Received March 5, 2017, accepted March 15, 2017, date of publication March 24, 2017, date of current version April 24, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2685531

An Efficient Activity Recognition Framework: Toward Privacy-Sensitive Health Data Sensing

SAMER SAMARAH[1], MOHAMMED GH. AL ZAMIL[1], AHMED F. ALEROUD[1], MAJDI RAWASHDEH[2], MOHAMMED F. ALHAMID[3], AND ATIF ALAMRI[3,4]

[1]Department of Computer Information Systems, Yarmouk University, Irbid 21163, Jordan
[2]Department of Management Information System, Princess Sumaya University for Technology, Amman 11941, Jordan
[3]Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[4]Research Chair of Pervasive and Mobile Computing, King Saud University, Riyadh 11543, Saudi Arabia

Corresponding author: S. Samarah (samers@yu.edu.jo)

This work was supported by King Saud University, Deanship of Scientific Research, Research Chair of Pervasive and Mobile Computing.

ABSTRACT Recent advances in wireless sensor networks for ubiquitous health and activity monitoring systems have triggered the possibility of addressing human needs in smart environments through recognizing human real-time activities. While the nature of streams in such networks requires efficient recognition techniques, it is also subject to suspicious inference-based privacy attacks. In this paper, we propose a framework that efficiently recognizes human activities in smart homes based on spatiotemporal mining technique. In addition, we propose a technique to enhance the privacy of the collected human sensed activities using a modified version of micro-aggregation approach. An extensive validation of our framework has been performed on benchmark data sets yielding quite promising results in terms of accuracy and privacy-utility tradeoff.

INDEX TERMS Internet of Things, data mining, data privacy, healthcare, smart home.

I. INTRODUCTION

In the era of IoT-based smart homes, ubiquitous computing devices have incorporated smartness, in the form of interconnected sensors and home appliances, into dwellings for providing variety of services such as security, safety, energy conservation, and healthcare. Advances in wireless communication and web technologies facilitate the remote monitoring of such systems for the purpose of detecting user behaviors and interactions with the IoT smart environment. The huge data collected from smart homes are analyzed as a collection of meaningful features and, then, used to build contextual models of recognized activities. The activities of home residents play a significant role in providing different services, which makes activity recognition (AR) an integral part of inferring contexts in smart homes. Figure 1 shows the architecture of the smart home and how it could be integrated with cloud infrastructure.

According to the World Health Organization [1], the number of people in the age 60 years or older will increase from 900 million to 2 billion between 2015 and 2050, shifting from 12% to 22% of the total global population. With this rapid change, there will be a significant increase in the healthcare cost and a high demand on health-related personnel that will be difficult to afford. A practical solution for this problem is to adopt smart technologies, within elderly's homes, to monitor their medical conditions and, hence, improving the quality of life in assisted living homes [2].

Healthcare monitoring systems are designed based on recognizing activities of humans in their environments depending on a set of sensed events. In fact, an activity is defined as a sequence of events (segment) in some predefined context. However, human behavior is likely to change at home, which complicates the process of detecting behavioral patterns of residents at smart homes. One solution is to make the inference model dynamic by incorporating new behaviors (activities) using an up-to-date training data. Indeed, the nature of smart home systems requires adopting spatial and temporal features to enhance the accuracy of inference models as human behaviors are interpreted using the occurrence time and/or the location in which it take place. Such features are required to temporally align different activities from multiple subjects performing similar behaviors.

Traditional data preprocessing, segmentation, feature extraction and classification techniques [3] are required to be adapted in an environment infrastructure that enforces nontraditional data acquiring mechanism, in which a sensor

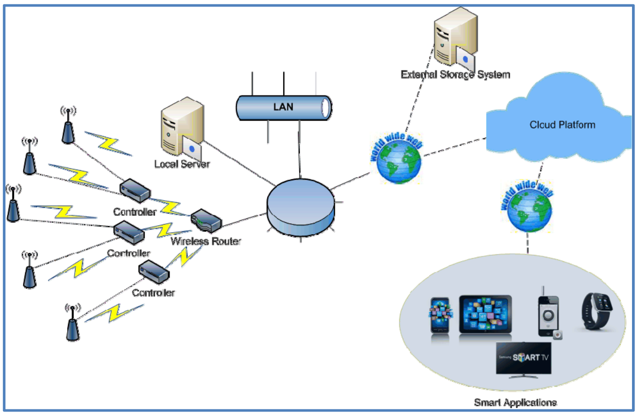


FIGURE 1. **Cloud-based infrastructure for IoT smart homes.**

event might be associated with more than one activity. For this reason, many solutions have been proposed to handle the situation in which overlaps exist during the segmentation task. In smart home environment, temporal and spatial characteristics are key features to perform such task. Figure 2 shows an example of some situations in which segments interleaving with each other's, where "Begin" and "End" indicate the start and the end of an activity.

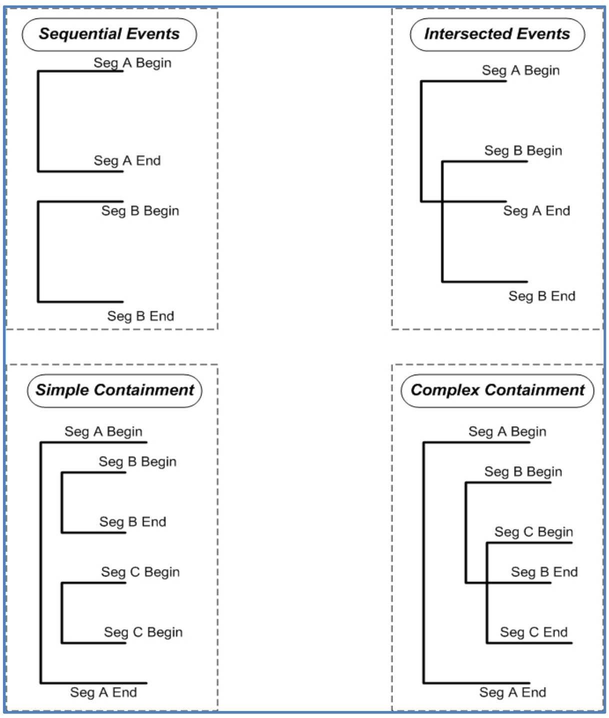


FIGURE 2. **Possible forms of segments in a smart home environment.**

Consider the following situation that represents simple containment segmentation: the patient has started watching the TV, and then she had to go to the bathroom. After this, the patient took his/her diabetes medication. The patient, then, came back and completes her favorite show on TV. Such scenario can be represented as shown in Figure 3:

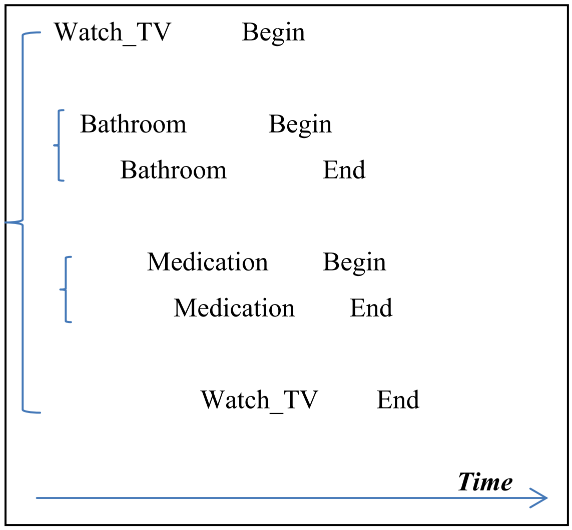


FIGURE 3. **Illustrative example.**

Another important issue that has a direct effect on the performance of data mining techniques, such as classification, is the extraction and representation of the significant features. During the implementation of the experiments, in this work, researchers found that not all entries from sensors are the same in terms of effectiveness; i.e. the effect of the sensor readings on recognizing some specific activities. For this reason, we developed a probabilistic model to weigh the readings from different sensors by taking into account the contextual features of data segments. Therefore, the effect of out-of-context events will be minimized.

On the other hand, sharing of personal data introduces many privacy concerns since several sensitive behaviors can be identified. The challenge is to find versions of health data that can preserve the privacy while still data mining tasks can be applied on it in an efficient manner. With today's high speed networks and health smart applications, it is becoming a necessity to collect, process, disseminate, anonymize and then store data as electronic medical records. From the privacy and security perspectives, two questions have to be answered: 1) is the collected data is private enough so that no one can misuse it? 2) How to examine whether the data is still useful enough to be used by researchers in the health domain if the data is de-identified? Achieving this balance between data usefulness (utility) for researchers and the privacy is a main objective of several e-health providers. It is not a straightforward task to anonymize health related data while keeping it useful for research objective when conventional encryption techniques are used.

This paper proposes an efficient classification technique that resolves the problem of segmenting health data that have been collected from smart homes and utilizes a probabilistic feature extraction technique that weight the readings from the sensory networks. Moreover, we propose a novel generalization technique, which aims to guarantee the utility of the anonymized health data. We proved that the proposed privacy technique works in an effective manner when applied to classify the anonymized health data without a significant degradation in the achieved accuracy.

The rest of this paper is organized as follows: Section II discusses and compares the related research with the proposed one. Section III describes the elementary components of the proposed framework and provides a formal description of preliminary definitions and equations that provide the baseline for the proposed methodology. Section IV describes the experiments that have been conducted to evaluate the output in terms of performance and privacy measurements. Section V discusses the result. Section VI concludes the paper and discusses the scope of future work.

II. LITERATURE REVIEW

Many smart home projects have been proposed aiming to improve the quality of life of residents; specifically for elderly people. Examples of such projects include: GatorTech [4], MavHome [5], PlcaeLab [6], CASAS [7], Sweet-Home [8], USEFIL [9]. These projects have focused on the deployment and the connectivity of the sensor devices for activity monitoring and, thus, many applications and services can be implemented. Activity Recognition techniques have been investigated in the literature based on two categories: visual sensing and data sensing techniques. In the first category, researchers applied image processing algorithms to detect actor's behaviors within a specific environment [10]–[12]. In the second category, data is analyzed using data mining and machine learning techniques to develop prediction models, which are, then, used to detect activities. The research in this paper focuses on the second category, where an effective technique is proposed to recognize smart home sensed activities while keeping high level of privacy in an e-health environment.

The use of wearable devices in healthcare systems has gained a significant attention in both research and industry during the last decade. The main advantage of this new technology is lowering cost and improving the quality of services provided to patients [13], [14]. For instance, a wearable ubiquitous healthcare monitoring system has been proposed in [15]. The proposed technique involves electrocardiogram (ECG), accelerometer and oxygen saturation (SpO2) sensors. The data is transmitted in wireless sensor network from on-body wearable sensor devices to a base-station, which is connected to a server. In addition Lee and Chung [16] designed a smart shirt which measures electrocardiogram (ECG) and acceleration signals for continuous and real time health monitoring. The shirt mainly consists of sensors for continuous monitoring of the health data and conductive fabrics to get the body signal as electrodes. Other techniques focus on utilizing smart phones apps to help improve patients health and fitness [17]. The apps collect patient's data, analyze it, and then provide frequent feedbacks.

A. ACTIVITY RECOGNITION

Data collected from daily activities can be helpful for assisting elderly people and improve their quality of life. Activity recognition techniques have been used to predict future events that act as an alert about the actions that should be accomplished or an early reminder for emergency.

In [18], the authors proposed a system that combines accelerated data with vital signs to achieve better accuracy in activity recognition. The system requires a single sensing device and a mobile phone for data collection. The reported accuracy has been based on recognizing five activities using three different window sizes.

Han *et al.* [19] proposed a framework for recognizing user's activities from sensor data in order to detect irregular and unhealthy patterns. These patterns are then sent to a doctor and a caregiver. The framework consists of three modules: activity recognition, activity pattern generation and lifestyle disease prediction.

A multimedia based activity recognition system was proposed in [20] and [21]. The system is designed based on features that have been extracted from video and audio data, captured in a multisensory environment, for state recognition of patients. Experimental results show that the combined modality achieves better accuracy than that using a single model to correctly classify the patient's condition.

There are many algorithms that have been proposed for activity recognition. In [22] a combination of machine learning models have been applied to analyze data collected from a small number of sensors planted in a kitchen environment. Further, researchers applied hierarchical clustering enriched with time and location of data records to recognize activities. Another research in [23], which applied Dynamic Bayesian networks (DBNs), extracts dependencies among features evolving over time. Their algorithm was able to detect activities of actors using hidden semi-Markov models (HSMMss). Both of the above techniques are not scalable; efficient, and they were applied on data collected from a small number of sensors.

Naive Bayesian network structures have shown to be efficient in virtual cloud systems. It was able to detect noises effectively compared to C4.5 decision tree algorithm. However, high-order temporal dependencies decrease the performance of such algorithm as it is depending on high-variance of data (high bias). Such assumption makes Naïve Bayesian Network performs poorly on real domains. Another version of Naïve Bayesian Networks is the Quantitative temporal Bayesian networks (QTBNs) [24]. It is based on tracking the actor's behavior as it is following a pre designated plan. It models fluent and quantitative temporal relationships (high-level), which require manual encoding among different activities. In [25] Tree-Augmented Bayesian Networks were used to recognize activities using encoded key dependencies. The disadvantage of this technique is that it requires feature weighting and encoding of temporal information, which might result in time-conflict and noisy results.

In [26] researchers proposed a hybrid methodology of Neural Networks and Lookup tables to predict what so called occupancy-mobility patterns. This technique introduces many difficulties when applied in systems that require prior knowledge (such as health systems) as neural networks cannot incorporate them dynamically.

Matching sequence of actions related to an inhabitant has been investigated in [27] and [28]. This methodology depends heavily on repetitive actions. Such technique is effective in specific domains (such as manufacturing) but it is difficult to apply in environments where the user behavior is not supposed to be repetitive. In addition it is not efficient in a noisy environment with multiple and heterogeneous sensors.

Traditional clustering and classification techniques for syntactic structure, time intervals, domain specific such as [29]–[34] are also difficult to be applied in this domain since in-home human activities are ad hoc. A more sophisticated approach has been proposed in [35] where a nearest neighbor and decision trees algorithms are applied for activity recognition. This approach is efficient but suffers from scalability problems, especially when large number of sensors is used. It requires incorporating all sensor features to apply prediction.

Many algorithms have been proposed to measure the effect of activity recognition on power consumption such as [36]–[38]. In addition several other techniques have been proposed to study how spatial and temporal feature selection affects the performance of activity recognition such as [39]–[41]. Another research direction is the embedding of domain knowledge to formulate ontology for the activities [42], [43]. Unfortunately, maintaining such a complex structure (with new training data) resulted in high time complexity.

B. PRIVACY PRESERVING DATA ANALYTICS

Research in e-Health has revealed that dealing with patients' data is by no means an easy task. Healthcare providers are investigating new techniques to share health-related data for research and practical experience. Nevertheless, when it comes to medical records, including Electronic Health Records (EHR), health sensors (e.g. ECG monitors) and demographic data, privacy and security are major concerns. In several countries such as the United States there are specific standards to protect electronic health records. These standards are used to provide guidelines for healthcare providers who may share patients' data. HIPAA is one of those standards from the US Department of Health and Human services [44].

Most of the proposed approaches focus mainly on the sensing and health monitoring services, but ignore the data privacy issues. There have been several techniques applied to encrypt (anonymize) medical data such as using Semantic Marginality [45] which has been proposed for nominal attributes; does not anonymized numerical attributes.

The use of sensing activities is important to monitor smart environments including smart homes. However, users' privacy is considered one of the major challenges that need to be considered [46]. For example several behavior inferences including occupancy detection can be discovered by monitoring smart meters [47]. Therefore, several researchers refer to the importance of designing privacy preserving data gathering techniques when collecting data in smart environments [48], [49].

There have been many approaches that address privacy issues in smart environments, in particular, activity recognition. A privacy preserving automatic fall detection approach was proposed in [50]. The technique used foreground mask for visualization to detect five types of activities using 3D depth information. Such approaches focus on introducing noise to the captured images.

A similar approach for training activity recognition systems in real-time was proposed in [51]. The approach utilized a face recognition scheme, which identifies users' faces to preserve privacy. While such approaches protect users' identity, they do not avoid inference-based privacy attacks [3]. In addition, they mainly focus on protecting the privacy in videos and images; not the features of the sensed data. There have been some approaches which work at the feature level to achieve privacy, for instance an approach in [52] detected coughs by monitoring users' voice streams on mobile devices. Cough activity was recognized by analyzing a feature set. To achieve privacy, the feature set was perturbed in order to avoid reconstructing the voice stream. However, the perturbation approach leads to significant information loss.

Yang *et al.* [53] proposed a hybrid solution for privacy preserving data sharing in cloud environment. Using such approach, different methods are combined to support multiple paradigms of medical data sharing with different privacy strengths. On the same vein, Li in [54] identify and sketch the policy implications of using health social networks and how policy makers and stakeholders should elaborate upon them to protect the privacy of online health data. Several authors discuss the challenges of sharing patient-specific health data. Khokhar *et al.* [55] develop an analytical cost model to measure trade-off between privacy and utility while demonstrates the effectiveness of the model in real-life. Private companies such as Hakeem [56] have introduced many techniques to preserve the privacy of patients data using techniques such as System Access Control, Audit logs, Encryption, Private and Closed LAN/WAN.

Some approaches utilize k-anonymity to achieve privacy when collecting data using sensor networks. For instance in [57] the authors utilized a k-Anonymity perturbation approach to enable k-indistinguishability when the data is aggregated in sensor networks. The approach focuses on the task of data aggregation, but it does not study the applicability of k-anonymity in activity recognition. Another approach in [58] used a generalization technique to protect users' contextual data such as his current activities. Such an obfuscation approach created several generalization levels to release data using an ontological description that identifies the generalization level of object type instances.

Few approaches introduced different privacy mechanisms by releasing high level statistical features which do not disclose sensitive information. In [59] Zhang *et al.* used mFingerprint, a mobile-based activity recognition system. This

27/2/2008	12:49:52.624433	M14	ON	Wash hands	begin
27/2/2008	12:49:53.802099	M15	ON		
27/2/2008	12:49:54.24004	M16	ON		
27/2/2008	12:49:55.470956	M17	ON		segment
27/2/2008	12:49:55.470956	M15	OFF		
27/2/2008	12:49:55.808938	M14	OFF		
27/2/2008	12:49:57.548709	M16	OFF		
27/2/2008	12:49:57.717712	M13	OFF	Wash hands	end

FIGURE 4. A sample dataset.

approach applied multimodal sensors which can effectively monitor user activities while maintaining privacy. While such approaches focused on anonymizing sensors data, other approaches focus on the feasibility of attacking the collected data in order to discover suspicious inferences. As an example the work in [60] presented an evidence of the several threats posed by shared mobile sensor data.

III. METHODOLOGY

This section discusses the methodology and the settings used for activity recognition along with a privacy preserving mechanism to identify users' activities and avoid inference attacks. Let $\{s_1, s_2, \ldots s_n\}$ be a sequence of sensor events generated from a smart home. The sensors are assumed to be of two states (e.g., Open-Close, On-Off ... etc.). The data set consists of a set of events in which each event is associated with a date, time, sensor id, and sensor status [45]. The data stream is annotated by experts at the beginning and the ending of each segment by an activity from a set of predefined activities $\{d_1, d_2, \ldots, d_m\}$. Figure 4 shows an example of the dataset format.

A. DATA PREPROCESSING AND ACTIVITY PROFILING

The datasets have been processed for the purpose of creating a profile for each activity (abstraction). The profile is defined as the set of sequences that identify every activity in the dataset. In order to perform this task, event records have to be labeled first to supervise the learning process. The labeling task involves assigning an activity label to every record in the dataset. Algorithm 1 shows a formal description for the activity labeling and profiling process.

Line (1) defines the cleaning process where some activities in the dataset start but never end. This situation arises due to noise or when extracting portions of the dataset without taking into account the embedded structure of activities; i.e. every activity starts must end. Line (2) defines an empty set as a stack of activities, where the highest index in this stack (n) is the one on the top. Line (3) starts the main loop in this algorithm that spans every record (R) in the dataset. Lines (4-7) define the action taken when the algorithm detects the beginning of an activity (a_i).

First, the algorithm pushes the activity ID in the stack and inserts a sequence record in the database with a specific format; $(Date, Time, SensorID, Value, a_i)$. Lines (8-11) define the opposite situation in which an activity declares the end of its events. The algorithm first inserts the sequence record into the database and then pops the activity from the stack; then the activity becomes inactive. Lines (12-16) define the situation where concurrency exists due to the interleaving among activities. When two or more activities are active at the same time, our algorithm assigns the labels of ALL active activities. Such situation produces redundancy in the database, but enhances the accuracy of pattern detection.

Notice that, when ($n = 1$), no concurrency exists among events. Finally, Line (17) defines the situation in which a data record has been detected with no active activities. Such a situation arises due to incorrect manual labeling of activities and the noise in the datasets. Our approach, to handle such situation, is to skip such records. For data tuples that reside between the end of a segment and the beginning of a consecutive one (i.e. tuples that are not related to any segment), the label "Others" is assigned. In other words, these tuples are not labeled in the original dataset and have not been assigned to any data segments. During the preprocessing phase, we prepared two versions of every dataset in which the "others" label is kept in the first version, while ALL "others" labeled tuples have been removed from the other version. Our purpose is to study the effect of such tuples on the performance of our approach.

Algorithm 1 Activity Labeling (Dataset D)

— Let $Active = \{a_0, a_1, \ldots a_n\}$ be a set of active activities ordered chronologically, i.e. activities that began but not finished yet.
— Let a_i is an activity label such as $a_i \in Active$

==**Begin**==
1| $\forall\,(i)$, $remove\ (record_i D)$
$\quad where\ \exists\,(record_i.begin) \bigwedge (record_i.end)$
2| $Activity = \emptyset$
3| $for\ each\ record\ R \in D$
4| $\quad if\ (R.Activity = begin)\ Then$
5| $\{\ Push\ (Activity\ a_i, Set\ Active)$
6| $\quad Insert\ (Date, Time, SensorID, Value, a_i)$
7| $\}$
8| $if\ (R.Activity = end)\ Then$
9| $\{\ Insert\ (Date, Time, SensorID, Value, a_i)$
10| $\quad Pop\ (Activity\ a_i, Set\ Active)$
11| $\}$
12| $if\ (R.Activity = Null)$
13| $\{\ for\ (i = n\ TO\ 0\ Step - 1)$
14| $\quad \{\ Insert\ (Date, Time, SensorID, Value, a_i)$
15| $\quad \}$
16| $\}$
17| $If\ |Activity| = 0\ Then\ Skip$
18| **End FOR**
==**End**==

After labeling the dataset, the records of each activity are converted into a set of sequences; creating what is called the

Activity Profile. Each activity segment, in the labeled data, will form a sequence consisting of the sensors' IDs that have a complete cycle of their states. For instance, the door sensor s_k will be considered in the sequence if it has an "Open" state followed later with a "Close" state in the same segment.

As an illustrative example, consider the sequence "M4 M3 M4", as one of the sequences in the Read activity profile. This sequence consists of two motion sensors and means that the sensor M4 has a complete cycle of its states (i.e., "ON" followed by "Off"), and then M3 has a complete cycle, followed by M4 again with a complete cycle; given that all these events occurred in the same segment.

B. FEATURE VECTOR GENERATION

In this section we discuss the process of converting the sequences in the activity profiles into a set of feature vectors. Each sequence seq_l in the activity a_i profile is converted into a fixed dimension vector x_j consisting of an entry for each sensor s_k in the system. The entry s_k is waited based on the following equation:

$$wait(s_k) = \frac{\sum |s_k \subseteq Seq_l(a_i)|}{|S| \subseteq Seq_l(a_i)} \times \frac{\sum |s_k \subseteq Seq_l(a_i)|}{|S| \subseteq D} \tag{1}$$

1. $\sum |s_k \subseteq Seq_l(a_i)|$ is the number of times the sensor s_k appears in $Seq_l(a_i)$
2. $|S| \subseteq Seq_l(a_i)$ is the total number of sensors in $Seq_l(a_i)$
3. $|S| \subseteq D$ is the total number of sensors in the dataset

Each x_j is tagged with the label a_i. The collection of vectors and the corresponding activity labels are then become the training data that is fed into a classifier to learn the activity models.

C. PRIVACY PRESERVING OF USER ACTIVITIES

Preserving the privacy of the physical activities, while maintaining the required level of data utility, is required to avoid several forms of inference attacks, which are initiated to recognize specific activities that are considered private. Overall, the generalization techniques are utilized to prevent these attacks by replacing the real values of each feature with more generic values. The data recipient can still receive an anonymized version of the data, which achieves the required level of anonymity but can be still used for classifying non-sensitive activities.

1) ANONYMIZATION OF SENSORS DATASETS

The proposed method for privacy preserving of user activities has some similarity with k-anonymity. It creates groups of records with at least k-records in each group. However, there are two main differences between our method and k-anonymity. First, instead of utilizing suppression, we applied a bucketization approach to the entire dataset before the clustering step. This makes our method less sensitive to noise when clustering the records. We implemented a per activity micro-aggregation approach. The typical micro-aggregation approaches do not consider the differences between activities to perform the de-identification.

In fact, one of the major limitations of the existing k-anonymity based approaches is the difficulty of creating groups of equal size. Second, the k-anonymity approaches divided the data into sensitive and non-sensitive attributes. While it is straightforward to categorize features in some domains, this task is not easy when the collected data is a sensor-based where the majority of features are numerical. The k-anonymity model divides the data into many equivalent classes such that the values of an identification attribute of any record in the dataset are similar to at least $k-1$ records.

Other data perturbation techniques such as micro-aggregation work on numerical attributes to create clusters and replace their data by summaries such as the averages of feature values within each cluster. We decided to address the issues above using a modified version of micro aggregation techniques. Our approach satisfies the following properties:

1. It protects privacy of users by anonymizing the features of the collected activities.
2. It has very minor impact on the quality of activity recognition methods

It can be applied for each category of activities, therefore, no need to categorize activities into sensitive and non-sensitive activities. A summary of the perturbation steps is presented in Figure 5.

Algorithm 2 Perturbation of Sensor Data (Dataset D, k)

1| Let k' be the bucketization factor, I_ab be the value of a feature in column a, record b)

==**Begin**==

2| Divide D into n datasets, such that d_i contains records with activity A_i

3| **FOR** $i = 0$ to n do

- a. Apply bucketization to each record to replace I_ab with $(I\_ab - Min\_a)/k'$
- b. Cluster the records in D into clusters with at least k records

FOR each cluster C_j **do**

- a. Compute the mean of each feature for all records in same cluster
- b. Replace the values of each feature using the mean

End FOR

4| **End FOR**

==**End**==

Formally, if $D = \{d_1, \ldots, d_n\}$ is the set of activities with probability $P(d_i)$ for each activity, and $F = \{f_1, \ldots, f_m\}$ is a feature vector F with a set of values that identify the weight of each feature f_i with activity d_i, the objective is to generalize the feature vectors of activities $d_1, \ldots, d_k$ with a guarantee that even if attackers know the real values of every record in the generalized data except single record, the attacker still cannot infer from the observed perturbed data the value of that one record. The process of adding privacy to the collected data is described in algorithm 2.

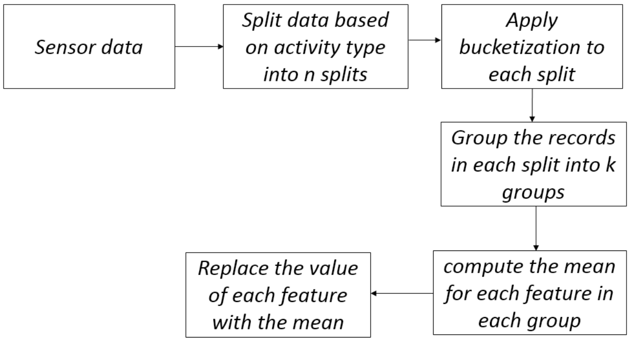


FIGURE 5. The micro-aggregation approach to anonymize sensors data.

As a precondition, we first split the pre-processed labeled sensor data D into n subsets, each subset contains the records that are labeled with an activity A_i. The objective is then to create cohesive clusters within each type of activities (e.g. minimizing intra cluster distance). Each cluster c_i must contain at least k records to achieve the required level of indistinguishability. In general, the greater the value of k, is the higher the amount of privacy. At the same time, the higher is the amount of information loss. Before the clustering step, we applied a bucketization function to minimize the effect of noise within each split. We applied k-means clustering to create clusters with at least k records. Since there could be clusters with less than k records, the smaller clusters are merged to create clusters with at least k records. Next, the value of each feature is replaced with the mean μ_f for a specific feature f in the cluster c_i.

IV. EXPERIMENTS AND RESULTS

This section presents the experiments that have been carried to evaluate the proposed activity recognition technique and its privacy preserving version. Three datasets from CASAS project have been used in the evaluation process [61].

TABLE 1. Dataset description.

Dataset	AcS	NAc	Dataset Size	Ns	Ts	Density Ratio=Ns/|D|
Milan	33	16	433665	1402	3577	≈ 39%
Tulum	20	11	1048576	471	1901	≈ 25%
Kyoto	25	5	64250	0	120	0

Acs: Active Sensors (Those that fire events during the experiment)
NAc: Number of Activities
Ns: Number of Sequences Labeled "Others
Ts:Total Number of Sequences

Table 1 shows the characteristics of datasets. Active Sensors represent the sensors that reported results within the dataset; some sensors are available in the environment in a sleep mode. The number of activities represents the number of distinct activities' names within the dataset. The dataset size is the total number of records in the dataset. "Ns" is the number of sequences in the activity profiles labeled with "Others". Finally, the ratio is the number of "Others" activities to the total number of records within datasets.

For Milan and Tulum datasets, the experiments have been carried with and without the "Others" activity. Note that only the active sensors have been used; as some sensors were not active in the test-bed during the data collection process.

A. EFFICIENCY OF MINING USER ACTIVITIES

To measure the effectiveness of the proposed activity profiling, the Naïve Bayes classifier (NBC) have been used and applied on the different datasets. The NBC has been used due to the weighting mechanism adopted in this paper. We computed the accuracy and F-measure according to the following formula [62]:

$$\begin{aligned} Accuracy\,(D, M) &= \frac{|Correctly\ Classified\ records|}{Total\ Number\ of\ Test\ records} \\ &= \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|} \end{aligned} \tag{2}$$

$$F\text{–}measure = \frac{2\,|TP|}{(2\,|TP| + |FP| + |FN|)} \tag{3}$$

- $|TP|$ (True Positive): is the number of correct predictions that an instance is positive.
- $|TN|$ (True Negative): is the number of correct predictions that an instance is negative.
- $|FP|$ (False Positive): is the number of incorrect of predictions that an instance positive.
- $|FN|$ (False Negative): is the number of incorrect of predictions that an instance negative.

TABLE 2. Accuracy and ***F***-measure.

Dataset	Accuracy	*F*-Measure
Milan (Without Others)	0.91	0.91
Milan (With Others)	0.88	0.89
Tulum (Without Others)	0.91	0.91
Tulum (With Others)	0.92	0.92
Kyoto	0.98	0.98

Table 2 summarizes the accuracy and the F-measure for the different datasets. As the table indicates, the accuracy is higher when the "Others" activity is excluded from the data.

TABLE 3. Accuracy reported for the different datasets [61].

Dataset	NBC	HMM	CRF
Milan	0.76	0.77	0.61
Tulum	0.59	0.75	0.79
Kyoto	0.78	0.78	0.97

Table 3 shows the results reported in [63] for the same datasets using three different classifiers: Naïve Bayes(NB), Hidden Markov Model (HMM) and Conditional Random Field(CRF). As shown, the activity profiling process, which is applied on the data, has increased the accuracy of the recognition. The enhancement reached to 10-12% compared to the worst accuracy achieved on Table 2.

B. PRIVACY PRESERVING OF USER ACTIVITIES

1) EVALUATION MEASURES

In order to evaluate the proposed method, we utilized the Conditional Privacy, a measure which is based on the differential entropy of a random variable. As shown on [64], the differential entropy of inferring the original attribute value A given the anonymized value B is

$$h(A \mid B) = -\int_{\Omega_{A,B}} f_{A,B}(a,b)\, log_2 f_{A|B=b}(a)\, da\, db \tag{4}$$

A is a variable that describes the data.
B is the variable that gives information on A.
$\Omega_{A,B}$ Identifies the domain of A and B.
Therefore the average conditional privacy of A given B, is:

$$\prod (A \mid B) = 2^{h(A \mid B)} \tag{5}$$

If A represent a feature in the original data and B is a feature in the anonymized data, then if the entropy between A and B is 0, this implies that there is no privacy preserved. As such the higher the value of conditional privacy is the higher the level of the protection.

Accuracy and F measure: We utilized accuracy and F measure to validate the effectiveness of our anonymization algorithm. In the experimental evaluation we proved that our method works reliably and is able to produce satisfactory accuracy values while preserving the privacy.

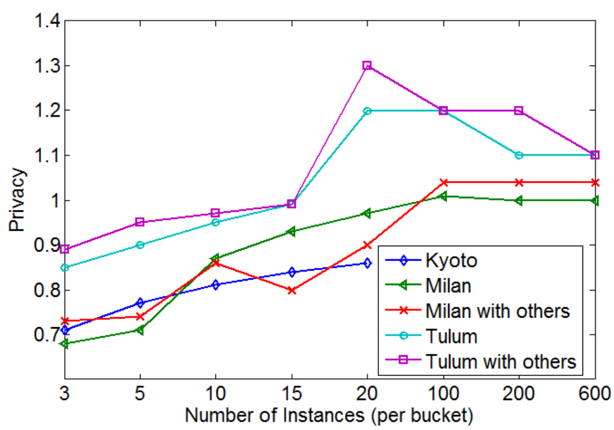


FIGURE 6. Average conditional privacy on different datasets.

2) PRIVACY RESULTS

Figure 6 shows the conditional privacy measures for the de-identified datasets using our approach. Our algorithm achieves satisfactory results. In particular, by increasing the number of instances per bucket, the Tulum dataset with others class achieves about 1.3 privacy compared to 0.89 when the number of the instances per bucket = 3. The privacy value is almost 1.1 when raising the value of k to 600.

To compare our results with existing approaches, we implemented a permutation function in Java, we then permutated the values of each feature and measure the privacy for perturbed data. The privacy values for Kyoto, Milan without Others, Milan with Others, Tulum, and Tulum With Others using the simple perturbation function are 0.31, 0.26, 0.34, and 0.45, and 0 respectively. Our approach has higher privacy values compared to such traditional perturbation techniques.

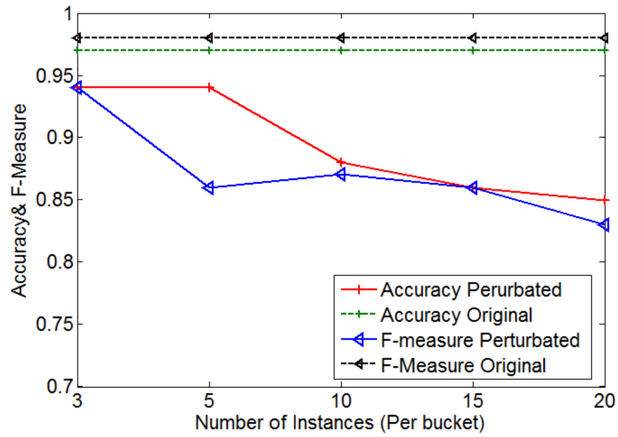


FIGURE 7. Naïve Bayes classification results on anonymized vs. original data (Kyoto dataset).

We ran several experiments to measure the accuracy on anonymized versus original data. We utilized Naïve Bayes classifier to run our experiments. This experiment is conducted in two steps:

1. Creating and testing the classification model using the original data.
2. Creating the classification model using the anonymized or perturbed data, then testing the model using the original data. We used accuracy and F-measure for the evaluation.

As shown in Figures 7, the classification results are not significantly affected when our perturbation mechanism is applied on the Kyoto dataset, in particular, when the number of records per bucket is less than 5. Due to the limited number of records in this dataset, the maximum number of records per bucket is 20. When the number of records per bucket is 5, the difference in terms of accuracy between the pretreated and the original dataset is less than 0.03, which shows that the perturbation approach does not have significant effects on the statistical characteristics of the features in this dataset. We have validated these results on other datasets, including the versions that contain the unknown activities, which are labeled as *others* in both Milan and Tulum datasets.

Figures 8, 9, 10, and 11 show that there is no significant difference between accuracy and F-measure values on the original and anonymized datasets. In addition, when increasing the number of instances per bucket we noticed a minor decline in accuracy and F-measure; this comes as a cost of achieving higher privacy to each dataset.

Table 4 shows an accuracy-based comparison between our approach and the random permutation which only adds a

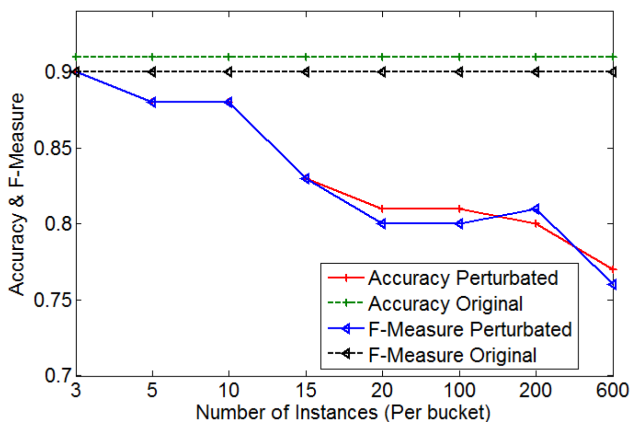


FIGURE 8. Naïve Bayes classification results on anonymized vs. original data (Milan dataset).

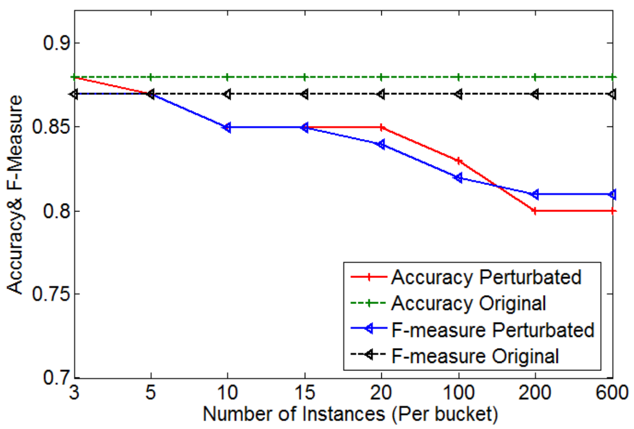


FIGURE 9. Naïve Bayes classification results on anonymized vs. original data (Milan with others dataset).

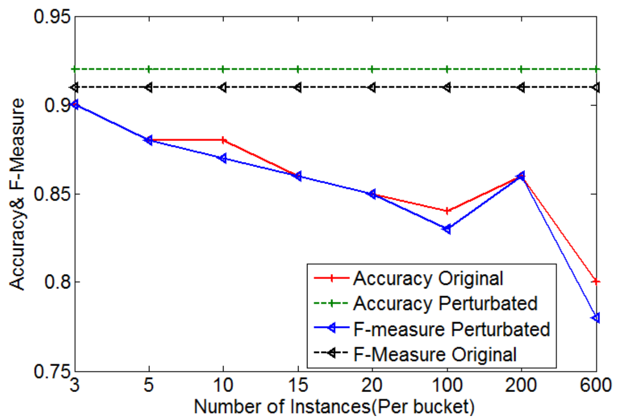


FIGURE 10. Naïve Bayes classification results on anonymized vs. original data (Tulum dataset).

random noise to the values of each feature. The values of the achieved classification accuracy using our approach are higher than using random perturbation, particularly, when the number of instances per bucket is small.

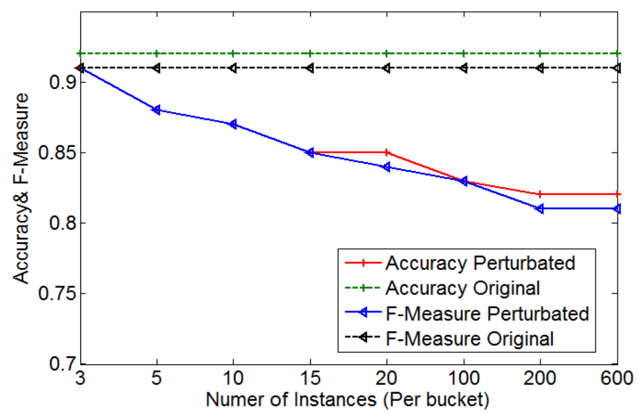


FIGURE 11. Naïve Bayes classification results on anonymized vs. original data (Tulum with others dataset).

TABLE 4. Accuracy-based comparison (our approach vs. random permutation).

Dataset/ Evaluation Measure	Accuracy/ original data	Accuracy/ anonymized data (min number of instances)	Accuracy/ Anonymized data (max number of instances)	Accuracy Using random permutation
Kyoto dataset	0.98	0.94	0.83	0.75
Milan dataset	0.91	0.90	0.76	0.70
Tulum	0.92	0.90	0.78	0.71
Milan (with others)	0.88	0.87	0.81	0.73
Tulum (with others)	0.92	0.91	0.82	0.72

V. DISCUSSION OF RESULTS

This section discusses the results we reported in section 6. The two-fold discussion will focus on justifying performance results in addition to the privacy one.

A. PERFORMANCE OF ACTIVITY RECOGNITION

Table 2 showed the performance of our proposed technique on every dataset. The experiments have been performed by considering the "Others" label and without incorporating it. Here we can see clearly that the accuracy of "Milan" dataset has been negatively affected by "Others", while it has a positive or has no impact on the accuracy of "Tulum" dataset.

Actually, such phenomena resulted from the percentage of "Others" segments or tuples as compared to the total number of tuples in the dataset, see Table 1. Because the density ratio of "Others" tuples is high as compared in "Tulum" dataset (by 14%), we observed that this ratio decreases the accuracy by (≈2%). "Kyoto" dataset does not contain any outliers ("Others" labels) and receive the highest performance among all datasets, which support our claim on the effect of the density of "Others" label on the performance of the classification task.

In addition, Table 3 provided a comparison with a highly reputable technique proposed in [61] with our proposed one. Our proposed technique outperformed them in terms of accuracy measure. The use of standard weighted model rather than data-driven one and the accuracy of the profiling task play a significant role on achieving such results.

B. PRIVACY PRESERVING OF USER ACTIVITIES

Anonymization process aims at achieving a balance between privacy and utility. We utilized a generalization approach to achieve this balance and minimize the amount of information loss when handling human-sensed data. Our approach performs perturbation on the data to make it k-indistinguishable. Several factors have affected the achieved privacy. First, the increasing number of records per group increases the level of privacy. The achieved level of privacy declines again when increasing the number of records per bucket to more than 20. Since our approach is tested on small datasets we noticed such a minor difference in the values of the achieved level of privacy. It is expected that there will be more significant difference when increasing the size of the datasets. Second, changing the dataset has no significant effects on the values of the achieved privacy. In addition, adding new types of activities does not have any significant effect on privacy values. The values of the achieved privacy when anonymizing Tulum, Tulum with others) is about the same. As such, the effectiveness of the proposed approach does not depend on the number and type of the activities used.

Third, the major usage of the proposed approach is to generate anonymized records that have similar statistical characteristics of the original ones. We proved this by using the anonymized data to create classification models. Yet, using the anonymized data for classification is not a common practice in the area of activity recognition. Our anonymization approach is designed and tested for a particular task, which is activity recognition. It can be also generalized and tested for other machine learning tasks and other statistical models. One particular objective is to apply our approach for distance-based data mining techniques. Fourth, compared to other perturbation-based techniques, our approach achieves better accuracy when the anonymized data is used to create the classification model. Even when increasing the number of records in each bucket to 600, our approach still outperforms the traditional approaches such as perturbation. Adding random noise to each record leads to a significant decline in the values of accuracy and F-measure on all datasets. For instance, when using the Tulum dataset, the achieved accuracy is about 0.90 (i.e., the number of instances per bucket is 3) compared to 0.71 when the random perturbation approach is used.

VI. CONCLUSION

We present a mining framework that is able to efficiently and privately recognize activities in smart homes environments. Our proposed framework takes advantage of a novel weighted profiling technique to achieve higher accuracy compared to other related techniques. In addition, a micro-aggregation technique has been proposed to enhance the privacy of the collected human sensed activities based on the category of each activity. The proposed privacy-preserving technique utilizes a modified version k-anonymity to replace the real value of each feature with the statistical characteristics of the original data. This work can be extended to consider applying privacy preserving techniques on vertically or horizontally distributed sensors data. In addition we plan to test our approach against popular inference attacks.

REFERENCES

[1] (2017). *10 Facts On Ageing And The Life Course*. [Online]. Available: http://www.who.int/features/factfiles/ageing/ageing_facts/en/

[2] G. Demiris, B. K. Hensel, M. Skubic, and M. Rantz, "Senior residents'perceived need preferences for 'smart home' sensor technologies," *Int. J. Technol. Assessment Health Care*, vol. 24, no. 1, pp. 120–124, 2008.

[3] Q. Fang, J. Sang, C. Xu, and M. S. Hossain, "Relational user attribute inference in social media," *IEEE Trans. Multimedia*, vol. 17, no. 7, pp. 1031–1044, Jul. 2015.

[4] S. Helal, W. Mann, H. El-Zabadani, J. King, Y. Kaddoura, and E. Jansen, "The Gator tech smart house: A programmable pervasive space," *Computer*, vol. 38, no. 3, pp. 50–60, Mar. 2005.

[5] D. J. Cook *et al.*, "MavHome: An agent-based smart home," in *Proc. PerCom*, 2003, pp. 521–524.

[6] B. Logan, J. Healey, M. Philipose, E. M. Tapia, and S. Intille, "A long-term evaluation of sensing modalities for activity recognition," in *Proc. Int. Conf. Ubiquitous Comput.*, 2007, pp. 483–500.

[7] D. J. Cook, A. S. Crandall, B. L. Thomas, and N. C. Krishnan, "CASAS: A smart home in a box," *Computer*, vol. 46, no. 7, pp. 62–69, 2013.

[8] M. Vacher *et al.*, "The sweet-home project: Audio technology in smart homes to improve well-being and reliance," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2011, pp. 5291–5294.

[9] Q. Ni, A. B. García Hernando, and I. P. de la Cruz, "The Elderly's independent living in smart homes: A characterization of activities and sensing infrastructure survey to facilitate services development," *Sensors*, vol. 15, no. 5, pp. 11312–11362, 2015.

[10] L. Xia and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Apr. 2013, pp. 2834–2841.

[11] R. Volner, P. Bores, and V. Smrz, "A product based security model for smart home appliances," in *Proc. 11th Int. Biennial Baltic Electron. Conf.*, 2008, pp. 221–222.

[12] M. S. Hossain and G. Muhammad, "Cloud-assisted industrial Internet of things (iiot)—Enabled framework for health monitoring," *Comput. Netw.*, vol. 101, pp. 192–202, Jun. 2016.

[13] A. Pantelopoulos and N. G. Bourbakis, "A survey on wearable sensor-based systems for health monitoring and prognosis," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 1, pp. 1–12, Jan. 2010.

[14] Y.-L. Zheng *et al.*, "Unobtrusive sensing and wearable devices for health informatics," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 5, pp. 1538–1554, May 2014.

[15] W.-Y. Chung, Y.-D. Lee, and S.-J. Jung, "A wireless sensor network compatible wearable U-healthcare monitoring system using integrated ECG, accelerometer and SpO_2," in *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBS)*, Aug. 2008, pp. 1529–1532.

[16] Y.-D. Lee and W.-Y. Chung, "Wireless sensor network based wearable smart shirt for ubiquitous health and activity monitoring," *Sens. Actuators B, Chem.*, vol. 140, no. 2, pp. 390–395, 2009.

[17] J. P. Higgins, "Smartphone applications for patients' health fitness," *Amer. J. Med.*, vol. 129, no. 1, pp. 11–19, 2016.

[18] D. Lara, A. J. Pérez, M. A. Labrador, and J. D. Posada, "Centinela: A human activity recognition system based on acceleration and vital sign data," *Pervasive Mobile Comput.*, vol. 8, no. 5, pp. 717–729, 2012.

[19] Y. Han, M. Han, S. Lee, A. Sarkar, and Y.-K. Lee, "A framework for supervising lifestyle diseases using long-term activity monitoring," *Sensors*, vol. 12, no. 5, pp. 5363–5379, 2012.

[20] M. S. Hossain, "Patient state recognition system for healthcare using speech and facial expressions," *J. Med. Syst.*, vol. 40, no. 12, p. 272, 2016.

[21] M. S. Hossain, "Patient status monitoring for smart home healthcare," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2016, pp. 1–6.

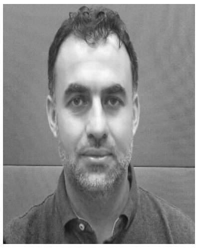
[22] T. Barger, M. Alwan, S. Kell, B. Turner, S. Wood, and A. Naidu, "Objective remote assessment of activities of daily living: Analysis of meal preparation patterns," Poster Presentation, Med. Autom. Res. Center, Health System, Univ. Virginia, Charlottesville, VA, USA, 2002.

[23] Y. Du, F. Chen, W. Xu, and Y. Li, "Recognizing interaction activities using dynamic Bayesian network," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, 2006, pp. 618–621.
[24] K. A. Tahboub, "Intelligent human-machine interaction based on dynamic Bayesian networks probabilistic intention recognition," *J. Intell. Robot. Syst.*, vol. 45, no. 1, pp. 31–52, 2006.
[25] M. G. Al Zamil and A. B. Can, "A model based on multi-features to enhance healthcare and medical document retrieval," *Informat. Health Social Care*, vol. 36, no. 2, pp. 100–115, 2011.
[26] M. Mozer, "The neural network house: An environment that adapts to its inhabitants', intelligent environments, papers from the AAAI spring symposium," AAAI Press, Palo Alto, CA, USA, Tech. Rep. SS-98-02, Mar. 1998.
[27] A. A. Chaaraoui, J. R. Padilla-López, P. Climent-Pérez, and F. Flérez-Revuelta, "Evolutionary joint selection to improve human action recognition with RGB-D devices," *Expert Syst. Appl.*, vol. 41, no. 3, pp. 786–794, 2014.
[28] S. Giroux, M. Castebrunet, O. Boissier, and V. Rialle, "A multiagent approach to personalization and assistance to multiple persons in a smart home," in *Proc. Workshops 28th AAAI Conf. Artif. Intell. (AAAI)*, 2014, p. 11.
[29] M. G. Al Zamil and A. B. Can, "ROLEX-SP: Rules of lexical syntactic patterns for free text categorization," *Knowl.-Based Syst.*, vol. 24, no. 1, pp. 58–65, 2011.
[30] M. G. A. Zamil and S. Samarah, "The application of semantic-based classification on big data," in *Proc. 5th Int. Conf. Inf. Commun. Syst. (ICICS)*, 2014, pp. 1–5.
[31] M. G. A. Zamil and S. Samarah, "Dynamic event classification for intrusion and false alarm detection in vehicular ad hoc networks," *Int. J. Inf. Commun. Technol.*, vol. 8, nos. 2–3, pp. 140–164, 2016.
[32] M. G. Zamil and S. Samarah, "Dynamic rough-based clustering for vehicular ad-hoc networks," *Int. J. Inf. Decision Sci.*, vol. 7, no. 3, pp. 265–285, 2015.
[33] S. Samarah, M. Zamil, and A. Saifan, "Model checking based classification technique for wireless sensor networks," *New Rev. Inf. Netw.*, vol. 17, no. 2, pp. 93–107, 2012.
[34] M. Rawashdeh, H.-N. Kim, and A. El Saddik, "Folksonomy-boosted social media search and ranking," in *Proc. 1st ACM Int. Conf. Multimedia Retr.*, 2011, p. 27.
[35] A. Alvi, U. Qamar, W. Muzaffar, and W. Butt, "A novel hybrid classifiers based model for mining in neuro-imaging," in *Proc. Int. Conf. Internet Things Cloud Comput.*, 2016, p. 13.
[36] Y.-X. Lai, C.-F. Lai, Y.-M. Huang, and H.-C. Chao, "Multi-appliance recognition system with hybrid SVM/GMM classifier in ubiquitous smart home," *Inf. Sci.*, vol. 230, pp. 39–55, May 2013.
[37] P. Cottone, S. Gaglio, G. Re, and M. Ortolani, "User activity recognition for energy saving in smart homes," *Pervasive Mobile Comput.*, vol. 16, pp. 156–170, Jan. 2015.
[38] M. S. Hossain, M. A. Rahman, and G. Muhammad, "Cyber–physical cloud-oriented multi-sensory smart home framework for elderly people: An energy efficiency perspective," *J. Parallel Distrib. Comput.*, vol. 103, pp. 11–21, Oct. 2016.
[39] C.-L. Wu, Y.-S. Tseng, and L.-C. Fu, "Spatio-temporal feature enhanced semi-supervised adaptation for activity recognition in IoT-based context-aware smart homes," in *Proc. IEEE Int. Conf. Cyber Phys. Soc. Comput.*, Aug. 2013, pp. 460–467.
[40] K. Wongpatikaseree, M. Ikeda, M. Buranarach, T. Supnithi, A. O. Lim, and Y. Tan, "Activity recognition using context-aware infrastructure ontology in smart home domain," in *Proc. 7th Int. Conf. Knowl., Inf. Creativity Support Syst. (KICSS)*, 2012, pp. 50–57.
[41] M. Bessho, N. Koshizuka, S. Kobayashi, and K. Sakamura, "Location systems for ubiquitous computing," *J. Inst. Electron. Inf. Commun. Eng.*, vol. 92, no. 4, pp. 249–255, 2009.
[42] K. Wongpatikaseree, A. O. Lim, M. Ikeda, and T. Yasuo, "High performance activity recognition framework for ambient assisted living in the home network environment," *IEICE Trans. Commun.*, vol. 97, no. 9, pp. 1766–1778, 2014.
[43] L. Chen, C. D. Nugent, and H. Wang, "A knowledge-driven approach to activity recognition in smart homes," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 961–974, Jun. 2012.
[44] U. S. D. o. H. H. Services. (2003). *U.S. Department of Health & Human Services. Health Information Privacy*. [Online]. Available: http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/
[45] J. Domingo-Ferrer, D. Sánchez, and G. Rufian-Torrell, "Anonymization of nominal data based on semantic marginality," *Inf. Sci.*, vol. 242, pp. 35–48, Sep. 2013.
[46] D. J. Cook and N. Krishnan, "Mining the home environment," *J. Intell. Inf. Syst.*, vol. 43, no. 3, pp. 503–519, 2014.
[47] D. Chen, S. Kalra, D. Irwin, P. Shenoy, and J. Albrecht, "Preventing occupancy detection from smart meters," *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2426–2434, Sep. 2015.
[48] P. M. Wightman, M. Zurbarán, M. Rodriguez, and M. A. Labrador, "MaPIR: Mapping-based private information retrieval for location privacy in LBISs," in *Proc. IEEE 38th Conf. Local Comput. Netw. Workshops (LCN Workshops)*, Oct. 2013, pp. 964–971.
[49] I. J. Vergara-Laurens, D. Mendez-Chaves, and M. A. Labrador, "On the interactions between privacy-preserving, incentive, and inference mechanisms in participatory sensing systems," in *Proc. Int. Conf. Netw. Syst. Secur.*, 2013, pp. 614–620.
[50] C. Zhang, Y. Tian, and E. Capezuti, "Privacy preserving automatic fall detection for elderly using RGBD cameras," in *Proc. Int. Conf. Comput. Handicapped Persons*, 2012, pp. 625–633.
[51] W. S. Lasecki, Y. C. Song, H. Kautz, and J. P. Bigham, "Real-time crowd labeling for deployable activity recognition," in *Proc. Conf. Comput. Supported Cooperat. Work*, 2013, pp. 1203–1212.
[52] E. C. Larson, T. Lee, S. Liu, M. Rosenfeld, and S. N. Patel, "Accurate and privacy preserving cough sensing using a low-cost microphone," in *Proc. 13th Int. Conf. Ubiquitous Comput.*, 2011, pp. 375–384.
[53] J.-J. Yang, J.-Q. Li, and Y. Niu, "A hybrid solution for privacy preserving medical data sharing in the cloud environment," *Future Generat. Comput. Syst.*, vol. 43, pp. 74–86, Feb. 2015.
[54] J. Li, "Privacy policies for health social networking sites," *J. Amer. Med. Informat. Assoc.*, vol. 20, no. 4, pp. 704–707, 2013.
[55] R. H. Khokhar, R. Chen, B. C. Fung, and S. M. Lui, "Quantifying the costs and benefits of privacy-preserving health data publishing," *J. Biomed. Informat.*, vol. 50, pp. 107–121, Aug. 2014.
[56] A. Hakeem. (2015). *Electronic Health Solutions (EHS)*. [Online]. Available: http://www.ehs.com.jo/en/content/about-us-0
[57] M. M. Groat, W. Hey, and S. Forrest, "KIPDA: k-indistinguishable privacy-preserving data aggregation in wireless sensor networks," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 2024–2032.
[58] F. Rahman, Md. E. Hoque, F. A. Kawsar, and S. I. Ahamed, "Preserve your privacy with PCO: A privacy sensitive architecture for context obfuscation for pervasive e-community based applications," in *Proc. IEEE 2nd Int. Conf. Soc. Comput. (SocialCom)*, Aug. 2010, pp. 41–48.
[59] H. Zhang, Z. Yan, J. Yang, E. M. Tapia, and D. J. Crandall, "Mfingerprint: Privacy-preserving user modeling with multimodal mobile device footprints," in *Proc. Int. Conf. Soc. Comput. Behavioral-Cultural Modeling Predict.*, 2014, pp. 195–203.
[60] N. D. Lane, J. Xie, T. Moscibroda, and F. Zhao, "On the feasibility of user de-anonymization from shared mobile sensor data," in *Proc. 3rd Int. Workshop Sens. Appl. Mobile Phones*, 2012, p. 3.
[61] WSU. W. C. Datasets, Ed. (2007). *CASAS Datasets*, [Online]. Available: http://casas.wsu.edu/datasets/
[62] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, vol. 5. San Francisco, CA, USA: Morgan Kaufmann, 2001.
[63] D. J. Cook, "Learning setting-generalized activity models for smart spaces," *IEEE Intell. Syst.*, vol. 27, no. 1, pp. 32–38, 2012.
[64] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proc. 20th ACM SIGMOD-SIGACT-SIGART Symp. Principles Database Syst.*, 2001, pp. 247–255.

SAMER SAMARAH received the Ph.D. degree in computer science from the University of Ottawa, Canada, in 2008. He served as an External Examiner for many theses. He is currently an Associate Professor with the Computer Information Systems Department, Yarmouk University, Jordan. He has many published journal and conference papers in the area of data mining and wireless networks. His research focuses on discovering behavioral patterns from data collected by wireless sensor networks and vehicular ad hoc networks. He is a Referee for many international journals and conferences.

MOHAMMED GH. AL ZAMIL received the B.Sc. and master's degrees in computer science from Yarmouk University (YU), Jordan, and the Ph.D. degree in information systems from Middle East Technical University, Ankara, Turkey, in 2010. He is currently an Associate Professor with the Department of Computer Information Systems, YU. His research interests include data mining, wireless sensor networks, model checking, software verification, and software engineering.

AHMED F. ALEROUD received the B.S. degree in software engineering from Hashemite University, Jordan, and the M.S. and Ph.D. degrees in information systems from the University of Maryland, Baltimore County. He was a Visiting Associate Research Scientist with the University of Maryland, where he was involved in cyber security research projects. He is currently an Assistant Professor of Computer Information Systems with Yarmouk University, Jordan. His research work focuses on cyber-security, data mining for privacy preserving network data analytics, and detection of social engineering attacks.



MAJDI RAWASHDEH received the Ph.D. degree in computer science from the University of Ottawa, Canada. He is currently an Assistant Professor with Princess Sumaya University for Technology, Jordan. His research interests include social media, user modeling, recommender systems, smart cities, and big data.

MOHAMMED F. ALHAMID received the Ph.D. degree in computer science from the University of Ottawa, Canada. He is currently an Assistant Professor with the Software Engineering Department, King Saud University, Riyadh, Saudi Arabia. His research interests include recommender systems, social media mining, big data, and ambient intelligent environment.

ATIF ALAMRI is currently an Associate Professor with the Information Systems Department, College of Computer and Information Sciences, King Saud University. Riyadh, Saudi Arabia. His research interests include multimedia-assisted health systems, ambient intelligence, and service-oriented architecture. He serves as a Program Committee Member for many conferences in multimedia, virtual environments, and medical applications. He was a Guest Associate Editor of the IEEE Transactions on Instrumentation and Measurement, a Co-Chair of the first IEEE International Workshop on Multimedia Services and Technologies for E-health, and a Technical Program Co-Chair of the 10th IEEE International Symposium on Haptic Audio Visual Environments and Games.