HealthXAI: Collaborative and Explainable AI for Supporting Early Diagnosis of Cognitive Decline

Elham Khodabandehloo[◇], Daniele Riboni[○*], Abbas Alimohammadi[◇*]

[◇] *Dept. of Geo-spatial Information System, K. N. Toosi University of Technology, Tehran, Iran*

[○] *Dept. of Mathematics and Computer Science, University of Cagliari, Italy*

Abstract

Our ageing society claims for innovative tools to early detect symptoms of cognitive decline. Several research efforts are being made to exploit sensorized smart-homes and artificial intelligence (AI) methods to detect a decline of the cognitive functions of the elderly in order to promptly alert practitioners. Even though those tools may provide accurate predictions, they currently provide limited support to clinicians in making a diagnosis. Indeed, most AI systems do not provide any explanation of the reason why a given prediction was computed. Other systems are based on a set of rules that are easy to interpret by a human. However, those rule-based systems can cope with a limited number of abnormal situations, and are not flexible enough to adapt to different users and contextual situations. In this paper, we tackle this challenging problem by proposing a flexible AI system to recognize early symptoms of cognitive decline in smart-homes, which is able to explain the reason of predictions at a fine-grained level. Our method relies on well known clinical models that consider subtle and overt behavioral anomalies, as well as spatial disorientation and wandering behaviors. In order to adapt to different individuals and situations, anomalies are recognized using a collaborative approach. We experimented our approach with a large set of real world subjects, including people with MCI and

*Corresponding authors

Email addresses: riboni@unica.it (Daniele Riboni[○]), alimoh_abb@mail.kntu.ac.ir (Abbas Alimohammadi[◇])

people with dementia. We also implemented a dashboard to allow clinicians to inspect anomalies together with the explanations of predictions. Results show that our system's predictions are significantly correlated to the person's actual diagnosis. To the best of our knowledge, this is the first work that explores data-driven explainable AI for supporting the diagnosis of cognitive decline.
Keywords: Pervasive healthcare, Explainable artificial intelligence, Cognitive decline, Sensor-based activity recognition.

1. Introduction

Declined fertility and increased longevity are determining a demographic shift that is considered one of the dominant phenomena of the 21^{th} century. Several recent studies forecast that in the near future the senior population is going to double as a percentage over the whole population, and this fact will have a strong impact on several fields of our societies [1]. In the healthcare domain, there is a growing interest in devising innovative strategies and techniques to prolong healthy and independent living in the elderly population. To this aim, the increasing adoption of Internet of Things (IoT) platforms in smart-homes, together with the integration of artificial intelligence (AI) agents, provide unprecedented opportunities for remotely monitoring the health status of seniors. In particular, it is necessary to implement effective tools to early detect symptoms of cognitive decline, in order to report them to clinicians, and to raise alarms when needed [2]. This need is particularly stringent for seniors living alone. Indeed, for this category of people, periodic screening by clinicians should be complemented by continuous cognitive assessment carried out using automatic tools for promptly detecting possible cognitive issues.

Different AI-based systems have been proposed in the last years to remotely assess the health status of seniors [3]. They can be divided in three categories: data-driven, knowledge-based, and hybrid approaches. Most data-driven methods rely on datasets of activities carried out by seniors and try to recognize abnormal behaviors based on deviations against the expected activity patterns [4].

However, given the high variability of activity execution due to contextual conditions, those methods may provide many false positives. Moreover, while they are effective in recognizing large deviations from the expected behavior, they are less suitable to recognize mild behavioral changes on the long term which may be indicators of cognitive decline. A further limitation of data-driven methods is that they do not provide explanations of their predictions to clinicians.

Knowledge-based approaches generally rely on manually defined sets of rules that model a given behavior as 'abnormal' according to some clinical model [5]. While those rules are easily interpretable by a human, manually defined rules can cover only a restricted set of possible behaviors. Moreover, rules must be carefully crafted considering the individual's profile. Hence, the approach can hardly scale with the number of monitored individuals. Hybrid approaches try to take the best of data-driven and knowledge-based ones, by mining a model of anomalies based on a labeled set of abnormal behaviors [6, 7]. Unfortunately, the acquisition of large labeled sets of abnormal behaviors is expensive and introduces severe privacy issues due to the presence of an observer in charge of annotating anomalies.

In this paper, we aim at devising an IoT system supporting the early diagnosis of cognitive decline, which provides caregivers with numerical scores reporting abnormal behaviors at a high level, together with natural language explanations of the predictions. We pursue this challenging goal by proposing HealthXAI, a novel collaborative sensor-based architecture empowered by explainable AI functionalities. Our system relies on well-known clinical models of cognitive decline, which define both abnormal behaviors and locomotion patterns. Regarding activities, we consider both a model of overt abnormal behaviors [8] and a model of subtle inefficiencies in the execution of Activities of Daily Living (ADLs) [9]. Regarding locomotion, we consider a model of wandering [10], and different indicators of spatial disorientation proposed in the literature [11, 12, 13]. HealthXAI acquires an anonymous dataset of activity data collaboratively gathered from smart-home inhabitants having a profile similar to the one of the target user. The system mines the dataset to instantiate

a personalized model of abnormal activities and locomotion episodes according to the clinical models. The platform collects activity and locomotion data from the smart-home infrastructure of the senior. Activity recognition is performed using existing activity recognition algorithms and is outside the scope of this paper. Based on the personalized model, HealthXAI detects possible anomalies and uses a decision tree machine learning algorithm to compute an anomaly score for each activity, as well as an overall anomaly score of the senior. The decision tree is parsed to produce natural language explanations of the predictions. A Web dashboard is available to clinicians, who can inspect anomalies and explanations at a fine-grained level.

We have implemented all the modules of HealthXAI and executed extensive experiments with a large set of real-world activity data acquired in instrumented smart-homes from 192 seniors, including people with Mild Cognitive Impairment (MCI) and people with dementia (PwD). The results indicate that HealthXAI predictions are significantly correlated to the actual diagnosis of the senior. Summarizing, the main contributions of our work are the following:

- We propose a novel collaborative IoT framework for recognizing symptoms of cognitive decline, which requires neither manual modeling, nor labeled datasets of abnormal behaviors.
- Compared to data-driven approaches, our framework provides natural language explanation of the detected anomalies which refer to well-known clinical models of cognitive decline.
- Compared to knowledge-based and hybrid approaches, our system can scale with the number of subjects and provides personalization thanks to the collaborative approach.
- We performed extensive experiments with several seniors, including cognitively impaired people, that show the potential of our system in supporting the diagnosis of cognitive decline.

The rest of the paper is structured as follows. Section 2 reports preliminary information and related work. The overall HealthXAI framework is introduced in Section 3. In Section 4 we present our methods for collaborative model-based

analysis of activities and motion, while in Section 5 we describe the technique for explainable detection of cognitive decline symptoms. In Section 6 we report and discuss the experimental results. Section 7 concludes the paper and outlines future work.

2. Preliminaries and related work

In this section, at first we present the clinical models adopted in this work. Those models have been proposed in the neuropsychology domain to characterize early symptoms of cognitive decline based on behavioral anomalies and locomotion patterns. Then, we review existing IoT-based methods to recognize abnormal behaviors and anomalous movements related to cognitive decline.

2.1. Clinical models of cognitive impairment based on behavioral anomalies

Several research studies show that it is possible to recognize early symptoms of cognitive decline based on subtle or overt anomalies performed by the elderly during the execution of his/her daily activities [14, 15]. Hence, different clinical models have been defined, which identify and classify these kinds of abnormal behaviors. In particular, for the sake of this work, we consider two models of behavioral anomalies, which are well known in the literature and proved to be effective in distinguishing cognitively impaired from cognitively healthy seniors in several studies. Even though the terminology in the clinical field may vary, we refer to those models as *clinical model of overt errors*, and *clinical model of subtle inefficiencies*, respectively.

2.1.1. Clinical model of overt errors

Overt errors, defined in [8], are classified in three main categories.

- *Omissions* are observed when a key step, or multiple steps of an activity are not performed by the elderly. A key step is an action that is necessary to correctly perform the activity. For instance, an omission occurs when an individual does not add salt to water for cooking pasta.

Sub-category of commission anomaly	Description	Example
Anticipation-omission	The individual performs actions in different order than appropriate	Putting pasta before boiling water
Perseveration	The individual repeats the same action more times than expected	Taking a medication more times than prescribed in a day

Table 1: The categories of Commission overt errors considered in HealthXAI.

- *Action-additions* are observed when an action not related to the current activity is executed. An action is not related if its execution has no effect on the outcome of the activity. For instance, an action-addition occurs when an individual who is cooking rice takes not only the container of rice, but also the one of pasta.
- Finally, *commissions* are observed when key actions of an activity are performed inaccurately. Commissions are further classified according to the kind of error performed by the individual. Table 1 reports the subcategories of commissions that we consider in this work.

Experiments with a large set of patients showed that the rate of overt errors performed by persons with MCI is significantly larger than the one of cognitively healthy seniors, and significantly lower than the one of PwD [16].

In order to recognize overt errors, the IoT system needs a model of activities and their key steps, and knowledge about current activity and executed actions.

2.1.2. Clinical model of subtle inefficiencies

Subtle inefficiencies, introduced by Seligman et al. in [9], are the result of subtle disruption of functional abilities in seniors who are still capable of completing instrumental activities of daily living (IADLs). The latter are daily living activities that involve the use of instruments, such as cooking, cleaning, laundry, and their execution is correlated to the cognitive functions of the individual [17]. Subtle inefficiencies are classified in five categories. For the sake of this work, we concentrate on the *reach-touch* category. Indeed, inefficiencies of that type can be monitored based on sensor data, while the other types of inefficiencies are hard to recognize without the use of cameras, which is out of

the scope of this paper.

A reach-touch anomaly is observed when an individual reaches for and touches an unwanted object during the execution of an activity. An unwanted object is an incorrect item for the execution of a given activity. For instance, that anomaly is observed when an individual reaches for and touches the sugar container when preparing pasta. The results of experiments with older adults suggest that subtle inefficiencies can be useful for the assessment of early functional decline in the elderly [9].

In order to recognize subtle inefficiencies, the IoT system needs a model of activities and unwanted objects, and knowledge about current activity and executed actions at a fine-grained level, including monitoring the reach and touch of objects.

2.2. Clinical models of cognitive impairment based on locomotion anomalies

Locomotion is a factor introduced by Algase [18] to describe the temporal phases of movement. A locomotion episode is defined as a rhythmical movement composed of walking phases followed by non-walking phases. We adopt the term 'locomotion' to refer to sequences of movements in the home. In the following, we describe the models adopted in this work.

2.2.1. Martino-Saltzman model

Among different models proposed for classifying and measuring wandering behaviors, one of the earliest and most widely accepted was proposed by Martino-Saltzman [19]. This model defines four distinctive classes of movement patterns, shown in Figure 1:

- **Direct**: a single straightforward path from a point to a destination, which is not complicated and does not diverge significantly from the most efficient path.
- **Pacing**: at least three consecutive repeated back-and-forth movements between two locations.

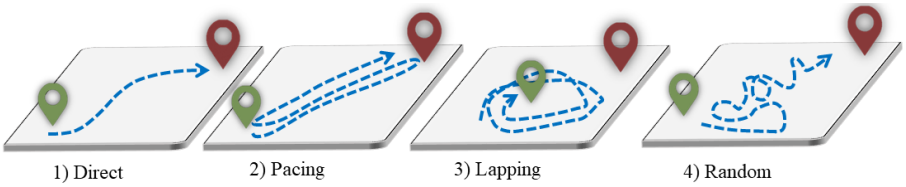

Figure 1: Travel patterns of people according to the Martino-Saltzman model of wandering behavior [19].

- **Lapping**: a circular repeated movement that passes across at least 3 distinct points and that is repeated at least twice.
- **Random**: an aimless movement across numerous locations, which is not direct.

Based on this model, *random*, *pacing* and *lapping* patterns are associated to cognitive decline. In our indoor context, a direct path is the shortest path between the start and the end of a trajectory in the home.

In order to recognize anomalies based on this model, the IoT system needs fine-grained information about the movement of the inhabitant in the home.

2.2.2. Low-level motion indicators

Few low-level motion indicators have been proposed in the literature, which proved to be effective in distinguishing cognitively healthy seniors from person with MCI and PwD [20]. For the sake of this work, we consider the indicators listed below, which can be easily monitored in a smart-home environment provided with fine-grained positioning technologies.

- **Jerk** [12] is the rate at which a person's acceleration changes with respect to time. Hence, it can be computed as the first time derivative of acceleration.
- **Sharp angles** [13] are defined as vector angles in a trajectory being equal to or more than 90 degrees.
- **Straightness** [11] is defined as the ratio of the distance between two consecutive trajectory segments and the distance between the start and end point of these segments.

2.3. IoT-based recognition of abnormal activities

In recent years, the application of AI methods to sensor data is gaining increasing interest in several applications areas [21, 22, 23, 24]. Internet of Things (IoT)-based solutions provide significant capabilities, such as continuous monitoring, cost effectiveness, and scalability, which make them a powerful solution for healthcare monitoring. For instance, Mining Mind is a project funded by the Korean Government to build a comprehensive platform for mining human's daily life data gathered from heterogeneous sources[1]. The platform provides several services for data integration and reasoning, including a module for human behavior quantification. That module adopts a mathematical model based on wellness guidelines [25].

Thanks to the emerging IoT adoption, smart-home technologies are significantly automating the tasks related to elderly care, by improving monitoring the wellness of the resident, especially in the case of medical emergencies. The use of smart-home systems enables continuous monitoring in seniors' residence without causing disturbance, while allowing long-term assessment of the inhabitants' behavioral patterns. A popular method for spatio-temporal activity tracking is based on motion sensing devices. Lam el al. propose a healthcare tool based on the Kinect platform and machine learning technologies, which supports activity tracking and monitoring for helping people with Alzheimer living independently [26]. Ota et al. used ultra-wideband impulse-radio (UWB-IR) monitoring sensors for recognizing different activities, including sleeping, sitting up in bed, wandering in the room, falling down, going in and out the room [27].

The importance of considering activity data acquired from smart-homes has been widely investigated. For example, the study reported in [28] shows that a person's walking velocity and activity patterns may change relevantly based on the current stage of cognitive impairment. Ishii et al. proposed an IoT platform to detect early symptoms of Alzheimer's disease by recognizing anomalous ac-

[1] `http://www.miningminds.re.kr`

tivities of persons living alone [29]. Suzuki et al. used position sensors mounted on the ceiling of rooms in a nursing home to monitor the inhabitants high-level ADLs [30]. Abnormal behaviors were recognized based on statistical deviation from the usual activity pattern. An IoT-based system for recognition of abnormal behaviors was proposed by Riboni et al. in [5]. That system relies on a set of manually defined rules that determine the recognition of behavioral anomalies at a fine-grained level. The considered anomalies are related to symptoms of MCI or dementia. While those rules are easily interpretable by a human, the system can hardly scale with the number of considered anomalies and different contextual conditions. Indeed, those rules strongly depend on the specific smart-home environment and on the person's characteristics. In order to address this problem, Haider Janjua et al. proposed a system to automatically mine behavioral anomaly rules from a dataset of labeled activities and abnormal behaviors [7]. However, the acquisition of a large labeled dataset of real-world anomalies is problematic. In this work, we aim at automatically training an explainable model of abnormal behaviors exploiting a collaborative approach, without the need of labeled datasets of anomalies.

2.4. IoT-based recognition of wandering and locomotion patterns

Real-time analysis of trajectories and early detection of wandering episodes is compulsory for reducing potential risks or damages to wanderers. In addition, real time wandering detection can increase the security and also decrease costs thanks to independent care [31]. In this sense, IoT technologies, smartphone sensors, and wearable devices, provide novel possibilities for increasing the effectiveness of remote healthcare services. Recently, different attempts have been made for wandering recognition in outdoor and indoor environments, by using positioning systems and location pattern mining methods [32]. Lin et al. proposed a method for wandering detection based on Martino-Saltzman model, by analysing outdoors GPS trajectories [33]. Another system for supporting secure and independent outdoor walking was proposed in [34]. The system relies on the analysis of human activity and it is adapted to individual locomotion

patterns. The system relies on a wandering detection framework, and includes a smart GPS tracker for real time outdoor positioning estimation. Schaat et al. investigated the feasibility of real-time detection of disorientation based on sensor data acquired from GPS and accelerometers for people with MCI and dementia [35], achieving promising results. Qiang Lin et al. propose a real time wandering detection method using GPS traces for detecting pacing and lapping of elders in outdoor environments [13]. The method relies on the analysis of changes in the shape of travel traces.

Recognizing wandering behaviors indoors is not straightforward, because of several contextual factors, including the current activity execution, and the presence of obstacles in the home. Kumar et al. showed the capability of evaluating the cognitive status using features extracted from trajectories, such as turning angle, path-efficiency, speed, and ambulation fraction, relying on Ultra-wideband real-time location data [36]. Vuong et al. developed an automatic method to classify travel patterns based on Martino-Saltzman model, using trajectories collected through RFID tags. They classified episodes through machine learning algorithms and showed that tree-based algorithms achieved good performance [37]. Kearns et al. acquired indoor trajectory data from both PwD and cognitively healthy subject in a common area of a residence for seniors. Position data was acquired using ultra-wide band radio technologies. The authors calculated path tortuosity by the Fractal Dimension (Fractal D) algorithm for discriminating between cognitively healthy subjects and PwD [38]. Other indoor localization technologies were proposed to support healthcare. Sun et al. used a pressure sensing system based on fiber-optic to track the movements of people [39], and a space encoding scheme to retrieve the inhabitants' position. They adopted mixture models on location information for recognizing high-level behaviors. All those systems recognize generic movements and activities, while in our work we tackle the recognition and explanation of fine-grained behavioral anomalies.

In another work, the authors adopted the Martino-Saltzman model for indoor monitoring of elderly patients with dementia [40]. Based on customized ac-

tive infrared sensors to gather human indoor motions, their system can identify wandering from repetitive locomotion episodes. Khan and Hassan introduced a framework for integrating two kinds of sensors: physiological sensors (such as blood pressure and heart rate monitors) as indicators of emotional and physiological arousal, and geo-location sensors for wandering detection based on the Martino-Saltzman model and the Algase wandering scale [41]. Another system, based on the smartphone sensors, provides two kinds of alarms in activity recognition and safe-zone geo-fencing [31]. Khodabandehloo and Riboni proposed a collaborative system to recognize early symptoms of cognitive decline based on indoor location traces analysed according to the Martino-Saltzman model [42]. Recognized anomalies are used to build feature vectors, and a machine learning method is used to predict the cognitive status of the individual. However, the system does not consider abnormal activities, and does not provide explanations of its predictions.

Another category of IoT technologies for tracking inhabitants in smart-homes consists in the use of camera and computer visions tools. For instance, the system proposed by Wang et al. [43] relies on cameras and computer vision for gait assessment of elderly people. However, camera-based systems determine relevant privacy issues, especially in the home. Moreover, the above mentioned works do not provide human-understandable explanations of their predictions.

3. HealthXAI system

In this section, we present the HealthXAI system and its modules.

3.1. HealthXAI architecture

The HealthXAI architecture is shown in Figure 2; the modules describing the core contribution of this paper are depicted as blue boxes. The architecture relies on a SMART-HOME SENSOR INFRASTRUCTURE which uses a STREAM PROCESSING SOFTWARE PLATFORM and a SEMANTIC INTEGRATION LAYER to integrate heterogeneous sensor data. The architecture includes modules for

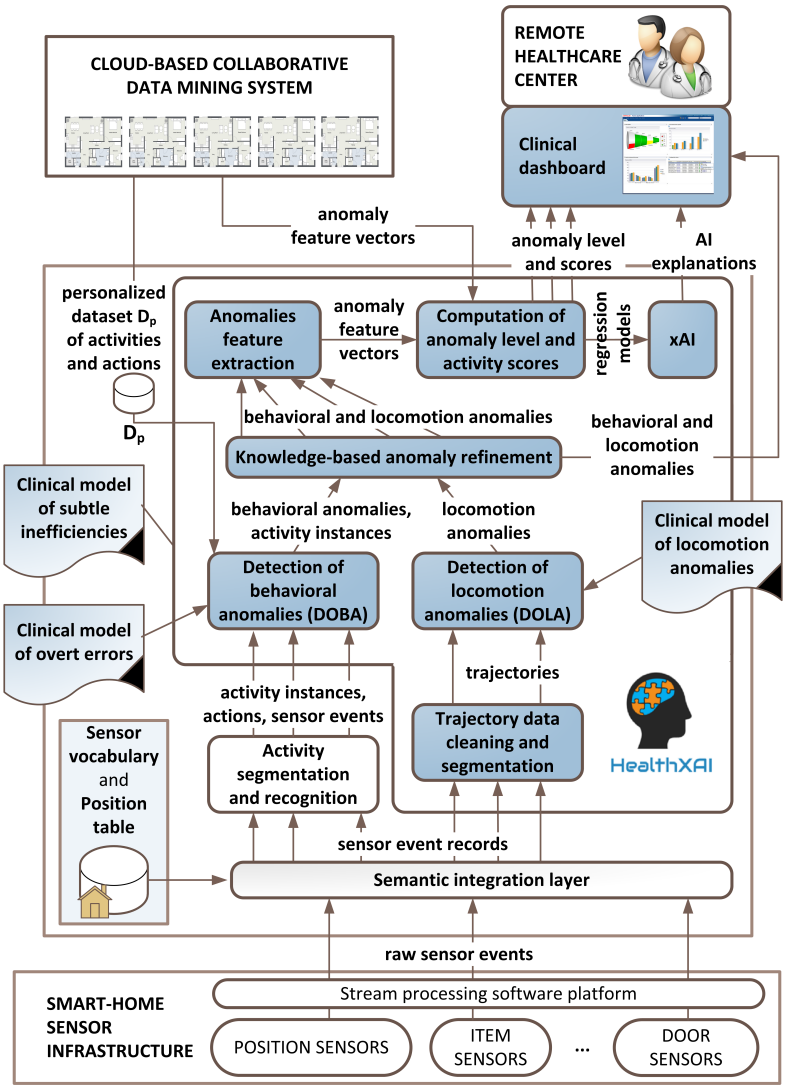

Figure 2: The overall HealthXAI architecture. Modules depicted in blue boxes represent the core technical contribution of this paper.

monitoring the inhabitant's activities (module ACTIVITY SEGMENTATION AND RECOGNITION) and movements (module TRAJECTORY DATA CLEANING AND SEGMENTATION). Detected activities include both high-level activities such as 'cooking' and fine-grained actions such as 'open fridge door'. Activities and motion data are analysed by the modules DOBA and DOLA, respectively, to detect

Module / data structure	Section
Activity instances and actions	4.1
Activity segmentation and recognition	3.5
AI explanation	5.2
Anomalies feature extraction	5.1
Anomaly feature vector	5.1
Anomaly level and scores	5.1
Behavioral anomalies	4.1
Clinical model of locomotion anomalies	2.2
Clinical model of overt errors	2.1.1
Clinical model of subtle inefficiencies	2.1.2
Computation of anomaly level and activity scores	5.1
Detection of behavioral anomalies (DOBA)	4.1
Detection of locomotion anomalies (DOLA)	4.2
Knowledge-based anomaly refinement	4.3
Locomotion anomalies	4.2
Personalized dataset D_p of activities and actions	4.1.1
Regression model	5.1
Semantic integration layer	3.3
Sensor event record	3.3
Sensor vocabulary and Position table	3.3
Stream processing software platform	3.2
Trajectory data cleaning and segmentation	3.4
Trajectory segments	3.4.2
xAI	5.3

Table 2: Lookup table of HealthXAI modules and data structures

anomalies according to the considered *clinical models*. A CLOUD-BASED COLLABORATIVE DATA MINING SYSTEM is in charge of collecting anonymous data about the activities and actions observed in different homes (named *personalized dataset* D_p), in order to calibrate the parameters of the DOBA module according to the context of the individual p. Detected activity instances and anomalies are analysed by the KNOWLEDGE-BASED ANOMALY REFINEMENT module to refine the locomotion anomaly predictions considering the context in which anomalies are observed. The latter module communicates refined anomalies to the module for ANOMALIES FEATURE EXTRACTION, which computes *anomaly feature vectors*. Those vectors are provided to the machine learning module COMPUTATION OF ANOMALY LEVEL AND ACTIVITY SCORES, which is in charge of providing anomaly scores for the different activities, as well as an overall anomaly level for the individual. The XAI module analyzes the trained *regression models* to compute detailed natural language explanations of the predictions, referring to the clinical models. The information about predictions, explanations, and fine-grained anomalies is communicated to a REMOTE HEALTHCARE CENTER, where clinicians can inspect the whole data through a user-friendly CLINICAL

DASHBOARD.

For the sake of readability, Table 2 indicates which sections explain the different modules and data structures of our system. In the rest of this section, we explain the modules for sensor data integration, and monitoring of activities and movements. The other modules, which are the core technical contribution of this work, are explained in Sections 4 and 5.

3.2. Smart-home sensor infrastructure

Since the focus of this paper in on processing activity and locomotion data, the core methods of our contribution are largely independent from the available sensor infrastructure. The HealthXAI system can be applied to a typical **SMART-HOME SENSOR INFRASTRUCTURE**, empowered with different kind of sensors to detect the position of the inhabitant and his/her interaction with furniture and appliances. Those sensors may include: passive infrared (PIR) motion sensors, or other sensors to detect the individual's position in the home; contact sensors to track the interactions with the apartment's doors or the use of furniture, such as cabinets or the fridge door; motion sensors attached to certain objects to detect their usage; power sensors to detect the use of certain electric appliances.

We assume that the smart-home system continuously acquires raw sensor data and communicates them to a STREAM PROCESSING PLATFORM (e.g., Apache Kafka) for raw integration and temporal synchronization. Each time a sensor fires, the platform sends a **raw sensor event** $rse = \langle t, s\_id, v \rangle$ to the HealthXAI system, where t is the timestamp of firing, s_id is the sensor's unique identifier, and v is the generated value. For the sake of this work, we assume that the home is inhabited by a single individual. This is a common situation for elderly people. Moreover, seniors living alone may have particular benefits from remote monitoring and assessment of cognitive functions. Hence, we are not interested in associating the sensor record to the person that triggered it.

3.3. Semantic integration layer

The **SEMANTIC INTEGRATION LAYER** is in charge of deriving higher-level information from raw sensor data. To this aim, it relies on a **sensor vocabulary**,

that provides the semantic description of each sensor given its identifier. We name the semantic description as *sensor type*. For instance, through the vocabulary, it may be possible to infer that a particular sensor identified as 'D01' is attached to the door of the fridge; i.e., its type is 'fridge door sensor'. Similarly, the vocabulary maps the emitted values to the semantics of sensed data (e.g., when a fridge door sensor emits the value '0', it means that the door is closed). Moreover, the system relies on a **position table** storing the relative position of each sensor in the home. A record of sensor position in that table is a triple: $\langle s\_id, x, y \rangle$, where (x, y) are the relative coordinates of the sensor identified by '*s_id*' in the home. Each time the semantic integration layer receives a raw sensor event, it joins the corresponding record with the sensor vocabulary and position table to obtain the (x, y) coordinates and the sensor semantics, producing a **sensor event record** $e = \langle p, t, s, v \rangle$, where $p = (x, y)$ are the relative coordinates of the sensor of type s that emitted a data value v at time t.

3.4. Trajectory data cleaning and segmentation

The smart-home system continuously collects the user's **position history** H; i.e., the temporal sequence of the user's positions within the home extracted from the sequence of sensor event records: $H = \langle p_1, p_2, \ldots, p_n \rangle$. The module for TRAJECTORY DATA CLEANING AND SEGMENTATION preprocesses the position history to reduce the noise level, and then partitions it in trajectory segments, which we denote as **trajectories** for short.

3.4.1. Trajectory data cleaning

In a real-world dense-sensing setup as the one considered in this work, position data may be affected by a high level of noise. For this reason, we adopt two methods to reduce the noise in H:

1. We assume a maximum possible velocity v of a person moving in the home. Hence, if the speed of movement computed between two consecutive sensors event records r_1 and r_2 is higher than v, we remove r_2 from H. For the sake of this work, we fix $v = 15\, m/s$.

2. In an indoor positioning system, we can assume a maximum resolution of position data. For instance, in our test-bed, the maximum distance between adjacent positioning sensors is 3 meters; hence, we can assume that, in the absence of noise, the distance between two consecutive positions in H should not exceed 3 meters. Based on that, we fix a threshold d (in our experiments, $d = 5\,m$), and when we observe two consecutive sensors event records r_1 and r_2 for which the distance exceeds d, we remove r_2 from H.

3.4.2. Segmentation

The module for trajectory segmentation is in charge of partitioning H to identify **trajectory segments**; i.e., temporally contiguous sequences of positions which correspond to a locomotion episode. To this aim, the segmentation module considers temporal information about the person's positions. The position history H is partitioned into a set S of non overlapping trajectory segments: $S = \{s_1, s_2, \ldots, s_m\}$, where each segment $s \in S$ is a temporal sequence of consecutive positions: $s_i = \langle p_j, p_{j+1}, \ldots, p_k \rangle$.

For applying segmentation, the algorithm identifies active locomotion phases, defined as parts of H in which the time interval between any two consecutive motion sensor activations is less than a threshold. For the sake of this work, we use a temporal threshold of 60 seconds. The first segment is initialized with the first entry of H, and continues until the temporal distance between two consecutive activations of motion sensors is below 60 seconds. The next trajectory is initialized with the last entry of the previous trajectory, and continues until the threshold condition is met, and so on until the end of H.

3.5. Activity segmentation and recognition

The module for **ACTIVITY SEGMENTATION AND RECOGNITION** is in charge of processing the continuous stream of sensor event records in order to recognize the activities that are occurring in the home as well as the actions that compose those activities. The module includes algorithms to accurately identify

the start and end time of activities; this task is called *activity segmentation* in the literature. Several research efforts have been spent in the last two decades to devise algorithms for activity recognition and segmentation based on sensor data. Different effective solutions to this problem have been proposed, which adopt data-driven [44, 45], knowledge-driven [46], or hybrid methods [47, 48]. Notably, some existing methods also rely on personalization features addressed to people with disabilities [49]. Hence, in this paper, we assume the existence of an effective module for action/activity segmentation and recognition, but we do not make any assumption about the actual implementation of that module. Indeed, the goal of this work is the analysis of activity data for anomaly detection, not the recognition of the activities. For this reason, in the experimental evaluation, this module relies on the ground truth about activity and action recognition available in the labeled dataset.

4. Collaborative model-based analysis of activities and motion

In this section, we illustrate the algorithms for collaboratively recognizing behavioral and locomotion anomalies based on clinical models.

4.1. Detection of behavioral anomalies

The module for DETECTION OF BEHAVIORAL ANOMALIES (DOBA) processes the stream of activity instances, actions, and sensor events, to compute a set of **behavioral anomalies** according to the **clinical model of subtle inefficiencies** or to the **clinical model of overt errors**. Since contextual conditions may affect the mode of activity execution, even in the absence of cognitive impairment, the parameters of those models are personalized, as explained in Section 4.1.1. As illustrated in Section 5, behavioral anomalies are later used by a regression algorithm to compute an overall anomaly level, as well as anomaly scores for activity instances. The HealthXAI algorithms also produce a natural language explanation for the predictions.

As discussed in Section 2.1, in this paper we concentrate on the recognition of overt errors (omissions, action-additions, and commissions), and 'reach-touch'

subtle inefficiencies. For the sake of clarity, in the rest of the paper we use the term **activity** to indicate high-level behaviors such as 'cooking' or 'taking medicines'. We name **actions** those simpler behaviors that are performed in order to execute an activity; e.g., 'turning on the stove', or 'opening the medicine cabinet'. We name **activity class** an abstract activity, such as 'cooking', while we name **activity instance** the actual occurrence of an activity of a given class during a certain time period.

4.1.1. Collaborative mining of personalized models of abnormal behaviors

In previous works, personalized behavioral anomalies were manually defined by domain experts based on common-sense knowledge using a rule-based language [5]. However, that approach could hardly scale with the number of considered individuals, number of activities, and variety of contextual conditions that may occur during the activity execution. In order to overcome this problem, in HealthXAI we adopt a collaborative approach to automatically derive the actions and parameters that determine an anomaly according to the considered clinical models for a certain individual. In particular, we rely on a **CLOUD-BASED COLLABORATIVE DATA MINING SYSTEM**, which collects anonymous trajectories, activity instances, actions, and sensor events from a set of collaborating smart-homes. The data are associated to the sanitized general characteristics of the inhabitant, such as age and physical condition. We do not consider the cognitive diagnosis of the individual, both for the sake of privacy, and because the diagnosis could be unavailable for several inhabitants. Those data are aggregated into a labeled dataset D. In order to derive the personalized models of an individual p, a subset of the dataset, named **personalized dataset D_p of activities and actions**, is communicated to the DOBA module of the HealthXAI system of p. That dataset contains only the activity data acquired from people having a profile similar to the one of p; i.e., similar age range and similar physical condition. Hence, the DOBA module of p mines D_p to extract personalized parameters used to detect p's behavioral anomalies.

4.1.2. Personalized recognition of omissions

An omission occurs when an individual does not perform a key action of an activity. As a consequence, omission recognition should rely on a model describing both the activities and the key actions that compose them. To this aim, in order to derive the personalized model of omissions for an individual p, we adopt a statistical approach, and we mine the co-occurrence frequency of $\langle$ activity, action $\rangle$ pairs from the dataset D_p. In particular, for each activity class a, and for each action type c, we compute the percentage of times that an instance of action c is executed at least once during the execution of an activity of class a in D_p. If the percentage exceeds a certain threshold t_o (e.g., $t_o = 95\%$), c is considered a key action for activity a for the individual p. Hence, when the DOBA module recognizes that during the execution of an instance of a the individual p did not perform c, it detects an omission.

4.1.3. Personalized recognition of action-additions

An action-addition occurs when, during an activity instance, the individual performs an action that is not related to that activity. In order to recognize this kind of anomalies, we take the same approach used for detecting omissions. Indeed, for each activity class a, and for each action type c, we compute the percentage of times that an instance of action c is executed at least once during the execution of an activity of class a in D_p. If the percentage is below a certain threshold t_{aa} (e.g., $t_{aa} = 2\%$), c is considered an action unrelated to activity a for p. Hence, when the DOBA module observes that during the execution of an instance of a the individual p performed the action c, it detects an action-addition.

4.1.4. Personalized recognition of anticipation-omissions

An anticipation-omission occurs when an individual performs actions in inappropriate order during a given activity instance. In order to infer whether a given temporal sequence of actions $s = \langle c_i, c_j \rangle$ is inappropriate for an activity of class a executed by p, we count the number of activity instances of class a

in D_p in which s is observed, and the number of activity instances of class a in which s is observed in reverse temporal order. We name ns the former number, and ns' the latter number. If both ns and ns' are equal to zero, we disregard the sequence s, because c_i and c_j were never observed in the same instance of activity a, despite their order of execution. Otherwise, we compute the ratio $r_s = \frac{ns}{ns + ns'}$. Hence, when the DOBA module observes the execution of $\langle c_i, c_j \rangle$ by p during an activity a, and the corresponding value of r_s is below a certain threshold t_{ao}, it concludes that p performed an anticipation-omission.

4.1.5. Personalized recognition of perseverations

A perseveration occurs when, during an activity instance, an individual repeats the same action more times than expected. We assume that the number of action executions during an activity follows a normal distribution in D_p. Hence, for each activity class a and for each action type c, we measure the standard deviation and average of times in which instances of c are executed during instances of a in D_p. When, during the execution of an instance of a by p, the DOBA module observes a number of repetitions of c that is statistically larger than expected according to the normal distribution considering a given threshold, it detects a perseveration episode.

4.1.6. Personalized recognition of 'reach-touch' subtle inefficiencies

A reach-touch subtle inefficiency occurs when an individual reaches for and touches an object that is not related to the execution of the current activity. In order to infer whether an object o is unrelated to a given activity class a for p, we measure the percentage of times that o is touched at least once during an instance of a in D_p. If the percentage is lower than a given threshold t_{rt}, we conclude that o is unrelated to a for p. Hence, when the DOBA module observes that during the execution of an instance of a the individual p reaches for and touches the object o, it detects a reach-touch subtle inefficiency.

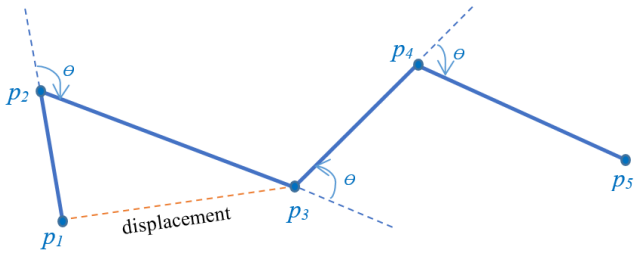

Figure 3: Example of part of a trajectory.

4.2. Detection of locomotion anomalies

The objectives of the **DETECTION OF LOCOMOTION ANOMALIES (DOLA)** module can be divided into **locomotion anomaly** detection and evaluation, where the aim of the latter is to characterise the detected anomalous trajectory according to the **clinical model of locomotion anomalies**. For each trajectory, we compute a set of indicators: temporal duration, traveled distance, and the anomaly indexes explained below.

4.2.1. Jerk

Jerk is defined as the rate of change in the acceleration of a person's trajectory. Jerk proved to be statistically correlated to the cognitive status of the subject [12]. In order to compute the jerk value of a trajectory s, we need to compute the jerk values of any three consecutive positions $\langle p_1, p_2, p_3 \rangle$ in s. Figure 3 illustrates a trajectory s composed of five consecutive points $\langle\ p_1, \ldots p_5\ \rangle$. Denoting the time difference between p_1 and p_2 as Δt, we calculate the following measures:

$$S_{p_1} = \frac{distance(p_1, p_2)}{\Delta t}, \tag{1}$$

$$Acc_{p_1} = \frac{S_{p_2} - S_{p_1}}{\Delta t}, \tag{2}$$

where S_p, Acc_p denote speed and acceleration, respectively. Then, we compute jerk as:

$$Jerk_{p_1} = \frac{Acc_{p_2} - Acc_{p_1}}{\Delta t}. \tag{3}$$

For computing the overall jerk of a trajectory $s = \langle p_1, p_2, \ldots, p_n \rangle$, we compute the values of jerk from $Jerk_{p_1}$ to $Jerk_{p_{n-2}}$. Then, we compute the overall jerk of s as:

$$Jerk_s = \frac{\sum_{i=1}^{n-2} \left| Jerk_{p_i} \right|}{T}, \tag{4}$$

where T is the temporal duration of s. This value is the ratio between the sum of absolute values of jerk and the time duration of the trajectory.

4.2.2. Sharp angles

According to [13], a sharp angle is defined as a vector angle in a trajectory being equal to or more than 90°. For instance, in the trajectory depicted in Figure 3, the angle corresponding to the movement $\langle\ p_1, p_2, p_3\ \rangle$ is a sharp angle, while the other ones are not. For every three consecutive positions $\langle\ p_{i-1}, p_i, p_{i+1}\ \rangle$ in a trajectory, we denote by θ the angle between the line connecting p_{i-1} with p_i and the line connecting p_i with p_{i+1}. Then, we compute the cosine of θ as:

$$\begin{aligned} cos\theta =& \frac{\overrightarrow{p_{i-1}p_i} \cdot \overrightarrow{p_i p_{i+1}}}{|\overrightarrow{p_{i-1}p_i}| \cdot |\overrightarrow{p_i p_{i+1}}|} = \\ &= \frac{(x_i - x_{i-1}, y_i - y_{i-1}).(x_{i+1} - x_i, y_{i+1} - y_i)}{\sqrt{((x_i - x_{i-1})^2 + (y_i - y_{i-1})^2)} \cdot \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2)}}. \end{aligned} \tag{5}$$

Based on their cosine value, we determine whether the three consecutive points $\langle\ p_{i-1}, p_i, p_{i+1}\ \rangle$ of s determine a sharp angle. If so, the value of $SharpAngle_{p_i}$ is 1; it is 0 otherwise.

In order to provide an overall measure of sharp angles for a trajectory s

$= \langle\ p_1, p_2, \ldots, p_n\ \rangle$, we compute the ratio of the number of sharp angles in a trajectory s and the temporal duration of s:

$$SharpAngle_s = \frac{\sum_{i=2}^{n-1} \left| SharpAngle_{p_i} \right|}{T}, \tag{6}$$

where T is the trajectory temporal duration.

4.2.3. Straightness

Straightness is defined as the ratio of the distance between two consecutive trajectory segments and the distance between the start and end point of these segments [11]. In the context of human trajectories, if a person does not change the orientation of movement, this ratio is 1. Otherwise, based on the amplitude of the turning angle, this indicator varies and indicates a person's tendency to erratic movements. The value of straightness for three consecutive points $\langle\ p_{i-1}, p_i, p_{i+1}\ \rangle$ is computed as:

$$Straightness_{p_i} = \frac{distance(p_{i-1}, p_i) + distance(p_i, p_{i+1})}{distance(p_{i-1}, p_{i+1})}. \tag{7}$$

In order to compute the overall value of straightness of a trajectory $s = \langle\ p_1, p_2, \ldots, p_n\ \rangle$, we compute the overall straightness of its points normalized by trajectory duration:

$$Straightness_s = \frac{\sum_{i=2}^{n-1} \left| Straightness_{p_i} \right|}{T}, \tag{8}$$

where T is the trajectory duration.

4.2.4. Martino-Saltzman indicators

The objective of this algorithm is to identify wandering episodes in trajectory segments. In a segment, a loop is defined as the largest continue sequence of positions $\langle\ p_1, p_2, \ldots, p_n\ \rangle$ such that the coordinates of p_1 and p_n are approximately the same. The method for finding loops relies on the computation of

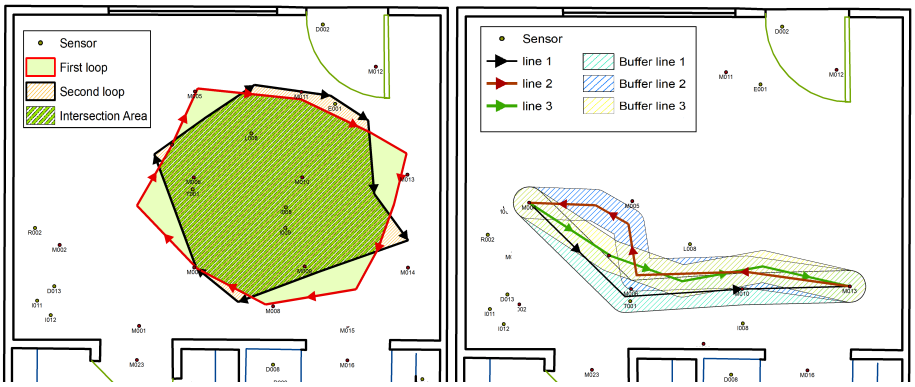

Figure 4: Detection of lapping and pacing episodes. Left figure (lapping): The arrows indicate the person's movements. The movements determine two loops, represented as first and second polygon. If the intersection of loops exceeds a given threshold, a lapping episode is detected. Right figure (pacing): The arrows indicate the person's movements from an origin to a destination. A buffer is built on the linear location traces. If the intersection of buffers exceeds a given threshold, a pacing episode is detected.

the linear misclosure of partial coordinates. In the first step, we calculate the sum of the positive and negative partial coordinates for x coordinates (Δx) and y coordinates (Δy) for each two consecutive positions. A sequence of locations is a loop if and only if the algebraic sum of both Δx and Δy is close to 0 up to an approximation error ϵ. In this work, we set ϵ to 0.3 m. The following steps summarize the algorithm for finding loops:

1. Calculate $\sum \Delta x$ and $\sum \Delta y$ (the algebraic sum of partial coordinates in episodes).

2. The sub-segment is a loop iff $\sum \Delta x < \epsilon$ and $\sum \Delta y < \epsilon$.

Lapping. For taking into account uncertainty in lapping episode detection, we consider the spatial overlap between consecutive loops. If the spatial overlap between loops exceeds a given threshold, we classify the episode as lapping. In this work, we fix the threshold to 70%. Figure 4 (left) illustrates our method. In particular, the consecutive loops extracted from a trajectory are compared by spatial intersection. If the latter is greater than the threshold, we predict a lapping episode according to the Martino-Saltzman model.

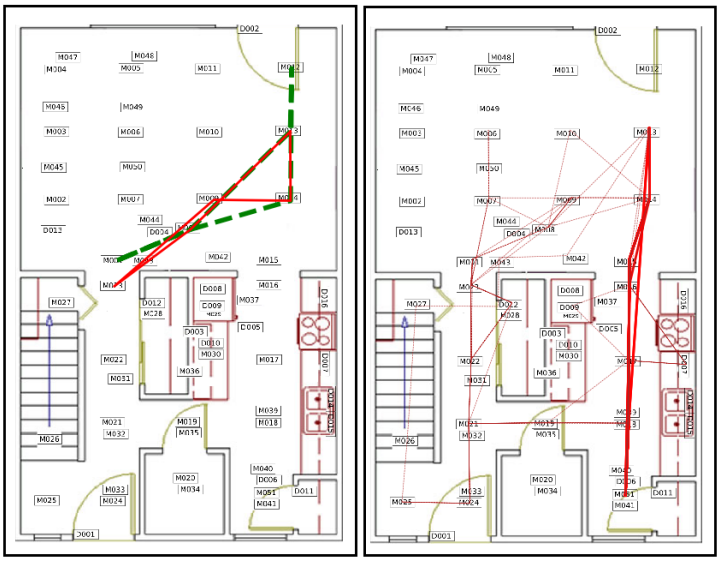

Figure 5: Example of *Pacing* and *Lapping* in trajectories. Lines represent a person's trajectories. The left image shows a lapping episode. The green dashed line shows the first polygon, and the red one shows the second polygon. It can be observed that they have more than 70% overlap. The right image is a sample of pacing from the fridge to a position in the dining room.

Pacing. Based on the Martino-Saltzman definition for pacing, a trajectory should have at least three back-and-forth movements between two locations. Hence, for recognizing pacing episodes in a trajectory, we need to consider the origin and destination of locomotion. However, because of the presence of obstacles in the home, the person's movements have some obligation. Hence, we must take into account uncertainty. For detecting pacing, a small deviation from the straight path (e.g., moving around a chair to go from the kitchen to the dining room) is not important and should be disregarded. For this purpose, we simplify the person's trajectory in order to smooth irrelevant deviations due to obstacles. To this aim, we use the well-know Douglas-Peucker polyline simplification algorithm [50].

The aim of the Douglas-Peucker algorithm is finding a curve similar to the original one but with fewer points. The algorithm relies on a point-to-edge distance tolerance value. Dissimilarity is measured based on Hausdorff Distance

between the original and simplified curve. This algorithm works by connecting the beginning vertex and end vertex in the trajectory, and calculating the distance between the other vertexes in the trajectory. If the maximum distance is less than the distance tolerance value ϵ, all the vertexes in the list are deleted. Otherwise, that vertex is retained, and the line is split into two curves, and the procedure is repeated. The value ϵ is the only input parameter of the algorithm. In this work, we set ϵ to 1.2.

After simplification, we can detect pacing episodes by considering the spatial overlap between the walked paths from origin to destination. In particular, we follow these steps:

1. Simplify trajectory by applying Douglas-Peucker algorithm.
2. Create a spatial buffer with a certain radius r for any two consecutive points in the episode (in this work, $r = 1\,m$).
3. Find any 3 consecutive sub-trajectories in the trajectory such that the spatial overlap between theirs buffer is higher than a threshold (in this work, the threshold is set to 60%).

Figure 4 (right) graphically illustrates the method, while Figure 5 shows two examples of pacing and lapping episodes found in a person's trajectory of the dataset used in this work.

4.3. Knowledge-based anomaly refinement

As explained before, behavioral and locomotion anomalies are *separately* detected by the DOBA and DOLA modules, respectively. However, in real world conditions, locomotion is naturally related to the ongoing activity. For instance, the normal execution of 'sweeping' may determine a random-like walk in the house, while the execution of 'setting up table' may determine pacing-like motion between the kitchen and the dining table. Hence, the DOLA module may classify as abnormal some locomotion episodes which are actually due to the normal execution of everyday activities.

In order to alleviate this problem, HealthXAI includes a module for **KNOWLEDGE-BASED ANOMALY REFINEMENT**, whose goal is to refine the locomotion anomaly predictions by considering the context in which they are observed. The module relies on a matrix M of *possible locomotion anomalies*, whose rows correspond to activity classes, and columns correspond to locomotion anomaly types. The value of $M_{i,j}$ equals to 1 if the anomaly corresponding to column j is a possible anomaly when observed during the activity corresponding to row i; $M_{i,j}$ equals to 0 otherwise.

In order to refine the set L of locomotion anomalies of a person p, the module relies on the set A of p's activity instances. For each $l \in L$ and $a \in A$, the module computes the intersection between the temporal intervals of both. If a non-null intersection exists, the module retrieves the value of M corresponding to l's activity class and a's anomaly type. If that value is 0, the anomaly is identified as not possible; hence, it is removed from L.

Currently, the matrix M is manually filled by domain experts based on external knowledge about activity patterns and locomotion anomalies, adopting a knowledge-based approach. Alternatively, the matrix M could be instantiated applying some data mining technique to a large dataset of activities and anomalies. However, due to the lack of such large dataset for evaluation, in this paper we pursue the knowledge-based approach. Section 6.4 reports the implementation of M used in our experiments.

5. Explainable detection of cognitive decline symptoms

In this section, we explain how HealthXAI computes activity anomaly scores and an overall anomaly level based on the history of detected candidate anomalies, and how it produces a natural language explanation of its predictions.

5.1. Computing the overall anomaly level

The detailed description of possible anomalies performed by an individual may provide useful information to clinicians for evaluating his/her cognitive

Feature name	Description
Additions	Number of action additions
Anticipation-Omissions	Number of anticipation-omissions
Omissions	Number of omissions
Perseverations	Number of perseverations
Reach-touch	Number of reach-touch subtle inefficiencies
Pacing	Number of pacing episodes
Lapping	Number of lapping episodes
Random	Number of random walk episodes
Jerk	Average jerk of trajectories
Straightness	Average straightness of trajectories
Sharp-points	Average number of sharp points in trajectories
Anomaly-level	Anomaly level in [01]

Table 3: Anomaly features

status. However, a global assessment of the cognitive status of the subject, automatically provided by the system through an overall **anomaly level**, may be important to raise alarms and to provide clinicians with a general overview of the patient's situation.

In order to compute the anomaly level for a person p, the module for COMPUTATION OF ANOMALY LEVEL AND ACTIVITY SCORES relies on a set of anonymous **anomaly feature vectors** gathered by the cloud-based collaborative data mining system from participants having a profile similar to the one of p. Those vectors are computed by the ANOMALIES FEATURE EXTRACTION module by querying the anomalies communicated by the module for KNOWLEDGE-BASED ANOMALY REFINEMENT. Anomaly feature vectors also include an anomaly score for each activity performed by the individual. In particular, each **anomaly feature vector** is composed of the features shown in Table 3.

The latter feature is computed based on the actual diagnosis of the cognitive status of the participant (i.e., cognitively healthy, MCI, or PwD), as explained in the following. The remaining features are computed based on the number of abnormal behaviors of the participant detected by HealthXAI in the last n days, where the value of the n parameter may be chosen by clinicians according to their requirements. In this work, due to the characteristics of the dataset used in our experiments, we set n to 1. The computed anomaly feature vectors are communicated to the HealthXAI system of p to infer his or her anomaly level.

In order to compute the anomaly level for an individual p, HealthXAI relies on supervised learning, and more specifically on regression [51]. We rely on regression instead of classification, because in the neuropsychology literature the classes of cognitively healthy persons and MCI persons, as well as those of MCI persons and PwD, are not strictly separated [14]. Hence, HealthXAI avoids providing a strict classification of the individual's cognitive status. Instead, it provides an anomaly level $l \in [0, 1]$, where the value 0 characterizes cognitively healthy subjects, and 1 characterizes PwD. According to the medical literature, people with MCI are still able to complete normal activities, while PwD are not [14]. As a consequence, it is reasonable to assume that the number of anomalies performed by people with MCI is closer to the one of cognitively healthy persons than to the one of PwD. Hence, in the feature vectors, we set the value $l = 0.3$ to characterize the anomaly level of individuals with MCI, while we set the anomaly level of cognitively healthy people to 0, and the one of PwD to 1.

In order to learn the **regression models**, HealthXAI uses the anomaly feature vectors and a Decision tree regression algorithm [52]. That algorithm is non-parametric, and is able to learn a model predicting the value of a target variable (in our case, the value l related to a given person) by constructing decision rules taking into account the feature values. Decision trees have large application in the medical domain [53]. In several studies, decision trees overcame other popular machine learning algorithms for different medical problems. For instance, in the diagnosis of MCI based on semantic information extracted from magnetic resonance imaging, decision trees achieved higher accuracy than Support Vector Machines, Bayesian networks, and backpropagation neural networks [54]. Overall, our experimental evaluation, reported in Section 6, shows that decision trees provide high accuracy for our problem compared to other machine learning algorithms.

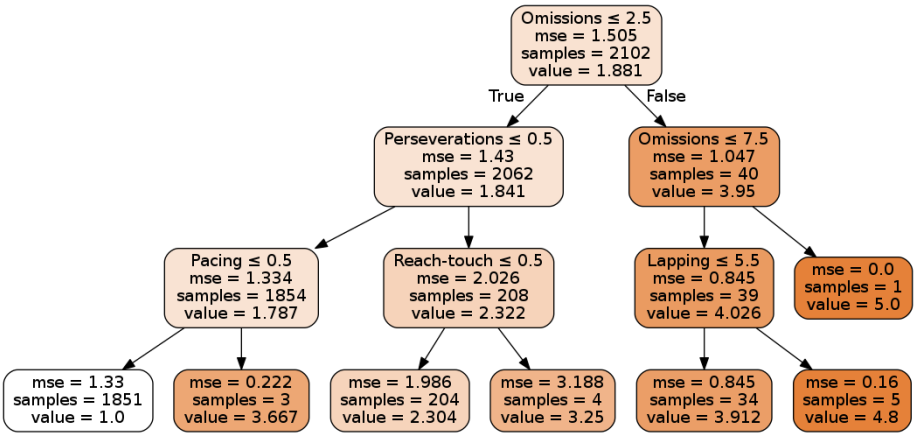

Figure 6: The learned model of a random tree used for the prediction of the anomaly score of an activity. The regression algorithm evaluates the rules conditions based on the feature vector of the patient's activity, starting from the root. If the condition is verified, it evaluates the condition of the left child; otherwise, it evaluates the one of the right child. The mechanism is repeated until a leaf is reached. The leaf contains the predicted value.

5.2. Computing the anomaly score of activities

In order to assess the cognitive status of a patient, it is useful to evaluate his/her abilities in performing activities of daily living [55]. Hence, when HealthXAI detects at least one candidate anomaly during a given activity instance ai, our system computes an **anomaly score** $s \in [1, 5]$ for ai, where 1 indicates no anomaly, and 5 indicates strong anomaly in the execution of ai. This refinement is performed in order to reduce the number of false positives; i.e., activity instances for which the AI algorithm detects some anomaly, which did not actually happen.

For each activity instance, HealthXAI computes a feature vector using the same features presented in the Section 5.1, plus an additional feature, which is the temporal duration of the activity instance. A Decision tree regression algorithm trained using the anomaly feature vectors is in charge of computing the anomaly score of an activity based on the corresponding feature vector.

5.3. Generating natural language explanations

Being based on a set of rules arranged in a hierarchy, decision trees have the advantage of being relatively easy to visualize and interpret. Figure 6 shows the

model of a decision tree regression algorithm that provides an anomaly score for a given activity based on the number of different kinds of anomalies recognized during its execution. Of course, the size of the tree strongly influences the readability of the model. In general, the deeper the tree, the more difficult is to understand the chain of reasons why a given prediction is computed. In our experimental evaluation, we found that the maximum depth of the most effective models is rather small. This fact allows the generation of easily understandable natural language explanations of HealthXAI predictions.

Considering the learned model and the feature vector values, the **xAI** module is in charge of generating a natural language description of each prediction of the decision tree, named **AI EXPLANATION**. At the beginning, considering the predicted value, the algorithm outputs a generic description of the corresponding anomaly level; e.g., 'The predicted anomaly level of the activity is 2.3. This is considered a mild anomaly level'. Then, the algorithm parses the learned decision tree to explain the reason of the prediction. The core of the xAI algorithm is shown in Figure 7. Formally, the algorithm takes as input the learned decision tree T and an anomaly feature vector v. T is a binary tree, represented as an array $[n_0, \ldots, n_k]$, where each n_i represents a node i. While leaves contain only their predicted value and some statistics, each non-leaf node contains these fields:

- $feat$ is the feature used for splitting the node;
- thr is the threshold value of the node;
- $child\_l$ is the index of the left child of the node;
- $child\_r$ is the index of the right child of the node.

For instance, in the tree shown in Figure 6, the root node has $feat =$ 'Omissions' and $thr = 2.5$. Hence, if the value of v[Omissions] is greater than 2.5, the decision path for v follows the path of $child\_r$; it follows the path of $child\_l$ otherwise.

At first (line 1), starting from the root node n_0, the algorithm traverses the tree structure to identify the decision path related to v. The nodes belonging to the decision path are stored in the array *indices*. The set of thresholds is

Algorithm ***xAI***(T, v):
Input: Decision tree T, anomaly feature vector v
Output: A string s of explanation

```
 1: indices = get_decision_path(T, v)
 2: thresholds = ∅
 3: for each feature feat do
 4:     thr_min = get_minimum_threshold(indices, feat, v)
 5:     thr_max = get_maximum_threshold(indices, feat, v)
 6:     thresholds = thresholds ∪ ⟨ feat, thr_min, thr_max ⟩
 7: end for
 8: s = 'Explanation: the individual did '
 9: for each ⟨ feat, thr_min, thr_max ⟩ ∈ thresholds such that thr_min is not null do
10:     s.append('at least ' thr_min feat)
11: end for
12: s.append('but ')
13: for each ⟨ feat, thr_min, thr_max ⟩ ∈ thresholds such that thr_max is not null do
14:     s.append('less than ' thr_min feat)
15: end for
16: return s
```

Figure 7: Algorithm for generation of natural language explanations.

initialized (line 2). Then (lines 3 to 7), the algorithm parses the *indices* nodes to determine the minimum and maximum thresholds (thr_min and thr_max, respectively) for each feature related to v's decision path. To this aim, it stores a triple $\langle feat, thr\_min, thr\_max \rangle$, where the latter two values are updated according to the traversal order. For instance, suppose that v[Omissions] = 10, and the T root condition is 'Omissions $\leq$ 2.5'. In this case, the algorithm computes the triple $\langle$Omissions$, 2.5, null\rangle$, meaning that the current decision path was determined by v[Omissions] being greater than 2.5. Then, suppose that the next node's condition is 'Omissions $\leq$ 7.5'. In this case, the algorithm updates the triple to $\langle$Omissions$, 7.5, null\rangle$, since the new thr_min value is more stringent than the one previously stored in the triple.

After computing thresholds, the algorithm initiates the explanation string s (*'Explanation: the individual did ... '*, line 8). Then, it parses the computed threshold triples to explain the reason of prediction. It starts with all triples having non-null values of thr_min (adding *'at least $thr_min $feat'*; e.g., *'at least 8 omissions'*, lines 9 to 11) and continues with all triples having non-null values of thr_max (adding *'but less than ... '*, lines 12 to 15). The actual algorithm also applies some fine-grained refinements, not reported in the pseudo-

code for lack of space, to improve the readability of the explanations (e.g., handling punctuation, rounding thresholds).

As an example, consider the model shown in Figure 6, and suppose that the current individual, during the activity 'cooking', performed one perseveration but no other anomaly. Based on that, the chain of decisions would bring to the third leaf, which would predict the anomaly level 2.304. In this case, the AI EXPLANATION would produce the following explanation: *'The predicted anomaly level of the activity is 2.3. This is considered a mild anomaly level. Explanation: the individual did less than 3 omissions, but at least one perseveration, and no reach-touch inefficiency'.*

6. Prototype and experimental evaluation

In this section, we illustrate the experimental evaluation that we carried out with real-world data acquired in smart-homes inhabited by a large set of people, including persons with MCI and PwD. The dataset was acquired and labeled by researchers of the Center for Advanced Studies in Adaptive Systems[2] (CASAS) of Washington State University. For the sake of privacy, before releasing the data, the researchers removed explicit identifiers of the individuals involved in the study. Moreover, quasi-identifier personal data such as age have been generalized to avoid identity disclosure. The exact age has been replaced by age ranges, such as 'middle age (45-59)', 'young-old (60-74)', and 'old-old (75+)'. The protocol of recruiting and data collection was approved by the Institutional Review Board of WSU [56].

6.1. Test-bed

The CASAS smart-home test-bed has been used in several studies on dementia and other cognitive disabilities, in order to investigate the relationship between ordinary activities and memory abilities [57]. The CASAS test-bed is a two-story apartment equipped with several passive infrared (PIR) motion

[2] `http://casas.wsu.edu/`

sensors mounted on the ceiling to track the user's position, item sensors for detecting the usage of selected items in the kitchen, door sensors, burner sensors, hot and cold water sensors, temperature sensors, and whole-apartment electricity usage meter. The average distance of PIR sensors is about one meter throughout the apartment. In total, the smart-home includes 52 motion sensors, 18 door sensors, and 10 item sensors. The first floor of the apartment consist of a kitchen, living room and a dining area, while bathroom and two bedrooms are located at the second floor. Figure 8 shows the layout of the CASAS smart-home. We use this test-bed dataset to reproduce the case of a residence for senior people, whose apartments have identical shape and sensors, but are inhabited by different persons.

6.2. Participants recruiting and cognitive evaluation

WSU researchers recruited the dataset participants through advertisements, physician referrals, and from people who took part to past studies in WSU laboratories. Recruitment and data collection were carried out over two years. In total, 400 individuals were recruited. After obtaining informed consent, participants underwent multidimensional clinical assessment, including both standardized and experimental neuropsychological tests. The cognitive health status of each participant was diagnosed by neuropsychologists. As a result of the evaluation, participants were classified in three categories, as shown below.

- Participants in the Dementia (category D) group met the DSM-IV-TR diagnostic criteria described in [58]. Those criteria include the observation of several cognitive deficits that negatively affect the normal execution of activities of daily living, and represent a decline from the previous health status of the senior. 36 Participants were classified in this group.
- Participants in the MCI category met the criteria defined by Petersen in [59] and those defined by the National Institute on Aging - Alzheimer's Association [60]. MCI is considered an intermediate stage between normal cognitive ageing and dementia. People with MCI have more memory problems than cognitively healthy seniors, but are still able to carry out their

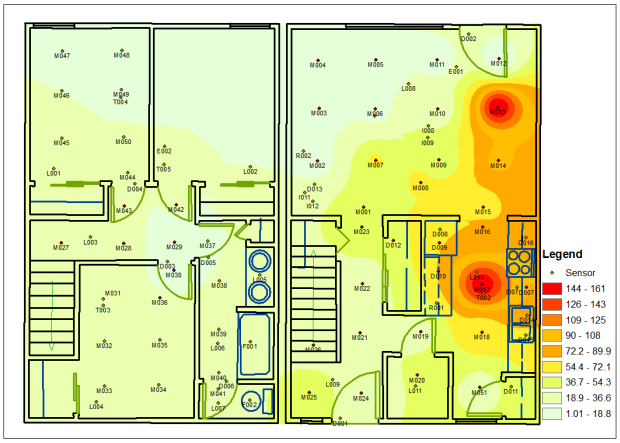

Figure 8: The CASAS testbed evnivorment [57]. The location of PIR and door sensors is represented by labeled dots. The heat map shows the distribution of activated sensors in the testbed for one of the patients in a day.

normal activities independently. This group included 59 participants.

- 255 participants were diagnosed as cognitively Healthy (category H). Among them, 37 were middle-aged persons (45-59 years old), 83 were young-old persons (60-74 years old), 44 were persons older than 75, and 91 were younger adults.

Among the remaining ones, 40 persons were diagnosed with "other conditions", and 10 persons had no diagnosis or were diagnosed as "at risk". Since we are interested in computing personalized models for elderly persons, we considered only data from seniors; i.e., participants in the categories D, MCI, and H whose age is 60 or older.

Moreover, some recruited participants, in particular PwD, were not able to execute activities in the test-bed for different reasons. Hence, we had to disregard them, since their activity data were absent from the dataset. We also disregarded patients who performed less than 6 activity instances, due to the lack of sufficient information for evaluating their cognitive status. Totally, we considered 192 subjects: 19 PwD, 54 people with MCI, 80 seniors aged from 60 to 74, and 39 elderly aged 75 or older.

6.3. Activity execution and data collection

After cognitive evaluation, participants were asked to execute so-called Day Out Tasks (DOTs) in the smart apartment. DOTs are defined as naturalistic tasks that require the ability to perform interleaved activities in order to reach a given objective. Those activities are at the core of competency in everyday life [61]. Activities were executed in a single day by each user according to a given script. Each DOT was described to the user as a complex activity composed by a number of simpler tasks. For instance, the complex activity 'Prepare a cup of soup using the microwave' consist of the following tasks: retrieve materials, measure water with cup, pour water to noodles, wait for water to simmer in the cup of noodles, retrieve and return pitcher of water from refrigerator, pour glass of water, and finally bring all items to the dining room table. The setup included 24 different DOTs. On average, the senior participants that we included in our experiments performed 15.2 DOTs (standard deviation = 1.8).

During the execution of DOTs, the smart-home infrastructure acquired the sensor data triggered by the activities and movements of the participant. For each sensor event, the dataset reports the following sensor record: $\langle\ t\ \ s\_id\ \ v\ \rangle$, where t is the timestamp of sensor firing, s_id is the sensor unique identifier, and v is the emitted value. The domain of the value v depends on the specific sensor; e.g., if a sensor is attached to a fridge door, the domain of v may be {'open', 'close'}. For instance, the record ⟨ 08:30:00.27063 M08 ON ⟩ represents the fact that the motion sensor M08 emitted the value ON at 08:30:00.27063. Note that, since each user executed his/her activities in a single day, the information about the day is not necessary. The detailed description of activities and tasks and the full dataset are available on the Web[3]. The setup of the smart-home is described in detail in [62].

[3] `http://casas.wsu.edu/datasets/assessmentdata.zip`

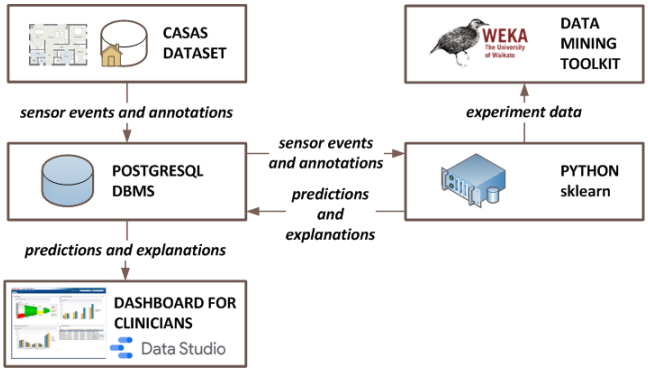

Figure 9: The software design of our experimental evaluation.

6.4. System prototype

The software design of our system prototype is depicted in Figure 9. The CASAS dataset, which is stored on different textual files, is imported into a PostgreSQL database through a Python script. PostgreSQL is a well known and widely adopted relational database management system (DBMS). The database stores very articulated and structured information, such as sensor events, sensor descriptions, participants, activities, actions, diagnoses, annotations, explanations. In total, the database schema includes 18 tables and 40 views. A small excerpt of the schema is reported in Figure 10.

We implemented the HealthXAI algorithms in Python. The code exploits the querying capabilities of PostgreSQL for efficiently computing the statistical information required by our algorithms. Through a Python script, we exported the data into the Weka [63] file format in order to evaluate different machine learning algorithms. Weka is a data mining toolkit that supports several algorithms, and allows rapid prototyping and experimentation of machine learning tasks. After the evaluation, we integrated the chosen machine learning algorithm in the Python code using the sklearn libraries[4]. Among the different variants of decision trees, which had similar accuracy in our experiments, we chose the

[4] `https://scikit-learn.org`

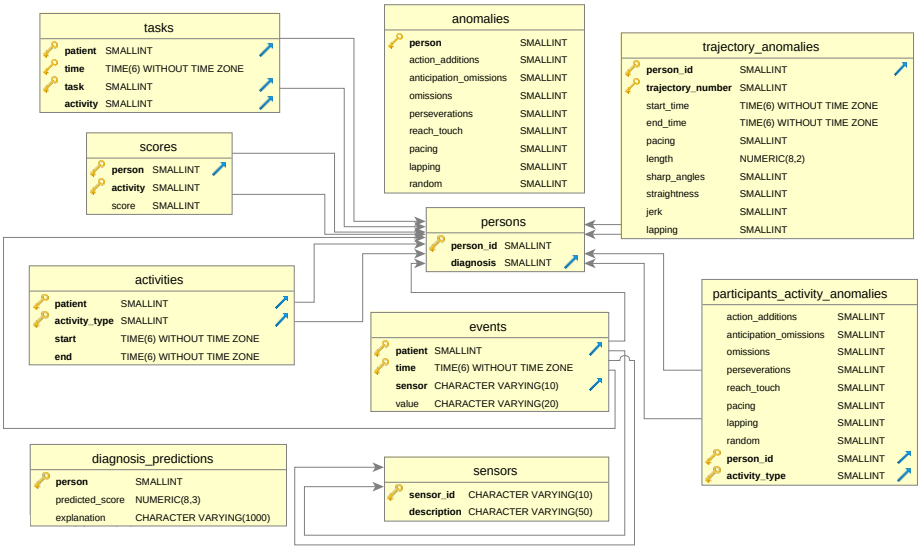

Figure 10: Part of the tables of HealthXAI database.

classical algorithm CART [52], which is known to be among the most effective tree-based learners. The M matrix for knowledge-based anomaly refinement was defined by a researcher by observing the typical locomotion patterns of cognitively healthy subjects. As a consequence, M was filled such that pacing, random walk, and lapping anomalies are considered not possible during the execution of the following activities: sweeping, water plants throughout the apartment, clean kitchen, sort and fold clothing, and prepare hot meal.

The results of HealthXAI (predictions and explanations) are stored in the database. An online **CLINICAL DASHBOARD** is deployed on the cloud using the Google Data Studio[5] platform. Through the dashboard, clinicians can query the database and inspect the relevant information and explanations.

In these experimental implementation, we used fixed values for the thresholds used for anomaly recognition explained in Section 4.1. In particular, we fixed the threshold t_o of omissions to 90%, the threshold t_{aa} of action-additions to 1%, the threshold t_{ao} of anticipation-omissions to 0.1. For recognizing perseverations,

[5] `https://datastudio.google.com`

Regressor	Correl. coeff.	Mean abs. err.	Root mean squared err.
5 Nearest Neigh.	0.253	1.122	1.646
Decision stump	0.431	1.025	1.229
Decision table	0.431	1.025	1.229
Linear regression	0.464	1.019	1.201
M5 model decision tree	**0.474**	1.014	**1.193**
Neural Net.	0.289	1.178	1.498
Random forest	0.454	**0.972**	1.224
Random tree	0.285	1.109	1.615
Red. Err. Prun. dec. tree	0.439	0.997	1.229
SVM	0.372	0.981	1.347

Table 4: Results of different regressors for the prediction of the anomaly score of activities. We considered five scores that vary from 1 (very accurate) to 5 (very inaccurate).

we used a threshold computed as $avg + 10 \cdot stddev$, where avg is the average number of action repetitions, and stddev is its standard deviation. Finally, we set the threshold t_{rt} of reach-touch inefficiencies to 10%.

The whole code of our system prototype, including the database, as well as the link to the dashboard that is freely accessible, is available on the Web[6]. For the sake of reproducibility, the code documentation includes the pseudo-identifiers of those individuals whose data have been used in our experiments.

6.5. Results of activity anomaly score prediction

In the CASAS dataset, the instances of 8 types of DOTs are annotated with a *score* that measures how accurately (considering correctness and completeness) they have been performed by the participant. The scores range from 1 (very accurate execution) to 5 (very inaccurate execution). The scores have been given by researchers observing the participant performing the activity. In total, the dataset includes 1208 activity instances with a score that were performed by the participants of our experiments. Among them, 716 include at least one anomaly according to the predictions of HealthXAI. Considering only the latter activity instances, the average score of seniors aged 60-74 is 1.87 ± 1.14; the one of elderly aged 75 or more is 2.18 ± 1.23; the one of MCI persons is 2.53 ± 1.36; the one of PwD is 3.84 ± 1.18. These statistics indicate that the score achieved by a person is significantly correlated to his/her cognitive health status.

[6] `https://sites.unica.it/domusafe/healthxai/`

We evaluated the accuracy of our method based on regression to compute the anomaly score of activities, as explained in Section 5.2. We used a leave-one-out cross validation approach, and we experimented with several among the most widely adopted regressors, including ones based on decision trees, Support Vector Machines (SVM), Neural Networks, lazy learners (i.e., kNN with $k = 5$), rules, and linear regression. Results are shown in Table 4. Among the evaluated regressors, the one achieving the best correlation coefficient r (i.e., the Pearson correlation coefficient [64]) was the M5' model decision tree algorithm [65]. With this algorithm, we obtained a correlation $r = 0.455$. This value indicates a moderate correlation between the predicted score and the ground truth. The mean absolute error was close to or below 1 for most regressors.

In general, we can observe that the decision tree regressors proved to be among the most effective regressors for this problem. Other popular algorithms, such as random tree, neural networks, and SVM, achieved considerably worse results. The relatively low correlation may be due to the intrinsic difficulty of the regression problem. Indeed, activity instances in the dataset are rather short, having an average duration of 3.19 ± 2.77 minutes. In more naturalistic conditions, many activities of daily living are carried out for longer. Hence, it is reasonable to expect achieving stronger correlation in more naturalistic setups. Moreover, while the achieved correlation may be insufficient to reliably evaluate the anomaly score of a single activity instance, it is reasonable to expect that on the long term the technique could provide more reliable predictions.

Classifier	**Accuracy**	**Precision**	**Recall**	F_1 **score**
5 Nearest Neigh.	0.499	0.497	0.499	0.49
Bayes net.	0.587	0.59	0.587	0.585
Decision table	**0.601**	**0.6**	**0.601**	**0.594**
Decision tree C4.5	0.564	0.567	0.564	0.547
Naive Bayes	0.57	0.577	0.57	0.549
Neural Net.	0.584	0.584	0.584	0.576
Random forest	0.575	0.575	0.575	0.569
Random tree	0.493	0.491	0.493	0.484
Red. Err. Prun. dec. tree	0.587	0.586	0.587	0.582
Ripper	0.557	0.568	0.557	0.543
SVM	0.589	0.59	0.589	0.586

Table 5: Results of different classifiers for the prediction of the anomaly score of activities. We considered three classes: no anomaly (score 1), mild anomaly (score 2 or 3), strong anomaly (score 4 or 5).

In other experiments, we evaluated a different approach to this problem. In particular, we treated the prediction of activity anomaly scores as a classification problem. We noticed that, in the dataset, the distribution of scores is not homogeneous. Indeed, 3 kinds of scores are frequent (score 1 having 774 instances; score 2 having 302 instances; score 4 having 308 instances), while the remaining ones are rather infrequent (score 3 having 57 instances; score 5 having 45 instances). Hence, initially we treated the problem as a 3-class classification problem, with one class 'no anomaly' composed of activity instances having score 1 (774 instances); one class 'mild anomaly' composed of activity instances having score 2 or 3 (359 instances); and one class 'strong anomaly' composed of activity instances having score 4 or 5 (353 instances). Results are reported in Table 5. The classifier achieving the best accuracy is the one based on decision table [66], with precision, recall, and F_1 measure of 0.603. In general, different classifiers (including the Decision tree that we used in our system) achieved similar performance, with F_1 measure close to 0.6.

Classifier	**Accuracy**	**Precision**	**Recall**	F_1 **score**
5 Nearest Neigh.	0.437	0.433	0.437	0.344
Bayes net.	0.521	N/A	0.521	N/A
Decision table	**0.539**	N/A	**0.539**	N/A
Decision tree C4.5	0.45	0.438	0.45	0.383
Naive Bayes	0.483	0.495	0.483	0.389
Neural Net.	0.535	**0.512**	0.535	**0.428**
Random forest	0.494	0.487	0.494	0.419
Random tree	0.425	0.415	0.425	0.34
Red. Err. Prun. dec. tree	0.527	0.496	0.527	0.366
Ripper	0.462	0.449	0.462	0.326
SVM	0.531	N/A	0.531	N/A

Table 6: Results of different classifiers for the prediction of the anomaly score of activities. We considered five classes that vary from 1 (very accurate) to 5 (very inaccurate).

We also performed experiments with five classes, that correspond to the original scores. In fact, there is a trade-off between granularity of scores and accuracy of predictions: with finer-grained scores (and consequently, larger number of classes), it is reasonable to expect lower recognition rates. Results with five classes of scores are reported in Table 6. We could not compute the value of precision and F_1 score for certain classifiers, since they never output a prediction for certain classes; hence, we marked the corresponding values as N/A (not

available). As expected, with this setup the accuracy of classifiers decreased, since more fine-grained classes are considered. The classifier achieving the best accuracy is the one based on multi-class logistic regression [67], with accuracy and F_1 measure close to 0.53. In general, different classifiers, including Decision trees, showed similar performance, with precision, recall, and F_1 measure close to the one of the logistic regression classifier.

Classifier	**Accuracy**	**Precision**	**Recall**	F_1 **score**
5 Nearest Neigh.	0.641	0.644	0.641	0.63
Bayes net.	0.687	0.688	0.687	0.676
Decision table	0.666	0.669	0.666	0.656
Decision tree C4.5	**0.709**	**0.709**	**0.709**	**0.697**
Naive Bayes	0.644	0.679	0.644	0.643
Neural Net.	0.683	0.681	0.683	0.668
Random forest	0.682	0.679	0.682	0.666
Random tree	0.648	0.645	0.648	0.631
Red. Err. Prun. dec. tree	0.701	0.697	0.701	0.679
Ripper	0.666	0.659	0.666	0.628
SVM	0.698	0.695	0.698	0.681

Table 7: Results of different classifiers for the prediction of the anomaly score of activities. We considered two classes: no anomaly (score 1), anomaly (score 2 to 5).

Finally, we performed experiments with only two classes of anomaly: no anomaly (that corresponds to the score 1), and anomaly (scores 2 to 5). Results are shown in Table 7. As we expected, with this setup the accuracy of classifiers increased. The algorithm achieving the highest accuracy was the decision tree classifier C4.5 [68], which achieved precision, recall, and F_1 measure higher than 0.7. In general, we can conclude that Decision treed are among the most effective machine learning algorithms both for regression and classification problems in the considered domain.

6.6. Results of personal anomaly level prediction

We carried out experiments to evaluate the HealthXAI mechanism for computing the overall anomaly level of individuals explained in Section 5.1. At first, we evaluated the prediction performance of different regressors using the whole set of participants. Results are reported in Table 8.

The regressor achieving the highest correlation coefficient ($r = 0.499$) is the one based on SVM. The M5' model decision tree and linear regression algorithms

Regressor	Correl. coeff.	Mean abs. err.	Root mean squared err.
5 Nearest Neigh.	0.232	0.226	0.373
Decision stump	0.328	0.207	0.292
Decision table	0.328	0.207	0.292
Linear regression	0.454	0.2	0.272
M5 model decision tree	0.43	0.206	0.276
Neural Net.	0.308	0.245	0.385
Random forest	0.399	0.207	0.278
Random tree	0.231	0.231	0.372
Red. Err. Prun. dec. tree	0.284	0.219	0.298
Simple linear regr.	0.487	0.196	**0.263**
SVM	**0.517**	**0.174**	0.276

Table 8: Results of different regressors for the prediction of the overall anomaly level of individuals. We considered three profiles of individuals: healthy seniors aged 60 or more (119 individuals, anomaly score = 0.0), elderly with MCI (53 individuals, anomaly score = 0.3), and PwD (19 individuals, anomaly score = 1.0).

Regressor	Correl. coeff.	Mean abs. err.	Root mean squared err.
5 Nearest Neigh.	-0.017	0.128	0.196
Decision stump	0.128	0.122	0.14
Decision table	0.128	0.122	0.14
Linear regression	**0.249**	0.119	**0.136**
M5 model decision tree	0.182	0.122	0.138
Neural Net.	0.136	0.139	0.188
Random forest	-0.013	0.13	0.146
Random tree	-0.053	0.134	0.2
Red. Err. Prun. dec. tree	-0.093	0.129	0.141
Simple linear regr.	-0.463	0.137	0.148
SVM	0.008	**0.101**	0.195

Table 9: Results of different regressors for the prediction of the overall anomaly level of individuals. We considered two profiles of individuals: healthy seniors aged 60 or more (119 individuals, anomaly score = 0.0), and elderly with MCI (53 individuals, anomaly score = 0.3).

achieved a correlation coefficient close to the one of SVM. These results indicate that, despite the relatively small number of participants, there is a significant correlation among the predicted anomaly level and the actual diagnosis of individuals. Also in this case, it is reasonable to assume that the correlation should increase by considering longer-term observations of the individuals' behaviors.

In another experiment, we considered only two typologies of individuals: cognitively healthy seniors, and people with MCI. Results are reported in Table 9. As expected, the achieved correlation is lower in this case, since the separation of the classes of healthy seniors and people with MCI is rather low. Indeed, people with MCI were still capable of completing most activities with few anomalies. In this setup, the regressor obtaining the largest correlation coefficient ($r = 0.182$) is the M5' model decision tree. The achieved r value indicates a weak correlation

Regressor	Correl. coeff.	Mean abs. err.	Root mean squared err.
5 Nearest Neigh.	0.288	0.192	0.366
Decision stump	-0.001	0.264	0.341
Decision table	-0.001	0.264	0.341
Linear regression	0.463	0.197	0.275
M5 model decision tree	0.508	0.192	0.265
Neural Net.	0.206	0.379	0.56
Random forest	0.384	0.223	0.288
Random tree	0.155	0.24	0.41
Red. Err. Prun. dec. tree	0.111	0.252	0.327
Simple linear regr.	**0.531**	**0.19**	**0.26**
SVM	-0.193	0.273	0.392

Table 10: Results of different regressors for the prediction of the overall anomaly level of individuals. We considered two profiles of individuals: elderly with MCI (53 individuals, anomaly score = 0.3), and PwD (19 individuals, anomaly score = 1.0).

among the predicted values and the ground truth. Among the other algorithms, the one based on linear regression achieved weak correlation, while the other ones essentially did not provide significant correlation.

Then, we considered two different typologies of individuals: elderly with MCI, and PwD. Results are reported in Table 10. In this case, the achieved correlation is larger, since there is more separation between the classes of people with MCI and PwD. The regressors achieving the highest r score are those based on linear regression and M5' model decision tree ($r = 0.49$). Some regressors, in particular those based on Neural networks and decision table, achieved essentially no correlation. The regressor based on Random forest was the only other algorithm achieving a correlation score larger than 0.4.

Moreover, we did experiments considering other two typologies of individuals: cognitively healthy seniors, and PwD. Results are shown in Table 11. In this case, the achieved correlation level is even larger than in the previous case, since the classes of cognitively healthy seniors and PwD are clearly separated. The M5' model decision tree algorithm obtained the highest correlation ($r = 0.601$). Most other regressors achieved significantly lower correlation values. Indeed, only regressors based on SVM and Random forest achieved a correlation score larger than 0.52. We also note that the regressor based on Neural networks achieved poor results in all our experiments, probably due to the relatively small size of the dataset, and because of the limited number of features.

Finally, we considered only two profiles of individuals: young-old cognitively

Regressor	**Correl. coeff.**	**Mean abs. err.**	**Root mean squared err.**
5 Nearest Neigh.	0.361	**0.141**	0.373
Decision stump	0.358	0.178	0.333
Decision table	0.358	0.178	0.333
Linear regression	0.486	0.2	0.312
M5 model decision tree	**0.581**	0.185	**0.281**
Neural Net.	0.351	0.224	0.435
Random forest	0.505	0.169	0.298
Random tree	0.26	0.188	0.43
Red. Err. Prun. dec. tree	0.408	0.183	0.328
Simple linear regr.	0.577	0.174	**0.281**
SVM	-0.024	0.146	0.382

Table 11: Results of different regressors for the prediction of the overall anomaly level of individuals. We considered two profiles of individuals: healthy seniors aged 60 or more (119 individuals, anomaly score = 0.0), and PwD (19 individuals, anomaly score = 1.0).

Regressor	**Correl. coeff.**	**Mean abs. err.**	**Root mean squared err.**
5 Nearest Neigh.	0.346	0.187	0.429
Decision stump	0.29	0.244	0.394
Decision table	0.29	0.244	0.394
Linear regression	0.579	0.218	0.329
M5 model decision tree	**0.707**	**0.161**	**0.279**
Neural Net.	0.594	0.213	0.354
Random forest	0.543	0.217	0.331
Random tree	0.309	0.217	0.463
Red. Err. Prun. dec. tree	0.546	0.186	0.333
Simple linear regr.	0.623	0.197	0.308
SVM	0.376	0.24	0.38

Table 12: Results of different regressors for the prediction of the overall anomaly level of individuals. We considered two profiles of individuals: healthy seniors aged 60 or more (119 individuals, anomaly score = 0.0), and PwD (19 individuals, anomaly score = 1.0).

healthy seniors aged 60-74, and PwD. Results are shown in Table 12. With this setup, we obtained the highest correlation $r = 0.693$ using the M5' model decision tree. This value of r essentially indicates a strong correlation among the predicted level and the diagnosis of the individual. We believe that the predictions' accuracy significantly improved because of a better model selection. Indeed, cognitively healthy seniors aged 75 or older, which were not considered in this setup, are more likely to suffer from physical problems or normal cognitive age-related decline that impact the way of execution of everyday activities. As a consequence, they may execute behaviors that resemble anomalies, even in the absence of cognitive issues, and this fact may actually confuse the regression algorithm. These results indicate the importance of carefully selecting the set of participants used for collaboratively training the model.

6.7. Dashboard for clinicians

As discussed before, while automatic tools for behavior monitoring and anomaly prediction may provide a useful support, the actual diagnosis must be provided by a clinician based on a detailed multi-dimensional examination of the patient. To this aim, HealthXAI provides a Web dashboard to allow clinicians inspecting anomalies, scores, and their automatically generated natural language explanations. The HealthXAI dashboard can be freely accessed on the Web[7].

A screenshot of the dashboard is shown in Figure 11. The visualization dashboard enables clinicians, who are the final users of HealthXAI, to analyze and visualize at fine-grained level the information stored in the HealthXAI database in a user-friendly way. The dashboard has been designed in order to help clinicians taking decision in shorter timescales. It includes various forms of data presentation (tables, plots, natural language explanations, numerical summaries) to ensure high flexibility for visualization of heterogeneous and sophisticated information. By selecting the current patient through a dropdown list on the

[7] `https://bit.ly/HealthXAI`

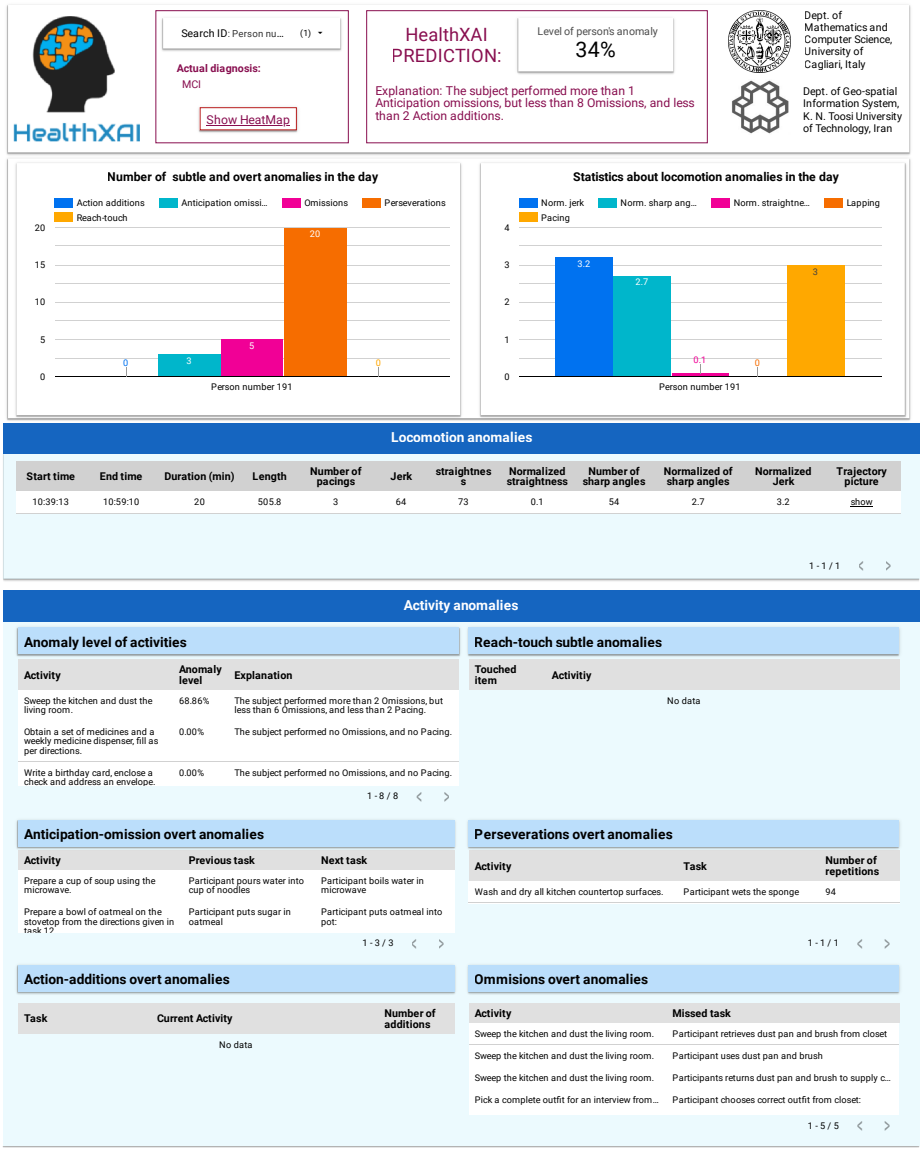

Figure 11: A screenshot of the HealthXAI dashboard.

top left, clinicians can inspect spatial disorientation, wandering behaviors, subtle and overt behavioral anomalies, together with the detailed explanations of predictions.

The dashboard contains three main sections. The first section (on top) reports overall information. In particular, the HealthXAI prediction index indicates the level of person anomaly in percentage, computed as explained in Section 5.1. Below the index, it is reported the natural language explanation for the index prediction. The actual diagnosis of the patient is reported below the dropdown list. Below, a hyperlink show heatmap allows inspecting the heatmap showing the patient's movement pattern in the home in the day, which is similar to the one shown in Figure 8. Two plots below illustrates the number of wandering episodes, subtle and overt anomalies in the day, as well as statistics about locomotion anomalies in the day. These plots allow clinicians having a quick overall vision of the anomalies executed by the patient, related to behaviors and movements, respectively.

The second section reports a table with details about locomotion anomalies. These refer to trajectories having pacing or lapping anomalies. For each anomalous trajectory, the table reports its start-time and end-time, duration, length, number of pacings, number of lappings, straightness and normalized straightness, jerk and number of sharp angles. In the last column, a hyperlink allows graphically inspecting the trajectory, as shown in Figure 3.

The third section shows details about activity anomalies. The first table reports the predicted anomaly level of activities together with the prediction's explanation. The reach-touch subtle anomalies table reports the activity during which the anomaly occurred, and the reached-and-touched item. Table anticipation-omission overt anomalies reports the anomalous activity class, and the tasks that have been executed in abnormal order. Table perseverations overt anomalies reports the activity, the task, and the number of repetitions of the task during the activity. Table action-additions overt anomalies reports the activity, the task unrelated to the activity, and the number of times that it was executed. Finally, Table omissions overt

ANOMALIES reports the activity and the necessary task that was not executed.

6.8. Discussion and limitations

Based on the experimental results, we can observe that, in general, there is a statistically significant correlation among the predictions of HealthXAI and the diagnosis of the cognitive health status of participants. This correlation is particularly evident when considering cognitively healthy subjects and PwD. On the contrary, we obtained a weak correlation when considering only persons with MCI and cognitively healthy seniors. This result is consistent with other studies in this field, such as those reported in [56], in which the separation between MCI and cognitively healthy subjects was hardly recognized using IoT data and artificial intelligence methods. We believe that this problem could be approached by using additional sensors to recognize a larger set of subtle inefficiencies. Indeed, based on the sensors available in our test-bed, we could only monitor one kind of subtle inefficiency. Other kinds could be detected using different sensors attached to everyday objects.

A key aspect that was not considered in this work is the long-term evaluation of cognitive decline based on the history of anomalies. In fact, despite the used dataset having been acquired from a large set of individuals, each individual was monitored only for a few hours. The lack of labeled real-world datasets acquired from large groups of individuals on the long term is a severe limitation for this field of study. We believe that it is reasonable to expect achieving higher correlation based on long-term monitoring, but this aspect should be confirmed by experiments on large trials in real-world conditions. Long-term monitoring should also be supported by specific techniques to adapt and update the learnt models with historic information.

7. Conclusion and future work

In this paper, we tackled the challenging issue of continuous remote monitoring of elderly people for supporting early detection of cognitive decline. Being

based on solid clinical models, and empowered by a collaborative approach, our solution has clear advantages with respect to the state of the art in terms of scalability and personalization. Moreover, being based on explainable AI, our system provides better support to clinicians in making a diagnosis. Large-scale experiments with real-world seniors show the potential of our system.

Several research directions should be investigated in future work. The accuracy of the system could be improved by considering additional kinds of subtle inefficiencies, possibly adopting different sensor infrastructures. The DOLA module may produce false positives due to contextual conditions such as the home shape or presence of obstacles; its algorithms could be refined by assuming some kind of external knowledge about the home. Personalization could be improved by considering individual's habits and physical conditions. Future research should also focus on assessment on Multi-Criteria Decision Making (MCDM) methods for predicting the overall anomaly level of individuals. For the sake of this work, we used fixed parameters set based on common sense knowledge; however, we aim at devising a method to automatically fine-tune parameters to increase accuracy. Techniques for long-term analysis of anomalies should also be investigated to improve the predictions. Moreover, different explainable AI methods should be considered to cope with the machine learning problems involved in this work. Finally, we aim at experimenting our system in fully naturalistic environments for longer periods of time.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments and suggestions for improving the technical content and presentation of this paper. This work was partially funded by the POR FESR Sardegna 2014-2020 project "MISTER: Match Information System and Technologies for the Evaluation of the Performance". The research visit of Elham Khodabandehloo at the University of Cagliari was supported by the Iranian Ministry of Science, Research, and Technology.

References

[1] D. E. Bloom, D. L. Luca, The global demography of aging: facts, explanations, future, in: Handbook of the economics of population aging, Vol. 1, Elsevier, 2016, pp. 3–56.

[2] Y.-H. Wu, Y.-C. Lu, Qualitative research on the importance and need for home-based telecare services for elderly people, Journal of Clinical Gerontology and Geriatrics 5 (4) (2014) 105–110.

[3] S. A. Graham, E. E. Lee, D. V. Jeste, R. Van Patten, E. W. Twamley, C. Nebeker, Y. Yamada, H.-C. Kim, C. A. Depp, Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: A conceptual review, Psychiatry research 284 (2020) 11–32.

[4] J. Yin, Q. Yang, J. J. Pan, Sensor-based abnormal human-activity detection, IEEE Trans. Knowl. Data Eng. 20 (8) (2008) 1082–1090.

[5] D. Riboni, C. Bettini, G. Civitarese, Z. H. Janjua, R. Helaoui, Smartfaber: Recognizing fine-grained abnormal behaviors for early detection of mild cognitive impairment, Artificial Intelligence in Medicine 67 (2016) 57–74.

[6] H. Sfar, A. Bouzeghoub, B. Raddaoui, Early anomaly detection in smart home: A causal association rule-based approach, Artif. Intell. Medicine 91 (2018) 57–71.

[7] Z. H. Janjua, D. Riboni, C. Bettini, Towards automatic induction of abnormal behavioral patterns for recognizing mild cognitive impairment, in: SAC, ACM, 2016, pp. 143–148.

[8] M. F. Schwartz, M. Segal, T. Veramonti, M. Ferraro, L. J. Buxbaum, The naturalistic action test: A standardised assessment for everyday action impairment, Neuropsychol Rehabil 12 (4) (2002) 311–339.

[9] S. C. Seligman, T. Giovannetti, J. Sestito, D. J. Libon, A new approach to the characterization of subtle errors in everyday action: Implications for mild cognitive impairment, Clin Neuropsychol 28 (2014) 97–115.

[10] D. Martino-Saltzman, B. B. Blasch, R. D. Morris, L. W. McNeal, Travel behavior of nursing home residents perceived as wanderers and nonwanderers, The Gerontologist 31 (5) (1991) 666–672.

[11] X. Li, Using complexity measures of movement for automatically detecting movement types of unknown gps trajectories, Am. J. Geogr. Inf. Syst 3 (2) (2014) 63–74.

[12] S. Dabiri, K. Heaslip, Inferring transportation modes from gps trajectories using a convolutional neural network, Transportation research part C: emerging technologies 86 (2018) 360–371.

[13] Q. Lin, D. Zhang, X. Huang, H. Ni, X. Zhou, Detecting wandering behavior based

on gps traces for elders with dementia, in: 2012 12th International Conference on Control Automation Robotics, IEEE, 2012, pp. 672–677.

[14] B. Winblad et Al., Mild cognitive impairment - beyond controversies, towards a consensus, J Intern Med 256 (3) (2004) 240–246.

[15] A. I. Sook, K. Ji-Hae, K. Seonwoo, C. W. Jae, K. Hyeran, S. H. S. Kang, K. D. Kwan, Impairment of instrumental activities of daily living in patients with mild cognitive impairment, Psychiatry Investig 6 (3) (2009) 180–184.

[16] T. Giovannetti, B. M. Bettcher, L. Brennan, D. J. Libon, M. Burke, K. Duey, C. Nieves, D. Wambach, Characterization of everyday functioning in mild cognitive impairment: A direct assessment approach, Dement Geriatr Cogn Disord 25 (2008) 359–365.

[17] E. Cornelis, E. Gorus, I. Beyer, I. Bautmans, P. De Vriendt, Early diagnosis of mild cognitive impairment and mild dementia through basic and instrumental activities of daily living: Development of a new evaluation tool, PLoS medicine 14 (3) (2017) 1–22.

[18] D. L. Algase, Wandering: a dementia-compromised behavior, Journal of Gerontological Nursing 25 (9) (1999) 10–16.

[19] D. Martino-Saltzman, B. B. Blasch, R. D. Morris, L. W. McNeal, Travel behavior of nursing home residents perceived as wanderers and nonwanderers, The Gerontologist 31 (5) (1991) 666–672.

[20] P. L. Sheridan, J. Solomont, N. Kowall, J. M. Hausdorff, Influence of executive function on locomotor function: divided attention increases gait variability in alzheimer's disease, Journal of the American Geriatrics Society 51 (11) (2003) 1633–1637.

[21] M. Chen, Y. Jiang, Y. Cao, A. Y. Zomaya, Creativebioman: A brain-and body-wearable, computing-based, creative gaming system, IEEE Systems, Man, and Cybernetics Magazine 6 (1) (2020) 14–22.

[22] F. Gerina, S. M. Massa, F. Moi, D. R. Recupero, D. Riboni, Recognition of cooking activities through air quality sensor data for supporting food journaling, Human-centric Computing and Information Sciences 10 (1) (2020) 1–26.

[23] M. Chen, Y. Cao, R. Wang, Y. Li, D. Wu, Z. Liu, Deepfocus: Deep encoding brainwaves and emotions with multi-scenario behavior analytics for human attention enhancement, IEEE Network 33 (6) (2019) 70–77.

[24] M. Chen, Y. Hao, Label-less learning for emotion cognition, IEEE Trans Neural Netw Learn Syst , Online ahead of print (2019) 1–11.

[25] H. S. M. Bilal, M. B. Amin, J. Hussain, S. I. Ali, S. Hussain, M. Sadiq, M. A. Razzaq, A. Abbas, C. Choi, S. Lee, On computing critical factors based healthy behavior index for behavior assessment, International Journal of Medical Infor-

matics (2020) 1–13.

[26] K.-Y. Lam, N. W.-H. Tsang, S. Han, W. Zhang, J. K.-Y. Ng, A. Nath, Activity tracking and monitoring of patients with alzheimer's disease, Multimedia Tools and Applications 76 (1) (2017) 489–521.

[27] K. Ota, Y. Ota, M. Otsu, A. Kajiwara, Elderly-care motion sensor using uwb-ir, in: 2011 IEEE Sensors Applications Symposium, IEEE, 2011, pp. 159–162.

[28] H. Dodge, N. Mattek, D. Austin, T. Hayes, J. Kaye, In-home walking speeds and variability trajectories associated with mild cognitive impairment, Neurology 78 (24) (2012) 1946–1952.

[29] H. Ishii, K. Kimino, M. Aljehani, N. Ohe, M. Inoue, An early detection system for dementia using the m2 m/iot platform, Procedia Computer Science 96 (2016) 1332–1340.

[30] R. Suzuki, S. Otake, T. Izutsu, M. Yoshida, T. Iwaya, Monitoring daily living activities of elderly people in a nursing home using an infrared motion-detection system, Telemed J E Health 12 (2) (2006) 146–155.

[31] J. Helmy, A. Helmy, The alzimio app for dementia, autism & alzheimer's: Using novel activity recognition algorithms and geofencing, in: 2016 IEEE International Conference on Smart Computing (SMARTCOMP), IEEE, 2016, pp. 1–6.

[32] Q. Lin, D. Zhang, L. Chen, H. Ni, X. Zhou, Managing elders' wandering behavior using sensors-based solutions: a survey, International Journal of Gerontology 8 (2) (2014) 49–55.

[33] Q. Lin, D. Zhang, X. Huang, H. Ni, X. Zhou, Detecting wandering behavior based on GPS traces for elders with dementia, in: 12th International Conference on Control Automation Robotics & Vision, IEEE, 2012, pp. 672–677.

[34] J. Ng, H. Kong, Not all who wander are lost: Smart tracker for people with dementia, in: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 2016, pp. 2241–2248.

[35] S. Schaat, P. Koldrack, K. Yordanova, T. Kirste, S. Teipel, Real-time detection of spatial disorientation in persons with mild cognitive impairment and dementia, Gerontology 66 (1) (2020) 85–94.

[36] A. Kumar, C. T. Lau, M. Ma, S. Chan, W. Kearns, Trend analysis in the trajectory of the dementia patients, in: 2017 21st International Computer Science and Engineering Conference (ICSEC), IEEE, 2017, pp. 1–5.

[37] N. Vuong, S. Chan, C. T. Lau, K. Lau, Feasibility study of a real-time wandering detection algorithm for dementia patients, in: Proceedings of the First ACM MobiHoc Workshop on Pervasive Wireless Healthcare, 2011, pp. 1–4.

[38] W. D. Kearns, J. L. Fozard, V. O. Nams, J. D. Craighead, Wireless telesurveil-

lance system for detecting dementia, Gerontechnology (2011) 90.

[39] Q. Sun, F. Hu, Q. Hao, Human movement modeling and activity perception based on fiber-optic sensing system, IEEE Trans. Human-Machine Systems 44 (6) (2014) 743–754.

[40] Q. Lin, W. Zhao, W. Wang, Detecting dementia-related wandering locomotion of elders by leveraging active infrared sensors, Journal of Computer and Communications 6 (05) (2018) 94.

[41] A. Khan, A. Z. Hassan, Framework to predict and identify wandering behavior in individuals with alzheimer's using physiological and kinect sensors, in: Proceedings of SAI Intelligent Systems Conference, Springer, 2018, pp. 387–398.

[42] E. Khodabandehloo, D. Riboni, Collaborative trajectory mining in smart-homes to support early diagnosis of cognitive decline, IEEE Transactions on Emerging Topics in Computing (Early access) (2020). doi:10.1109/TETC.2020.2975071.

[43] F. Wang, E. E. Stone, M. Skubic, J. M. Keller, C. Abbott, M. Rantz, Toward a passive low-cost in-home gait assessment system for older adults, IEEE J. Biomedical and Health Informatics 17 (2) (2013) 346–355.

[44] J. Wang, Y. Chen, S. Hao, X. Peng, L. Hu, Deep learning for sensor-based activity recognition: A survey, Pattern Recognition Letters 119 (2019) 3–11.

[45] A. Keshavarzian, S. Sharifian, S. Seyedin, Modified deep residual network architecture deployed on serverless framework of iot platform based on human activity recognition application, Future Gener. Comput. Syst. 101 (2019) 14–28.

[46] M. Rawashdeh, M. G. H. al Zamil, S. Samarah, M. S. Hossain, G. Muhammad, A knowledge-driven approach for activity recognition in smart homes based on activity profiling, Future Gener. Comput. Syst. 107 (2020) 924–941.

[47] G. Civitarese, C. Bettini, T. Sztyler, D. Riboni, H. Stuckenschmidt, *newNECTAR*: Collaborative active learning for knowledge-based probabilistic activity recognition, Pervasive and Mobile Computing 56 (2019) 88–105.

[48] D. Riboni, M. Murtas, Sensor-based activity recognition: One picture is worth a thousand words, Future Gener. Comput. Syst. 101 (2019) 709–722.

[49] A. Matassa, D. Riboni, Reasoning with smart objects' affordance for personalized behavior monitoring in pervasive information systems, Knowl. Inf. Syst. 62 (4) (2020) 1255–1278.

[50] D. H. Douglas, T. K. Peucker, Algorithms for the reduction of the number of points required to represent a digitized line or its caricature, Cartographica: the international journal for geographic information and geovisualization 10 (2) (1973) 112–122.

[51] D. A. Freedman, Statistical models: theory and practice, Cambridge University

Press, 2009.

[52] D. Steinberg, Cart: classification and regression trees, in: The top ten algorithms in data mining, Chapman and Hall/CRC, 2009, pp. 193–216.

[53] H. C. Koh, G. Tan, et al., Data mining applications in healthcare, Journal of healthcare information management 19 (2) (2011) 65.

[54] X. Zhang, B. Hu, X. Ma, P. Moore, J. Chen, Ontology driven decision support for the diagnosis of mild cognitive impairment, Computer Methods and Programs in Biomedicine 113 (3) (2014) 781–791.

[55] K. Jekel, M. Damian, C. Wattmo, L. Hausner, R. Bullock, P. J. Connelly, B. Dubois, M. Eriksdotter, M. Ewers, E. Graessel, et al., Mild cognitive impairment and deficits in instrumental activities of daily living: a systematic review, Alzheimer's research & therapy 7 (1) (2015) 17.

[56] A. M. Seelye, M. Schmitter-Edgecombe, D. J. Cook, A. Crandall, Naturalistic assessment of everyday activities and prompting technologies in mild cognitive impairment, J Int Neuropsychol Soc 19 (4) (2013) 442–452.

[57] D. J. Cook, A. S. Crandall, B. L. Thomas, N. C. Krishnan, CASAS: A smart home in a box, Computer 46 (7) (2013) 62–69.

[58] American Psychiatric Association, Diagnostic and statistical manual of mental disorders, American Psychiatric Pub, 2013.

[59] R. C. Petersen, J. C. Morris, Mild cognitive impairment as a clinical entity and treatment target, Archives of neurology 62 (7) (2005) 1160–1163.

[60] M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen, et al., The diagnosis of mild cognitive impairment due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease, Alzheimer's & dementia 7 (3) (2011) 270–279.

[61] P. W. Burgess, Strategy application disorder: the role of the frontal lobes in human multitasking, Psychological research 63 (3-4) (2000) 279–288.

[62] P. Rashidi, D. J. Cook, L. B. Holder, M. Schmitter-Edgecombe, Discovering activities to recognize and track in a smart environment, IEEE Trans. Knowl. Data Eng. 23 (4) (2011) 527–539.

[63] E. Frank, M. A. Hall, G. Holmes, R. Kirkby, B. Pfahringer, WEKA - A machine learning workbench for data mining, in: The Data Mining and Knowledge Discovery Handbook, Springer, 2005, pp. 1305–1314.

[64] J. Lee Rodgers, W. A. Nicewander, Thirteen ways to look at the correlation coefficient, The American Statistician 42 (1) (1988) 59–66.

[65] J. R. Quinlan, et al., Learning with continuous classes, in: 5th Australian joint

conference on artificial intelligence, Vol. 92, World Scientific, 1992, pp. 343–348.

[66] R. Kohavi, The power of decision tables, in: ECML, Vol. 912 of Lecture Notes in Computer Science, Springer, 1995, pp. 174–189.

[67] S. le Cessie, J. van Houwelingen, Ridge Estimators in Logistic Regression, Applied Statistics 41 (1) (1992) 191–201.

[68] S. Salzberg, Book review: C4.5: programs for machine learning, Machine Learning 16 (3) (1994) 235–240.