Smart home systems

1 Introduction

Due to the advancements in internet-of-things (IoT) technologies and their infrastructures, the proliferation of smart home services is gaining greater momentum across several ICT industries. IoT-based smart home systems combine heterogeneous ubiquitous devices and appliances that are connected together in order to provide smart services to homes' inhabitants [1]. For instance, such smart services can monitor power consumption, control smart appliances, or recognize residents' activities to provide healthcare services or detect critical medical conditions. As homes become more intelligent, more complex and technology dependent services appear to exist. Figure 1 shows an architectural design of a smart home system, in which multiple heterogeneous sensors are installed to recognize elderly residents' activities.

The home area network (HAN) defines the connection topology, in which sensors report their readings back to some central unit such as a server or a cloud. Several HAN network technologies are used in existing smart homes, such as X10, ZigBee, and Z-wave. X10 is a protocol that allows for remotely controlling appliances at smart homes [2]. It involves short-distance radio frequency that enables fast communication between transmitters and receivers. On the other hand, ZigBee and Z-wave support mesh network topology, in which there is no single path for the message to get to its designated destination [3, 4]. As technology converges and the cost of connectivity decreases dramatically, HANs are always connected to the internet via wireless mediums. However, such always-on connectivity raises Cybersecurity threats where HANs could be vulnerable to intruders' attacks [5, 6].

Existing wireless network technologies are vulnerable to espionage or vandalism threats. This may lead to exploit this vulnerability to recognize sensors' identities since the wireless communication mediums are public [7]. Therefore, it allows for identifying the life patterns of smart homes' residents. Consequently, this may result in a cybersecurity breach that affects individuals' privacy or leads to cybercrimes. Specifically, intruders try to learn the behavior of inhabitants by identifying the functionalities of sensors that have been installed at homes. This

may lead to significant privacy issues that might cause dangerous crimes, facilitates whaling (a phishing attack that targets high profile individuals), or leaks sensitive information about homes' residents. Sensors' identity theft in smart homes is the process of identifying the types of the installed sensors, which results in discovering the events or activities that have been performed by a homes' residents. Such knowledge allows intruders to commit fraud or other types of crimes. Furthermore, knowing sensors' identities could manipulate the integrity of broadcasting data.

Current literature paid more attention to research solutions that preserve peoples' data privacy while ignoring the devices that people use in their daily activities in smart homes. Such ubiquitous equipment can easily lead to useful knowledge, which breaches the privacy of people as well. For instance, in [8] researchers investigated the preservation of users' privacy in social big data. Other research proposed a novel technique for securing users' privacy in smart mobile applications [9]. While in [10], researchers utilized an authentication method to preserve the privacy of users who benefit from IoT services.

However, a rare interest has been noticed in preserving sensors' identities that are embedded in smart devices and appliances. Research in [11–14] have focused on several authentication methods to mitigate the risk of data breaches. While authentication and authorization paradigms have proven their effectiveness in preventing unauthorized access to network activities, surpassing such techniques becomes easier as intruders interrupt wireless signals and sniff data packets from the open space, not edge nodes.

This paper introduces a framework for protecting sensors' identities in smart homes using a novel data-driven technique. The proposed methodology defines the sensor's identity problem in smart home environment as a binary classification problem, which measures how likely an intruder can predict the identities of smart home sensors. Our proposed approach relies on defining an extra data-level, which partition sensor readings into a set of scrambled signals that, in turn, can be aggregated into a meaningful and readable record at the destination. The proposed protocol preserves the identity of the data source while keeping data granularity at lower levels. In other words, it does not add extra load on the network medium.

Given a set of sensors that are connected via HAN in a smart home such as $H = \{s_1, s_2, \dots, s_n\}$ and a set of readings that have been resulted from each type of these sensors such as $S_k = \{d_{k,1}, d_{k,2}, \dots, d_{k,l}\}$, the set of all possible data that might be generated from this network is defined by the function $H \cdot S$. Furthermore, we define the function as follows:

$$f(H \cdot S_k) = H \times \bigcup_{(i=1)^N} d_{k,i} \quad (1)$$

Through this research, we will prove that $f(H \cdot S_k)^{-1}$ is the inverse of the original function with extremely low probability of being detected or predicted by intruders. Moreover, we will test our proposed methodology using well-known and efficient binary classifiers to measure its performance in terms of misclassification rate, loss function, and the sensitivity of our proposed methodology to running time, detection rates, enhancements, and its effect on each classifier.

This paper is organized as follows: Sect. 2 discusses the related work in the literature and highlights the contribution of this research as compared to existing ones. Section 3 introduces and explains the proposed methodology. Section 4 illustrates the experiments and the anticipated results. Finally, Sect. 5 concludes the research work.

2 Literature Review

This section discusses the related research from three perspectives: the communications among smart home devices and appliances, the existing research on identity theft, and the application of data-driven solutions to the identity theft problem in smart homes.

Home area networks have been developed to define the operational connectivity among devices and appliances at smart homes. The main differences of such interconnectivity are the need to connect several equipment with fast transmission medium, reliable connection, low-load, and ad-hoc connectivity that allows for multiple and heterogeneous nodes [15]. HANs have been successfully implemented to monitor the daily operations of home appliances, such as turning the light on or off, controlling the home temperature, or providing voice command system to monitor home appliances.

Recent advances of IoT technologies added another level of complexity; it is the need to collect data and make the decision based on the behavioral patterns of home inhabitants. In other words, the operations of different appliances became no longer independent. In addition, homes are also connected to other homes, hospitals, schools, cars, and other data sources to formulate smart cities [16]. Such complexity evolves the concept of HAN to include extra services to cope with current technology and the emerging need for smart services.

Connectivity in an ad-hoc environment, in which new appliances can be added, and their locations are changing over time, has been handled through embedding appliances with wireless sensors [17]. Consequently, the open spectrum medium of wireless sensors communications raises the concerns of security and privacy issues [18]. For this reason, communication protocols for HANs have been developed to provide the required functionality, specifications, and preserve the privacy of such networks.

ZigBee is a bidirectional radio frequency protocol that adopts the wireless networking standards of IEEE 802.15.4. ZigBee technology has been widely used in developing HANs as it provides low data transmission communication and, consequently, long-life battery [19]. Since ZigBee has been articulated on the top of wireless mediums, its designated architecture makes it subject to intrusion attacks as all appliances are connected via a single coordinator (controller). In addition, connecting ZigBee networks to an external internet connection or Wi-Fi requires extra equipment and complexity, which makes ZigBee not suitable for IoT connectivity standards.

On the other hand, Wi-Fi provides high transmission rates as compared to ZigBee; permitting appliances that require streaming, synchronous communications,

and fast response to function in smart homes [20]. Wi-Fi is also a bidirectional radio frequency protocol; implements IEEE 802.11 standards. While Wi-Fi is considered one of the most reliable and trusted connectivity medium, it is prone to interference due to the open spectrum environment of wireless communications.

To overcome intrusions, several researches have investigated the application of a secure layer on RFID (Radio Frequency Identification) technology [21], in which identification is performed through RFID tags. Unfortunately, it is hard and inefficient to replace sensors embedded in appliances with RFID tags to formulate an ad-hoc network. Furthermore, IoT connectivity seems impossible with RFID protocol, which relies on EPC (Electronic Product Code) protocol that has been developed to track items rather than facilitate communication among devices.

The process of identity theft attack (ITA) is based on scanning the networks to detect unsecured and weakly configured connections. Once detected, the attacker copies the identity and uses it to access private information. The purpose of attacks is varying [22] and can be classified into the following: obstruction of data, counter international cyber security measures, retardation of decision making, denial in providing public services, abatement of public confidence, and other goals.

Our problem is complicated in terms of securing the type of sensor (device) rather than securing the communicating data; protecting sensor identity. Specifically, in smart homes, if the attacker detects the functionality of a sensor, it will be easy to understand the behavior of smart home's resident [23]. Our goal is to protect the identity of the sensors rather than protecting the data generated by them.

There are several solutions in the literature with techniques to detect such types of Cybersecurity attacks. Unfortunately, most of these techniques are not applicable for securing sensor identity, which is the primary constituent of modern smart homes. For instance, the well-known solution that keeps track of neighbors with their locations so that the designated base station can regulate the communication [24], is not applicable in this domain. Such a traditional solution secures the communicating data rather than the identities of sensors [25].

TOR-based anonymous communication among smart home appliances has been proposed in [26], in which, authentication phase has been omitted. This approach is based on public-key cryptography, which is very expensive in terms of processing time and memory. Another interesting solution is the lightweight authentication sessions that have been proposed in [27]. It relies on establishing a token-based protocol to legitimate the identity of a smart device in which a centralized state table is kept to manage this process. The proposed solution guarantees communication security while ignoring anonymity.

Santoso and Vun [28] have proposed a more specific solution to IOT systems by considering the user convenience aspect. The proposed solution is based on establishing a shared key among different IOT sensors using Elliptic Curve Diffie Hellman (ECDH) primitive. In [29] and [30], multi-level and multi-tier schemes were implemented to expand the ECDH functionality. Neither ECDH nor its expanded versions guarantee to keep the type of sensor secure from attackers.

The easiest way to detect the type of wireless sensors embedded in smart devices at modern smart homes is the analysis of the data that are generated by these sensors [31]. For instance, a motion sensor on the restroom door can be interpreted as

the inhabitant need to count the number of times this room is used. Another sensor is measuring the level of insulin of a specific resident. Simply, both facts are correlated, since insulin can be easily mapped to diabetes, in which going to the restroom is a major symptom. Therefore, the attacker can conclude that the resident is infected by diabetes.

Data-layer technologies can provide simple and efficient solutions to handle the identity theft of sensors in smart homes. The ultimate goal of such solutions is to hide the semantic of communicating data; rather than hiding their values [32, 33]. However, detecting the semantic of communicating data is not a simple task. Recent advances in machine learning techniques (such as deep learning) facilitate detecting behaviors through data history.

Another important direction is the adoption of time sensitive networks [34], which can be utilized to centralize the control and management of traffic streams as a scheduling problem. Such configuration enhancement would positively minimize the overall communication time and allows for enforcing security rules among all connected sensors. Although this research adopted a distributed algorithmic methodology, it could be upgraded, in the future, to more centralized architecture.

Table 1 summarizes the related work and compares among existing methods in terms of their contributions and research directions.

3 Sensors Identity Protection

This section introduces our methodology to protect sensors' that are installed in a home area network. The proposed methodology is driven by a verification phase, in which it has been verified during the modeling of every phase against reliability issues such as concurrency. First, we describe the common communication model in the home area network, which clarifies the implementation environment. Next, we illustrate how the attacker can benefit from such environment to identify the sensors and then learn some meaningful information. Finally, we provide algorithmic descriptions to our proposed technique that detailed the execution of different phases in addition to the way we verify each of them.

3.1 Communication Model

Home Area Networks (HANs) are described as networks in which several smart home appliances are connected to communicate fine-tune messages. While the communication topology and installation architecture are not a measure contribution of this research, this section shows the basic components of HANs that affect our perspective toward preserving the privacy of the smart sensors as part of such network. Figure 2 shows the general architecture of HAN that consists of connected smart appliances and the gateway, through an internal modem, to the internet.

Given a set of appliances in a smart home, where every appliance is equipped with a special-purpose sensor and maintains a look-up table. The set of sensors that are attached to home appliances are defined as $D = \{d_1, d_2, \dots, d_n\}$. Once an

Table 1 Summary of related work

Research articles	Communications among smart (IoT) devices and appliances	Protecting Sensors' identity theft	Application of data driven solutions to the identity theft problem in smart homes	Smart devices (IoT) Security	Securing data transmission
Al Ridhawi et al. (2020)	✓	X	X	✓	X
Zamil et al. (2019)	✓	X	X		X
Aloqaily et al. (2019)	✓	X	X	✓	X
Eschenauer and Gligor (2002)	✓	X	X	X	X
Al Ridhawi and Koth (2018)	✓	X	X	✓	X
Hoang and Pishva (2015)	X	X	X	✓	X
Kumar et al. (2015)	X	✓	X	✓	X
Santoso and Yun (2015)	X	X	X	X	✓
Ayday and Rajagopal (2013)	X	✓	X	X	X
Logue et al. (2013)	X	X	X	✓	✓
Nadi et al. (2019).	X	X	✓	X	X
Al Zamil and Betin Can (2011)	X	X	✓	X	X
Zuhairy et al. (2018)	X	X	✓	X	X

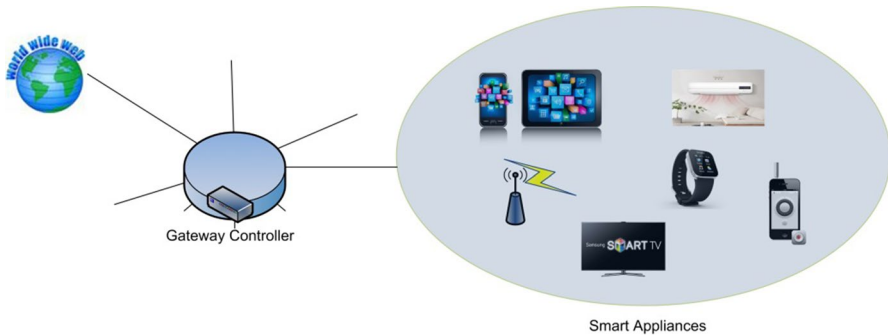


Fig. 2 Smart area network (HAN)

appliance joins the network, the gateway controller assigns a unique number that uniquely identifies each appliance in the network (network address). The set of identifiers is defined as follows: $ID = \{id_1, id_2, id_n\}$. Furthermore, during the initialization phase, an appliance constructs a look-up table that maintains information about other home appliances, type of messages, and the responding protocol.

As shown in Fig. 2, two communication schemes apply to this network: internal and external. Internal communications follow the IoT standard communications, in which devices are communicated directly (P2P). On the other hand, external communications are performed via the HAN gateway to connect the home network with external ones (the Internet). Communication messages, in this network, are following a predefined protocol. A message carries on information from the sender that is well-known to the receiver. For instance, a sensor broadcasts the temperature of the room. Once this message has been received by the air-conditioner, it will cool or heat the room, or even turn the power off. Accordingly, every device has its own reaction against broadcasted messages at the smart home.

The semantic of the data attached in each communicating message is totally dependent on the source sensor; i.e. the device ID. In other words, the receiver cannot interpret the message correctly and pick-up the appropriate reaction without knowing the ID. The traditional privacy preserving model relies on encrypting the communicating messages using the well-known public-private keys framework. Since every appliance has a public key and its own private one, it can easily identify the message content. Formally, given a message m_{ID} and public key C_k , the sender applies $Enc(m_{ID}, C_k)$ to encrypt the message, while the receiver applies $Dec(m_{ID}, C_k)$ to reverse the encryption. Figure 3 simplifies the secured communication paradigm in HANs.

3.2 Sensor Identity Attack (Use-Case)

Each home appliance has unique characteristics that lead to identifying its functionality. Data analysis and machine learning tools make it easy to analyze the patterns of communicating data to identify their semantics, which raise the vulnerability that attackers may exploit it. Although encrypting communicating

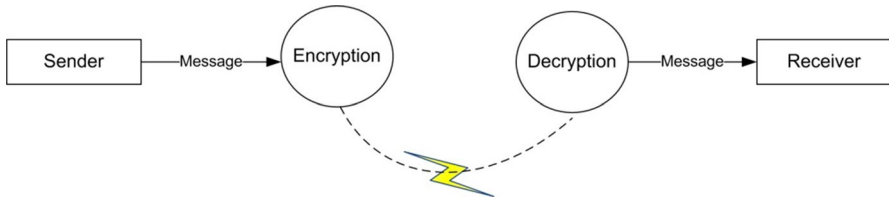


Fig. 3 Secured framework for communicated messages

messages helps preserve the data privacy, it will not prevent an attacker from learning their identity as the source location of these sensors cannot be hidden. Sensor identity, in this context, is defined as the distinguished job that a sensor provides when it is actively functioning. For instance, it is not that hard for attackers to predict the identity of home appliances in the kitchen. It is a matter of time to identify every device and its associated sensors once the attacker knows the source of messages. In this case, an intruder can collect useful information that is not directly related to exchanged data; it is enough for the attacker to know that there are some appliances have turned-on in the kitchen in a specific time to conclude that the resident is at home or the other rooms are empty. Figure 4 explains the way an intruder can attack the network through the cordless communication scheme.

To formulate the use-case, we are looking for a technique to frequently change sensors' identity overtime. While, at the same time, all other home appliances are still able to identify each other. In other words, we are trying to benefit from the fact that internal communications can be highly controlled and manipulated so that external intruders cannot understand their paradigms.

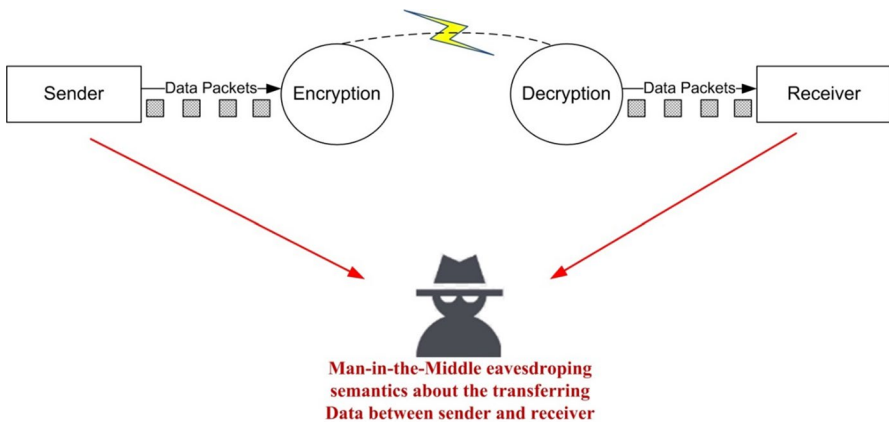


Fig. 4 Identity attacks

3.3 Identity Preserving Protocol

This section explains our proposed 3-phase technique for setting up the environment and harmonizes the implementation of an identity preserving algorithm among other communication components. Mainly, the three phases are: initialization, concealment, and communication. The following subsections provide a detailed algorithmic description of each phase.

3.3.1 Initialization Phase

This phase presents the initial actions to set up the sensors into the home area network. The process involves defining the essential parameters and actions that should be executed to deliver the required services. Algorithm 1 defines a queue that maintains information about every sensor in the network. This queue holds the actual information about every sensor and is required to be updated every specific amount of time. Line 2 defines the private key of the current sensor as a function of the public key and certificate. It follows the traditional IEEE standards of defining public-private keys. Line 3 defines the time variable; which is associated with the current sensor and does not require to be synchronized with similar variables in the sensors network. For these lines (2, 3, and 4), the time complexity that is required to execute them is $O(1)$.

Line 4 defines a join function that allows the sensor to join the home network. The joining process involves three actions. The first one is to set up networking addresses and a dedicated connection to the network gateway. The second action is to collect information about neighboring sensors and push them into the sensor's queue. The third action is to establish a permanent control connection in which commands and acknowledgments are passing through. The time complexity to run these actions is $O(N)$, where N is the total number of sensors in the home area network.

Initialization Phase	
Input:	<i>def</i> Q_{ID} : Queue of existing sensors C_k : Certified Key $S = \{s_1, s_2, \dots, s_n\}$: home sensors
Output:	Acknowledgements
Begin:	
1	<i>While</i> (s_k not Join)
2	<i>init</i> Private_Key = Prv(Publi $c_{Key}(C_k)$)
3	Time : $t_{ID} = 0$
4	Join(H)
5	$t_{ID} = \text{random}(0, \text{threshold})$
6	Push($t, Q_{ID} []$)
7	$ID = \text{Enc}(t, C_k)$
8	Broadcast(Ack(Enc(ID, C_k)))
9	loop
End	

Algorithm 1: The initialization phase as sensors join the home area network

To protect sensor identity, the initialization algorithm uses a random integer at each time. Lines 5, 6, and 7 are responsible for picking up a random integer that will be used later to encrypt the identity of the sensor. Consequently, this will hide the pattern of the sensor's activities. The time complexity that is required to execute both lines is $O(1)$. The threshold variable defines the range of numbers that can be assigned to the time variable. The higher the value of the threshold variable implies a lower chance to recognize the identity of the sensor. However, it plays a significant role in the performance of the communication, since the correlation between the threshold value and the communication performance is strongly negative. Line 8 broadcasts an acknowledgement packet confirming the new arrived information to other neighboring sensors with time complexity $O(1)$.

3.3.2 Concealment Phase

During this phase, every sensor is used to hide its functionality, broadcast, and receive others' information. This thread is executed concurrently and is strictly dependent on the environment settings. In Algorithm 2, Lines 1 to 3 are repeatedly used to generate a random time value, encrypt sensor's identity, and broadcast information to other neighboring sensors.

The while-loop in this algorithmic pseudo code is used to receive identities' acknowledgements from other sensors, decrypt them, and update the sensor's queue. Furthermore, the sensor acknowledges every packet to guarantee reliability. The time complexity (for N sensors) that is required to execute this thread is $O(N^2)$, since the function 'receive' requires N times and the function 'push' requires N times as well.

Concealment Phase	
Input:	<i>threshold</i>
	C_k : Certified Key
Output:	t_{ID} , Acknowledgement s
Begin:	
1	$t_{ID} = \text{random}(0, \text{threshold})$
2	$ID = \text{Enc}(t_{ID}, C_k)$
3	$\text{Broadcast}(\text{Ack}(\text{Enc}(ID, C_k)))$
4	$\text{while}(\text{receive}(\text{Ack}(t_{ID})) \{$
5	$t_{ID} = \text{Dec}(t_{ID}, C_k)$
6	$\text{Push}(t_{ID}, Q_{ID} [])$
7	$\text{send}(\text{Ack}(t_{ID})) \}$
8	<i>loop</i>
End	

Algorithm 2: Time-based Scrambling of Sensors IDs

3.3.3 Communication Phase

The main communication module defines how sensors execute the whole three phases sequentially and/or concurrently. Further, it consists of three threads that are responsible for sensing, receiving, and synchronizing the maintenance of system queues. As shown in Algorithm 3, at the very beginning, the module is initializing the environment and setting the time-threshold value. This value is globally accepted. The first thread is responsible for sensing the environment and reporting back the results in an encrypted message. The receiving thread is responsible for waiting until other neighboring sensors send their information. The synchronization thread, finally, maintains the system queues.

Communication Phase	
Input:	<i>threshold</i> C_k : Certified Key
Output:	t_{ID} , Acknowledgements
Begin:	
1	<i>initialization</i> (<i>threshold</i> , C_k , S)
2	if $t - 1_{ID} = t_{ID}$)
3	$t_{ID} = \text{ranodm}(0, \text{threshold})$
4	<i>Thread Sensing</i> ::
5	begin
6	<i>Sense</i> _{ID} (m)
7	$m = \text{Enc}(m, C_k)$
8	<i>Unicast</i> (m, Q_{ID})
9	end
10	<i>Thread Receiving</i> ::
11	begin
12	<i>receive</i> (m)
13	$m = \text{Dec}(m, C_k)$
14	<i>sender</i> = <i>get</i> _{ID} (t_{ID}, m)
15	<i>get_data</i> (m)
16	<i>process_data</i> (m)
17	end
18	<i>Thread Synchronize</i> ::
19	begin
20	<i>Scramble</i> _{IDS} (Q_{ID})
21	<i>set_queue</i> (Q_{ID})
22	end
End	

Algorithm 3: Main Communication algorithm

The concurrent implementation of these three threads raises the problem of interleaving, in which a thread may access the same resource while it is used by another one. Threads, in this context, are triggered according to the sensing environment. For instance, the sensors are set to read the environment every amount of time. Each

sensor has its own setting. On the other hand, once a sensor reports a change of its identity, others should synchronize their queues. Thus, the interleaving among threads may often occur during the lifetime of the network.

3.4 Model Verification

Each phase of the proposed protocol is designed to be implemented as an embedded code in every cooperative sensor. Therefore, sensors may run similar threads simultaneously, which raises the problem of how main parameters can be synchronized. For instance, queues' values must be known for other sensors at the end of the sensing phase.

Since concurrency is a major issue, in this technique, model verification is necessary for ensuring that the system handles interleaving among its running threads. For this reason, we use model checking as a tool to ensure that the proposed technique achieves the overall system specifications in terms of temporal aspects. As described in [35, 36], model checking verification is a tool that can detect whether a model conforms to a specific requirement specification or not. It cannot tell that the system is error free, but, on the other hand, it can tell with a counter example that the system violates a given requirement.

To involve model checking in our design, we use UPPAAL [37] to formally describe the interactions among threads in our proposed technique. UPPAL defines the system as a set of states and transitions.

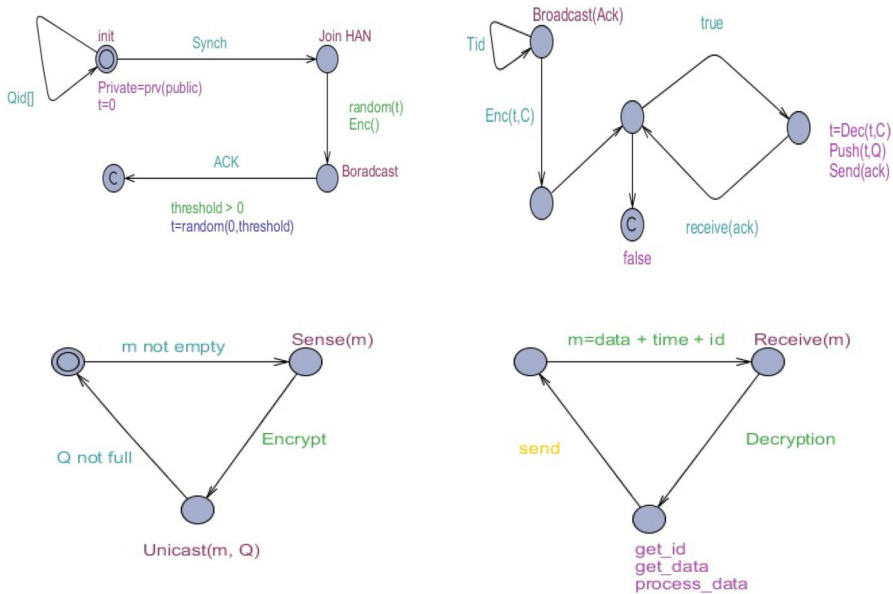


Fig. 5 UPPAL models of systems threads

Table 2 System temporal specifications

#	Requirement	Temporal specification
1	Mutex	$A[] (i.id) A[] (j.id) Q[i] \& Q[j] \text{ imply } i == j$
2	Deadlock free	$A[] \text{ no deadlock}$
3	Starvation free	$\text{Sense}(i).req \rightarrow \text{Sense}(i).wait$
4	Liveness requirement	$\text{Sense}(i).send \text{ imply } \text{sense}(j).ack$
5	Active sensing	$A[] (i.id) \text{ imply } E[] \text{ Sense.Synch}$
6	Reachable property	$A[] (i.id) E[] \text{ Sense}(m) \text{ imply } \text{Sense.get_data}$
7	Reachability property	$A[] (i.id) E[] \text{ Sense}(m) \text{ imply } \text{Sense.Enc}$
8	Reachability property	$A[] (i.id) E[] \text{ Sense}(m) \text{ imply } \text{Sense.Dec}$
9	Safety property	$A[] \text{ Sense.init imply Sense.receive}$
10	Safety property	$A[] \text{ Threshold } <> 0 \& t \leq \text{Threshold}$

concurrently and tests their verifiability in terms of a given set of temporal specifications. Figure 5 shows the finite-state automata for every thread.

For this reason, we designed 10-specifications, shown in Table 2, to verify that our proposed technique maintains reliability in terms of concurrency, resource sharing, and safety. Our modeling of the whole technique was controlled by checking the temporal specifications of this system. Eventually, the proposed technique proved its ability to maintain concurrency during execution. The properties were used are: Mutex, Deadlock, Starvation, Liveness, Activation, Reachability, and Safety.

4 Experiments and Results

This section describes and explains a set of experiments that have been conducted to evaluate the performance of the proposed technique. There are mainly two types of experiments: performance evaluation and sensitivity analysis. The first experiment is focusing on testing how our proposed technique enhances the misclassification rates based on the assumption that sensors identities would be hard to be identified if existing data features do not clearly point to them. The second experiment, on the other hand, shows how our proposed technique is sensitive to the threshold value range and queue size.

Table 3 Description of datasets

Dataset name	Size	# Activities	# Sensors
Tulum	1,048,576	16	20
Cairo	158,409	10	25
Milan	433,665	15	33
Kyoto	64,250	5	25

To perform such experiments, we applied our technique on four well-known datasets that have been collected from smart homes in four different countries obtained from CASAS project [38]: France-Tulum, Egypt-Cairo, Italy-Milan, and Japan-Kyoto. Every dataset comprises instances covering a finite set of activities. Actions are generated using motion, temperature, or detection sensors. Table 3 briefly describes the datasets. Note that every dataset has an attached map that shows the location of sensors, which can be interpreted as the location where a specific action triggers.

Instances in these datasets represent the daily activities of a single resident; they have been collected and labeled manually, so that experiments could be supervised using already annotated instances. Note, only active sensors have been used, because some sensors were not active in the testbed during the data collection process. Figure 6 shows the planted sensors in the Milano smart home.

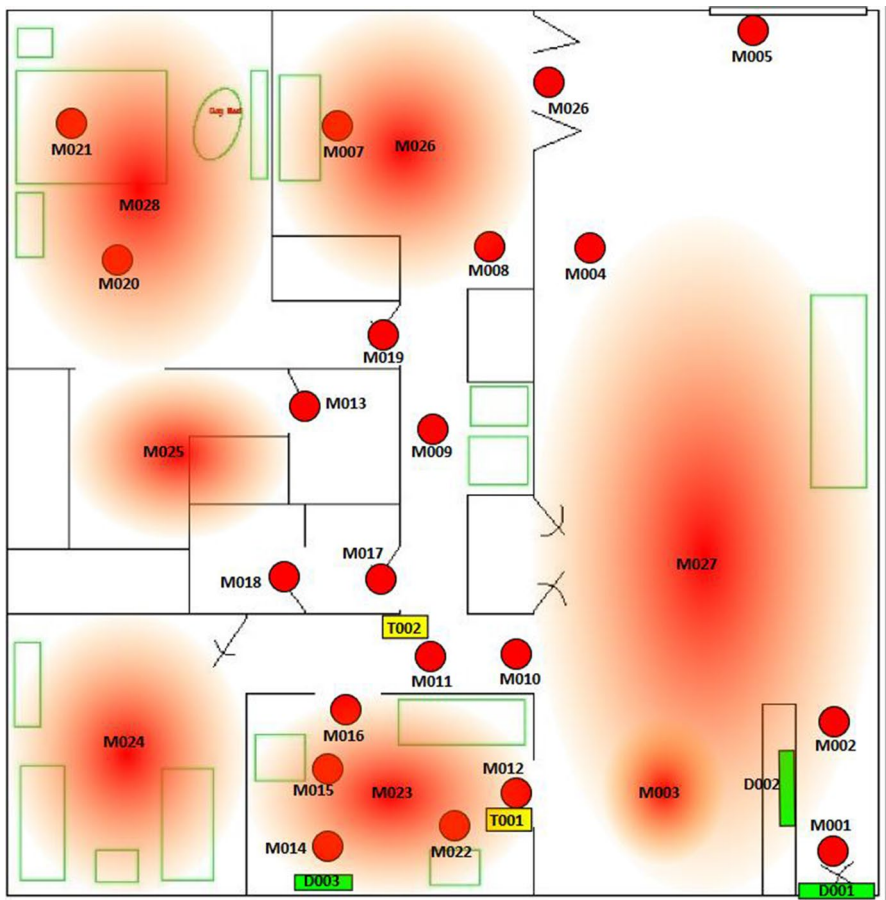


Fig. 6 The distribution of the sensors in the smart home (Milano-Italy)

Moreover, we also conducted the preprocessing task to assemble all actions that are related to each activity; since each activity consists of a set of actions. This preprocessing task is known as data segmentation, in which each segment is a record that aggregate information from multiple actions to represent a specific activity.

Finally, all experiments have been conducted using Python 2.2 and the machine specifications are as follows: Intel(R) Core(TM) i5-4200 M CPU@2.50 GHz, Physical RAM4.00 GB, Windows 7 Professional Edition. Moreover, data visualization and results analysis have been developed using MATLAB.

4.1 Performance Analysis

In this section, we present our experiments to measure the performance of the proposed technique. We designed the experiment as a binary classification problem to measure the ability of different classifiers to identify the sensors' identities (class labels). For this reason, we chose the misclassification rate (MC) as a measure of the classifiers' disability for recognition (i.e. the higher the MC value implies the higher the disability of a classifier to identify the sensor identity). The misclassification rate is defined as follows:

$$MC_{rate} = (FP + FN)/(TP + FN + TN + FP) \quad (2)$$

Where TP is the true positive value, FN is the false negative value, FP is the false positive value, and TN is the true negative value. The misclassification rate shows how our approach can confuse these classifiers (intruders) in recognizing the sensors' identities. First, we ran the classifiers on the original data and measure the misclassification rates. Next, we applied our technique on the original data at different threshold values and feed the classifiers with the new datasets to measure the misclassification rates after applying our proposed technique. To implement this experiment, we chose six-classification algorithms: K-Nearest-Neighbor (KNN), Hidden Markov Model (HMM), Support-Vector Machine (SVM), Decision Tree base classifier (J48), Naïve Byes (NB), and Conditional Random Field (CRF). These algorithms were widely applied in similar research such as [1, 5, 16, 17]. Moreover, we used to split every dataset into training and testing sets (2:1 ratio) in which both of them were fixed; to prevent testing the classifiers on different testing sets.

As shown in Table 4, the misclassification rates have been increased significantly as the threshold value increased. This indicates that at high threshold time interval, the proposed technique was able to achieve high misclassification rates, which implies its ability to decrease the detection rates. In other words, an intruder who owns the features or the sensors data has a lower chance to recognize the source identity.

We performed additional experiment to ensure that the overall performance of the proposed technique satisfies our goal; minimizing the detection rate of given classifiers. In this experiment, we used to apply the loss rate formulas to provide an indicator of how it is complex to identify an entity using existing features. Indeed, the higher the loss rate values the better for protecting the class labels (sensor identity).

Table 4 Misclassification Rates of Data Classifiers at different threshold values

Algo.	Dataset	<i>MC org</i>	<i>MC</i> <i>Th = 100</i>	<i>MC</i> <i>Th = 200</i>	<i>MC</i> <i>Th = 300</i>	<i>MC</i> <i>Th = 500</i>	<i>MC</i> <i>Th = 700</i>	<i>MC</i> <i>Th = 1000</i>
KNN	Tulum	0.40	0.50	0.62	0.78	0.85	0.94	0.95
	Cairo	0.33	0.41	0.51	0.64	0.71	0.85	0.93
	Milan	0.38	0.47	0.59	0.73	0.81	0.89	0.97
	Kyoto	0.27	0.34	0.43	0.53	0.58	0.70	0.77
HMM	Tulum	0.29	0.36	0.45	0.56	0.62	0.74	0.82
	Cairo	0.26	0.33	0.41	0.51	0.56	0.68	0.96
	Milan	0.28	0.35	0.44	0.55	0.60	0.72	0.79
	Kyoto	0.21	0.26	0.32	0.40	0.61	0.73	0.80
SVM	Tulum	0.18	0.22	0.28	0.34	0.52	0.62	0.88
	Cairo	0.14	0.18	0.23	0.28	0.42	0.76	0.84
	Milan	0.15	0.19	0.24	0.30	0.45	0.80	0.88
	Kyoto	0.13	0.16	0.20	0.26	0.38	0.69	0.98
J48	Tulum	0.33	0.41	0.51	0.64	0.70	0.84	0.93
	Cairo	0.29	0.36	0.45	0.56	0.62	0.74	0.82
	Milan	0.31	0.39	0.48	0.60	0.66	0.79	0.87
	Kyoto	0.27	0.33	0.41	0.52	0.57	0.68	0.97
NB	Tulum	0.23	0.29	0.36	0.45	0.68	0.82	0.90
	Cairo	0.09	0.11	0.14	0.18	0.27	0.48	0.68
	Milan	0.18	0.23	0.28	0.35	0.53	0.64	0.90
	Kyoto	0.11	0.14	0.17	0.21	0.32	0.58	0.82
CRF	Tulum	0.26	0.32	0.40	0.50	0.55	0.66	0.94
	Cairo	0.25	0.31	0.39	0.48	0.72	0.87	0.96
	Milan	0.19	0.24	0.29	0.37	0.55	0.66	0.94
	Kyoto	0.09	0.11	0.14	0.17	0.26	0.46	0.66

The loss rate is defined as the square root of the TP complement and the FP value as follows:

$$Loss = \sqrt{([1 - TP]^2 + [FP]^2)} \quad (3)$$

Note that the value of the loss rate is greater than or equal zero; the maximum might exceed 1 as it depends on the TP value. The lower the TP value the more chance of loss rate to exceed 1.

As shown in Table 5, for all classification algorithms and datasets, the loss rates have been increased significantly. This implies that applying the proposed technique significantly complicated the process of recognizing the class labels (sensors' identities). On the other hand, we noticed that the loss rate affected by the shape of the datasets; the dataset with more activities and sensors achieves higher loss rates as compared to others.

Table 5 Average loss rates at different threshold values

Algorithm	Dataset	<i>Loss in Original Data</i>	<i>Loss in Converted Data</i>
KNN	Tulum	0.716	0.920
	Cairo	0.678	0.993
	Milan	0.707	0.985
	Kyoto	0.646	0.936
HMM	Tulum	0.654	0.963
	Cairo	0.642	0.941
	Milan	0.652	0.995
	Kyoto	0.610	1.018
SVM	Tulum	0.592	0.968
	Cairo	0.575	0.971
	Milan	0.579	1.003
	Kyoto	0.568	1.063
J48	Tulum	0.677	0.935
	Cairo	0.655	0.984
	Milan	0.665	0.964
	Kyoto	0.641	0.978
NB	Tulum	0.623	0.991
	Cairo	0.547	1.043
	Milan	0.595	1.016
	Kyoto	0.557	1.034
CRF	Tulum	0.637	0.925
	Cairo	0.631	0.993
	Milan	0.599	1.040
	Kyoto	0.545	1.034

Finally, we conducted statistical testing to measure the significance of the differences that have been achieved after applying the proposed 3-phase algorithm. At 99% confidence, the differences (enhancement and loss rates) that have been achieved were statistically significant at $p < 0.001$.

4.2 Sensitivity Analysis

This section discusses how the threshold value affects the overall performance of the proposed technique in terms of time cost, detection rates, enhancements, and its effect on each classifier. The proposed technique has benefited from the threshold parameter to expand the range of scrambled values that could be assigned as sensors used to encrypt their identities over a time ranges from zero to the threshold value. For this reason, we consider how this parameter affects the physical running time in terms of milliseconds. As shown in Fig. 7, the actual time that is required to execute the proposed algorithms increases exponentially until a certain threshold value, then

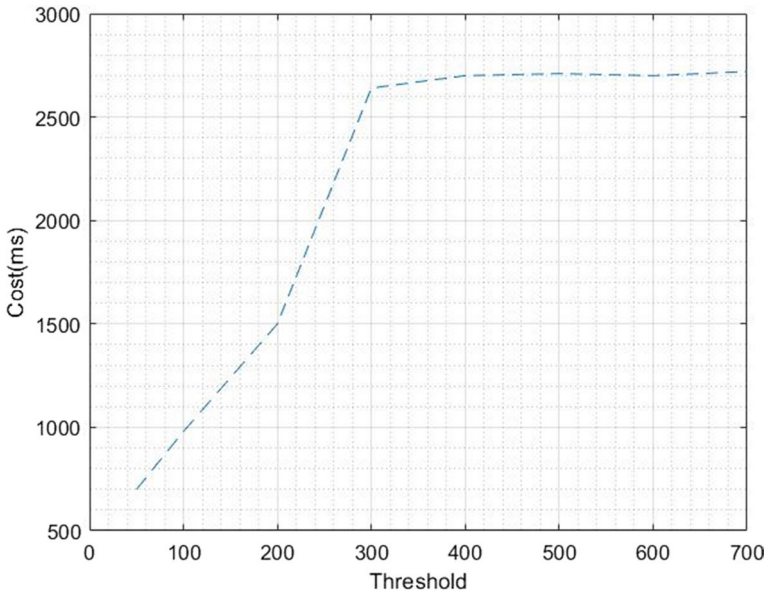


Fig. 7 Sensitivity of time cost to threshold values

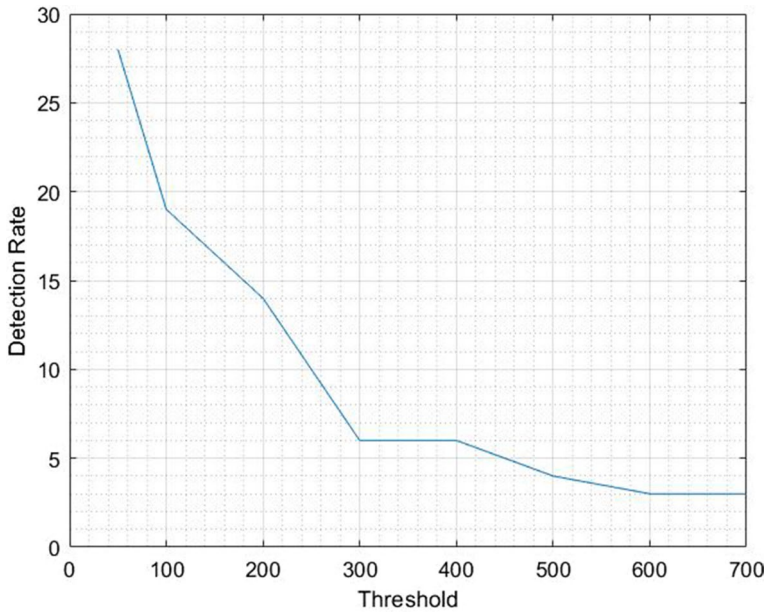


Fig. 8 Sensitivity of detection cost to threshold values

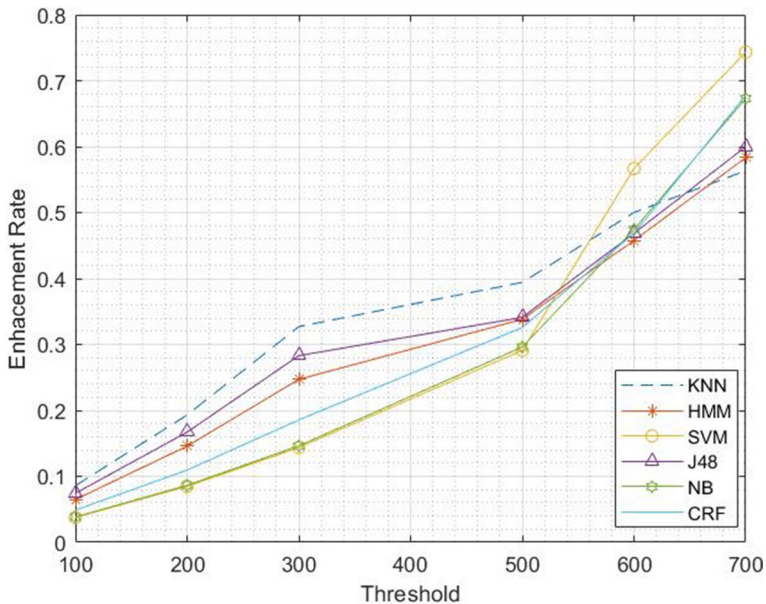


Fig. 9 The effect of threshold value on the enhancement rates

it became stable as the threshold value increases. This implies that at high threshold value, the required time to execute the 3-phase algorithm is linear.

On the other hand, we noticed that the effect of the low threshold value on the detection rates is low but enhanced significantly as the threshold value increased. Figure 8 shows that at average threshold value, the detection rate decreased significantly; making the protection from sensor identity theft efficient and effective.

In addition, we investigated the enhancement rates that have been achieved by each classifier to notice their behavior against different threshold values. According to Fig. 9, all classifiers achieved enhancements as the threshold value increased. The SVM classifier achieved a higher enhancement at the high threshold value, while KNN achieved a higher enhancement at the lower threshold value.

Finally, we investigated how the datasets specifications (such as number of activities, number of sensors in the home area network, etc.) affect the performance of different classifiers by considering different threshold values. This experiment is important as it gives valuable knowledge on how to pick up the best threshold value for specific smart home settings.

As shown in Fig. 10, KNN tends to perform better as the size of the dataset is smaller. The misclassification rate of Kyoto dataset (smallest size) was lower among other datasets, while it was the highest for Tulum (larger size). On the other hand, HMM (Fig. 11) showed no clear pattern on how it performs against the dataset's specifications. We noticed that most datasets achieved similar MC rates at the high threshold value.

As it was the lowest performance in terms of MC rate, SVM (Fig. 12) tends not to be affected by the specifications of the datasets. We believe that converting the features,

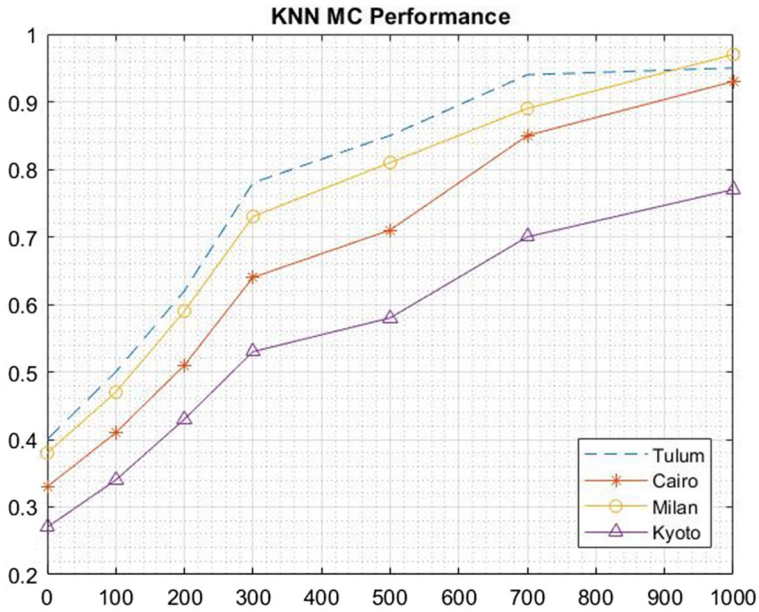


Fig. 10 KNN MC performance

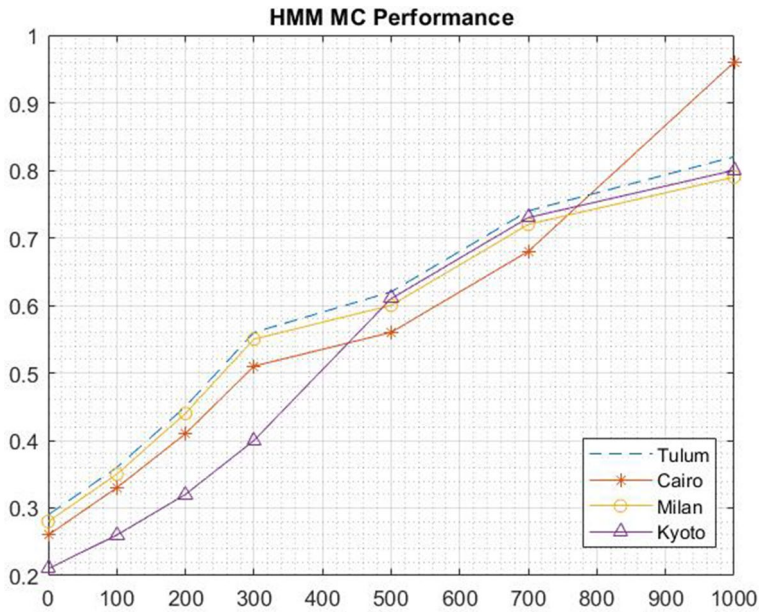


Fig. 11 HMM MC performance

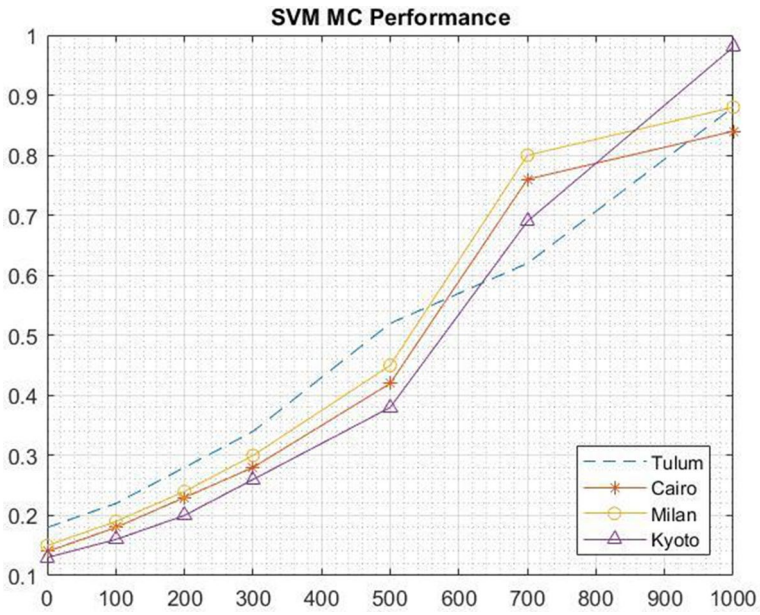


Fig. 12 SVM MC performance

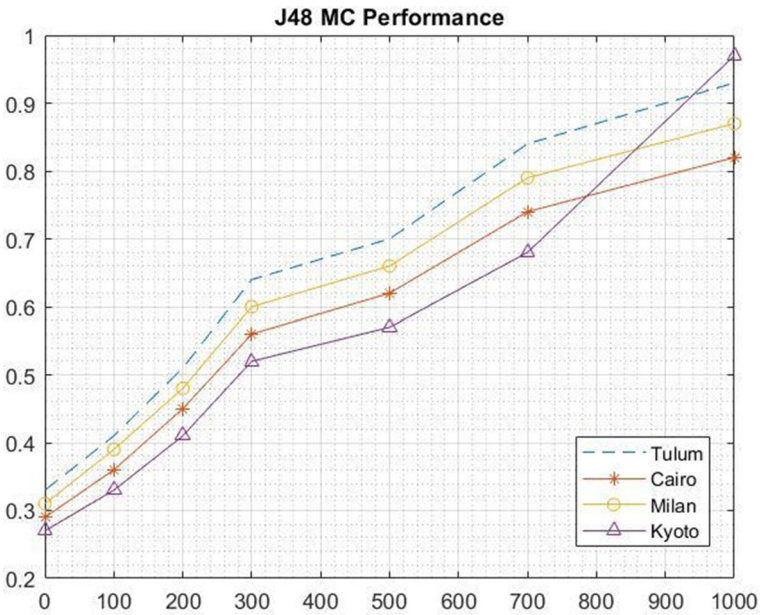


Fig. 13 J48 MC performance

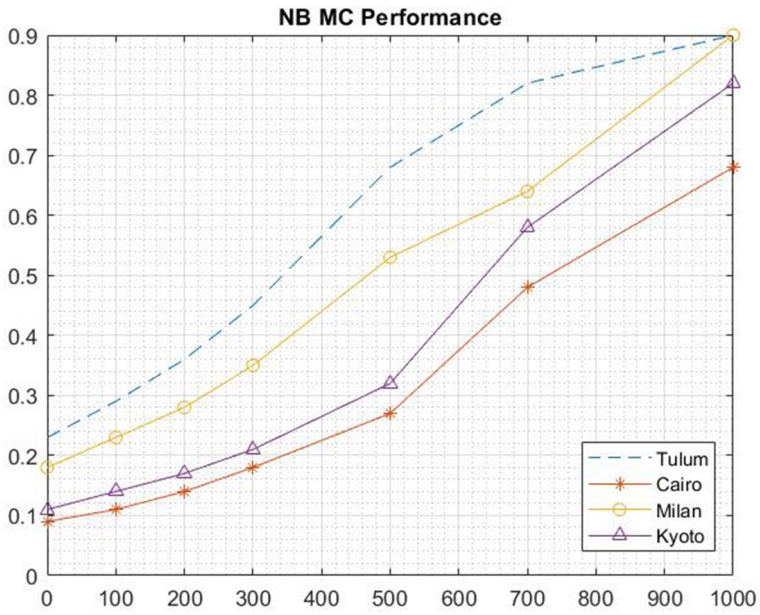


Fig. 14 NB MC performance

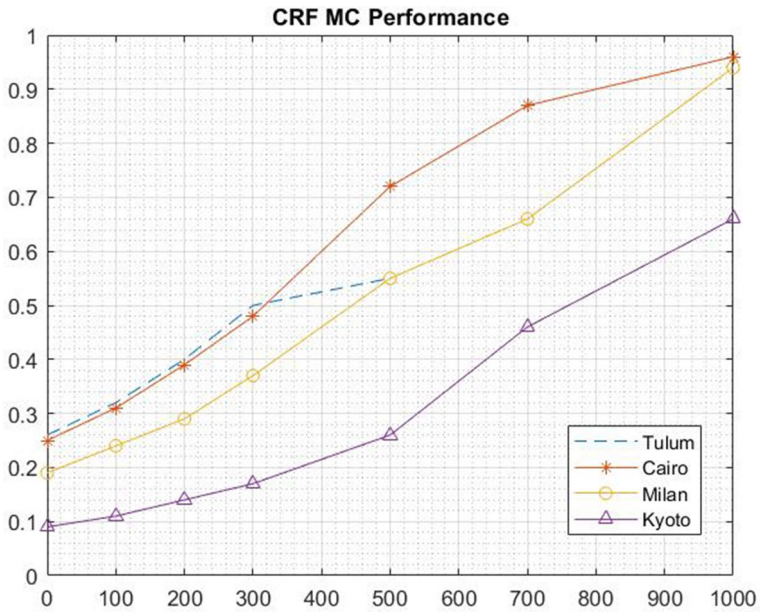


Fig. 15 CRF MC performance

during the preprocessing, to formulate vectors of frequencies was the reason behind these results. Figure 13 shows that J48 achieved high detection as the dataset size is small; except in the case of Kyoto dataset at a high threshold value.

Figure 14 explains the performance of NB classifier as it appears to be affected by the dataset's specifications for the largest datasets (Tulum and Milan), while the effect was opposite at the lowest ones (Cairo and Kyoto). CRF (Fig. 15) shows no pattern on how the classifier is affecting the dataset's specifications, since the performance fluctuated as the threshold values increased.

Accordingly, the results indicated that there are not enough proofs to conclude the effect of datasets specifications on the performance of the classifiers; except in few cases that cannot be generalized. Therefore, the proposed technique was able to perform well in protecting the sensors identities in smart home regardless of how the collected features are processed.

5 Conclusion

This paper introduces a novel approach to protect the sensors' identities of smart homes. The proposed model aims to increase the security level of smart home devices and appliances, which reduces the risk of identifying sensors' functionalities. The proposed approach applied three-phase technique that manages a synchronized space among connected sensors and prevents the identification of sensors from outsiders. Furthermore, the proposed solution preserved the linearity of time required to manage the protection of the home network sensors' identities. The empirical results showed significant performance enhancements in protecting sensors' identities in smart home area networks. Additionally, the experiments highlighted the impact of the threshold value that defines the time interval for each sensor on the model performance.

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by YA, AARB, MALZ and SS. The first draft of the manuscript was written by YA and MAZ and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Not Applicable.

Compliance with Ethical Standards

Conflict of interest No conflict of interest neither competing interests.

References

1. Zamil, M.G.A.: Multimodal daily activity recognition in smart homes. In: 2019 6th International Conference on Control, Decision and Information Technologies (CoDIT), pp. 922–927. IEEE (2019)

2. Iqbal, M.A., Asrafuzzaman, S.K., Arifin, M.M., Hossain, S.A.: Smart home appliance control system for physically disabled people using kinect and X10. In: 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), pp. 891–896. IEEE (2016)
3. Di Francesco, M., Anastasi, G., Conti, M., Das, S.K., Neri, V.: Reliability and energy-efficiency in IEEE 802.15. 4/ZigBee sensor networks: an adaptive and cross-layer approach. *IEEE J. Sel. Areas Commun.* **29**(8), 1508–1524 (2011)
4. Al Ridhawi, I., Otoum, S., Aloqaily, M., Boukerche, A.: Generalizing AI: challenges and opportunities for plug and play AI solutions. *IEEE Netw.* (2020). <https://doi.org/10.1109/MNET.011.2000371>
5. Al Zamil, M.G., Samarah, S., Rawashdeh, M., Hossain, M.S., Alhamid, M.F., Guizani, M., Alnu-sair, A.: False-alarm detection in the fog-based internet of connected vehicles. *IEEE Trans. Veh. Technol.* **68**(7), 7035–7044 (2019)
6. Naoui, S., Elhdhili, M.E., Saidane, L.A.: Lightweight and secure password based smart home authentication protocol: LSP-SHAP. *J. Netw. Syst. Manag.* **27**(4), 1020–1042 (2019)
7. Kumar, P., Braeken, A., Gurtov, A., Inatti, J., Ha, P.H.: Anonymous secure framework in connected smart home environments. *IEEE Trans. Inf. Forensics Secur.* **12**(4), 968–979 (2017)
8. Qiu, M., Gai, K., Xiong, Z.: Privacy-preserving wireless communications using bipartite matching in social big data. *Future Gener. Comput. Syst.* **87**, 772–781 (2018)
9. Gai, K., Choo, K.K.R., Qiu, M., Zhu, L.: Privacy-preserving content-oriented wireless communication in internet-of-things. *IEEE Internet Things J.* **5**(4), 3059–3067 (2018)
10. Wu, F., Xu, L., Kumari, S., Li, X.: A privacy-preserving and provable user authentication scheme for wireless sensor networks based on internet of things security. *J. Ambient Intell. Humaniz. Comput.* **8**(1), 101–116 (2017)
11. Yi, X., Willemson, J., Nait-Abdesselam, F.: Privacy-preserving wireless medical sensor network. In: 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, pp. 118–125. IEEE (2013)
12. Alami, A., Benhlima, L., Bah, S.: An overview of privacy preserving techniques in smart home wireless sensor networks. In: 2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA), pp. 1–4. IEEE (2015)
13. Li, C.T., Lee, C.C.: A novel user authentication and privacy preserving scheme with smart cards for wireless communications. *Math. Comput. Model.* **55**(1–2), 35–44 (2012)
14. Lekshmy, P.L., Rahiman, M.A.: Hybrid approach to speed-up the privacy preserving kernel K-means clustering and its application in social distributed environment. *J. Netw. Syst. Manag.* **28**(2), 398–422 (2020)
15. Al Ridhawi, I., Otoum, S., Aloqaily, M., Jararweh, Y., Baker, T.: Providing secure and reliable communication for next generation networks in smart cities. *Sustain. Cities Soc.* **56**, 102080 (2020)
16. Samarah, S., Zamil, M.G.A., Rawashdeh, M., Hossain, M.S., Muhammad, G., Alamri, A.: Transferring activity recognition models in FOG computing architecture. *J. Parallel Distrib. Comput.* **122**, 122–130 (2018)
17. Zamil, M.G.A., Samarah, S., Rawashdeh, M., Karime, A., Hossain, M.S.: Multimedia-oriented action recognition in Smart City-based IoT using multilayer perceptron. *Multimed. Tools Appl.* **78**(21), 30315–30329 (2019)
18. Aloqaily, M., Otoum, S., Al Ridhawi, I., Jararweh, Y.: An intrusion detection system for connected vehicles in smart cities. *Ad Hoc Netw.* **90**, 101842 (2019)
19. Dou, N., Mei, Y., Yanjuan, Z., Yan, Z.: The networking technology within smart home system-Zig-Bee technology. In: 2009 International Forum on Computer Science-Technology and Applications, vol. 2, pp. 29–33. IEEE (2009)
20. Liu, B.: Wireless broadband networks in complex mine environment. In: International Conference on Information Computing and Applications, pp. 82–90. Springer, Berlin (2013)
21. Juels, A.: RFID security and privacy: a research survey. *IEEE J. Sel. Areas Commun.* **24**(2), 381–394 (2006)
22. Rid, T., Buchanan, B.: Attributing cyber attacks. *J. Strateg. Stud.* **38**(1–2), 4–37 (2015)
23. Hachmi, F., Boujenfa, K., Limam, M.: Enhancing the accuracy of intrusion detection systems by reducing the rates of false positives and false negatives through multi-objective optimization. *J. Netw. Syst. Manag.* **27**(1), 93–120 (2019)
24. Eschenauer, L., Gligor, V.D.: A key-management scheme for distributed sensor networks. In: Proceedings of the 9th ACM Conference on Computer and Communications Security, pp. 41–47 (2002)
25. Al Ridhawi, I., Kotb, Y.: A secure service-specific overlay composition model for cloud networks. *Softw. Netw.* **2018**(1), 221–240 (2018)

26. Hoang, N.P., Pishva, D.: A TOR-based anonymous communication approach to secure smart home appliances. In: 2015 17th International Conference on Advanced Communication Technology (ICACT), pp. 517–525. IEEE (2015)
27. Kumar, P., Gurtov, A., Inatti, J., Ylianttila, M., Sain, M.: Lightweight and secure session-key establishment scheme in smart home environments. *IEEE Sens. J.* **16**(1), 254–264 (2015)
28. Santoso, F.K., Vun, N.C.: Securing IoT for smart home system. In: 2015 International Symposium on Consumer Electronics (ISCE), pp. 1–2. IEEE (2015)
29. Ayday, E., Rajagopal, S.: Secure device authentication mechanisms for the smart grid-enabled home area networks (No. REP_WORK) (2013)
30. Logue, J.D., Supramaniam, S., Hardison, O.B., Luxemburg, J.A.: U.S. Patent No. 8,539,567. U.S. Patent and Trademark Office, Washington, DC (2013)
31. Nadi, R.A., Zamil, M.G.A.: A profile based data segmentation for in-home activity recognition. *Int. J. Sens. Netw.* **29**(1), 28–37 (2019)
32. Al Zamil, M.G., Betin Can, A.: A model based on multi-features to enhance healthcare and medical document retrieval. *Inform. Health Soc. Care* **36**(2), 100–115 (2011)
33. Zuhairy, R.M., Al Zamil, M.G.: Energy-efficient load balancing in wireless sensor network: an application of multinomial regression analysis. *Int. J. Distrib. Sens. Netw.* **14**(3), 1550147718764641 (2018)
34. Nasrallah, A., Balasubramanian, V., Thyagaturu, A., Reisslein, M., ElBakoury, H.: Reconfiguration algorithms for high precision communications in time sensitive networks: time-aware shaper configuration with IEEE 802.1 qcc (extended version). arXiv preprint [arXiv:1906.11596](https://arxiv.org/abs/1906.11596) (2019)
35. Clarke Jr., E.M., Grumberg, O., Kroening, D., Peled, D., Veith, H.: Model Checking. MIT Press, Cambridge (2018)
36. Al Zamil, M.G., Samarah, S.: Application of design for verification to smart sensory systems. In: Qatar Foundation Annual Research Conference Proceedings Volume 2014 Issue 1, Vol. 2014, No. 1, p. ITPP0366. Hamad bin Khalifa University Press (HBKU Press) (2014)
37. Larsen, K.G., Pettersson, P., Yi, W.: UPPAAL in a nutshell. *Int. J. Softw. Tools Technol. Transf.* **1**(1–2), 134–152 (1997)
38. <http://casas.wsu.edu/datasets/>. Accessed 20 July 2020

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.


Yazan Alshboul is an assistant professor in Information Technology department and the coordinator of the cybersecurity program at Yarmouk University. He earned his Ph.D in information systems/information assurance and security from Dakota State University in USA in 2017. His research interest is in cybersecurity and data analysis.

Abdel Al Raouf Bsoul is an associate professor at Computer Science Department at Yarmouk University. Dr. Bsoul obtained his Ph.D. degree in Computer Science from Virginia Commonwealth University, Richmond, VA, USA in 2011. Dr. Bsoul's current research interests include biomedical signal and image processing, wireless sensor networks and security, natural language processing, and computational intelligent systems.

Mohammed AL Zamil is a Professor in the department of computer science at Yarmouk University (YU) in Jordan. He obtained his Ph.D degree in Information Systems from Middle East Technical University, Ankara, Turkey (2010). His research interests include Data Analysis and Segmentation for smart systems, Edge-computing, and data-driven security applications.

Samer Samarah is a full professor at Information Technology Department. He obtained his PhD in Computer Science from University of Ottawa, Canada in 2008. Samarah research interests focus on data analysis techniques for smart environments, Internet of Things, and security applications. Samarah has many publications in his area; specifically, Activity Recognition in smart homes.

Authors and Affiliations

Yazan Alshboul¹  · Abdel Al Raof Bsoul²  · Mohammed AL Zamil²  ·
Samer Samarah³ 

Yazan Alshboul
Yazan.shboul@yu.edu.jo

Abdel Al Raof Bsoul
raoofbsoul@yu.edu.jo

Samer Samarah
Samers@yu.edu.jo

- ¹ Department of Information Technology, Yarmouk University, Irbid, Jordan
- ² Department of Computer Science, Yarmouk University, Irbid, Jordan
- ³ Department of Information Systems, Yarmouk University, Irbid, Jordan